TECHNIQUES AND DEVICES
FOR
HIGH-RESOLUTION ADAPTIVE OPTICS
by
ARTHUR DOUGLAS FISHER
M.S., EECS, Massachusetts Institute of Technology (1976)
Submitted in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY
at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY
August 1981
© Massachusetts Institute of Technology 1981
Signature of Author Department of Electrical Engineering and Computer Science August 14, 1981
Certified by.................. Cardinal Warde Thesis supervisor
Accepted by.................. Arthur C. Smith Chairman, Department Committee on Theses
NOV 13 1981
TECHNIQUES AND DEVICES
FOR
HIGH-RESOLUTION ADAPTIVE OPTICS
by
ARTHUR DOUGLAS FISHER
Submitted to the Department of Electrical Engineering
and Computer Science on August 14, 1981
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy

ABSTRACT

Techniques and devices have been developed for the construction of
high-resolution adaptive optical systems, containing thousands to
millions of spatial resolution elements. These systems potentially offer
the resolution, as well as sensitivity and speed, required for real-time
wavefront correction in such applications as optical communications
through low-visibility or turbulent atmospheric conditions, imaging through
the atmosphere, and high energy laser propagation.

Unlike the previous technology, consisting of arrays of individual
phase estimators hard-wired to modulator arrays, an "all-optical" parallel
processing approach has been pursued. In operation, wavefront spatial
phase variations, e.g. due to atmospheric propagation, are transformed by
an all-optical phase estimator into intensity variations in a control beam.
This control beam drives a monolithic optically-addressed spatial phase
modulator to compensate the corrupted wavefront. This system employs a
new closed-loop phase measurement and compensation technique, called an
Interference Phase Loop (IPL), and a novel optically-addressed phase modu-
lator, called a Microchannel Spatial Phase Modulator (MSLM).

In the IPL technique, the output intensity from an elementary inter-
ferometric phase sensor (e.g. Zernike phase-contrast, homodyne, or hetero-
dyne interferometer) is detected and employed to drive a phase-only
modulator in the path of the wavefront being measured. It is shown
theoretically and experimentally that this configuration can continuously
estimate and compensate phase with no phase-quadrant ambiguity over multi-
ple-π radians of dynamic range, even in the presence of wavefront amplitude
variations. The basic feasibility and operational theory of the IPL were
demonstrated in a laboratory test system containing a single resolution
element. Wavefront shaping and phase compensation were demonstrated with
a nineteen element discrete-channel test system.

The Microchannel Spatial Light Modulator was developed from a non-
functional prototype to a usable modulator with a half-wave exposure sen-
sitivity of 2.2nJ/cm² (@635nm), a framing rate of a few Hz, and a spatial
resolution of 2 lp/mm. Quantum-limited sensitivities, better than kilo-
hertz framing rates, and spatial resolutions approaching 20lp/mm are ulti-
mately expected. An MSLM was successfully employed to demonstrate basic
phase compensation with an all-optical MSLM/IPL adaptive system.

Thesis Supervisor: Cardinal Warde
Title: Associate Professor of Electrical Engineering and Computer Science
ACKNOWLEDGEMENTS

It is with sincere gratitude that I thank my supervisor, Professor Cardinal Warde, whose valuable guidance and support made this thesis possible. I would also like to thank my thesis readers, Professor David J. Epstein and Professor Jeffrey H. Shapiro, for taking time to carefully read this manuscript and making many valuable suggestions to enhance it.

Thanks are also due to an army of graduate and undergraduate students who polished electrooptic crystals and provided other materials fabrication assistance for the experimental portions of this work; these persons include: John Thackara, Bob Dillon, Aryeh Weiss, Mohammed Burmawi, Dennis Cocco, Mark Tannen, Mark Smith, Barbara Messinger, Lisa Bahler, and Curtis McMillan. Aryeh Weiss in particular deserves credit for succeeding where others had failed in finding an outside company capable of polishing electrooptic crystals to our exacting standards. Bob DeCesaris made valuable contributions in the study and fabrication of photoconductors. Special thanks are also due to John Thackara for his valuable assistance in many experiments.

I am also indebted to John McCormick, Teruo Hiruma, and many other very capable people at Hamamatsu Corporation, whose great enthusiasm and technical skill enabled the success of our collaborative efforts to produce a microchannel spatial light modulator with a visible photocathode. Bob Dillon also deserves thanks for the
impressive perseverance and skill he devoted to his attempts to create visible photocathodes at MIT.

In addition, I would like to thank Linda Sayegn for her excellent typing of most of the first draft of this thesis and Yoshiko Ryu for her aid in drafting many of the figures.

Finally, I am very grateful for the financial support which made this work possible, including: the Vinton Hayes Fund of MIT, National Science Foundation contract ENG-78-21603, and Air Force Office of Scientific Research Contract AFOSR-77-3328. Much of this work made use of the facilities of MIT's Center for Material Science and Engineering, which is supported by NSF grant DMR-78-24185.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>3</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>16</td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>17</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>17</td>
</tr>
<tr>
<td>1.2 Phase estimation</td>
<td>21</td>
</tr>
<tr>
<td>1.3 Phase modulators</td>
<td>25</td>
</tr>
<tr>
<td>1.4 Applications for high-resolution adaptive optics</td>
<td>28</td>
</tr>
<tr>
<td>1.5 Thesis organization</td>
<td>42</td>
</tr>
<tr>
<td>II. PHASE ESTIMATION</td>
<td>45</td>
</tr>
<tr>
<td>2.1 Overview of phase measurement techniques</td>
<td>45</td>
</tr>
<tr>
<td>2.1.1 Elementary interferometric phase measurement</td>
<td>46</td>
</tr>
<tr>
<td>2.1.2 Improved phase measurement systems</td>
<td>48</td>
</tr>
<tr>
<td>2.1.3 Real-time high-resolution phase measurement</td>
<td>55</td>
</tr>
<tr>
<td>2.2 The Interference Phase Loop (IPL)</td>
<td>66</td>
</tr>
<tr>
<td>2.2.1 General formulation for modelling IPL behavior</td>
<td>67</td>
</tr>
<tr>
<td>2.2.2 Basic static behavior of the IPL</td>
<td>74</td>
</tr>
<tr>
<td>2.2.3 Basic dynamic behavior of the IPL</td>
<td>87</td>
</tr>
<tr>
<td>III. ADVANCED OPERATIONAL THEORY OF THE INTERFERENCE PHASE LOOP</td>
<td>99</td>
</tr>
<tr>
<td>3.1 A more detailed look at IPL performance</td>
<td>99</td>
</tr>
<tr>
<td>3.1.1 Insensitivity to signal amplitude fluctuations</td>
<td>99</td>
</tr>
<tr>
<td>3.1.2 Initial phase acquisition</td>
<td>108</td>
</tr>
<tr>
<td>3.1.3 Multistability and 2m degeneracy</td>
<td>112</td>
</tr>
<tr>
<td>3.1.4 Phase compensation performance</td>
<td>114</td>
</tr>
<tr>
<td>3.1.5 Spatial resolution</td>
<td>113</td>
</tr>
<tr>
<td>3.1.6 Other loop filters</td>
<td>125</td>
</tr>
<tr>
<td>3.1.6.1 Integrating type-one loop filter</td>
<td>126</td>
</tr>
<tr>
<td>3.1.6.2 Two-pole loop filters</td>
<td>138</td>
</tr>
<tr>
<td>3.1.6.3 Summary</td>
<td>144</td>
</tr>
</tbody>
</table>
3.1.7 IPL performance in the presence of noise 148
3.1.8 IPL performance with extraneous coherent illumination 165
3.1.9 Irradiance required by the IPL 167
3.2 Alternative interferometric implementations 175
3.2.1 The heterodyne IPL 175
  3.2.1.1 The heterodyne IPL with an integrating loop filter 176
  3.2.1.2 The heterodyne IPL with a low-pass loop filter 183
  3.2.1.3 Conclusions 192
3.2.2 The phase-contrast IPL 198
  3.2.2.1 General formulation 199
  3.2.2.2 Gain degradation due to spatial phase fluctuations 203
  3.2.2.3 Behavior of a single spatial element 204
  3.2.2.4 Global behavior 206
  3.2.2.5 Simulated behavior 210
  3.2.2.6 Simplifying approximations 212
  3.2.2.7 Behavior with a low-pass loop filter 215
  3.2.2.8 Behavior with other loop filters 219
  3.2.2.9 Stability 222
  3.2.2.10 Spectral bandwidth 224
  3.2.2.11 Summary of phase-contrast IPL behavior 239

3.3 Summary 240

IV MONOLITHIC SPATIAL LIGHT MODULATORS 246
  4.1 Available modulators 246
  4.2 The Microchannel Spatial Light Modulator (MSLM) 252
    4.2.1 Principles of operation 253
      4.2.1.1 Circuit models 253
      4.2.1.2 Continuous mode 261
      4.2.1.3 Framed modes 262
    4.2.2 Experimental MSLM systems 274
    4.2.3 MSLM performance: theory, measurement, enhancement 276
      4.2.3.1 Sensitivity 276
      4.2.3.2 Time response 282
4.2.3.3 Phase modulation dynamic range
4.2.3.4 Resolution and optical quality of images
4.2.4 MSLM summary
4.3 MSLM-based high-resolution adaptive systems
  4.3.1 Continuous mode MSLM-based IPL
  4.3.2 Electron-deposition framed mode MSLM-based IPL
  4.3.3 Inverted-secondary framed mode MSLM-based IPL
  4.3.4 Heterodyne MSLM-based IPL
  4.3.5 MSLM-based IPL summary
4.4 Conclusions

V EXPERIMENTAL DISCRETE-CHANNEL OPTICALLY-ADDRESSED PHASE MODULATORS
  5.1 Single-channel system
  5.2 Nineteen-channel system
  5.3 Twenty-channel large dynamic range system
  5.4 A proposed medium-resolution discrete-channel system

VI EXPERIMENTAL ADAPTIVE OPTICAL SYSTEMS
  6.1 Single-channel system
    6.1.1 Aberrators for performance evaluation
    6.1.2 Experimental performance
  6.2 Nineteen-channel system
    6.2.1 Techniques for performance evaluation
    6.2.2 Experimental performance
  6.3 "All-optical" phase compensation with the MSLM
    6.3.1 Experimental performance of timed-integration MSLM-based IPL
    6.3.2 Experimental performance of inverted-secondary MSLM-based IPL
    6.3.3 Attempted high-resolution MSLM-based IPL

VII SUMMARY AND SUGGESTIONS FOR FUTURE WORK
APPENDICES

<table>
<thead>
<tr>
<th>Appendices</th>
<th>Details</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>INTERFERENCE PHASE LOOP DETAILS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A.1 Average value of a class of stochastic phasors</td>
<td>398</td>
</tr>
<tr>
<td></td>
<td>A.2 Single lens optical convolution</td>
<td>398</td>
</tr>
<tr>
<td></td>
<td>A.3 The Zernike phase-contrast filter</td>
<td>401</td>
</tr>
<tr>
<td></td>
<td>A.4 The discrete linearized Zernike phase-contrast IPL</td>
<td>404</td>
</tr>
<tr>
<td></td>
<td>A.5 Stability of the linearized Zernike phase-contrast IPL</td>
<td>411</td>
</tr>
<tr>
<td>B</td>
<td>MICROCHANNEL SPATIAL LIGHT MODULATOR DETAILS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B.1 MSLM circuit modelling details</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td>B.2 Secondary emission in the MSLM</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td>B.3 Parallel crystal Fabry-Perot interferometer</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>B.4 Special functions achievable with the MSLM</td>
<td>456</td>
</tr>
<tr>
<td></td>
<td>B.5 MSLM sensitometry</td>
<td>462</td>
</tr>
<tr>
<td></td>
<td>B.6 Proximity focussing in the MSLM</td>
<td>473</td>
</tr>
<tr>
<td></td>
<td>B.7 Experimental development of the MSLM</td>
<td>477</td>
</tr>
<tr>
<td>C</td>
<td>DESIGN AND OPERATIONAL DETAILS OF THE DISCRETE-CHANNEL SYSTEMS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C.1 Single-channel system</td>
<td>497</td>
</tr>
<tr>
<td></td>
<td>C.2 Nineteen-channel system</td>
<td>497</td>
</tr>
<tr>
<td>D</td>
<td>BEHAVIOR OF THE STREHL RATIO IN ADAPTIVE OPTICAL SYSTEMS</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>THE MULTIPLE FORWARD SCATTER MODEL FOR LOW-VISIBILITY OPTICAL COMMUNICATIONS</td>
<td>524</td>
</tr>
<tr>
<td>F</td>
<td>PHASE COMPENSATION BY REAL-TIME HOLOGRAPHY AND NONLINEAR OPTICAL MIXING</td>
<td>531</td>
</tr>
<tr>
<td>G</td>
<td>BIBLIOGRAPHIC COMMENTS ON PHASE MEASUREMENT AND RELATED TOPICS</td>
<td>537</td>
</tr>
<tr>
<td>REFERENCES</td>
<td></td>
<td>541</td>
</tr>
<tr>
<td></td>
<td></td>
<td>554</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Conceptual diagram of an all-optical high-resolution adaptive optical system.</td>
<td>20</td>
</tr>
<tr>
<td>1.2</td>
<td>The Microchannel Spatial Light Modulator (MSLM).</td>
<td>27</td>
</tr>
<tr>
<td>2.1</td>
<td>Some techniques for high-resolution phase estimation.</td>
<td>57</td>
</tr>
<tr>
<td>2.2</td>
<td>An &quot;all-optical&quot;, high-resolution Interference Phase Loop.</td>
<td>64</td>
</tr>
<tr>
<td>2.3</td>
<td>General homodyne-heterodyne implementation of the IPL.</td>
<td>69</td>
</tr>
<tr>
<td>2.4</td>
<td>Graphical solution for the steady-state of the type-zero IPL.</td>
<td>77</td>
</tr>
<tr>
<td>2.5</td>
<td>Phase estimation performance ($\phi_m$ versus $\phi_t$).</td>
<td>78</td>
</tr>
<tr>
<td>2.6</td>
<td>Phase estimation error as a function of input phase ($\phi_e$ versus $\phi_t$).</td>
<td>79</td>
</tr>
<tr>
<td>2.7</td>
<td>Modulator phase state-space ($\dot{\phi}_m$ versus $\phi_m$).</td>
<td>89</td>
</tr>
<tr>
<td>2.8</td>
<td>Error phase state-space ($\dot{\phi}_e$ versus $\phi_e$).</td>
<td>91</td>
</tr>
<tr>
<td>3.1</td>
<td>Gain $G_1$, offset $G_0$, and $G_0 \geq G_1$ as a function of relative input amplitude $</td>
<td>E_1</td>
</tr>
<tr>
<td>3.2</td>
<td>Dependence of error phase $\phi_e$ on relative input amplitude $</td>
<td>E_1</td>
</tr>
<tr>
<td>3.3</td>
<td>$\phi_m$ after initial phase acquisition as a function of $\phi_t$.</td>
<td>110</td>
</tr>
<tr>
<td>3.4</td>
<td>Graphical construction for reacquisition phase after reaching modulator limits.</td>
<td>110</td>
</tr>
<tr>
<td>3.5</td>
<td>Simulated spatial wavefront compensation.</td>
<td>117</td>
</tr>
<tr>
<td>3.6</td>
<td>Basic behavior of the IPL with an integrating loop filter.</td>
<td>127</td>
</tr>
<tr>
<td>3.7</td>
<td>State-space characteristic with &quot;dead regions&quot;.</td>
<td>131</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td>Phase estimation performance ($\phi_m$ versus $\phi_i$) with &quot;dead regions&quot;.</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>Modulator state-space with a two-pole, type-zero loop filter.</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td>Modulator state-space with a two-pole, type-one loop filter.</td>
<td></td>
</tr>
<tr>
<td>3.11</td>
<td>State-space behavior of the heterodyne IPL with an integrating loop filter.</td>
<td></td>
</tr>
<tr>
<td>3.12</td>
<td>State-space behavior of the integrating heterodyne IPL with an out of range frequency.</td>
<td></td>
</tr>
<tr>
<td>3.13</td>
<td>The Interference Phase Loop as a phase-locked loop.</td>
<td></td>
</tr>
<tr>
<td>3.14</td>
<td>State-space behavior of the low-pass heterodyne IPL with $\Delta w=0.5\pi\lambda_1$.</td>
<td></td>
</tr>
<tr>
<td>3.15</td>
<td>State-space behavior of the low-pass heterodyne IPL with $\Delta w=-0.5\pi\lambda_1$.</td>
<td></td>
</tr>
<tr>
<td>3.16</td>
<td>State-space for the low-pass heterodyne IPL with $\Delta w=\lambda_1 G_1$.</td>
<td></td>
</tr>
<tr>
<td>3.17</td>
<td>Relative phase error as a function of time.</td>
<td></td>
</tr>
<tr>
<td>3.18</td>
<td>The phase-contrast IPL.</td>
<td></td>
</tr>
<tr>
<td>3.19</td>
<td>Simulated behavior of the phase-contrast IPL.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Four-pole circuit model.</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Simplified two-pole circuit model.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>MCP current density, J, as a function of gap voltage, $V_g$.</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Transient response of the crystal and gap voltages to a step in $V_b$ with $R_L&lt;&lt;R_x$.</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Transient response of the crystal and gap voltage to steps in $V_b$ and $I$ with $R_L&gt;R_x$.</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>The photoconductive-erasure framed mode.</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>The secondary-erasure framed mode.</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>J versus $V_g$ during secondary emission charge removal.</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>The inverted-secondary framed mode.</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.10 Inverted-secondary crystal voltage, $V_x$, as a function of irradiance, $I$.

4.11 Demountable MSLM.

4.12 Sealed visible-photocathode MSLM.

4.13 Imaging with the visible-photocathode MSLM.

4.14 Modulator state-space ($\phi_m$ versus $\phi_m$) for the continuous mode IPL without threshold.

4.15 Phase error, $\phi_e$, versus modulator phase, $\phi_m$, for an IPL without threshold.

4.16 State-space characteristics for the electron-deposition MSLM-based IPL.

4.17 Phase error, $\phi_e$, as a function of time for the electron-deposition MSLM-based IPL.

4.18 Fraction, $f(\Delta\phi_e)$, of initial phase errors compensated to within $\pm\Delta\phi_e$ of $\phi_{eu}$ in the "timed-integration" MSLM-based IPL.

4.19 State-space characteristic for the inverted-secondary framed mode MSLM-based IPL.

4.20 Active phase error, $\phi_{e1}$, as a function of signal amplitude, $|E_i|$, in the inverted-secondary MSLM-based IPL.

5.1 Single-channel, optically-addressed phase modulator.

5.2 Typical modulator state-space ($\phi_m$ versus $\phi_m$) of the single-channel IPL.

5.3 Nineteen-element spatial phase modulator array.

5.4 Nineteen-element detector array.

5.5 Thirty-element spatial phase modulator array.

5.6 Alignment of glass slabs.

5.7 Interferograms of thirty-element modulator array.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>Simple discrete-channel system.</td>
<td>340</td>
</tr>
<tr>
<td>6.1</td>
<td>Single-channel homodyne interference phase loop.</td>
<td>344</td>
</tr>
<tr>
<td>6.2</td>
<td>Basic phase estimation and compensation behavior of the IPL.</td>
<td>347</td>
</tr>
<tr>
<td>6.3</td>
<td>Behavior of the open-loop interferometer.</td>
<td>347</td>
</tr>
<tr>
<td>6.4</td>
<td>IPL performance with air turbulence from a hot-plate.</td>
<td>350</td>
</tr>
<tr>
<td>6.5</td>
<td>IPL performance with wideband phase fluctuations.</td>
<td>351</td>
</tr>
<tr>
<td>6.6</td>
<td>IPL performance with large phase variations.</td>
<td>353</td>
</tr>
<tr>
<td>6.7</td>
<td>IPL immunity to signal amplitude fluctuations.</td>
<td>355</td>
</tr>
<tr>
<td>6.8</td>
<td>Experimental IPL state-space trajectories.</td>
<td>356</td>
</tr>
<tr>
<td>6.9</td>
<td>Modulator state-space ($\phi_m$ versus $\phi_m$) with inverted modulation.</td>
<td>356</td>
</tr>
<tr>
<td>6.10</td>
<td>Nineteen-element laboratory IPL (version I-1).</td>
<td>359</td>
</tr>
<tr>
<td>6.11</td>
<td>Interferogram from nineteen-element IPL (I-1).</td>
<td>362</td>
</tr>
<tr>
<td>6.12</td>
<td>Improved nineteen-element laboratory IPL (version I-2).</td>
<td>362</td>
</tr>
<tr>
<td>6.13</td>
<td>Fourier transform of the compensated beam.</td>
<td>364</td>
</tr>
<tr>
<td>6.14</td>
<td>Manual phase compensation by image sharpening.</td>
<td>370</td>
</tr>
<tr>
<td>6.15</td>
<td>IPL phase compensation of mostly static wavefront aberration.</td>
<td>371</td>
</tr>
<tr>
<td>6.16</td>
<td>Photographs of Fourier transforms of aberrated and IPL compensated waves.</td>
<td>372</td>
</tr>
<tr>
<td>6.17</td>
<td>IPL phase compensation of dynamic wavefront distortions.</td>
<td>373</td>
</tr>
<tr>
<td>6.18</td>
<td>Far field image (Fourier hologram) for image compensation experiments.</td>
<td>376</td>
</tr>
<tr>
<td>6.19</td>
<td>All-optical IPL employing the MSLM.</td>
<td>378</td>
</tr>
<tr>
<td>6.20</td>
<td>Initial-error state-space for the timed-integration framed IPL with a Fabry-Perot interferometer.</td>
<td>380</td>
</tr>
</tbody>
</table>
Figure 6.21 Phase compensation results with the timed-integration framed IPL.

6.22 Error state-space for the inverted-secondary framed IPL.

6.23 Phase compensation results with the inverted-secondary framed IPL.

6.24 Interferograms of wavefront reflected from MCP in visible-phocathode MSLM.

A.1 Dual lens optical convolution.

A.2 Single lens optical convolution.

A.3 Zernike phase-contrast system.

B.1 Secondary emission coefficient, \( \delta' \), as a function of primary electron energy, \( q\bar{V}_e \).

B.2 MSLM secondary emission geometry.

B.3 Gap current, \( i_g \), as a function of gap voltage, \( V_g \), with mono-energetic electrons.

B.4 Special cases of \( i_g \) versus \( V_g \).

B.5 Idealized MCP electron energy distribution.

B.6 Energy distribution of secondary electrons.

B.7 Comparison of theoretical and experimental \( i_g \) versus \( V_g \) secondary emission characteristics.

B.8 Dependence of the high-energy secondary emission coefficient, \( \delta_H \), on the average energy, \( q\bar{V}_a \), of high-energy MCP electrons.

B.9 Estimation of MSLM secondary emission parameters from its \( i_g \) versus \( V_g \) characteristic.

B.10 MSLM secondary emission geometry with a grid.

B.11 Theoretical \( i_g \) versus \( V_g \) characteristic with a grid.
Figure | Page
---|---
B.12 Geometry and definition of terms for parallel-crystal Fabry-Perot. | 457
B.13 Output intensity modulation, $I_{\text{out}}$, of a parallel-crystal Fabry-Perot. | 459
B.14 Output phase modulation, $\phi_f$, of a parallel-crystal Fabry-Perot. | 459
B.15 Intensity produced by a parallel-crystal Fabry-Perot in an external interferometer. | 461
B.16 Transfer function for stored image thresholding. | 464
B.17 Possible example of MSLM optical digital logic. | 469
B.18 Contrast reversal. | 472
B.19 Image addition. | 472
B.20 Image subtraction. | 472
B.21 Analogue and digital thresholding. | 472
B.22 Contrast enhancement. | 472
B.23 Logical AND operation. | 472
B.24 Experimental configuration for evaluating demountable MSLMs. | 482
B.25 Experimental configuration for evaluating the sealed visible-phocathode MSLM | 483
B.26 Demountable MSLM imaging experiments. | 484
B.27 Time response measurements. | 486
B.28 Secondary emission behavior. | 488
B.29 Microchannel plate (MCP) measurements. | 492
C.1 Schematic diagram of the single-channel high voltage amplifier circuit. | 498
C.2 Equivalent circuit for frequency response of the single-channel high voltage amplifier. | 499
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.3</td>
<td>Generalized photodiode op-amp model.</td>
<td>501</td>
</tr>
<tr>
<td>C.4</td>
<td>Relations between $\phi_e, I, v_o, v_x$, and $\phi_m$.</td>
<td>505</td>
</tr>
<tr>
<td>C.5</td>
<td>Circuit diagram of one channel of the twenty-channel high voltage amplifier.</td>
<td>508</td>
</tr>
<tr>
<td>C.6</td>
<td>Response of the high voltage vacuum tube amplifier.</td>
<td>511</td>
</tr>
<tr>
<td>C.7</td>
<td>Operating details of the twenty-channel high voltage amplifier.</td>
<td>515</td>
</tr>
<tr>
<td>C.8</td>
<td>Photographs of the twenty-channel high voltage amplifier</td>
<td>516</td>
</tr>
<tr>
<td>C.9</td>
<td>Interferometric readout of modulator phase, $\phi_m$, as grid voltage, $V_G$, varies.</td>
<td>521</td>
</tr>
<tr>
<td>F.1</td>
<td>Real-time holographic adaptive phase compensation with an SLM.</td>
<td>538</td>
</tr>
<tr>
<td>F.2</td>
<td>Closed-loop real-time holographic phase compensation.</td>
<td>538</td>
</tr>
<tr>
<td>F.3</td>
<td>Real-time holographic adaptive phase compensation with a photorefractive material-1.</td>
<td>539</td>
</tr>
<tr>
<td>F.4</td>
<td>Real-time holographic adaptive phase compensation with a photorefractive material-2.</td>
<td>539</td>
</tr>
<tr>
<td>F.5</td>
<td>Phase compensation by four-wave mixing.</td>
<td>539</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.1</td>
<td>Comparison of IPL behavior with various loop filters.</td>
<td>145</td>
</tr>
<tr>
<td>3.2</td>
<td>Time to first lose phase lock.</td>
<td>161</td>
</tr>
<tr>
<td>3.3</td>
<td>Computer simulation of phase-contrast IPL.</td>
<td>213</td>
</tr>
<tr>
<td>4.1</td>
<td>Phase-modulating optically-addressed SLMs.</td>
<td>248</td>
</tr>
<tr>
<td>B.1</td>
<td>Minimum MCP-to-crystal voltage, $V_{gm}$, for proximity focus.</td>
<td>480</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

1.1 Background

The past decade has seen a growing interest in the study of adaptive optical systems to overcome the limitations imposed by propagation over optical paths distorted by such time-varying effects as turbulence, scattering particles, thermal blooming, or thermally or gravitationally deformed optical components. Real-time wavefront compensation is expected to offer considerable improvement in performance to such applications as optical communications\textsuperscript{1-8}, astronomy\textsuperscript{9-16}, optical radar and laser beam steering\textsuperscript{17-20}, high energy beam propagation\textsuperscript{19-31}, laser-cavity mode control\textsuperscript{30-32}, figure control of large optical components\textsuperscript{33,34}, and general imaging through turbulent media\textsuperscript{19,35-46}.

An adaptive optical system generally functions by sensing wavefront distortions in real time and using this information to control a spatial light modulating device, which modifies the wavefront in order to optimize some aspect of system performance. Most adaptive optical systems can be classed as either: outgoing-wave systems which predistort a transmitted beam in order to maximize the power delivered to a distant target or received-wave systems which correct a distorted beam arriving at a receiver aperture. Since it is usually sufficient to compensate only phase distortions\textsuperscript{2,22,46}, an adaptive optical system is generally composed of two major components: a phase sensor-estimator structure and a spatial phase modulating device.
Historically, the first reference to adaptive optics was probably Babcock's\textsuperscript{9,10} proposal in the 1950's to employ predetection processing in astronomy. The adaptive microwave systems\textsuperscript{47} of the early 1960's also laid significant foundations, and adaptive optics has been recognized as an independent field since the 1966 Woods Hole Conference on the restoration of atmospherically degraded images.\textsuperscript{35} The earliest systems demonstrated include: active figure control of a three segment telescope mirror in 1969\textsuperscript{33,34}, a two-channel transmitted beam compensator in 1970\textsuperscript{22}, and a nine-channel compensated heterodyne receiver in 1971\textsuperscript{2}. Important more recent developments in adaptive optics include: a Hughes "multidither-COAT" transmitted beam compensator with eighteen channels\textsuperscript{18-20}, an ITEK "RTAC" image compensating receiver with twenty-one channels\textsuperscript{45}, "image sharpening" astronomical correction systems with six\textsuperscript{11,12}, nineteen\textsuperscript{14,15}, and eight\textsuperscript{46} channels, and demonstrations of wavefront conjugation by nonlinear optical mixing.\textsuperscript{48} The study of adaptive optics has grown so rapidly that there are too many references to conveniently cite; those mentioned above are just a small sampling. There is a variety of books\textsuperscript{49}, review articles\textsuperscript{30,31,50-52}, journal issues\textsuperscript{53}, and conference proceedings\textsuperscript{16,35,44,54-57} which provide a good overview.

Most previous adaptive optical systems have employed an array of complicated discrete phase estimators hard-wired to an array of discrete phase modulators. The complexity of this "hard-wired"
approach has limited adaptive optical systems to a few hundred resolution elements at most, whereas many applications such as atmospheric optical communications, image compensation, and high energy optical systems could potentially benefit from resolutions orders of magnitude greater. In particular, it has been suggested that laser communications through low-visibility atmospheric conditions, as in the presence of smoke, clouds, precipitation, or fog, could be considerably enhanced by adaptive systems containing millions or even billions of resolution elements. Motivated by the potential benefits of high-resolution adaptive optical systems, some of which will be discussed in more detail in section 1.4, this thesis research attempts to extend the state-of-the-art of adaptive optics to make high-resolution applications feasible. Many of these applications also make speed and sensitivity demands which the previous technology has difficulty meeting.

Unlike the previous "hard-wired" systems, a considerably more attractive approach for implementing a high-resolution adaptive optical systems is an "all-optical", parallel processing configuration. Figure 1.1 is a conceptual diagram of such a system, pictured as the front-end of a phase-compensated direct detection communications receiver. The spatially distorted input wavefront at $A'$ passes through an optically addressed, monolithic, spatial phase modulator which adds the appropriate phase, $\phi_m(\vec{r},t)$, to produce a compensated output wavefront at $A$. Part of the compensated wavefront is fed
Fig. 1.1 Conceptual diagram of an all-optical high-resolution adaptive optical system. (The lens $L_1$, aperture stop, and detector at the right constitute a diffraction-limited direct-detection receiver. The radius of the aperture stop opening is $r = 0.66 \lambda f / r_R$, where $f$ is the lens' focal length, $\lambda$ is wavelength, and $r_R$ is the input aperture radius.)
back into the box labeled "phase estimator", whose output is an optical control signal, which drives the optically-addressed spatial phase modulator to the desired phase distribution, $\phi_m(r,t)$. This system thus consists of two major components: a phase estimator which transforms spatial phase variations in its input beam into appropriate intensity variations in its output control beam, and an optically-addressed light modulator whose spatial phase modulation is controlled by the spatial intensity distribution of the control beam. This thesis is primarily concerned with techniques and devices for actually implementing these phase estimator and modulator structures. This "all-optical" approach allows systems containing a million or more resolution elements to be constructed, in principle, with only a negligible increase in complexity; only an increase of the resolution and/or active area of the modulator is required.

It should be mentioned that nonlinear optical mixing can provide an alternate means for implementing high-resolution adaptive systems. However, these nonlinear techniques generally require fairly large input signals and have only been applied to phase-conjugate transmitter predistortion applications. (However, as suggested in Appendix F, nonlinear techniques may also be applicable to compensated receiver applications.)

1.2 Phase estimation

Historically, a variety of techniques have been suggested and employed to measure wavefront phase aberrations in adaptive optical systems.
Use of direct phase measurement with a two-beam homodyne or heterodyne interferometer (e.g., Twyman-Green, Mach-Zehnder) has generally been limited to geometries where a stable local reference beam is easily obtained, such as optical figure correction systems \(^{33,34}\) or phase-conjugate transmitter predistortion systems \(^{22,23}\). In the latter application \(^{22,23}\), part of the transmitted laser beam reflects from a distant target and returns to the transmitter, where it can interfere with another wavefront from the same laser source.

Elementary two-beam interferometry produces a phase signal proportional to \(a \cdot \sin \phi\) in response to the wavefront \(a(x,y,t)e^{i\phi(x,y,t)}\). The \(a \cdot \sin \phi\) signal results in: an inability to distinguish wavefront phase from amplitude fluctuations, a dynamic range of only \(\pi\) radians, a nonlinear response to \(\phi\), and phase quadrant ambiguities in the inferred phase. In practice these difficulties have been overcome by inducing an optical frequency shift between the signal and reference beams, detecting the resulting heterodyne interferogram, and employing sophisticated electronic postprocessing in each spatial channel, such as synchronous detection and electronic phase measurement. Two-beam interferometry also suffers from a limitation to monochromatic light.

Direct self-interference phase measurement, where the reference wavefront is derived from the input beam, is more attractive in applications, such as astronomical image compensation, in which a stable external reference source and/or strictly monochromatic operation are not feasible. The use of such techniques as radial-shearing interferometry \(^{62}\) and Zernike phase-contrast \(^{13,60}\) has been proposed, often in conjunction with ac modulation of the interferogram \(^{60,62}\) and/or electronic postprocessing \(^{13,60,62}\) to remove the \(a \cdot \sin \phi\) phase signal
limitations.

A considerably broader spectral bandwidth, which is essentially white light, has been obtained by visualizing the local slope or spatial derivative, rather than absolute phase, of the wavefront. Arrays of Hartmann sensors\textsuperscript{61} or differential lateral-shearing interferometers\textsuperscript{63-65} have been employed for slope sensing, usually with ac modulation and complicated electronic postprocessing to reconstruct the phase distribution from the slope measurements.

A variety of other techniques, dependent on even more sophisticated post detection computation, have been proposed. One group of these, aperture synthesis techniques\textsuperscript{66,67}, is a variant of a Michelson stellar interferometer, in which wavefront phase aberrations are computed by processing interferograms made with the light passing through pairs of small subapertures in the input aperture. Another group of techniques is iterative algorithms\textsuperscript{68} which compute phase by measuring the intensity in the input aperture and/or image plane and employing known constraints, such as non-negative field amplitude or finite size of a distant source (e.g. a star).

Many adaptive systems have employed a completely different approach in which no direct measurements are made on the corrupted wavefront. Instead, a "hill climbing" procedure is employed in which the spatial phase modulator elements are systematically varied in order to maximize some figure of merit, such as the output of a single detector viewing the power reflected from a glint on a distant target. In the "multidither" COAT (Coherent Optical Adaptive Techniques), transmitter-predistortor systems\textsuperscript{18-22}, each phase compensator spatial element is modulated at a distinct "dither" frequency (frequency division multiplexing). The
output of a single detector is processed in order to derive a spatial distribution of modulator phases which maximize the light reflected from a target glint. In the "image sharpening" receiver systems 11,12,14,46, modulator elements are varied sequentially, (time division multiplexing), in order to maximize, for example, the power seen by a diffraction-limited, Fourier-plane detector.

Nonlinear optical mixing is another indirect technique, in that a predistorted transmitter beam (phase conjugate) is generated without explicitly estimating the wavefront distortion.

Most existing phase measurement techniques, such as those outlined above, either have undesirable performance limitations or are too complicated to implement the million or more resolution elements required by the high-resolution "phase estimator" box in Fig. 1.1. The ideal high-resolution phase estimator should be all-optical, employing only passive optical elements and possibly monolithic spatial light modulators which are addressed by optical beams to control other beams. Although essentially all-optical, such elementary phase visualization techniques as Zernike phase-contrast, heterodyne, homodyne, shearing, or polarization interferometry are by themselves not suitable. They cannot distinguish wavefront phase from amplitude variations and suffer from a nonlinear dynamic range of only $\pi$ radians as well as from $2n\pi$ and phase quadrant ambiguities. The configurations which have usually been employed to overcome these limitations are far too complex for a high-resolution phase detector.

This thesis develops a new phase measurement technique called an Interference Phase Loop (IPL) 69,70 which outperforms most existing
systems and is particularly well suited to high-resolution adaptive optics. More specifically, in the Interference Phase Loop technique, the output intensity from an elementary interferometer (e.g. Zernike phase-contrast, homodyne, etc) is detected and used to drive a phase-only modulator in the path of the wavefront being measured. The system in Fig. 1.1 becomes an example of an IPL when the box labelled "phase estimator" is replaced by an elementary interferometer. It will be shown both theoretically and experimentally that the Interference Phase Loop technique not only largely ignores amplitude fluctuations, but also has no phase quadrant ambiguity, and can continuously estimate phase over multiple \( \pi \) radians of dynamic range. In many respects, the IPL can be considered as an optical phase-locked loop which directly tracks optical wavefront phase.

1.3 Phase modulators

Previous adaptive optical systems have employed three major types of phase modulators: deformable mirrors, electrooptic phase modulating materials, and acoustooptic Bragg cells. The "deformable mirrors" have consisted of either an array of discrete reflectors\(^{12,19,20,34}\), whose height and/or tilt could be electronically controlled; or a continuous slab\(^{45,71}\), plate\(^{14,72}\), or membrane\(^{73}\) deformable by an array of electronic actuators. The electrooptic modulators\(^{2,74}\) and Bragg cells\(^{22,23,75}\) have also been employed in discrete array formats.

The "all-optical" parallel processing configuration of Fig. 1.1 requires a monolithic optically-addressed spatial light modulator (SLM). Although such devices have been proposed and demonstrated, mostly in the laboratory, none of the existing\(^{76,77}\) optically-addressed, phase-only
SLMs are suitable for implementing the desired all-optical adaptive system. A substantial portion of this thesis research program has therefore been devoted to the development of an SLM suitable for high-resolution adaptive applications.

Effort was concentrated on a promising new modulator called the Microchannel Spatial Light Modulator (MSLM)\textsuperscript{78-84}. The MSLM is potentially more sensitive and faster than most other optically-addressed SLMs. In addition, rudimentary prototype MSLMs had previously been constructed\textsuperscript{78,79} in our laboratory.

The MSLM, which was first proposed by C. Warde and J.H. Shapiro, is illustrated in Fig. 1.2. In operation, a control beam (coherent or incoherent light) incident on a photocathode creates an electron image which is amplified by a microchannel array plate (MCP) and proximity focused onto a dielectric mirror. The MCP is an array of tiny (approximately 10\,\mu m) semiconducting-glass lined pores, each functioning as a continuous-dynode multiplier of electrons "bouncing" down the pores. The charge distribution on the dielectric mirror produces a spatially varying electric field in an electrooptic plate coated with a transparent electrode on the other side. Since the electrooptic plate is constructed from a material whose refractive index is a function of the applied electric field, a spatially varying refractive index is induced which modulates the phase of the reflected read beam. (The electrooptic plate will also be called the "crystal", since it is usually a cut and polished crystal of an electrooptic material such as LiNbO\textsubscript{3}, LiTaO\textsubscript{3}, or KD\textsuperscript{*}P.) The MSLM is ultimately expected to be capable of multiple $\pi$ radian phase modulation dynamic range at kilohertz rates, with high resolution (20 lp/mm) and quantum-limited sensitivity.
Fig. 1.2 The Microchannel Spatial Light Modulator
1.4 Applications for high-resolution adaptive optics

This section will attempt to provide some perspective into practical applications and performance requirements for high-resolution adaptive optical systems.

The term "high-resolution" is employed in a slightly unconventional manner here in order to concisely convey the notion of a system containing a large number of resolution elements in an unspecified aperture area. This is more applicable than the conventional usage, which is in terms of resolution element density (e.g. cycles/mm or line-pairs/mm) or absolute size (mm$^2$/element); because with a given number of resolution elements, the adaptive aperture can be arbitrarily (within broad bounds) reimaged to another area. As in conventional usage, the start of the "high-resolution" regime is relative to the application, (e.g. 1000 lp/mm is high resolution for photographic film, but 20 lp/mm is for real-time SLMs). In this adaptive optical context, systems with more resolution cells than can be practically implemented by discrete-channel, hard-wired technology will be considered to be high resolution. This regime starts at a few hundred resolution elements$^{49-57}$. (This author is not aware of any discrete systems with more than about 350 elements$^{59}$). As a practical matter, the discrete-channel experience of this thesis research (which is reported in Chapters 5 and 6) suggests that the "high-resolution" techniques to be discussed can considerably simplify the fabrication, and the electronic and optical alignment, of adaptive systems employing even tens of resolution elements.
Low-visibility optical communications

It has been suggested\textsuperscript{1,3-8,85-87} that high-resolution adaptive phase compensation may considerably improve the performance of optical communications links through low-visibility weather. If reliable operation could be assured through such strongly scattering conditions as snow, rain, fog, or smoke, atmospheric optical communications appears attractive for such applications as\textsuperscript{1,3,4,8,87} satellite-to-earth links (high bandwidth and antenna gain with a tiny, low-power transmitter), temporary or frequently reconfigured links (e.g., compact, secure, directional ship-to-ship link or optical "walkie-talkie"), or non-line-of-sight optical broadcast.

As an optical beam propagates through the low-visibility atmosphere, its direct line-of-sight intensity decays exponentially with distance due to absorption and scattering losses. At visible wavelengths the absorption is usually small (high albedo),\textsuperscript{98} and since the particles are significantly larger than a wavelength, much of the scattered radiation is forward directed.\textsuperscript{1,3-8,88} In fact, the scattered beam is often much more intense than the direct line-of-sight beam.\textsuperscript{1,3-8,89-93} The potential therefore exists for improving optical communications performance by effectively utilizing the scattered light in addition to the direct beam.

The multiply scattered light arrives from a wide range of angles (broad angular spectrum) and hence is not collected by conventional narrow field-of-view (FOV) receivers, such as heterodyne\textsuperscript{2,94-98} or diffraction-limited direct-detection\textsuperscript{8,94-96}. One solution, as illustrated in Fig. 1.1, is to precede the narrow-FOV receiver with an adaptive optical system, which senses and "straightens out" the
scattered wavefront distortions. The scattered light is then converted to an approximate planewave, which can be collected by a diffraction-limited FOV receiver.

Actually, merely increasing the FOV is not sufficient to justify the additional complexity of an adaptive optical system. In a direct-detection receiver, which consists of a lens followed by a detector in its Fourier plane, the FOV can be enlarged to receive the scattered light by merely increasing the detector area. However, background light also arrives with a broad angular spectrum and hence often renders this simple wide-FOV ("photon bucket") receiver inoperable, particularly during the daytime. The real advantage of the adaptive receiver in Figure 1.1 is that it combines wide-FOV signal reception with the inherent discrimination against background light of a diffraction-limited receiver.

From the Fourier transform relationship between the angular spectrum and mutual coherence function99 (i.e., Cittert-Zernike Theorem100), one would expect that the broad angular spectrum scattered light is coherent over a very small area. This partially coherent light in the receiver aperture can be treated as a spatial array of tiny uniform "coherence cells"89-93 of radius $\rho_0$3,4,8,89-93 (see Eq. (E.2) in Appendix E). Low-visibility propagation studies suggest that $\rho_0$ is often so small that an adaptive system must be able to resolve a million or more coherence cells in order to compensate a receiver aperture only a few centimeters in diameter.

The expected performance and system requirements of low-visibility adaptive receivers can be explored in greater detail with the aid of
a low-visibility propagation model known as the "multiple-forward scatter" (MFS) model. Some of the basic results of that model are reviewed in Appendix E. According to the MFS model, when the direct line-of-sight beam is attenuated by $-e^{-ksL}$ over a propagation link of length $L$, the scattered radiation collected by a wide-FOV receiver is only attenuated by $-e^{(1-\alpha)ksL}$, where $\alpha = 0.57$. Wide-FOV reception thus offers a signal gain of $-e^{\alpha ksL}$.

Without adaptive phase compensation, a wide-FOV direct-detection receiver requires a Fourier-plane detector area of $A_{dw} = \frac{\lambda^2 f^2}{(4\pi p_0^2)}$ (where $F =$ lens focal length). The compensated receiver of Fig. 1.1 has a diffraction-limited detector area of $A_{dd} = \frac{(1.22\pi)^2 \lambda^2 f^2}{(4\pi p_0^2)}$, where $r_R$ is the input-aperture radius. Since background light has an approximately uniform angular spectrum, it is uniformly distributed in the detector (i.e., Fourier) plane. The adaptively compensated receiver thus suppresses background light by a factor of $A_{dd}/A_{dw} = (p_0/r_R)^2 \approx 1/M$. Note that $M = (r_R/p_0)^2$ is the number of coherence cells in the input aperture, which is often in the range of $10^6$ to $10^9$.

Over very long pathlengths, the line-of-sight beam may be so completely attenuated that only scattered light is seen, even by a diffraction-limited receiver. Since the broad angular spectrum of scattered light is approximately uniform over a small Fourier-plane detector ($A_d < A_{dw}$), over long paths the signal advantage of a wide-FOV receiver relative to a diffraction-limited receiver is approximately $A_{dw}/A_{dd} = M$, (see Eq. E.1b), instead of $e^{\alpha eL}$.

As mentioned in Appendix E, adaptive phase compensation is superior to simple wide-FOV direct-detection when the wide-FOV en-
hanced signal \( x e^{\alpha_k s} L \) or \( x M \) is not large enough to offset the detrimental effects of the wide-FOV background \( x M \). For a given transmitter power and level of background radiance, this condition (Constraint (3) in Appendix E) places a minimum constraint on the link length \( L_{\text{min}} \), which \( L \) must exceed in order for phase compensation to be potentially attractive.

A variety of feasibility constraints can determine the maximum useable compensated link length, \( L_{\text{max}} \). One constraint ((4) in Appendix E) is that the compensated signal-to-noise ratio (SNR) be sufficient for reliable communications, given the desired data rate, transmitter power, and maximum acceptable error probability. Another ((5) in Appendix E) is that there be sufficient signal to drive the phase compensator; it will be seen in Section 3.1.9 that at least a few photons must be collected in every coherence cell during the "coherence time" \( T_C \) of the scattering medium. \( T_C \) is defined as a time short enough that the dynamic scattered-wave aberrations appear to be approximately static. There is a lack of good theoretical models or experimental measures for the low-visibility medium coherence time; it is probably not usually larger than a few \( \mu \text{sec}^8 \).

Appendix E shows that when simple wide-FOV direct-detection is limited by background light, the signal light available to drive the phase compensator is also small relative to background light. This is very restrictive, but Eq. (E.6) suggests that phase compensation can still be both desirable and feasible, i.e. \( L_{\text{min}} = L_{\text{max}} \), if

\[
\left[ \frac{\eta^2 N_\lambda}{n_0} \Delta \lambda \right] T_C > 5
\]

\((N_\lambda \text{ and the other factors in this expression are defined in Appendix E).} \)
This can be interpreted as a minimum constraint on the scattering channel coherence time, which $T_c$ must exceed in order for adaptive low-visibility communications to be viable. The above relation also reflects the fact that phase compensation is not justified when alternative background light suppression techniques are employed, such as a narrowband spectral filter for small $\Delta \lambda$ or a low-background communications wavelength (e.g., 0.2 to 0.3$\mu$m solar-blind region) for small $N_\lambda$. (In practice filters which are simultaneously wide-FOV and suitably narrowband are not available. There is also a lack of sources and modulators for mid-UV communications.)

Another important consideration in a low-visibility communications link is "multipath spread", wherein scattering increases the path length and thereby time required for light to travel from the transmitter to receiver. Thus a transmitted impulse is received as a broadened pulse with a decaying tail, with width $T_p$ (e.g., $e^{-1}$ intensity); and the communications bandwidth (e.g., maximum date rate) of the optical channel is reduced to the "coherence bandwidth", $w_c = 1/T_p$. A variety of experimental measurements$^{89-91}$ and theoretical expressions$^{8,89-91,101}$ are available for $w_c$. In many instances $w_c > 50$ MHz has been measured$^{8,89}$; although under severe conditions $w_c$ as small as a few MHz or less has been reported$^{90,91}$. The scattering medium should be underspread$^{1,4}$, i.e. $T_c > T_p = w_c^{-1}$. Otherwise, if the compensator has enough signal, there is almost surely enough energy in each signal pulse, of width $T_p > T_c$, to communicate without compensation.

In conclusion, a wide-FOV receiver can significantly improve the performance of a low-visibility optical communications link if the additional wide-FOV background light is somehow suppressed. Sources for solar blind operation or sufficiently narrowband filters currently do
not exist, which leaves adaptive phase compensation as a potential background suppression technique. However, suitable adaptive phase compensation hardware, capable of: extreme resolution (10^5-10^8 spatial elements), fast response (\(\mu\)sec), and extreme sensitivity (quantum limited), needs to be developed. The techniques and devices discussed in this thesis are attempting to close the wide gap between these demanding specifications and existing adaptive technology. More fundamentally, additional experimental and theoretical characterization of the low-visibility coherence time is required to determine how often the channel varies slowly enough to allow the adaptive system to collect sufficient energy.

**Optical communications through atmospheric turbulence**

In many respects adaptive optical communications through turbulence \(^2,30,31,86,87,94-98,102-106\) is similar to the low-visibility case, but the effects of turbulence are less severe. The direct line-of-sight planewave, however, is not heavily attenuated as in the low-visibility case, but is converted by atmospheric refractive index fluctuations from a planewave into a beam with a broader angular spectrum. Eq.(D.21) in Appendix D is an expression for the coherence cell radius, \(\rho_o\), of the corresponding distorted wavefront in the receiver aperture; typically \(1\text{cm} < \rho_o < 10\text{cm}\) \(^{95-98,102-106}\). For \(r_R < \rho_o\) the received energy increases in proportion to \(r_R^2\) \(^{95-98,102-105}\). Increasing \(r_R\) beyond \(\rho_o\), however, offers little additional increase in the signal energy collected by a diffraction-limited receiver \(^{95-98,102-105}\). In a diffraction-limited direct-detection receiver, for example, most of the additional signal falls outside the detector aperture stop in the
Fourier plane. The full signal could be collected by a simple wide-FOV direct-detection receiver (photon-bucket), but the collected wide-FOV background light is often excessive. The adaptive phase-compensated receiver of Fig. 1.1 is thus applicable here, allowing simultaneous wide-FOV reception, with the signal proportional to $r_R^2$, and diffraction-limited background rejection.

To the extent that the central region of the turbulence-broadened angular spectrum is flat, phase-compensated reception results in approximately $(r_R/\rho_0)^2 = M$ times more signal collection than uncompensated diffraction-limited reception. Actually, as mentioned in Appendix D, the improvement can be even more dramatic, being more accurately approximated by $\exp(1.031 \cdot M^{5/6})$. This assumes that the phase compensator can resolve $\rho_0$ (i.e., has $\geq M$ resolution elements). Phase compensation reduces the background light by a factor of about $1/M$ relative to a simple wide-FOV photon bucket.

Besides the phase distortions (broadened angular spectrum) mentioned above, turbulence also introduces signal amplitude fluctuations (scintillation) across the receiver aperture. Perfect adaptive amplitude compensation$^{2,31,46}$, in addition to phase compensation, only provides a signal advantage of about $1.8^{31}$. In fact, phase-only compensation can be superior to phase and amplitude compensation. For example, Appendix D shows that in some cases the intensity seen by a diffraction-limited receiver is proportional to $<|E|^2>e^{-\sigma^2}$, where the brackets denote spatial averaging, $\sigma = \sqrt{\langle \phi^2 \rangle - \langle \phi \rangle^2}$, and $|E(x,x)|e^{j\phi(x,y)}$ is the spatial field distribution in the input aperture. Phase compensation results in $\sigma = 0$, but any attempt to
compensate $|E_1|$ with a typical attenuating amplitude modulator will reduce the overall signal. (Similar arguments can be advanced for phase-only compensation in the low-visibility case discussed previously).

Due to amplitude fluctuations, a large receiver aperture ($r_R > r_0$) is desirable in order to employ "aperture averaging" to reduce the system noise contributed by scintillation. In order to effectively utilize the additional signal collected by this larger aperture) an adaptive phase compensator containing a relatively large number of resolution elements is required, e.g., $M=4.9 \times 10^3$ for $r_R = 70$ cm and $r_0 = 1$ cm.

A large portion of the increased angular spectrum due to turbulence is due to large-scale, dynamic (1 Hz to 1 kHz) angle-of-arrival fluctuations. After compensation with a tilt-tracking receiver, however, there is still significant higher-order structure with a coherence cell size of about 3.4 $r_0$, which requires a fine-scale phase compensator.

Turbulence also results in spreading and steering of the transmitter beam, which can cause a large amount of power to completely miss the receiver. These effects provide an additional application for adaptive phase compensation. The phase compensator can be moved to the transmitter in order to predistort the transmitted beam with the conjugate of the turbulence phase distortion it will encounter in propagation. Approximately the same coherence cell size, $r_0$, and number of resolution elements, $M$, are required for transmitter predistortion as for receiver compensation.
The coherence bandwidth of the turbulent channel is generally very broad ($\omega_c > 10^{10}$ Hz)$^{98}$ and usually does not limit performance. The turbulence-induced phase fluctuations usually vary relatively slowly, with the coherence time, $T_c$, generally ranging from $1$ msec to $100$ msec.$^{95-98,102-105}$.

In conclusion, adaptive phase compensation offers the potential of significantly improving the performance of communications links through atmospheric turbulence. The required number of resolution elements, $10^2$ to $10^4$, is often larger than can be reasonably implemented with the previous hard-wired discrete-channel technology, but is well within the realm of the high-resolution techniques and devices discussed in this thesis. Due to the relatively long coherence time of turbulence, $T_c > 1$ msec, these high-resolution adaptive systems should experience no difficulty operating in real-time and collecting sufficient signal energy.

**Imaging through atmospheric turbulence**

It has long been recognized$^{9-16,35-46,98}$ that such applications as astronomical imaging, high altitude surveillance of the earth, and general long-distance imaging on the earth's surface are significantly degraded by atmospheric turbulence. The dominant effect of turbulence is resolution reduction to a "coherence cell" limited resolution ($\sim 1.22\lambda/2\rho_o$)$^{9-16,35-46,98}$, rather than the diffraction limit of the physical aperture ($\sim 1.22\lambda/2r_p$). Since $\rho_o$ rarely exceeds $10$ cm$^{9-16,35-46,98}$ the long-exposure resolution of the best earth-based telescopes is generally limited to about $1$ arcsec. A short exposure ($< 10$ msec)$^{15,39}$ can result in a slightly better resolution of
about $1.22\lambda/6.8\rho_o^{37,40,42}$, by removing the smearing due to image dancing (i.e., angle-of-arrival fluctuations). Vastly improved resolution, however, can be obtained by compensation$^{9-16,38-40,43,48}$ of the spatial phase distortions in the input aperture with an adaptive optical system which can resolve the coherence cells.

Since a large aperture is generally employed in order to collect sufficient light, reduce amplitude scintillation (i.e., twinkling), and achieve high compensated resolution, relatively high resolution adaptive systems, such as those developed in this thesis, are required. For example, $M = (r_p/\rho_o)^2 = 2.1\times10^3$ resolution elements are required to compensate a 36 inch telescope most of the time ($\rho_o = 1\text{cm}$), or $M = 6.5\times10^4$ for the 200 inch Hale telescope.

When a discrete-element adaptive phase compensator is employed even greater resolutions may be required, because discrete-element modulation of the input aperture results in multiple repeating images.$^{100,107,108}$ These can degrade resolution by overlapping when the Nyquist sampling theorem is not satisfied.$^{107-109}$ The spacing of the modulation elements, $P$, should generally be less than half of the smallest detail, $d_f$, in the input aperture. For example, if a planet of diameter $D$ is being imaged over a distance $L$, $d_f = \lambda L/D$ $^{100,107,108}$. In order to resolve planet details of size $d_p$, the input aperture should have a diameter of at least $2r_R = \lambda L/d_p$; hence the compensator requires $-(2r_R/P)^2 = (4r_R/d_f)^2 = 4(D/d_p)^2$ resolution elements. Thus a discrete-element system should have more resolution elements than the number of image points $(D/d_p)^2)$ being observed in the object.

Generally, only portions of an object experiencing the same phase distortions, i.e., lying in the same "isoplanatic patch"$^{13-15,31,98}$,
can be compensated. (Thus, in the above discussion D should be the isoplanatic patch size, rather than the total object size). In astronomical applications the isoplanatic patch is about 10 arcsec\textsuperscript{14} (Jupiter extends 40 arcsec).\textsuperscript{14} Actually, the isoplanatic patch size can be increased by compensating the various layers of phase aberration with multiple modulators in series.\textsuperscript{13,51,31}

One potential difficulty with adaptive image compensation is that atmospheric phase distortions cannot generally be distinguished from the phase of a coherently illuminated object\textsuperscript{98}. This is not a problem if one has apriori information, for example that the object has a spatially uniform phase distribution. More generally, the object can be accompanied by a planar reference beam, such as an unresolved adjacent point source or reflecting glint.

Image compensation is thus an important potential application of the high-resolution adaptive systems discussed in this thesis. These systems should have no difficulty achieving the resolutions (>10\textsuperscript{3} elements), speeds (<1KHz), and sensitivities required for most atmospheric imaging applications. It will also be seen (in Sec. 3.2.2.9) that some of these systems can perform well in white light, which is necessary, for example, in astronomical applications. In fact, previous white light adaptive imaging systems, such as those employing shearing interferometry,\textsuperscript{45,58,63-65} Hartmann sensors,\textsuperscript{61} or image sharpening\textsuperscript{11,12,14,15,46} have generally been too complicated to achieve the resolutions required to fully realize the potential benefits of phase compensated imaging. In addition, the image sharpening\textsuperscript{11,12,14,15,46} approach is fundamentally less sensitive,\textsuperscript{50,110} by a factor of about M, than most other approaches, including those of this thesis.
High energy laser systems

High energy laser systems\textsuperscript{24-32,111} in such applications as laser fusion, long-haul power distribution, and weaponry encounter a variety of degrading effects which can be effectively corrected by adaptive optical systems.

Adapative phase-compensated laser cavities\textsuperscript{30-32} have been utilized to remove aberrations generated by such effects as gas flow instabilities and thermal distortions, and to control the laser mode structure. Although tens of resolution elements can offer considerable improvement, in some instances a hundred or more may be justified\textsuperscript{30-32}.

Absorption of energy from the optical beam can cause very serious thermally-induced distortions in optical components. A large portion of these distortions are due to high spatial-frequency "Fresnel Ripples"\textsuperscript{30,31} generated by near-field diffraction of the optical beam. High-resolution adaptive systems, with more than $10^3$ elements, are generally required to compensate these Fresnel Ripple aberrations.

A high-energy beam passing through the atmosphere generates thermally induced refractive index gradients, which in turn cause the beam to bloom or spread.\textsuperscript{24-31} The energy loss due to thermal blooming can sometimes be reduced by adaptive predistortion of the transmitter beam with the conjugate of the thermal phase distortions.\textsuperscript{22,23,30,31} In strong turbulence, however, this technique can become unstable. Multi-dither systems,\textsuperscript{18-21} which maximize the power on a glint at the target, tend to perform better. Moderate resolutions, of tens of elements, appear to be adequate for thermal blooming compensation. Another means for reducing the effects of thermal blooming and thermal optical component distortions is to spread the beam energy over a large
aperture area of several meters. 28,111

Finally, atmospheric turbulence is, of course, a major source of aberration. The resulting beam steering and spreading can be removed by a phase-conjugate, transmitter-predistortion adaptive system. Large numbers of compensator elements, \( 10^4 \) or more, are often required, particularly when large apertures are employed to reduce thermal blooming effects. The high-resolution adaptive systems of this thesis can be applied there. It will be shown in Chapter 2 that the basic adaptive system on the left half of Fig. 1.1 can be employed for transmitter conjugation if a weak probe beam or glint reflection is available from the distant target.

Most high energy laser applications are thus dependent on adaptive phase compensation for success, and the phase compensator must be able to resolve large numbers of resolution elements \((10^3 \text{ to } 10^4)\) in order to compensate such spatially diverse aberrations as Fresnel Ripple thermal distortions and atmospheric turbulence. The high-resolution techniques of this thesis are applicable, but further work is required to develop compatible modulators which can tolerate the high optical powers involved.

Other important potential applications for high-resolution adaptive optical systems include low-power laser radar and tracking systems, which require maximization of the beam intensity on a distant target, and figure correction of large telescopes in space.

Summary

Such applications as atmospheric optical communications, atmospheric imaging, and high energy laser systems cannot realize their full potential without the aid of adaptive optical systems containing
thousands, or even more than a million, resolution elements. While the need for these high-resolution systems has been recognized \(^{30,31,56,111}\), the previous hard-wired, discrete-channel technology has prevented the demonstration of any adaptive systems with more than a few hundred resolution elements. The techniques and devices of this thesis should ultimately make it possible to construct adaptive systems with sufficient resolution, along with the speed and sensitivity required by the above-mentioned applications.

1.5 Thesis organization

Following a brief review of a wide variety of phase measurement techniques, most of which are not well suited for implementing sensitive high-resolution adaptive systems, Chapter 2 introduces the basic operational theory of the Interference Phase Loop. This theory is extended in Chapter 3, where attention is focussed on most of the more important aspects of IPL performance, and a few alternative interferometric configurations are discussed. The reader who is not particularly interested in the detailed theory of IPL operation can skip Chapter 3 without a great loss of continuity. Alternatively, the Table of Contents can be used as a guide to focus on particular aspects of IPL theory which are of interest.

After a short overview of alternative existing modulators, Chapter 4 deals with the basic operational theory, developmental work, and performance of the MSLM. Chapter 4 also describes some of the special aspects of attempting to implement an Interference Phase Loop with a monolithic spatial light modulator, in particular the MSLM.

Although a large portion of the thesis research program was concerned with MSLM development, only a portion of that work is reported
here, due to space limitations and the peripheral nature of that work relative to adaptive optics. Some of the more important contributions of the MSLM work are documented in Appendix B. These include a theoretical description of the secondary electron emission behavior of the MSLM (important in the functioning of the MSLM in an IPL adaptive system) and a description of special functions, such as thresholding, contrast enhancement, contrast reversal, and binary logic operations, which the MSLM is intrinsically capable of performing.

Chapter 5 describes a variety of optically-driven phase modulating systems which were constructed to implement and investigate the proposed adaptive optical techniques. These discrete-channel systems allow IPL performance to be studied in more depth; and they can simulate the response characteristics of a large variety of SLMs.

Finally, experimental adaptive optical systems, which were constructed and evaluated in the laboratory are discussed in Chapter 6. These systems demonstrated the basic practicality of the IPL technique in phase estimation and compensation applications. Much of the basic operational theory of Chapters 2 and 3 was verified in a system containing a single spatial resolution element. Wavefront shaping and phase compensation were demonstrated with a nineteen-element, discrete-channel system; and an MSLM was successfully employed to demonstrate "all-optical" phase compensation.

The appendices support the discussion in the main text by: reviewing required background results, often from a perspective not usually seen in the literature; providing alternative derivations; and presenting some
of the more tedious details. Appendix A is associated mainly with Chapter 3, B with Chapter 4, and C with the experimental results of Chapters 5 and 6. Appendix C attempts to provide details which would allow the reader to reproduce and extend the experimental work. Appendix D presents some important results concerning the Strehl ratio, which are utilized in a number of places. Appendix E reviews the multiple forward scatter model for low-visibility optical communications. Appendix F presents alternative high-resolution phase compensation techniques utilizing nonlinear optical processes. Finally, Appendix G provides bibliographic background on phase measurement techniques.
II. PHASE ESTIMATION

Aside from measuring the atmospheric phase distortion to be compensated in an adaptive optical system, techniques for real-time measurement of optical phase have been developed for such diverse applications as: studying refractive index fluctuations in plasmas or gases\cite{112,114,115,116,117,118,33,34,119,122,126,123,124}, visualizing transparent structures (e.g., biologic cells)\cite{117,118}, extracting phase information from a coherent optical processor\cite{33,34,119,122,126}, and characterizing surface nonuniformities\cite{123,124}. This existing phase measurement technology was reviewed and a variety of new modifications and techniques were considered for implementing the phase estimator in the high-resolution, adaptive configuration of Fig. 1.1. Designing a suitable phase estimator is complicated by the fact that the amplitude of the wavefront to be compensated usually also varies in space and time. Besides ignoring amplitude variations, the ideal phase estimator should also linearly estimate phase over a multiple $\pi$ radian dynamic range and operate without $2\pi$ or phase quadrant ambiguities. The phase estimate must also be compatible with an all-optical format or at least be simple enough to allow the implementation of many parallel channels. Effort was finally concentrated on a new technique called an Interference Phase Loop\cite{69,70}, which meets the criteria outlined above and performs better than most existing phase measurement systems. The basic static and dynamic behavior of the IPL are introduced in section 2.2 of this chapter, and a more comprehensive analysis can be found in Chapter 3.

2.1 Overview of phase measurement techniques

Space limitations do not allow a truly comprehensive overview of
phase measurement technology to be presented here; however Appendix G provides a more extensive, although by no means all-inclusive, bibliography of existing phase measurement techniques.

2.1.1 Elementary interferometric phase measurement

Elementary interferometry, in which a reference beam is added to the wavefront being measured and the resulting intensity is detected, is perhaps the most straightforward and well-known technique for phase measurement. Mach-Zehnder, Twyman-Green, and Fizeau two-beam interferometers are commonly employed configurations in which the reference and aberrated beams follow separate paths. There are also a large variety of self-interference configurations in which the reference wavefront is derived directly from the aberrated wavefront. As mentioned in Section 1.2, self-interference techniques are often useful in a wider range of applications and have a broader spectral bandwidth (at least a few hundred angstroms) than separate path or external reference techniques.

Zernike phase-contrast is a well-known self-interference technique, in which the input wavefront is Fourier transformed with a lens, and the zero order of the transform is multiplied by a tiny diffraction-limited spot. The inverse transform then contains an image of the input wavefront added to an approximately planar reference wavefront which is diffracted from the spot. Zernike phase-contrast will be discussed in more detail in Section 3.2.2 and appendix A.3, (see Fig. A.3).

Zernike phase-contrast is just one of a large variety of phase visualization techniques which multiply the Fourier transform of the measured wavefront by a specialized frequency-plane Fourier filter.
When the filter has opaque regions, such as a knife-edge or opaque spot, these are often referred to as Schlieren techniques. Other Fourier filters which have been utilized include: multiple diffraction-limited spots, birefringent plates, segmented polarizers, steps of attenuation, linear transmission gradients, nonlinear transmission gradients, and gratings. Some of these systems produce an output proportional to the spatial derivative of phase, rather than the desired phase estimate.

In the radial shearing self-reference techniques, a small portion of the aberrated wavefront is expanded, e.g. with a small lens, to become a reference wavefront. Double-focus polarization interferometry is a related technique in which a birefringent lens divides the aberrated beam into extraordinary and ordinary wavefronts which are imaged to different points. In the image plane of one wavefront, the other wavefront serves as a quasiplanar reference beam.

A lateral-shearing interferometer generally employs a grating or a birefringent polarizing element to divide an input wavefront into two laterally shifted wavefronts, which can interfere with each other. With large shear, this becomes a "total doubling" interferometer and visualizes the phase shift between pairs of points in the wavefront. With small shear, this becomes a "differential" interferometer and visualizes the local slope or phase derivative.

In a scatterplate interferometer wavefronts are imaged through a special weakly scattering screen which adds a scattered reference beam to the image. (In practice, the configuration is more complicated, with two passes through the scattering screen being required).

The elementary interferometers considered here are essentially
all-optical, high-resolution systems, however as mentioned previously, their performance is seriously limited by a phase estimate of the form $a \cdot \sin \phi$ (e.g. sensitive to amplitude variations, dynamic range of only $\pi$ radians, $2n\pi$ and phase quadrant ambiguities).

2.1.2 Improved phase measurement systems

A variety of sophisticated configurations have been employed to overcome the limitations of elementary interferometry.

Many of these systems produce an ac phase signal of the form $a \cdot \sin(\omega t + \phi)$ and employ electronic post-processing to extract the phase. Methods for obtaining the ac phase signal include: heterodyne interferometry, where the optical frequency of one path of a separate-path interferometer is shifted; homodyne interferometry, with a tilted reference beam and detection with a scanning sensor (e.g. vidicon tube); and conversion of phase shift information into a local rotation of linear polarization direction and following this polarized beam with a rotating analyzer. The last technique is also applicable to obtaining an ac output from a self-reference Zernike phase-contrast interferometer. Optical frequency shifts in the heterodyne interferometers have been produced by such means as Bragg cells, rotating gratings, rotating $\lambda/2$ or $\lambda/4$ plates, sinusoidally driven electrooptic phase modulators, and axial-shifting or vibrating mirrors.

A variety of electronic means, such as zero crossing measurements relative to a reference sinusoid, can be employed to determine the phase of an ac interference signal. With two quadrature reference signals (i.e. $\cos(\omega t)$ and $\sin(\omega t)$) and auxiliary digital circuitry, phase quadrant and $2n\pi$ ambiguities can be eliminated. Another technique
is to multiply the phase signal by \( \sin(\omega t) \) and \( \cos(\omega t) \), followed by low-pass filtering to generate the signals \( S = a \cdot \sin \phi \) and \( C = a \cdot \cos \phi \) (quadrature synchronous detection). A digital or analogue circuit can then compute \( \phi = \tan^{-1}(S/C) \). A variation of this technique is to sequentially sample the ac intensity \( I(t) \), and digitally compute \( S = \sum_{i} I(t_i) \sin(\omega t_i) \) and \( C = \sum_{i} I(t_i) \cos(\omega t_i) \) (discrete Fourier transform coefficients) and the phase \( \phi = \tan^{-1}(S/C) \).

An interesting closed-loop heterodyne interferometer was recently demonstrated in which phase was measured by controlling the reference beam path length (by axially moving the same vibrating mirror which generated the heterodyne optical frequency shift, \( \Delta \omega \)) in order to null or maximize the detected interferogram intensity. An error signal was generated by differencing (to eliminate phase quadrant ambiguities) two bandpass filtered versions of the detector intensity separated in time by half a cycle (\( \Delta \omega / 4\pi \)). The integral of this error signal controlled the reference beam path modulator. A fringe counting circuit eliminated \( 2n\pi \) ambiguities by discharging the integrator and registering a count whenever the phase changed by a full cycle. This system employed complicated analogue and digital electronics, rather than exploiting the inherent capabilities of closed-loop operation, which will be discussed in Sec. 2.2.

As mentioned in Chapter 1, another class of high-performance phase estimation systems produces an optical signal which is a function of the local slope or spatial derivative of the wavefront phase. Electronic postprocessing is then employed to reconstruct the two-dimensional spatial phase distribution. This approach usually removes phase quadrant and \( 2n\pi \) ambiguities, allowing phase estimation over multiple \( \pi \) radians of dynamic range. Frequency plane filters of the form: \( a + bx \), \( \sqrt{a + bx} \), or a three-
step approximation to \( a + bx \), have been employed to obtain a spatial derivative; but they do not perform well when the signal amplitude also varies. Better performance can be obtained with polarization or grating differential lateral-shearing interferometers \(^{45,64,132}\).

Although most of these slope visualization techniques can operate with fairly broad spectral bandwidth light (>100Å), some have been demonstrated which perform well with incoherent white light. These include: Hartmann sensors \(^{50,61}\), a Moire effect system \(^{141}\), and a modified lateral-shear grating interferometer \(^{63}\). The Hartmann sensor divides the aperture into an array of subapertures, each containing a lens followed by a quadrant detector in its focal plane; local wavefront tilt moves the focal spot over the quadrant detector. In the Moire system \(^{141}\), a Moire pattern produced by a wavefront passing through two close-spaced gratings is modified by the local wavefront direction of travel (i.e. slope).

In order to enhance the isolation from amplitude variations, it is usually desirable to employ a slope sensor which produces an ac signal of the form \( a \cdot \sin(wt + \alpha \frac{d\phi}{dx}) \), where \( \alpha \) is a constant. This is easily accomplished in a grating lateral-shear interferometer by employing a rotating radial (sunburst pattern) grating \(^{45,63-65}\). An ac signal can be generated in a Hartmann sensor \(^{61}\) by wobbling the input wavefront tilt with a nutating mirror; it is also desirable to add an undistorted reference beam, before the nutator, to generate a reference signal \(^{61}\).

Many of these slope visualization systems produce only a one-dimensional derivative, and hence require two identical orthogonally oriented two-dimensional sensor arrays operating in parallel. A single array of Hartmann sensors, however, is usually sufficient, due to the quadrant
detector in each channel.

Phase estimation techniques based on the analytic functional properties of the optical field and Fourier transform properties of optical propagation have received considerable attention in recent years. (Appendix G contains a bibliography of some of the major work in this area). This area of study is often termed "the phase problem" and generally consists of attempting to determine the spatial phase distribution of a complex field (e.g. optical field or mutual coherence function) based purely on modulus (e.g. optical intensity) measurements. For example, it is well known that the phase of a "minimum phase" field is uniquely determined (within a spatial constant) by the Hilbert transform of its modulus. The optical fields of interest, unfortunately, are not usually minimum phase, and there are generally a large, but finite, number of phase distributions which result in identical intensity distributions.

In the optical "phase problem", the Fourier transform $F = \mathcal{F}(ae^{j\phi})$ of the input aperture distribution ($f = ae^{j\phi}$) is also of interest, especially since the input aperture can be Fourier transformed with a lens and the resulting intensity distribution, $|F(u,v)|^2$, can be detected. If $\phi_F$ could be determined, the inverse transform could be computed to obtain $ae^{j\phi}$ and hence $\phi$, because $a^2$ is directly measureable. Unfortunately, $\phi_F$ cannot be uniquely determined from $|F|$. The one-dimensional case of the Fourier transform $F(u)$ resulting from the finite aperture distribution $f(x)$ ($f(x) = 0$ for $-\infty < x < x_1$ and $x_2 < x < \infty$, where $x_1 < x_2$) has been widely investigated. It has been found that there are $2^M$ possible solutions for $\phi_F$, given $|F(u)|$, where $M$ is an integer dependent on the exact form
of $F(u)$ or $f(x)$. (The solution for $\phi_F$ depends on $M$ complex constants, which can be determined from $|F|$ except for the sign of their imaginary parts. Thus, all $2^M$ possibilities for $\phi$ can in principle be computed, but the correct one cannot generally be identified.)

Robinson\textsuperscript{142} considered the possibility of computing $\phi$ (or $\phi_F$) from measurements of both the aperture intensity $|f|^2$ and Fourier-plane intensity $|F|^2$. He concluded that, in general, $\phi$ cannot be uniquely determined. Other researchers, however, have obtained much more positive results. Greenaway\textsuperscript{144} has suggested that in the 1-D case mentioned above, an opaque mask placed in part of the finite aperture (i.e., $f(x)=0$ for $-\infty<x<x_1$, $x_3<x_4$, and $x_2<x<\infty$; where $x_1<x_3<x_4<x_2$) can result in $M=0$, and hence a unique solution for $\phi_F$ given $|F|$. Gonsalves\textsuperscript{145} demonstrated two numeric techniques (an iterative and a parameter regression procedure) for computing $\phi$ given $|f|$ and $|F|$, and he encountered an ambiguity between $\phi(x)$ and $\phi_0-\phi(-x)$ only, where $\phi_0$ is a constant. Fienup\textsuperscript{68} developed an algorithm employing only $|f|$ and very general information about $F$ (e.g., that the unaberrated distant image is characterized by $F > 0$). He obtained impressive results and was able to compute complicated two-dimensional $\phi(x,y)$ distributions with no ambiguity. Theoretical results have also been reported which indicate that phase ambiguity is much less of a problem in two dimensions than in one dimension.

In many adaptive optical applications it may not matter if there are unresolvable ambiguities in $\phi$. For example, if all possibilities (e.g. $2^M$) can be enumerated, one could in principle successively try them until a figure of merit, such as Strehl ratio in the $F$ plane, is optimized or at least exceeds a threshold. If optimization of the figure of merit
only occurs for the correct $\phi(x,y)$ distribution, this is then a procedure to determine $\phi$ uniquely. However, if many $\phi(x,y)$ distributions optimize the desired figure of merit, any of them is sufficient for that application.

Actually, the phase problem, i.e., can $\phi$ be uniquely determined from $|f|$ and/or $|F|$, is often an academic issue in optics, because practical approaches for the unique computation of $\phi$ from intensity measurements $^{128,147,148}$ have been demonstrated. These approaches generally involve introducing known modifications into $f$ or $F$ and measuring the modified intensities $^{147,148}$ $|f'|^2$ or $|F'|^2$. For example, a tilted reference beam $e^{-j\omega_0 x}$ can be added to $f$ and the resulting hologram can then be digitally reconstructed in a computer. More specifically, the detected intensity can be numerically Fourier transformed (e.g. with a 2-D FFT); resulting in an array of complex numbers representing $G(u,v)=(a^2+b^2)\delta(u,v)+bf(u+u_0,v)+bf(u-u_0,v)$. Assuming that $F(u,v)$ is effectively bandlimited to the range $-u_B<u<u_B$, all values of $G(u,v)$ outside of the strip $u_0-u_B<u<u_0+u_B$ can be zeroed. The inverse Fourier transform is then an array of complex numbers, which is a unique approximation to $a(x,y)e^{j\phi(x,y)}$. In image compensation applications (e.g. astronomy), a similar procedure can be applied directly to $|f|$ when a "reference" source, unresolved by the input aperture, happens to exist in the same isoplanatic patch as the image of interest.

Another technique involves digital reconstruction of off-axis Zernike phase-contrast interferograms. More specifically, a diffraction-limited opaque filter is placed off-axis in the F plane. The F plane is followed by a second lens that transforms the signal back to a modified image, which includes a tilted reference wave. The intensity of this interferogram is recorded. Another interferogram is then recorded with
the off-axis opaque spot in the F plane rotated 180° about the origin. Finally, the intensity of |f| is recorded. These three spatial intensity distribution then allow unique computation of ae^jφ.

Based on a procedure suggested by Kohler and Mandel, an interesting one-dimensional approach may be practical. The f(x,y) plane, i.e. input aperture, is sampled by a slit, \( \pi(\frac{y-y_0}{y}) \), which extends over the full aperture in the x direction and is too narrow to resolve any details in the y direction. The F(u,y) plane, i.e. 1-D Fourier plane, is multiplied by a filter with transmission e^{uy}; and a succeeding optical Fourier transform of this plane results in the new field f'(x,y). The phase \( \phi(x,y_0) \) is then determined by a computation (e.g. digital) of the form

\[
\phi(x,y_0) = -\frac{\partial}{\partial y_0} \int_0^\infty dx' \ln(|f'(x,y)|) \bigg|_{y=y_0}
\]

Other less direct computational phase estimation procedures include the Michelson interferometry or aperture synthesis approach mentioned in Section 1.2 and the iterative algorithms mentioned above and in Section 1.2. It should also be noted that in the single-detector "image sharpening" and "multidither" approaches, the array of drive signals to the modulator is often an estimate of the spatial phase aberration to be compensated. Finally, it should be mentioned that Robinson and Dyson have theoretically discussed optimum phase estimation structures, which minimize the mean-squared error of the estimate in closed-loop adaptive systems.

Summarizing, a large variety of optical and electronic techniques have been developed for improving phase measurement performance beyond the limited capabilities of elementary interferometry. These can be broadly classed as: ac systems, in which the phase of an ac interferometric signal is electronically detected; systems which electronically reconstruct
wavefront phase from a spatial array of slope measurements; systems
in which phase is calculated from spatial arrays of dc intensity measure-
ments; and a variety of systems specifically developed for adaptive optics,
such as "multidither" techniques.

2.1.3 Real-time high-resolution phase measurement

High-resolution phase measurement presents the challenge of attaining
the high performance of the sophisticated systems discussed above, but
in configurations uncomplicated enough to allow a million or more
parallel resolution elements, or better yet all-optical implementations,
to be realized.

Configurations involving millions of parallel hard-wired electronic
channels interconnecting a detector array to a light modulator array are
clearly not worth pursuing. One compromise for implementing the phase
estimation box in Fig. 1.1 is to employ an internally scanned detector,
such as a CCD array or a vidicon, to drive a single electronic channel.
This channel may, for example, contain a computer and could implement
one of the phase estimation techniques of section 2.1.2. The final
phase estimate could drive a scanned-electron-beam addressed spatial
amplitude modulator, e.g. an Eidophor tube, to produce a transmission
proportional to \( \sqrt{\phi} \) (i.e. intensity proportional to \( \phi \)). More efficiently,
the phase estimator and modulator functions in Fig. 1.1 could be united
by having the serial channel drive an electron beam addressed spatial
phase modulator directly. Actually, this serial scheme is not usually an
attractive alternative; because as the number of resolution elements
increases, the available time per element generally becomes impractically
short. For example with a phase fluctuation bandwidth of 1 KHz and a
thousand resolution elements, the time per estimate computation is
is already below one microsec!

A superior approach would seem to be to increase the sophistication of the all-optical system as much as possible and attempt to limit any active processing to only optically-addressed spatial light modulators (SLMs) similar to the MSLM of Fig. 1.2.

Figure 2.1a depicts one possible, although nonoptimal, system. A homodyne interferogram with a tilted reference beam (i.e. a hologram) is incident on a hard-clipping SLM, whose reflected output amplitude modulation \( t_M \) as a function of input interferogram intensity \( I_D \) is:

\[
t_M = \begin{cases} 
1 & \text{for } I_D \geq I_t \\
0 & \text{for } I_D < I_t 
\end{cases}
\]

where \( I_t \) is a spatially uniform threshold. (The MSLM can be operated in a hard-clipping mode). The readout beam \( E'_t \) (coherent or incoherent) is modulated by the resulting square-wave fringes and passed through a sawtooth amplitude filter with period \( P = \lambda / \sin \theta \). This is a "detour-phase" approach which utilizes the fact that the position, \( x \), of a fringe is proportional to phase. For example, fringe maxima occur at \( x = (\phi - 2n\pi)P/2\pi \). The output phase-image intensity is approximately \( I_\phi(x,y) = A(\phi(x,y) + B) \), where \( A \) and \( B \) are constants. (This assumes that \( P \) is much less than the smallest spatial details in \( \phi \)). This system removes quadrant ambiguity and diminishes amplitude fluctuation effects, but it suffers from \( 2n\pi \) ambiguities and some residual contamination from amplitude fluctuations. Amplitude variations cause the fringe width to vary (one possible solution may be to set the threshold to obtain very narrow fringes which can be low-pass filtered to a uniform width).

Sprague and Thompson\textsuperscript{129} describe an interesting system which is depicted in Fig. 2.1b. That system employs optical differentiation and
General notation:

\[ I_D(x,y,t) = \text{Intensity seen by detector D.} \]

\[ t_M(x,y,t) = \text{Amplitude modulation produced by modulator M. (The reflected readout beam is multiplied by } t_M). \]

\[ E_i, E_r = \text{Optically-addressed, reflective-readout spatial light modulator. (e.g. MSLM)} \]

\[ E_i, E_r = \text{Input wavefront whose phase is to be measured.} \]

\[ E_r', E_r'' = \text{Flat reference wavefronts.} \]

\[ BS = \text{Beamsplitter} \]

\[ m = \text{mirror} \]

\[ L = \text{lens} \]

\[ f = \text{focal length of a lens.} \]

a) Detour-phase system for phase estimation. (H is a grating with a periodic-sawtooth amplitude transmission profile).

\[ H_1(x) \text{ and } H_2(x) \text{ are transmission filters:} \]

\[ H_1(x) = (x_0 - x)B \]

\[ H_2(x) = \begin{cases} iB_2 & \text{for } |x| < x_1 \\ B_3/x & \text{for } |x| \geq x_1 \end{cases} \]

b) Differentiation and integration system for phase estimation.

Fig. 2.1 Some techniques for high-resolution phase estimation.
integration to visualize phase without quadrant or $2\pi$ ambiguities. Unfortunately, it cannot tolerate amplitude fluctuations and ignores phase variations which are exactly perpendicular to the 1-D filters $H_1$ and $H_2$. Assuming a phase-only input $E_1 = e^{i\phi}$, differentiation and detection results in $I_D = |d\phi/dx|^2$. The SLM then produces amplitude modulation

$$t_M = B_0 \sqrt{I_D} = B_0 d\phi/dx,$$

which after integration results in $E_\phi = B_0 \phi + B_1$.

Here $B_0$ and $B_1$ are constants, and with large $B_1$ ($B_1 > 2\pi$ is realizable) $I_\phi = B_1^2 + B_1 B_0 \phi$ results. (The effective dynamic range for phase estimation is $B_1/B_0 > 2\pi$).

Other sophisticated systems for phase measurement have been reported in the literature, such as a Hilbert transform system$^{151}$, which may have practical utility for high-resolution phase estimation. It may be possible to optically implement some of the computational approaches mentioned in section 2.1.2, such as Kohler and Mandel's exponential filtering technique$^{147}$ or some of the iterative algorithms$^{68,145}$. In fact, an optical feedback system has been proposed$^{152}$ for implementing a similar iterative algorithm. There are also a variety of techniques for interferometry with a dynamic range in excess of $2\pi$ (see bibliography in appendix 6.9).

There are a few techniques for possibly generating the $t_M = \sqrt{I_D}$ nonlinearity required by the system in Fig. 2.1b$^{152-155}$. For example, a hard-clipping SLM could be employed in conjunction with a new half-tone thresholding technique$^{153-155}$. This half-tone technique is quite general and may even be generalizable to other useful nonlinearities, such as $t_M = \sqrt{\tan^{-1}I_D}$ or $t_M = 1/\sqrt{I_D}$. The latter nonlinearity would allow wavefront amplitude fluctuations to be removed, as depicted in Fig. 2.1c. (Another means for obtaining the same behavior would be to find
c) System for amplitude-compensated phase estimation.
(B is a spatial and temporal constant. BS₂ and m₁ are required so that the image at the modulator is not mirror-inverted relative to the image at the detector. R is a quarter-wave plate, which allows BS₃ to be a polarizing beamsplitter for higher efficiency.)

d) Hybrid amplitude-compensated phase estimator. (eSLM₁ is an electron-beam addressed spatial light modulator. SF₁ is a spatial filter.)

Fig. 2.1 Some techniques for high-resolution phase estimation.
a material, e.g. photochromic, which darkens with $I_D$, such that its transmission varies as $I_D^{-1/2}$). This $I_D^{-1/2}$ filtering could be extremely useful as a preprocessor to remove the amplitude dependence from a phase visualization system, such as that in Fig. 2.1b or an elementary interferometer. Unfortunately, this technique is extremely inefficient, since it attenuates all input intensities to the level of the lowest intensity which the modulator can sense, (at which point $t_M$ is at its maximum value of unity).

Casasent\textsuperscript{156} described an alternative system for relatively high-resolution phase estimation without amplitude fluctuations, which is sketched in Fig. 2.1d. A tilted-reference homodyne interferogram is recorded by a TV camera. The resulting sinusoidal-fringe serial signal is hardclipped to become a square-wave and normalized in width with a monostable element. The resulting square-wave grating is written on an electron-beam-addressed spatial amplitude modulator (eSLM\textsubscript{1} in Fig. 2.1d). This grating produces multiple order reconstructions of a constant amplitude wave with the phase to be measured. One order is isolated by the spatial filter SF\textsubscript{1}, and the phase signal $I_\phi = B_0^2 + B_1 B_0 \phi$ is produced by the optical differentiation and integration system of Fig. 2.1b.

The method of Fig. 2.1e is a potentially superior means for removing amplitude fluctuations from the phase estimate. There, an external source is employed to read out the SLM with gain. This is possible since: 1) amplitude correction is implemented after the interferogram has already been formed, so that the phase of the readout beam is inconsequential; and 2) all terms in the interferogram have the same amplitude dependence. These conditions hold in Fig. 2.1e, because the interferometer reference
e) Amplitude-compensated conjugate-interferometer phase visualization. 
\[ I_\phi = B(\alpha_1^2 + \alpha_2^2 + 2\alpha_1 \alpha_2 \cos(2\phi_i + \phi_a)) \]
where \( B, \alpha_1, \) and \( \alpha_2 \) are spatial and temporal constants. \( \phi_a \) is the phase of \( \alpha_1 \alpha_2^* \). \( B \) is real. \( M_5 \) removes mirror inversion.

Using stimulated Brillouin Scattering

Using 4-wave mixing

Using 3-wave mixing

Using the photorefractive effect

48, 157, 158, 252, 253

f) Nonlinear-mixing conjugate interferometers. (In each case a conjugate beam is reflected from, and/or transmitted through, the nonlinear material. \( \alpha, \alpha', \alpha'', \) and \( \alpha''' \) are constants.)

Fig. 2.1 Some techniques for high-resolution phase estimation.
beam is the conjugate of the input beam. The conjugate wave is generated by reconstructing a real-time hologram in SLM₁. A strong Eᵣ beam provides effective optical gain.

Nonlinear optical processes provide an alternative to an SLM for obtaining the conjugate of the input wave. A large variety of suitable nonlinear processes have been experimentally demonstrated⁴⁸, (see bibliographic comments in appendix G.16). These include: stimulated Brillouin scattering, four-wave mixing, three-wave mixing (optical parametric amplification), the photorefractive effect, and a photon-echo effect. In fact, self-reference interferometry by conjugate wavefront generation has been reported in the literature¹⁵⁷,¹⁵⁸. Some potential nonlinear mixing geometries for generating the Eᵢ+Eᵣ interferogram are sketched in Fig. 2.1f.

Actually, all of the high-resolution phase estimation systems outlined so far are unnecessarily complicated, because closed-loop configurations similar to Fig.1.1 possess inherent properties which simplify phase estimation. By exploiting these closed-loop properties, an alternative phase estimation technique called an Interference Phase Loop⁶⁹,⁷⁰ (IPL) was developed. More specifically, in the Interference Phase Loop technique the output intensity from an elementary interferometer (e.g. Zernike phase-contrast, homodyne, heterodyne, polarization, shearing, or one of the others in Sec. 2.1.1) is detected and fed back to drive a phase-only modulator in the path of the wavefront being measured. With an elementary interferometer as the phase estimator, Fig. 1.1 becomes an example of an interference phase loop. An "all-optical", high-resolution implementation of the IPL, employing homodyne interferometry and the MSLM
as the spatial phase modulator (reflection readout), is depicted in Fig. 2.2.

It will be shown that in operation, the drive signal to the IPL phase modulator, and in many cases the output intensity from the elementary interferometer itself, become representations of the phase estimate. It will also be seen that the IPL modulates the wavefront being measured with its conjugate. The IPL technique is thus particularly well suited to high-resolution phase compensation applications.

The IPL technique is much simpler to implement and enjoys performance superior to most other existing phase estimation techniques. It is shown both theoretically and experimentally in subsequent sections of this thesis that the interference phase loop largely ignores amplitude fluctuations, has no phase quadrant ambiguity, and can continuously estimate phase over multiple \( \pi \) radians of dynamic range. An elementary interferometer can thus assume the role of the phase estimator in Fig. 1.1, without imposing the limitations of elementary interferometry on overall system performance. When self-interference is employed (e.g. Zernike or Schlieren phase-contrast, or shearing or polarization interferometry), strictly monochromatic light is not required; this may extend the usefulness of this phase sensing technique to a variety of astronomical and image compensation applications.

Although this thesis will focus on issues related to the IPL technique, it should be noted that many of the conjugate-generation processes in Figs. 2.1 e and f (e.g. real-time holography, nonlinear optical mixing) may provide an alternative means for high-resolution phase compensation. These conjugate processes often produce an output proportional to \( E^* E_r \), where \( E_r \) is a pump or reference wave. When
Fig. 2.2 An "all-optical" high-resolution Interference Phase Loop
E_r=E_i$, a phase-compensated wave proportional to $E_i^*E_i$ can result. Appendix F illustrates some potential geometries for implementing this phase compensation technique. Some of these configurations are related to the IPL. For example, the closed-loop system of Fig. F.2 becomes operational and is a homodyne IPL when $E_r$ is coaxial with the feedback beam ($\theta=0$) and a phase-only SLM is employed. Some of the major advantages and disadvantages of the nonlinear compensation techniques, relative to the IPL, are mentioned in appendix F.
2.2 The Interference Phase Loop

Reiterating, in the Interference Phase Loop phase estimation and compensation technique, the output intensity from an elementary interferometer is detected and used to drive a phase-only modulator in the path of the wavefront being measured.

Whereas most previous researchers\textsuperscript{2,3,4,5,140} viewed closed-loop configurations similar to Figure 1.1 as a means for keeping the error to the phase detector small; a major contribution of this work is the realization that the closed-loop forms an integral part of, and greatly enhances and simplifies the phase estimator. For example, without any additional complexity (the acq is also not required), the system in Fig. 2.2 can produce an intensity distribution in plane B which is a real-time continuous (no fringes) grey-tone-image representation of the phase of $E_i$ over multiple $\pi$ radians; even when the amplitude of $E_i$ varies widely in space and time. The closed-loop phase measurement system\textsuperscript{140} mentioned in section 2.1.2, which employed an ac interferometer with feedback to a phase modulator in its local oscillator beam, resembles the IPL; but that system employed very complicated analogue and digital electronics, rather than exploiting the inherent properties of optical phase feedback.

McDowell's\textsuperscript{2} four-channel heterodyne communications receiver with electronic feedback to a spatial phase modulator in the L.O. beam is related to the IPL. However, he did not exploit the inherent capabilities of this technique. He constrained his system to constant beam amplitude, small initial phase error, and less than $2\pi$ radians of dynamic range.

Principles similar to those employed in the IPL are implicit in some of the adaptive-optical, transmitter-predistortion systems with
feedback derived from a target reflection, e.g. COAI\textsuperscript{18-21} and Cathey et al.\textsuperscript{22,23}. Aside from a very different geometry, the above mentioned systems employ synchronous or heterodyne detection and more complicated electronics to separate the phase from amplitude variations. The IPL is also somewhat related to Dicke's\textsuperscript{13} proposed astronomical image compensator with digital feedback, closed-loop bistable Fabry-Perot systems\textsuperscript{159}, and image processors with feedback\textsuperscript{160}.

One may also view the IPL as a phase-locked loop directly tracking optical phase. In fact the specific implementation depicted in Fig. 3.13 is essentially identical to a conventional electronic phase-locked loop.

2.2.1 General formulation for modelling IPL behavior

A variety of analysis techniques can be employed to determine such essential operating features of the IPL as: steady state error, dynamic response, bandwidth, stability, phase tracking range, and the effects of noise, amplifier threshold and saturation, and signal amplitude fluctuations. For small errors, classical control theory\textsuperscript{161} can be applied to the linearized system. Modern control theory\textsuperscript{162} is particularly relevant, and there are a variety of applicable techniques from the theory of nonlinear systems\textsuperscript{163}. There is also a great wealth of relevant phase-lock loop literature\textsuperscript{164,165}.

In the following sections, the homodyne implementation will be utilized to illustrate the basic static and dynamic operational characteristics of the IPL. In Chapter 3 a few aspects of IPL behavior will be discussed in more detail and the analysis will be extended to include the effects of noise and the additional features introduced by the Zernike phase-contrast and heterodyne implementations. (The implementation is classified by the type of elementary interferometer employed in the loop).
The generalized homodyne-heterodyne implementation of the IPL can be represented by the schematic system of Fig. 2.3. The spatially distributed detector (DET), amplifier (AMP), and phase modulator (MOD) combination models either a discrete channel hard-wired system or an "all-optical" system employing an SLM. For example, in the specific case of the MSLM: DET corresponds to the photocathode, AMP to the MCP, MOD to the electrooptic crystal, and the array of electronic links to the spatially distributed electron current.

It is assumed that the incident field, \( E_i(x,y,t) = |E_i(x,y,t)| \exp[j(\phi_i(x,y,t)-w_i t)] \), is quasi-monochromatic and that the system resolves (continuous SLM), or satisfies the Nyquist sampling theorem (discrete wired system) on, the spatial bandwidth of \( E_i \). For atmospheric phase compensation the required spatial resolution is approximately the "coherence cell" size.

In Fig. 2.3, the modulator imparts a constant phase shift \( \phi_0(x,y) \) and a controlled phase shift \( \phi_m(x,y,t) \) to the input wave. The spatial dependence of \( \phi_0(x,y) \) is due to modulator nonuniformities; e.g., an electrooptic plate whose thickness is uneven. The modulated beam is assumed to suffer an attenuation \( \alpha(x,y) \) and phase shift \( \phi_1(x,y) \) in traveling from the modulator to the detector. At the face of the detector (z=0 plane), the local oscillator (L.O.) field, \( E_L = |E_L(x,y,t)| \exp[j(\phi_L(x,y,t)+\pi/2-w_L t)] \), and modulator field, \( E_m = |E_i(x,y,t)| \exp[j(\phi_i(w,y,t)-\phi_0(x,y)-\phi_m(x,y,t)-\phi_1(x,y)-w_i t)] \), combine to produce the intensity, \( I(x,y,t) \). Suppressing the spatial and temporal dependencies,
Figure 2.3 General homodyne-heterodyne implementation of the IPL

\[ E_{l} = |E_{l}|e^{j\left(\phi_{l} + \pi/2 - \omega t\right)} \]

\[ E_{m} = \alpha_{l} |E_{l}|e^{j\left(\phi_{m} - \phi_{l} - \phi_{i} - \omega t\right)} \]

Spatial array of electronic links

L.O. BEAM

\[ E_{l} = |E_{l}|e^{j(\phi_{l} + \pi/2 - \omega t)} \]

\[ E_{i} = |E_{i}|e^{j(\phi_{i} - \phi_{l} - \omega t)} \]

\[ I_{o} + I_{i} \sin \phi_{e} \]

AMP \[ g_{1}H(s) \]

DET \[ g_{0} \]
\[ I = I_0 + I_1 \sin \phi_e \quad (2.1) \]

where

\[ I_0 = \frac{1}{2Z_0} \left( \alpha^2 |E_i|^2 + |E_s|^2 \right) \quad (2.2a) \]

\[ I_1 = \frac{1}{2Z_0} (2\alpha |E_i||E_s|) \quad (2.2b) \]

and the loop phase error is

\[ \phi_e = \phi_i - \phi_m - \phi_0 - \phi_s - (w_i - w_s) t \equiv \phi_i - \phi_m - \phi_b \quad (2.3) \]

\[ Z_0 = \sqrt{\mu_0 / \varepsilon_0} \] is the intrinsic impedance of free space. In obtaining Eqs. (2.1) - (2.3), the fact that \( E_i, E_s, \) and \( E_m \) are vector fields has been neglected, since they are assumed to be polarized parallel to each other. In addition, the fact that the actual physical fields are the real part of the phasors \( E_i, E_s, \) and \( E_m \) does not affect these results, since the detector is assumed insensitive to optical carrier frequencies, (e.g. \( w_i, w_s, 2w_i, 2w_s, w_i + w_s \)).

The phase \( \phi_b \), defined by Eq. (2.4), is the intrinsic background phase of the system. In the homodyne case \( \phi_b \) is time independent, so it is convenient to take \( \phi_b = \phi_i + \phi_0 + \phi_s \) as the zero reference of phase by introducing \( \phi'_i \)

\[ \phi'_i = \phi_i - \phi_b \quad (2.4) \]

Note that \( \phi'_i = 0 \) when the phase being measured, \( \phi'_i \), is equal to \( \phi_b \). The loop phase error, Eq. (2.3), can now be written

\[ \phi_e = \phi'_i - \phi_m \quad (2.5) \]

For generality, the detector in Fig. 2.3 is modelled as exhibiting threshold and saturation by
\[ i_o = \begin{cases} 0 & I < I_t \\ g_o (I - I_t) & I_t \leq I \leq I_s \\ g_o (I_s - I_t) & I > I_s \end{cases} \] (2.6)

Here \( i_o \) is the detector output signal and \( g_o \) is a proportionality constant; for example if \( i_o \) is a current, \( g_o \) is the detector responsivity times the area of a resolution cell (amps \( \times \) \( \text{m}^2/\text{watts} \)). Such physical devices as the MSLM (see eqs. (4.32), (4.38c), (4.18)) or a photodiode driving an operational amplifier (see Eq. (5.2a)) can fit into this model, exhibiting saturation and an adjustable threshold.

It will be seen that detector threshold often enhances IPL performance.

Most physical phase modulators also have a limited dynamic range; assuming that \( 0 \leq \phi_m \leq \phi_s \),

\[ \phi_m = \begin{cases} 0 & i_1 \leq 0 \\ g_2 i_1 & 0 < i_1 < \phi_s / g_2 \\ \phi_s & i_1 \geq \phi_s / g_2 \end{cases} \] (2.7)

Here \( i_1 \) is the modulator drive signal and \( g_2 \) is a constant of proportionality (for example, radians/volt). The frequency response of the detector and modulator can be lumped into the amplifier transfer function, \( \tilde{H}(s) \), where the tilde is used to represent the Laplace Transform. In the Laplace domain,

\[ \tilde{y}_1(s) = g_i \tilde{H}(s) \tilde{y}_0(s) \] (2.8)

For most systems of interest \( \tilde{H}(s) \) can be adequately expressed as the ratio of polynomials

\[ \tilde{H}(s) = \frac{\sum_{i=0}^{N_k} k_i s^i}{\sum_{i=0}^{N_k} \ell_i s^i} \] (2.9)

(usually \( N_\ell < N_k \))
Owing to system nonlinearities, it is easier to analyze the loop's operation in the time domain. Employing \( \frac{d^i}{dt^i} \) as an operator representing \( i \)th order differentiation, the transform of \( \hat{H}(s) \) can be written:

\[
\hat{H}(s) \leftrightarrow \sum_{i=0}^{N_k} \frac{k_i d^i}{\sum_{i=0}^{N_k} l_i d^i} \]

Equations (2.1) and (2.6)-(2.10) result in the following differential equation for the generalized IPL; when the detector and modulator are in their linear regions:

\[
(N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i}) \phi_m = g_0 g_1 g_2 (N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i})(I_0 - I_t + I_1 \sin \phi_e) \quad (2.11)
\]

Using Eq. (2.5), Eq. (2.11) can be rewritten as differential equations for the two major system observable phases \( \phi_m \) and \( \phi_e \):

\[
(N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i}) \phi_m = (N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i})(G_0 - G_1 \sin(\phi_m - \phi_e)) \quad (2.12)
\]

\[
(N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i}) \phi_e = (N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i}) \phi_m - (N_k \sum_{i=0}^{N_k} \frac{d^i}{dt^i})(G_0 + G_1 \sin \phi_e) \quad (2.13)
\]

These are the fundamental equations describing IPL operation. Using Eqs. (2.2) and (2.11), the lumped variables \( G_0 \) and \( G_1 \) can be written:

\[
G_0 = g_0 g_1 g_2 (I_0 - I_t) = \frac{1}{2 z_0} g_0 g_1 g_2 [(a^2 |E_i|^2 + |E_\perp|^2) - I_t^2] \quad (2.14a)
\]

\[
G_1 = g_0 g_1 g_2 I_1 = \frac{1}{z_0} g_0 g_1 g_2 |E_i| |E_\perp| \quad (2.14b)
\]
Here $I'_t = 2Z_{0} I_{t}$. The system operational limits of Eq. (2.6) impose the constraint:

$$0 \leq G_{0} + G_{1} \sin \phi_{e} \leq G_{s} = g_{0} g_{1} g_{2} (I_{s} - I_{t})$$  

(2.14c)

In Eqs. (2.14) $g_{0} g_{1} g_{2}$ has units of radians/(watts/cm²), $\hat{H}$ in Eqs. (2.8) through (2.9) is dimensionless, and $G_{0}$ and $G_{1}$ are expressed in radians. It will be seen in chapters 4, 5, and 6 that the detector, modulator, and amplifier models of Eqs. (2.6)-(2.8) can conform to a wide variety of physical devices, with the proper choice of parameter values. Typical values of $G_{1}$ ranged from $2\pi$ to 50$\pi$ radians and $G_{0}$ ranged from $-2.5\pi$ to $2.5\pi$ in the experimental IPL systems demonstrated in chapter 6. As a specific example, the MSLM in the all-optical IPL of section 6.3.1 was characterized by (from Eq. (4.38b)),

$$g_{0} g_{1} g_{2} = \frac{\pi}{S} t_{w} = 5 \times 10^{9} \text{ cm}^{2} \text{ radians} \text{ watt}$$

Here $S_{\pi}$ is the half-wave exposure sensitivity, i.e. power in Joules/cm² required for $\pi$ radians modulation, and $t_{w}$ is the system integration time. That system saturated at $I_{s} = 2.5$ nW/cm², corresponding to a maximum $G_{1}$ of $G_{s} = 12.5\pi$. (By adding a second MCP, the MSLM gain should be increaseable to $g_{0} g_{1} g_{2} = 1.7 \times 10^{12} t_{w} \pi$). Specific hardware-determined expressions for the single-channel and nineteen-channel IPL experiments are given in Eqs. (5.5) and Eqs. (5.15)-(5.17) respectively.

In order to further evaluate the behavior of the IPL, it is necessary to specify $\varepsilon_{i}$ and $k_{i}$ of the system transfer function, $\hat{H}(s)$. 
2.2.2. **Basic Static Behavior of the IPL**

The simplest system transfer function is $\hat{H}(s)=1$, which unrealistically implies that the system components have an infinite bandwidth. Eqs. (2.12) and (2.13) then result in the following implicit relations for $\phi_m$ and $\phi_e$:

$$\phi_m = G_0 - G_1 \sin (\phi_m - \phi_i')$$  \hspace{1cm} (2.15)

$$\phi_e = \phi_i' - G_0 - G_1 \sin \phi_e$$  \hspace{1cm} (2.16)

More realistically, if a low-pass filter ($\hat{H}(s) = \frac{1}{\lambda_1 \lambda_2 / [(s+\lambda_1)(s+\lambda_2)]}$), two-pole filter ($\hat{H} = \lambda_1 \lambda_2 / [(s+\lambda_1)(s+\lambda_2)]$), or other type-zero* transfer function is assumed in Eqs. (2.12) and (2.13), and all the derivatives of $\phi_m$ or $\phi_e$ respectively are set to zero; the steady-state relations of Eqs. (2.15) and (2.16) result. This will be demonstrated more rigorously in the next section and in Chapter 3. Since many real SLMs or discrete component systems are adequately described by a type-zero transfer function; equations (2.15) and (2.16) describe the general steady-state behavior of a large class of IPL implementations.

---

*A type n transfer function is of the form:161*

$$\hat{H}(s) = \frac{1}{s^n} \prod_{i=1}^{N} \frac{1}{(s-b_i)}$$

$$\prod_{i=1}^{N} \frac{1}{(s-a_i)}$$

where all $a_i$ and $b_i$ are nonzero.
It is convenient to rewrite Eqs. (2.15) and (2.16) as

\[ \phi_m = \phi_i' - \sin^{-1}\left(\frac{\phi_m - G_0}{G_1}\right) - 2n\pi = \phi_{mss} \]  

(2.17a)

or

\[ \phi_m = \phi_i' + \sin^{-1}\left(\frac{\phi_m - G_0}{G_1}\right) + (2n-1)\pi \]  

(2.17b)

and

\[ \phi_e = \sin^{-1}\left(\frac{\phi_m - G_0}{G_1}\right) + 2n\pi = \phi_{ess} \]  

(2.18a)

or

\[ \phi_e = -\sin^{-1}\left(\frac{\phi_m - G_0}{G_1}\right) + (2n-1)\pi \]  

(2.18b)

(Here \(-\frac{\pi}{2} \leq \sin^{-1}x \leq \frac{\pi}{2}\))

It is apparent from the above results that with large gain \((G_1 \to \infty)\), the modulator phase estimates the input phase within an integer multiple of \(\pi\), i.e., \(\phi_m = \phi_i' + n\pi\). The error phase, \(\phi_e = n\pi\), is the error of this estimate. Actually, if the system bandwidth is not infinite, one would expect the \((2n-1)\pi\) solutions to be unstable, since a small deviation of \(\phi_m\) from equilibrium in Eq. (2.17b) results in an increased deviation. This conclusion is verified by the dynamic analysis in the next section.

Thus with large gain, the modulator phase estimates the input phase modulo \(2\pi\); i.e., \(\phi_m = \phi_i' + 2n\pi\).

With smaller gain, the IPL behavior is not so obvious since Eqs. (2.17) and (2.18) are still implicit transcendental equations in \(\phi_m\) or \(\phi_e\). Of particular interest is the behavior of the dependent observables \(\phi_m\) and \(\phi_e\) as a function of the independent input variable \(\phi_i'\). Graphical techniques provide a powerful tool for finding and studying the solutions to these implicit transcendental equations, particularly with such
complications as cutoff and saturation. In Fig. 2.4 $\phi_{m_1} = \phi_1^i - \phi_e$ and $\phi_{m_2} = G_o + G_1 \sin \phi_e$ (from Eqs. (2.5) and (2.15)) are plotted as a function of $\phi_e$. Here $\phi_{m_2}$ has been truncated to include the effects of limited modulator range, as described by Eq. (2.7). Solutions occur where $\phi_{m_1}$ and $\phi_{m_2}$ intersect each other. By shifting $\phi_{m_1}$ vertically with $\phi_1^i$, the complete equilibrium behavior of $\phi_m$ vs. $\phi_1^i$ (Fig. 2.5) or $\phi_e$ vs $\phi_1^i$ (Fig. 2.6) can be determined. The unstable $(2n-1)\pi$ trajectories have been omitted from Fig. 2.5. In Fig. 2.5 it was really only necessary to represent $-\pi \leq \phi_1^i \leq \pi$, since $\phi_m$ is periodic in $\phi_1^i$. This can be seen by inserting $\phi_1^i + 2n\pi$ into Eq. (2.15).

Notice in Figs. (2.4) & (2.5) and Eqs. (2.17) & (2.18) that, for a given $\phi_1^i$, multiple stable values of $\phi_m$ are possible. This multistable behavior will be discussed in more detail later in this chapter and in chapter 3. As $\phi_1^i$ continuously increases or decreases the physical limits of the modulator are ultimately encountered and then $\phi_m$ eventually jumps to another stable equilibrium in the useful operating region. This behavior is represented in Fig. 2.5 by the dotted one-way vertical segments and in Fig. 2.4 by the $\phi_{m_1}$ curve traversing and jumping off the end of a flat portion of the $\phi_{m_2}$ curve. With imperfect modulator limits, e.g. curvature in the "flat" parts of $\phi_{m_2}$ in Fig. 2.4, it will be shown in section 2.4.3 that the jump to another trajectory occurs at the point where $\phi_{m_1}$ just separates from $\phi_{m_2}$. That point also divides the stable and unstable sides of $\phi_{m_2}$.

Equations (2.15) or (2.16) could have been solved by other graphs such as: the intersection of $\phi_{m_1} = \phi_m$ and $\phi_{m_2} = G_o + G_1 \sin(\phi_1^i - \phi_m)$ both plotted as a function of $\phi_m$; or $\phi_e = \phi_e$ and $\phi_{e_2} = \phi_1^i - G_o - G_1 \sin \phi_e$ both versus $\phi_e$. 
Fig. 2.4 Graphical solution for steady-state behavior of the IPL

Equation are at the intersections of $\phi_m = \phi_i - \phi_e$ and $\phi_m = G_0 + G_1 \sin \phi_e$.

The stable solutions are circled.

($\phi_i = 2\pi$, $G_0 = 1.8\pi$, $G_1 = 7.3\pi$, $\phi_e = 3.8\pi$)
Fig. 2.5 Phase estimation performance ($\phi_m$ versus $\phi_i'$)
Fig. 2.6 Phase estimation error as a function of input phase ($\phi_e$ versus $\phi_i$).
(For the trajectory through point A in Fig. 2.5)
The parts of Figure 2.5 for which the modulator is in its active region could also have been reverse plotted from

\[ \phi_i' = \phi_m - \sin^{-1}\left(\left(\phi_m - G_0/G_1\right)/2\pi\right) - 2n\pi \]  

(2.19)

Equation (2.17a) can also be solved numerically for \( \phi_m \), given \( \phi_i' \), by substituting a value of \( \phi_m \) in the left side and iteratively resubstituting the refined \( \phi_m \) value from the right side until \( \phi_m \) converges.

The phase tracking behavior of the IPL can be described more precisely in terms of \( \frac{d\phi_m}{d\phi_i} \), which is the incremental phase estimation transfer function, and \( \frac{d\phi_e}{d\phi_i} \), which is the sensitivity of the phase error to the input phase. Differentiating Eqs. (2.17a) and (2.18a) yields:

\[ \frac{d\phi_m}{d\phi_i} = \frac{G_1a}{1 + G_1a} \]

\[ \frac{d\phi_e}{d\phi_i} = \frac{1}{1 + G_1a} \]

where \( a = \sqrt{1 - \left(\frac{\phi_m - G_0}{G_1}\right)^2} = \cos \phi_e \)

These results are the same as those for a linear type-zero control system with gain \( G_1 \). However, in this system the gain decreases as the error increases. With large gain, \( G_1 \gg G_0 \), these expression exhibit the ideal behavior:

\[ \frac{d\phi_m}{d\phi_i} \rightarrow 1 \]

\[ \frac{d\phi_e}{d\phi_i} \rightarrow 0 \]
Insensitivity to amplitude fluctuations

It is evident from Figure 2.4 that the system exhibits the desired behavior of much greater sensitivity to phase than amplitude. Amplitude fluctuations cause the offset $G_0$ and gain $G_1$ to vary (Eqs. (2.14)). Amplitude induced variations in $G_0$ cause $\phi_{m2}$ in Figure 2.4 to move up and down relative to $\phi_{m1}$; this does not effect large gain operation as long as cutoff or saturation are not encountered. In addition, the effects of amplitude fluctuations on $G_0$ can be partially masked by other terms, such as $|E_2|^2$ in Eq. (2.14a). Provided that the gain, $G_1$, remains large, Figure 2.4 suggests that amplitude induced fluctuations in gain are also inconsequential; only the error, which is already very small, will vary. The effects of amplitude fluctuations are discussed in more detail in Chapter 3.

Optical Phase Image

It was seen that the modulator phase is an estimate of the input phase, i.e., $\phi_m = \phi_1 + 2n\pi$. However, Fig. 2.3 and Eqs. (2.14) and (2.15) show that, in a type-zero IPL, $\phi_m$ is also an amplified version of the interferometic intensity at the detector, I. Thus the intensity at C in Fig. 2.3 is also an estimate of $\phi_1$, and, with large gain, ignores variations in $|E_1|$

More specifically, substituting $\phi_e$ given by Eq. (2.18a) into I of Eq. (2.1), results in

$$I = I_0 + I_1 \frac{\phi_m - G_0}{G_1}$$

A study of Fig.2.4 reveals that the following condition always holds:

$$-1 \leq \frac{\phi_m - G_0}{G_1} \leq 1$$

(2.21)
Employing Eqs. (2.14) with Eq. (2.20) results in the following exact expression for $I$

$$I = I_t + \frac{1}{g_0 g_1 g_2} \left( \phi_i^1 - \phi_e \right) \quad (2.22a)$$

With large gain, $\phi_e \rightarrow 2n\pi$ and

$$I = I_t + \frac{1}{g_0 g_1 g_2} \left( \phi_i^1 - 2n\pi \right) \quad (2.22b)$$

Remember that $g_0 g_1 g_2$ and $I_t$ are system constants. This phase image could be recorded with high-contrast film or a detector array with threshold.

The fact that the interferometer intensity, $I = I_0 + I_1 \sin \phi_e$, is both a representation of the error $\phi_e$ (particularly for small $\phi_e$) and an unamplified version of the phase estimate (e.g., Eq. 2.23) is not a contradiction. In analogy with a classical type-zero control system, the error ($\sin \phi_e$) is amplified to drive the output $\phi_n$.

By staying on one trajectory in Fig. 2.5, the IPL is capable of giving a continuous grey-scale intensity readout of phase over a very large range (modulators capable of $> 10\pi$ radians are available); a feat almost unheard of among phase visualization techniques.

In fact, if $\phi_m$ cutoff or saturation is not encountered, the input phase variation corresponding to the full interferometer intensity range of $2I_1$ (Eq. 2.1) is extended from $\Delta \phi_i^1 = \pi$ radians in the elementary interferometer to $\Delta \phi_i^1 = 2G_1 + \pi$ radians in the IPL system. This is easy to show: using Eq. (2.21) the minimum value of $\phi_m$ is $G_0 - G_1$, at which point Eq. (2.17a) yields $\phi_i^1 = G_0 - G_1 - \frac{\pi}{2} + 2n\pi$. The maximum value of $\phi_m$, $G_0 + G_1$, results in $\phi_i^1 = G_0 + G_1 + \frac{\pi}{2} + 2n\pi$; thus the maximum variation of $\phi_i^1$ on one trajectory (constant $n$) is

$$\Delta \phi_i^1 = 2G_1 + \pi \quad (2.23a)$$
The large phase tracking range in $\phi_i$ tends to decrease the contrast of the phase image in Eqs. (2.22). However by specifying a maximum acceptable error, $\phi_{em}$, and making $G_1$ no larger than required to obtain that error, reasonable contrast can be obtained.

The modulator dynamic range, $(0 \leq \phi_m \leq \phi_s)$, is generally chosen to encompass the range of phases to be visualized, $(\Delta \phi_i = \phi_s + \phi_{em})$. Although $G_o$ can vary with input amplitude, a good choice for its average value is $G_o = \phi_s / 2$. $(I_t$ or $|E_2|^2$ can be appropriately adjusted, or possibly even controlled by an automatic gain control (agc) loop). It then follows from Eq. (2.18a) $(\phi_e = \sin^{-1}((\phi_m - G_o)/G_1))$ that the largest error, which occurs at the extremes of modulator phase $(\phi_m = 0$ or $\phi_m = \phi_s)$, is minimized, and the error is zero in the center where $\phi_m = \phi_s / 2 = G_o$. The maximum error is $|\phi_e| \leq \phi_{em} = \sin^{-1}(\frac{\phi_s}{2G_1})$ (2.23b)

Using Eq. (2.1) the maximum contrast can be written as $C = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} = \frac{I_1}{I_0} \sin \phi_{em} = \frac{I_1}{I_0} \frac{\phi_s}{2G_1}$

The last two expressions on the right specify $G_1$ and hence $g_0 g_1 g_2$ in terms of the fixed systems parameters. As a specific example $G_1 = 1.18 \phi_s$, $\lambda / 10$ accuracy, and $|E_1| = 0.8 |E_2|$ results in $C = 57\%$. With $\lambda / 20$ accuracy, $C = 30\%$.

When the phase image is detected with a device possessing intensity threshold, $I_{td}$, the intensity seen follows from Eqs. (2.22) and (2.14) as $I_d = I_t - I_{td} \frac{I_1}{G_1} (\phi_i - \phi_e) = I_t - I_{td} \frac{I_1}{G_1} \phi_m$. The resulting contrast is

$$C = \frac{I_1 \phi_s / G_1}{I_t - I_{td} + I_1 \phi_s / G_1}$$
(Here $I_{td}$ is the threshold of a spatial detector array employed to sense the phase image; this is distinct and independent of the IPL detector array with threshold $I_t$). By adjusting $I_t = I_{td}$, 100% contrast can be obtained with arbitrarily small error. Depending on the physical hardware, $I_t$ and $I_{td}$ are often adjustable parameters. If $I_{td}$ is not controllable (e.g. high-contrast film) it can be effectively adjusted by passing the interferoqram through a variable attenuator.

**Linearized Steady-State Relations**

If the gain is made large enough, $G_1 \gg \phi_m - G_0$ will hold over the whole active region ($0 \leq \phi_m < \phi_S$); and Eq. (2.19a) can be linearized to obtain an explicit approximation for $\phi_m$

$$\phi_m \approx \frac{G_1}{1+G_1} \left( \phi_i' + \frac{G_0}{G_1} + 2n\pi \right) \quad (2.24)$$

An important subtlety is that $n$ is not arbitrary. Given $\phi_m$, $\phi_i'$ and the principal value of $\sin^{-1}$, the only stable value of $n$ which satisfies Eqs. (2.17a) is that for which $|\phi_m - \phi_i' - 2n\pi| \leq \frac{\pi}{2}$. More explicitly

$$n = \text{INT} \left( \frac{\phi_m - \phi_i' + \pi}{2\pi} \right) \quad (2.25)$$

Here $\text{INT}(x)$ produces the greatest integer less than or equal to $x$, e.g. $\text{INT}(2.3) = 2$ and $\text{INT}(-1.4) = -2$. It is more informative to rewrite $\phi_m$ as

$$\phi_m = \frac{G_1}{1+G_1} \left( \phi_i' + \frac{G_0 - 2n\pi}{G_1} \right) + 2n\pi \quad (2.26a)$$

or

$$\phi_m = \phi_i' + 2n\pi + \frac{G_0 - 2n\pi - \phi_i'}{1+G_1} \quad (2.26b)$$
Equations (2.26) are very interesting in that they are linearized and at the same time preserve important subtleties of IPL behavior. For a given $\phi_1$, note that $\phi_{m_{SS}}$ is multistable at solutions near $\phi_1 + 2n\pi$ and that the size of the estimation error depends on the n value. It should be noted that the n value in Eq. (2.25) has physical meaning; for example, as n in Eq. (2.25) changes by $\pm 1$, the drive voltage to an actual phase modulator changes by $2V_\pi$, where $V_\pi$ is the voltage required for half-wave modulation. The multistability of the IPL will be discussed further in chapter 3.

Adaptive Phase Compensation with the IPL

An adaptive phase compensator should produce an output phase whose wavefront shape is independent of time and the input wavefront shape.

In the schematic IPL of Figure 2.2, the phase of the output beam at A is $\phi_A = \phi_1 - \phi_m - \phi_o$; but using Eq. (2.4) shows $\phi_A$ is also equal to

$$\phi_A = \phi_e + \phi_1 + \phi_\perp$$

(2.27a)

With large gain: $\phi_e \sim -2n\pi$ and $\phi_A$ becomes

$$\phi_A \sim \phi_1 + \phi_\perp - 2n\pi$$

(2.27b)

Thus the output wavefront at A is compensated to the same shape as the system phase $\phi_1 + \phi_\perp$. Note that in the homodyne IPL $\phi_1$ and $\phi_\perp$ are static with time and can be adjusted to any desirable shape (e.g., plane-wave) with a static phase corrector such as a hologram. Another way of interpreting this phase compensation behavior is that since a phase variation of $\Delta\phi_1 = 2G_1 + \pi$ is required to change $\phi_e$ by $\pi$, moderate changes in in $\phi_1$ ($\Delta\phi_1 < G_1$) produce only very small changes in $\phi_e$ and hence $\phi_A$.
It is also noteworthy that $\phi_A$ does not depend on the static modulator phase $\phi_0$; the IPL based phase compensator thus also corrects the spatial nonuniformities of the modulator. (As will be seen in section 3.1.2, this modulator correction will usually be modulo $2\pi$).

Amplitudes far below the average will eventually cause $G_1$ (Eq. (2.14b)) to become small, and will not be compensated well. This should not be serious, since these components make only a small constructive or destructive contribution in many applications (e.g. at the communications detector in Fig. 1.1).

In phase compensation applications, it should be noted that the physical division between the estimator and modulator of Fig. 1.1 is not really valid for the IPL. The whole closed loop, including the interferometer and modulator, is an integral part of the phase estimator.

The IPL configuration of Fig. 2.2 can also generate the wavefront conjugate of $\phi_i(x,y)$ if another beam is injected at $A$. The conjugate can be separated from the input beam with a beam splitter. Alternatively, Fig. 2.2 can be viewed as an adaptive-optical transmitter-predistorter system; where $E_i$ is a probe beam received from the target and another beam, injected at $A$, becomes the predistorted transmitter beam.

Gain

It has been seen that with "large" gain, $G_1$, the type-zero homodyne IPL accurately estimates and compensates phase, even in the presence of amplitude fluctuations. According to Eq. (2.17a), the large gain condition corresponds to $G_1 \gg |\phi_m - G_0|$. Operation with $\phi_m = G_0$ can satisfy this condition, however it is usually desirable to have small error over the full multiple-$\pi$ radian operating range of $\phi_m$ ($0 \leq \phi_m \leq \phi_S$). More general conditions are $G_1 > G_0$ or $G_1 > \phi_S$. As an example,
Eq. (2.32), with $G_0 = \phi_s / 2$ and $\phi_s = 4\pi$, shows that $\phi_0 = 6.5\pi$ results in an error which does not exceed $0.1\pi (\lambda/20)$; with $G_1 \approx 32\pi$, $|\phi_e| < 0.02\pi (\lambda/100)$. Eq. (2.23) implies that the minimum usable gain is $G_1 \approx \pi/2 = 1.6$, which allows a full cycle of $\phi_0$ to be acquired (assuming $G_0 > \pi/2$).

Since intensity must be positive, an interferometer always results in $I_o > I_1$, (e.g. in Eqs. (2.1) and (2.2)). If the detector lacks threshold Eq. (2.14) reveals that $G_0 > G_1$ results, which violates the $G_1 > G_0$ gain condition. Detector threshold thus enhances IPL operation, but is not essential. Alternatives to threshold include electronic offset of the detector output signal or the inclusion of extra dynamic range in $\phi_m$ in order to maintain $G_1 > |\phi_m - G_0|$. Other techniques for obtaining satisfactory IPL performance with an SLM lacking intensity threshold are discussed in chapter 4 (section 4.3). As a result of the $G_1 > G_0$ condition, it is often not desirable to operate with an intense L.O. beam ($|E_2| >> |E_1|$); this is in contrast to conventional homodyne or heterodyne detection, where $|E_2| >> |E_1|$ usually holds.

2.2.3. Basic dynamic behavior of the IPL

The homodyne implementation with a low-pass loop filter of the form in Eq. (2.28) will be used to illustrate the basic dynamic behavior of the IPL. It will be seen in later chapters that many physical detector-amplifier-modulator systems are adequately modelled by such a dominant single pole. Some of the effects introduced by employing other loop filters or elementary interferometers in the IPL will be discussed in Chapter 3.
With a transfer function of the form

$$\hat{H}(s) = \frac{\lambda}{s+\lambda}$$  \hspace{1cm} (2.28)

the system differential equations (Eqs. (2.12) and (2.13)) become:

$$\dot{\phi}_m = \lambda(-\phi_m + G_0 - G_1 \sin(\phi_m - \phi_i))$$  \hspace{1cm} (2.29)

and

$$\dot{\phi}_e = \dot{\phi}_i - \lambda(\phi_e - \phi_i + G_0 + G_1 \sin \phi_e)$$  \hspace{1cm} (2.30)

Here $\dot{\phi} = \frac{d\phi}{dt}$.

Graphical techniques\textsuperscript{162} provide significant insight into the behavior of these nonlinear differential equations. The modulator state-space plot ($\dot{\phi}_m$ versus $\phi_m$) of Fig. 2.7 is a graphical representation of Eq. (2.29) with $\dot{\phi}_i = 0$. Limited modulator dynamic range ($0 < \phi_m < \phi_s$) and detector threshold and saturation, from Eq. (2.6), are also represented. Whenever the derivative is positive ($\dot{\phi}_m > 0$) the modulator phase increases, following a trajectory to the right in Fig. 2.7. Whenever $\dot{\phi}_m < 0$, the trajectory goes to the left. Stable equilibria (circled point) occur where trajectories converge from above and below onto the $\dot{\phi}_m = 0$ axis.

According to Eq. (2.29), the whole sinusoidal trajectory of Fig. 2.7 shifts laterally in proportion to $\dot{\phi}_i$. The stable equilibrium points of $\phi_m$ thus follow the trajectories in Fig. 2.5 of the last section, continuously tracking $\phi_i$ variations within the system bandwidth. The fact that these trajectories are the same as those for $\hat{H}(s) = 1$ should not be surprising. According to Fig. 2.7, $\dot{\phi}_m = 0$ at equilibrium which causes Eq. (2.29) to become
Fig. 2.7 Modulator phase state-space ($\dot{\phi}_m$ versus $\phi_m$)

From $\dot{\phi}_m = \lambda (-\phi_m + G_o - G_1 \sin(\phi_m - \phi_i^l))$ with $\phi_i^l = 0$, $G_o = 1.8\pi$, $G_1 = 7.3\pi$, and $\phi_s = 3.8\pi$
Eq. (2.15). The error behavior of Fig. 2.6, could have also been obtained from Fig. 2.7 as the change in $\phi_m$ between where the trajectory crosses the lines $\dot{\phi}_m = -\lambda(G_0 - \phi_m)$ and $\dot{\phi}_m = 0$ (equilibrium). Notice that the trapezoidal operating region of Fig. 2.7 does not change with $\phi_i$.

The reader should study Fig. 2.7 very carefully; it is incredibly descriptive, conveying most of the essential behavior of the low-pass homodyne IPL. All the static results of the previous section as well as a large number of dynamic conclusions can be derived from it.

Also important, but not quite as informative, is the error state-space plot ($\phi_e$ versus $\phi_e$) of Fig. 2.8, which is a graph of Eq. (2.30) with $\dot{\phi}_e = 0$ and $\phi_i = 0$. The whole curve, including operating region boundaries, shifts up as $\phi_i$ increases; but the $\phi_e$ equilibria on the $\phi_e$ axis (where $\dot{\phi}_e = 0$) do not change much. The stable equilibria at $\phi_e = 2\pi n$ are circled. The $\phi_e$ operating range, $(\phi_i - \phi_s \leq \phi_e \leq \phi_i)$, shifts laterally with $\phi_i$, lying between the intersections on the $\phi_e$ axis of the dashed lines marked "determines $\phi_m$ limits".

Notice that the $\phi_e$ equilibria in Fig. 2.8 are insensitive to shifts in offset ($G_0$), gain ($G_1$), and input phase $\phi_i$; provided $G_1$ remains large and saturation or cutoff are not encountered. The insensitivity of $\phi_m$ to variations in $G_1$ and $G_0$ induced by input intensity fluctuations is also apparent in Fig. 2.7 ($\dot{\phi}_m$ versus $\phi_m$). As long as $G_1$ is large, the whole trajectory can shift up and down significantly without greatly perturbing the equilibria at $\dot{\phi}_m = 0$.

Bandwidth of the single-pole IPL

The maximum rate at which $\phi_i$ can change and still be successfully tracked by the type-zero IPL is of great interest. Instantaneous jumps in
Fig. 2.8 Error phase state-space ($\dot{\phi}_e$ versus $\phi_e$).

(From $\dot{\phi}_e = \lambda(-\phi_e + \phi'_1 - G_0 G_1 \sin \phi_e)$; with $\phi'_1 = 0$, $G_1 = 7.3\pi$, $G_0 = 1.8\pi$, $\phi_s = 3.8\pi$.)
\( \phi_i \) are tracked with a settling delay, (i.e., "step response"). In Fig. 2.7, when \( \phi_i' \) suddenly changes by \( \Delta \phi_i' \), the whole trajectory instantaneously shifts laterally by \( \Delta \phi_i' \). If the IPL were in equilibrium (with \( \phi_m = \phi_{m0}, \dot{\phi}_m = 0 \), and \( \phi_i' = \phi_{i0}' \)) before the shift; after the shift \( \phi_{m0} \) results in a nonzero \( \dot{\phi}_m \) on the shifted trajectory, which drives \( \phi_m \) to its new equilibrium. Immediately after the shift the derivative is: 
\[
\dot{\phi}_{m0} = (-\phi_{m0} + G_o - G_1 \sin(\phi_{m0} - \phi_{i0}' - \Delta \phi_i')).
\]
The return to equilibrium is approximately an exponential decay, since straight lines (see Fig. 2.7) on a state-space plot (e.g., \( \dot{\phi} = a - b\phi \)) correspond to exponential decay. More specifically, if the system were initially in lock with small error and \( \Delta \phi_i' \) is small, \( \phi_e - \phi_m \) will remain small and Eq. (2.29) can be linearized, becoming
\[
\dot{\phi}_m + \lambda (G_1 + 1) \phi_m = \lambda (G_o + G_1 \phi_i' - G_1 2\pi n) \tag{2.31}
\]
where
\[
n = \text{INT}[ (\phi_e + \pi) / 2\pi ]
\]
With initial condition \( \phi_{m0} \), the general solution to Eq. (2.31) is
\[
\phi_m(t) = \phi_{m0} e^{-\lambda (G_1 + 1) t} + \phi_{mss}(1 - e^{-\lambda (G_1 + 1) t}) \tag{2.32}
\]
where
\[
\phi_{mss} = \frac{G_1}{1 + G_1} (\phi_{i0}' + \Delta \phi_i' + \frac{G_o + 2\pi n}{G_1}) - 2\pi \tag{2.33}
\]
Note that \( \phi_{mss} \) is the general linearized, type-zero, steady-state discussed previously in Eqs. (2.19). With large gain, \( \phi_{mss} \) equals the new input phase \( \phi_{i0}' + \Delta \phi_i' \).

Equations (2.31) and (2.32) suggest that the effective tracking
bandwidth of the single-pole IPL is

\[ \text{BW} = \lambda (G_1 + 1) = \lambda G_1 \]  

(2.34)

The loop should smoothly track phase fluctuations for which

\[ \dot{\phi}_i < \lambda (G_1 + 1) \dot{\phi}_i \]  

(2.35)

When Eq. (2.35) is satisfied, the loop converges faster than \( \phi_i \) varies.

In most practical applications \( \phi_i(t) \) has a complicated time dependence and it is useful to decompose \( \phi_i(t) \) into its spectral components, \( \hat{\phi}_i(w) \). To the extent that the IPL is not perturbed much from equilibrium and the linear approximation of Eq. 2.31 remains valid, the response at each frequency superimposes, i.e.,

\[ \phi_m(t) = \int_{-\infty}^{\infty} \phi_m(w)e^{jwt} \, dw \]

At frequency \( w \), the responses of \( \hat{\phi}_m \) and \( \hat{\phi}_e \) follow from Eq. (2.31)

\[ \hat{\phi}_m(w) = \frac{G_1 \delta(w)}{1 + G_1} \left( \frac{G_0}{G_1} - 2\pi \right) + \frac{\lambda G_1}{jw + \lambda (G_1 + 1)} \hat{\phi}_i(w) \]  

(2.36)

\[ \hat{\phi}_e(w) = -\frac{G_1 \delta(w)}{1 + G_1} \left( \frac{G_0}{G_1} - 2\pi \right) + \frac{jw + \lambda}{jw + \lambda (G_1 + 1)} \hat{\phi}_i(w) \]  

(2.37)

Here \( \delta(w) \) is the impulse function.

Assuming \( G_1 >> 1 \), successful phase tracking \((\hat{\phi}_m(w) = \hat{\phi}_i(w) + \text{const})\) and phase compensation \((\hat{\phi}_e(w) = \text{const})\) occur for spectral components satisfying \( w < \lambda (G_1 + 1) \). When this condition, which is the same as Eq. (2.35), is not satisfied, the spectral components of the phase fluctuations which are filtered out of the modulator signal are passed through in the phase error.
i.e., \( \phi_e = \phi_i \) and \( \phi_m = 0 \) (for \( w \neq 0 \)). Notice that \( \phi_e(w) + \phi_m(w) = \phi_i(w) \) at all frequencies. According to Eq. (2.5), this division of phase fluctuations between the modulator and phase error holds for any stable transfer function.

The bandwidth constraints, Eqs. (2.34) and (2.35), are most useful for \( \phi_i(t) \) which is randomly fluctuating in no systematic way; so that \( \phi_e(t) \) remains small. As mentioned previously, the physical limits of the modulator (0s\( \phi_m \leq \phi_s \)) restrict the active error range to (using Eq. (2.18a))

\[
-sin^{-1} \frac{G_o}{G_1} \leq \phi_e - 2n\pi \leq sin^{-1} \frac{\phi_s - G_o}{G_1}
\]

(2.38)

If the active \( \phi_e \) does become large, the linear approximation fails. Then the frequency modulated \( \sin(\phi_e(w)e^{jwt}) \) term cross-couples many harmonics of \( w \); and the simplifying advantages of the frequency domain analysis are destroyed. It will be seen in Chapter 3 that expressions can still be obtained for the bandwidth with large error in some instances (e.g., with an integrating loop filter). As \( \phi_e \) increases toward \( (3-4n)\pi/2 \), it will be seen that the bandwidth generally decreases, since the approach to \( \dot{\phi}_e = 0 \) in the \( \phi_e \) versus \( \phi_e \) state-space (e.g., Fig. 2.8) becomes less steep.

Stability

It should be noted that the linearized closed-loop transfer function of the low-pass IPL, \( H(w) = \frac{G_1\lambda}{jw+\lambda(1+G_1)} \), is stable for any value of gain. This is consistent with the more general state-space trajectory of Fig. 2.7, which reveals that \( \phi_m \) is unconditionally stable after a step in \( \phi_i \). The modulator will always be driven to an equilibrium on the \( \dot{\phi}_m = 0 \) axis, even if \( G_o > G_1 \). The stability with other loop filters and interferometric implementations is discussed in Chapter 3.
Operational Limits

The effects of detector threshold and saturation ($i_{\text{o cutoff}} < i_0 < i_{\text{osat}}$) and the limited dynamic range of real modulators ($0 < \phi_m < \phi_s$) are included in Fig. 2.7. As $\phi_i$ increases, $\phi_m$ eventually saturates; then $\phi_m = \phi_s$ and $\phi_m = 0$ until further increases in $\phi_i$ cause an unstable equilibrium of Fig. 2.7 to reach $\phi_s$. From there the system follows the trajectory which causes $\phi_m$ to jump to the next-lowest stable equilibrium point ($\phi_m = \phi_s - \pi$). Similarly, with decreasing phase $\phi_m$ does not stay pinned at its minimum value, but eventually returns to its phase-tracking region. These effects are illustrated in the $\phi_m$ versus $\phi_i$ graph of Fig. 2.5. Expressions are derived in Chapter 3 for the recovery value of $\phi_m$ after reaching the $\phi_m = 0$ or $\phi_m = \phi_s$ limits.

In Fig. 2.7, the detector limitations on $i_o$ (from Eq. 2.6) constrain $\phi_m$ and hence lengthen the time it takes $\phi_m$ to move from the vicinity of an unstable equilibrium to a stable equilibrium. The time constant on the linear segment is $\lambda$. If the modulator had a larger dynamic range, Fig. 2.7 reveals that detector threshold and saturation would also constrain $\phi_m$, to $0 < \phi_m < G_s / \lambda$ ($G_s$ is defined in Eq. (2.14c)). Fig. 2.5 retains the same shape, but $\phi_{m\text{max}} = G_s$ replaces the upper limit. This limitation occurs because the steady-state value of $\phi_m$ is proportional to the detector signal in an IPL with a type-zero loop filter.

Other nonlinearities can also be represented in Fig. 2.7. For example, if the modulator exhibits imperfect saturation, Eq. (2.7) becomes $\phi_m = a_2(\phi_m) \cdot i_1$, where $a_2(\cdot)$ is a nonlinear decreasing function of $\phi_m$, for $\phi_m > \phi_s$. According to Eqs. (2.14), a reduction in $a_2$ diminishes $G_0$ and $G_1$ and hence causes the envelope ($G_0 \pm G_1 - \phi_m$) of the sinusoidal trajectory.
in Fig. 2.7 to shift down and to constrict for $\phi_m > \phi_S$. This "squashed" trajectory will allow $\phi_m$ to increase with $\phi_1^*$ until the saturated values of $G_0$ and $G_1$ ($G_{os}$ and $G_{ls}$ respectively) satisfy $G_{os} + G_{ls} \sin(\phi_1^* - \phi_m) - \phi_m = 0$, which occurs at $\phi_m$ somewhere beyond $\phi_S$. Then $\phi_m$ is driven by a negative trajectory to its next smallest equilibrium. (In Fig. 2.4 the jump condition, $G_{os} + G_{ls} \sin(\phi_0^* - \phi_m)$, occurs where locus $\phi_m$ is just tangent to the now-curved top of locus $\phi_m^2$).

In the absence of modulator or detector limitations, Fig 2.7 reveals that the dynamic range of $\phi_m$ is

$$G_0 - G_1 - \frac{\pi}{2} \leq \phi_m \leq G_0 + G_1 + \frac{\pi}{2}$$

(2.39)
as was suggested previously. For example when the upper limit is exceeded, $\phi_0^* = \pi/2$ and the tip of the sinusoidal trajectory drops below the $\phi_m = 0$ axis. Then $\phi_m$ follows a decreasing trajectory to the next lowest equilibrium.

**Time response**

Although Eqs. (2.28) and (2.30) cannot be solved for closed-form expressions for the time behavior of $\phi_e$ and $\phi_m$, a numerical solution was obtained by using difference equations similar to Eqs. (3.58) in chapter 3. The convergence along the approximately linear portions of the trajectories in Figs. 2.7 and 2.8 (corresponding to $\sin \phi_e < 0.5$) was found to be well characterized by Eqs. (2.32), i.e. an exponential response with time constant $\tau_a = 1/[(\lambda(1 + G_1))]$, particularly for $G_1 >> G_0$. For larger error, $\tau_b = 1/[(\sqrt{G_1^2 - G_0^2} + 1)\lambda]$ was found to be a better approximation to the time constant. In the absence of limits on $i_o$, the time to traverse a complete loop of the sinusoid from within 10% of an unstable equilibrium to 10% of a stable equilibrium was found to be less than $4\tau_a$ (with $|\phi_m - G_0| < G_1$).
This relatively fast response occurs because the sinusoidal loop passes through regions of large derivative. This time is comparable to the rise time along a linear segment ($\approx 2.2 \tau_a$). Hence $\tau_a$ is a good estimate of the system phase acquisition time constant, even when the initial error is large or phase-lock is being reacquired after the system has reached the modulator's limits.

**Concluding Remarks**

The interference phase loop appears to be simple enough to be practical for implementing high-resolution systems and yet capable of performance superior to that of most other phase estimation techniques. It was shown that the modulator phase can continuously estimate and compensate input phase fluctuations within the system bandwidth, ($\approx \lambda_1 G_1$), even in the presence of signal amplitude fluctuations. Since there is no phase quadrant ambiguity, it always apparent whether $\phi_i$ is increasing ($\phi_m$ increases) or decreasing ($\phi_m$ decreases). In phase estimation applications, the control signal to the modulator or the irradiance of the interferogram at the detector can be taken as the phase estimate. The dynamic range for real-time phase tracking is limited only by the system gain and physical limits of the system components. Most of the important features of the dynamic and static behavior of the low-pass homodyne IPL can be inferred from the $\phi_m$ versus $\phi_m$ state-plot of Fig. 2.7. The linearized expressions in Eqs. (2.25) and (2.26) summarize the important aspects of the steady-state behavior; these expressions are valid when the last term in Eq. (2.26b) is less than about $\pi/6$, which usually holds over the full dynamic range of the modulator.

Many of the basic operational characteristics of the IPL
introduced in this chapter are treated more thoroughly in Chapter 3. Although the homodyne implementation with a low-pass loop filter was used to motivate the IPL discussions of this chapter, it will be seen in Chapter 3 that other interferometric implementations (e.g., heterodyne or Zernike phase-contrast) and other loop filters (e.g., integrator or two-pole filter) do not greatly modify the conclusions presented here. In fact the reader who is not interested in the more detailed aspects of IPL performance, can skip Chapter 3 without a great loss of continuity.
III. ADVANCED OPERATIONAL THEORY OF THE INTERFERENCE PHASE LOOP

Whereas the previous chapter introduced the essential operational behavior of the IPL, this chapter further develops the more important IPL characteristics and introduces other interferometric implementations. Topics covered in this chapter include: IPL performance in the presence of amplitude fluctuations, phase acquisition, multistability and $2n\pi$ degeneracy, spatial resolution requirements, spatial phase compensation with the IPL, the dependence of IPL behavior on the kind of filter incorporated into the loop, the effects of noise on IPL performance, and the minimum irradiance required for successful phase tracking. The behavior of IPLs employing heterodyne interferometry or self-interference (Zernike phase-contrast), rather than homodyne interferometry, are also discussed.

3.1. A more detailed look at IPL performance

Since the reader is most familiar with the homodyne IPL, much of the discussion in this section will continue to be couched in terms of that implementation. However, most of the analysis techniques and conclusions are applicable to other implementations.

3.1.1. Insensitivity to amplitude fluctuations

Equations (2.14) of Chapter 2 showed that fluctuations in the input signal amplitude, $|E_i|$, cause the IPL offset, $G_0$, and gain, $G_1$, to vary. However Figs. 2.4, 2.7, and 2.8 illustrated that the IPL phase estimate, $\phi_m$, and phase error, $\phi_e$, are not strongly influenced by variations in $G_0$ and $G_1$ as long as $G_1 > |\phi_m - G_0|$ continues to hold.

Explicit expressions for the sensitivity of the phase estimate and phase error to variations in $G_0$ and $G_1$ can be found by differentiating the system operating relations with respect to $G_0$ and $G_1$. For example in
the homodyne IPL with a type-zero loop filter, it follows from Eqs. (2.5), (2.17a), and (2.18a) that

\[
\frac{\partial \phi_m}{\partial G_0} = -\frac{\partial \phi_e}{\partial G_0} = \frac{1}{1+G_1a} \tag{3.1a}
\]

\[
\frac{\partial \phi_m}{\partial G_1} = -\frac{\partial \phi_e}{\partial G_1} = \frac{\sqrt{1-a^2} \sin \phi_e}{1+G_1a} \tag{3.1b}
\]

where \( a = \sqrt{1 - \left( \frac{\phi_m - G_0}{G_1} \right)^2} = \cos \phi_e \)

(Remember, from Eq. (2.5), that \( \phi_m = \phi_i - \phi_e \).)

With large gain, \( G_1 \gg G_0 \), the desired insensitivity is exhibited:

\[
\frac{\partial \phi_m}{\partial G_0} \rightarrow 0 \tag{3.2a}
\]

\[
\frac{\partial \phi_m}{\partial G_1} \rightarrow 0 \tag{3.2b}
\]

The full effect of amplitude, \( |E_i| \), fluctuations is dependent on the specific implementation as expressed through \( \frac{\partial G_0}{\partial |E_i|} \) and \( \frac{\partial G_1}{\partial |E_i|} \). Using Eqs. (3.1), (assuming \( \frac{\partial \phi}{\partial |E_i|} = 0 \)):

\[
\frac{\partial \phi_m}{\partial |E_i|} = -\frac{\partial \phi_e}{\partial |E_i|} = \frac{\frac{\partial G_0}{\partial |E_i|} + \frac{\partial G_1}{\partial |E_i|} \sin \phi_e}{1 + G_1 \cos \phi_e} \tag{3.3}
\]

In the homodyne implementation \( G_0 \) and \( G_1 \) are defined in terms of \( |E_i| \) through Eqs. (2.14), which can be rewritten:

\[
G_0 = g(\alpha^2 |E_i|^2 + |E_i|^2 - I_0^2) \tag{3.4a}
\]
and

\[ G_1 = 2g\alpha |E_i||E_\perp| \quad (3.4b) \]

where

\[ g = \frac{g_0 g_1 g_2}{2Z_o} \quad (3.4c) \]

and \( I_t^* = 2Z_0 I_t \).

The modulator phase is \( \phi_m = \phi_1 - \phi_e \), where the error phase follows from Eqs. (2.18a) and (3.4) as: (modulo \( 2\pi \))

\[
\phi_e = \sin^{-1}\left( \frac{\phi_m - G_0}{G_1} \right) = \sin^{-1}\left( \frac{1}{2|E_\perp|} \left( \frac{A}{\alpha |E_i|} - \alpha |E_i| \right) \right) - 2n\pi
\]

\[ = \sin^{-1}\left( \frac{gA - g\alpha^2 |E_\perp|^2}{2g\alpha |E_\perp|^2} \right) - 2n\pi \quad (3.5) \]

where

\[ r = \alpha |E_i| / |E_\perp| \]

and

\[ A = \frac{\phi_m}{g} - |E_\perp|^2 + I_t^* = (\phi_1^* - \phi_e - G_0(0)) / g \quad (3.6a) \]

with

\[ G_0(0) \equiv \lim_{|E_i| \to 0} G_0 = g(|E_\perp|^2 - I_t^*) \quad (3.6b) \]

\( (I_t^* < (|E_i| + |E_\perp|)^2 \text{ must hold}) \)

Eq. (3.3) now becomes

\[
\frac{\partial \phi_m}{\partial |E_i|} - \frac{\partial \phi_e}{\partial |E_i|} = \frac{\alpha}{2} \left( \frac{A}{|E_i|^2 - \alpha^2} + 1 \right)
\]

\[ \quad - \frac{1}{2g\alpha |E_i|} + \sqrt{|E_\perp|^2 - 4 \left( \frac{A}{\alpha |E_i|} - \alpha |E_i| \right)^2} \quad (3.7) \]

The gain \( g \), local oscillator amplitude \( |E_\perp| \), and threshold \( I_t^* \) are fixed parameters of the IPL, chosen by the system designer.

When the gain \( G_1 \) is made large by increasing the local oscillator amplitude \( |E_\perp| \), while also increasing the threshold \( I_t^* \) in order to
maintain $|E_2|^2 - I_t'$ at a constant value, the IPL becomes insensitive to fluctuations in $|E_i|$; i.e.,

$$\lim_{|E_2| \to \infty} \frac{\partial \phi_m}{\partial |E_i|} = 0 \quad (3.8)$$

This is essentially a restatement of the suggestion in chapter in chapter 2 that the IPL exhibits good immunity to amplitude fluctuations when $G_1 > G_0$ holds. Note that as $g$ increases, $A$ and hence $\phi_e$ and $\partial \phi_m/\partial |E_i|$ also become less sensitive to the value of $\phi_m$.

With arbitrary system parameters $|E_2|$, $I_t'$, and $g$, the dependence of $\phi_m$ on $|E_i|$ as represented by Eqs. (2.5) and (3.4) - (3.7) is not obvious. For example, the limits $|E_2| \to 0$ or $|E_i| \to \infty$ cannot be directly taken because $A$ in Eq. (3.6a) is a function of $\phi_m = \phi'_i - \phi_e$, and $\phi_e$ is a function of $|E_i|$ by way of Eq. (3.5). It will be seen that there are generally minimum and maximum field amplitudes between which a given operating point is valid, i.e. the $G_1 - |\phi_m - G_0| \leq 0$ condition in Eq. (3.5) is satisfied between these limiting amplitudes. Between these extremes, $|E_i|$ can vary by more than an order of magnitude with negligible effect in $\phi_e$ or $\phi_m$.

Useful insights can be gained by referring back to Fig. 2.4; remember that steady-state values of $\phi_m$ and $\phi_e$ occurred where the two curves $\phi_{m_1} = \phi'_i - \phi_e$ and $\phi_{m_2} = G_0 + G_1 \sin \phi_e$ intersected. In Fig. 2.4: $G_1$ is the size of the $\phi_{m_2}$ sinusoid; $G_0$ is the distance from the "zero-crossing" of the sinusoid (at $\phi_e = 2\pi n$) to the $\phi_m = 0$ axis; and $G_0 + G_1$ defines the envelope of the sinusoid. The behavior of $G_1$, $G_0$, and $G_0 + G_1$ as $|E_i|$ varies is sketched in Fig. 3.1. Notice from Eqs. (3.4) that when $|E_i|$ is small (i.e., $\alpha |E_i| < |E_2|$), $G_1$ increasing in proportion to $|E_i|$, grows faster than $G_0$, which is proportional to $|E_i|^2$. 


Fig. 3.1 Gain $G_1$, offset $G_0$, and $G_0 \pm G_1$ as a function of relative input amplitude $|E_i|/|E_x|$. (For $g|E_x|^2=40\pi, g_{1t}=38.2\pi$, and $\alpha=1$.)
As $|E_i|$ increases, eventually the distance from the "zero crossing" to the $\phi_m$ solution in Fig. 2.4 (i.e., $\phi_m - G_0$) exceeds $G_1$; after passing through $-\pi/2$ the error jumps to another operating point. As $|E_i|$ decreases, eventually $G_0$ stabilizes at $g(|E_i|^2 - I_i')$ and $G_1$ decreases until $G_1 < |G_0 - \phi_m|$; then the system must jump to another operating point. The associated dynamic behavior can be seen by studying Fig. 2.7. In the special case of $\phi_i' = G_0(0)$, examination of Fig. 2.4 reveals that the error continues to decrease as $|E_i|$ gets arbitrarily small; (i.e. $\lim_{|E_i| \to 0} |\phi_e| = 0$).

Figure 3.2 illustrates the behavior of the phase error, $\phi_e$, as a function of input amplitude, $|E_i|$. (These results were computed by iterating Equations (3.5) and (3.6) until $\phi_e$ converged.) Note that $\phi_e$ remains small over an order of magnitude of $|E_i|$. As $g$ increases, the dynamic range extends further in the small $|E_i|$ direction.

The three different types of behavior for small $|E_i|$ in Fig. 3.2 can be understood from Eq. (3.5): 1) when $A < 0$ for small $|E_i|$, decreasing $|E_i|$ results in $\phi_e \to -\pi/2 + 2n\pi$; 2) when $A > 0$ for small $|E_i|$, $\phi_e \to \pi/2 + 2n\pi$; and 3) when $\lim_{E_i \to 0} A < 0$, $|\phi_e| < \pi/2 + 2n\pi$ results. The actual case depends on $\phi_i'$ and the multistability of the trajectory, which can be seen from Eqs. (3.6) where $A(0) = \lim_{E_i \to 0} A = \phi_i'(0) - G_0(0)$. In case 1, $A(0) < 0$ and $\lim_{|E_i| \to |E_i|_{\text{min}}} \phi_e = -\pi/2 + 2n\pi$ implies

(Case 1) \[ G_0(0) - \phi_i' - 2n\pi > \frac{\pi}{2} \] \hspace{1cm} (3.9a)

Similarly in cases 2 and 3:

(Case 2) \[ G_0(0) - \phi_i' - 2n\pi < \frac{\pi}{2} \] \hspace{1cm} (3.9b)

(Case 3) \[ \frac{\pi}{2} \leq G_0(0) - \phi_i' - 2n\pi \leq \frac{\pi}{2} \] \hspace{1cm} (3.9c)

In Eqs. (3.9) $n$ is given by Eq. (2.25)

Case 3 is the most desirable, because the error remains finite as
a) $\phi_e'$ versus $|E_i|/|E_\lambda|$, illustrating the variation of $\phi_e$ as $\phi_i'$ varies over one full cycle of $2\pi$ radians. For $|E_i|$ under the solid curves the error does not exceed $\pm 0.1\pi$. The solid curves are for $g|E_\lambda|^2=40\pi$ and $gI_t'=38.2\pi$. The dotted curves are for $g|E_\lambda|^2=80\pi$ and $gI_t'=78.2\pi$ ($\alpha=1$).

b) $\phi_e'$ versus $|E_i|/|E_\lambda|$, illustrating the maximum variation of $|E_i|$ for which a given trajectory (n value) is valid. Here, $\phi_e'$ is $\phi_e$ modulus $2\pi; \phi_i'=-0.2\pi; g|E_\lambda|^2=40\pi; gI_t'=38.2\pi$; and $\phi_m=\phi_i'-\phi_e'+2nm$ where $n=0, 1, 2, \text{ or } 4$.

Fig. 3.2 Dependence of error phase $\phi_e$ on relative input amplitude, $|E_i|/|E_\lambda|$
\(|E_i|\) goes all the way to zero. However, since \(\phi_i'\) (modulo \(2\pi\)) ranges from \(-\pi\) to \(+\pi\), all three cases can occur in an actual system. Careful examination of Fig. 2.7 as \(|E_i|\to 0\) provides further insight. Eqs. (3.4) and Fig. (3.1) show that as \(|E_i|\to 0\), \(G_o\) remains constant at about \(G_o(0)\) and \(G_i\) linearly decreases. Assuming no \(\phi_i\) limits, stable equilibria to the left of \(\phi_m=G_o(0)-\pi\) go to an error of \(\sin \phi_e = -1\) (case 1) at a minimum value of \(|E_i|\); all values of \(\phi_i'\) (modulo \(2\pi\)) will remain in case 1. If \(|E_i|\) drops below the minimum value, the system follows an increasing sinusoidal trajectory to another value of \(\phi_m\) closer to \(G_o(0)\). Similarly, case 2 occurs for \(\phi_m>G_o(0)+\frac{\pi}{2}\), and case 3 occurs only on the specific trajectory satisfying \(-\pi<\phi_m-G_o(0)<\pi\). In the limit of \(|E_i|\to 0\), Fig. 2.7 reveals that \(\phi_m(0)-G_o(0)\) and \(\phi_e(0)\to \phi_i'-G_o(0)\).

The above discussion suggests that the best immunity to amplitude fluctuations is obtained when \(\phi_m\) is operated near \(G_o(0)\). This is consistent with earlier statements that best performance is obtained for \(G_i>|\phi_m-G_o|\). It is shown in the next section that the initially acquired phase of \(\phi_m\) is \(\phi_{mo}\), where \(\phi_{mo}\) is the initial condition at \(t=0\). Thus optimum suppression of amplitude fluctuations is obtained by occasionally restarting the IPL with \(\phi_{mo}=G_o(0)\).

The actual range of \(|E_i|\) for which a given trajectory is valid can be found by setting the argument of the arcsine in Eq. (3.5) equal to \(\pm 1\) as appropriate and solving the resulting quadratic equation in \(|E_i|\).

For case 1:

\[
|E_i| - \sqrt{\frac{\phi_i' + \pi}{2 - 2n\pi}} + I_t' \leq |E_i| \leq |E_i| + \sqrt{\frac{\phi_i' + \pi}{2 - 2n\pi}} + I_t' \tag{3.10}
\]

With large \(g\) and \(I_t' = |E_i|^2\) the range of \(\alpha |E_i|\) is almost from zero to \(2|E_i|\), which is twice the local oscillator (L.O.) amplitude. There is an inter-
mediate value of $|E_i|$ where $\phi_e$ reaches a minimum; when $\frac{\partial \phi_e}{\partial |E_i|} = 0$ in Eq. (3.7):

$$\alpha |E_i| = \sqrt{A} \quad (3.11)$$

and

$$\phi_{e\min} = -\sin^{-1} \left( \frac{\sqrt{A}}{|E_\lambda|} \right) \quad (3.12)$$

($A < 0$ always holds in this case).

For case 2 the error ranges from $-\pi/2$ at the lower operational limit of $|E_i|$ to $+\pi/2$ at the upper limit. From Eq. (3.5)

$$\sqrt{\frac{\phi'_i - \pi/2 - 2n\pi}{g}} + I'_t - |E_\lambda| \leq \alpha |E_i| \leq \sqrt{\frac{\phi'_i + \pi/2 - 2n\pi}{g}} + I'_t + |E_\lambda| \quad (3.13)$$

and $\phi_e = 0$ when:

$$\alpha |E_i| = \sqrt{A} \quad (3.14)$$

($A > 0$ holds in this case).

The dynamic range of $\alpha |E_i|$ is again twice the L0.amplitude; i.e., $2|E_\lambda|$.

In case 3, the IPL continues to track $\phi'_i$ as $|E_i|$ goes all the way to zero, at which point the error is:

$$\phi_e(|E_i|=0) = \phi'_i - G_o(0) \quad (3.15)$$

The operating range is:

$$0 \leq \alpha |E_i| \leq |E_\lambda| + \frac{\sqrt{\frac{\phi'_i + \pi/2 - 2n\pi}{g}} + I'_t}{\sqrt{A}} \quad (3.16)$$
In the special case of $\phi_i^*=G_0(0)$

$$0 \leq |E_i| \leq 2 |E_\tilde{i}|$$

(3.17a)

In conclusion, the IPL shows good immunity to amplitude fluctuations of a few orders of magnitude, as long as the local oscillator amplitude is made larger than the largest expected value of $\alpha|E_i|$. (Remember that $\alpha|E_i|$ is the signal amplitude reaching the IPL detector after attenuation by $\alpha$.) Actually, $\phi_e$ can be zeroed for any value of $r = |E_i|/|E_\tilde{i}|$ ($0 < r < \infty$) by readjusting $I_\downarrow$ and $n$, but the new $I_\downarrow$ and $n$ will reduce the range of tolerable fluctuations, $\Delta|E_i|/|E_i|$, in input amplitude. It is also desirable to have a large system gain, $g$, and to operate with $\phi_m = G_0(0)$. This discussion was for a homodyne IPL with a type-zero loop filter, but similar behavior is obtained with other interferometric implementations and loop filters. The minimum irradiance required by the IPL in the presence of noise will be considered in section 3.1.9.

3.1.2. Initial Phase Acquisition

The transient behavior of the IPL between initial turn on and successful phase tracking is most easily seen with the aid of state-space plots such as $\dot{\phi}_m$ versus $\phi_m$, or $\dot{\phi}_e$ versus $\phi_e$. Before turn-on the loop is generally open and $\phi_m$ is preset at an initial value, $\phi_{m0}$, with $\dot{\phi}_{m0} = 0$. For example, in the system of Fig. 2.3: the amplifier could be off with a constant output, e.g. $\phi_{m0} = 0$; or a mirror could be rotated to block the optical feedback path and reflect a
calibrated intensity into the detector; or $E_i$ could be blocked, with the detector seeing $|E_2|^2$. At turn on the state-space characteristic instantaneously becomes valid and describes the transient response to equilibrium.

In the specific case of the homodyne IPL with a low-pass filter, the value of $\phi_m$ after phase acquisition is generally within $\pm \pi$ radians of $\phi_{mo}$. At the instant of turn on, Fig. 2.7 ($\phi_m$ versus $\phi_m$) becomes valid, and $\phi_m$ exhibits a step response with initial derivative

$$\dot{\phi}_{mo} = \lambda (\phi_{mo} - G_o G_1 \sin(\phi_{mo} - \phi_1)).$$

(If $i_0$ cutoff or saturation occurs, the initial derivative is $\dot{\phi}_{mo} = -\lambda \phi_{mo}$ or $\dot{\phi}_{mo} = \lambda (G_s - \phi_{mo})$. The turn-on phase corresponding to any $\phi_1$ can be found by laterally shifting the trajectory in Fig. 2.7. The turn-on value of $\phi_m$ as a function of $\phi_1$ is illustrated in Fig. 3.3 for two values of $\phi_{mo}$. Figure 2.7 reveals that the $2\pi$ changes in $\phi_m$ occur when $\phi_{mo}$ is at an unstable equilibrium, i.e.,

$$\phi_1 = \phi_{mo} - \sin^{-1}(\frac{\phi_{mo} - G_o}{G_1}) + (2n+1)\pi$$

(3.18)

where $n$ is specified by Eq. (2.25). As long as the modulator limits are not encountered (i.e., $0 < \phi_m < \phi_s$), Eqs. (2.17a) and (3.18) show that the turn-on value of $\phi_m$ is always in the range:

$$\phi_{mo} - \pi - \sin^{-1}(\frac{\phi_{mo} - G_o}{G_1}) < \phi_m < \sin^{-1}(\frac{\phi_{mo} - G_o}{G_1}) < \phi_{mo} + \pi - \sin^{-1}(\frac{\phi_{mo} - G_o}{G_1})$$

(3.19)

With large gain

$$\phi_{mo} - \pi \leq \phi_m \leq \phi_{mo} + \pi$$

(3.20)
Fig. 3.3 $\phi_m$ after initial phase acquisition as a function of $\phi_i'$. (Shown for two initial values of $\phi_{mo}'$.)

Fig. 3.4 Graphical construction for reacquisition phase after reaching modulator limits. The bold lines are $\phi_m = G_0 + G_1 \sin \phi_e$. The dashed lines are $\phi_m = \phi_i' - \phi_e$, $\delta_s = \pi - \phi_{es}$ and $\delta_c = \pi - \phi_{ec}$, where $\phi_{es}$ and $\phi_{ec}$ are given by Eq. (3.21).
As the system tracks $\phi_i$, $\phi_m$ will range over multiple $\pi$ radians and can reach its limits, $\phi_m=0$ vs. $\phi_m=\phi_s$ in some resolution elements. In many applications occasionally restarting the system by illuminating the detector with a calibrated light source can improve system performance. The restart phase error is minimized for $\phi_{m_0}=G_o$, since $\phi_e=\sin^{-1}\left(-\frac{\phi_m-G_o}{G_1}\right)$. The system should also be designed with $G_o=\phi_s/2$ to maximize the distance to the modulator limits.

**Phase Reacquisition after reaching modulator limits**

As mentioned in Chapter 2, when the IPL is allowed to reach the modulator limits ($\phi_m=0$ or $\phi_s$), it does not stay pinned, but eventually reacquires the input phase. Just before recovery $\phi_m$ is at an unstable equilibrium and the error is at the maximum or minimum extreme of its active range,

$$\pi-\sin^{-1}\left(-\frac{G_o}{G_1}\right) \leq \phi_e \leq \phi_{es} \leq \pi-\sin^{-1}\left(-\frac{\phi_s-G_o}{G_1}\right) \quad \text{(modulo } 2\pi) \quad (3.21)$$

The reacquisition values of $\phi_m$ ($\phi_{ms}$ and $\phi_{mc}$ after the $\phi_m=\phi_s$ or $\phi_m=0$ limits respectively) can be determined by assuming that $\phi_i$ stays constant during the recovery transient.

After $\phi_m=\phi_s$:

$$\phi_i' = \phi_m + \phi_{es} = \phi_s + \pi - \sin^{-1}\left(-\frac{\phi_s-G_o}{G_1}\right) \quad (3.22a)$$

The corresponding reacquisition value of $\phi_m$ after saturation is:

$$\phi_{ms} = \phi_s - \sin^{-1}\left(-\frac{\phi_s-G_o}{G_1}\right) - \sin^{-1}\left(-\frac{\phi_{ms}-G_o}{G_1}\right) - 2\pi \quad (3.22b)$$
After $\phi_m = 0$:

$$\phi' = \pi \sin^{-1} \left( \frac{G_2}{G_1} \right) - 2\pi$$

(3.23a)

and

$$\phi_{mc} = -\pi \sin^{-1} \left( \frac{G_2}{G_1} \right) - \sin^{-1} \left( \frac{\phi_{mc} - G_2}{G_1} \right) + 2\pi$$

(3.23b)

The implicit transcendental equations in $\phi_{ms}$ and $\phi_{mc}$ can be solved with the aid of graphical constructions, such as $\phi_e$ versus $\phi_m$ in Fig. 3.4. Alternatively, numeric iteration can be employed, or if $G_1$ is large enough these equations can be linearized.

Numerical simulations of IPL dynamic behavior have shown that the reacquisition time is generally less than $4/\lambda(1+G_1)$, which is almost as fast as the initial convergence time ($\text{rise time} \approx 2/\lambda(1+G_1)$).

3.1.3 Multistability and $2n\pi$ degeneracy

Due to the transcendental nature of its operating equations, the IPL can exhibit multistability. This is apparent in the $n$ dependence of Eqs. (2.17) and (2.18) and in Figs. 2.4, 2.5, 2.7, and 2.8. In particular, notice in Fig. 2.5 that for a given value of $\phi_1$ multiple values of $\phi_m$ are possible. The total number of equilibria is determined by the maximum gain, $G_1$, or the physical limits of the system hardware. This multistable behavior can be observed in the phase estimate $\phi_m$, the phase error $\phi_e$, and the interferometer intensity $I$. This behavior may have application to building two-dimensional optical digital processing elements.

In phase estimation or compensation applications, there need not be multistability of $\phi_m$, since the initial value...
of modulator phase, \( \phi_{m0} \), determines which trajectory in Fig. 2.5 is followed. (With large gain, the value of \( \phi_m \) after phase acquisition was seen to be in the range \( \phi_{mo} - \pi \leq \phi_m < \phi_{mo} + \pi \)). For example if \( \phi_{mo} \) is occasionally reset to \( 2\pi \) and \( -\pi \leq \phi_i < \pi \) then all the points on the wavefront will continue to operate on the \( n=1 \) trajectory, where

\[
n = \text{INT} \left( \frac{\phi_n - \phi_i + \pi}{2\pi} \right),
\]

(3.24)

unless the physical limits of the modulator are encountered. Methods for occasionally resetting \( \phi_{mo} \) were mentioned in the last section.

However when the initial lock is acquired in the presence of input phase variations over multiple cycles (\( \phi_i = \phi_i + 2\pi m \), where \( -\pi \leq \phi_i \leq \pi \)), there is a \( 2\pi \) degeneracy in the phase estimate. The IPL will lock with

\[
\phi_m = \phi_i + 2\pi = \phi_i + 2(n+m)\pi = \phi_i + 2\pi
\]

(3.25a)

where

\[
\xi = n + m = \text{INT} \left( \frac{\phi_m + \pi}{2\pi} \right)
\]

(3.25b)

The \( \xi \) value of the initially acquired trajectory depends on the initial value, \( \phi_{mo} \), and is independent of \( m \). Using Eqs. (3.25a) and (3.20):

\[
\text{INT} \left( \frac{-\phi_{mo}}{2\pi} \right) \leq \xi \leq \text{INT} \left( \frac{\phi_{mo}}{2\pi} \right) + 1
\]

(3.26)

The initial trajectory is thus characterized by \( n(x,y) = \xi - m(x,y) \), where \( \xi \) is a predetermined constant and \( m \) depends on the input phase. In some instances, there may be an advantage to varying \( \xi \) spatially, instead of employing a uniform value for \( \phi_{mo} \).

Since the IPL can have a very large dynamic tracking range, it
is possible to avoid this $2m\pi$ degeneracy in the estimate of $\phi'_i$. For example, a smooth reference wavefront, with $2(m-1)\pi \leq \phi'_i \leq 2m\pi$, could be employed during initial phase acquisition, and then the corrupted signal could be introduced at a rate slow enough to be tracked. In an atmospheric compensation application, this amounts to having the system in operation before low-visibility conditions occur. In laboratory phase measurement applications, the phase of a thick, (multiple $\pi$ radian), transparent object could be visualized with a continuous grey level rather than fringes by sliding it into the input beam with the IPL already in operation. Some of the other more complicated phase estimators, such as those which reconstruct phase from a spatial derivative across the aperture $^{45,61,63-65}$, can also remove this degeneracy.

In many applications, such as adaptive phase compensation, it is not desirable to remove this degeneracy from the estimate; because a modulator with a large enough dynamic range to conjugate the maximum number of cycles of phase difference, which can possibly appear in the aperture, is not practical. In addition, if only the power at a diffraction-limited detector in the Fourier plane is of interest, perfect compensation may offer no advantage; since the Fourier transform of a plane wave is identical to the transform of a phase front with cells differing by $2m\pi$ radians. (However, with a monolithic modulator, the finite resolution will generally cause a phase error in the transition region between two resolution cells differing by $2\pi$)$^{31}$.

3.1.4 Spatial phase compensation performance of the IPL

Some of the effects of IPL imperfections on spatial phase
compensation will be mentioned in this section. One measure of wavefront compensation performance is the Strehl ratio, Q. As discussed in Appendix D, the Strehl ratio is approximately the intensity, $I_d$, seen by the direct detection communications receiver at D in Fig. 1.1, divided by the intensity which would be seen with perfect phase compensation. Utilizing Eqs. (D.6), (D.4), (A.2a), and (D.7) from the appendices, the Strehl ratio can be written:

$$Q = \frac{|E_i| \cos(\phi_i - \phi_m - \phi_0)^2 + |E_i| \sin(\phi_i - \phi_m - \phi_0)^2}{|E_i|^2}$$  \hspace{1cm} (3.27)

The brackets denote spatial averaging (e.g., over the exit aperture of the modulator). With perfect phase compensation, $Q=1$ results.

Simulated behavior

The IPL compensation behavior for wavefront fluctuations within the system's bandwidth can be found by assuming a wavefront $|E_i(x,y)e^{j\phi_i(x,y)}$ and using the IPL steady-state relations, e.g., Eq. (2.17a), to compute $\phi_m(x,y)$, given the parameters of the specific implementation (e.g., $\phi_0$, $\phi_1$, $\phi_2$, $G_o$ and $G_1$). The Strehl ratio can be computed from Eq. (3.27) by evaluating the appropriate integrals over the area, $A_m$; e.g.

$$\langle |E_i| \rangle = \frac{1}{A_m} \iint_{A_m} |E_i| \, dx \, dy$$  \hspace{1cm} (3.28)

When the system can be treated as having $N$ discrete resolution elements, the averages become summations;

$$\langle |E_i| \rangle = \frac{1}{N} \sum_{k=1}^{N} E_{i_k}$$  \hspace{1cm} (3.29)
As a specific example, Figs. 3.5 summarize the simulated phase compensation performance of a nine resolution element homodyne IPL with a type-zero loop filter. Eqs. (2.17a) and (3.27) were employed, with summations to compute the averages. An arbitrarily chosen, uncompensated wavefront is depicted in Fig. 3.5a along with the corresponding IPL compensated wavefront. The finite error of the compensated wavefront is not visible relative to the large phase range covered by Fig. 3.5a. Fig. 3.5b shows the computed evolution of the Strehl ratio as the nine channels are turned on one by one. The specific spatial location of the resolution elements has no effect on the Strehl ratio; only the relative phase of the resolution elements matters. Turning on the elements in the numbered sequence of Fig. 3.5a results in a decrease of Q before it increases as more channels become operational.

Approximate behavior

Appendix D shows that, with the introduction of various assumptions, more explicit statements can be made on the behavior of the Strehl ratio. For example, if $|E_i|$ and $\phi_i$ are statistically independent, $\phi_i$ is spatially ergodic, and $\phi_i$ is characterized by a Gaussian probability distribution with mean $<\phi_i>$ and variance $\sigma_i^2=<\phi_i^2>-<\phi_i>^2$; the corrupted Strehl ratio is:

$$Q = e^{-\sigma_i^2}$$

(3.30)

After compensation by a high gain homodyne IPL with a type-zero loop filter, Appendix D shows:

$$Q = \exp(-\frac{\sigma_i^2}{(1+G_1)^2})$$
Nine element homodyne IPL ($G_0 = 1.8\pi$, $G_1 = 70\pi$, initial value: $\phi_{m0} = 2\pi$ in all elements)

(a) Initial and compensated phase in each resolution element

(b) Evolution of Strehl ratio as elements are sequentially locked
Notice that the phase compensation is essentially perfect as long as the gain is much larger than the rms phase variation; i.e., $G_1 \gg \sigma_1$. The behavior of $Q$ with other IPL implementations can be treated in a similar manner.

3.1.5 Spatial Resolution

The discussions up to this point have assumed that the IPL completely resolves the input wavefront aberrations to be estimated and compensated. When the phase aberration spatial detail exceeds the system resolution, the conjugate impressed on the input wavefront is imperfect and overall IPL performance is degraded.

Monolithic IPL

The effects of limited spatial resolution in the detector-amplifier-modulator system of Fig. 2.3 can be represented by an integral of the form

$$\phi_m(x, y, t) = \int dt \int dx' dy' g(x, x', y, y'; t-t')(I_0 + I_1 \sin \phi_e)$$

(3.31a)

Here $g$ is a spatio-temporal response function which includes: spatial spreading of the interferogram at the detector due to resolution limitations of the optical system and misalignment, electronic crosstalk between spatial locations, and coupling between spatial locations at the modulator (i.e., the modulator "influence function")$^{50,166}$.

In many instances the temporal impulse response, $h(t)$, is spatially uniform, and at regions removed from the spatial boundary
of the system $g$ is spatially stationary (i.e., not a function of the actual position). It is then convenient to write $g$ as the product of an open-loop impulse response $h(t)$, point spread function (psf) $f(x, y)$, and gain constant $q_o$ (with units of radians cm$^2$/watt). The open-loop response integral presented above now becomes a spatio-temporal convolution:

$$
\phi_m(x, y, t) = [h(t)f(x, y)] * [G_o + G_1 \sin \phi_e]
$$

(3.31b)

where

$$
G_o = g_o I_o \quad \text{and} \quad G_1 = g_o I_1
$$

With perfect resolution $f(x, y)$ is a spatial impulse, $\delta(x)\delta(y)$, and passes through the convolution. Eq. (3.31b) then becomes a temporal-only convolution and represents the IPL dynamic behavior discussed in previous sections. For arbitrary resolution, additional insight can be obtained by taking the spatio-temporal Fourier transform of Eq. (3.31b):

$$
\tilde{\phi}_m(u, v, w) = \tilde{h}(w)\tilde{f}(u, v)\left[\tilde{G}_o(u, v, w) + \tilde{G}_1(u, v, w) * \tilde{f}(\sin \phi_e(x, y, t))\right]
$$

(3.31c)

Based on the previous infinite resolution results, it is reasonable to expect that as long as $G_1 > 1$ holds, $\phi_e$ should usually become small enough to allow linearization of the sinusoid. Using the relation $\phi_e = \phi_1 - \phi_m$ and assuming that $G_o$ and $G_1$ are approximately spatially and temporally constant, the linearized
closed-loop response is

\[ \hat{\phi}_m(u,v,w) = \frac{\hat{H} \hat{F} G_1}{1 + \hat{H} \hat{F} G_1} \left[ \hat{\phi}'_i(u,v,w) + \left( \frac{G_0}{G_1} - 2n\pi \right) \delta(u,v,w) \right] \] (3.31d)

\[ \hat{\phi}_e(u,v,w) = \frac{\hat{\phi}'_i(u,v,w) - (G_0 - 2n\pi) \delta(u,v,w)}{1 + \hat{H} \hat{F} G_1} \] (3.31e)

Equations (3.31d,e) are a more general statement of the temporal-frequency response results of Eqs. (2.36) and (2.37). Thus, spatial frequencies within the spatial bandwidth of the system (for which \( \hat{G}_1 \hat{F}(u,v) > 1 \)) are tracked by the modulator and removed from the phase error; i.e. \( \hat{\phi}_m(u,v) \approx \hat{\phi}'_i(u,v) \) and \( \hat{\phi}_e(u,v) \approx 0 \). Conversely, phase aberration spatial frequencies which exceed the spatial resolution of the system are tracked with diminished closed-loop gain, \( \hat{H} G_1 \hat{F}(1 + \hat{H} G_1 \hat{F}) \approx \hat{H} G_1 \hat{F}(u,v) + o \), and are not compensated from the phase error; i.e.,

\[ \hat{\phi}_e(u,v) = \hat{\phi}'_i(u,v)/(1 + H G_1 \hat{F}(u,v)) \approx \hat{\phi}'_i(u,v). \]

Equation (2.31c) also provides insight into the effects of tilt misalignment between the input and L.O. beams. Since tilt introduces spatial ramp terms into \( \hat{\phi}'_i \), it is convenient to write

\[ \phi_e \equiv \phi_{eo} + u_0 x + v_0 y, \] (when there is no tilt, \( \phi_e = \phi_{eo} \)).
Using
\[ \mathcal{H}(\sin \phi_0) = \frac{1}{2j} (\hat{W}_+ (u, v) - \hat{W}_- (u, v)) \text{ where } \hat{W}_\pm = \mathcal{F}(e^{\pm j\phi_0}), \]

Eq. (2.31c) becomes:
\[ \hat{\phi}_m = \mathcal{H} \mathcal{F}(G_0 + G_1 \ast [\hat{W}_+ (u - u_0, v - v_0) - \hat{W}_- (u + u_0, v + v_0)]) / 2j \]

The spatial spectrum of the interferogram is thus shifted by the spatial carrier frequencies, \( u_0, v_0 \), introduced by the tilt. (For example a tilt of \( \theta \) in the \( x \) direction between the L.O. and signal beams introduces the spatial carrier \( u_0 = 2\pi \sin \theta / \lambda \)). The IPL will estimate and compensate the tilt, but with degraded performance as \( u_0 \) and \( v_0 \) become large enough to move the spatial spectrum outside of the spatial bandwidth of \( \mathcal{H}(u, v) \).

It should be noted that the spatial variation in the visible interferogram provides an indication of the spectral content of \( \sin \phi_0 \), which the IPL is required to resolve. Consider, for example, a sinusoidal grating plus tilt, \( \phi_0 = u_0 x + b \sin(u_1 x) \). The one-dimensional spectrum seen by the IPL in Eq. (2.31b) is then
\[ \mathcal{H}(\sin \phi_0) = \frac{1}{2j} \sum_{n=-\infty}^{\infty} J_n(b)[\delta(u - (u_0 + n u_1)) - \delta(u + (u_0 + n u_1))] \]

where \( J_n(b) \) is an \( n^{th} \) order Bessel function. If \( b << 1 \), only \( J_0 \) is important and the period of the spatial carrier,
is directly visible in the interferogram. With \( u_0 = 0 \) and \( b < \pi \), the approximation \( ^{100} \sin \phi_e \approx -jJ_1(b)[\delta(u-u_1) - \delta(u+u_1)] \) holds, and a low-contrast periodic fluctuation with period \( 2\pi/u_1 \) is visible. With \( u_0 = 0 \) and \( b > \pi \), \( ^{100} \sin \phi_e \approx -jJ_m(b)[\delta(u - mu_1) - \delta(u + mu_1)] \) where \( m = \text{INT}(b) \) holds, and a dominant period of \( \approx 2\pi/bu_1 \) can be seen in the interferogram.

An interesting implication of Eqs. (2.31d,e) is that if \( G_0 \) is spatially uniform (e.g., \( |E_x| > |E_y| \) and \( |E_y| \) is spatially uniform) and \( \hat{H}(u,v) \) has a spatial passband characteristic with negligible response at zero frequency, the offset term, \( G_0 \), can be suppressed. Operation with \( G_1 > G_0 \) generally improves such IPL performance characteristics as immunity to amplitude fluctuations, residual phase error, and phase acquisition time. Low frequency information in \( \phi_1 \) will not be lost if a tilted L.O. beam is used to modulate the spectrum of \( \sin \phi_e \) into the bandpass. Some actual monolithic optically-addressed phase modulators, such as the Ruticon\(^{188-190}\) or devices employing the photorefractive effect\(^{198}\), have a spatial passband characteristic.

When \( \phi_e \) is large, additional insight can be gained by assuming a specific \( \hat{H} \hat{F} \) response and transforming Eq. (3.31c) back into a differential equation in the time-space domain. For example with \( \hat{H} = w_0/(w_0 + jw) \) and \( \hat{F} = u_0 v_0/(u_0 + ju)(u_0 + jv) \) (a choice of \( \hat{F} \) with cylindrical symmetry may be more reasonable), the dynamic equation corresponding to Eq. (3.31b) is:
\[
\frac{\partial^3 \phi_m}{\partial x \partial y \partial t} + \frac{\partial^2 \phi_m}{\partial y \partial t} + \frac{\partial^2 \phi_m}{\partial x \partial t} + \frac{\partial^2 \phi_m}{\partial x \partial t} + \frac{\partial^2 \phi_m}{\partial y \partial t} + \frac{w_0}{u_0} \frac{\partial \phi_m}{\partial x} + \frac{w_0}{u_0} \frac{\partial \phi_m}{\partial y} + \frac{w_0}{u_0} \frac{\partial \phi_m}{\partial t} = w_0 u_0 v_0 (-\phi_m + G_o + G_1 \sin(\phi_i - \phi_m))
\]

Partial differential equations of this form can sometimes result in spatially distributed large-scale modes which are also a function of the initial \((t=0)\) state and spatial boundary conditions. This system is damped; and it can be shown that it eventually reaches a steady-state characterized by \(\frac{\partial \phi_m}{\partial t} = 0\) and the spatial partial differential equation

\[
\frac{\partial^2 \phi_m}{\partial x \partial y} + \frac{u_0}{\partial x} \frac{\partial \phi_m}{\partial x} + \frac{v_0}{\partial y} \frac{\partial \phi_m}{\partial y} = u_0 v_0 (-\phi_m + G_o + G_1 \sin(\phi_i - \phi_m))
\]

(3.31f)

This equation describes the steady-state spatial distribution of the modulator and is solved by the convolution \(\phi_m(x,y) = f \ast (G_o + G_1 \sin \phi_e)\).

The linearized result of Eq. (3.31c-d) suggests that the closed-loop spatial resolution can be improved by increasing \(G_1\). For example, the linearized response of Eq. (3.31f) to a step \(\phi_{i0} U(x)\) in the \(x\) direction (with \(\phi_m(x = 0) = 0\)) is

\[
\phi_m(x > 0) = \frac{G_1}{1 + G_1} (\phi_{i0} + \frac{G_o}{G_1} + 2 \pi n)(1 - e^{-x/L}),
\]

where \(L = [u_0(1 + G_1)]^{-1}\) is a measure of the resolution scale-size.
Discrete-Channel IPL

In a discrete channel IPL, the spatial integration in Eq. (3.31a) becomes a summation of small-area integrals over each detector element. The spatial dependence of $g$ expresses effects which induce coupling between channels, such as the influence function of each modulator element. In many instances $g$ does not cause strong coupling to adjacent channels; then the Nyquist sampling theorem imposes the primary resolution constraint. The sampling theorem essentially requires that the distance, $P$, between modulator elements should be less than half of the minimum spatial period to be compensated; or that $1/P > 2 u_m$, where $u_m$ is the maximum significant spatial frequency in $\pi / (\sin \phi_e)$. This condition often amounts to requiring two detector elements per fringe in the visible interferogram, so that each detector element sees a uniform shade of grey rather than fringes.

Spatial frequency components in excess of $1/2P$ are generally averaged out by the detector integrals of Eq. (3.31a), and hence invoke little response in $\phi_m$ and remain uncompensated in $\phi_e$. The major exception to this statement is uniform fringes across each detector element, such as those induced by a tilted L.O. beam. In his "wavefront alignment function" discussion, McDowell\textsuperscript{2} shows that the detector integrals of Eq. (3.31a) remain sensitive to the untitled phase $\phi_{eo}(\equiv \phi_e - u_0 x - v_0 y)$, but the system gain is reduced by a factor similar to $\text{sinc}(d_0 u_0/2\pi)$, where $d_0$ is the diameter of a detector element. There is no response when an integral number of fringes appears across a detector element, because as these fringes
shift with $\phi_{eo}$, there is no change in the spatially integrated intensity. In order to function with L.O. tilt it is still required that the spatial spectrum of $\phi_{eo}$ not exceed $1/2P$ and that the tilt be uniform (within $1/2P$) across the whole detector array (since individual modulator elements cannot correct tilt).

In summary, a monolithic SLM estimates and compensates phase fluctuations for which $G^* H F(u,v) > 1$, where $F(u,v)$ is the optical transfer function (OTF) of the detector-to-modulator subsystem. Spatial spectral components outside this spatial bandwidth remain uncompensated in $\phi_e$. A discrete-element IPL is limited by the sampling theorem to spectral components less than $1/2P$, where $P$ is the spacing between modulator elements.

3.1.6. Other loop filters

The discussion up to this point has been couched in terms of an IPL with a type-zero loop filter, and in particular, a low-pass filter. This section discusses some of the new aspects of IPL behavior introduced by a type-one integrating filter and by two-pole filters.
3.1.6.1. Integrating type-one loop filter

With the integrating loop filter of Eq. (3.32), the general homodyne operating Equations, (2.12) and (2.13), become Eqs. (3.33).

\[ \dot{H}(s) = \frac{\lambda_1}{s} \quad (3.32) \]

\[ \dot{\phi}_m = \lambda_1 (G_o - G_1 \sin(\phi_m - \phi_i^t)) \quad (3.33a) \]

and

\[ \dot{\phi}_e = \dot{\phi}_i^t - \lambda_1 (G_0 + G_1 \sin \phi_e) \quad (3.33b) \]

Note that with a large time constant, \( \tau = 1/\lambda \), the low-pass filter of Eq. (2.28) becomes an integrator. Much of the integrating behavior can thus be viewed as a special case of the low-pass behavior, in the limits \( \lambda \rightarrow 0 \), \( G_1 \rightarrow \infty \), and \( G_0 \rightarrow \infty \), with the constraints \( \lambda G_0 \rightarrow \lambda_1 G_0 \) and \( \lambda G_1 \rightarrow \lambda_1 G_1 \). Eqs. (3.33) then follow from the corresponding low-pass relations, Eqs. (2.29) and (2.30).

The state-space plot (\( \dot{\phi}_m \) vs \( \phi_m \)) of Fig. 3.6a reveals that the homodyne IPL with an integrating loop filter can have equilibria at:

\[ \dot{\phi}_m = \phi_i^t + \sin^{-1} \left( \frac{G_0}{G_1} \right) - 2n\pi \quad (3.34a) \]

and

\[ \phi_e = -\sin^{-1} \left( \frac{G_0}{G_1} \right) + 2n\pi \quad (3.34b) \]

where

\[ n = INT \left( \frac{\phi_e^t + \pi}{2\pi} \right) \quad (3.34c) \]

Notice that the explicit steady-state expressions of Eqs. (3.34) can be obtained without resorting to graphical techniques. As the phase estimation performance plot of Fig. 3.6b shows; \( \phi_m \) is directly equal to \( \phi_i^t \) plus an offset phase which is independent of \( \phi_i^t \). Unlike the low-
Fig. 3.6a Modulator state-space ($\phi_m$ versus $\psi_m$) with an integrating loop filter.

Fig. 3.6b Phase estimation performance with an integrating loop filter.

Fig. 3.6 Basic behavior of the IPL with an integrating loop filter.
pass case, the IPL with an integrator becomes unstable with $G_1 < G_0$. Then according to Fig. 3.6a, $\dot{\phi}_m = 0$ cannot occur; and $\phi_m$ moves off to the right to lock up at $\phi_m = \phi_s$.

**Insensitivity to amplitude fluctuations**

Equations (3.34) and Fig. 3.6a suggest that as long as $G_1 \gg G_0$ holds, the integrating IPL is not sensitive to amplitude fluctuations.

Changes in $G_0$ and $G_1$ will only perturb the already small error term. As $|E_i|$ decreases, eventually $G_0 > G_1$ can occur and $\phi_e$ becomes unstable, unless the IPL is designed with $I'_t = |E_{x\perp}|^2$ in Eq. (3.4a); the error then becomes

$$\phi_e = \sin^{-1}\left(\frac{|E_i|}{2|E_{x\perp}|}\right) \quad (3.35a)$$

$$\quad (\text{for } I'_t = |E_{x\perp}|^2) \quad (3.35b) \)$$

By employing a strong enough L.O. beam, i.e. $|E_{x\perp}| > |E_i|$, the effects of amplitude fluctuations can thus be made minimal. This behavior is similar to the low-pass IPL behavior in Figs. 3.2 (case 3); however the immunity of the integrating IPL to amplitude variations is superior since Eqs. (3.35) continue to remain valid for all values of $\phi'_i$ and $\phi_e$. Note that $\phi_e \to 0$ for $|E_i| \to 0$.

In practice, the ability to adjust $I'_t = |E_{x\perp}|^2$ is very dependent on the specific hardware configuration. (It will also be seen that it is often more desirable for $I'_t$ to be an electronic offset rather than a threshold). Some of the experimental systems in chapters 5 and 6 can be operated to satisfy Eqs. (3.35). The minimum required irradiance, $|E_i|^2$, is then determined by other considerations, such as noise performance and convergence time.
Linearized behavior

When the gain is large and the system is near equilibrium, the phase error of Eq. (3.34b) becomes small and the dynamic equations of the integrating IPL can be linearized. Equation (3.33a) becomes

\[ \dot{\phi}_m + \lambda_1 G_1 \phi_m = \lambda_1 G_1 (\phi'_i + G_0/G_1 - 2n\pi) \]  \hspace{1cm} (3.36)

With small error, the approximate bandwidth of the IPL with an integrating loop filter is thus

\[ BW = \lambda_1 G_1 \]

The combined step and ramp responses of the modulator and error to the input \( \phi'_i = \phi'_{i0} + \Delta wt \)

\[ \phi_m = \phi_{m0} e^{-\lambda_1 G_1 t} + \phi_{mss} (1 - e^{-\lambda_1 G_1 t}) + \Delta wt \]  \hspace{1cm} (3.37a)

\[ \phi_e = (\phi'_{i0} - \phi_{m0}) e^{-\lambda_1 G_1 t} + (\phi'_{i0} - \phi_{mss}) (1 - e^{-\lambda_1 G_1 t}) \]  \hspace{1cm} (3.37b)

Here \( \phi_{m0} \) is the initial modulator phase and

\[ \phi_{mss} = \phi'_{i0} + G_0/G_1 - 2n\pi - \frac{\Delta W}{\lambda_1 G_1} \]  \hspace{1cm} (3.37c)

(For the homodyne IPL, \( \Delta W = 0 \).)

State-space "dead regions"

In order to obtain Fig.3.6a, Equation (2.8) in the basic derivation was modified to include a negative offset in the amplifier:

\[ \hat{I}_1(s) = g_1 \hat{H}(s) \hat{I}_0(s) - \frac{i_{ocutoff}}{s} \]  \hspace{1cm} (3.38)

In practice, this may introduce excessive complexity into a high-resolution adaptive optical system.
The state-space plot corresponding to the simple system of Eqs. (2.6)-(2.8) is shown in Fig. 3.7a for two values of $G_1$. This IPL is afflicted with "dead regions", where $\phi_m=0$ and $\phi_m$ remains trapped until $\phi_i$ changes enough. In practice $\phi_m$ may shift slowly to the left if the integrator "leaks" (i.e., is a low-pass filter with large time constant), or to the right if there is dark current or enough stray light to exceed the detector threshold. In most instances the "dead region" drift can be neglected since it is characterized by a time-constant much longer than the time during which the IPL is operated.

The phase estimation performance ($\phi_m$ vs. $\phi_i$) of this system is depicted in Figs. 3.8 for the initial step response and phase tracking with two values of $G_1$. This system can estimate a step if the initial value of $\phi_m$ is in an active region, and can track increasing phase until the modulator limit is reached. However decreasing phase cannot be tracked; and with bidirectional fluctuations in $\phi_i$, $\phi_m$ will eventually lock-up in saturation, with $\phi_m=\phi_s$. The low-gain state-space trajectory of Fig. 3.7 results in superior phase estimation performance, as is evident in Fig. 3.8b. The phase error, $\phi_e = \pi/2$, is relatively large; but being spatially uniform, it can be included in the reference phase. Unfortunately, since $G_0 = G_1$, the low-gain system performs poorly when $|E_i|$ also varies. The heterodyne discussion of section 3.2.1 will show that adding a continuously increasing component to the input phase improves the performance of this system.

Since the integrator never sees a negative input signal; its output can only increase and never decrease. In practice, the integrator could be occasionally reset, resulting in a discontinuous framed-mode of operation. In the framed mode, during the rapid initial convergence
a) Modulator state-space ($\dot{\phi}_m$ versus $\phi_m$).

b) Error state-space ($\dot{\phi}_e$ versus $\phi_e$).

Fig. 3.7 State-space characteristics with "dead regions".
Fig. 3.8 Phase estimation performance ($\phi_m$ versus $\phi_i'$) with "dead regions".
step response, \( \phi_i^* \) appears static. The phase estimate (and compensated output) is then held constant during a storage interval in which \( \phi_i^* \) continues to remain fairly constant; this is easily implemented by interrupting the input signal to the integrator. The frame ends by rapidly discharging the integrator. This framed mode requires a modulator dynamic range of only \( 2\pi \) radians; since it must only initially acquire, and not track, the input phase.

**Time response expressions**

The homodyne IPL with an integrating loop filter is characterized by the differential equation (3.33a), which can be rewritten

\[
\dot{\phi}_m = G'_0 + G'_1 \sin \phi_e
\]  

(3.39)

where \( G'_0 = \lambda_1 G_o, \ G'_1 = \lambda_1 G_l \) and \( \phi_e = \phi_i^* - \phi_m^* \).

There are generally limits on \( \dot{\phi}_m \) and \( \phi_m^* \); i.e., \( G'_c \leq \dot{\phi}_m^* \leq G'_s \) and \( 0 \leq \phi_m^* \leq \phi_s \). Fig. 3.6a is an example of a \( \dot{\phi}_m^* \) vs \( \phi_m^* \) state-space plot.

The time required for \( \phi_m^* \) to transit between the two phases \( \phi_a \) and \( \phi_b \) is:

\[
t_{ab} = \frac{\int_{\phi_a}^{\phi_b} d\phi_m}{\dot{\phi}_m} = \frac{\int_{\phi_a}^{\phi_b} d\phi_m}{\dot{\phi}_m} G'_0 + G'_1 \sin (\phi_i^* - \phi_m^*)
\]  

(3.40)

Since the transit time is usually much faster than fluctuations in \( \phi_i^* \), it is reasonable to assume \( \dot{\phi}_e = \dot{\phi}_i^* - \dot{\phi}_m = -\dot{\phi}_m \), which allows Eq. (3.40)
to be rewritten as

\[ t_{ab} = \int_{a}^{b} \frac{d\phi}{G_{0} + G_{1} \sin \phi} \]  \hspace{1cm} (3.41)  

where \( a = \phi_{ea} = \phi_{i} - \phi_{a} \) \hspace{1cm} (3.42a), \hspace{1cm} and \hspace{1cm} \( b = \phi_{eb} = \phi_{i} - \phi_{b} \) \hspace{1cm} (3.42b)

This integral is readily evaluated.\(^{167}\)

With stable equilibria

When \( G'_{0} < G'_{1} \), Eq. (3.39) and Fig. 3.6a show that stable equilibria \((\phi_{m} = 0)\) can occur at \( \phi_{e} = \phi_{es} = -\sin^{-1}(G_{0}/G_{1}) + 2n\pi \). With \( G'_{0} < G'_{1} \), the time for \( \phi_{e} \) to transit from \( \phi_{ea} \) to \( \phi_{eb} \) follows from Eqs. (3.41) and (3.42) as

\[ t_{ab} = \frac{1}{\lambda_{0}} \ln \left[ \frac{G'_{0} \tan \frac{a}{2} + G_{1} - \lambda_{0}}{G'_{0} \tan \frac{a}{2} + G_{1} + \lambda_{0}} \right] \cdot \ln \left[ \frac{G'_{0} \tan \frac{b}{2} + G_{1} + \lambda_{0}}{G'_{0} \tan \frac{b}{2} + G_{1} - \lambda_{0}} \right] \]  \hspace{1cm} (3.43a)

Here

\[ \lambda_{0} = \sqrt{G_{1}^{2} - G_{0}^{2}} = \lambda_{1} \sqrt{G_{2}^{2} - G_{0}^{2}} \]  \hspace{1cm} (3.43b)

This is only valid if no equilibria are included in the range \( b < \phi < a \), and the dynamic limits of \( \phi_{m} \) and \( \phi_{m} \) are not encountered.

This expression can be simplified by rewriting it about \( \phi_{es} \) with the new variables \( a' = \phi_{es} - \phi_{ea} \) and \( b' = \phi_{es} - \phi_{eb} \), where \( |a'| > |b'| \); then
\[ t_{ab} = \frac{1}{\lambda_o} \ln \left[ \frac{\cot \frac{b'}{2} - \frac{G_0'}{G_o}}{\cot \frac{a'}{2} - \frac{G_0'}{G_o}} \right] \]  

(3.44)

The following relations, employing trigonometric identities, were used to obtain Eq. (3.44)

\[-\tan \frac{\phi_{es}}{2} = \tan \left( \frac{1}{2} \sin^{-1} \frac{G_0'}{G_1} \right) = \frac{G_0'}{G_1 + \lambda_o} = \frac{G_1 - \lambda_o}{G_1 + \lambda_o} = \frac{G_1 - \lambda_o}{\sqrt{G_1^2 + \lambda_o^2}} \]  

(3.45)

Eq. (3.44) can be rewritten as an exact expression for the error time response from \( \phi_e(t=0) = \phi_{ea} \) to \( \phi_e(t) = \phi_{eb} \):

\[ \phi_e(t) = \phi_{es} - 2\cot^{-1} \left[ \cot \left( \frac{\phi_{es} - \phi_{ea}}{2} \right) e^{\frac{\lambda_o t}{\lambda}} \right] + \frac{G_0'}{\lambda} (1 - e^{-\lambda_o t}) \]  

(3.46)

The corresponding modulator response is \( \phi_m(t) = \phi_i - \phi_e(t) \).

An important case is the final convergence to \( \phi_{es} \). Assuming an initial condition near \( \phi_{es} \), \( \cot(a'/2) = 2/a' \) and Eq. (3.46) can be approximated by:

\[ \phi_e(t) \approx \phi_{es} - \frac{\left( \phi_{es} - \phi_{ea} \right) e^{\frac{-\lambda_o t}{\lambda}}}{1 + \frac{1}{2} \left( \phi_{es} - \phi_{ea} \right) \frac{G_0'}{\lambda} (1 - e^{-\lambda_o t})} \]  

(3.47)

The convergence to equilibrium is thus exponential and takes infinite time; however after a few time constants, \( \tau = \lambda_o^{-1} \), the IPL is for all practical purposes at equilibrium. With large gain, \( G_1 >> G_0, \lambda_o \) in Eq. (3.43) becomes \( \lambda_1 G_1 \), and \( \tau \) is identical to the time constant found by linearizing Eq. (3.33a).

If small perturbations from equilibrium occur less rapidly than \( \tau \), the system remains in quasi-equilibrium and tracks these perturbations. Thus \( \lambda_o \) can be considered the effective bandwidth of the IPL with an integrating
loop filter. Note that the bandwidth, $BW$, decreases with increasing phase error, i.e.,

$$BW = \omega_0 = \sqrt{G_1^2 - G_0^2} = G_1 \sqrt{1 - \sin^2 \phi_s}$$  \hfill (3.48)

In regions of the state-space trajectory where intensity threshold ($\phi_m = G_c$) or saturation ($\phi_m = G_s$) are encountered;

$$t_{ab} = \frac{\phi_{eb} - \phi_{ea}}{G_c}$$  \hfill (3.49a)

or

$$t_{ab} = \frac{\phi_{ea} - \phi_{eb}}{G_s}$$  \hfill (3.49b)

respectively; and the time response is $\phi_e(t) = \phi_{ea} + G_c t$ or $\phi_e = \phi_{ea} - G_s t$.

With no stable equilibria

When $G_0 > G_1$, there are no stable equilibria and Eq. (3.41) is more conveniently evaluated as

$$t_{ab} = \frac{2}{w} \tan^{-1} \left( \frac{w \cdot \sin \left( \frac{a-b}{2} \right)}{G_1 \cos \left( \frac{a-b}{2} \right) + G_0 \sin \left( \frac{a+b}{2} \right)} \right)$$  \hfill (3.50a)

where

$$w = \sqrt{G_0^2 - G_1^2}$$  \hfill (3.50b)

The time behavior of $\phi_e(t)$ from initial condition $\phi_e(t=0) = \phi_{ea}$ is

$$\phi_e(t) = 2 \tan^{-1} \left[ \frac{\frac{\phi_{ea}}{2} - \frac{1}{w} (G_1 + G_0 \tan \frac{\phi_{ea}}{2}) \tan \frac{wt}{2}}{1 + \frac{1}{w} (G_1 + G_0 \tan \frac{\phi_{ea}}{2}) \tan^2 \frac{wt}{2}} \right]$$  \hfill (3.51a)

From an initial condition of $\phi_{ea} = 2n\pi$: 
\[ \phi_e(t) = -2\tan^{-1}\left(\frac{G'_0 \tan\frac{wt}{2}}{w + G'_1 \tan^2\frac{wt}{2}}\right) \] (3.51b)

These expressions assume that the dynamic limits of \( \phi_n \) and \( \dot{\phi}_m \) are not exceeded. Examination of Eq. (3.59) reveals that \( \phi_e \) is periodic, with period

\[ T_c = \frac{2\pi}{w} = \frac{2\pi}{\sqrt{G'_0^2 - G'_1^2}} \] (3.52)

The behavior of \( \phi_e \) as a function of time is illustrated in Fig. 4.17a of the next chapter. Examining the state-space plot \( (\dot{\phi}_e \text{ vs } \phi_e) \) of Fig. 4.16b, it is not surprising that the time response is symmetric about \(-\pi/2\) and \(-3\pi/2\). In particular:

\[ t_0 \text{ to } -\frac{\pi}{2} = t \frac{\pi}{2} \text{ to } -\pi = \frac{T_c}{\pi} \tan^{-1}\left(\sqrt{\frac{G'_0 + G'_1}{G'_0 - G'_1}}\right) \] (3.53a)

\[ = \frac{T_c}{2} \text{ for } G_0 \rightarrow G_1 \]

and

\[ t_{-\pi} \text{ to } -\frac{3\pi}{2} = t \frac{3\pi}{2} \text{ to } -2\pi = \frac{T_c}{\pi} \tan^{-1}\left(\sqrt{\frac{G'_0 - G'_1}{G'_0 + G'_1}}\right) \] (3.53b)

\[ = 0 \text{ for } G_0 \rightarrow G_1 \]

With \( G_0 \rightarrow G_1 \), most of the time is spent in the vicinity of \( \phi_{eu} \equiv (3-4n)\pi/2 \).

The time to transit from a phase \( a = \phi_{eu} + \Delta \) to \( b = \phi_{eu} - \Delta \) is

\[ t_{+\Delta} \text{ to } -\Delta = 2t_{\Delta} \text{ to } \phi_{eu} = \frac{2T_c}{\pi} \tan^{-1}\left(\sqrt{\frac{G'_0 + G'_1}{G'_0 - G'_1}} \frac{1 + \tan\left(\frac{\Delta}{2} - \frac{\pi}{4}\right)}{1 - \tan\left(\frac{\Delta}{2} - \frac{\pi}{4}\right)}\right) \] (3.54)

When \( \Delta \) is small, this expression simplifies to

\[ t_{\Delta} \text{ to } -\Delta \sim \frac{2T_c}{\pi} \tan^{-1}\left(\sqrt{\frac{G'_0 + G'_1}{G'_0 - G'_1}} \frac{\Delta}{2 + \Delta}\right) \] (3.55)
\[ \tan\left(\frac{\Delta}{2} - \frac{\pi}{4}\right) = -\frac{\cos \Delta}{1 + \sin \Delta} \]

The time expressions developed here will be employed in the next chapter, in the context of IPLs constructed with integrating spatial light modulators.

### 3.1.6.2. Two-pole loop filter

Loop filters with more complicated frequency responses can often be adequately characterized by their two most dominant poles. A general type-zero, two-pole loop filter is:

\[ H(s) = \frac{\lambda_1 \lambda_2}{(s + \lambda_1)(s + \lambda_2)} \quad (3.56) \]

The corresponding operating equations of the homodyne IPL follow from Eqs. (2.12) and (2.13) as

\[ \ddot{\phi}_m + (\lambda_1 + \lambda_2) \dot{\phi}_m + \lambda_1 \lambda_2 \phi_m = \lambda_1 \lambda_2 (G_0 + G_1 \sin(\phi_i - \phi_m)) \quad (3.57a) \]

and

\[ \ddot{\phi}_e + (\lambda_1 + \lambda_2) \dot{\phi}_e + \lambda_1 \lambda_2 \phi_e = \dot{\phi}_i + (\lambda_1 + \lambda_2) \dot{\phi}_i + \lambda_1 \lambda_2 (\phi_i - G_0 - G_1 \sin \phi_e) \quad (3.57b) \]

As one would expect, the equilibria (where \( \ddot{\phi}_m = \dot{\phi}_m = 0 \)), of this higher order system are identical to the basic type-zero equilibria of Eqs. (2.17) and (2.13). The stability of these solutions can be investigated by studying state-space plots, such as \( \dot{\phi}_m \) versus \( \phi_m \). Unlike the single-pole systems, the dynamic equations, Eqs. (3.57), do not specify a single unique state-space trajectory, but rather an infinite family of trajectories. These trajectories only intersect at equilibria and can be uniquely identified by specifying a non-equilibrium point on them, such as the initial condition \( \dot{\phi}_m^{(0)} \) and \( \dot{\phi}_m^{(0)} \) at \( t=0 \). The \( \phi_m \) state-
space trajectories can be investigated with the aid of the slope information contained in Eq. (3.57a), i.e.

\[
S = \frac{d\phi_m}{d\phi_m} = \frac{\ddot{\phi}_m}{\dot{\phi}_m} = -\left(\lambda_1 + \lambda_2\right) + \lambda_1 \lambda_2 \left(\frac{-\dot{\phi}_m + G_c G_1 \sin(\phi_t - \phi_m)}{\dot{\phi}_m}\right)
\]

The first term corresponds to exponential decay; and analytic geometry reveals that the second term can result in a spiral into a type-zero equilibrium, where the numerator and denominator of the second term go to zero and \( S \) becomes indeterminant.

Eq. (3.57a) cannot be directly solved for the state-space or time behavior of \( \phi_m \), however numeric procedures such as iterating through the difference equations \( \dot{\phi}_m = \dot{\phi}_m + S \Delta \phi_m \) and \( \phi_{m+1} = \phi_m + \Delta \phi_m \) can be used. More information can be gained by employing time as a parameter. Expanding \( \dot{\phi}_m \) and \( \phi_m \) in a Taylor series,

\[
\phi_{m+1} = \phi_m + \ddot{\phi}_m \Delta t + \frac{\dddot{\phi}_m}{2} \Delta t^2 \tag{3.58a}
\]

\[
\dot{\phi}_{m+1} = \dot{\phi}_m + \ddot{\phi}_m \Delta t \tag{3.58b}
\]

\[
t_{n+1} = t_n + \Delta t \tag{3.58c}
\]

With \( t \) normalized in terms of \( \lambda_1 \), i.e., \( t_n = n / \lambda_1 \), \( \ddot{\phi}_m \) follows from Eq. (3.57a) as

\[
\ddot{\phi}_m = -(1 + \lambda) \dot{\phi}_m + \lambda \left( -\phi_m + G_c + G_1 \sin(\phi_t - \phi_m) \right) \tag{3.58d}
\]

where \( \lambda = \lambda_2 / \lambda_1 \).
State-space trajectories, resulting from computer programs implementing Eqs. (3.58) and related numeric procedures such as the Runge-Kutta Nystrom method, are illustrated in Figs. 3.9.

The state-space trajectories in Figs. 3.9 are for system parameters, \((G_o=1.8\pi\text{ and } G_1=4\pi)\), similar to those of the single-pole trajectory in Fig. 2.7. The dashed curve in Fig. 3.9a is the corresponding single-pole trajectory. The time evolution of \(\phi_m\) is represented on a few trajectories as dots separated by the time interval \(0.1/\lambda_1\). Fig. 3.9b shows the effect of the relative values of the two poles. The \(\lambda=10\) curve is close to the single-pole trajectory, since with large \(\lambda_2\) Eq. (3.56) becomes a single-pole filter. The \(\lambda=0.1\) curve is close to a single-pole trajectory, characterized by \(\lambda_2=0.1\lambda_1\).

For the given gain \(G_1\) and offset \(G_o\), the trajectories in Figs. 3.9 are always stable. The initial phase acquisition, however, may require \(\phi_m\) to skip cycles, varying over more than \(2\pi\) radians. The time dots in Fig. 3.9a show that the convergence time is generally longer than in the single-pole case, where the approximately linear segments of the trajectory have a time constant of about \((G_1+1)\lambda_1\). (The interval between the time dots in Fig. 3.9 corresponds to 1.36 single-pole time constants).

Trajectories were computed for a variety of additional \(G_1, G_o\), and \(\lambda\) combinations. These always reached stable equilibria. In general as \(G_1\) was increased: the convergence time decreased; there was less tendency to skip cycles; more stable equilibria came into existence; and the spiral to equilibrium contained larger \(\phi_m\) excursions. The last effect is expected from the behavior of the linearized (small \(\phi_e\)) dynamic equations. The root locus of a linear two-pole feedback system goes from overdamped
Fig. 3.9 Modulator state-space with a two-pole type-zero loop filter. ($\lambda = \frac{\lambda_0}{\lambda_1}$. Dots are spaced at $0.1/\lambda_1$ in time. $G_0 = 1.8\pi$, $G_1 = 4\pi$.)
(no state-space spiral) to larger underdamped oscillation frequencies as the system gain increases\textsuperscript{161}.

A type-one, two-pole loop filter of the form

$$\hat{H}(s) = \frac{\lambda_1 \lambda_2}{s(s+\lambda_1)}$$

results in significantly different behavior. The dynamic equation of the modulator becomes

$$\ddot{\phi}_m + \lambda_1 \dot{\phi}_m = \lambda_1 \lambda_2 (G_0 + G_1 \sin(\phi_i - \phi_m)) \quad (3.59)$$

Since this equation is periodic in $\phi_m$, the $\dot{\phi}_m$ vs $\phi_m$ state-space can be fully represented by the behavior in a $2\pi$ radian range of $\phi_m$.

The state-space trajectories in Fig. 3.10 were computed using Eq. (3.59) in a program similar to that employed to obtain Figs. 3.9. With large gain $G_1$, Fig. 3.10a shows that the trajectories spiral into the basic type-one equilibria of Eqs. (3.34). As the gain decreases, stable equilibria continue to exist as long as $G_1 > G_0$ holds; the bold line in Fig. 3.10b, however, reveals that a stable limit-cycle in $\dot{\phi}_m$ also develops. Along this limit-cycle $\dot{\phi}_m(\phi_m) = \dot{\phi}_m(\phi_m + 2\pi)$, and the average value of $\dot{\phi}_m$ is approximately $G_0$. This $\dot{\phi}_m$ limit-cycle is an undesirable instability, since, rather than estimating a static input phase, the modulator phase grows without bound. In Fig. 3.10b all modulator states $(\phi_m, \dot{\phi}_m)$ within the dashed boundaries are stable; but there is a range of unstable $\phi_m$ values which are driven to the $\dot{\phi}_m$ limit-cycle. As $G_1$ continues to diminish, the unstable range of $\phi_m$ widens until, as shown in Fig.3.10c, all trajectories go to the $\dot{\phi}_m$ limit-cycle (when $G_1 < G_c$). The interested
Fig. 3.10 Modulator state-space with a two-pole type-one loop filter. (λ_2=1)
reader can derive the minimum gain required to obtain stability from all initial modulator states (as in Fig. 3.10a), by employing techniques similar to those developed by Viterbi for phased-lock loop analysis.

The consequences of detector threshold and saturation, which constrain the $G_0 G_1 \sin \phi_e$ term in Eq. (3.59), were also investigated. In general, these limitations can shift the equilibria from those in Eq. (3.34) and can cause the $\dot{\phi}_m$ limit-cycle to flatten.

3.1.6.3 Summary

Table 3.1 compares a few of the major characteristics of IPLs incorporating the loop filters discussed in this section. Notice that the stable steady-state step-response depends on the filter type (number of poles at the origin of complex frequency plane) rather than the total number of poles. The stability of the steady-state, however, may be degraded by adding more poles to the filter. The phase tracking error with a type-one filter is independent of $\phi_i'$ and $\phi_m$, which can offer advantages. For example, the type-one system can be designed to operate with small error, regardless of the value of $\phi_i'$, as $|E_i|$ diminishes all the way to zero. The type-zero system offers similar performance only over a restricted range of $\phi_i'$. In some instances the $\phi_m$ dependence of the type-zero filter is more desirable, since it allows the error to be nulled by operation with $\phi_m = G_0$.

When $\phi_i'$ has a continuously increasing or decreasing ramp component, the type-zero filter results in a phase tracking error which grows with time, whereas the type-one system exhibits a finite error in response to a ramp. The implications of this on heterodyne operation, which will be seen to introduce a ramp component into $\phi_i'$, are discussed in section 3.2.1.
<table>
<thead>
<tr>
<th>M(s) (Open-loop filter)</th>
<th>Type</th>
<th>( \hat{P}(s) ) (Linearized closed-loop filter)</th>
<th>( \hat{P}_m ) (Steady-state response to a step in ( \hat{P}_m ))</th>
<th>( \hat{P}_e ) (Linearized steady-state error following a step and ramp in ( \hat{P}_e ))</th>
<th>BW (Phase tracking bandwidth) (approx is for ( G \gg 1 ))</th>
<th>( \tau ) (convergence time constant) (for ( G \gg 1 ))</th>
<th>STABILITY</th>
<th>( \Theta_F ) (Loop-noise bandwidth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{s^2} ) (low-pass)</td>
<td>0</td>
<td>( \frac{10G}{s^2 + (G_1 + 1)} )</td>
<td>( 1 - \sin^{-1} \left( \frac{G_1}{G_2} \right) ) + 2( \pi )</td>
<td>( \frac{G_1 + \frac{1}{2} \frac{1}{s} \omega + \frac{1}{2} \frac{1}{s} \omega \text{if } \omega &gt; 1 )</td>
<td>( \frac{1}{\omega} \left( 1 - \left( \frac{G_1}{G_2} \right)^2 \right) )</td>
<td>( \frac{1}{10G} )</td>
<td>Always stable</td>
<td>( \frac{G_1}{1 + G_1 + \frac{1}{4} } )</td>
</tr>
<tr>
<td>( \frac{1}{s} ) (integrator)</td>
<td>1</td>
<td>( \frac{10G_1}{s + \frac{G_1}{G_1}} )</td>
<td>( 1 - \sin^{-1} \left( \frac{G_1}{G_2} \right) ) + 2( \pi )</td>
<td>( \frac{\omega}{\omega - \omega} \left( \frac{G_1}{G_2} \right) + \frac{1}{\omega} )</td>
<td>( \frac{1}{10G} )</td>
<td>( \frac{1}{1 + G_1} )</td>
<td>Stable for ( G_1 &gt; G_0 )</td>
<td>( \frac{1}{10G} )</td>
</tr>
<tr>
<td>( \frac{1/2}{s + 1} ) (two-pole)</td>
<td>0</td>
<td>( \frac{10G_1}{s^2 + (1/2)s + 1} )</td>
<td>( 1 - \sin^{-1} \left( \frac{G_1}{G_2} \right) ) + 2( \pi )</td>
<td>( \frac{G_1 + 1/2 \frac{1}{s} \omega + 1/2 \frac{1}{s} \omega \text{if } \omega &gt; 1 )</td>
<td>( \sqrt{1 + \frac{1}{2} \frac{1}{s} \omega + \frac{1}{2} \frac{1}{s} \omega \text{if } \omega &gt; 1 )</td>
<td>( \frac{1}{1 + G_1} )</td>
<td>Always stable</td>
<td>( \frac{G_1}{1 + G_1} )</td>
</tr>
<tr>
<td>( \frac{1/2}{s + 1} ) (two-pole)</td>
<td>1</td>
<td>( \frac{10G_1}{s^2 + (1/2)s + 1} )</td>
<td>( 1 - \sin^{-1} \left( \frac{G_1}{G_2} \right) ) + 2( \pi )</td>
<td>( \frac{\omega}{\omega - \omega} \left( \frac{G_1}{G_2} \right) + \frac{1}{\omega} )</td>
<td>( \sqrt{1 + \frac{1}{2} \frac{1}{s} \omega + \frac{1}{2} \frac{1}{s} \omega \text{if } \omega &gt; 1 )</td>
<td>( \frac{1}{1 + G_1} )</td>
<td>Always stable</td>
<td>( \frac{G_1}{1 + G_1} )</td>
</tr>
</tbody>
</table>

(set \( \omega = 0 \) to see step response error)
The phase tracking bandwidth in Table 3.1 is taken as the 3 db rolloff of the single-pole closed-loop frequency response, or the frequency of peak response (after which a rolloff begins) of the two-pole systems. The two-pole bandwidth tends to increase with the square root of \( G_1 \) and hence is usually narrower than the single-pole bandwidth. With large gain, \( G_1 \), the two-pole systems exhibit behavior similar to an underdamped linear system, with decaying oscillations to equilibrium. The frequency of these oscillations generally increases with \( G_1 \). Although \( \phi_m \) may first cross its equilibrium value much sooner in the two-pole than the single-pole case, the decay time constant of the two-pole oscillations is independent of \( G_1 \) and hence tends to be slower than the single-pole convergence time. The modulator phase generally converges to within a few percent of its steady-state phase after a few time constants, even when a cycle is skipped on the way to convergence (e.g. when reacquiring phase after reaching the modulator limits). The type-zero filters always converge to a stable equilibrium. The type-one filters, however, become unstable and exhibit a limit-cycle in \( \phi_m \) if \( G_1 \) is not large enough. The single-pole integrator requires \( G_1 > G_0 \), and may have state-space "dead regions" which interfere with stability. The two-pole type-one system requires even larger \( G_1 \) for stability.

The loop-noise bandwidth anticipates the noise presentation of the next section and will be discussed more thoroughly there. It is generally desirable to make \( B_F \) as small as possible in order to minimize the effects of noise on IPE performance.

The intent of this section was not to specify a particular loop filter as being superior, but rather to provide a perspective of the
types of behavior which the IPL exhibits. An IPL possessing any of the
loop filters discussed here can be made to perform adequately, exhibiting
$\phi_m = \phi_i$ and $\phi_e = 0$. In practice the loop filter often depends on constraints
imposed by the available hardware, particularly in high-resolution
applications implemented with a monolithic SLM. Many SLMs $^{80,189,192}$, such
as the MSLM described in the next chapter, can be adequately character-
ized by an integrating or low-pass, single-pole frequency response. A
system exhibiting a more complicated frequency response can often be
approximately characterized by a single dominant pole or its two most
dominant poles.

In low-resolution, discrete-channel implementations, it is
sometimes possible to specify the loop filter, or at least to modify
the intrinsic detector and modulator responses. In those instances the
simpler frequency responses, and in particular the single-pole responses,
are preferable. A simple response is not only usually easier to
implement, but also facilitates the analysis and adjustment of the IPL.
This is an important consideration, because the usual closed-loop
design aids such as root-locus or bode plots are not applicable to
this nonlinear system.
3.1.7 IPL performance in the presence of noise

An analysis of the IPL in the presence of noise conveniently subdivides into two components: identifying and characterizing specific physical noise sources, and gaining insight into the general effects of noise on IPL performance. The former topic, sources of noise in optical systems, is well developed in the optical communications literature\(^1\),\(^3\),\(^8\)-\(^8\),\(^7\),\(^9\)-\(^9\),\(^6\),\(^1\),\(^0\),\(^1\),\(^4\),\(^9\); a few pertinent examples will be briefly presented. In this section emphasis will be placed upon the latter topic, providing insight into the effects of noise on IPL behavior, including quantifying the resulting performance degradation and establishing maximum acceptable noise level bounds. In a specific application, the noise bounds can be interpreted in terms of the relevant noise sources to evaluate the overall performance.

It will be seen in section 3.2.1.1 (see Fig. 3.13) that the IPL is closely related to an electronic phase-locked loop (PLL) except that part of the feedback path is optical. As such, many of the results in this section will be taken from the PLL literature\(^1\),\(^6\),\(^4\),\(^1\),\(^6\) rather than rederived.

**General formulation**

Most sources of noise in the IPL can be treated as white (wideband relative to the system frequency response), zero-mean, Gaussian random processes, \(W(t)\), added to the signal at the input to the loop filter\(^1\),\(^6\); i.e.,

\[
\phi_m(t) = h^*(G_0 + G_1\sin \phi_e + W(t))
\]  

(3.60)

This result follows from Eq. (2.10) and (2.12) with \( h \) as the open-loop impulse response (i.e., inverse Laplace transform of \( \hat{H}(s) \)) and the \( * \) representing convolution. It is further assumed that \( W(t) \) is wide-sense stationary and uncorrelated with the signal. The behavior of \( \phi_e \) can be found from \( \phi_e = \phi_i - \phi_m \). The discussion in this section will pertain to a
single spatial resolution element.

The correlation function of \( W(t) \) is

\[
R_W(\tau) = \overline{W(t+\tau)W(t)} = \frac{N_0}{2} \delta(t-\tau) \tag{3.61a}
\]

The overbar denotes ensemble averaging and \( N_0 = 2S_n(w) \) is the one-sided power spectrum of the fluctuations in \( W \). Since the noise sources are assumed independent, they sum in \( N_0 \), e.g. \( N_0 = N_{\text{thermal}} + N_{\text{shot}} + N_{\text{dark}} + \ldots \). \( N_0 \) is also related to the variance of \( W(t) \), since

\[
\sigma_W^2 = \overline{W^2 - \overline{W}^2} = \overline{W^2} = R_W(0)
\]

Employing the Wiener-Kinchine theorem \( (R(\tau) = \mathcal{F}^{-1}(S(w))) \),

\[
\sigma_W^2 = R_W(0) = \int_{-\infty}^{\infty} S_n(w) \frac{dw}{2\pi} \quad (= N_0 \int_{0}^{\infty} \frac{dw}{2\pi}) \tag{3.61b}
\]

(For white noise \( \sigma_W^2 \to \infty \)). The probability of a given noise amplitude at a given time, assuming that \( W(t) \) is ergodic, is

\[
P(W) = \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left(-\frac{W^2}{2\sigma_W^2}\right) \tag{3.61c}
\]

Note that zero-mean Gaussian white noise is essentially fully characterized by \( N_0 \); (the joint probability distribution \( P(W(t), W(t+\tau)) \) is often required to provide a more complete description of the basic amplitude and temporal properties of other types of noise).

Sources of noise

This model is more or less consistent with the noise associated with such effects as background irradiance, detector dark current, thermal noise generated in the detector and electronic portions of the loop, and shot noise. A few specific examples of the form of the component added to \( N_0 \) by various noise sources follows.

The one-sided power spectrum of thermal noise (or Johnson noise)
\[ N_t = q_1^2 q_2^2 4kT_e/R_e \] (3.62a)

Here \( q_1 \) and \( q_2 \) are constants defined in Eqs. (2.6) and (2.7) respectively and \( k \) is the Boltzmann constant. \( T_e \) and \( R_e \) are the effective noise temperature and resistance of the detector and electron processing parts of the loop referenced back to the detector current, \( i_0 \), of \( E_q \). (2.6).

The number of photons seen by the IPL in a given time is a conditional inhomogeneous poisson process \[1,86,95,110,170,171\]. Usually, the duration of the system impulse response (e.g. \( h(t) \)) greatly exceeds the time between photon arrivals (high density shot noise), in which case the fluctuations in the detector current, \((I-I)g_o\), can be treated as zero-mean Gaussian white noise. (This noise is approximately stationary if the average intensity, \( I \), varies only very slowly). Taking \( I \) as \( I_0 + I_1 \sin e \), from Eq. (2.1), these fluctuations constitute signal and local oscillator induced shot noise, which contributes
\[ N_q = 2aeq_o q_1^2 q_2^2 I \] (3.62b)

to the one-sided power spectrum, \( N_0 \). Here \( a \) is a constant dependent on the detector, \( a = h\nu q_o / (e\eta A_r) \) (=1 for a photocathode), \( e \) is the charge on an electron, \( I \) is the average irradiance, \( q_o q_1 q_2 \) is defined by Eqs. (2.6) to (2.8), \( h\nu \) is the energy of a photon, \( n \) is the quantum efficiency, and \( A_r \) is the area of a resolution element.

Background radiation from natural and artificial sources introduces a field \( E_b(r,t) \) at the detector, which can usually \[1,86,95\] be modelled as a zero-mean complex Gaussian random process with mutual coherence function \( \Gamma(r_1,r_2,t_1,t_2) = E_b(r_1,t_1)E_b^*(r_2,t_2) = I_r^2 \delta(t_1-t_2) \delta(r_1-r_2) \), where \( I_r/2 \) is the radiance \((W/m^2HzSr)\) of a polarization component of the background at optical frequency \( c/\lambda \). The average background power
collected by an optical system with aperture area $A_a$, field of view (FOV) $\Omega$, and optical bandwidth $c\Delta\lambda/\lambda^2$ is

$$
\bar{P}_b = I_b A_d = I_\lambda \Omega A_a \Delta \lambda
$$

(3.63a)

Here $I_\lambda = c I_f/\lambda^2$ has units of $W/m^2$Srum and $\bar{I}_b$ is the effective background irradiance over the detector area, $A_d$. Typical values of $I_\lambda$ range from $10$ (day) to $10^{-3}$ $W/m^2$Srum (night).[8]

As a specific example, a lens of area $A_a$ followed by a diffraction limited detector in its Fourier plane has an FOV of $\Omega = (\frac{1.22 \pi \lambda}{2})^2/A_a = \lambda^2/A_a$ and hence collects

$$
\bar{P}_b = I_\lambda \lambda^2 \Delta \lambda
$$

(3.63b)

This is also the average background power seen in the IPL. For example, if the input aperture in a self-interference implementation (e.g. Zernike phase-contrast) is imaged through a spatial filter which restricts the FOV to $\Omega = \lambda^2/A_r$, where $A_r$ is the area of a detector resolution element, the total aperture of area $A_a$ sees $\Omega A_a = \lambda^2 A_a/A_r$, but each resolution element sees only $A_r/A_a$ of this power.

The average background power does not contribute to noise fluctuations and merely causes a small change in $G_o$, which can be treated as a shift in $I_t$ of $\bar{P}_b/A_d$. The fourth order statistics of $E_b$ can contribute to the noise, $(N_A = 3(G_0 q_1 q_2)^2 P_b^2/(2 A_r \Delta \lambda))$, but this effect is usually negligible. The most important effect of background radianc is the generation of additional shot noise in the detector. In a resolution element of a self-interference IPL (e.g. Zernike phase-contrast), this contribution is

$$
N_b = 2 a e q_1 q_2 P_b^2 / A_r
$$

(3.63c)

In the homodyne IPL this contribution becomes[1,8,8,95]
\[ N_{bh} = 2(g_0 g_1 g_2)^2 I_f \lambda^2 |E_2|^2 (2Z_0 A_r) \] (3.63d)

This is usually negligible relative to the local oscillator shot noise. (The ratio of background to L.O. shot noise is \( I_f \lambda^2 / h \nu \) which is less than \( 10^{-8} \) during the day).

Dark current can be treated as an equivalent input dark irradiance \( I_d \) which generates the shot noise

\[ N_d = 2aeg_0 q_1^2 q_2^2 \frac{I_d}{164} \] (3.64)

Viterbi shows that stochastic variations in \( G_0 \) and \( G_1 \), (for example, caused by scattering induced fluctuations in \( |E_f| \)), can also be approximated by adding another component to \( \mathcal{W}(t) \) and hence \( N_0 \).

Another noise source, which has been observed to be significant in some applications, is manmade interference coupled into the electronic part of the IPL through stray radio frequency and powerline coupling. This can be approximately treated by adding systematic components to \( G_0 \) and a zero-mean random component to \( \mathcal{W}(t) \) and \( N_0 \).

**Noise performance of the linearized IPL**

It has been seen that with large gain \( G_1 > G_0 - \phi_m \), the IPL tends to operate with \( \phi_e = 2n\pi \). Useful insights can thus be obtained by studying the noise performance of the IPL with the linearization \( \sin \phi_e = \phi_e \); a presentation of the more general nonlinear case will follow this discussion.

The following linearized noise analysis is very similar to that of Viterbi\textsuperscript{164}. The noise performance of linearized closed-loop optical systems similar to the IPL has also been studied by Robinson\textsuperscript{3,4,149} and by Dyson\textsuperscript{110}. They studied optimal linearized systems which estimate phase with minimum mean-squared error. They also considered the effect of photon-photon correlations which arise in closed-loop adaptive systems because each
collected photon changes the phase modulator and thereby the probability of detecting succeeding photons.

The basic linearized operational equation follows from Eq. (3.60) as,

\[ \phi_m = f^* \left[ (\phi'_i + G_0) + \frac{W(t)}{G_1} \right] \]  \hspace{1cm} (3.65a)

Here \( f \) is the closed-loop impulse response, which is the inverse Laplace transform of

\[ \hat{f} = \frac{G_1 \hat{r}}{1 + G_1 \hat{r}} \]  \hspace{1cm} (3.65b)

The effect of noise is to introduce randomness into \( \phi_m \) through the \( W(t) \) term in Eq. (3.65a). The average phase estimate, however, remains unmodified, i.e.

\[ \bar{\phi}_m = h^* (\phi'_i - G_0/G_1) \]

because \( W \) has zero mean. With Gaussian statistics the randomness in \( \phi_m \) can be fully characterized by \( \bar{\phi}_m \) and the variance \( \sigma^2_{\phi} = \bar{\phi}_m^2 - \bar{\phi}_m^2 \). Physically, \( \sigma_{\phi} \) is the rms magnitude of the fluctuations in \( \phi_m \). Since Eq. (3.65a) is linear and \( W(t) \) is independent of the first term, the variance of the two terms superimpose. For example, (if \( F \) is a stable transfer function), all the variance of \( \phi_m \) in the steady-state following a step in \( \phi'_i \) is due to \( W(t) \); since \( \phi'_i \) is constant, the variance of \( \phi_e \) is also \( \sigma^2_{\phi} \). The variance can be computed from the power spectrum of \( \phi_m \) through Eq. (3.61b); in the steady-state the power spectrum of \( \phi_m \) due to noise is

\[ S_\phi(w) = \left| \frac{\hat{F}(w)}{G_1} \right|^2 \cdot S_n(w) = \left| \frac{\hat{F}(w)}{G_1} \right|^2 \cdot \frac{N_o}{2} \]  \hspace{1cm} (3.66a)

Thus,

\[ \sigma^2_{\phi} = \frac{N_o}{2G_1^2} \int_{-\infty}^{\infty} |\hat{F}|^2 \frac{dw}{2\pi} = \frac{N_o B F}{G_1^2} \]  \hspace{1cm} (3.66b)
Here
\[ B_F = \frac{\int_0^\infty |\tilde{F}|^2 \frac{dw}{2\pi}} {G_1} \quad (3.66c) \]
is the "loop-noise bandwidth" of the closed-loop transfer function.

For example in the low-pass homodyne IPL, \( \tilde{H} = \frac{w_0}{(jw_0 + G_1)} \) and \( \tilde{F} = \frac{1}{G_1} \frac{w_0}{(jw_0 + G_1)} \) follows from Eq. (3.65b); the integral in Eq. (3.66c) is readily evaluated to reveal that
\[ B_F = \frac{w_0 G_1^2}{4(1+G_1)} \quad (3.67a) \]
and from Eq. (3.66b)
\[ \sigma_\phi^2 = \frac{w_0 N_0}{4(1+G_1)} = \frac{G_1}{1+G_1} \frac{N_0}{4G_1} \quad (3.67b) \]

With the integrating filter \( \tilde{H} = \frac{w_0}{jw_0} \), \( \tilde{F} = \frac{G_1}{w_0} \frac{w_0}{(jw_0 + G_1)} \) and
\[ B_F = \frac{G_1 w_0}{4} \quad (3.67c) \]
and
\[ \sigma_\phi^2 = \frac{N_0 w_0}{4G_1} \quad (3.67d) \]
The \( B_F \) expressions above along with those for a couple of two-pole filters appeared in Table 3-1. With a type-one two-pole filter \( \sigma_\phi^2 \) has the form in Eq. (3.67d); with a type-zero two-pole filter
\[ \sigma_\phi^2 = \frac{G_1}{1+G_1} \frac{N_0}{4G_1} \frac{w_0 w_1}{w_0 + w_1} \quad (3.67e) \]
The low-pass case has a slightly narrower noise bandwidth and smaller variance than the integrating case. In practice \( G_1 \gg 1 \) is usually true, so the rms noise fluctuation in both systems is essentially \( \sigma_\phi \) of Eq. (3.67d). Notice that as the filter gain, \( w_0 \), increases the noise-bandwidth and hence \( \sigma_\phi^2 \) increases.

Equations (3.67) do not provide a clear indication of the effect of changing \( G_1 \), because as seen in Eqs. (3.62)-(3.64), \( N_0 \) includes a \( g_0 g_1 g_2 \) gain factor. It is thus convenient to reference \( N_0 \) back to the input irradiance, \( I \). Comparison of Eqs. (2.11) and (3.60) reveals that
\[ S_{\psi}(w) = |\hat{H}(w)|^2 N_0 = \left(\frac{q_0 q_1 q_2}{2}\right)^2 |\hat{H}(w)|^2 N_I \]
where
\[ N_I = \frac{N_0}{(q_0 q_1 q_2)^2} \tag{3.68a} \]
is the power spectrum of an input noise irradiance which includes the
effects of all the noise sources. Using Eqs. (2.2b) and (2.14b), Eq.(3.67d)
now becomes:
\[ \sigma_{\phi}^2 = \frac{q_0 q_1 q_2 w_0 N_I}{4I_1} \tag{3.68b} \]

Thus increasing \( G_1 = q_0 q_1 q_2 I_1 \) by increasing the signal irradiance
\( I_1 \) improves the noise performance; but increasing the electronic gain
parameter \( q_0 q_1 q_2 w_0 \) degrades the noise performance!

Eq. (3.66b) can also be rewritten
\[ \sigma_{\phi}^2 = \frac{N_IB_F}{I_1^2} = \frac{q_0^2 N_IB_F}{2q_0^2 I_1^2} \]

note that \( q_0^2 N_IB_F = \sigma_i^2 \equiv (\Gamma - 1)^2 \) is proportional to the detector
noise power which passes within the system bandwidth, and \( q_0^2 I_1^2 \) is the
detector signal power due to the phase term. Thus
\[ \sigma_{\phi}^2 = \frac{1}{(S/N)} = \frac{1}{(s/n)^2} \tag{3.68c} \]
where \((s/n)\) is the ratio of detector signal to noise current.

The more general nonlinear analysis reveals that, for many loop
filters when the linear approximation is valid (i.e. \( \sigma_{\phi} < \pi/6 \)), \( \phi_m \) in Eq.(3.60)
and \( \phi_e \) are approximately Gaussian distributed. The probability of the
error not exceeding \( \phi_1 \) is thus
\[ P(|\phi_e - \bar{\phi}_e| < \phi_1) = \int_{-\phi_1}^{\phi_1} \exp\left(-\frac{(\phi_e - \bar{\phi}_e)^2}{2\sigma_{\phi}^2}\right) d\phi \tag{3.69} \]

(Following a step in \( \phi; \) \( \bar{\phi}_e = G_0/G_1 + 2n\pi) \)
This equation is also valid for \( \phi_m \), with \( \bar{\phi}_m = \phi_i \).
Eq. (3.69) in conjunction with $\sigma_\phi^2 = 1/(S/N)$ from Eq. (3.68) provides considerable insight into the performance of the IPL, given only $S/N$, the signal-to-noise ratio. For example, with $S/N=3.6$, the error fluctuations during 68% of the time will not exceed $\sigma_\phi = \pi/6$, and during 95% of the time the $\phi_e$ fluctuations do not exceed $\pi/3$. If $S/N$ is increased to 10, the 68% bound, i.e. $\sigma_\phi$, decreases to $\pi/10$.

The analysis in a practical application pivots about the signal-to-noise ratio. Without assuming any specific noise sources Eqs. (3.68) and (3.69) can be utilized to determine the maximum tolerable variance, $\sigma_{\phi_m}$, and hence the minimum usable $S/N$. The $S/N$ for the specific system configuration and noise sources ($N_0$) can then be evaluated to find the overall constraints imposed by noise, such as the minimum number of signal photons required for phase compensation.

**General nonlinear noise analysis**

Since the IPL is closely related to a phase-locked loop, the following analysis is very similar to the nonlinear noise analysis of a PLL. The reader is thus referred to the PLL literature, for example Viterbi's discussion in Ref.164, for a more detailed treatment of the issues outlined here.

The integrating homodyne IPL is one of the easiest cases to discuss. Using $H(s) = w_0/s$ and Eq. (3.60) with $\phi_e = \phi_1 - \phi_m$, the step response of $\phi_e$ can be described by

$$\dot{\phi}_e = -w_0(G_o + G_1 \sin \phi_e + W)$$  \hfill (3.70)

For a given $\phi_e$, $\dot{\phi}_e$ is a Gaussian random process with mean $-w_0(G_o + G_1 \sin \phi_e)$ and variance $w_0^2 \sigma_\phi^2$. Eq. (3.70) also implies that $\phi_e$ is a Markov random process, where the transition probability is a function of only the present value of $\phi_e(t)$ and is independent of the previous history of $\phi_e$. 

The dynamic and steady-state behavior of this system is completely described by \( P(\phi_e, t) \), the probability of a given \( \phi_e \) at a given time. For example, \( R(\tau), S(w), \phi_e, \sigma^2_\phi \), and the steady-state distribution \( \lim_{t \to \infty} P(\phi_e, t) \) can be determined from \( P(\phi_e, t) \).

Utilizing the theory of Markov processes\(^{164,171,172}\) one can derive a differential equation for \( P(\phi_e, t) = P(\phi_e, t) \)

\[
\frac{\partial P(\phi_e, t)}{\partial t} = w_o \frac{\partial}{\partial \phi_e} [(G_0 + G_1 \sin \phi_e) P(\phi_e, t)] + \frac{w_o^2 N_o}{4} \frac{\partial^2 P(\phi_e, t)}{\partial \phi_e^2}
\]  

(3.71)

(Viterbi also provides the general technique for deriving a differential equation for \( P(\phi_e, t) \) when \( W \) in Eq. (3.70) is not zero-mean Gaussian noise).

Eq. (3.71) is a Fokker-Planck equation. If the \( \sin \phi_e \) factor were a constant this would be an ordinary diffusion equation with drift\(^{168,172} \), whose solution is a Gaussian with mean \( \phi_e = -(G_0 + G_1 \sin \phi_e) t \) and variance \( \sigma^2 = w_o^2 N_o t / 4 \).

Thus one may guess that the mean solution will stop changing when \( \phi_e = -\sin^{-1}(G_0 / G_1) + 2n\pi \) (i.e., the noiseless steady-state), and that noise causes the variance to grow continuously with time.

With the \( \sin \phi_e \) variable coefficient, Eq. (3.71) is periodic and its general solution\(^{164} \) reveals that \( P(\phi_e, t) \) tends to spread out from an initial value of \( P(\phi_e, 0) = \delta(\phi_e - \phi_{eo}) \) to local maxima at higher and higher \( n \) values of \( \phi_e = -\sin^{-1}(G_0 / G_1) + 2n\pi \); but with a constant variance \( \sigma^2_\phi \) about each \( \phi_e(n) \).

The phase modulator in a noisy IPL thus initially converges to \( \phi_m = \phi_1 + \sin^{-1}(G_0 / G_1) + 2\eta_0 \pi \) with a variance of \( \sigma^2_\phi \) (where \( \eta_0 = \text{INT}((\phi_m - \phi_1 + \pi) / 2\pi) \)) as in the linear approximation. However, with time, there is a finite probability that the noise will knock the IPL out of equilibrium and cause \( \phi_m \) to relock at another \( n \) value. Since \( \phi_m \) has a limited dynamic
range, this can be a serious degradation. If \( \phi_m \) had an unlimited dynamic range, the probability of any given value goes to zero for \( t \rightarrow \infty \), due to the continuous spreading of \( P(\phi, t) \).

The steady-state distribution about any value can be found by solving for \( P(\phi) = \lim_{t \rightarrow \infty} P(\phi, t) \), with the periodic boundary condition

\[
P(\pi) = P(-\pi)
\]

and normalization

\[
\int_{-\pi}^{\pi} P(\phi_e) \, d\phi_e = 1
\]

(This normalization corresponds to the infinite initial condition

\[
P(\phi_e, 0) = \delta(\phi_e - \phi_0 - 2n\pi).
\]

The steady-state of Eq. (3.71) can be found by setting \( \dot{P}(\phi_e, t) = 0 \), which results in a linear first order differential equation with variable coefficients:

\[
\frac{\partial P(\phi_e)}{\partial \phi_e} + \frac{4}{w_o N_o} (G_0 + G_1 \sin \phi_e) P(\phi_e) = \text{constant}
\]

This equation is readily solved. The constant is determined by Eq. (3.72a) and the solution must be normalized by Eq. (3.72b). In the special case of \( G_0 = 0 \), it can be shown that:

\[
P(\phi_e) = \frac{\exp(h \cos \phi_e)}{2\pi I_0(b)}
\]

where

\[
b = \frac{4G_1}{w_o N_o}, \quad \frac{G_1^2}{N_o B_F} = \frac{I_F^2}{N_F} = \frac{S}{N}
\]

Here \( I_0 \) is a modified bessel function of order zero, \( B_F \) is the loop noise bandwidth from Eq. (3.67c), and \( S/N \) is the detector signal to noise power ratio.

In the linear approximation the variance (in Eqs. (3.41b), (3.42d), and (3.68)) equaled \( 1/b \); and, in fact, for \( b = S/N > 4 \), \( I_0(b) \) and \( \cos \phi_e \) can be expanded to arrive at the approximate expression:

\[
S/N = \frac{1}{2} \left( \frac{S}{N} \right) \frac{1}{b} + \frac{1}{2} \left( \frac{S}{N} \right) \frac{1}{b^2}
\]
\[
P(\phi_e) = \frac{1}{\sqrt{2\pi} \sigma_\phi} \exp\left(-\frac{\phi_e^2}{2\sigma_\phi^2}\right)
\]  
(3.74a)

where

\[
\sigma_\phi^2 = \frac{1}{2\pi} \int_0^\pi \phi_e^2 \exp(b \cos\phi_e) \, d\phi_e
\]  
(3.74b)

\[
= \frac{1}{b} = \frac{1}{S/N} = \frac{I_1^2}{N_1 B_F}
\]  
(3.74c)

This is the same as the linear approximation; hence for S/N>4, Eq.(3.69) gives the percentage of the time that the absolute value of the error does not exceed \(\phi_i\), (in applications where 2\(\pi\) multistability does not matter).

With small S/N, the exact behaviors of Eqs. similar to (3.73a),(3.74b), and \(\int P(\phi)\,d\phi\) are graphed in Viterbi. Even with S/N=1, the behavior is similar to Eq. (3.74a) with \(\sigma_\phi^2=1.6/b\). For S/N=0, \(P(\phi)=1/2\pi\) and \(\sigma_\phi^2=\pi^2/3\), i.e. when the noise is infinite \(\phi_m\) and \(\phi_e\) become random with no tendency toward any particular value (uniformly distributed).

When 2\(\pi\) multistability is of concern, the effects of cycle skipping can be characterized by the average time \(T_1(2\pi)\) required for \(\phi_e\) to first reach \(\phi_e=\pm2\pi\). (This is analogous to the "mean first passage time" to first reach an absorbing boundary in a random walk problem, or to the average time to go broke in the gambler's ruin problem.)\textsuperscript{72} The frequency of skipping cycles, \(f_1=1/T_1(2\pi)\), i.e. the rate at which cycles are skipped, is also of interest. In some applications the first passage times to other values of \(\phi_e\), such as those corresponding to the modulator limits, are also useful for evaluating the performance of a noisy IPL.

The probability density function \(P(\phi_e, t)\) for an IPL constrained to operate with \(|\phi_e|<\phi_L\) can be found by solving Eq. (3.71) with initial condition \(P(\phi_e, 0)=\delta(\phi_e-\phi_0)\) and boundary condition \(P(\phi_L, t)=P(-\phi_L, t)=0\).
(P(ϕe,t)=0 for all |ϕe|>ϕL). Then

\[ q(t) = \int_{-ϕL}^{ϕL} P(ϕe,t) \, dϕe \]

is the probability that ϕe has not reached ±ϕL during time t. It is also useful to define r(t) through

\[ q(t) = 1 - \int_0^t r(t') \, dt'. \]

Here r(t') is the probability that t' is a sufficiently long interval for |ϕe| to reach ±ϕL for the first time, i.e. q(t) is the probability density function of the time required to reach ±ϕL for the first time. The expected time to reach ±ϕL for the first time is thus

\[ T_1(ϕL) = \int_0^∞ t \, r(t) \, dt = -\int_0^∞ \frac{3q}{3t} \, dt = tq|_0^∞ + \int_0^∞ q \, dt = \int_0^∞ q \, dt \]

\[ = \int_{-ϕL}^{ϕL} \int_0^∞ P(ϕe,t) \, dϕ \, dt \]

Following the procedure outlined above one finds

\[ T_1(ϕL) = \frac{b}{4B_F} \int_{-ϕL}^{ϕL} (\int_0^∞ \exp[b(\cosϕ-\cosx)]dx) \, dϕ \]

where b is given by Eq. (3.73b)

In the specific case of the time to skip cycles:

\[ T_1(2\pi) = \frac{π^2 b F_0(b)}{2B_F} \] (3.75a)

For b=S/N > 4, good approximations are \( σ^2 = 1/b \) and

\[ T_1(2\pi) = \frac{π^2}{4B_F} e^{2/σ^2} = πτe^{2/σ^2} \] (3.75b)

where \( τ = G_1 w_0 \) is the convergence time of the noiseless integrating IPL.

Also of interest is \( T_1(π) \) which is the time for the IPL to first reach its worst-case error of π; Viterbi shows that \( T_1(π) = T_1(2π)/2 \).
$T_1(\pi)$ can be taken as a measure of the time to first lose frequency lock. Table 3.2 shows the increase of $T_1(\pi)$ with $S/N$, based on Eq. (3.75a).

<table>
<thead>
<tr>
<th>b = $S/N$</th>
<th>$T_1(\pi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>$2\tau$</td>
</tr>
<tr>
<td>0.5</td>
<td>$5\tau$</td>
</tr>
<tr>
<td>1</td>
<td>$15\tau$</td>
</tr>
<tr>
<td>2</td>
<td>$100\tau$</td>
</tr>
<tr>
<td>3</td>
<td>$720\tau$</td>
</tr>
<tr>
<td>4</td>
<td>$6500\tau$</td>
</tr>
<tr>
<td>5</td>
<td>$40000\tau$</td>
</tr>
</tbody>
</table>

$\tau = \omega_0 G_1$ is the IPL convergence time constant $T_1(2\pi) = 2T_1(\pi)$ is the average time to first move one full cycle (i.e. lose and reacquire lock).

The worst acceptable value of $S/N$ and hence $T_1(\pi)$ depends on the particular application, such as whether the IPL is framed or how fast $\phi_1$ varies. However, since complete convergence of $\bar{\phi}_m$ to its steady-state mean value takes about $5\tau$ and $T_1(\pi) = 5\tau$ for $b=S/N=0.5$ ($\sigma_\phi=\pi/2$), $S/N=0.5$ can be taken as a signal-to-noise threshold for IPL operation. Table 3.2 shows that this threshold interpretation is particularly meaningful since $T_1(\pi)$ increases very rapidly for $S/N>0.5$.

With other loop filters, similar behavior results. In fact, to a high degree of accuracy, Eq.(3.73a) gives $P(\phi_1)$ and Eq.(3.75a) gives $T_1(2\pi)$ with $b$ computed via Eq.(3.73b) by substituting in the value of $B_F$ (from Eq.(3.66c)) appropriate to the particular loop filter. When $G_0 \neq 0$, there is a nonzero steady-state error and the distributions become centered about the mean value of $\bar{\phi}_e$. With the proper $B_F$, the approximations in Eqs. (3.69),(3.74c), and (3.75b) also remain valid.

**Discussion of noise performance**

IPL operation in the presence of white Gaussian noise thus introduces
approximately Gaussian fluctuations into $\phi_e$ and $\phi_m$. The mean values of $\phi_e$ and $\phi_m$ in the noisy system tend to converge to the noiseless stable equilibria, i.e. $\bar{\phi}_m = \phi^*_1 - 2n\pi$ and $\bar{\phi}_e = 2n\pi$, and hence estimate and compensate $\phi^*_1$. The phase estimate and error, however, are now characterized by a variance, $\sigma^2_\phi$.

The general nonlinear analysis has revealed that there is effectively an absolute minimum S/N threshold of 0.5, in order that the IPL converge faster than $\phi_m$ and $\phi_e$ fluctuate. A more conservative threshold condition is to demand that the rms magnitude of the phase fluctuations, $\sigma_\phi$, does not exceed a specified maximum. In his linearized analysis, Robinson\textsuperscript{3,149} has suggested 0.5 (≈0.16π radians (∼λ/13)) as a practical performance criterion. For $S/N \geq 2$, $\sigma^2_\phi = (S/N)^{-1}$ is valid and this condition amounts to specifying $S/N \geq 4$. With this condition satisfied, Eq. (3.69) shows that $|\phi_e - \bar{\phi}_e| < 0.16\pi$ 68% of the time and $|\phi_e - \bar{\phi}_e| < 0.26\pi$ 90% of the time. Another interpretation of this condition is that the IPL can distinguish $2\pi/2\sigma_\phi$ (∼6) distinct phase values, with $S/N \geq 4$. The stability with $S/N \geq 4$ is good, since according to table 3.2 an average interval of more than 6000π occurs between times when the fluctuations cause loss of phase lock to occur.

According to the noise model of this section, the noise performance in a specific application is completely characterized by only knowledge of $S/N = I^2_1/N_1B_F$, where $I_1$ is established by the optical system, $B_F$ is a property of the electronic link, and $N_1$ is the power spectrum of the important optical and electronic noise sources. The actual $S/N$ in conjunction with the minimum tolerable $S/N$ discussed in the preceding paragraph defines the overall noise-imposed operating constraints of the application.
An example of special interest is IPL operation under ideal circumstances, where the dominant noise source is only the shot noise due to the intensity $I = I_0 + I_1 \sin \phi_e$. Using Eqs. (3.62b) (for a photocathode) and (3.73b),

$$S/N = \frac{L^2}{2h\nu \eta_{\tau} B_F (I_0 + I_1 \sin \phi_e)} \quad (3.76a)$$

Eq. (3.67d) reveals that $1/B_F = 4\tau$, where $\tau = 1/\omega_0 g_1$ is the convergence time constant. Using Eqs. (2.2) with $\alpha = 1$ and $r = |E_i|/|E_\lambda|$, the SNR becomes

$$S/N = 4\eta_1 \frac{r}{1 + r^2 + 2r \sin \phi_e} > \frac{4\eta_1 r}{(1 + r)^2} \quad (3.76b)$$

where

$$\eta_1 = \frac{n_1 I_1}{hv} = \frac{n_1 I_1}{h\nu \omega_0 g_1} = \frac{n_1}{h\nu \omega_0 g_1 g_2} \quad (3.76c)$$

is the average number of photons collected from the $I_1$ signal per system time constant per resolution element. Notice that $\eta_1$ is a system constant independent of $|E_i|$ or $|E_\lambda|$ and that $S/N$ depends only on the ratio of $|E_i|$ to $|E_\lambda|$ and not on their actual magnitudes! This occurs because the photon collection time in Eq. (3.76c), $\tau = 1/(\omega_0 g_1 g_2 I_1)$, increases as $I_1$ decreases; the SNR thus depends on the total energy or number of photons collected, but not on the irradiance level. Eq. (3.76c) can also be expressed in terms on the number of photons collected from $E_i$ by noting that

$$\eta_1 = \frac{2 n_1 r |E_i|^2}{r h\nu 2Z_o} \quad (3.76d)$$

As the local oscillator strength increases, $S/N$ passes through a maximum and then decreases; for $|E_\lambda| >> |E_i|$, Eq. (3.76b) becomes $S/N = 4\eta_1 |E_i|/|E_\lambda|$. This is in contrast to the optimum (minimum mean-square error) linearized estimators discussed by Robinson\textsuperscript{149} and Dyson\textsuperscript{150}; where the SNR improved to an asymptotic limit as $|E_\lambda|$ was increased. In the
estimator \(66,110,149,173\), the open-loop gain factor \((q_o q_1 q_2 w_0)\) varies in proportion to \(1/(|E_z||E_x|)\) which causes \(B_F\) in Eq. (3.76a) to be independent of \(|E_x||E_z|\). In that case, as \(|E_x|\to\infty\), Eq. (3.76a) approaches

\[
S/N = \frac{2nA\tau}{h\nu} \frac{|E_z|^2}{B_F 2Z_0} \overset{1}{=} 2n_i
\]

(3.77a)

which is independent of \(|E_x|\). The optimum gain and open-loop bandwidth are also functions of the statistics of \(\phi_i\). It is unlikely that an all-optical SLM-based adaptive system could be configured as an optimum estimator; however, the homodyne IPL with \(r\) chosen to maximize Eq. (3.76b) (or \(q_1\) scaled to preserve the maximum SNR as \(r\) varies) approximates the optimum homodyne estimator.

With small error \((\bar{\phi} = 2\pi\)), Eq. (3.76b) is maximum for \(r = 1\), becoming

\[
S/N = 2n_1 = n_i \quad (= 1/\sigma_i^2)
\]

(3.77b)

Here

\[
n_i = \frac{nA\tau}{h\nu} \frac{|E_z|^2}{2Z_0}
\]

(3.77c)

is the average number of photons collected from the input irradiance incident on a resolution element during the time \((4\tau = 1/R_F)\) which the IPL takes to converge to within 2% of its steady-state.

Eqs. (3.77) can be utilized to specify the minimum number of photons, \(n_i\), required per resolution cell during the phase estimate convergence time. For example, Robinson's \(\sigma_i < 0.5\) radian criterion \((S/N > 4)\) suggests \(n_i > 4\) is required, which agrees with Robinson's result for his linearized system. Alternatively, it was suggested that the IPL will operate after a fashion down to a threshold of \(S/N = 0.5\), which corresponds to \(n_i > 0.5\) or requiring at least one photon every other convergence interval.
Actually, the above results, e.g. Eq. (3.77), are not rigorously correct in this low photon flux regime, since high density shot noise was assumed. With low photon flux the shot is a conditional inhomogeneous Poisson random process, rather than Gaussian white noise. The discussion on pages 85-87 of Viterbi provides insight on generalizing the analysis to low photon flux and suggests that the above results do not require significant modifications. This conclusion is further supported by the similarities between the above results and the linearized, more rigorous low-photon flux results of Robinson and Dyson.

It should be noted that since $n_s=2K_1$ is independent of $|E_1|$ for $r=1$, specifying a minimum value of $n_s$ does not constrain $|E_1|$, but rather restricts the maximum usable IPL gain factor, $w_0g_0g_1g_2$.

Additional comments on noise performance will be made in section 3.1.9 in the context of determining the minimum signal irradiance, $I_1=|E_1|^2/2Z_0$, required in a particular application.

3.1.8 IPL performance with extraneous coherent illumination

In a practical IPL system reflections from beamsplitter surfaces or other paths may result in additional beams at the detector which are coherent with $E_1$ and $E_2$. With the extraneous beam $E_a=|E_a|e^{i\phi_a}$, the interferogram irradiance seen at the detector in Fig. 2.3 (with $a=1$) can still be expressed as $I=I_0+|I_1|\sin\phi_e$; however now,

$$I_0 = \frac{1}{2Z_0} (|E_1|^2 + (1+\epsilon^2)|E_2|^2 + 2\epsilon|E_2|^2\cos(\phi_e-\phi_a)) \quad (3.78a)$$
\[ I_1 = \frac{1}{Z_0} |E_i| |E_l| \sqrt{1 + \epsilon^2 + 2\epsilon \cos(\phi_l - \phi_a)} \]  

(3.78b)

and

\[ \phi_e = \phi_i - \phi_l - \phi_o - \phi_m - \phi_e = \phi'' - \phi_m - \phi_e \]  

(3.78c)

where

\[ \epsilon = \frac{|E_a|}{|E_l|} \quad (<<1 \text{ usually}) \]  

(3.78d)

and

\[ \phi_e = \tan^{-1} \left[ \frac{\sin \phi - \cos \phi a}{\cos \phi + \epsilon \sin \phi_a} \right] \]  

(3.78e)

These expressions suggest that the extraneous beam can be treated as a perturbation to the local oscillator beam. If \( \epsilon \) is small and \( \phi_a \) is spatially uniform, \( E_a \) causes a small shift of \( \phi_e \) away from \( \phi_l \), and hence a small shift in the system reference phase. Spatial variations in \( \phi_a \) appear as a small spatially varying error in the phase estimate and in the compensated beam. Eq. (3.78a) suggests that in equilibrium \( (\phi_e = 0) \), low contrast fringes may still be visible in the interferogram. If \( \epsilon \) approaches unity, the L.O. beam will be completely cancelled where
\( \phi_L - \phi_c = (2n-1)\pi \) holds, and \( I = \frac{E_i d^2}{2Z_0} \) will convey no phase information to the IPL. With \( \epsilon = 1 \), \( E_a \) can also make a significant contribution to the system shot noise.

In conclusion, extraneous coherent beams are essentially ignored by the homodyne IPL as long as they are much weaker than the local oscillator beam.

3.1.9 Irradiance required by the IPL

The minimum signal irradiance, \( I_i = |E_i|^2 / 2Z_0 \), required by the IPL depends on noise considerations as well as the performance specifications of the application.

It was suggested in section 3.1.7 that noise alone does not generally place a lower bound on \( I_i \); for example when \( r = |E_i| / |E_L| \) is held constant, the SNR as given by Eq. (3.76b) was found to be independent of \( I_i \). On the other hand, many aspects of IPL performance such as phase tracking error or bandwidth are degraded as \( G_1 \) is reduced. Since \( G_1 = g' |E_i| / |E_L| / Z_0 = q'I_1 \) (from Eq. (2.14b) with \( \alpha = 1 \) and \( g' = g_0 q_1 q_2 \)), specifying a minimum gain \( G_1 > G_m \) places a lower bound on \( |E_i| \); i.e.

\[
|E_i| > \frac{G_m Z_0}{g' |E_L|}
\]  \hspace{1cm} (3.79a)

This can also be written as

\[
I_i > \frac{G^2}{g' Z_L I_L} \hspace{1cm} (3.79b)
\]

or

\[
I_1 > \frac{G_m}{g'} \hspace{1cm} (3.79c)
\]

where \( I_L \equiv |E_L|^2 / 2Z_0 \).

For example, the low-pass homodyne IPL requires \( G_1 > G_m \pi \) in order to properly
converge with any value of $\phi'_1$.

When the electronic gain $g'$ is limited by hardware considerations, Eq. (3.79b) directly gives the minimum required input irradiance. However, Eqs. (3.76b) and (3.76c) suggest that requiring a minimum S/N, i.e.

$$S/N > m ,$$

(3.80)

may impose a constraint on $g'$ in Eqs. (3.79) which is below the hardware bound. This is most likely in high-resolution applications, where $A_r$ is small.

The choice of $G_m$ in Eqs. (3.79) is dependent on the desired performance in a given application. For example, if the input phase fluctuations have bandwidth $w_B$, an integrating ($\hat{H}=w_o/s$) or low-pass ($\hat{H}=w_o/(s+w_o)$) homodyne IPL will require: (assuming $G_m >> 1$)

$$G_1 > w_B/w_o = G_m$$

(3.81a)

The modulator dynamic range should be large enough to encompass the range of expected fluctuations in $\phi'_1$. Utilization of the full modulator dynamic range $\phi_S$ in a low-pass homodyne IPL requires

$$G_1 > \phi_S/2 = G_m$$

(3.81b)

The maximum error in the low-pass homodyne IPL generally occurs at the extremes of $\phi_m$ (i.e., $0 \leq \phi_m \leq \phi_S$). When the system is designed with $G_o = \phi_S/2$, it follows from Eq. (3.21) that the estimation error or the phase compensation flatness will be better than $\phi_{em}$, i.e., $|\phi_e| < \phi_{em}$, if

$$G_1 > \phi_S/2\sin\phi_{em} = G_m$$

(3.81c)

In some applications it may be advantageous to specify the system performance in terms of the minimum acceptable compensated Strehl ratio, $\eta_m$. Under a variety of assumptions outlined in Appendix D, it follows from Eq. (3.31) that
\[
G_1 > \frac{\sigma_i}{\sqrt{-\ln(\eta_m)}} = G_m
\]  
(3.81d)

is required. Here \(\sigma_i\) is the rms magnitude of the phase fluctuations, and \(G_1 \gg 1\) is assumed.

In practice, the maximum value of \(G_m\) resulting from consideration of all the relevant performance constraints should be employed.

The constraint in Eq. (3.81c) is an oversimplification when \(I_1\) fluctuates over a large dynamic range, because \(G_0\) is dependent on \(I_1\). Given a maximum acceptable error of \(\phi_{em}\), Eqs. (3.5) and (3.6) reveal that

\[
|E_1| > |E_2| \left( \sqrt{-\cos^2 \phi_{em} + (\phi_i - \phi_{em})/g' I_t - I_{em} I_t - \sin \phi_{em}} \right)
\]  
(3.82a)

is required. This relation is consistent with the discussion in section 3.1.1, and results in the lower bound of Eq. (3.10) (case 1) when \(\phi_{em} = \frac{\pi}{2} + 2n\pi\), the bound in Eq. (3.13) (case 2) when \(\phi_{em} = \frac{\pi}{2} + 2n\pi\), and according to Eq. (3.16) (case 3) \(|E_1|\) has no lower bound if \(\phi_{em} > |\phi_i - g'(I_e - I_t)|\). Since \(|E_1|\) is assumed small, Eqs. (3.5) and (3.6) show that the simpler approximation

\[
|E_1| > \frac{Z_0}{|E_2| \sin \phi_{em}} \frac{\phi_i - \phi_{em} + I_t - I_e}{g'}
\]  
(3.82b)

is often valid. Given \(\phi_{em}, I_t, I_e\) and a \(2\pi\) range of \(\phi_i\), Eqs. (3.82) have no influence if they bound \(|E_1|\) at a lower value than Eqs. (3.79) and (3.81).

If Eqs. (3.82) constrain \(|E_1|\) to larger values and the term dependent on \(g'\) is dominant, Eqs. (3.82) and (3.80) determine the maximum value of \(g'\) and the minimum value of \(|E_1|\). When Eqs. (3.82) constrain \(|E_1|\) to larger values, but the \(g'\) dependence is negligible, Eqs. (3.82) alone determine the minimum required irradiance.

Based on section 3.1.7, the noise constraint, \(S/N > m\), can be stated more explicitly. For a homodyne IPL with a type-one (one or more pole)
or low-pass (with $G_1 > 1$) loop filter, Eqs. (3.62)-(3.64) show that Eq. (3.80) becomes:

$$S/N = 4n_{11} + r^2 + 2rsin\phi_e + I_n/I_\xi > m$$  \hspace{1cm} (3.83)

where

$$I_n = I_d + \frac{n\lambda^2 I}{h\nu} + \frac{2h\nu k T_e}{e^2 n A_r R_e}$$

is the noise contribution due to dark current and background shot noise, and thermal noise. The dark current, $I_d$, generally dominates in a homodyne IPL with a photocathode detector (e.g., using an MSLM). From Eq. (3.76c),

$$n_1 = \frac{n\lambda^2 I}{h\nu}I_1 = \frac{n\lambda}{h\nu w_0} q = \frac{n_w}{G_1}$$  \hspace{1cm} (3.84a)

As mentioned previously, $n_1$ is independent of $I_1$ and hence $|E_1|$. $n_w$, which depends on $|E_1|$, is defined by

$$n_w = \frac{n\lambda}{h\nu} w_0^{-1} I_1$$  \hspace{1cm} (3.84b)

and is the average number of photons collected in a resolution element during the time $w_0^{-1}$; where $w_0^{-1}$ is, for example, the open-loop time constant of a low-pass loop filter. With an integrating filter, $w_0^{-1}$ can be taken as the frame-time $t_w$ during which the integrator is operated.

$w_0^{-1} = t_w$ can be seen by noting that Eq. (2.11), with $I_t = 0$, $\dot{\gamma} = \omega_0/s$, and approximately constant irradiance during $t_w$, results in $\phi_m = w_0 g'I = \Delta \phi_m/t_w$ and Eqs. (2.6)-(2.8) show that $g' = \Delta \phi_m/I$. Further consideration reveals that $g'$ is proportional to $t_w$, (e.g. Eq. (4.38b)), so that the closed-loop response, $w_0 G_1 = g'I_1/t_w$, is independent of $t_w$, as expected.)

A few examples of the minimum irradiance required by the IPL operating under a variety of noise and other performance constraints follow. In some low-resolution applications it may be feasible to track $I_1$ with $I_{\xi}$ and hence operate with constant $r = |E_1|/|E_{\xi}|$. For example, assuming that
\[ r \gg I_1 / I_0 \] and allowing for large \( \Phi_e \), Eq. (3.83) can be written

\[ S/N = 4n_r r / (1+r)^2 > m \text{ or} \]

\[ g' < \frac{n A_r}{\text{mhw}_0} \frac{4r}{(1+r)^2} \]

A performance constraint of the form in Eqs. (3.79c) then specifies the minimum level of \( I_1 \) as

\[ I_1 > \frac{mG_{\text{mhw}_0} (1+r)^2}{n A_r} \frac{4r}{4r} \]

Greater intuitive insight can be gained by utilizing Eq. (3.84b) to rewrite this condition as

\[ n_w > \frac{m G_m (1+r)^2}{4r} \]

or

\[ n_{w_1} > \frac{m G_m (1+r)^2}{8} \]

where

\[ n_{w_1} = \frac{n A_r}{n_{\text{hw}_0}} I_1 = \frac{r}{2} n_w \quad (3.84c) \]

For example, holding \( r = 1 \) and requiring \( G_1 > 2\pi = G_m \) and \( S/N > 4 = m \), specifies that

\[ n_{w_1} > 4 \pi \approx 12.6 \text{ photons} \]

the corresponding minimum irradiance is thus that which provides about 13 photons in a resolution element during the open-loop time constant. (This amounts to only 2 photons during the closed-loop time constant.)

In high-resolution applications, it is generally more practical to fix the L.O. intensity and allow \( r \) to diminish as \( |E_1| \) decreases. With \( r \ll 1 \), the \( S/N \) restriction of Eq. (3.83) becomes

\[ S/N = 2n_1 I_1 / (I_1 + I_n) = n_1^2 / (n_1 + n_n) = \frac{8 n_1^2}{r^2 n_1 + n_n} > m \quad (3.85) \]

where \( n_1 \) and \( n_n \) are the average number of L.O. photons and noise counts (e.g. dark current) collected, respectively, during \( \tau = 1/G_1 w_0 \) in a resolution element of area \( A_r \). As expected for a noisless process, the shot noise is
proportional to \( n_x + n_n \). The noise constraint can be rewritten as

\[
I_1 > \frac{m g' h w_0 (I_x + I_n)}{2 \nu A_r}
\]

This condition in conjunction with the performance constraint of Eq. (3.79c) reveals that there is an optimal value of \( g' \), which allows the SNR and other performance constraints to be met with minimum \( I_1 \) or \( |E_1| \).

The optimum \( g' \), which occurs where the two boundary conditions intersect (consider both conditions plotted on a graph of \( I_1 \) versus \( g' \)), is

\[
g' = \sqrt[2nA_r]{\frac{2 G_m }{m h w_0} (I_x + I_n)}
\]

The corresponding minimum \( I_1 \) irradiance is

\[
I_1 > \sqrt{\frac{G_m h w_0 \nu}{2nA_r} (I_x + I_n)}
\]  \hspace{1cm} (3.86a)

or

\[
n_w > \frac{G_m h w_0}{2nA_r} \sqrt{n_w + n_w}
\]  \hspace{1cm} (3.86b)

where \( n_w + n_w = \frac{nA_r}{h v w_0} (I_x + I_n) \)

When L.O. shot noise dominates, i.e. \( I_x >> I_n \), Eq. (3.86b) specifies the minimum required input irradiance as

\[
I_i > \frac{m G_m h w_0}{8nA_r}
\]

or

\[
n_w > \frac{m G_m}{8}
\]  \hspace{1cm} (3.87a)

With optimal \( g' \), S/N>4, and \( G_1 > 2 \pi \); \( n_w > \pi \) implies that about 3 photons are required in a resolution cell during the open-loop time constant.

When \( I_x << I_n \), e.g. when dark current shot noise dominates, Eq. (3.86b) becomes

\[
n_w > \sqrt{G_m n w_0 / 2} \quad \text{or} \quad n_w > \frac{r G_m n w_0 / 2}{2}
\]  \hspace{1cm} (3.87b)
It should be noted that these conditions can often be satisfied when the IPL irradiance, \( I_0 + I_1 \sin \phi_e \), is considerably less than the effective noise irradiance \( I_n \).

When the desired tracking bandwidth \( w_B \) is the dominant constraint, Eq. (3.81a) showed that \( G_{s} > w_B / w_0 = G_m \) is required. Eq. (3.86b) then becomes

\[
\frac{n_B}{h} w_B^{-1} I_1 > \sqrt{m/2} \sqrt{n_B^2 + n_B^n} = \sqrt{n_B^2 + n_B^n} \tag{3.88a}
\]

where

\[
\frac{n_B}{h} w_B^{-1} (I_L + I_n)
\]

Here \( n_B \) is the average number of photons collected from the \( I_1 \) signal in a resolution element during the shortest time during which \( I_1 \) remains constant (e.g., the phase "coherence time"). When L.O. shot noise dominates, this constraint becomes

\[
\frac{n_B}{h} w_B^{-1} I_1 > \frac{m}{8} \tag{3.88b}
\]

In cases where Eq. (3.82a) is the dominant condition for the values of \( I_0, I_t \), and \( \phi_{em} \) in a particular application, Eqs. (3.82a) and (3.83) can be solved for the optimum \( q' \) and associated minimum \( |E_q| \), however the resulting cubic equation can be difficult to solve.

In summary, the minimum irradiance required by the IPL is a function of the signal-to-noise ratio along with a variety of other performance criteria. These constraints often suggest an optimal value of electronic gain \( q' \) at which the system should be operated for maximum sensitivity. Under ideal circumstances, where L.O. shot noise dominates, the minimum irradiance is that which produces a few photons in each resolution element during the integration time, \( t_w \), or low-pass open-loop time constant \( w_0^{-1} \). When the bandwidth \( w_B \) of the input phase fluctuations
is large ($w_B \gg w_o$), this condition becomes a few photons per phase
coherence time, $T_\phi = w_B^{-1}$. In instances where other noise sources such
as dark current shot noise dominate, the average number of photons collec-
ted from the $I_1$ irradiance should exceed the square root of the average
number of noise counts; this can allow operation with signal irradiances
below the effective noise irradiance.
3.2 Alternative interferometric implementations

The homodyne IPL, which has been utilized to introduce all of the IPL operational theory up to this point, is only one of a variety of possible interferometric implementations; such as those incorporating a heterodyne, Zernike phase-contrast, polarization, or shearing interferometer into the loop. The study of these other systems not only provides a greater depth of understanding into the operational theory of the IPL, but is of practical importance for choosing an IPL implementation well suited to a particular application.

The heterodyne and one of the self-interference implementations, Zernike phase-contrast, have been chosen for consideration in this section; they provide a rounded perspective of the variability of IPL behavior. The heterodyne implementation is of particular interest in adaptive receiver applications because it may be easier to maintain a stable receiver local oscillator near, rather than exactly at, the optical frequency of the distant transmitter laser. The heterodyne analysis is also applicable to a homodyne IPL with a slowly drifting local oscillator (L.O.). The phase-contrast configuration may also be useful in some adaptive receiver applications. It offers the simplification of no local oscillator, with attendant wider optical bandwidth; but it can be less sensitive, particularly during initial phase acquisition. It will be seen that these two implementations share much in common with each other and with the homodyne IPL. They also exhibit significant differences, which become advantages or disadvantages depending on the application.

3.2.1 The heterodyne IPL

In the heterodyne implementation there is an offset, $\Delta w = w_2 - w_1$, between
the optical frequency of the L.O. beam, at \( w_o \), and the input beam, at \( w_i \).

It is thus convenient to rewrite the input phase \( \phi_i \), which was defined by Eqs. (2.3) and (2.4), as
\[
\phi_i = \phi_{i0} + \Delta \omega t \quad ; \tag{3.89}
\]
where \( \phi_{i0} \) is a function of space but is constant or only slowly varying with time.

The details of heterodyne behavior are dependent on the choice of loop filter. Two varieties of single-pole filters, a type-one integrator and a type-zero low-pass filter, will be considered here. These are applicable to systems which can be characterized by a single dominant pole, including many SLMs. The integrating heterodyne case is a little easier to follow and will be discussed first.

3.2.1.1 The heterodyne IPL with an integrating loop filter

When the integrating filter \( \tilde{H}(s) = \lambda_1/s \) is employed in the loop the basic dynamic equations, Eqs. (2.12) and (2.13), become
\[
\dot{\phi}_m = \lambda_1 (G_0 - G_1 \sin(\phi_m - \phi_{i0} - \Delta \omega t)) \tag{3.90a}
\]
and
\[
\dot{\phi}_e = \Delta \omega - \lambda_1 (G_0 + G_1 \sin \phi_e) \tag{3.90b}
\]
The \( \dot{\phi}_e \) versus \( \phi_e \) state-space plot corresponding to Eq. (3.90b) is illustrated in Fig. 3.11a. This figure is essentially identical to \( \dot{\phi}_e \) versus \( \phi_e \) for the homodyne IPL with an integrating filter, which was given by Eq. (3.33b); the major difference being that the offset term becomes \( \Delta \omega - \lambda_1 G_0 \) in the heterodyne case.

Fig. 3.11a shows that stable equilibria can occur at
\[
\phi_e = \sin^{-1}\left(\frac{\Delta \omega/\lambda_1 - G_0}{G_1}\right) + 2n\pi = 2n\pi \tag{3.91a}
\]
Using \( \phi_m = \phi_i - \phi_e \), the corresponding modulator steady-state is
Fig. 3.11 State-space behavior of the heterodyne IPL with an integrating loop filter. \((G_1=2\pi, G_o=1.8\pi, \Delta w=0.5\pi\lambda_1, G_c=0\pi, G_s=3.3\pi, \phi_{io}=0\pi)\)

The dots in (b) are spaced at a time interval of \(0.2/\lambda_1\). (From Eq. (3.72b) the time constant of the system is \(0.209/\lambda_1\).)
\[
\phi_m = \phi_{10}^* + \Delta wt - \phi_e = \phi_{10}^* + \Delta wt - 2n\pi 
\]  
(3.91b)

The approximations on the right of Eqs. (3.91) follow for large \( G_1 \). The time behavior of \( \phi_e \) during convergence is given by Eqs. (3.43b) and (3.47) with \( G_0' \) replaced by \( \lambda_1 G_0 - \Delta w \), and Eqs. (3.49). When the error remains small during convergence, the integrating heterodyne behavior is well described by the linearized expressions in Eqs. (3.37). The linearized solution and more general steady-state behavior of Eqs. (3.91) reveals that the modulator exactly tracks the ramp \( \Delta wt \), as expected for a type-one system. The additional static error due to the ramp is negligible if \( \Delta w \ll \lambda_1 G_1 \).

The modulator state-space is not immediately obvious, because \( \phi_m \) in Eq. (3.90a) is directly dependent on time. However, given an initial condition \( \phi_{m0} = \phi_m(t=0) \), a \( \phi_m \) versus \( \phi_e \) state-space trajectory can be plotted by using \( \phi_e(t) \) from Eq. (3.46), (3.49), or (3.51) in conjunction with Eq. (3.90a) and the relation \( \phi_m = \phi_{10}^* - \phi_e \). Fig. 3.11b illustrates the modulator state-space corresponding to the error state-space in Fig. 3.11a. This state-space is meaningful since a steady-state behavior is eventually reached which is independent of the initial condition. Examination of Eq. (3.90a) shows that Fig. 3.11b can be interpreted as resulting from laterally shifting, with \( \Delta wt \), the homodyne state-space, e.g. Fig. 3.6a, during convergence. In equilibrium, \( \phi_m \) continues to track \( \Delta wt \) and hence converges to \( \dot{\phi}_m = \Delta w \), rather than \( \dot{\phi}_m = 0 \). If the maximum modulator phase is \( \phi_s \), the system will eventually reach and lock-up at \( \phi_m = \phi_s \) with \( \dot{\phi}_m = 0 \).

Notice that the heterodyne implementation removed the "dead regions" from the integrating homodyne state-space (e.g., Figs. 3.7). According to Eq. (3.90b) and Fig. 3.11a, with proper choice of \( \Delta w \gg 0 \), the \( \phi_e \) trajectory can be vertically shifted to include \( \dot{\phi}_e = 0 \). This is consistent with the
earlier observation that even with dead regions \( \phi_m \) can track increasing \( \dot{\phi}_1 \). In both the homodyne and heterodyne implementations the integrator output, and hence the modulator phase, always increases. However in the heterodyne case, the phase estimate, \( \phi_m \), decreases relative to \( \dot{\phi}_1 \) when \( \dot{\phi}_m \) increases more slowly than \( \Delta \omega t \).

With the detector limitations, \( 0 \leq G_0 + \dot{G}_1 \sin \phi_e \leq G_s \), the trackable rate of change in \( \dot{\phi}_{10} \) is limited to
\[
-\Delta \omega < \dot{\phi}_{10} < G_s - \Delta \omega
\]  
(3.92a)

In the absence of detector limits the tracking bandwidth of \( \dot{\phi}_{10} \) becomes
\[
|\dot{\phi}_{10}| < BW = \lambda_1 \sqrt{G_1^2 - (G_0 - \Delta \omega / \lambda_1)^2} = 1/\tau
\]  
(3.92b)

Although the linear system of Eq. (3.37) will exactly track any \( \Delta \omega t \) ramp, the actual nonlinear system is limited to
\[
\lambda_1 (G_0 - \dot{G}_1) < \Delta \omega < \lambda_1 (G_0 + \dot{G}_1)
\]  
(3.92c)

(assuming \( \dot{\phi}_{10} = 0 \)). With larger gain, \( G_1 >> G_0 \), this becomes simply \( |\Delta \omega| < \lambda_1 G_1 \).

The inclusion of detector limits reduces this range to
\[
0 < \Delta \omega < \lambda_1 G_s
\]  
(3.92d)

Examples of \( \phi_e \) and \( \phi_m \) state-spaces for an out of range frequency are illustrated in Figs. 3.12, (with no detector or modulator limitations).

This system is unstable, exhibiting a limit-cycle in \( \dot{\phi}_e \) which is described by Eqs.(3.51) with \( G'_0 = \lambda_1 G_0 - \Delta \omega \), and has period
\[
T_c = 2\pi / \lambda_1 \sqrt{(G_0 - \Delta \omega / \lambda_1)^2 - G_1^2}
\]  
(3.93a)

There are no stable equilibria where the \( \dot{\phi}_m \) versus \( \phi_m \) trajectory crosses the \( \dot{\phi}_m = 0 \) axis because the trajectory does not have a negative slope above and below the axis.
Fig. 3.12 State-space behavior of the integrating heterodyne IPL with an out of range frequency offset. With $G_1=2\pi$, $G_o=1.8\pi$, $\Delta w=-0.5\pi_1$, $\phi_{i0}=0$, $\phi_{mo}=0$. 

a) Error state-space

b) Modulator state-space
The heterodyne offset shift from \( \lambda_1 G_0 \) to \( \lambda_1 G_0 - \Delta w \) in Eq. (3.90b) and Fig. 3.11a tends to improve IPL performance. The relative gain can be increased by choosing \( \Delta w = \lambda_1 G_0 \), so that \( G_1 >> G_0 - \Delta w / \lambda_1 \). When this condition holds the \( \phi_e \) trajectory in Fig. 3.11a crosses the \( \phi_e \) axis with its maximum slope, \( (\lambda_1 G_1) \). The corresponding phase tracking error in Eq. (3.91) goes to zero (modulo 2\( \pi \)), the immunity to amplitude fluctuations is improved, and the phase acquisition time and tracking bandwidth are optimized.

Performance with wide-ranging amplitude fluctuations is best when the IPL can be designed with

\[
\Delta w = q(|E_\perp|^2 - I_i^*) \tag{3.93b}
\]

Then from Eq. (3.35a), \( \phi_e = \sin^{-1}(0.5q|E_\perp|/|E_i|) \); and \( |E_i| \) fluctuations have little effect when a strong L.O. beam \( |E_i| > |E_i| \) is employed.

Since Eq. (3.93b) can be satisfied with \( I_i^* = 0 \), no intensity threshold or offset is required to achieve this behavior.

Fig. 3.13 illustrates that the heterodyne IPL with an integrating loop filter is essentially identical to a conventional electronic phase-locked loop. The frequency of the compensated input beam after the modulator, \( W_L + W_m \), becomes locked to the local oscillator optical frequency, \( W_L \).

Actually, with the appropriate choice of loop filter, \( H_i(s) \), all the IPL implementations can be viewed as phase-locked loops; although some, such as the homodyne implementation, are initially in frequency lock.

The continuous growth of \( \phi_m \) at about the rate \( \Delta w \) causes an actual integrating IPL to eventually lock up at \( \phi_m = \phi_s \). In practice the integrator and hence the modulator can be periodically reset to an initial value, \( \phi_{mo} \), resulting in a discontinuous framed mode of operation. This is feasible since initial phase acquisition, occurring with time constant \( \tau \) in Eq. (3.92b) is usually much more rapid than the drift rate, \( \Delta w \). By
Fig. 3.13 The Interference Phase Loop as a phase-locked loop.

Here $H'(s)=sH(s)=1$ for the integrating IPL. The integrator and MODulator combination corresponds to a voltage controlled oscillator (VCO) with center frequency $w_i$ and output frequency $w_i - w_m$. In frequency lock $\phi_m = \phi_i = \omega = w_m - w_i$ and the output of the VCO is locked to $w_i$, the frequency being tracked.
limiting the frame time to about \( T_f = (\phi_s - \phi_{mo} - \pi) / \Delta w \), \( \phi_m \) can be prevented from reaching \( \phi_s \). With the detector induced limits, \( 0 < G_o + G_1 \sin \phi_e < \pi / G_s \), Fig. 3.11a suggests that the worst case acquisition time increases, but will not exceed the larger of \( \pi / \Delta w \) or \( \pi / G_s \). The fraction of time spent usefully tracking phase can be made large by employing a modulator with a dynamic range of multiple \( \pi \) radians.

Alternatively, if a large dynamic range modulator is not available, a sample and hold framed mode can be employed. The IPL is turned on for a period of about \( \pi / \Delta w \) seconds to allow acquisition of all possible input phases. The input to the detector is then blocked, and the integrator holds the phase estimate, \( \phi_m \), for the rest of the frame time. The framing rate is determined by how fast \( \phi_{io} \) varies, and the required modulator dynamic range is about \( 3\pi \) radians. These framed modes will be discussed further in section 4.3.4 in the context of implementing an IPL with an integrating version of the MSLM.

3.2.1.2 The heterodyne IPL with a low-pass loop filter

Some dominant-single-pole filters, including those intrinsic to a few SLMs, are better described by a type-zero transfer function of the form \( \hat{H}(s) = \lambda_1 / (s + \lambda_1) \). The corresponding dynamic equations follow from Eqs. (3.89), (2.12), and (2.13) as:

\[
\dot{\phi}_m = \lambda_1 (-\phi_m + G_o - G_1 \sin(\phi_m - \phi_{io} - \Delta wt)) \quad (3.94a)
\]
and

\[
\dot{\phi}_e = \Delta w - \lambda_1 (\phi_e - \phi_{io} - \Delta wt + G_o + G_1 \sin \phi_e) \quad (3.94b)
\]

The underlined terms in Eqs. (3.94) distinguish these expressions from the integrating dynamic equations of Eqs. (3.90).
The state-space behavior associated with Eqs. (3.94) is not immediately apparent; as neither relation is periodic in $\phi_e$ or $\phi_m$, and both exhibit a direct time dependence. Nevertheless, state-space trajectories can be computed by specifying the initial values $\phi_{mo}$ and $\phi_i(t=0)$ and employing Eqs. (3.58a,b,c). Computed error ($\dot{\phi}_e$ versus $\phi_e$) and modulator ($\dot{\phi}_m$ versus $\phi_m$) state-space trajectories are illustrated in Figs. 3.14 for the same parameter values ($G_1=2\pi$, $G_0=1.8\pi$, $\Delta w=0.5\pi\lambda_1$) as those employed for the integrating trajectories in Fig. 3.11, but with no detector or modulator limits.

When the initial value of modulator phase is small ($\phi_{mo}<G_1$), Eqs. (3.94) are the same as Eqs. (3.90) for the integrating filter case. Hence the initial segment of the state-space trajectories from $\phi_{mo}=0$ in Figs. 3.14 (bold segments) essentially follow the integrating trajectories of Figs. 3.11. However as $\phi_m$ grows, the low-pass behavior differs dramatically. The error state-space converges to a state-space periodic in $\phi_e$ and the modulator develops a stable limit-cycle oscillation in $\phi_m$. The periodicity of $\dot{\phi}_e$ versus $\phi_e$ is not surprising considering that $\ddot{\phi}_e$ is periodic in $\phi_e$ and time independent; i.e. taking the time derivative of Eq. (3.94b) results in

$$\dddot{\phi}_e = \lambda_1 (\Delta w - (1+G_1 \cos \phi_e) \dot{\phi}_e)$$

The maxima and minima of the $\dot{\phi}_e$ trajectory occur where $\dddot{\phi}_e = 0$, which corresponds to the curves

$$\dot{\phi}_e = \frac{\Delta w}{1+G_1 \cos \phi_e}$$

One cycle of this curve is dotted in Fig. 3.14a. This constraint reveals that the minimum possible value of $\dot{\phi}_e$ on the limit-cycle is $\Delta w/(1+G_1)$ and that an equilibrium with $\dot{\phi}_e=0$ and $\dddot{\phi}_e=0$ is not possible.
Fig. 3.14 State-space behavior of the low-pass heterodyne IPL with $\Delta w = 0.5\pi \lambda_1$. For $G_1 = 2\pi$, $G_0 = 1.8\pi$, $\phi_{i0} = 0$.

(A occurs at time 0, B at $0.5/\lambda_1$, C at $6.6/\lambda_1 + 4n/\lambda_1$, D at $(9.0 + 4n)/\lambda_1$)
This constraint also tends to raise the low-pass trajectory above that of the integrating case in Fig. 4.11a. (The corresponding integrating constraint is more lenient, with \( \dot{\phi}_e = -\lambda_1 \dot{\phi}_e G_1 \cos \phi_e \).) Thus for parameter values which lead to a stable error in the integrating case, the error in the low-pass case grows continuously.

The modulator state-space can be understood by noting from Eq. (3.94a) that the heterodyne IPL is the same as a homodyne IPL (Eq. (2.29)) with continuously increasing input phase (\( \phi_0' = \phi_0 + \Delta \omega t \)). Referring back to Fig. 2.7, there is an initial rapid convergence along a homodyne trajectory toward \( \dot{\phi}_m = \dot{\phi}_i' \). The modulator then attempts to track \( \dot{\phi}_i' \), causing the trajectory in Fig. 2.7 to shift to the right with \( \Delta \omega t \), and \( \dot{\phi}_m \) to converge to \( \dot{\phi}_m = \dot{\phi}_i' - \Delta \omega \) rather than zero. Unlike the integrating case, \( \dot{\phi}_m \) is not exactly equal to \( \Delta \omega \), as evidenced by the fact that \( \dot{\phi}_e = \Delta \omega - \dot{\phi}_m \) in Fig. 3.14a doesn't stabilize at zero. Corresponding points on the \( \dot{\phi}_e \) and \( \dot{\phi}_m \) trajectories in Figs. 3.14 (for \( \phi_{i0}' = 0 \), \( \phi_{m0} = 0 \)) are denoted by the letters A, B, C, D. In the absence of detector and modulator physical limits, the top of the sinusoidal trajectory in Fig. 2.7 eventually drops below \( \dot{\phi}_m \approx \Delta \omega \), (approximately when \( \dot{\phi}_m = \lambda_1 (G_0 + G_1 - \phi_m) \times \Delta \omega \) occurs), and a decreasing sinusoidal trajectory is followed back to a smaller value of \( \phi_m \). This decreasing trajectory is not much distorted from the homodyne case by the simultaneous change in \( \Delta \omega t \), as long as \( \Delta \omega t \lambda_1 (G_1 + 1) \) holds; note that the trajectory touches the line \( \dot{\phi}_m = \lambda_1 (G_0 - G_1) \times \Delta \omega \). The modulator phase then increases until \( \dot{\phi}_m \approx \Delta \omega \) again, and thus enters an endless limit-cycle oscillation.
The effects of detector or modulator physical limitations are easily incorporated into the programs utilized to compute state-space trajectories. Eq. (3.94a) suggests that the limits can be drawn on the modulator state-space as a trapezoid in the same manner as for the homodyne IPL in Fig. 2.7. The detector limits flatten the sinusoidal trajectory, and whichever limit restricts the maximum excursion of \( \phi_m \) will give rise to a limit-cycle oscillation. For the specific limits in Fig. 2.7, \( \phi_m \) increases until \( \phi_m = \phi_s \) ; pauses at \( \phi_m = \phi_s \) with \( \dot{\phi}_m = 0 \); and then follows a decreasing trajectory to a value of \( \phi_m \) about \( \pi \) radians smaller. This oscillation will then continue as long as the IPL operates. In contrast, \( \phi_m \) in the integrating IPL normally rises until becoming pinned at \( \phi_m = \phi_s \) with \( \dot{\phi}_m = 0 \), since \( \dot{\phi}_m < 0 \) cannot occur. However, when the amplifier includes an additional offset, as in Eq. (3.38) and Fig. 3.6, \( \phi_m \) can decrease and a limit-cycle oscillation with an upper bound of \( \phi_s \) can arise in the integrating case also.

The time \( T_c \) of one period of the limit-cycle in Figs. 3.14 can be found by applying the constraint \( \dot{\phi}_e(\phi_e + 2n\pi, t+nT_c) = \dot{\phi}_e(\phi_e, t) \) to Eq. (3.94b), with the result that

\[
T_c = \frac{2\pi}{|\Delta w|} \quad (3.95)
\]

This is in fact the observed period of the oscillations in the numerically computed behavior.
An expression can also be written for the approximate time, $t_0$, to traverse the trajectory from A to $\phi_e=0$ in Fig. 3.14a. When $\phi_e=0$, Eq. (3.74a) with $\phi_m=\Delta w/\lambda_1$ reveals that $\phi_m=G_0-\Delta w/\lambda_1$ and hence $\phi_e=\phi_{i0}+\Delta w t_0-\phi_m = 0 = \phi_{i0}+\Delta w t_0+\Delta w/\lambda_1-G_0$. The time interval is thus $t_0 = (G_0-\Delta w/\lambda_1-\phi_{i0})/\Delta w$. (Care must be taken to resolve the $2\pi$ multistability of $\phi_m$ so that $t_0>0$ always results). This expression for $t_0$ agrees with the numeric results.

Figs. 3.15 illustrate computed error and modulator state-space trajectories for the same parameter values which resulted in the unstable integrating trajectories of Figs. 3.12. Note that the low-pass error trajectory is quite similar to the integrating trajectory in Fig. 3.12a. The period for a cycle in $\phi_e$, however, differs significantly between the two cases; the low-pass period follows from Eq. (3.95) as $4/\lambda_1$, while the integrating period follows from Eq. (3.93) as $1.76/\lambda_1$. The low-pass modulator trajectory is also markedly different, exhibiting a limit-cycle oscillation at small $\phi_m$.

With large frequency offset, $\Delta w > \lambda_1(G_1+1)$, the homodyne state-space shifts faster than it can be transversed. Fig. 3.16 illustrates that the behavior is significantly modified.

Additional insight into the behavior of the low-pass heterodyne IPL can be gained by examining the linearized ($\sin\phi_e=\phi_e$) behavior in the vicinity of $\phi_e=2\pi$. Linearizing the system dynamic relations, Eqs. (3.94), and solving the resulting differential equations results in
Fig. 3.15 State-space behavior of the low-pass heterodyne IPL with $\Delta w = -0.5\pi \lambda_1$. With $G_1 = 2\pi$, $G_0 = 1.8\pi$, $\phi_{i0} = 0$, $\phi_{m0} = 0$. 
Fig. 3.16 State-space behavior of the low-pass heterodyne IPL with $\Delta \omega \gg \lambda_1 G_1$.
($G_1 = 2\pi$, $G_0 = 1.8\pi$, $\Delta \omega = 4\pi \lambda_1$, $\phi_{10} = 0$, $\phi_{m0} = 0$)

(a) Error state-space

(b) Modulator state-space
\[
\phi_m = \phi_{m0} e^{-\lambda_0 t} + \phi_{ms} (1-e^{-\lambda_0 t}) + \frac{G_1}{1+G_1} \Delta w t \tag{3.96a}
\]

and
\[
\phi_e = (\phi_{i0} - \phi_{m0}) e^{-\lambda_0 t} + \phi_{es} (1-e^{-\lambda_0 t}) + \frac{1}{1+G_1} \Delta w t \tag{3.96b}
\]

where
\[
\lambda_0 = \lambda_1 (1+G_1) \equiv 1/\tau_0 \tag{3.96c}
\]

\[
\phi_{ms} = \frac{G_1}{1+G_1} (\phi_{i0} + \frac{G_0}{G_1} - \frac{\Delta w}{\lambda_0} + 2n\pi) \tag{3.96d}
\]

and
\[
\phi_{es} = \frac{\phi_{i0}}{1+G_1} - \frac{G_1}{1+G_1} (\frac{G_0}{G_1} - \frac{\Delta w}{\lambda_0} + 2n\pi) \tag{3.96e}
\]

Unlike the integrating case, the modulator frequency in Eq. (3.96a) is not exactly equal to the input frequency, \(\Delta w\). This inability to track a ramp is expected for a type-zero system, and results in the error growing continuously with time, rather than stabilizing with \(\phi_e = 0\) as in the integrating case.

The steady-state linearized behavior is a good description of the segment of the state-space trajectory from B to C in Figs. 3.14. There \(\phi_e = 0\), \(\dot{\phi}_e = \frac{\Delta w}{1+G_1}\), and \(\dot{\phi}_m = \frac{G_1}{1+G_1} \Delta w\) are observed in the computed behavior.

The low-pass heterodyne IPL can perform useful phase estimation along this segment. Although the error grows continuously, a spatially uniform, time-varying error term does not degrade the phase compensation and spatially-relative phase estimation performance of the IPL. In fact, the growth of \(\phi_m\) at a slower rate than \(\Delta w\) has the advantage of increasing the useful tracking time until \(\phi_m\) reaches its operational limits. Along the B to C segment in Fig. 3.14, slow changes in \(\phi_{i0}\) are tracked by \(\phi_m\), as long as \(\dot{\phi}_i\) is within the system bandwidth; i.e.,

\[
|\dot{\phi}_{i0} + \Delta w| < \lambda_1 (1+G_1) \tag{3.97a}
\]
The maximum trackable frequency offset of the low-pass heterodyne IPL is thus, (for \( \phi_{10} = 0 \)),

\[
|\Delta w| < \lambda_1 (1 + G_1) = \lambda_1 G_1
\]

(3.97b)

Fig. 3.16 shows that when this condition is not met a useful phase tracking segment does not occur.

Summarizing, the low-pass heterodyne IPL (with \( |\Delta w| < \lambda_1 (1 + G_1) \)) has three successive regimes of operation. There is an initial rapid convergence of \( \phi_m \) to \( \phi_{10} + \text{const}(x,y) \) occurring with time constant \( \tau = 1/\lambda_1 (1 + G_1) \). This is followed by a useful real-time phase tracking interval \( T_t \) during which \( \phi_m \) includes a spatially uniform growth term. In the absence of detector or modulator limits, \( \phi_m \) drifts during the tracking time, \( T_t \), from \( \phi_{10} \) until \( G_0 + G_1 - \frac{\Delta w}{1 + G_1} \), (e.g., see Fig. 3.14), hence

\[
T_t = \frac{G_0 + G_1 - \phi_{10} + 2n\pi}{\frac{1}{1 + G_1} - \frac{\Delta w}{\lambda_1}} - \frac{1}{\lambda_1}
\]

(3.98a)

where

\[
n = \text{INT} \left( \frac{\phi_{10} - \phi_{mo}}{2\pi} \right) + \frac{1}{\lambda_1}
\]

(3.98b)

and \( \phi_{mo} \) is the initial value of \( \phi_m \). By increasing \( G_1 \) (and \( \phi_s \)) relative to \( \phi_{mo} \), \( T_t \) can be made arbitrary large. After the tracking interval, \( \phi_m \) enters into a limit-cycle oscillation with period \( 2\pi/\Delta w \).

In practice, the low-pass heterodyne IPL could be forced to operate in its useful tracking regime by employing a framed operating mode, with \( \phi_m \) being reset to \( \phi_{mo} \) at intervals of \( T_t \) in order to avoid limit-cycle oscillations. Another potential operating mode is a sample and hold framed mode, where \( \phi_m \) is allowed to converge for about \( 5\tau_0 \) seconds and then frozen for the rest of a frame-time interval during which \( \phi_{10} \) remains approxi-
mately constant.

A third possibility is to continuously operate in the limit-cycle regime. This would be feasible if the cyclic variation of \( \phi_m \) is spatially uniform and thus preserves a spatially-relative estimate of \( \phi_{i0} \). Fig. 3.17 is a representative computation of the time evolution of the phase estimation error relative to the error with \( \phi_{i0} = 0 \). (The actual parameter graphed is \( \phi_{i0} - [\phi_m(\phi_{i0}) - \phi_m(\phi_{i0} = 0)] = \phi_e(\phi_{i0}) - \phi_e(\phi_{i0} = 0) \) versus time.) This figure reveals that for a given value of \( \phi_{i0} \), \( \phi_m \) is a valid relative estimate for a portion of the cycle time, \( T_c \). During that time slow variations in \( \phi_{i0} \) are also tracked. Limit-cycle operation is thus useable to an extent, but has serious shortcomings.

All three stages of low-pass heterodyne operation; convergence, tracking, and limit-cycle operation; are evident in Fig. 3.17.

3.2.1.3 Conclusions

For both of the loop filters considered, the heterodyne IPL exhibits similar basic behavior. There is an initial rapid convergence to the relative spatial input phase, \( \phi_{i0} \), with a time constant of \( \approx \lambda \frac{1}{2} G_1 \). This is followed by a continuous increasing or decreasing drift of the modulator phase at about the rate \( \Delta \omega \), during which slow variations in \( \phi_{i0} \) (\( \phi_{i0} < \Delta \omega \)) can be estimated and compensated. A third type of behavior occurs when the IPL reaches its phase-tracking dynamic range limits, either limit-cycle oscillations or lock-up at the maximum modulator phase (integrating case).

During the modulator drift stage of operation, the IPL attempts to frequency-lock the output beam from the modulator to the local oscillator optical frequency. This permits IPL operation with an L.O. offset or imperfectly stabilized L.O., for example in adaptive receiver applications.
Fig. 3.17 Relative phase error as a function of time.

For $G_1=3.9\pi$, $G_0=1.8\pi$, $\Delta w=0.5\pi \lambda_1$, $T_c=2\pi/\Delta w=4/\lambda_1$, $\tau=1/\lambda_1 (G_1+1)=0.065/\lambda_1$, and $\phi_{mo}=2\pi$. (The flat tracking interval would be about $4/\lambda_1$ longer if $\phi_{mo}=0$ were used. The $n$ value of a given curve is constant.)
The maximum $\Delta w$ offset which the IPL can tolerate is given by Eq. (3.92a), (3.92d), or (3.97d), depending on the specific system configuration.

As illustrated by the frequency-locking behavior, the heterodyne IPL is very similar to a conventional phase-locked loop (PLL). The PLL literature\textsuperscript{164,165} thus provides additional insights into its operation. For example, Ref.\textsuperscript{164} shows that the fundamental behavior is essentially unmodified with two-pole and other more complicated loop filters; but initial frequency-lock and phase acquisition can take longer, and the constraints on trackable $\Delta w$ can become more restrictive.

The growth of $\phi_m$ at a slower rate than $\Delta w$ in the low-pass filter case is often superior to the perfect frequency tracking of the integrating case (where $\phi_m = \phi_{10} + \Delta wt - \phi_e$ and $\phi_e = \text{const}(t) + f(t)$), because the system can track $\phi_{10}$ longer before $\phi_m$ reaches its operational limits. The time-varying but spatially-uniform error with imperfect frequency-lock does not impair the spatially-relative phase estimation or compensation performance of the IPL. Operation with $\phi_m = \phi_{10} - \text{const}(t)$ and $\phi_e = \text{const}(t) + \Delta wt$ may be even better; $\phi_m$ then takes very long to reach its operational limits. This effectively corresponds to the sample and hold mode of operation; except that changes in $\phi_{10}$ are not tracked during the hold time.

An even more desirable operating mode may be to imperfectly track $\Delta wt$ in such a way that $\phi_{10}$ is always tracked (within a time-varying, spatially-uniform error), and the modulator never reaches its dynamic limits. An example is a bipolar chopping mode, where the modulator alternates with a 50% duty cycle between 0 and $\pi$ radians at the rate $\Delta w$; elementary Fourier analysis reveals that 80% of the modulated signal
power is frequency shifted by $\pm \Delta w$. The low-pass limit-cycle is an approximation to this bipolar chopper; there, the portion of the modulated signal spectrum not shifted by $\pm \Delta w$ results in a significant spatial error during part of a cycle. FM modulation theory suggests that sinusoidal modulation at $\Delta w$ can also shift a significant portion of the signal spectrum by $\pm \Delta w$. These techniques can track unlimited dynamic range in $\phi'_{10}(x,y,t)$ by modifying the relative phase of their oscillations in each spatial resolution cell. Frequency modulation techniques capable of producing only one side-band of modulation ($+\Delta w$ or $-\Delta w$), such as Bragg cells, rotating gratings, or rotating half-wave plates, should result in superior performance. Unfortunately those techniques are not readily incorporated into a monolithic configuration with a large number of spatial resolution elements.

An alternative approach is to employ two modulators in series. A separate single-channel IPL loop, driven by a discrete detector in the detector plane, could drive a spatially uniform frequency modulator in front of the spatial phase modulator of the IPL system tracking $\phi'_{10}$, (for example the speed of a rotating half-wave plate could be controlled.) The frequency modulator removes spatially-uniform "piston" phase components in $\phi'_{10}$ and the $\Delta w t$ term; thereby greatly diminishing the dynamic range requirements placed on the spatial IPL system. This auxiliary single-channel IPL effectively locks the signal beam to the L.O. frequency. A superior configuration, particularly in adaptive receiver applications, is to place the spatially uniform frequency modulator in the L.O. beam; thereby directly employing the heterodyne IPL to lock the L.O. laser to the transmitter laser wavelength. (Eq. (2.3) shows that the IPL analysis
is unmodified with the modulator in the L.O. beam.) The spatially-variant system then becomes a homodyne IPL, (although it must track spatially-distributed doppler shifts).

In some applications heterodyne operation can offer additional advantages which should not be suppressed. In particular, the heterodyne IPL can increase the effective system gain, \( G_1 \), relative to the offset, \( G_0 \), by reducing the offset to \( G_0 - \Delta w/\lambda_1 \). This can result in diminished phase tracking error, improved immunity to amplitude fluctuations, faster phase acquisition, and larger tracking bandwidth. The additional offset term can also overcome the "dead region" problem which arises with an integrating loop filter. The final decision of where to employ the heterodyne implementation depends on the overall system specifications and constraints of the application.
3.2.2. The phase-contrast IPL

The phase-contrast implementation of the IPL is sketched in Fig. 3.18. As is the usual case for Zernike phase-contrast,\textsuperscript{126-128,206} \( H(u,v) \) is an unresolved (i.e., smaller than Airy disk of lens \( L_1 \)) filter in the center of the Fourier transform plane of lens \( L_1 \). When this filter blocks the center of the transform plane or the filter is a knife edge, this is known as Schlieren phase-contrast.

The phase-contrast approach is appealing because it is potentially simpler to implement, and the incident beam does not have to be nearly as monochromatic as in the homodyne or heterodyne IPL. No stable local oscillator must be maintained; and since lens \( L_1 \) is usually necessary to image the feedback path anyway, the major complexity may be the filter \( h(u,v) \). In some instances, this could be as simple as a tiny aperture stop. A potential disadvantage is less gain than is possible in implementations employing a local oscillator.

3.2.2.1. General Formulation

The field at the exit face of the modulator in Fig. 3.18 is

\[
E_m(x,y,t) = a|E_1(x,y,t)|e^{j(\phi_1(x',y',t)-\phi_m(x',y')-\phi_0(x',y'))}
\]  

(3.99)

Here \( \phi_m \) is the controlled modulator phase, \( \phi_0 \) is the intrinsic modulator phase, and \( a \) is a spatially uniform complex factor, which includes the optical carrier \( e^{-j\omega_0 t} \). In plane \( P_2 \) the Fourier transform of \( E_m \), \( F(u,v) = \mathcal{F}(E_m) \) (times a quadratic phase factor), is multiplied by the filter

\[
H(u,v) = \alpha[1 + (be^{j\phi_h} - 1)P(u,v)]
\]  

(3.100)

where \( P(u,v) \) is an unresolved aperture in the center of the Fourier plane.
\[ E_i = |E_i| e^{i(\phi_i - \omega_it)} \]

\[ E_m = q |E_i| e^{i(\phi_i - \phi_m - \phi_0)} \]

**Spatial Array of Electronic Links**

\[ F(U_1V) = \mathcal{H}(E_m) \]

\[ H(U_1V) = I_0 + I_1 \sin \phi_e \]

\[ i_1(x,y,t) \]

**Fig. 3.18 The phase-contrast IPL.**

(E\textsubscript{m} is defined at the exit face of the modulator)
The effect of the size of \( P(u,v) \) is an important practical aspect of an actual implementation and is discussed in Appendix A.2. There it is shown (based on Eqs. (A.31) with (3.100)) that the field at the detector in plane \( P_2 \), which is the retransformation of \( F(u,v) H(u,v) \), is approximately:

\[
E_d = c_0 [E_m(-x,-y) + A_f A_m(b e^{j\phi h - 1}) a < |E_i| e^{j(\phi_i - \phi_m - \phi_o)} >] \tag{3.101}
\]

The angle brackets denote spatial averaging; \( A_m \) is the area of the exit aperture of the modulator; and \( A_f \) is the area in frequency space (units of meters\(^{-2} \)) of \( P(u,v) \), (the physical area follows from Eq. (A.30) as \((\lambda f)^2 A_f \)). The complex constant \( c_0 \) represents the attenuation and phase shift incurred in passing through the phase-contrast system.

The single lens in Fig. 3.18 performs two Fourier transforms, from \( P_1 \) to \( P_2 \) and \( P_2 \) to \( P_3 \). Appendix A.2 shows that, in this application, this single lens system is equivalent to the more common dual lens convolution. In some cases another lens may be required to reinvert the image. Using Eqs. (A.16) from Appendix A.2:

\[
c_0 = \frac{\alpha}{4\pi^2} \frac{d_1}{d_2} \exp[jk(d_1 + d_2 + \frac{x^2 + y^2}{d_2 - f})] \tag{3.102}
\]

It is convenient to rewrite the spatial average in Eq. (3.101) as:

\[
< |E_i| e^{j(\phi_i - \phi_m - \phi_o)} > \equiv c_0^* c \tag{3.103a}
\]

It is also useful to define

\[
c' \equiv \frac{c''}{< |E_i| >} \leq 1 \tag{3.103b}
\]
The properties of $c'$ and $\phi_c$ are discussed in Appendix A.1; the inequality in Eq. (3.78b) follows from Eq. (A.4). The behavior of $c'$ is further discussed in Appendix A.4, where it is shown that $c'^2$ is the Strehl ratio of the compensated IPL wavefront.

Employing Eqs. (3.100) and (3.103), the detector field (Eq. (3.101)) can be rewritten:

$$E_d = c_o a e^{j\phi_c} (|E_i| e^{j(\phi_i - \phi_m - \phi_0 - \phi_c)} + |E_r| e^{-j(\phi_1 - \pi/2)})$$

(3.104)

where

$$|E_r| = c_o a A_m A_f \sqrt{1 + b^2 - 2b \cos \phi_h}$$

(3.105a)

and

$$\phi_1 = \tan^{-1} \frac{b \cos \phi_h - 1}{b \sin \phi_h}$$

(3.105b)

(Note: $\tan(\theta - \frac{\pi}{2}) = 1/\tan \theta$)

The second term in Eq. (3.104) is an effective interferometric reference wave; and the irradiance at the detector is:

$$I = \frac{1}{2Z_0} |E_d|^2 = I_0 + I_1 \sin \phi_2$$

(3.106)

($Z_0$ is the intrinsic impedance of free space)

Here

$$I_0 = \frac{|c_o a|^2}{2Z_0} (|E_i|^2 + |E_r|^2)$$

(3.107a)

$$I_1 = \frac{|c_o a|^2}{Z_0} |E_i||E_r|$$

(3.107b)

and $\phi_c = \phi_i - \phi_m - \phi_0 - \phi_c + \phi_1 = \phi'' - \phi_m - \phi_c + \phi_1 = \phi_1 - \phi_m$

(3.107c)
Here \( \phi''_i \) lumps the input phase \( \phi_i \) and modulator distortion \( \phi_o \), and \( \phi'_i \) is the input phase with \( \phi_c - \phi_1 \) as its zero reference. Note that \( \phi_1 \), \( b \), and \( \phi_h \) of Eqs.(3.105) are constants which are independent of time and space and specifiable by the system designer.

The basic equations describing the phase contrast IPL intensity, (Eqs.(3.106) and (3.107)), look identical to those for the homodyne IPL (in Eqs. (2.1) and (2.2)). The same basic operating equations, (Eqs. 2.13)), thus also follow, with Eqs.(3.107) used to determine \( G_o \) and \( G_1 \). For example, with the low-pass type-zero filter \( H_w = w_o/(s + w_o) \) in the loop, the system equations very much like those of the low-pass IPL, Eqs. (2.29) and (2.30), discussed previously:

\[
\phi_m = w_o (\phi_m + G_o - G'_1 c' \sin(\phi_m - \phi'_i)) \tag{3.108a}
\]

\[
\dot{\phi}_e = \phi'_i - w_o (\phi_e - \phi'_i + G_o + G'_1 c' \sin \phi_e) \tag{3.108b}
\]

where

\[
G_o = g_o g_1 g_2 \frac{|c_{oa}|^2}{2Z_0} (|E_i|^2 + |E_p|^2 - I'_t) \tag{3.109a}
\]

\[
G_1 = G'_1 c' = g_o g_1 g_2 \frac{|c_{oa}|^2}{2Z_0} \left| \frac{E_i}{c} \right| \left| E_p \right| < |E_i|> c' \tag{3.109b}
\]

\[
\phi''_i = \phi'_i - \phi_c + \phi_1 \tag{3.109c}
\]

\( \phi_c \) and \( c' \) are defined in Eqs.(3.103) and \( \phi''_i = \phi'_i - \phi_o \).

This low-pass loop filter case will be employed to illustrate many of the aspects of phase-contrast IPL behavior in the discussions which follow.

Although Eqs.(3.109) look very similar to those of the homodyne IPL (Eqs. (2.29) and (2.30) discussed previously), there are important differences. The parameters \( c' \) and \( \phi_c \) result in spatial coupling between the resolution elements of the system. If the IPL operates in the usual
manner, $\phi_m = \phi_i'$ is the modulator steady-state; however the parameter $\phi_c$ causes the reference phase ($\phi_o + \phi_c - \phi_1$) to depend on the overall $\phi_i$ distribution. In addition, $c'$ results in gain attenuation which is dependent on the degree of corruption of the input wavefront.

3.2.2.2. Gain degradation due to spatial phase fluctuations

The $c'$ dependence can cause the initial gain to be low when the spatial fluctuations in $\phi_i'$ are large; however as the IPL converges $c'$ approaches unity. The system should be designed with $G_i'$ large enough to operate with the smallest expected value of $c'$. When $\phi_i - \phi_o - \phi_m$ is spatially uniform, $c'=1$ follows from Eqs. (3.103). Appendix D shows that in some instances $c'$ increases from $e^{-\sigma_i^2/2}$ to $e^{-\frac{1}{2}(\frac{\sigma_i^2}{1+G_i'})^2} = 1$ as the IPL converges, where $\sigma_i^2 = <\phi_i'^2> - <\phi_i'^2>$.

It should be noted that $c'$ gain degradation also results in a tradeoff between the maximum spatial corruption and the maximum rate of temporal fluctuations in $\phi_i$ which can be tracked. This arises because, as shown in Eq. (2.35), the maximum trackable rate of $\phi_i$ variation depends on the system gain:

$$\dot{\phi_i} < w_o G_i' c' \phi_i$$

(3.110a)

where $c'$ is a function of the degree of spatial corruption. In some cases $c'$ and hence the bandwidth is explicitly a function of $\sigma$, the rms phase fluctuation: $BW = w_o G_i' c' = w_o G_i' e^{-\sigma_i^2/2}$. The IPL will never converge if a significant portion of the input phase fluctuations have a bandwidth greater than $w_o G_i' \exp(-\sigma_i^2/2)$.

Noise considerations place a more fundamental limitation on IPL performance. Under ideal circumstances, where only signal-induced shot noise is important, it follows from Eqs. (3.76) and (3.104) to (3.107)
(With $\phi_1=0$, $\phi_\text{h}=\pi$, $A_mA_x=1/4$, $c_0=1$, and $c^*=e^{-\sigma^2/2}$) that the signal-to-noise ratio before convergence is

$$
\frac{\text{S/N}}{\text{S/N}} = \frac{nA_r e^{-\sigma^2/2} |E_1| |E_1|<(1+b)}{h \omega |g_o| \delta^2 |E_1|^2 + |E_1|^2 \exp(-\sigma^2/2) (1+b)/4^2} \quad (3.110b)
$$

Using Eq. (3.76c), the approximate average S/N can be written as

$$<\text{S/N}> = 2n_1 e^{-\sigma^2/2}$$

before convergence and $\text{S/N}=1.6n_1$ after convergence.

(Interestingly, the after convergence S/N of the phase-contrast IPL can exceed the S/N of the homodyne IPL which is $\text{S/N}=4n_1 |E_1|/|E_\text{g}|$ for $|E_\text{g}| >> |E_1|$, from Eq. (3.76b). This occurs because, although the L.O. beam enhances the gain of the homodyne IPL, it also contributes to shot noise.)

Even when $\phi_1(x,y)$ is static in time, the discussion of Section 3.1.7 suggests that the IPL has no hope of converging unless $\text{S/N}>0.5$ holds.

Employing the methodology of Section 3.1.9 with $\text{S/N}>0.5$ and $G_m=\sigma_1$

suggests that the average number of photons $<n_\text{w1}>$ collected from $E_1$
in a resolution element during $\omega_1^{-1}$ must exceed $<n_\text{w1}> > \sigma_1 e^{\sigma_1/4}$. After convergence $<n_\text{w1}> > \sigma_1/3.2$ is required. (This is not much less sensitive than the homodyne IPL which requires $n_\text{w1} > \sigma_1/16$, using Eq. (3.87a) with $G_m=\sigma_1$ and $m=0.5$).

The phase-contrast IPL with an input aperture size of a few

inches is expected to converge with $\sigma_1$ produced by atmospheric turbulence,

($\sigma_1$ increases with aperture diameter in turbulence). In low-visibility

applications, however, as scattering becomes more severe eventually

$\sigma_1$ will become large enough to preclude initial convergence. An

investigation of the most severely scattering channel which the phase-

contrast IPL can compensate, based on Eq. (3.110b), is left for future work.

3.2.2.3 Behavior of a single spatial resolution element

The behavior of a single spatial resolution element will be studied
by assuming all-but-one of the resolution elements are turned off or at equilibrium with fixed \( \phi_m \). If the input phase distribution is also static with time, the phase \( \hat{\phi}_m^1 \) in Eqs. (3.108) will be approximately constant in the active channel. Then \( \hat{\phi}_m \) vs \( \phi_m \) or \( \hat{\phi}_e \) vs \( \phi_e \) can be plotted for that channel in state-space plots similar to Figures 2.7 or 2.8 respectively.

To the extent that \( \hat{\phi}_1 \) is static, it can be concluded that after a step of \( \phi_1 \) in one spatial resolution element; that element converges to the usual type-zero IPL equilibrium, with

\[
\phi_m = \phi_1^i - \phi_e \tag{3.111a}
\]

and

\[
\phi_e = \sin^{-1} \left( \frac{\phi_m^i - G_0}{G_1} \right) + 2n\pi \tag{3.111b}
\]

where

\[
n = \text{INT} \left( \frac{\phi_e^i + \pi}{2\pi} \right) \tag{3.111c}
\]

The phase \( \phi_1^i = \phi_i - \phi_o + \phi_1 - \phi_c \) is approximately static, because \( \phi_i \), \( \phi_o \), and \( \phi_1 \) are assumed static with time and the average quantity \( \phi_c \) is only weakly dependent on the active resolution element. For example, with \( N \) spatial resolution cells, the influence of the kth cell is attenuated by \( 1/N \); e.g. (if \( |E_1| \) is uniform and \( \theta_k = \theta_i - \theta_o - \theta_m \))

\[
c_i^e e^{j\phi_c} = \langle e^{j\theta} \rangle = \frac{1}{N} \sum_{j \neq k} e^{j\theta_j} + \frac{e^{j\theta_k}}{N} \tag{3.112}
\]

This argument can also be extended to changes over a few resolution elements, as long as \( \phi_c \) is not perturbed much.

The state-space plot for \( \hat{\phi}_e \) vs \( \phi_e \) (Fig. 2.8) suggests that even if \( \phi_c \) varies, the IPL can still be stable. Once \( \phi_e \) has converged to the vicinity of \( 2n\pi \), \( \phi_c \) fluctuations only cause the curve to move up and down. With large gain, vertical motion does not perturb the \( \phi_e \) equilibrium much.
The stability of the phase contrast IPL, the steady state values of \( \phi_m \) and \( \phi_e \) finally reached, and other aspects of IPL operation are explored more rigorously in the sections which follow.

3.2.2.4. Global behavior

The global behavior of the overall phase-contrast IPL differs significantly from that of small groups of spatial resolution elements.

**Average behavior**

When all the modulator resolution elements are operational and have converged to their steady-state, (e.g., \( \phi_m = 0 \) in Eq.(3.108a) at all spatial locations) the average equilibrium modulator phase with a type-zero loop filter is

\[
<\phi_m> = <G_0> + <G_1 \sin \phi_e>
\]  (3.113)

In order to simplify the average, some general relations will be developed. Using Eq.(3.109b), it is useful to define

\[
G_1'' = g_0 g_1 g_2 |c_o a|^2 |E_1|/(c_0 Z_0),
\]

which is independent of \( |E_1| \); then

\[
G_1 = G_1' c' = G_1'' |E_1| <\left| E_1 \right| c' = G_1'' |E_1| c''
\]  (3.114)

Also using \( \phi_e = \phi''_i - \phi_m - \phi_c + \phi_1 \) from Eq.(3.107 c), and noting that \( \phi_1 \) and \( \phi_c \) are always spatially uniform (see Eqs.(3.103a) and (3.105b)), results in:

\[
<G_1 \sin \phi_e> = <G_1'' c' <\left| E_1 \right| > [<\left| E_1 \right| \sin(\phi''_i - \phi_m) > \cos(\phi_1 - \phi_c) \\
+<\left| E_1 \right| \cos(\phi''_i - \phi_m) > \sin(\phi_1 - \phi_c)]
\]  (3.115)

Eq.(3.103a) and (A.2) result in the additional general identities:
\[<|E_i| \sin(\phi_i^m - \phi_m^e)| = c' \sin \phi_c \] 
\[= c' \cos \phi_c \] 
\[c'' = \sqrt{|E_i| \cos(\phi_i^m - \phi_m^e)|^2 + |E_i| \sin(\phi_i^m - \phi_m^e)|^2} \] 
\[\phi_c = \tan^{-1}\left(\frac{|E_i| \sin(\phi_i^m - \phi_m^e)|}{|E_i| \cos(\phi_i^m - \phi_m^e)|}\right) \]

Eqs. (3.114)-(3.116) produce the extremely useful identity

\[<G_1 \sin \phi_e> = <G_i^0>|E_i|^2 c'^2 \sin \phi_1 = <G_1>c' \sin \phi_1 \] 

No assumptions have been made on the spatial variations of \(|E_i|\) or \(\phi_i^e\); and Eq. (3.117) is an identity, true at all times and for any loop filter \(\hat{H}(s)\), even when the IPL is not in equilibrium. The only significant assumption underlying this result is that \(G_i^0\) is statistically independent of \(|E_i|\) and \(\phi_i^e\); alternatively \(G_i^0\) can be assumed spatially uniform with very little loss of generality.

Employing Eq. (3.117), the average type-zero modulator phase of Eq. (3.113) becomes

\[<\phi_m^e> = <G_0> + <G_1>c' \sin \phi_1 \] 

This result is somewhat paradoxical. Although previous consideration of the state-space plots (Fig. 2.7 and 2.8 from Eqs. (3.108)) suggested that the phase of \(\hat{H}\) individual resolution element of the modulator is an estimate of \(\phi_i^e\) (i.e. \(\phi_m(x,y) = \phi_i^e(x,y) + 2\pi m\)), the average phase estimate, \(<\phi_m^e>\), is independent of the average input phase, \(<\phi_i^e>\). The phase
$\phi_1$ of Eq.(3.105b) is a specified constant of the Zernike filter; and $c'$ is generally dependent on higher order statistics, e.g., $c'=e^{-\sigma^2/2}$ in Appendix D. In the steady-state, for which Eq.(3.118) applies, it can be shown that $c'=1$ usually holds.

In addition, the $n$ value of the multistable $\phi_m$ equilibria can be set in a single spatial element by restarting that element with an appropriate initial value of $\phi_{mo}$; however $\langle \phi_m \rangle$ is not multistable! If all elements have the same $n$ value, it is given by

$$n_0 = \text{INT} \left( \frac{\phi'_1 - \langle G_0 \rangle - \langle G_1 \rangle c' \sin \phi_1 + \pi}{2\pi} \right)$$

(3.119)

even if they were all started at another $n$ value, $n=\text{INT}[\phi'_1 - \phi_{mo} + \pi)/2\pi]$. (The $c'$ dependence of Eq.(3.119) cannot reflect a uniform change in $\phi_{mo}$, since Eqs.(3.103) reveal that adding a uniform phase, such as $2m\pi$, to $\phi_m$ does not modify $c'$.)

Equation (3.117) also implies that it is not possible to have a small error, $\phi_e=0$, unless $\phi_1=0$. According to Eq. (3.111b), a small error also requires $\phi_m=\langle G_0 \rangle$ which implies $n=\text{INT}(\phi'_1 - \langle G_0 \rangle + \pi)/2\pi]$.

**Example:** Spatially uniform input phase

One of the most pathological cases is spatially uniform input phase, $\phi_i''$; using Eqs.(3.103) and (3.118), the phase in every resolution cell is

$$\phi_m^{ss} = \langle \phi_m \rangle = \langle G_0 \rangle + \langle G_1 \rangle \sin \phi_1$$

(3.120)

and $c'=1$. Since $\phi_i''-\phi_m$ is spatially uniform, Eq.(3.116d) reveals that

$$\phi_c = \phi_i'' - \phi_m$$

(3.121a)

and hence

$$\phi_e = \phi_i'' - \phi_m - \phi_c + \phi_1 = \phi_1$$

(3.121b)
Note that the modulator phase is independent of the input phase; and the error phase does not approach 2nπ, but rather φ₁. In fact φₘ and φₑ can be adjusted arbitrarily over a large range by the choice of φ₁ in the Zernike filter. The phase-contrast IPL cannot estimate a uniform input phase because, as Eqs.(3.121) demonstrate, the φᵽ term of the reference phase cancels φ₁ and φₘ. Since φᵽ-φₘ is spatially uniform, the system does successfully, albeit trivially, perform phase compensation. With uniform φᵽ, G₀, and G₁, the dynamic equation for φₘ in Eq. (3.108) becomes simply

\[ \dot{φₘ} + \lambda φₘ = φₘ \]

where φₘ is given by Eq. (3.120).

General qualitative global behavior

The more general case of a spatially varying but temporally constant input phase is more difficult to explain. If all the resolution cells are started at initial value φₘ₀ and then turned on one by one, Eq.(3.108a) and the state-space plot (φₘ vs φₘ) of Fig. 2.7 shows that each cell will turn on with an initial modulator phase within ±π of φₘ₀. (If this is not clear, refer to Section 3.1.2). However when all the elements have been turned on, Eqs.(3.108) and Fig. 2.7 must somehow result in a φₘ distribution with mean \( <φₘ> = <G₀> + G₁ c' \sin φ₁ \); although φₘ₀ and \( <φₘ> \) may differ by many cycles, i.e., φₘ₀ = \( <φₘ> + 2mπ \). (The changes in c' as the system converges only affect the gain which has little bearing on the equilibrium value of φₘ in Fig. 2.7.)

The drift to \( <φₘ> \) can be understood by examining the case where the elements are turned on one by one. As each element converges φₑ is slightly changed, perturbing \( φᵽ = φᵽ - φₑ + φ₁ \) to drift toward \( <φₘ> \). As \( φᵽ \) drifts, the whole trajectory of Fig. 2.7 shifts laterally and carries the equilibrium values of all the previously turned on elements toward \( <φₘ> \).
By the time the last element is turned on $\phi_i$ has drifted far enough, possibly over many cycles, to result in the proper value of $\langle \phi_m \rangle$.

These notions can be made more precise by numerically solving or simulating the behavior of the nonlinear, multistable, spatially-coupled, transcendental, differential equation (3.108a) together with Eqs. (3.109c), (3.116c), and (3.116d). Alternatively, assumptions can be introduced which simplify the system enough to obtain an analytic solution. For example, if the system can be broken into N resolution cells, eqs. (3.108a) (3.116c), and (3.116c) (with eq. (3.109c)) constitute a set of N+2 coupled equations in the N+2 variables $c'$, $\phi_c'$, and N values of $\phi_m$. When Eqs. (3.116c), (3.116d), and the steady-state relations (3.111) are all solved together, the solution does in fact satisfy the $\langle \phi_m \rangle$ condition in Eq. (3.118)! Both numeric simulations and approximate analytic solutions are employed in the discussions which follow. 3.2.2.5. Simulated behavior

A variant to numerically solving Eq.(3.108a) over the whole modulator area, is to repeatedly iterate over the array, turning on one resolution cell at a time. The steady-state value of $\phi_m$ in each resolution cell becomes its initial value for the next iteration over the array. A computer program took initial vectors of $\phi_i$, $\phi_o$, $\phi_{m0}$, $G_o$, and $G_i$ for N resolution cells and employed (Eqs.(3.116c),(3.116d),(3.111a),(3.111b), and (3.111c) to compute the evolution of $c'$, $\phi_c$ $\phi_i$ and $\phi_m$ as modulator elements were successively turned on. Summations similar to that in Eq.(3.112) were used to compute the averages in Eqs. (3.116).

Fig. 3.19a shows the evolution of $\phi_m$ in each resolution cell, $\phi_c$, and $c'$ for a three spatial element IPL. The final value of $\langle \phi_m \rangle$ satisfied Eq.(3.118), drifting from $2\pi$ to $\langle \phi_m \rangle = 1.8 + 7.3\pi \sin(\pi/8) = 4.59\pi$. In
Fig. 3.19 Simulated behavior of the phase-contrast IPL. Calculated evolution of $\phi_m$, $\phi_c$, and $c'$ for a 3 spatial element IPL, with $\phi_i=(\pi/6, \pi/3, \pi/2)$, $\phi_i=0.125\pi$, and spatially uniform $G_i=1.8\pi$ and $G_i=7.3\pi$. Each small unit on the horizontal axis corresponds to one iteration over all three modulator elements.
Fig. 3.19b, the initial values of \( \phi_m \) were separated from each other by 2\( \pi \), the final values satisfied Eq. (3.118). The initial and final states of four trials, including the two graphed above are summarized in Table 3.3.

Aside from \( \langle \phi_m \rangle \) satisfying Eq. (3.118), \( \phi_m \) in each resolution cell was also an estimate of \( \phi_1' \), in accordance with the steady-state relations in Eqs. (3.111). The error could be large (\( \phi_e = \phi_1' \)), but was approximately spatially constant when each modulator element had the same value of \( \varepsilon = \text{INT}(0.5(\phi_m + \pi)/\pi) \). Thus, the phase-contrast IPL can indeed estimate and compensate phase, but \( \phi_m \) is constrained by Eqs. (3.116). The reference phase of the estimate, \( \phi_0 + \phi_c - \phi_1' + \phi_e \), is a function of the overall input phase distribution; and the compensated phase, \( \phi_c - \phi_1' + \phi_e \), can be made spatially uniform.

The approach of stepping through the elements brings \( c' \) to unity rapidly. Since \( \phi_c \) does not drift much during a cycle, all the elements are essentially compensating after the first cycle. Examination of Table 3.3 reveals the approximate numeric relations \( \phi_c = \langle \phi_1' \rangle - \langle \phi_m \rangle \) and \( \langle \phi_e \rangle = \phi_1' \).

3.2.2.6. Simplifying assumptions and approximations

With the introduction of various assumptions, explicit expressions can be obtained to describe the behavior of the phase-contrast IPL.

When \( \phi_e \) is small and \( \phi_1 \) of Eq. (3.105b) was designed to be small, the general identity of Eq. (3.117) can be linearized, resulting in

\[
\langle G_1 \phi_e \rangle = \langle G_1 \rangle c' \phi_1
\]
Table 3.3 Computer simulation of phase-contrast IPL

<table>
<thead>
<tr>
<th>Trial</th>
<th>Resolution cell</th>
<th>System Parameters</th>
<th>Initial Conditions</th>
<th>Final Values</th>
<th>Comp. Phase &lt;φ²_e-φ_m&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>G_o G_i</td>
<td>θ_i θ_i0 φ_c c'</td>
<td>φ_m φ_c c' φ²_e-φ_m</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>all</td>
<td>3.18x 7.3x 0.125x</td>
<td>-1.67x 0.91x 2x</td>
<td>-4.26x 0.9998 4.59x 4.59x 0.125x</td>
<td>.26x</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n/6</td>
<td>2x</td>
<td>4.43x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>n/3</td>
<td>2x</td>
<td>4.59x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>n/2</td>
<td>2x</td>
<td>4.75x</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>all</td>
<td>3.18x 7.3x 0.125x</td>
<td>-1.67x 0.91x 2x</td>
<td>-4.16x 0.964 4.49x 4.49x 0.129x</td>
<td>.26x</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n/6</td>
<td>0x</td>
<td>2.43x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>n/3</td>
<td>2x</td>
<td>2.50x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>n/2</td>
<td>4x</td>
<td>6.55x</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>all</td>
<td>3.18x 7.3x -0.01x</td>
<td>-1.67x 0.91x 2x</td>
<td>-1.24x 0.999 1.57x 1.17x -0.01x</td>
<td>.76x</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n/6</td>
<td>2x</td>
<td>1.41x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>n/3</td>
<td>2x</td>
<td>1.57x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>n/2</td>
<td>2x</td>
<td>1.73x</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>all</td>
<td>9.18x 10x 0.01x</td>
<td>-2.02x 0.296x 2x</td>
<td>-2.13x 0.9998 2.11x 2.11x 0.01x</td>
<td>-.15x</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.67x</td>
<td>2x</td>
<td>1.48x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.56x</td>
<td>2x</td>
<td>1.59x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.38x</td>
<td>2x</td>
<td>1.76x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.13x</td>
<td>2x</td>
<td>2.00x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.03x</td>
<td>2x</td>
<td>2.15x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.18x</td>
<td>2x</td>
<td>2.30x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.29x</td>
<td>2x</td>
<td>2.41x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.50x</td>
<td>2x</td>
<td>2.61x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.63x</td>
<td>2x</td>
<td>2.74x</td>
<td></td>
</tr>
</tbody>
</table>

In these four trials G_o and G_i are spatially uniform. N is the number of resolution cells.
φ²_e = φ_c + φ_i - φ_m. φ_m = G_o G_i sin_i. φ²_e - φ_m is the "compensated" phase.
Using \( \phi_e = \phi'' - \phi_m + \phi_1 \) from Eq. (3.107c), with the above relation, results in an approximate expression for \( \phi_c \)

\[
\phi_c = \frac{<G_1 \phi''_i> - <G_1 \phi_m>}{<G_1>} + (1-c')\phi_1
\]  
(3.122)

When \( c' = 1 \), which is usually true near equilibrium and \( G_1 \) is statistically independent of \( \phi_i'' \) and \( \phi_m \); Eq. (3.122) simplifies to

\[
\phi_c = <\phi_i''> - <\phi_m>
\]  
(3.123a)

Also, averaging Eq. (3.107c) results in \( <\phi_e> = <\phi_i'' - \phi_m> - c' \phi_1 \)  
(3.123b)

and hence:

\[
<\phi_e> = \phi_1
\]  
(3.123c)

These relations, Eqs. (3.123) can be very general and were observed in the simulation run results of Table 3.1. For example, if \( \theta = \phi_i'' - \phi_m \) is a (1) temporally deterministic, but (2) spatially ergodic random process, (3) characterized by a probability distribution, \( P(\theta) \), which is symmetric about \( <\theta> \); and (4) \( E_i \) and \( \theta \) are statistically independent; it follows from Appendix A that

\[
\phi_c = <\theta> = <\phi_i''> - <\phi_m>
\]

When \( G_0 \) and \( G_1 \) are approximately spatially uniform and \( \phi_i'' \) satisfies the four conditions outlined above, Appendix D shows that \( \theta \) also satisfies those conditions. In particular, when \( \phi_i'' \) is Gaussian distributed, Appendix D demonstrates (as mentioned previously) that as phase lock is acquired the gain with a type-zero loop filter increases from \( G_1 = G_i e^{-\sigma_i^2/2} \) to

\[
G_1 = G_i e^{-\sigma_i^2} = G_i \exp(-\frac{1}{2} (\frac{\sigma_i^2}{1+G_1})^2)
\]
Here $\sigma^2_i = \langle \phi^2_i \rangle - \langle \phi_i \rangle^2$ and $\sigma^2_e = \langle \phi^2_e \rangle - \langle \phi_e \rangle^2$.

The previously discussed independence of $\langle \phi_e \rangle$ and $\langle \phi_m \rangle$ on $\phi'_i$ can be understood from Eqs. (3.123), which show that on the average $\phi_c$ cancels $\phi'_i$ and $\phi_m$. It should be noted that $\langle \phi_e \rangle \rightarrow 0$ for $\phi_1 \rightarrow 0$ is consistent with $\phi_e = \sin^{-1}(\phi_m - G_0)/G_1$, since according to Eq. (3.118), $\langle \phi_m \rangle \rightarrow G_0$.

3.2.2.7. Behavior with a low-pass loop filter

When the simplifications $\phi_c = \langle \phi'_i - \phi_m \rangle$ and $c' = 1$ are valid, the differential equation, Eq. (3.108a), for the phase-contrast IPL becomes:

$$\dot{\phi}_m = w_o (-\phi_m + G_o + G_1 \sin(\phi_i^m - \phi_m - \phi'_i + \phi_m + \phi_1))$$

(3.124)

$(\phi'_i = \phi'_i - \phi_c + \phi_1)$ from Eq. (3.109c) was also employed.

The $\langle \phi_m \rangle$ term causes this to be a system of spatially coupled nonlinear differential equations. One solution approach is to treat only the case of small $\phi_e$, where Eq. (3.124) can be linearized. Then if $N$ discrete spatial resolution cells are also assumed, $\langle \phi_m \rangle$ becomes a summation. The linearized form of Eq. (3.124) can then be recast as a vector differential equation and solved by standard matrix techniques of modern linear systems theory. A variant of this approach, in the Laplace domain, is discussed in Section 3.2.2.8 and Appendix A.4.

There is a much more elegant approach which does not assume small error or discrete resolution cells. Employing the general relation of Eq. (3.118) for $\langle \phi_m \rangle$ automatically decouples Eq. (3.124)! The independent differential equation at each spatial location becomes

$$\phi'_m = w_o (-\phi_m + G_o + G_1 \sin(\phi_i^m - \phi_m - \phi'_i + \phi_1 + G_o + G_1 \sin \phi_1))$$

(3.125)
This is essentially identical to the extensively discussed homodyne equation of Eq. (2.29). A state-space plot similar to Fig. 2.7 reveals that stable equilibria occur at

\[
\phi_m = \phi_1'' + <G_o> + <G_1> \sin \phi_1 + \phi_1 - \sin^{-1} \left( \frac{\phi_m - G_o}{G_1} \right)\ - \ 2n\pi
\]

(3.126)

This result could have been obtained directly from the general type-zero loop filter steady-state of Eqs. (3.111) with the use of Eq. (3.118).

Eq. (3.126) can be solved for \( \phi_m \) vs \( \phi_1 \) by graphical techniques such as Figure 2.4, but care must be taken to resolve the multistability of \( \phi_m \) to be consistent with \( <\phi_m> \) of Eq. (3.118). For example if \( n \) has the same value at all locations, using Eq. (3.111c) and \( <\phi_e> = \phi_1 \), this value is \( n = <n> = \text{INT}((\phi_1 + \pi)/2\pi) \). Generally \( |\phi_1| < \frac{\pi}{2} \), from Eq. (3.105b) and overall stability considerations, and \( <n> = 0 \).

When \( \phi_e \) is small, which according to Eq. (3.123) requires that \( \phi_1 \) be small, Eq. (3.126) can be linearized to obtain

\[
\phi_m = \frac{G_1}{G_1 + 1} \left( \phi_1'' + <\phi_1''> + <G_1 + 1> \phi_1 + \frac{G_o}{G_1} + <G_o> - 2n\pi \right)
\]

(3.127)

With large gain, \( G_1 \gg G_o \), the modulator phase becomes

\[
\phi_m = (\phi_1'' + <\phi_1''> - 2n\pi) + (<G_o> + <G_1> \phi_1)
\]

(3.128)

Equations (3.126)-(3.128) summarize the basic behavior of the phase-contrast IPL with a type-zero loop filter. These results show that the phase-contrast IPL does in fact estimate \( \phi_1'' = \phi_1 + \phi_o \). From Eqs. (3.128) and (3.117), the reference phase of this estimate is approximately
\[ \phi_b = \langle \phi_i'' \rangle - \langle G_0 \rangle - \langle G_1 \rangle - \langle |E_1| \rangle^2 c^2 \sin \phi_1 \]

(\(\phi_m = 0\) when \(\phi_i'' = \phi_b\)).

Since \(\phi_1\) and all spatial averages are spatial constants, this is a valid \textit{spatially uniform} reference phase. However \(\phi_b\) does vary in time with \(\langle \phi_i \rangle\) and \(\langle |E_1| \rangle\), and hence with the overall spatial distribution of the input field. One consequence of this is that \(\phi_m\) cannot track spatially uniform ("piston") phase changes in \(\phi_i''\). Like the homodyne IPL, the phase-contrast IPL does not distinguish the modulator aberration \(\phi_0(x,y)\) from \(\phi_1\); a uniform modulator is required in phase estimation applications.

In adaptive phase compensation applications, the compensated phase is \(\phi_b\) of Eq. (3.129). Since \(\phi_b\) is a spatial constant, the compensated phase-front is a plane wave, which is desirable for most applications. The phase-contrast IPL also corrects the intrinsic modulator aberration \(\phi_0(x,y)\), as evidenced by the fact that \(\phi_b\) is independent of \(\phi_0\). The fluctuations of \(\phi_b\) with \(\langle \phi_i \rangle\) and \(\langle |E_1| \rangle\) cause the overall "piston phase" of the compensated plane wave to vary with time. This phase is not seen in a direct detection communications receiver.

The \(\phi_i'' - \phi_i''\) component of \(\phi_m\) in Eq. (3.128) tends to keep \(\phi_m\) near \(\langle G_0 \rangle + \langle G_1 \rangle \phi_1\), (or more generally \(\langle G_0 \rangle + \langle G_1 \rangle c \sin \phi_1\)), even as \(\phi_i''\) varies over many cycles. The convergence of \(\phi_m\) to \(\langle G_0 \rangle + \langle G_1 \rangle \phi_1\) has the positive effect of restricting the 2\(\pi\) multistability of \(\phi_m\) (e.g. \(<n>=0\)), and preventing \(\phi_m\) from reaching the physical limits of the modulator as frequently as it would in the independent loop implementations, (e.g., homodyne IPL). With \(\phi_1\) designed to be zero, \(\phi_m\) will be driven toward \(\langle G_0 \rangle\), which makes the IPL "self-centering" to its minimum-error operating
point, since $\phi_e = \sin^{-1}\left(\left(\phi_m - G_0\right)/G_1\right)$. (Also recall from Eq. (3.123b) that when $\phi_1 = 0$, $\langle \phi_e \rangle = 0$). The value of $\langle G_0 \rangle$ can be controlled with $I_0'$ of Eq. (2.14a).

If $\phi_1$ is too large a modulator with an extremely large dynamic range, e.g., $\langle G_0 + G_1 \rangle$, may be required, where most of that range is rarely used. Large $\phi_1$ also accentuates the $\langle |E_i| \rangle^2$ dependence (see Eq. (3.117)) of the $\langle G_1 \rangle \cdot \sin \phi_1$ term. Fortunately this term is spatially uniform and hence does not destroy the "immunity" to amplitude variations of the relative phase estimate.

The above discussions suggest that there are advantages to specifying $\phi_1 = 0$ in the design of an actual phase-contrast IPL. According to Eq. (3.105b), this can be accomplished by employing a Zernike filter with either $b \cdot \cos \phi_n = 1$ or $\phi_n = \pi$. In the former case $b > 1$ is achievable by attenuating the region around the center spot of the Zernike filter; (i.e., when $\alpha < 1$, $0 < b - 1/\alpha$ can exceed unity). However, this filter is difficult to fabricate and wasteful of intensity, with an effective reference beam amplitude of (from Eq. (3.105a)):

$$|E_r| = c'A_m A_f \sqrt{1+b^2 -2 \langle |E_i| \rangle} \quad (3.130a)$$

In the $\phi_n = \pi$ case, the reference beam amplitude is

$$|E_r| = c'A_m A_f (1+b) \langle |E_i| \rangle \quad (3.130b)$$

The latter case is more efficient, being able to function with no Zernike filter attenuation ($b=1$, $\alpha=1$), and providing larger $|E_r|$ and $G_1$ (Eq. (3.109b)) with a given $b$. Some comments on the construction of such filters are given in Appendix A.3.
Eqs. (3.130) and (3.109b) show that an additional feature of the Zernike phase-contrast IPL is an automatic gain control (agc) behavior in which $|E_r|$ tracks $|E_i|$. This tends to decrease $\phi_e$ and improve immunity to amplitude fluctuations. As will be discussed in Section 4.3.1, this behavior is particularly advantageous in systems with no intensity threshold. (Those systems perform best when $|E_r|=|E_i|$; on the average, Eq. (3.105a) becomes $|E_r|\approx |E_i|$ when $b=(1-A_m A_f)/A_m A_f$ and $\phi_h=\pi$). The overall effect of $|E_i|$ variations on IPL performance can be evaluated with technique similar to those employed in Section 3.1.1, but with $|E_r|$ replacing $|E_i|$.

3.2.2.8. Behavior with other loop filters

With small $\phi_1$ and small $\phi_e$, the general IPL dynamic equation (2.12) can be linearized. Also assuming that $G_0$, $G_1$, and $c'$ are approximately constant in time; the Laplace transform of the linearized dynamic equation can be taken, resulting in:

$$\dot{\phi}_m = \hat{H} \left( \frac{G_0}{s} + G_1 (\hat{\phi}_1^* + \hat{\phi}_m - \hat{\phi}_1^* \right)$$

(3.131)

This is a general transfer function for perturbations from an equilibrium with small $\phi_e$. Eq. (3.131) can be rewritten:

$$\dot{\phi}_m = \hat{\phi} \left( \hat{\phi}_1^* + \hat{\phi}_e + \hat{\phi}_1^* \right)$$

(3.132)

where

$$\hat{\phi} = \frac{\hat{H} G_1}{1 + \hat{H} G_1} = \text{closed-loop transfer function}$$

(3.133a)

$$\hat{\phi} = \frac{G_0}{s G_1}$$

(3.133b)
and
\[ \tilde{\phi}_1 = \frac{\phi_1}{s} \quad (3.133c) \]

Since \( \phi_e \) and \( \phi_1 \) are assumed small, using the general relation (3.122) for \( \phi_c \) introduces no new assumptions. In particular the spatial distributions of \( \phi_i^\nu \), \( |E_i| \), and \( \phi_m \) are unrestricted. With (3.122), Eq. (3.132) becomes:
\[ \tilde{\phi}_m = \tilde{\rho} (\tilde{e} + \tilde{\phi}_1) - \frac{\langle G_1 \phi^\nu \rangle}{\langle G_1 \rangle} + c \tilde{\phi}_1 - \frac{2\pi n}{\langle G_1 \rangle} + \frac{\langle G_1 \phi_m^\nu \rangle}{\langle G_1 \rangle} (3.134) \]

The last term of this equation results in spatial cross-coupling.

If the IPL is treated as consisting of \( N \) spatial resolution elements, the spatial averages can be expressed as summations over resolution cells. Using lower subscripts to index resolution elements, Eq. (3.134) can be written as one of \( N \) equations in \( N \) variables:
\[ \tilde{\phi}_m = \rho_k (\tilde{e}_k + \tilde{\phi}_1) - \frac{\langle G_1 \phi^\nu_i \rangle}{\langle G_1 \rangle} + c \tilde{\phi}_1 - \frac{1}{N \langle G_1 \rangle} \sum_{j=1}^{N} G_{1j} \phi_m^\nu_j - \frac{2\pi n_k}{s} (3.135) \]

This system of equations is solved by matrix techniques in Appendix C to obtain an uncoupled equation for each \( \phi_m^\nu_k \). However, a more elegant approach to the same solution, which does not require assuming discrete resolution cells, will be presented here.

Multiplying both sides of the continuous equation (3.134) by \( G_1 \) and taking the spatial average result in an expression for \( \langle G_1 \phi_m^\nu \rangle \):
\[ \langle G_1 \phi_m^\nu \rangle = \frac{\langle G_1 \rangle}{(1 - \tilde{\rho} G_1)} [\langle G_1 \tilde{e} \rangle + \langle G_1 \phi_1^\nu \rangle - \frac{\langle G_1 \phi_1^\nu \rangle}{\langle G_1 \rangle} + \langle G_1 \phi_m^\nu \rangle + \langle F G_1 \rangle c \tilde{\phi}_1] \quad (3.136) \]

Substitution of Eq. (3.136) into Eq. (3.134) results in a spatially decoupled expression for \( \phi_m^\nu \). After some manipulation and remembering that
the average of a sum is the sum of the averages, one obtains the general
Laplace domain response near equilibrium:

\[
\hat{\phi}_m = \hat{F}(\hat{\phi}_1 + \hat{\phi}_1') + \frac{<G_0 \hat{\phi}_1 + <G_1 \hat{\phi}_1> + <G_1 \hat{\phi}_1> c \phi_1>}{<(1-\hat{F})G_1>} - \frac{2\pi}{s}
\]  (3.137)

This is identical to the discrete result, Eq. (A.51), obtained in
Appendix A.4 by solving the system of coupled equations.

The linearized Laplace domain response of Eq. (3.137) provides
important insights into the behavior of the phase-contrast IPL. In almost
all instances the open-loop filter \(\hat{A}\) is a system parameter, statistically
independent of \(|E_1|\) and \(\phi_1\) are also independent, Eq. (3.137)
becomes

\[
\hat{\phi}_m = \hat{F}(\hat{\phi}_1 + \hat{\phi}_1') + \hat{\phi}_1 + \frac{<G_0 \hat{\phi}_1 + <G_1 \hat{\phi}_1> c \phi_1>}{<(1-\hat{F})G_1>} - \frac{2\pi}{s}
\]  (3.138)

Note that the only spatially variant terms on the right are \(\hat{\phi}_1\), \(\hat{\phi}_1\'), and
\(n\). When \(\hat{A}\), \(G_0\), and \(G_1\) are spatially uniform:

\[
\hat{\phi}_m = \hat{F}(\hat{\phi}_1 + \hat{\phi}_1') - \frac{2\pi}{s} + \frac{\hat{A}}{s} (G_0 + G_1 c \phi_1)
\]  (3.139)

Equation (3.137) - (3.139) reveal that, much like the spatially
decoupled implementations of the IPL, the phase-contrast IPL estimates
\(\phi_1\) with the closed-loop transfer function \(\hat{F} = \hat{A}G_1/(1+\hat{A}G_1)\). However unlike
other implementations, the modulator phase in the phase-contrast IPL
also converges to the vicinity of \(<\phi_m = <G_0> + <G_1> c \phi_1>\) with a response
characterized by the open-loop transfer function \(\hat{A}\).

Note that, with a low-pass loop filter the bandwidth of \(\hat{F}, ((w_c+1)G_1),\)
is generally greater than that of \(\hat{A}, (w_0)\). This implies behavior similar
to the simulation of Fig. 3.19; each resolution element rapidly converges
to $\phi_m = \phi_i'$, which results in a slower drift of $<\phi_m>$ to $<G_0> + c'sin\phi_1$.

When $\hat{H}$ is a normalized type-zero transfer function, i.e., $\hat{H}(s-o) = 1$, the steady-state response of Eq. (3.137) to a step in $\phi_i'$ is

$$
\phi_m = \lim_{s \to 0} S_{\phi_m}(s) = \frac{G_1}{1+G_1} \left[ \phi_i' + \frac{G_0}{G_1} \langle \phi_i'' \rangle + \frac{G_1 G_0}{1+G_1} \right] - 2n\pi
$$

(3.140)

As one would expect this is essentially the same as the low-pass result of Eqs. (3.126) - (3.128). Equation (3.140) is more general; the use of Eq. (3.123a) to derive Eq. (3.125) assumed that $|E_i|$ and $\phi_i$ are statistically independent and that $\phi_i$ has certain additional statistical properties.

3.2.2.9. Stability

The discussion of Sections 3.2.2.3 and 3.2.2.5 has shown that the phase-contrast IPL with a low-pass filter can be brought to the steady-state of Eqs. (3.111). However with all the elements operating, it is conceivable that spatial coupling may introduce local instabilities or undamped cooperative modes.

If the IPL is brought to the vicinity of an equilibrium at all spatial locations, for example by turning on one element at a time; a perturbation analysis can be carried out about this equilibrium. If the system always returns to the assumed equilibrium after a perturbation, this is a self-consistent proof that the equilibrium can exist and is stable.
For example, with a type-zero loop filter and large gain, the equilibrium of Eq. (3.111) results in spatially uniform \( \phi_i' - \phi_m' \) and hence \( c' = 1 \). Also assuming that \( |E_i| \) and \( \phi_i \) are statistically independent and that \( \phi_i \) satisfies the assumptions of Appendix D (or that \( \phi_e \) is small), allows Eq. (3.125) to be used to describe the low-pass loop-filter dynamic response of a single spatial location to a perturbation from equilibrium. Eq. (3.125) is essentially identical to the stable homodyne result of Eq. (2.29). Since there is no spatial coupling, there is no reason to expect the equilibria to be unstable or develop into cooperative modes.

More generally, Eq. (3.137) describes the return to equilibrium in the case of small \( \phi_e \) (and small \( \phi_i \)). Equation (3.137) is uncoupled, and Appendix A.5 shows that it is stable if the open-loop transfer function \( \hat{H}_K \) and closed-loop transfer function \( \hat{P}_K \) are both stable in every resolution cell (for the current values of gain, \( G_1 \)). By way of comparison, the homodyne IPL only demands that \( \hat{P} \) be stable. The dependence on both \( \hat{H} \) and \( \hat{P} \) is especially evident in Eq. (3.139); where phase tracking occurs with \( \hat{P} \), and the convergence to \( G_0 + G_1 c' \phi_i \) occurs with \( \hat{H} \). The direct dependence on the poles of the open-loop transfer function effectively limits \( \hat{H} \) to type-zero transfer functions. Otherwise the steady-state of \( \phi_m' \) (\( \lim_{s \to 0} s \hat{\phi}(s) \)), is unbounded.

The stability of \( <\phi_m> \) is also easy to investigate. Applying the general identity (3.117) to the spatial average of the IPL dynamic equation (2.12) removes the nonlinear term from the average response. Assuming only that \( c' \) is approximately constant with time, allows the Laplace transform of the average response to be taken; resulting in
\[ \langle \phi_m \rangle = \langle \hat{H}(s) \rangle \frac{1}{s} (\langle G_0 \rangle + \langle G_1 \rangle c' \sin \phi_1) \] (3.141)

This is a surprisingly lenient condition for the stability of \( \langle \phi_m \rangle \); requiring only that the average open-loop rather than closed-loop transfer function be stable. Following the methodology of Eq. (A.57) in Appendix A.5, it is easy to show that \( \langle \hat{H} \rangle \) is stable if \( \hat{H}_K \) is stable for all \( K \). The constraint that \( c' \) be constant could be met if \( \langle \phi_m \rangle \) is responding to a perturbation from equilibrium. In the low-pass case, the corresponding time domain equation is:

\[ \dot{\phi}_m = -w_0 (\langle -\phi_m \rangle + \langle G_0 \rangle + \langle G_1 \rangle c' \sin \phi) \]

Of course \( \dot{\phi}_m = 0 \) does not rule out cooperative modes or even assure that \( \phi_m \) is stable, unless additional information such as \( \phi_m > 0 \) is available.

3.2.2.10 Spectral Bandwidth

As mentioned previously, the phase-contrast IPL can operate with polychromatic light. This feature may potentially extend IPL utility to such applications as astronomical image compensation.13

With polychromatic light, the field in the input aperture can be written as a frequency superposition of monochromatic complex phasor components.
Here \( E \) and \( E_w \) are a Fourier transform pair, and the actual physical field is the real part of \( E \). (Note that \( w \) in Eq. (3.143) has the opposite sign from the usual inverse Fourier transform). Transmission through a spatial phase modulating layer, such as the IPL phase modulator, introduces a time delay into \( E \), resulting in the modulated field \( E(x,y,t-T(x,y)) \). By the time-delay theorem of Fourier analysis, the spectrum of this delayed field becomes \( E_w(x,y,w)e^{j\omega T(x,y)} \). The phase \( \phi = wT \) can also be expressed in terms of the Optical Path Distance (OPD), \( D = nL \), where \( L \) is the physical thickness and \( n \) is the refractive index of the phase modulator; then

\[
\phi = wT = \frac{w}{c}D = \frac{2\pi}{\lambda}D
\]  

(3.144)

(In this section, \( \lambda \) is used exclusively to represent optical wavelength).
Assuming that the system is time invariant over many cycles of \( w \) and linear, the previously derived monochromatic results can be linearly superimposed through Eq. (3.143). Using OPD instead of phase, a spectral component of the field at the IPL detector follows from Eqs. (3.101) through (3.105) as

\[
E_{wd}(x,y,w) = c_{o}(w) \left[ |E_{wi}(x,y,w)| \exp\left( j \frac{w}{c} (D_{i} - D_{m} - D_{o}) \right) e^{j\phi_{2}} \right.
\]

\[
+ |E_{wr}(w)| \exp\left( - j \frac{w}{c} (D_{1} - D_{c} - \frac{\pi}{2}) \right) \]  \hspace{1cm} (3.145a)

where

\[
|E_{wr}| = \frac{w^{2}}{4\pi^{2}} c_{w}(w) \frac{A_{m} A_{h}}{\sqrt{1 + b^{2} - 2b \cos \left( \frac{N_{c} D_{n}}{c} \right)}} \]  \hspace{1cm} (3.145b)

and

\[
c_{w}(w)e^{j\omega D_{c}/c} = \langle |E_{wi}| \exp(j \frac{w}{c} (D_{i} - D_{m} - D_{o}))e^{j\phi_{2}} \rangle \]  \hspace{1cm} (3.145c)

Here: the angle brackets represent spatial averaging, \( A_{h} \) is the physical area (in \( m^{2} \)) of the central part of the Zernike filter, and \( \phi_{2}(x,y,w) \) is the phase of \( E_{wi}(x,y,w) \) due to spatially-variant temporal fluctuations in \( E_{i}(x,y,t) \). The \( D_{1} \) and \( D_{c} \) OPDs are not functions of \( x \) and \( y \), but may depend on \( w \). Assuming that all the media are nondispersive (i.e. \( n \) is not a function of \( w \)) and that the OPDs vary only very slowly with time (relative to \( 2\pi/w \)),
\( D_i, D_m, \) and \( D_0 \) are then only functions of \( x \) and \( y \) but not \( w \).

Spatial light modulators which employ a deformable mirror\(^{50,71-73}\) to control \( D_0 + D_m \) satisfy the nondispersive assumption, however most electro-optic phase modulators do not.

The instantaneous intensity seen at the detector is

\[
I_d(x,y,t) = |E_d(x,y,t)|^2,
\]

where the ensemble average (denoted by an overbar) has been taken, since \( E_d \) is generally a stochastic process in time and space. In terms of the frequency spectrum, \( E_{wd} \),

\[
\bar{I}_d = \int \int \int \int dw_1 dw_2 E_{wd}(w_1) E_{wd}(w_2) e^{-j(w_1-w_2)t}
\]  

(3.146a)

Assuming that the detector is preceded by a spectral filter with field transmission \( B_\lambda(w) \), which includes the spectral sensitivity of the detector itself, the filtered intensity is

\[
\bar{I}_d = \int dw_d e^{jw_d t} \int B_\lambda(w_c + \frac{w_d}{2}) B_\lambda^*(w_c - \frac{w_d}{2}) \Gamma_0(w_d, w_c)
\]

(3.146b)

Here: \( w_d = w_1 - w_2, \) \( w_c = (w_1 + w_2)/2 \), and \( \Gamma_0(w_d, w_c) = \frac{E_{wd}(w_1) E_{wd}^*(w_2)}{E_{wd}(w_1) E_{wd}^*(w_2)} \) is the two-frequency mutual coherence function\(^{99,101}\) of \( E_{wd} \) evaluated at the single point \( x,y \).

The output current from the detector is given by the convolution of \( \bar{I}_d \) with the detector impulse response \( h_d(t) \),
assuming that detector threshold or saturation are not encountered:

\[ i_0(x,y,t) = g_0 \int_{-\infty}^{\infty} h(t-t') \overline{I_d(t')} \, dt' \]  

(3.147a)

Using Eq. (3.146b) results in

\[ i_0(x,y,t) = g_0 \int_{-\infty}^{\infty} dw_d \left[ \int_{-\infty}^{\infty} dw_c e^{jw_d t} H_d(w_d) B_\lambda(w_c + w_d) B_\lambda^*(w_c - w_d) \Gamma_0(w_d,w_c) \right] \]

(3.147b)

The bandwidth of the detector transfer function, 
\[ H_d(w) \approx \frac{1}{\tau_d}(h_d(t)) \]  

is generally much narrower \((\leq 10^9 \text{ Hz})\) than the optical carrier \((> 10^4 \text{ Hz})\), so that 
\[ w_c \pm w_d/2 \approx w_c \]  holds within the detector bandwidth. Thus 
\[ H_d(w_d) \]  is effectively a delta function, i.e. 
\[ H_d(w_d) \approx \delta(w_d)/\tau_d \], where \( \tau_d \) is the time constant of \( h_d(t) \). (For example, \( h_d(t) = [\exp(-t^2/\tau_d^2)]/\tau_d \) results in

\[ H_d(w) = \left[ \frac{\tau_d}{2} \exp(-w(\tau_d/2)^2) \right]/\sqrt{\pi} \tau_d \to \delta(w)/\sqrt{\pi} \tau_d \]  for large \( \tau_d \). The detector current can thus be written as

\[ i_0 = g_0 \int_{-\infty}^{\infty} dw_c |B_\lambda(w_c)|^2 \frac{|E_{wd}(w_c)|^2}{\tau_d} \]  

(3.147c)

\[ = g_0 \int_{-\infty}^{\infty} dw_c |B_\lambda|^2 I_{wd} \]  

(3.147d)

where

\[ I_{wd} = \frac{|E_{wd}|^2}{\tau_d} \]
Notice that $i_0$ in Eq. (3.147c) is no longer a function of time. Actually, Eq. (3.147b) shows that $i_d$ will change on a time scale of $\tau_d$ or less and it is useful to consider $I_{wd}(w,t)$ as a slowly varying spectral density, which is constant, unless for example $D_i$, $D_m$, or $|E_{wi}|^2$ in Eq. (3.145) change. These temporal fluctuations will be further filtered by the loop filter $\tilde{H}(w)$, which is assumed to be much narrower ($<10^6$ Hz) than the diode filter $H_d(w)$.

If ergodic temporal statistics are assumed, the ensemble average becomes a time average (i.e. $\bar{I} = \frac{1}{T} \int_{-T/2}^{T/2} I \, dt$), and a spectral intensity superposition similar to Eqs. (3.147c,d) follows directly from Eq. (3.146a). In Eq. (3.147c) $I_{wd} = |E_{wd}|^2/r_d$ is effectively the "local-time" power spectrum of the input field. (To a more accurate approximation, the power spectrum should be taken as $S(s) = \lim_{\tau_d \to \infty} I_{wd}$.)

The spatial and temporal coherence properties of $E_i(x,y,t)$ enter into Eq. (3.146) through the $|E_{wi}|^2$ dependence of Eq. (3.145). In general, a partial coherence analysis in terms of the spatio-temporal mutual-coherence function

$$E_i(x_1,y_1,t_1)E_i^*(x_2,y_2,t_2)$$

of the input field is required; however the mathematics are too unwieldy to conveniently present here. The Zernike phase-contrast IPL does not perform well unless $E_i$ is spatially coherent across the input aperture. Assuming spatial coherence, but leaving the temporal coherence unspecified, $E_i$ can
be written as

\[ E_i(x, y, t) = |E_i'(x, y)| f(t - T_d(x, y)) \]  

(3.148)

where most of the time variation is in \( f \), and \( |E_i| \) is temporally constant over times much longer than the coherence time of \( R_i(\tau) \equiv f(t) f^*(t + \tau) \). Then it is not difficult to show that \( I_{wd} \) follows from Eq. (3.145a) as

\[ I_{wd} = I_o(w) + I_1(w) \sin(w D_e/c) \]  

(3.149a)

where

\[ I_o(x, y, w) = |B_\lambda(w)|^2 S_i(w) \frac{|c_o|^2}{2\tau_0} (|E_i'(x, y)|^2 + |E_{wr}(w)|^2) \]  

(3.149b)

\[ I_1(x, y, w) = 2|B_\lambda(w)|^2 S_i(w) \frac{|c_o|^2}{2\tau_0} |E_i'(x, y)| |E_{wr}(w)| \equiv |B_\lambda|^2 I'_1 \]  

(3.149c)

and

\[ D_e(x, y) = D_i - D_m = D_i - D_m - D_o - D_c + D_l \]

In Eq. (3.149b,c) \( S_i(w) \) is the power spectrum of \( E_i \) and is equal to the Fourier transform of \( R_i(\tau) \). For example, white light is temporally incoherent and is characterized by \( S_i(w) = S_w \),
where \( S_w \) is a constant independent of frequency, \( w \). The corresponding autocorrelation of \( f(t) \) is impulsively correlated, i.e. \( R_i(\tau) = S_w \delta(\tau) \). (Note that the average signal in Eq. (2.147c) is proportional to \( R_d(\tau = 0) \), where \( R_d \) is the autocorrelation of the spectrally filtered detector field). Light which is temporally incoherent but spatially coherent over the input aperture is a reasonable model in applications where \( E_i \) arises from a small, distant while light object, such as in compensation of astronomical images; \( E_i \) then exhibits very rapid temporal fluctuations (on the order of a few times \( 2\pi/w \)), which are correlated throughout the input aperture.

Referring back to the system diagram in Fig. 3.18 and using Eqs. (3.149), allow a differential equation to be written for the behavior of the modulator OPD in the polychromatic phase-contrast IPL. With the specific low-pass loop filter \( \tilde{H}(w) = w_0/(jw + w_o) \) \((w < < \) optical frequencies), the differential equation for \( D_m \) is:

\[
\dot{D}_m = w_0 (-D_m + G''_o - \int_{-\infty}^{\infty} G_w(w) \sin\left(\frac{w}{c} (D_m - D_i')\right) dw \\
(3.150a)
\]

where

\[
G''_o = g_o g_1 g_2 \left( \int_{-\infty}^{\infty} I_0(x,y,w) \ dw - I_t \right) \\
(3.150b)
\]

and
\[ G_w = g_1 g_2 I_t(x,y,w) |B_{\lambda}(w)|^2 \equiv |B_{\lambda}|^2 G_w' \]  

(3.150c)

Here \( g_2' \) is the proportionality constant between modulator OPD and the modulator drive signal, and \( I_t \) is the detector threshold intensity.

**Linearized Behavior**

In order to gain insight into the behavior of Eq. (3.150a) it is useful to examine the linearized case, where the argument of the sine is small for all optical wavelengths, \( \lambda = 2\pi c/w \), passed by the spectral filter \( |B_{\lambda}|^2 \) in Eqs. (2.149b,c), i.e.

\[ \left| \frac{w}{c} D_e \right| < 2\pi x \]  

(for example \( x = 1/12 \))  

(3.151)

Then Eq. (3.150c) becomes

\[ D_m + w_0 (1 + G_1'') D_m = w_0 (G'' + G_1'') D' \]  

(3.152a)

where

\[ G''_1 \equiv \int_{-\infty}^{\infty} |B_{\lambda}|^2 G_w \frac{w}{c} dw \]  

(3.152b)
If $|B_\lambda|^2$ has a square spectral passband between $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$:

$$G''_1 = \int \frac{2\pi c/\lambda_{\text{min}}}{2\pi c/\lambda_{\text{max}}} G_w w dw$$

(3.152c)

Note that $D'_i$ and $D_m$ in Eq. (3.152a) have passed out of the spectral integrals. The steady-state reached by Eq. (3.152a) is

$$D_m = \frac{G''_1}{1 + G''_1} \left( D'_i + \frac{G''_0}{G''_1} \right)$$

(3.153a)

The steady-state OPD error is:

$$D_e = \frac{D'_i - G''_0}{1 + G''_1}$$

(3.153b)

For $G''_i > G''_0$ and $G''_i >> 1$, $D_m = D'_i$ and $D_e = 0$. In the more general case of arbitrary open-loop transfer function, $\sim(s)$, the linearized Laplace-domain response is

$${\sim}D_m = \frac{\sim H G''_1}{1 + H G''_1} \left( D'_i - \frac{G''_0}{G''_1} \right)$$

(3.154)

Equation (3.152-3.154) are significant results, which show that in the linear approximation the IPL successfully performs polychromatic phase compensation, tracking and compensating OPD rather than phase!
In fact, Eq. (3.152c) suggests that performance can be improved by increasing the spectral bandwidth, because $G_i''$ is increased. However as the maximum spectral frequency $w_{\text{max}} = 2\pi c/\lambda_{\text{min}}$ is increased, Eq. (3.151) shows that the maximum tolerable error, $D_e$, for validity of the linear approximation, must be decreased. Self-consistency demands that

$$\lambda_{\text{min}} > \frac{1}{xG_i''} D_i' \quad (3.155a)$$

or

$$D_i' < x G_i'' \lambda_{\text{min}} \quad (3.155b)$$

which follows from Eqs. (3.151) and (3.153) with $G_i'' > 1$ and $G_i'' > G_o''$, (according to Eq. (3.150a) the threshold $I_t$ can be utilized to reduce $G_o''$). For example, taking: $\lambda_{\text{min}} = 0.43 \mu m$ (approximately the minimum visible wavelength), $G_i'' = 40\pi$, and $x = 1/12$, OPD distortions as large as 14 $\mu m$ ($\phi = 65\pi \approx .43 \mu m$) can be compensated. In practice, saturation of the phase modulator at its maximum OPD will limit the maximum compensatable $D_i'$. Eq. (3.151) places no maximum constraint on $\lambda$; in practical applications dispersion and other system limitations may limit $\lambda_{\text{max}}$.

As in the monochromatic case, Eq. (3.150a) is multistable, also having linearized solutions where
\[ |\frac{W}{C} D_{\text{e}} - 2n\pi| = \left| \frac{2\pi D_{\text{e}}}{\lambda} - 2n\pi \right| < 2\pi x \tag{3.156} \]

The corresponding linearized Laplace domain result is

\[ D_{\text{m}} = \frac{\sum_{n=1}^{\infty} \frac{G_n''}{G_n' + G_n''} \left[ D_i' - \frac{G_i'}{G_n'} - 2n\pi \frac{G_n''}{G_n'} \right]}{1 + \sum_{n=1}^{\infty} \frac{G_n^2}{G_n}} \tag{3.157} \]

where

\[ G_2 = \int_{-\infty}^{\infty} |B_{\lambda}|^2 G_w dw \tag{3.158} \]

Eq. (3.156) places both minimum and maximum constraints on \( \lambda \), i.e.,

\[ \lambda_{\text{min}} = \frac{D_{\text{e}}}{n + x} < \lambda < \frac{D_{\text{e}}}{n - x} = \lambda_{\text{max}} \]

The spectral bandwidth for which the linearized solution is valid is thus greatly diminished for the higher \( n \)-value solutions, becoming

\[ \Delta \lambda = \lambda_{\text{max}} - \lambda_{\text{min}} = \frac{2x}{n^2 - x^2} D_{\text{e}} = \frac{2x}{n} \lambda_0 \tag{3.159} \]

where \( \lambda_0 = \frac{\lambda_{\text{max}} + \lambda_{\text{min}}}{2} = \frac{nD_{\text{e}}}{n^2 - x^2} \approx D_{\text{e}}/n \)
For example, with $\lambda_0 = 0.6 \mu m$, $x = 1/6$, and $n = 1$, the spectral bandwidth is $\Delta \lambda = 0.2 \mu m$.

**Nonlinear Behavior**

The nonlinear behavior of the polychromatic phase-contrast IPL with a low-pass loop filter is described by the differential equation (3.150). The corresponding $D_m$ versus $D_m$ state-space characteristic can be plotted by numerically evaluating the integral in Eq. (3.150a) with an appropriate functional form for $G_w(w)$. For example, when: the assumptions leading to Eq. (D.11) in appendix D are satisfied, $S(w) = S_w$, and $|B_\lambda|^2 = 1$ within a spectral passband:

$$G_w(w) = \text{constant} \cdot w^2 \sqrt{1 + b^2 - 2b \cos(wD_H/c)} \cdot \exp\left(-\frac{w^2}{2c^2} \sigma_D^2\right)$$

where $\sigma_D^2 = <(D_i - D_m - D_o)^2> - <D_i - D_m - D_o>^2$ (Note that $G_w(w)$ is maximum near $\lambda = \pi \sigma_D/\sqrt{2}$).

Numerical experiments have suggested that the $D_m$ versus $D_m$ behavior of Eq. (3.150a) is somewhat related to that of the low-pass homodyne IPL in Fig. 2.7. To understand this, note that the integral in Eq. (3.150a) always equals zero when $D_m = D_i^l$ and changes sign when the relative values of $D_m$ and $D_i^l$ change. As $D_m$ moves from its initial condition, of for example zero, a stable white-light equilibrium thus results at $D_m = D_i^l$, (assuming that $G_o^2 < G_i^2$ of Eq.(3.152b) ). The numerically obtained value of $D_m$ was found to agree closely with the linear approximation of Eq. (3.153a). In practice achieving $D_m = D_i^l$ may be difficult if $D_i^l$ is large, however the fact that $D_c$ of Eq. (3.145c) is often equal
to the spatial average \(<D_i - D_m - D_o\) tends to keep \(D_i^r\) small.

Numeric integration reveals that the integral in Eq. (3.150a) is characterized by decreasing oscillations as \(D_m - D_i^r\) increases, much like a sinc function. Nulls occur near \(n(\lambda_{\min} + \lambda_{\max})/2\) when \(|B_\lambda|^2\) has a square bandpass between \(\lambda_{\min}\) and \(\lambda_{\max}\). If \(G_i^r\) is not too large, multiple stable nulls in \(\tilde{D}_m\) occur near the linearized solutions in Eq. (3.157). The allowable spectral bandwidth of these solutions is significantly larger than suggested by Eq. (3.159). For small \(n\) values the spectral bandwidth can be extended to cover more than the whole visible spectrum, \((\approx 0.43-0.68 \mu m)\), by making \(G_i^r\) sufficiently larger than \(G_o^r\).

There are only a finite number of multistable solutions, because, as in the LPF homodyne IPL, the \(-D_m\) term in Eq. (3.150a) introduces an overall negative slope into the \(\tilde{D}_m\) versus \(D_m\) state-space. Notice, also in analogy with Fig. 2.7, that the whole trajectory shifts to the right and downward as \(D_i^r\) increases.

In operation, the polychromatic phase-contrast IPL is thus expected to converge to a stable equilibrium near the initial value of \(D_m\), and then track slow variations in \(D_i^r\) until \(D_m\) either reaches its physical limits or the equilibrium becomes unstable. As the spectral bandwidth, \(\Delta \lambda\), is increased, the OPD estimation error of the \(n=1\) solutions tends to grow, and the total number of multistable solutions diminishes; the overall \(D_i^r\) tracking range of a given \(n\)-value equilibrium also tends to decrease.

For a given \(\Delta \lambda\), the performance, in terms of OPD error and \(D_i^r\) tracking range, tends to improve as \(G_i^r\) is increased, (e.g. as the electronic gain is increased). It should be stressed that even with completely white light, a solution appears to always exist at \(D_m = D_i^r\) (\(n=0\)) when the system is
designed with $G_0^<G_0^>$. ($G_0^>$ can be controlled by appropriate choice of $I_t$ in Eq. (3.150b)).

In conclusion, the phase-contrast IPL performs surprisingly well in broad-spectrum or white light, becoming an OPD estimator rather than a phase estimator. A major reason for the good performance is that the IPL tends to null the OPD error, $D_e$.

This analysis was subject to the restrictions that the irradiance be spatially coherent across the input aperture and that the modulator be nondispersive. These restrictions are reasonable for astronomical image compensation with a deformable mirror or a nondispersive electrooptical material (if such a material exists). In fact, Dicke$^{13}$ had previously proposed the use of Zernike phase-contrast in an astronomical image compensation system with a deformable mirror. Analysis of the performance degradation suffered by the phase-contrast IPL as a function of relaxing the spatial coherence and dispersion assumptions is left for future work.

When $\sigma_{D_i} = \langle D_i^2 \rangle - \langle D_i \rangle^2$ is large the initial gain is often reduced by a factor similar to $\exp(-\sigma_{D_i}^2/2)$ and may be too small for convergence, particularly when there is additional gain reduction due to a broad spectral bandwidth. Another consideration is whether the system has enough dynamic range to actually reach the OPD null $D_i' = D_m$; the fact that $D_i'$ includes a $D_i'-\langle D_i \rangle$ term enhances the likelihood of achieving this null.

The ultimate feasibility of employing the phase-contrast IPL in a given white light application will depend upon such application-dependent considerations as the rms OPD variation, $\sigma_{D_i}$, the temporal rate of fluctuations in $D_i$, and the levels of background and signal induced shot noise.
3.2.2.11 Summary of phase-contrast IPL behavior

The phase contrast IPL was shown to perform phase estimation and compensation, with \( \phi_m = \phi'_i \text{-const}(x,y) \) and \( \phi_e = \text{const}(x,y) \). This behavior is similar to that of the homodyne IPL, with the exception that the reference phase of the estimate and loop gain at each spatial location are a function of the overall spatial phase distribution in the input aperture. This spatial coupling leads to the result that: although \( \phi_m \) is an estimate of \( \phi_i \) at each point, the average value of \( \phi_m \), \( \langle \phi_m \rangle = <G_o> + <G_i> \cdot \sin \phi_i \), is independent of \( \phi_i \) ! This seeming paradox was resolved by showing that the phase is estimated relative to a reference phase which includes the average input phase \( \langle \phi_i \rangle \), hence the average relative phase \( \langle \phi_i \rangle \) is independent of \( \phi_i \). (More specifically, \( \phi'_i = \phi_i - \langle \phi_i \rangle - \phi_0 - \langle \phi_0 \rangle + <G_o> + <G_i> \cdot \sin \phi_i \)).

The behavior above results in "self centering" of \( \phi_m \), with \( \phi_m \) being maintained near \( <G_o> + <G_i> \cdot \sin \phi_i \). This tends to: keep \( \phi_m \) away from the physical limits of the modulator, maintain an average error (e.g., \( \phi_e = 0 \), for \( \phi_i = 0 \)), and diminish \( 2\pi \cdot \) ambiguities. As time passes only spatially variant, and not uniform "piston", phase aberrations are tracked. The phase-contrast IPL also has an intrinsic "automatic gain control" property, since the amplitude of the effective reference beam is proportional to the spatial average of the input field amplitude. This enhances the immunity to amplitude fluctuations. When a nondispersive phase modulator is employed, the IPL can potentially perform well in white light. It then estimates and compensates OPO, rather than phase.

After convergence with \( G_i \cdot \sigma_i \), the sensitivity of the phase-contrast IPL is comparable to that of the heterodyne or homodyne IPL. Before convergence, however, the gain is diminished by approximately \( e^{-\sigma_i^2/2} \), which can be serious when the rms phase fluctuations, \( \sigma_i \), are large.
The dynamic response of the phase contrast IPL is characterized by a rapid convergence to the local spatially-relative phase estimate, with the closed-loop transfer function; followed by a slower drift of $\phi_m$ to the vicinity of $\langle G_0 \rangle + G_1 \sin \phi_1$, with the open-loop transfer function. If a given hardware system has stable open-loop and closed-loop transfer functions at all spatial locations in an independent-loop IPLs (e.g., homodyne), the phase-contrast IPL will also be stable.

The behavior summarized above was supported by theoretically investigating: the response to an input phase change in a single spatial location; the global numerically simulated behavior; the linearized behavior, assuming only that $\phi_e$ and $\phi_i$ are small; and the general behavior of the nonlinear low-pass loop filter case. The last result is particularly informative: Eqs. (3.125)-(3.128) represent the general low-pass solution for the behavior of this spatially-coupled, nonlinear (transcendental), multistable system.

In conclusion, the phase-contrast IPL shows promise for implementing high-resolution adaptive systems. Not requiring a stable local oscillator, it is fairly simple to construct and looks particularly attractive for such nonmonochromatic applications as astronomy. The principal disadvantage of the phase-contrast IPL is gain reduction during initial convergence. This can make the phase-contrast IPL unattractive in some instances, depending on: the rms phase fluctuation $\sigma_i$, the signal level $|E_i|$, the rate of phase fluctuations, and the spectral bandwidth of the application.

3.3 Summary

Chapter 2 employed the homodyne IPL with a low-pass loop filter to develop the basic static and dynamic behavior of the IPL. It was
shown that the IPL can stably estimate and compensate phase in real-time over multiple $\pi$ radians of dynamic range without phase quadrant ambiguity, even in the presence of significant fluctuations in signal amplitude.

This chapter examined IPL behavior in greater depth and extended the scope of the analysis to include other interferometric implementations and loop filters. Performance issues addressed here included: the effects of signal amplitude fluctuations on IPL performance, phase acquisition transient response, multistability, spatial resolution, IPL performance in the presence of noise, minimum signal requirements, and spatial phase compensation with the IPL. Some of the major conclusions of this chapter are outlined below.

When the local oscillator is designed to be greater than the largest expected signal, it was shown that the homodyne IPL can be made insensitive to over an order of magnitude of signal amplitude fluctuation. The immunity to amplitude fluctuations with an integrating loop filter is potentially superior to that with a low-pass filter. In some instances, the integrating case can be adjusted to operate as the intensity diminishes all the way to zero, with any value of input phase.

Multistability and $2\pi$ degeneracy of $\phi_m$ can be controlled by exploiting the fact that the IPL generally begins operating within $\pm \pi$ of its initial modulator phase, $\phi_{mo}$. In some instances, optimal performance is obtained for $\phi_{mo} = 0$, particularly when the offset $G_0$ is in the center of the operating range, i.e. $G_0 = \phi_s/2$. Multiple-$\pi$ spatial phase fluctuations can be tracked without ambiguity if they are introduced after the IPL is already in operation.

The IPL must spatially and temporally resolve the phase distortions
to be estimated and compensated, in order to modulate the aberrated wave with its conjugate. Aberration components outside the IPL spatial or temporal bandwidth are tracked with diminished closed-loop gain and contribute to the phase error of the compensated wavefront. A discrete-element IPL should satisfy the Nyquist sampling theorem in the spatial domain. This often amounts to requiring that each detector element see a uniform shade of grey in the open-loop interferogram.

In phase compensation applications, the IPL can enhance the intensity seen at a diffraction-limited detector (e.g., in a direct-detection communications receiver) from approximately $e^{-\sigma_i^2}$ to $e^{-\left(\sigma_i / G_1\right)^2}$. Operation with $G_1 > \sigma_i$ is thus desirable.

The behavior of IPL systems characterized by various loop filter time responses, including those with: a single dominant time constant (low-pass and integrating) and two dominant time constants (type-zero and type-one), were compared in table 3.1. The steady-state behavior was shown to depend on the loop-filter type (e.g., zero or one) rather than the total number of time constants. The type-one systems are characterized by a step-response error independent of $\phi_i$ and $\phi_m$ and a constant heterodyne ramp error. The type-zero step-response error is dependent on $\phi_i$ and $\phi_m$; this sometimes decreases the allowable range of amplitude fluctuations, but allows error minimization by operation with $\phi_i = G_0$.

The continuously increasing type-zero ramp error is spatially uniform and hence does not degrade spatially-relative phase estimation and compensation performance.

Exact expressions were obtained for the nonlinear transient response of the integrating system, with time constant $\tau = \frac{1}{\lambda_1 \sigma_1 G_0^2 - G_0^2} = \frac{1}{\lambda_1 G_0^2}$. The approx-
imate time constant of the low-pass system is $[\lambda (1 + G)]^{-1}$. Numerical tech-
niques were employed to investigate the transient response of the two-pole
systems. The type-zero systems are unconditional stable in response to
a step, but the type-one systems are sometimes afflicted with state-space "dead regions" and can exhibit a limit-cycle in $\phi_m$ if $G_1$ is not large
enough.

The loop filter is often fixed by hardware constraints. When the fil-
ter can be designed, however, the single-pole filters are preferable;
since more complicated systems are usually much more difficult to analyze
and adjust but do not generally offer significant performance advantages.

Section 3.1.8 showed that extraneous coherent illumination, e.g.
extra beamsplitter reflections, need not have serious consequences.

Potential sources of noise in IPL systems include: shot noise due
to signal irradiance, local oscillator irradiance, background light,
and dark current; and thermal noise. The signal and/or local oscillat-
or shot noise is generally unavoidable and dominates under ideal circum-
stances. The major consequence of noise is to introduce fluctuations
into $\phi_e$ and $\phi_m$ about their noiseless equilibrium values ($\bar{\phi}_e = 2n\pi$, $\bar{\phi}_m = \phi_1 - 2n\pi$). These fluctuations are approximately Gaussian distributed (assuming high
density shot noise), and can be characterized by their rms magnitude, $\sigma_{\phi} = \frac{1}{S/N}$.
A nonlinear analysis revealed that noise fluctuations also cause the IPL
to occasionally lose phase lock and jump ($\pm 2\pi$ radians) to another multi-
stable $\phi_m$ or $\phi_e$ equilibrium. Consideration of the mean time to first
lose phase lock suggests that the IPL signal-to-noise ratio must exceed
a threshold of approximately $S/N > 0.5$, in order to converge faster than
phase lock is lost. A larger $S/N$ is often desirable to keep $\sigma_{\phi}$ small.
Noise considerations alone are generally insufficient to specify the minimum irradiance required by the IPL, because the system photon integration time tends to increase as the signal decreases. Noise and performance constraints taken together, however, can generally be employed to determine a minimum irradiance and optimum system gain for a particular application. Ideally, the minimum signal is on the order of a few photons in a resolution cell per system integration time. When dark current or background light is significant, the signal photon count should generally exceed the square root of the dark current or background count.

The three interferometric realizations of the IPL; heterodyne, homodyne, and self-interference (Zernike phase-contrast); discussed in this chapter provide a rounded perspective of IPL behavior. The heterodyne IPL was shown to have three successive operating regimes: initial rapid convergence, real-time phase tracking, and either limit-cycle oscillations in \( \phi_m \) or lock-up at the maximum modulator phase. Practical operating schemes exploiting this behavior were suggested, including a sample-and-hold framed mode, a phase-tracking framed mode, and a limit-cycle continuous mode. Much like a conventional PLL, the phase-tracking mode tends to lock the compensated signal-beam frequency to the local oscillator optical frequency. This reduces the stability and frequency tolerance demands on the local oscillator laser. Another potential advantage of heterodyne operation is reduction of the \( G_0 \) offset to \( G_0 - \Delta w/\lambda_1 \), which can offer a variety of advantages, including: improved immunity to amplitude fluctuations, faster phase acquisition, larger bandwidth, diminished tracking error, and removal of the integrating loop filter "dead regions".
The Zernike phase-contrast self-interference IPL configuration presented the challenge of analyzing a multi-stable, spatially-coupled nonlinear system. Numeric, perturbational, linearized, as well as general (e.g., Eqs. (3.126)-(3.127)) solutions were obtained. The transient response includes a rapid phase estimation and compensation convergence followed by a slower "self-centering" spatially-uniform drift of $\phi_m$. This "self-centering" feature tends to: decrease $2\pi\text{m}$ ambiguity, maintain an average error independent of $\phi_i$ (e.g., zero), and keep $\phi_m$ away from the physical limits of the modulator. The phase-contrast IPL also exhibits an "automatic gain control" property which enhances its immunity to signal amplitude fluctuations. The primary shortcoming of the phase-contrast IPL is gain reduction by approximately $e^{-\sigma_i^2}$ during initial convergence, where $\sigma_i$ is the rms spatial phase fluctuation of the uncompensated input wavefront. This self-interference implementation looks particularly attractive for such applications as astronomical image compensation, since it requires no external reference beam and can operate in white light.

This chapter in conjunction with Chapter 2 attempted to address most of the major aspects of IPL performance. This material provides a broad theoretical base which can be employed to design, characterize, and optimize IPL systems, given a particular set of hardware and application constraints.
IV. MONOLITHIC SPATIAL LIGHT MODULATORS

Although adaptive phase compensation can be accomplished by employing many hard-wired modulators in parallel, that approach is too combersome to be worthy of serious consideration in high-resolution applications. A much more practical approach is an all-optical parallel processing configuration employing an optically-addressed monolithic spatial phase modulator (SLM). Adequate performance in many high-resolution applications requires a phase-only SLM with a million or more resolution elements, better than kHz operating rates, a dynamic range of at least $2\pi$ radians, and extreme (e.g., quantum limited) sensitivity. Since no existing optically-addressed modulator meeting these performance objectives was available, it was necessary to develop a relatively new modulator, the MSLM, for this application.

The next section reviews most of the existing optically-addressed SLMs. The section after that, which accounts for the majority of this chapter, is concerned with the operating principles, performance, and experimental development of the MSLM. Because many of the important contributions of this thesis research were in the realm of MSLM development, this discussion goes beyond the minimal description of the MSLM required to understand its operation in an IPL adaptive system. The last section of this chapter narrows the discussion back to some of the special considerations of attempting to implement an interference phase loop with the MSLM.

4.1. Available modulators

Over the past decade and a half a variety of optically-addressed SLMs have been reported in the literature, most having been developed
for display, digital mass storage or block data (page composer) applications. Refs. 76, 77, 177, and 178 present good overviews of these devices and the major characteristics of a few are summarized in Table 4.1.

Most existing optically-addressed SLMs combine a photoconductive layer and a voltage controlled amplitude and/or phase modulating (e.g., electrooptic) material in a sandwich structure. With no write beam, a voltage applied to the sandwich is dropped mostly across the unilluminated photoconductor by capacitive voltage division. In illuminated regions the photoconductor shunts the voltage to the modulating layer, which in turn impresses the spatial modulation on a readout beam. A specific sandwich often employed is: a write-beam window with a transparent electrode, the photoconductor, an opaque layer to isolate the write and read beams, a reflecting layer to allow readout by reflection, the modulating material, another transparent electrode, and the readout window.

In the Phototitus \(^{179}\) (or photo DKDP \(^{180}\) SLM, the modulating layer is KDP, an electrooptical crystal. The liquid crystal light valve \(^{181-186}\) (LCLV) employs a liquid crystal; and the FERPIC, CERAMPIC, and FERICON devices \(^{187}\) use a PLZT ceramic.

The Ruticons \(^{186-190}\) have a slightly different structure, with an electrostatically-deformable elastomer placed directly on the photoconductor. Electrical contact is made to the other side of the elastomer in one of three ways: a conducting liquid (α-Ruticon) a gas discharge (β-Ruticon), or a reflective conducting membrane (γ-Ruticon). The thermoplastic (TP) SLM is very similar to the β-Ruticon, except that
<table>
<thead>
<tr>
<th>Name</th>
<th>Modulating Material</th>
<th>Light Sensor</th>
<th>Resolution (lp/mm)</th>
<th>Sensitivity (µW/cm²)</th>
<th>Phase Dynamic Range (Radians)</th>
<th>Operating Voltage (Volts)</th>
<th>Time Write (msec)</th>
<th>Response Erase (msec)</th>
<th>Store Developed At</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo-titus</td>
<td>KDP</td>
<td>amorphous Se</td>
<td>6.5</td>
<td>10</td>
<td>0.5</td>
<td>250 (-51°C)</td>
<td>0.01</td>
<td>0.03</td>
<td>&lt;1 hr (-51°C)</td>
<td>France</td>
</tr>
<tr>
<td>LCLY</td>
<td>Twisted nematic</td>
<td>Cds</td>
<td>33.4</td>
<td>5</td>
<td>1.7</td>
<td>5 - 15 (ac)</td>
<td>10</td>
<td>15</td>
<td>15msc</td>
<td>Hughes</td>
</tr>
<tr>
<td></td>
<td>liquid crystal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FERPIC</td>
<td>Strain-biased</td>
<td>ZnCdS</td>
<td>20.4</td>
<td>5</td>
<td>&gt;3e</td>
<td>525 (1mm thick)</td>
<td>10</td>
<td>10</td>
<td>1000</td>
<td>Bell</td>
</tr>
<tr>
<td>(FERICON)</td>
<td>PLZT</td>
<td>PVK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scandia</td>
</tr>
<tr>
<td>Ruticon</td>
<td>Deformable</td>
<td>amorphous Se</td>
<td>10-45</td>
<td>30</td>
<td>0.2</td>
<td>300 (for 0.1µm)</td>
<td>5</td>
<td>4</td>
<td>15min</td>
<td>Xerox</td>
</tr>
<tr>
<td></td>
<td>elastomer, PVK:TNF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROM</td>
<td>Bi₁₂SiO₂₀</td>
<td>Bi₁₂SiO₂₀</td>
<td>25.0</td>
<td>5</td>
<td>&gt;2e</td>
<td>3900e</td>
<td>5</td>
<td>1</td>
<td>&lt;2hrs</td>
<td>Itek</td>
</tr>
<tr>
<td>MLM</td>
<td>Deformable membrane</td>
<td>Sl</td>
<td>3</td>
<td>2</td>
<td>2e</td>
<td>40-100e</td>
<td>0.001</td>
<td>10</td>
<td></td>
<td>Perkin-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Elmer</td>
</tr>
<tr>
<td>Thermoplastic</td>
<td></td>
<td>PVK:TNF</td>
<td>20-120</td>
<td>5</td>
<td>&lt;e</td>
<td>700-900</td>
<td>10</td>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oil-film SLM</td>
<td>Oil-film</td>
<td>heat</td>
<td>51.0</td>
<td>100</td>
<td>14e</td>
<td>0 (passive)</td>
<td>100</td>
<td>&lt;100</td>
<td>&lt;0.1 sec</td>
<td>Switzerland</td>
</tr>
<tr>
<td></td>
<td>(Polysiloxan)</td>
<td>absorption in plastic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEMLM</td>
<td>Deformable</td>
<td>photocathode</td>
<td>5</td>
<td>0.001</td>
<td>0.5e</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>100 sec</td>
<td></td>
<td>England</td>
</tr>
<tr>
<td>MSLM</td>
<td>Electroptic</td>
<td>photocathode</td>
<td>20</td>
<td>0.002</td>
<td>10⁻⁶</td>
<td>5³</td>
<td>250-8000³</td>
<td>&lt;1³</td>
<td>&lt;1³ weeks</td>
<td>MIT</td>
</tr>
</tbody>
</table>

**Notes:**
- Best reported values unless otherwise noted
- a) Best quoted value in near future
- b) Spatial bandpass MTF
- c) Resolution at 50% contrast
- d) Resolution at 10% contrast
- e) For half-wave modulation in reflection
- f) For e⁻ response
- g) Commercially available
the elastomer is replaced by a plastic which is deformable only when heated. The membrane light modulator (MLM) resembles the $\gamma$-Ruticon, but instead of the elastomer there is rigid dielectric containing holes over which a metallic membrane is stretched. Electrostatic forces deflect the membrane into the holes.

The Pockels Readout Optical Modulator (PROM) does not require a photoconductor, since the modulating element, $\mathrm{Bi}_{12}\mathrm{SiO}_{20}$, is a photoconductive (blue sensitive) electrooptic crystal. The PROM sandwich consists of two transparent electrodes which are isolated from the crystal by thin dielectric layers. Photoferroelectric crystals also do not require a photoconductor, but they are only useful at high spatial frequencies (600 to 2000 lp/mm) and have an extremely small dynamic range ($<\pi$ radians).

A new oil-film SLM has recently been reported. This device contains an oil-film which is deformed in an image-wise manner by the heat generated when the write-beam is absorbed in an underlying plastic layer.

The MSLM, which is illustrated in Fig. 1.2, does not rely on photoconductivity. A similar SLM, the photoemitter-membrane light modulator (PEMLM) with a membrane stretched directly across the MCP output, has also been reported in the literature.

The IPL adaptive system requires a phase-only SLM. The deformable modulators meet this constraint, but they respond only in a relatively narrow spatial passband. This drawback can be overcome by the complexity of adding a spatial carrier to the write beam.

Phase modulation proportional to the log of the write-beam irradiance
has been reported in the twisted nematic liquid crystal of the LCLV. This modulation saturates when all the molecules have become aligned but increases with increasing liquid crystal layer thickness. Unfortunately, there is an unavoidable polarization rotation which usually results in simultaneous phase and amplitude modulation.

Of the PLZT devices, the CERAMPIC operates by scattering and is useless for phase modulation, however the strain-biased FERPIC exhibits useful birefringence and the FERICON exhibits useful surface deformation. The phase dynamic range of the FERPIC depends on the amount of strain bias, more than $3\pi$ radians ($1\ \text{mm thick, 633 nm}$) has been reported (with only 3% scattering).

Devices which employ an electrooptic crystal (e.g., phototitus, PROM, MSLM) can produce phase-only modulation with a large dynamic range, the ultimate limit being dielectric breakdown of the crystal or other layers. Thick uniform photoconductive layers are extremely difficult to fabricate and have long response times; hence the photoconductor in the phototitus is only $12\ \mu\text{m}$ thick corresponding to $120\ \text{volt}$ breakdown and a dynamic range of $\sim\pi/2$ radians. In the PROM, high resolution demands a thin crystal ($800\ \mu\text{m}$) which forces the dielectric layers to be thin (6 $\mu\text{m}$) in order to obtain proper voltage division. The vacuum gap in the MSLM can withstand very large voltages. Electrooptic crystals also offer an extremely rapid intrinsic time response ($\sim 1\ \text{psec}$), unlike liquid crystals ($\sim 10\ \text{msec}$), PLZTs (50 to 400 $\mu\text{sec}$), deformable elastomers (1 $\text{msec}$), or deformable membranes (0.5 $\mu\text{sec}$). The spectral dispersion of an electrooptic crystal can be a disadvantage in non-monochromatic applications.
The first seven modulators in Table 4-1 employ photoconductivity, which limits their ultimate sensitivity and response time. The time response of a photoconductor is limited primarily by carrier transit time across the photoconductive layer, and the carrier decay time (lifetime) to return to a high resistance state before the next image can be written (> 50 μsec). In comparison, the MSLM and PEMLM are much more sensitive and have a time response limited only by the maximum current which is available to charge the modulating material.

Most SLMs exist only as laboratory prototypes and some are not even being actively developed anymore. Only the PROM and LCLV are available commercially, costing about $30,000 each; and only a few tens of each have been manufactured. The PROM requires different wavelength illumination for writing (blue) than for readout (red) and hence cannot be employed in a closed-loop IPL adaptive system. The LCLV is also not suitable, since it simultaneously modulates both phase and amplitude. Attempts at borrowing prototypes of other SLMs were unsuccessful. Thus even a crude laboratory demonstration of an all-optical adaptive system required that an SLM also be constructed.

Previous chapters showed that the SLM employed in an IPL should have a dynamic range of at least 2π radians. Of the remaining SLMs in table 4.1, only the FERPIC, MLM, oil film SLM, and MSLM qualify; all of which require further developmental work. The history of MSLM technology at MIT75-84 coupled with the MSLMs superior ultimate speed and sensitivity performance capabilities, makes the MSLM the clear choice.
4.2. The Microchannel Spatial Light Modulator (MSLM)

The basic construction and operation of the MSLM were introduced in the discussion of fig. 1.2 in Chapter 1. The earliest MSLM's, \textsuperscript{78,79} employed distorted electro-optic crystals, were incapable of imaging, and at best exhibited localized patches of modulation in response to a uniform write beam. These early devices were also insensitive, containing a photocathode sensitive to only UV with a quantum efficiency of $10^{-8}$ at 254 nm. Their corresponding half-wave sensitivity (i.e. exposure required for $\pi$ radians of phase modulation) was about 1 mJ/cm$^2$.\textsuperscript{78,79} Their write time exceeded 1 sec for a maximum dynamic range of $\frac{2}{3} \pi$ radians, and their erase time exceeded 60 sec.

A large theoretical and experimental research program was undertaken to understand and improve the MSLM's operating characteristics; significant performance improvements were achieved. These include construction of a vacuum-sealed MSLM with a high quantum efficiency visible photocathode ($\eta = 6\% @ 655$ nm) and high sensitivity ($S_\pi = 2.2$ nJ/cm$^2 @ V_{\text{MCP}} = 1.5$ kV). That modulator employed an electro-optic crystal flat to $\lambda/2$ and exhibited a resolution of 2 lp/mm. Other MSLMs were produced with half-wave write and erase times of $\sim 200$ msec. This research also included the first successful imaging, and secondary emission electron removal experiments with the MSLM. Specific approaches have been developed for upgrading the sensitivity, dynamic range, speed, resolution, and optical quality of future MSLMs. The MSLM is expected to ultimately have a half-wave sensitivity of better than 1 pJ/cm$^2$, a framing rate of better than 1 kHz, and a spatial resolution of approximately 20 line pairs/mm.

A large variety of photocathodes and electro-optic materials are available to increase the versatility of the MSLM. The control-signal
wavelength is limited only by the response of existing photocathodes; no photocathode is required for extreme UV or soft x-rays. The readout light is bounded by the spectral transmittance of electro-optic materials (middle UV to middle IR). The MSLM is primarily a phase modulator; intensity modulation can be produced by employing an external interferometer, or by the Fabry-Perot interferometer resulting from the reflection from both surfaces of a parallel electro-optic crystal. Amplitude modulation can be produced by employing an electro-optic material with induced birefringence, such as KDP or Bi_{12}SiO_{20}, between crossed polarizers.

4.2.1. Principles of Operation

The modulator can be operated either in a framed mode with discrete write, read, and erase periods, or in a continuous mode. In the framed mode of operation, the control optical signal is integrated and stored in the form of an electron image; after the modulation is read out, the electron image is erased. In the continuous mode of operation, the "instantaneous" modulation is proportional to the control optical image intensity for all temporal variations within the bandwidth of the system.

Greatly simplified (see fig. 4.2), the framed mode can be modelled as a current source \( i_g \) charging the electro-optic plate capacitance, i.e.,

\[ V_x = \frac{1}{C_x} \int i_g \, dt \]

and the continuous mode is essentially a current source driving the electro-optic plate resistance, i.e., \( V_x = R_x \, i_g \)

4.2.1.1. Circuit Models

The performance of the MSLM can be understood in more detail by referring to the circuit model of fig. 4.1. This model pertains to a
Fig. 4.1 Four-pole circuit model

$I = \text{photoinduced output current from the MCP}$
$g = \text{gap between the MCP and dielectric mirror}$
$T = \text{layers of the dielectric mirror (e.g. TiO}_2\text{ layers)}$
$Q = \text{layers of other material in dielectric mirror (e.g. SiO}_2\text{)}$
$X = \text{electro-optic crystal}$
$R_L = \text{effective resistance from a point on the dielectric mirror to ground}$
$R_S = \text{discrete resistor in series with } V_b$
$R_D = \text{leakage resistance around the mirror and crystal}$

(E = effective crystal-mirror parameters)

Fig. 4.2 Simplified two-pole circuit model
spatial resolution cell of area $A_r$, which is determined by the overall resolution of the MSLM. Although $i_g = JA_r$, $C_g$, $R_L$, $C_T$, $R_T$, $C_Q$, $R_Q$, $C_x$ and $R_x$ are all effective values for a resolution cell, it is often convenient to use the total modulator area for $A_r$, since the $A_r$ dependence will cancel out of most of the operating equations. It is also easy to rescale the parameters to a new value of $A_r$. Typical measured and calculated device parameters for a one inch diameter modulator ($A_r = 5.07$ cm$^2$) are:

- $C_x = 2 \times 10^{-10}$ F (LiNbO$_3$, 0.25 mm thick), $C_Q = 2 \times 10^{-8}$ F (SiO$_2$).
- $C_T = 10^{-6}$ F (TiO$_2$), $R_x = 5 \times 10^{13}$ $\Omega$, $R_Q = 2 \times 10^{10}$ $\Omega$, and $R_T = 10^3$ $\Omega$. The resistance $R_L$ represents the resistance between a resolution cell on the surface of the dielectric mirror and ground when the edge of the mirror is grounded. In general, $R_L$ can vary over several orders of magnitude, depending on the type of coating, if any, deposited on the electron surface of the crystal. The discrete resistor $R_b$ in series with $V_b$ is generally less than $10^6$ $\Omega$. The gap capacitance $C_g$ depends on the gap width; e.g. $10^{-11}$ F for a 0.1 mm gap with $A_r = 5.07$ cm$^2$.

The intensity-controlled output current $i_g = JA_r$ of the MCP is modelled as a dependent source. The observed relationship between the MCP output current density, $J$, and gap voltage, $V_g$, is idealized in fig. 4.3. The MCP electrons generally create secondary electrons at the electron reception surface, or "target", e.g. dielectric mirror. The exact shape of the $J$ vs. $V_g$ characteristic depends on MSLM geometry, the energy distribution of MCP electrons, and the secondary emission characteristics of the target. At small $V_g$, $J$ decreases due to the collection of some secondary electrons from the target (and some primary MCP electrons) at the output face of the MCP and other grounded surfaces, rather than at the target. It should
Fig. 4.3 MCP current density, $J$, as a function of gap voltage, $V_g$.

$I_1$ is an irradiance level. From Eq. (4.5), $J_1 = \frac{V_b}{(R_s + R_E)A_r}$, and $V_{g1} = \frac{V_b R_L}{(R_s + R_L + R_E)}$. 
be noted that the current densities are almost always too low

\[ J < J_{\text{s.c.}} = \frac{4}{9} \varepsilon_0 \sqrt{\frac{2e}{m}} V_g^{3/2} \bar{\lambda}_g^{-2} = 23 \text{ ma/cm}^2 \text{ at } V_g = 1 \text{ volt and a gap width of } \bar{\lambda}_g = 1 \text{ mm} \]  

for space charge effects to be significant. With negative gap voltages, \( V_g < 0 \), many MCP electrons and all the secondaries from the target are repelled back to the MCP, often reversing the direction of current flow. These secondary emission effects are discussed in more detail in Appendix B.2; including the derivation of an equation for the \( J \) vs. \( V_g \) curve in fig. 4.3 as a function of MSLM material parameters.

As long as \( V_g > V_{g_c} \) in fig. 4.3, all the MCP electrons and secondaries are collected at the target and \( J \) is just the primary MCP current density \( J_p \), which is controlled by the write-beam irradiance \( I \) through:

\[ J_p = \frac{\eta e GI}{h \nu} = r_I I \]  

(4.1)

Here \( \eta \) is the photocathode quantum efficiency, \( e \) is the electron charge, \( G \) is the MCP gain, \( h \) is Planck's constant, \( \nu \) is the optical frequency, and \( r_I \) is the "current responsivity." At high optical irradiance, \( J \) begins to saturate and \( G \) decreases as \( J \) approaches the MCP "strip current" level, \( J_{\text{strip}} \). (The "strip current" flows in the walls of the pores, and is the source for the amplified MCP output current.)

The Laplace frequency response \( \tilde{V}_x(s) \) and time domain behavior corresponding to the circuit model of fig. 4.1 are presented in Appendix B.1. The results are in agreement with circuit intuition. There is an initial rapid capacitive division of any voltage changes between the crystal, gap and mirror with time constant:

\[ \tau_1 \approx (R_L || R_s)(C_g || C_x || C_T || C_Q) \]  

(4.2)
(Here $||$ is a mathematical operator: $x||y||z = (x^{-1} + y^{-1} + z^{-1})^{-1}$. This is followed by a slower transition to a resistive-dominated steady-state with:

$$\tau_2 \approx (R_S + R_L)(R_X + R_Q + R_T)[C_G + C_X || C_Q || C_T]$$

(4.3)

The resistive-dominated steady-state values of crystal and gap voltage are given by:

$$V_{xs} \approx \frac{R_X (V_b + R_L J A_r)}{R_S + R_L + R_E}$$

(4.4)

and

$$V_{gs} \approx \frac{R_L [V_b - (R_S + R_E) J A_r]}{R_S + R_L + R_E}$$

(4.5)

Here $R_E = R_X + R_Q + R_T$. An example of the transient response of $V_g$ and $V_x$ to a step in $V_b$ is presented in fig. 4.4 for the case of $R_L << R_X$. The expected response to steps in both $V_b$ and intensity is illustrated in fig. 4.5 for $R_L > R_X$.

The other two time constants of the four-pole model are associated with the dielectric mirror: $\tau_3 \approx R_Q C_Q \approx 400$ sec for SiO$_2$, and $\tau_4 \approx R_T C_T \approx 10^{-3}$ sec for TiO$_2$. They do not strongly influence the MSLM time response because the very thin layers of the dielectric mirror are characterized by high capacitance and low resistance and thus account for only small voltage drops. In fact, the full solution of Appendix B.1 reveals that the simplified circuit model of fig. 4.2 is sufficient for most purposes, with the identifications: $R_E = R_X + R_Q + R_T$ and $C_E = C_X || C_Q || C_T$. (Of course with no dielectric mirror or other coating on the crystal; $R_E = R_X$ and $C_E = C_X$.) The Laplace frequency response (eqs. (B.1) and (B.2)) and dynamic behavior (eqs. (B.3)) of the
Fig. 4.4 Transient response of the crystal and gap voltages to a step in $V_b$ with $R_L \ll R_x$.
(In the absence of control light, I)
Fig. 4.5 Transient response of the crystal and gap voltages to steps in $V_b$ and $I$ with $R_L > R_x$. The driving function is $V_b U(t) + I[U(t-t_1) - U(t-t_2)]$, where $U(t)$ is the unit step function.
simplified model are also presented in Appendix B.1. It is shown in Appendix B that according to both models the MSLM is always over-damped, characterized by only real decaying time constants.

For typical device parameter \((R_s \ll (R_x, R_L) \ll R_p, (C_x C_g \ll C_Q, C_T)\) the two dominant time constants can be further simplified to:

\[
\tau_1 \approx R_s (C_x | C_g) \approx 10^{-6} \text{ sec} \tag{4.6a}
\]

and

\[
\tau_2 \approx (R_x | R_L) (C_x + C_g) \approx 10^{-3} \text{ sec to } 10^3 \text{ sec} \tag{4.6b}
\]

(depending on \(R_L\))

Since \(\tau_2\) is generally much larger than \(\tau_1\), it is the dominant time constant; and the step responses of \(V_g\) and \(V_x\) can be further approximated by the single pole equation:

\[
V_x = V_{xs} (1 - e^{-t/\tau_2}) + \frac{C_g}{C_g + C_E} V_b e^{-t/\tau_2} \tag{4.7a}
\]

\[
V_g = V_{gs} (1 - e^{-t/\tau_2}) + \left[ \frac{C_E}{C_g + C_E} (V_b - R_s J_A_r) \right] e^{-t/\tau_2} \tag{4.7b}
\]

\(V_{xs}\) and \(V_{gs}\) are given by eqs. (4.4) and (4.5).

4.2.1.2. Continuous Mode

In the continuous mode, the crystal voltage \(V_x\) is proportional to the input irradiance \(I\) for all temporal variations within the system bandwidth, \(1/\tau_2\). The transfer function for \(V_x\) as a function of \(I\) (Fig. 4.5) can be derived from the steady-state relations, Eqs. (4.4) and (4.5), utilizing the latter as a load-line on fig. 4.3. The continuous mode cannot be utilized when \(R_L \ll R_x\); since according to eq. (4.4) the crystal charges to \(V_b\) with no input irradiance \(I\). When \(R_L \gg R_x\, \text{V}_x\) is
essentially the voltage dropped by $i_g = JA_r$ in flowing through the crystal resistance; i.e. $V_x \approx R_x JA_r$. The crystal voltage saturates when $I$ becomes large enough to cause $V_g \approx V_b - V_x$ to drop below $V_{gc}$ in fig. 4.3; unless MCP saturation is encountered first ($J \approx J_{strip}$), then $V_{x_{sat}} \approx R_x J_{strip} A_r$.

In the linear operating region (with $V_g > V_{gc}$), $J = J_P$ of eq. (4.1) and:

$$V_x \approx \left(R_x A_r \frac{enG}{h\nu}\right) I = R_x A_r r_1 I$$

(4.8a)

Since crystal voltage and hence modulator phase is proportional to $I$ it is useful to define a half-wave irradiance:

$$I_\pi = \frac{V_\pi}{(R_x A_r r_1)}$$

(4.8b)

High bandwidth operation calls for a low resistance electro-optic material (e.g. if $1/\tau_2 = 2\pi$ kHz and $C_x = 2 \times 10^{-10}$ F, $R_x = 8 \times 10^5$ $\Omega$) is required.

When $i_g > 0$ in fig. 4.3, the continuous mode relies on deposition of primary electrons; and when $i_g < 0$ (e.g. $V_b < 0$) the gap current $i_g$ is due to secondary emission of electrons from the dielectric mirror (or crystal) surface. The contrast in the two cases is reversed relative to each other; in the latter case $V_x$ decreases as irradiance, $I$, increases.

4.2.1.3. Framed Modes

The framed operating modes, which depend on how electrons are deposited and removed, increase the versatility of the MSLM. The **photoconductive-erasure framed mode** begins, in Figure 4.6, with the application of $V_b$. After approximately $5\tau_1$ seconds $V_b$ has capacitively divided between the gap and crystal, resulting in $V_g \approx V_b$.
Fig. 4.6 The photoconductive-erasure framed mode
and $V_x \approx 0$ (since $C_g \ll C_x$). Then the control or write light is applied for the write period $t_w$. In practice, $V_b$ and $I$ can be applied simultaneously, since $\tau_1 \ll \tau_2$. The image is stored until the frame is ended by turning off $V_b$ and uniformly illuminating a photoconductive coating on the dielectric mirror. The photoconductor decreases $R_L$, allowing rapid charge removal, with time constant $\tau_{2e}$.

Unlike the continuous mode, the framed modes can be utilized when $R_L \ll R_x$. However, the combined write and storage time ($t_2$ in fig. 4.6) should be significantly less than $\tau_2$ from eq. (4.3), which approximates the time constant for relaxation of the electron image. In practical framed systems, $R_x$ and $R_L$ can be very large, storage times of weeks have been observed.

Since the write time is much shorter than the storage time (i.e. $t_w \ll \tau_2$), the approximate time response of eq. (4.7a) can be linearized:

$$V_x \approx V_x(1 - e^{-t_w/\tau_2}) \approx V_x \frac{t_w}{\tau_2} \quad (4.9)$$

Using eq. (4.4) for $V_x$ and eq. (4.6b) for $\tau_2$, with typical parameter values ($JA_x \gg V_b/R_L; R_Q,R_T,R_s \ll R_x,R_L$; and $C_g \ll C_x \ll C_Q,C_T$), results in:

$$V_x \approx \frac{A_r t_w}{C_x} J \quad \text{(for } V_x < V_b) \quad (4.10)$$

Voltage saturation occurs as $V_x$ nears $V_b$, because $J$ in fig. 4.3 decreases when $V_g \approx V_b - V_x$ drops below $V_{gc}$. In practice, the system is designed with $t_w$ short enough to keep $V_x$ far below $V_b$ in order to maintain a proximity
focussing voltage in the gap. (Proximity focussing is discussed in Appendix B.8.) With $V_x < V_b$ and $J < J_{strip}$, $J$ is given by $J_p$ of eq. (4.1), and eq. (4.10) becomes:

$$V_x = \left( \frac{A}{C_x} \frac{r_w}{\eta G} \frac{n e G}{h \nu} \right) I$$

(4.11)

Usually, $t_w$ is a constant; and, in analogy with fig. 4.5 of the continuous mode, $V_x$ is proportional to irradiance.

Equation (4.11) also gives the time required to write a given amount of phase modulation. For example, the time, $t_\pi$, to write the voltage for half-wave phase retardation, $V_\pi$, is:

$$t_\pi \approx \frac{h \nu}{e n G} \cdot \frac{C_x V_\pi}{A r} \cdot \frac{1}{I} = \frac{\sigma_\pi}{r_I} \cdot \frac{1}{I} = \frac{S_\pi}{I}$$

(4.12)

Here $\sigma_\pi = \frac{C_x V_\pi}{A r}$ is the half-wave crystal surface charge density and $S_\pi$ (in Joules/cm$^2$) is the half-wave sensitivity (exposure required to write a phase-shift of $\pi$ radians). For the longitudinal electro-optic effect in $z$-cut LiNbO$_3$, $\sigma_\pi = \varepsilon_0 \varepsilon \lambda / (2n_0^3 r_{13}^2 \lambda_x)$ in reflection; here $\varepsilon_0$ is the permittivity of LiNbO$_3$, $n_0$ is the ordinary refractive index, $r_{13}$ is the electro-optic coefficient, and $\lambda_x$ is the crystal thickness. With $\lambda = 6328 \, \text{Å}$ and $\varepsilon_x = 1$ mm, $\sigma_\pi = 1.6 \times 10^{-7}$ C/cm$^2$.

The secondary-erasure framed mode is characterized by the sequence of events in Fig. 4.7. The write and store cycles are identical to those in Fig. 4.6. However, the erase cycle commences by turning on a uniform erase beam $I_{\text{erase}}$, followed by a controlled decrease or ramping off of $V_b$; during which secondary emission removes the crystal
Fig. 4.7 The secondary-erasure framed mode
charge. The important details of secondary emission are summarized in the \( J \) vs \( V_g \) characteristic of Fig. 4.8.

When the strong \( I_{\text{erase}} \) beam is turned on, a large MCP current rapidly charges the crystal to saturation voltage \((V_x = V_b + V_g = V_b)\) and decreases the gap voltage along the \( I_{\text{erase}} \) trajectory in Fig. 4.8 from \( V_g > V_{g_c} \) to the stable load-line equilibrium at \( V_g = V_{g_1} = 0 \). (The load-line, \( V_g = V_b - R_x J_{A_r} \), from Eq.(4.5), is very flat because \( R_x \) is assumed large.) When \( V_b \) begins to decrease, the load-line moves to

\[ J_{A_r} \approx C_x V_b + \frac{V_b - V_{g_1}}{R_x} \approx C_x \dot{V}_b \]  

(4.13)

This follows from the same approximations used to derive Eq.(4.10), \( (J_{A_r} = C_x \dot{V}_x) \), with the recognition that since \( V_g = 0 \) in equilibrium, \( \dot{V}_x = \dot{V}_b - \dot{V}_g = \dot{V}_b \). (This is a stable equilibrium. When \( V_g > V_{g_2} \), \( J \) is too small to erase \( V_x \) at the rate at which \( V_b \) is falling; so the decrease of \( V_b \) is capacitively coupled into the gap as a decrease of \( V_g \). When \( V_g_3 < V_g < V_{g_2} \), \( J \) is too large and the excess secondary charge increases \( V_g \).) At the end of the erase cycle in Fig. 4.7, \( V_b = 0, \dot{V}_b = 0, V_g \) returns to \( V_{g_0} \), and \( V_x = -V_{g_0} = 0 \).

In the inverted-secondary framed mode of Fig. 4.9, the image is written by secondary emission and erased with primary electrons. The cycle commences with a fully charged crystal \((V_x = V_b - V_{g_1} = V_b, V_g = V_{g_1} = 0 \) in Fig. 4.8). Then, with the write-image irradianc applied, \( V_b \) is ramped down. Much of the image (e.g. \( I_1 \) in Fig. 4.8) may lack sufficient intensity to operate on the \( \dot{V}_b \) load-line; i.e., in Fig. 4.8, \(|\dot{V}_b| > J_{\text{max}} |A_r/C_x| \). As \( V_b \) falls; capacitive coupling diminishes \( V_g \) in
Fig. 4.8 $J$ versus $V_g$ during secondary emission charge removal.

The load-lines are approximately described by

$$J_{Ar} = C_x V_b + (V_b - V_g)/R_x.$$  

$I_{erase} > I_1$ are two write-beam irradiance levels.
Fig. 4.9 The inverted-secondary framed mode.
these regions. Since no electrons have energy greater than the MCP voltage \( V_m \); when \( V_g \) decreases below \(-V_m\), a "lock-out" condition occurs in which electron flow ceases with only partial removal of the \( V_x \) charge. The charge removal is more complete in areas of greater image irradiance. After the storage time, the image is removed by applying \( V_b = V_{b0} \) to recover from "lock-out" \((V_g = V_{g0} + V_{b0} > -V_m)\), and then applying a uniform control beam.

As depicted in Fig. 4.9, the dynamic range can be extended with smaller supply voltage, \(|V_b|\), by employing bipolar \( V_b \). The photoconductive-erasure and secondary-erasure modes can also be operated with bipolar \( V_b \).

An approximate expression for crystal voltage, \( V_x \), as a function of irradiance, \( I \), can be derived for the inverted-secondary framed mode. Using the approximations that lead to Eq. (4.10), \( dV_x \frac{JAR}{C_x} dt \) follows. Integrating both sides, and using \( \dot{V}_x = \dot{V}_b - \dot{V}_g = \frac{JAR}{C_x} \) and \( dt = dV_g / \dot{V}_g \) results in

\[
\int_{V_{x0}}^{V_x} dV_x = \int_0^t \frac{JAR}{C_x} dt = \int_{V_{g1}}^{V_m} \frac{JAR}{C_x} \frac{dV_g}{\dot{V}_g} = \int_{V_{g1}}^{V_m} \frac{JAR}{C_x V_b - JAR} dV_g
\]

(4.14)

As long as \( J < J_{strip} \), \( J \) is proportional to irradiance \( I \); and the \( J \) vs. \( V_g \) relation (e.g. Fig. 4.8) can be formally written \( J = F(V_g)I \). Using \( V'_g = -V_g \), \( V'_g = -V_{g1} \), and \( F(V_g) \); Eq. (4.14) becomes:

\[
V_x = V_{x0} - \left[ \int_{V_{g1}}^{V_m} dV_g \left( 1 - \frac{C_X \dot{V}_b}{JARF(V'_g)} \right)^{-1} \right]
\]

(4.15)

This integral can be evaluated numerically using a \( J \) vs. \( V_g \) characteristic such as Fig. 4.8. Alternatively, an explicit expression for \( F(V_g) \) is
derived in Appendix B.2.4; and the resulting integral is evaluated.

From Eqs. B.52 and B.54, \( V_x \) can be approximated by:

\[
V_x \approx V_x^0 + \bar{V}_a \ln(1 - aI) \quad (4.16)
\]

where

\[
a = \frac{F(V'_g p)_{A_F}}{C_x \bar{V}_b} = \frac{1}{\delta_x} F(V'_g p) \approx \frac{n e G}{\delta_x \hbar \nu} f_H(1 - \delta_H) e^{-V'_g p / \bar{V}_a} \quad (4.17)
\]

Here \( f_H \) is the fraction of MCP electrons with a Maxwellian energy (e\( V_a \))
distribution (i.e., \( f_H \left( \frac{V_a}{\bar{V}_a} \right) \)); \( \delta_H \), from Eq. (R.30), is the secondary
emission coefficient of these electrons; and e\( V_a \) is their average energy.
e\( V_{m} \) is the maximum ejection energy, and \( V_g p \) is defined in Fig. 4.8.

Figure 4.10a depicts this \( V_x \) versus I transfer characteristic
for a small dynamic range of modulator voltage (\( \Delta V_x = 4\bar{V}_a \)). When small
\( \Delta V_x \) is sufficient (\( \Delta V_x < \bar{V}_a \approx 60 \) volts) only low intensities are required,
i.e., \( aI \ll 1 \) in Eq. (4.16). In that instance Eq. (4.15) approximates to
\( V_x = V_x^0 - \bar{V}_a aI \) and the inverted-secondary framed mode becomes linear in I.

On the other hand, many electro-optic materials require larger
voltages (e.g., \( V_\pi = 3kV \) for LiNbO\(_3\)). The resulting transfer characteristic
in Fig. 4.10b is characterized by an extremely large gain with intensity
threshold and saturation, which is essentially "hard clipping" at
threshold intensity:

\[
I_t = \frac{1}{a} = \frac{C_x}{A_F(V'_g p) / \bar{V}_b} \quad (4.18)
\]

(During secondary emission charge removal \( \dot{V}_b < 0 \) and \( F(V'_g p) < 0 \); hence \( I_t > 0 \).

Another interpretation of Eq. (4.18) is that, if the irradiance
level is fixed at I, the maximum erasure rate is
\( \dot{V}_b \leq V_{b_{\text{max}}} = \frac{A_F(V_g p)}{C_x} I_t \).
\[ I(\text{irradiance}) \]

a) With a small dynamic range, \( \Delta V_x \), in modulator voltage (\( V_a = 60 \) volts)

\[ I(\text{irradiance}) \]

b) With a large dynamic range, \( \Delta V_x \), in modulator voltage ("hard clipping" at \( I_t \))

\[ I(\text{irradiance}) \]

Fig. 4.10 Inverted-secondary crystal voltage, \( V_x \), as a function of irradiance, \( I \).

\[ I_t = \frac{C_x}{A_{rF}} \dot{V}_h \]
Since $J=FI$, this is consistent with $C_x \frac{V}{V_{bm}} < J_{\text{max}} A_r$ in Fig. 4.8. The corresponding best secondary electron removal time is:

$$t_s = \frac{\Delta V_x}{V_{bm}} = \frac{C_x \Delta V_x}{F(V_g')A_r I} = \frac{\Delta \phi_m t}{\pi} \frac{V_g' V_a'}{\dot{V}}$$

(4.19)

The above expression defines the half-wave charge removal time, $t_s$. Using Eqs. (B.50) of Appendix B.2 for $F$:

$$t_s = \frac{C_x V_{\pi}}{A_r I} \frac{V_g' V_a'}{\dot{V}} = \frac{t_{\pi}}{e \pi G_f H(1-\delta_H)} \frac{V_g' V_a'}{\dot{V}}$$

(4.20)

Here $t_{\pi}$ from Eq.(4.12) is the half-wave write time for electron deposition. The corresponding half-wave sensitivity is:

$$S_{s\pi} = \frac{V_g' V_a'}{t_{s\pi} I} = \frac{e \pi G_f H(1-\delta_H)}{f_H(1-\delta_H)}$$

(4.21)

where $S_{s\pi}$ from Eq.(4.12) is the electron deposition half-wave sensitivity.

The choice of which operating mode actually to employ is a function of the application and available materials. The continuous mode is only feasible when an electro-optic material with a low bulk resistivity is available. Suitable materials exist and materials such as Bi$_{12}$SiO$_{20}$, which are photoconductive and insensitive to red light, could possibly be biased to the appropriate conductivity with a uniform blue light source. Conversely, the framed modes demand that the resistivity be high enough to store the image during the write time. In the following discussions the photoconductive-erasure and secondary-erasure framed modes will often be referred to as the electron-deposition
framed modes, since the image is written by depositing electrons. This is in contrast to the inverted-secondary framed mode, where the image is written by removing electrons. The modes which employ secondary emission perform best with MSLM hardware configurations which include a grid between the MCP and crystal.

4.2.2 Experimental MSLM systems

Most of the important operating characteristics of the MSLM have been experimentally measured. A detailed discussion of these measurements is too much of a diversion to be included here, however, a few of the measured results are presented in the performance discussion of the next section. Additional details of this MSLM experimental development work are reported in other documents 80,81,211,212 and Appendix B.7.

The majority of MSLM development was carried out with MSLMs operated in a demountable metal vacuum system, which is diagrammed in Fig. 4.11. A variety of electro-optic crystals of various sizes, polish quality, electron surface coatings (SiO₂+TiO₂ dielectric mirror, Ba+BaO, Si+SiO+SiO₂, high resistance epoxy, and uncoated), and material (LiTaO₃, LiNbO₃, and KD³P) were employed. Besides crystals, electron targets of graphite and Al were utilized to study MSLM electron dynamics; and a phosphor screen was used to examine imaging and proximity focusing properties. Photocathodes of CsI, Inconel, Ba,BaO, and Au were also employed.

Most of the greatest improvements in performance were realized with the vacuum-sealed MSLM illustrated in Fig. 4.12. This device employed a parallel LiTaO₃ crystal (0.5mm thick x 20mm diameter) with-
Fig. 4.11 Demountable MSLM
out a dielectric mirror, and a bialkali photocathode sensitive to visible light. Due to the difficulty of photocathode fabrication, this MSLM was constructed in collaboration with Hamamatsu Corporation; using the electron optics of a modified image intensifier in place of proximity focusing between the photocathode and MCP. This researcher was responsible for the basic design specifications and all of the initial evaluation measurements on this device. Employing a flat parallel crystal supplied by us, Hamamatsu Corporation fabricated the photocathode and assembled the device.

This MSLM contained an electron permeable Al barrier on the MCP input to prevent feedback of ions from contaminating the photocathode. At the expense of some gain reduction, this opaque barrier also served to isolate the photocathode from the readout light, and to prevent contamination of the crystal surface by conductive photocathode materials (e.g. Cs).

4.2.3 MSLM Performance: Theory, Measurements, Enhancement

The theoretical and experimental investigations of the MSLM have resulted in a good understanding of its limitations and the improvements which are required to realize its ultimate performance potential.

4.2.3.1 Sensitivity

Due to its large electron gain and its ability to integrate charge, the MSLM is potentially capable of achieving the high photon sensitivities of image intensifiers. A single MCP can provide spatially uniform (±5% variation) electron gains of $10^4$, or two MCP's can be cascaded for a gain of $10^7$. The reader is referred to Appendix B.5 for a more careful discussion of the sensitivity measures used here.
Fig. 4.12 Sealed visible-photocathode MSLM
As found in section 3.1.9, the minimum usable irradiance is generally a function of the signal-to-noise ratio and other performance requirements of the application, such as speed or dynamic range. In the framed modes, the minimum irradiance \( I_f \) required to achieve a given dynamic range \( \Delta \phi_m \) and write-time \( t_w \) (or frame-rate), or phase-modulation frequency \( w_m = \Delta \phi_m / t_w \), can be expressed in terms of the half-wave sensitivity. Since \( \Delta \phi_m = \pi I_f t_w / S \), the minimum required irradiance is

\[
I > I_f = \frac{S}{\pi} \frac{\Delta \phi_m}{t_w}
\]  

(4.22a)

When secondary emission is employed, \( S_{se} \) of Eq. (4.21) should be used instead of \( S \).

Since phase is directly proportional to irradiance in the continuous mode [e.g., Eq.(4.8a)], it appears that the nominal irradiance, \( I_c \), depends on only dynamic range, \( \Delta \phi_m \). In actuality, there is also an implicit response speed specification, since the effective risetime (10–90%) of the continuous mode is \( t_w = 2.2 \tau_2 = 2.2 R_x C_x \). Based on Eqs.(4.8), (4.22), and (P.68),

\[
\Delta \phi_m = \pi I_c = \frac{\pi I_c t_w}{2.2} = \frac{\pi I_c}{S} \frac{\pi}{S} \frac{t_w}{w_m}
\]  

(4.22b)

where \( w_0 = 1/\tau_2 \) is the 3db bandwidth of the continuous mode MSLM. The required irradiance is

\[
I > I_c = 2.2 \frac{S}{\pi} \frac{\Delta \phi_m}{t_w} = 2.2 I_f = \frac{S}{\pi} \frac{w_0 \Delta \phi_m}{\pi}
\]  

(4.22c)

For a given dynamic range and speed, the continuous mode may thus require about twice as much irradiance as the electron-deposition framed modes.

Eq. (8.65) shows that \( S = h \nu \sigma_{\pi} / (n e G) \), hence the minimum required
irradiance decreases as the MCP gain, $G$, increases. In high-resolution applications noise often limits the maximum usable gain and there is an optimum gain (and hence $S_\pi$) which minimizes the intensity required to simultaneously satisfy the noise and other performance constraints. In lower resolution applications the maximum possible $G$ may be limited by the available MCP(s) to values smaller than the optimum $G$, then Eqs.(4.22a) or (4.22c) directly give the minimum required irradiance.

The primary noise sources in the MSLM are signal and dark current induced shot noise. Based on Eqs. (3.61b), (3.62b), and (3.64), with $g' = \frac{g_0 g_1 g_2}{S_\pi w_\pi}$ (framed mode) or $\pi/(S_\pi w_0)$ (continuous mode), the noise-induced phase variance of the open-loop continuous-mode MSLM is

$$\sigma^2_\phi = g'^2 \frac{\hbar \nu w_0}{2 A_r n} (I + \bar{I}_d) = \frac{1}{2} \left( \frac{\pi G}{A_r \sigma_\pi} \right)^2 (n_w + n_{wd})$$

(4.23a)

where

$$n_w + n_{wd} = \frac{n A}{\hbar \nu w_0} (I + \bar{I}_d)$$

(4.23b)

Here $\bar{I}_d$ is the effective input irradiance or "noise equivalent power" ($\text{NEP} = \bar{I}_d = 10^{-15} \text{W/cm}^2$) corresponding to the MCP dark current. For the framed modes $w_0$ in Eqs. (4.23) should be replaced by $2t_w$.

Eqs. (4.22a and c), which take the form $I > \Delta \phi_m/g'$, in conjunction with specifying the maximum tolerable rms fluctuations due to noise, i.e. $\sigma^2_\phi < \sigma^2$, results in an optimal value of $g'$:

$$g' = \frac{n A}{\hbar \nu w_0} \sqrt{\frac{\Delta \phi_m}{2} + \frac{2m' n_{wd}}{2}}$$

(4.24a)

If $g'$ is smaller, greater irradiances are required to meet the speed and dynamic range constraints; if $g'$ is larger the noise induced fluctuations become excessive. Eq. (4.24a) also specifies the optimal MCP gain as
or exposure sensitivity as $S_{\pi} = \pi / \omega_0 w_0$, (replace $w_0$ by $1/t_w$ for framed modes). The associated minimum irradiance satisfies

$$\sigma^2 = \frac{\Delta \phi_m^2}{2} \frac{n_w + n_{wd}}{n_w^2} > m'$$

or more explicitly

$$n_w > \frac{\Delta \phi_m^2}{2m'} \left( \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{2m'}{\Delta \phi_m^2} n_{wd}} \right)$$  \hfill (4.24b)$$

When $n_{wd} \ll 1$ for the bandwidth and resolution of the application, signal shot noise dominates and Eq. (4.24b) becomes $n_w > \Delta \phi_m^2 / 2m'$. With $n_{wd} \gg 1$, $n_w > \sqrt{n_{wd} \Delta \phi_m^2 / 2m'}$ is required; it should be noted that when $n_{wd} \gg 1$, $\sqrt{n_{wd}} \ll n_{wd}$ and hence the MSLM can operate with $I$ considerably below $I_d$.

When the MSLM is employed in a closed-loop homodyne IPL, the minimum irradiances and optimal values of $q'$ developed in section 3.1.9 should be employed.

As a specific numeric example, with $\sigma < 0.5$ radian ($m' = 0.25$) and $\Delta \phi_m = \pi$, Eq. (4.24b) shows that $n_w > 2\pi^2 (0.5 + \sqrt{0.25 + n_{wd} / 2m'})$ is required. Assuming 20 lpi/mm resolution, $h \nu = 2$ eV, $\omega_0 = 2\pi$ kHz, and $I_d = 10^{-15}$ W/cm²; results in $n_{wd} = 1.2 \times 10^{-5}$ which is negligible and hence $n_w > 2\pi^2 = 20$ photons. The corresponding optimal $S_{\pi}$ is 2.5 nJ/cm² and the optimal gain is $G = 1.3 \times 10^6$ (assuming $\sigma = 1.6 \times 10^{-7}$ C/cm²).

Under similar conditions, Eq. (3.87c) (with $m = 4$ and $G_m = \pi$) reveals that a closed-loop homodyne IPL employing the MSLM is more sensitive, requiring only $n_{wi} > \pi / 2 = 1.5$ photons. The optimal $S_{\pi}$ is $S_{\pi} = \sqrt{n_{wi}} \times 10^{-13}$ J/cm² and $G = 3.1 \times 10^7 \sqrt{n_{wi}}$. (For an L.O. producing $n_w = 20$ photons, $S_{\pi} = 0.45$ pJ/cm² and $G = 7 \times 10^6$)

An MSLM employing a single MCP is hardware limited to $G = 10^4$ and $S_{\pi} = 0.3$ nJ/cm², and hence requires significantly more irradiance than the
minimum values in the preceding examples. On the other hand, an MSLM with
two MCPs in the above examples should be operated with less than its
maximum gain \((G=10^7, S_\pi=0.3\text{pJ/cm}^2)\) for best results. In some lower
resolution applications it may be advantageous to employ a third MCP;
for example with 210\text{nm/mm} resolution, a gain of \(G=1.3\times10^8\) is required to
obtain maximum open-loop sensitivity.

**Experimental Sensitivity**

The early MSLMs\(^7_8\text{--}^{80}\) were insensitive due to the incompatibility
of high quantum efficiency photocathodes (e.g. Cs activated) with
demountable metal vacuum systems\(^2_{15,216}\) (pressure of \(10^{-6}\) to \(10^{-7}\) torr)
A CsI photocathode, \((\eta=10^{-8} \text{ with } \lambda=2537 \text{ Å, } \eta=0 \text{ with visible } \lambda)\) was
employed initially, since it could tolerate periodic atmospheric
exposure. Inconel and Au resulted in some improvement; and Ba and BaO
photocathodes increased the UV quantum efficiency by \(10^3\).

Since a visible photocathode is critical to most closed-loop
adaptive optical applications, significant effort was expended in that
direction. The sealed visible-photocathode MSLM illustrated in Fig. 4.12
was constructed in collaboration with Hamamatsu Corporation. In-house
visible photocathode fabrication is extremely difficult; but a system to
fabricate photocathodes and incorporate them into an indium-sealed MSLM
was also designed. R. Dillon refined this design and is currently in the
process of constructing a photocathode transfer system.

The bialkalai photocathode of the Hamamatsu MSLM had a
quantum efficiency of \(\eta=6\% \text{ at } \lambda=655 \text{ nm (LED); the corresponding half-}
wave sensitivity was } S_\pi=2.2 \text{ nJ/cm}^2 \text{ with } V_{\text{MCP}}=1500\text{V. This is probably }
the best sensitivity every achieved in an optically-addressed real-time
SLM.
4.2.3.2. Time Response

The half-wave rise time, or time required to add or remove enough charge to change the modulation phase by \( \pi \) radians, is reviewed below for the various modes. (\( I_w \) is the control-beam irradiance). For primary \( e^- \) deposition, from Eq.(4.12):

\[
t_{\pi} = \left( \frac{h \nu}{e \eta G} \cdot \frac{C^* V_\pi}{A^*} \right) \cdot \frac{1}{I_w} \cdot \frac{S_{\pi}}{I_w} \tag{4.25a}
\]

For secondary emission \( e^- \) removal, from Eq.(4.20):

\[
t_{S_{\pi}} > \frac{S_{\pi} \exp(V'_a/V)}{f_H(1-\delta_H)} \cdot \frac{1}{I_w} \tag{4.25b}
\]

(larger when \( V_b < V_{b_{\text{max}}} \) is chosen).

In the continuous mode, from Eq.(4.24)

\[
t_{C_{\pi}} > 2.2 S_{\pi} \cdot \frac{1}{I_w} \tag{4.25c}
\]

(larger when \( \Delta \phi_m > \pi \) is possible at \( I_w \); i.e. \( R_x > \frac{V_{\pi}}{r_A I_w} \)).

The overall time response depends on the operational details.

In terms of "bandwidth", \( BW \), or the maximum control beam frequency at which the MSLM can operate without serious performance degradation, the framed cycle-time should be \( \leq 1/2BW \) (quasi Nyquist criterion); and in the continuous mode \( \tau_c > \frac{1}{2 \pi BW} \) (3db criterion).

Since the write time is inversely proportional to the control-beam irradiance in Eqs.(4.25) the time response is usually limited by either the maximum available irradiance or the saturation level of the MCP emission current. Kilohertz framed operation requires an MCP emission current density of \( J_p = 6 \times 10^{-4} A/cm^2 \) (LiNbO_3, \( x = 0.5 \) mm, \( t_{\pi} = 0.5 \) msec).
Although steady-state $J_p$ cannot exceed the MCP strip current density $J_{\text{strip}}$, flowing in the pore walls (typically $J_{\text{strip}} = 1\mu A/cm^2$ with $V_{\text{mcp}} = 1kV$), limited-charge MCP pulses approaching $1A/cm^2$ have been reported.\textsuperscript{218} $J_{\text{strip}}$ can be increased by operating at higher bias voltage $V_m$; however, ion feedback noise becomes serious. MCP's can also be constructed to have enhanced strip current; often at the expense of reduced gain. For example, the deposition of Ba in the MCP pores doubled $J_{\text{strip}}$ and reduced the gain by over an order of magnitude. MCP's with strip currents of greater than 100 $\mu A$ at 1kV are expected to be commercially available soon.\textsuperscript{219} Future high speed MSLM's will probably employ two MCPs, a normal MCP as a preamplifier and a high strip current MCP as a power amplifier.

A variety of photoconductors including Se, PVK:TNF, CdS, Bi$_{12}$SiO$_{20}$, Cd$_{1-x}$Zn$_x$S, and a few organic materials were considered for charge removal in the photoconductive mode. The ideal photoconductor should possess: large dark resistance $\rho_D$ for image storage, a high dark-to-light resistance ratio $\rho_D/\rho_L$, much greater sensitivity to the erase wavelength (e.g. blue or UV) than the control and read beams, fast photoconductive initiation and recovery, good vacuum characteristics, high breakdown voltage, and stability under electron bombardment. Photoconductors with $\rho_D$ from $10^5$ $\Omega$ cm to $10^{15}$ $\Omega$ cm, $\rho_D/\rho_L$ up to $10^5$, and msec initiation and recovery times exist.\textsuperscript{203} In a typical application ($C_x = 10^{-10} F$, $\rho_D = 5 \times 10^8 \Omega$ cm, $\rho_D/\rho_L = 10^4$), a 50$\mu$m thick layer of photoconductor would result in an erase time constant of $\tau_{2e} = 10^{-3}$ sec and storage time of $\tau_2 = 10$ sec. A photoconductor sensitive to the control and readout illumination can be employed if the MCP is coated
with an opaque membrane (e.g. Hamamatsu MSLM), and (as suggested by A. Weiss) the dielectric mirror blocks the readout beam but transmits the erase wavelength.

In the secondary-erasure and inverted-secondary modes, the charge removal times have tended to be long due to the limited number of high energy MCP electrons. Increasing the MCP bias, $V_m$, has resulted in some increase of electron energy; and Appendix B.2.3 suggests that interposing an acceleration grid between the MCP and crystal should greatly improve the secondary time response. When sensitivity is not critical, direct use of an electrostatic lens (e.g. in Hamamatsu MSLM), without an MCP, can provide a similar optimization of the electron energy.

The secondary emission characteristics also depend on the details of the electron surface (e.g. crystal or dielectric mirror) condition and preparation. Significant differences were observed between different samples of the same material (e.g. LiTaO$_3$) and even between the removal of intrinsic ($V_x < 0$) and deposited ($V_x > 0$) electrons. Reproducibility, spatial uniformity, and overall secondary emission efficiency might be further enhanced by the deposition of an efficient secondary emission layer (e.g. BaO or MgO) on the dielectric mirror (or crystal).

Secondary emission is potentially faster than photoconductivity for $e^-$ removal, due to the initiation and recovery times associated with photoconductivity. In addition, inserting the acceleration grid required for effective utilization of secondary emission should be more straightforward than finding and uniformly applying a suitable photoconductive
coating. It should ultimately be possible to obtain a net gain in secondary current, resulting in faster electron removal than deposition.

Image storage applications demand large values of \( \tau_2 \) \([= (R_x || R_L) C_x \) from Eq. (4.3)]. Uncoated LiNbO\(_3\) and LiTaO\(_3\) have very large surface \( R_L \) and bulk \( R_x \) resistivities \((-10^{16} \Omega \text{cm})\). The lateral resistance for charge leakage is approximately the surface sheet resistance; i.e., \( R_L = p/t \), where \( t \) is the effective layer thickness. With multiple layers on the crystal surface (e.g. dielectric mirror), the layer with the smallest value of \( p/t \) generally dominates, since the layer to layer resistance \( (p/t A_r) \) is usually small. For example, with a TiO\(_2\)/SiO\(_2\) dielectric mirror the TiO\(_2\) layers dominate, with \( R_L = 1.5 \times 10^{13} \Omega \) corresponding to \( \tau_2 = 2000 \) sec.

Experimental Time Response Results

The fastest measured primary mode write rate in the visible-
photocathode MSLM was \( \dot{V}_x \) = 20KV/sec corresponding to \( t_\pi = 400 \) msec \((V_\pi = 8kV, J_p = 1.8 \mu A/cm^2, V_m = 1.5kV, \) irradiance = 16 nW/cm\(^2\)). The associated secondary emission charge removal time was \( t_{s\pi} = 2.2 \) minutes (in Eq. (4.25b), \( f_H(1-\epsilon_H) = .003 \)). For fear of damaging the MCP, no attempt was made to increase \( J_p \) further. Employing a thicker crystal \((x = 4\) mm\) in the demountable system to reduce \( C_x \) resulted in \( t_\pi = 100 \) msec, limited by maximum UV write irradiance). This crystal also exhibited improved secondary emission characteristics with \( f_H(1-\epsilon_H) = .02 \) and \( t_{s\pi} = 13 \) sec \((J_p = 0.2\mu A/cm^2, V_\pi = 5kV)\). A. Weiss recently obtained \( t_\pi = 3\) msec and \( t_{s\pi} = 100 \) msec \((V_m = 1.9kV \) and \( i_{\text{strip}} = 60\mu A)\).

Resistive erasure experiments were also conducted, in which semi-
conducting coatings consisting of Si+SiO₂ or Ba+BaO were deposited on the
dielectric mirror, resulting in charge leakage time of $\tau_{2e} \approx 100$ msec.
Although not yet employed in the MSLM, photoconductive layers of PVK:TNF
and amorphous Se were fabricated by R. DeCesaris. The Se layers
were very promising exhibiting $\rho_D = 2 \times 10^9 \Omega \text{cm}$, $\rho_D/\rho_L = 10^3$, and msec photo-
conductive initiation and recovery times.

Continuous mode experiments were attempted in the demountable
system with a KD*P crystal exhibiting $t_{cm} \approx 2.2 \tau_2 = 1.3$ sec. Faster
operation is expected with other electro-optic materials such as DCDA,
(with $\rho = 10^{10} \Omega \text{cm}$, $\varepsilon_r = 60$, $\tau_2 \approx \tau_e \approx 0.05$ sec, $V_\pi = 2$ kV) or uniformly illuminated
$\text{Bi}_2\text{SiO}_3$, which is photoconductive.

Images stored in the sealed MSLM were observed over periods as
long as 13 days, with hardly any image degradation, implying
$R_x \gg R_L \gg 6 \times 10^{15} \Omega$, or $\rho_x \gg 10^{17} \Omega \text{cm}$. These resistivities are somewhat larger
than expected and may be due to high energy electrons becoming bound
into long-lived trap states. Other charge decay experiments suggested
$\rho_x = 10^{15} \Omega \text{cm}$ in LiNbO₃.

4.2.3.3. Phase modulation dynamic range

A large (multiple $\pi$ radian) phase modulation dynamic range is
desirable in adaptive phase compensation applications. In the framed
modes the dynamic range increases with the bias voltage swing, $\Delta V_b$, until
dielectric breakdown of the crystal occurs. In the continuous
mode, the maximum modulator voltage is limited by the maximum available
MCP current \( \Delta V_x < R_x A_r J_{strip} \). The 0.75 mm thick crystal of \( L_i NbO_3 \) \( (V_\pi = 3kV) \), used in many demountable MSLM experiments, is expected to be breakdown limited to 37kV (which corresponds to a crystal leakage current density of \( <10^{-11} \) A/cm\(^2\)). Breakdown of dielectric mirror layers or other very thin coatings should not be a limiting factor, since they drop only small voltages.

In practice, field-emission breakdown of the MCP-to-crystal gap has been the major limitation. Strong point-emission currents begin to appear around \( V_g = \pm 5kV \) depending on the MCP and crystal surface conditions and the gap width, \( \varepsilon_g \). This effect limited the dynamic range of the visible photocathode MSLM to about \( \pi \) radians \((-3kV < V_g < 5kV, \varepsilon_g = 0.5mm, V_\pi = 8kV)\). Unfortunately, proximity focusing (see Appendix B.8) demands a narrow gap. Continued operation at large \( V_g \) "burns" away many of the emission points, however, the large current densities can damage the MCP or internal MSLM wiring. Twice as much dynamic range for a given maximum \( |V_g| \) can be obtained in the framed modes by operating between negative and positive bias voltages, \( V_b \).

Further improvements in dynamic range can be obtained by employing other electro-optic materials, e.g.: \( V_\pi = 225 \) for KD*P at \( T = -51 \) °C, \( V_\pi = 2kV \) for DCDA, \( V_\pi = 1.15kV \) for \( LiNbO_3 \) cut at \( 55^\circ \) to the \( z \) axis, and \( V_\pi = 530V \) for some PLZT strain-biased ceramics (<3% scattering, 0.4 msec switching time). The use of a shorter wavelength, e.g. the 4420Å line of an HeCd laser would also increase the available depth of phase modulation.
4.2.3.4. Resolution and Optical Quality of Images

The spatial resolution of the MSLM is usually limited by fringing fields in the electro-optic crystal. Roach\textsuperscript{223} has shown that the crystal resolution $N_x$ (line-pairs/mm) is determined primarily by the crystal thickness $\ell_x$, and the ratio of its dielectric constant along the field $\varepsilon_n$ to that perpendicular to the field $\varepsilon_\perp$, i.e. (for \~10% contrast)

$$N_x \approx \frac{1}{\ell_x} (\varepsilon_n/\varepsilon_\perp)^{1/2}$$

Thus LiTaO$_3$ ($\varepsilon_n/\varepsilon_\perp)^{1/2} = 0.94$, $N_x = 1.9$ l.p/mm @ $\ell_x = 0.5$ mm) offers a small improvement over LiNbO$_3$ ($\varepsilon_n/\varepsilon_\perp)^{1/2} = 0.64$, $N_x = 1.3$ l.p/mm @ $\ell_x = 0.5$ mm). Other materials such as cooled KD*P ($\varepsilon_n/\varepsilon_\perp)^{1/2} = 3.2$ @ -51 °C, $N_x = 6.4$ l.p/mm @ $\ell_x = 0.5$ mm) are expected to offer resolution comparable to that of scanning electron beam SLMs.\textsuperscript{76} Future MSLMs are expected\textsuperscript{221,228} to employ crystals as thin as 0.1mm to obtain resolutions of \~10 l.p/mm. The crystal resolution is not expected to exceed the ultimate limit imposed by single MCP pore spacing(50 l.p/mm for 10μ pores); but two cascaded MCP's may limit resolution (\~25 l.p/m).

As discussed in Appendix 8.6, proximity focusing does not degrade resolution as long as a minimal gap voltage $V_{gm}$ is always maintained. This is implemented by limiting the exposure in the e$^-$ deposition framed modes or the maximum MCP current in the continuous mode so that the crystal does not charge fully, i.e., $V_x \leq V_b - V_{gm}$. The required gap voltage can be decreased by shortening the gap width, $\ell_g$. However, the minimum $\ell_g$ is limited by field emission breakdown of the gap, and by the requirement $C_g \ll C_x$ (e.g. for LiNbO$_3$ with $\varepsilon_n = 30$ and $\ell_x = 0.5$ mm, $C_g \ll C_x$ implies $\ell_g > 0.14$ mm). Appendix 8.6 suggests that in the visible-photocathode
MSLM, ($\lambda = 0.5 \text{mm}$), $V_{gm} > 113$ volt is required for 5 lp/mm resolution, which exceeds the crystal resolution of 2 lp/mm. (For 20 lp/mm $V_{gm} = 1800$ volts is required.) As mentioned in appendix B.2.3, the inclusion of a grid between the MCP and crystal allows the crystal to fully charge without degradation of proximity focusing.

The maintenance of a minimal longitudinal gap field also inhibits relaxation of the deposited charge due to lateral fields on the crystal surface.

When the image is written by secondary emission in the inverted-secondary framed mode, $V_g$ becomes negative (e.g., Fig. 4.9); but useful resolution can still be obtained. This is due to a few major effects: (1) MCP electrons with high enough energy to overcome the negative gap field tend to be well collimated. (2) As electrons are removed from a region on the crystal, that area becomes positive relative to surrounding charged regions and tends to receive the MCP electrons ejected at it (positive feedback). Conversely in the $e^-$ deposition modes, (e.g. primary framed), as electrons collect additional MCP electrons tend to be repelled to surrounding regions. (c) The positive charge may be less mobile and hence laterally relax less than the the excess electrons in the $e^-$ deposition modes. (4) The reverse field ($V_g < 0$) prevents redistribution of secondaries to other regions on the crystal; redistribution is a problem with forward field focusing.

**Optical Quality**

Spatially uniform photocathodes and microchannel plates with less than $\pm 5\%$ variation are readily obtained. However, the optical quality has been severly limited by the extreme difficulty of fabricating
optically flat (<\lambda/5 variation/inch) electro-optical crystals. The quest for a flat crystal has significantly delayed MSLM development. (A perfectly flat crystal is not required in the IPL application, since the IPL can self-compensate crystal distortions which do not exceed the spatial resolution of the modulator or 2π radians of dynamic range (in order to avoid 2π ambiguities). Until recently, however, none of the available crystals came close to meeting these constraints.)

Another source of spatial nonuniformity is flexing of the crystal and MCP, due to electrostatic forces induced by \( V_b \). Crystal flexing can be greatly reduced by placing the transparent \( V_b \) electrode on the readout window rather than the crystal surface, or actually gluing the crystal to the readout window. (High vacuum glues may exist.)

Although crystal flexing can be a serious problem in phase modulation applications, it is not important in birefringent amplitude or Fabry-Perot intensity modulation applications.

A wedged crystal or a dielectric mirror is necessary in phase modulation applications to remove the strong intensity modulation (see Appendix B.3) due to Fabry-Perot interference between the reflections from the two crystal surfaces. An optically flat readout window is also required.

Modulator quality is also a function of the spatial uniformity of any coatings deposited on the electron surface of the crystal. A major aspect of future MSLM development will be the investigation of surfaces to receive the electron image, which have more suitable dielectric, secondary emissive, resistive, and/or photoconductive properties. An amorphous semiconductor layer may give the the MSLM the ability to threshold the control image in the framed modes.
Fig. 4.13 Imaging with the visible-photocathode MSLM

(a) Uniformly discharged to $V_x = -2\text{kV}$
(b) Uniformly charged to $V_x = 5\text{kV}$
(c) "MIT" written by electron deposition
(d) Vertical bar written by inverted-secondary framed mode
(e) "MIT MSLM" written by electron deposition
Experimental Images

The visible-photocathode MSLM exhibited a spatially uniform resolution of ~2 lp/mm, which is the expected crystal limited resolution. To expedite fabrication, an uncoated parallel crystal was employed. Figures 4.13a, b show the uniformly discharged ($V_x = -2kV$) and uniformly charged ($V_x = 5kV$) crystal. The fine fringe pattern is interference between the MCP and crystal reflections. The large concentric fringe, corresponding to interference between the reflections from the front and back crystal surface, provides an intensity readout of the crystal phase. Figure 4.13c shows the word "MIT" written by an electron deposition framed mode ($V_b = -2kV$, $V_b = 5kV$, $\Delta V_x$ of image = 3kV). The line widths of the pattern at the crystal face are about 0.45 mm, which is about the width of the imaged electron distribution. The M and I are separated by about 0.8 mm at their closest approach. A vertical bar written by the inverted-secondary framed mode is illustrated in Fig. 4.13c. A more detailed $e^-$ deposition pattern recently written by J. Thackara, is illustrated in Fig. 4.13d. Additional images are displayed in appendix B.5.

4.2.4 MSLM Summary

The MSLM has been shown to be capable of at least three distinct modes of operation, the continuous mode, the electron-deposition framed modes, and the inverted-secondary framed mode. The choice of operating mode is dependent upon the application and specific fabrication details, such as the choice of electro-optic crystal or employment of a grid between the MCP and crystal. A given hardware configuration generally does not exhibit all three operating modes. The interested reader is referred to appendix B for additional operational details, such as specialized image processing functions which the MSLM is inherently
capable of performing.

This research program has resulted in significant improvements in MSLM performance, bringing the MSLM from a preliminary prototype stage to a device usable for a variety of applications. A sealed MSLM with a sensitive visible photocathode (n = 6% at 655 nm), write times of hundreds of msec, storage times of weeks, and crystal limited resolution of 2 lp/mm, (corresponding to -800 resolving elements over the 16 mm diameter useable crystal area) is available. The secondary emission charge-removal time of this device is very long, with t_s > 2.2 minutes. Faster demountable MSLMs have been demonstrated with write and erase times of ~200 msec. Another student has recently demonstrated secondary-erasure framed mode cycle times of 50 msec, by employing an acceleration grid between the MCP and crystal.

The MSLM is ultimately expected to be capable of: better than msec cycle times, with the major speed limitation being the maximum available MCP current; a fringing-field limited resolution of better than 20 lp/mm; quantum limited sensitivities of a few photons per resolution cell in a write-time interval; and a dynamic range in excess of 5π radians, constrained primarily by high voltage breakdown effects.

4.3. MSLM-Based High-Resolution Adaptive Systems

As mentioned previously, an "all-optical" configuration, employing only optical components and an SLM (e.g. fig.2.1), appears to be the ideal approach for implementing high-resolution adaptive systems. Having obtained an SLM with good basic resolution, sensitivity, speed, and dynamic range characteristics; the system designer is faced with the additional challenge of making the adaptive system function, given the fixed properties of the
modulator. Unlike a discrete channel wired system, where the properties can be electronically tailored, only very simple spatially uniform operations can be implemented in an SLM-based system.

This section deals mostly with the specific operational characteristics of the MSLM; some of these considerations will also be applicable to other SLMs. It will be seen that the major limitation of the MSLM is not its intrinsic filter transfer function but rather that its operating modes do not all possess an adjustable intensity threshold; and, as was suggested previously, intensity threshold or offset can significantly enhance IPL performance.

4.3.1. Continuous Mode MSLM-based IPL

The continuous mode MSLM (and most other SLMs) lack intensity threshold; this may change in the future. The continuous mode MSLM-based homodyne IPL is characterized by:

\[ \phi_m = w_0 (-\phi_m + G_0 + G_1 \sin \phi_e) \]  \hspace{1cm} (4.27)

where

\[ G_0 = \frac{g^1}{2Z_0} (|E_1|^2 + |E_2|^2) \]  \hspace{1cm} (4.28a)

\[ G_1 = \frac{g^1}{Z_0} |E_1| |E_2| \]  \hspace{1cm} (4.28b)

\[ g^1 = \frac{\pi}{V \pi} V_x = \frac{\pi}{V \pi} R_x A_r \frac{neG}{h \nu} = \frac{\pi}{S \pi w_0} \]  \hspace{1cm} (4.29)
and

\[ w_0 \approx 1/R_x C_x \quad (4.30) \]

These relations follow from eqs. (2.12), (2.14), (4.6), (4.7), (4.8a) and (8.65). The applied bias voltage \( V_b \) and MCP strip current density \( J_{\text{strip}} \) impose constraints on \( \phi_m \) and \( G_0 + G_1 \sin \phi_e \):

\[ 0 \leq |\phi_m| \leq |\phi_s| = \left| \frac{\pi}{V_b} \right| \quad (4.31a) \]

and

\[ 0 \leq |G_0 + G_1 \sin \phi_e| \leq |G_s| \frac{\pi}{\sigma_\pi} \frac{J_{\text{strip}}}{w_0} \quad (4.31b) \]

(Here \( \sigma_\pi = C_x V_{\pi}/A_x \), and taking the absolute value makes eqs. (4.31) valid for \( V_b < 0 \).) Equation (4.31b) corresponds to the active irradiance limits:

\[ 0 \leq I \leq I_s = \frac{h\nu}{n\varepsilon_0} J_{\text{strip}} \quad (4.32) \]

A representative \( \phi_m \) vs. \( \phi_m \) state-space plot, corresponding to eq. (4.27), is illustrated in fig. 4.14. Since \( G_0/G_1 > 1 \), due to the lack of intensity offset, the phase error, \( \phi_e \), tends to be large and the immunity to signal amplitude fluctuations is poor.

Utilizing eqs. (4.27) and (4.28), the steady-state error can be expressed as

\[ \phi_e = \sin^{-1}\left[\frac{\phi_m}{G_0/G_1}\right] \quad (4.33) \]

where
Fig. 4.14 Modulator state-space ($\dot{\phi}_m$ versus $\phi_m$) for continuous mode IPL without threshold ($G_0 = 3.5\pi$, $G_1 = 3.42\pi$, $\phi_s = 4\pi$)

Fig. 4.15 Phase error, $\phi_e$, versus modulator phase, $\phi_m$, for an IPL without threshold ($G_0 > G_1$). The solid curves show the effect of increasing $G_1$ and $G_0$ while $r = |E_i|/|E_z| = 0.9$ remains constant.
\[
\frac{G_0}{G_1} = \frac{1}{2} \left( \frac{1}{r} + r \right) \quad \text{with} \quad r = \frac{|E_i|}{|E_x|}
\]  

(4.34)

For all values of \(r\), \((0 \leq r \leq \infty)\), \(G_0/G_1 \geq 1\) holds. The performance of the system is optimal when \(G_0/G_1\) is minimized, which occurs when \(|E_i| \approx |E_x|\).

An automatic gain control (agc) could perhaps track the average signal inadinance, \(<|E_i|^2>\), with the L.O. irradiance, \(|E_x|^2\). This intensity balance is automatically obtained in the Zernike phase-contrast implementation, where from Eq. (3.105)

\[
|E_x| = |E_r| = c' A_m A_f \sqrt{1 + b^2 - 2b \cos \phi_h} <|E_i|> \quad (4.35)
\]

Here \(c' \leq \sqrt{2}\) is approximately unity when in lock, \(A_m A_f \approx \frac{1}{4}\), and \(b\), \(\phi_h\) are fixed parameters of the Zernike Fourier-plane filter; (See Eq. (3.100)). Typically, \(\phi_h = \pi\) and eq. (4.35) becomes

\[
|E_x| \approx <|E_i|> \quad (4.36)
\]

when \(b = 3\).

Actually, the IPL can operate with small error even when \(G_0 >> G_1\), since from eq. (4.33) \(\phi_e\) can be made small by maintaining \(\phi_m \approx G_0\). As discussed in section 3.2.2, this is automatic in the Zernike phase-contrast implementation! In other implementations, e.g. homodyne, the IPL could be periodically restarted at \(\phi_m \approx G_0\) with a calibrated control-beam irradiance. For example, if the system in fig. 4.14 is restarted at \(\phi_m = 3\pi\), the error will largely remain in the range \(\sin^{-1}\left(\frac{2-3.5}{3.4}\right) <\phi_e < \sin^{-1}\left(\frac{4-3.5}{3.4}\right)\), (i.e., \(-.15 \pi < \phi_e < .05 \pi\)).
Alternatively, in adaptive phase compensation applications it is not important that $\phi_e$ be small, but rather that $\phi_e$ be stable with respect to changes in $\phi_i$ and $\phi_m$. Figure 4.15, which depicts $\phi_e$ vs. $\phi_m$ for fixed $|E_i|/|E_x|$, reveals that the operational ($0 \leq \phi_m < \phi_s$) variation of $\phi_e$ is diminished for large $G_1$ and $G_0$; although $G_0 > \phi_s$ prevents $\phi_e = 0$ from being achievable. Best results are obtained when $\phi_s$ is very large and $\phi_m$ is occasionally restarted within a few $\pi$ radians of $\phi_s$. In the Zernike phase-contrast implementation, arbitrary $<\phi_m>$ can be maintained by appropriate choice of $\phi_i$ in Eq. (3.105).

It may be possible to incorporate intensity thresholding into the continuous mode MSLM by employing a strain-biased PLZT electro-optic plate. For example, PLZT 10/65/35 requires a threshold voltage of $V_{X_t} \approx 1.9 \varepsilon_{X} V_{\pi}$, where $\varepsilon_{X}$ is the PLZT thickness in mm, before any phase modulation occurs, i.e.,

$$
\phi_m = \begin{cases} 
\frac{\pi V}{\varepsilon_{X}} (V_X - V_{X_t}) = g'(I - I_t) & I > I_t \\
0 & I \leq I_t 
\end{cases}
$$

(4.37)

Where $g'$ is given by eq. (4.29) and $I_t = \frac{\pi V_{X_t}}{g'}$. Unfortunately, the resistivity ($\rho \approx 10^{13} \Omega \text{cm}$) of these PLZTs is too large for high speed operation; possibly doping or other compositional modifications may result in lower resistivity.

Another approach for achieving intensity thresholding is to precede the write-beam side of the MSLM with a two-dimensional planar semiconductor laser. The interferogram then serves as a pump beam. In spatial locations where $I_0 + I_1 \sin \phi_e$ exceeds the lasing threshold, the laser output drives the MSLM. Image thresholding has actually been reported with a
2-D GaAs platelet laser. Such a configuration can produce additional gain, but the threshold may be too high or the output too noisy to be practical.

4.3.2. Electron-deposition framed mode MSLM-based IPL

The electron-deposition framed mode MSLM is an example of an SLM which integrates its write-beam irradiance. An IPL implemented with such an SLM, lacking in intensity threshold, is characterized by

\[ \phi_m = \frac{G'_0 + G'_1 \sin \phi_e}{t_w} \]  

(4.38a)

In the MSLM, \( G_0 \) and \( G_1 \) are given by Eqs. (4.28) with

\[ g' = \frac{\pi}{\sqrt{v}} \frac{A}{C} \frac{r}{t_w} \frac{\nu G}{h\nu} = \frac{\pi}{S} \frac{r}{t_w} \]  

(4.38b)

\( = q_1 q_2 \) in Eqs. (2.14)

Here \( t_w \) is the write time; and Eq. (4.38b) follows from Eqs. (4.11) and (8.65). The erase and write bias voltage levels, \(-V_{bw}\) and \(V_{bo}\) respectively, bound \( \phi_m \) to

\[ -\frac{\pi}{V} v_{b_B} \leq \phi_m \leq \frac{\pi}{V} v_{bo} \]

and from Eq. (4.32)

\[ 0 \leq \phi_m \leq G'_s = \frac{\pi}{V} J_{strip} = g' I_s / t_w \]  

(4.38c)

Figure 4.16a is a representative state-space (\( \phi_m \) vs \( \phi_m \)) characteristic for \(|E_i|=80\%|E_k|\). The value of \( G_0 + G_1 \) relative to \( G_s \) can be controlled by adjusting the MCP gain or by attenuating the MSLM control irradiance.

(In the visible-phocathode MSLM, \( I_s=2.5 \) nW/cm\(^2\)). When \( \phi_m \) is approximately constant during the write time: \( \phi_e = \phi_m = 0 \); and the error state-space (\( \phi_e \) vs \( \phi_e \)) takes the form in Fig. 4.16b.
Fig. 4.16 State-space characteristics for electron-deposition MSLM-based IPL. (with $|E_i| = 80\%|E_2|$)
Although there are no stable equilibria, if the IPL is designed with $G_0/G_1=1$, small values of $\phi_m$ occur near $\phi_{e_u}=(3-4\pi)/2$. As shown in Fig. 4.17, the IPL spends more time near these "quasi-equilibrium" points than in transit between them, and hence is compensating $\phi_e$ to about $\phi_{e_u}$ during most of the time. In a continuously integrating IPL, given $G_0/G_1$ and $G_s$, the fraction of time $\phi_e$ spends within $\pm \Delta \phi$ of $\phi_{e_u}$ is (see Fig. 4.16b)

$$P(\Delta \phi) = \frac{T_e}{T_c} = \frac{\phi_e \text{ transit time from } \phi_{e_u} + \Delta \phi \text{ to } \phi_{e_u} - \Delta \phi}{\phi_e \text{ transit time from } \phi_{e_u} \text{ to } \phi_{e_u} - 2\pi} \quad (4.39)$$

Assuming that $\phi_i'$ is constant during $T_e$ and $G_0+G_1<G_s$, the desired result follows directly from Eqs. (3.50a) and (3.52)

$$P(\Delta \phi) = \frac{2}{\pi} \tan^{-1} \left( \frac{\sqrt{G_0/G_1 + 1}}{G_0/G_1 - 1} \left( 1 + \frac{\Delta \phi}{\frac{2}{\pi} \cdot \frac{\pi}{4}} \right) \right) \quad (4.40a)$$

with small $\Delta \phi$

$$P(\Delta \phi) \approx \frac{2}{\pi} \tan^{-1} \left( \frac{\sqrt{G_0/G_1 + 1}}{G_0/G_1 - 1} \frac{\Delta \phi}{\frac{2+\Delta \phi}{\pi}} \right) \quad (4.40b)$$

$P(\Delta \phi_e)$ can be interpreted as the probability that the IPL error is bounded by $\phi_{e_u} - \Delta \phi_e < \phi_e < \phi_{e_u} + \Delta \phi_e$ in a given resolution cell at a given time. When $\phi_i'(x,y)$ has a uniform probability distribution in space, $P(\Delta \phi_e)$ is the fraction of resolution cells flat to within $2\Delta \phi_e$ of each other. For example with $|E_1|=0.9|E_2|$ ($G_0/G_1$ is given by Eq. (4.31)), the fraction of
Fig. 4.17 Phase error as a function of time for electron-deposition MLSM-based IPL.

Fig. 4.18 Fraction of initial phase errors, \( f(\Delta \phi_e) \), compensated to within \( \pm \Delta \phi_e \) of \( \phi_{eu} \) in the "timed-integration" MLSM-based IPL.
cells flat to $\frac{\lambda}{5}$ is $P(\pi/5) = 0.898 \approx 90\%$, (note: $2\Delta \phi_e = \frac{\lambda}{5}$, $2\pi$) ; and the fraction flat to $\frac{\lambda}{10}$ is $P(\pi/10) = 0.795 \approx 80\%$.

When $\dot{\phi}_m$ saturation is encountered (i.e., $G_o + G_t > G_s$), the unsaturated state-space trajectory is followed in the interval

$$-\frac{\pi}{2} + \theta \leq \phi_e \leq \theta + \frac{\pi}{2}$$

(4.41a)

where

$$\theta = \sin^{-1} \left( \frac{G_s' - G_o'}{G_1} \right) + \frac{\pi}{2}$$

(4.41b)

The rest of the trajectory is along $\dot{\phi}_m = -G_s$. Using Eqs. (4.40), (3.49), and (3.52), with $\dot{\phi}_m$ saturation $P(\Delta \phi_e)$ becomes

$$P_s(\Delta \phi_e) = \frac{P(\Delta \phi_e)}{\pi + 2\theta \sqrt{\frac{G_1^2 - G_0^2}{G_s}} + P(\theta)}$$

(4.42)

Using (4.40) and (4.41b) the $P(\theta)$ term can be written

$$P(\theta) = \frac{2}{\pi} \tan^{-1} \left( \sqrt{\frac{G_0 + G_1}{G_0}} G_s' G_0 + \sqrt{G_s^2 - (G_s - G_0)^2} \right)$$

$$\sqrt{G_1 G_s' G_0 + \sqrt{G_1^2 - (G_s - G_0)^2}}$$

Saturation of $\dot{\phi}_m$ generally degrades $P_s(\Delta \phi_e)$ relative to $P(\Delta \phi_e)$; for example when $G_s = 1.5G_0$, $P(\pi/5) = 0.90$ drops to $P_s(\pi/5) = 0.87$, (with $|E_i| = 0.9|E_\lambda|$).

In an actual phase compensation application, the MSLM integrates for a limited write time $t_w$, stores the estimate for $t_s$, and completes a frame.
by erasing $\phi_m$ to its minimum limit during $t_e$. Ideally $\phi_i$ does not change appreciably during a frame. More specifically, the frame period, $t_e + t_w + t_s$, should satisfy the Nyquist criterion with respect to the bandwidth of $\phi_i$, $BW_i$; i.e., $t_e + t_w + t_s < 1/2BW_i$. In addition, only a small fraction of the time should be wasted acquiring the phase, i.e., $t_e + t_w << t_s$. Saturation at $\phi_m = \phi_s$ can be avoided by keeping $t_w$ short enough, i.e., $t_w < (\phi_s/2\pi) T_c$. A minimum dynamic range of $2\pi$ radians is required to track any input phase, $\phi_i$; however a larger $\phi_s$ is preferable. The approximate probability for a spatial resolution element being compensated during $t_s$ is $P(\Delta\phi_e)$.

The minimum write time is $t_w = T_c - T_e$, which is the time required to sweep all the uncompensated initial errors into quasi-equilibrium (within $\pm \Delta \phi_e$ of $\phi_{e_u}$ in Fig. 4.16b). An alternative "timed-integration" framed IPL mode is to operate with $t_w = T_c - T_e$ exactly. This is long enough to compensate most input phases and allows only a small fraction of input phases to "escape" from the $\phi_{e_u} \pm \Delta\phi_e$ range. The range of initial phase errors ($\phi_{e_u} = \phi_i$; for $\phi_{e_u} = 0$) which are not correctly compensated can be found by solving Eq. (3.50a) for $a$, given $t_{ab} = T_c - T_e$ and $b = -\pi/2 - \Delta\phi_e$.

(This is equivalent to solving for $b$, given $a = -\pi/2 + \Delta\phi_e$.) The uncompensated interval is then $\Delta\phi = |b-a|$; and the fraction of initial input phases compensated to within $\pm \Delta \phi_e$ is $f(\Delta\phi_e)$

$$
f(\Delta\phi_e) = 1 - \frac{|a-b|}{2\pi} = 1.25 - \frac{\Delta\phi_e}{2\pi} + \frac{1}{\pi} \tan^{-1} \left[ \frac{-2G_1 \gamma^3 + 3G_0 \gamma^2 - G_0}{-G_0 \gamma^3 + 3G_0 \gamma + 2G_1} \right] \quad (4.43a)
$$
where

$$\gamma = \tan \left( \Delta \phi_e / 2 - \pi / 4 \right) = \frac{-\cos \Delta \phi_e}{1 + \sin \Delta \phi_e} = \frac{-1}{1 + \Delta \phi_e} \quad (4.43b)$$

This follows from Eq. (3.51a), using \( \psi_e = \frac{-\pi}{2} + \Delta \phi_e \) and

$$t = t_w = T_c - T_e = T_c (1 - P(\Delta \phi_e)) \quad (4.44)$$

(and assuming intensity saturation does not occur, i.e., \( G_s > G_0 + G_1 \)).

The behavior of \( f(\Delta \phi_e) \) as a function of \( \Delta \phi_e \) and \( |E_i| \), as illustrated in Fig. 4.18, is a little complicated. Since all error phases lie within \( \pm \pi \), when \( \Delta \phi_e = \pi, P(\Delta \phi_e) = 1 \); resulting in \( t_w = 0 \) (Eq. 4.44) and \( f(\Delta \phi_e) = 1 \). As \( t_w \) is increased, corresponding to decreasing the \( \Delta \phi_e \) tolerance, more initial phases "escape" from the \( \phi_e \pm \Delta \phi_e \) range and \( f(\Delta \phi_e) \) decreases. However as \( \phi_e \pm \Delta \phi_e \) nears \( \phi_u \), the "escape" rate becomes so slow that \( f(\Delta \phi_e) \) actually increases. At extremely small \( \Delta \phi_e \): \( P(\Delta \phi_e) \to 0 \), \( t_w \to T_c \), the uncompensated interval exceeds \( 2\Delta \phi_e \), and \( f(\Delta \phi_e) \) diminishes to 0 at \( \Delta \phi_e = 0 \).

Fig. 4.18 is useful for finding the optimum write time \( t_w \) for the timed-integration framed IPL. A value of \( t_w \) near the small \( \Delta \phi_e \) rolloff of \( f(\Delta \phi_e) \) should be chosen. For example when \( |E_i| = 0.9|E_x| \), \( t_w = T_c (1 - P(\pi/25)) \) compensates 97% of all possible input phases to within a flatness of \( \lambda/25 \). With \( |E_i| = 0.8|E_x| \), \( t_w = T_c (1 - P(\pi/15)) \) compensates 93% of input phases to a flatness of \( \lambda/15 \); and with \( |E_i| = 0.5|E_x| \) 80% of input phases can be compensated to a flatness of \( \lambda/5 \).

This system exhibits surprisingly good performance; the short integration time overcomes its inherent instability. Besides allowing
fast, efficient \( t_w < t_s \) framing, minimizing the integration time also reduces the sensitivity to background light. This system can operate with a minimum modulator dynamic range of only \( 2\pi \). Unfortunately, as evident in Fig. 4.18, degradation is suffered as \( G_0/G_1 \) decreases; an automatic gain control, causing \( |E_x| \) to track \( <|E_z|> \) may help.

4.3.3. Inverted-secondary framed mode MSLM-based IPL

Unlike the modes discussed previously, the inverted-secondary framed-mode of the MSLM exhibits an intensity threshold characteristic. As discussed in Section 4.2.1.3, this mode is characterized by an adjustable "hard-clipping" threshold, with essentially no response \( (\dot{\phi}_m=0) \) for \( I<I_t \) and a full response \( (\dot{\phi}_m=\pi \dot{V}_b/V_{\pi}) \) for \( I>I_t \) (see Fig. 4.10). From Eqs. (4.17), (4.18) and (4.21), the threshold intensity is approximately

\[
I_t = \frac{\chi \dot{V}_b}{A_r} \frac{h\nu e^{-V_g p/V_a}}{\eta eG f_H(1-\delta_H)} = \frac{S_{\pi}}{V_{\pi}} \dot{V}_b
\]

(4.45)

\( I_t \) can be adjusted by controlling \( \dot{V}_b \), the rate at which the MSLM crystal bias is reduced; or \( G_0 \), the MCP gain.

A homodyne IPL implemented with an inverted-secondary MSLM is characterized by

\[
\dot{\phi}_m = -\frac{\pi}{V_{\pi}} |\dot{V}_b| U(I_0+I_1 \sin \phi_e-I_t)
\]

(4.46)

Here \( U(I) \) is the unit step function, i.e. \( U(I>0)=1 \) and \( U(I<0)=0 \); and \( I_0 \) and \( I_1 \) are given by Eqs. (2.2) and (2.3). Examples of the \( I \) vs \( \phi_m \) and modulator state-space (\( \dot{\phi}_m \) vs \( \phi_m \)) characteristics are shown in Fig. 4.19.
Notice that the \( \dot{\phi}_m \) vs \( \phi_m \) state-space has "dead" regions where \( \dot{\phi}_m = 0 \); the effects of these regions are discussed in Section 3.1.6.1. The modulator dynamic range includes negative phases, i.e., \(-\phi_m < \phi_m < \phi_s\). (The electron deposition frame modes can also be operated with negative modulator phase.)

A compensation frame begins by rapidly initializing \( \phi_m \) to \( \phi_s \) (during \( t_e \)). This is followed by an inverted secondary write period of \( t_w \), which is short enough for \( \phi_i \) to appear constant. A minimum dynamic range of \( \phi_m + \phi_n > 2\pi \) and write period of \( t_w > 2V_b/|\dot{V}_b| \) is required to compensate all values of \( \phi_i \). The frame concludes with a storage period \( t_s \). The frame period, \( t_e + t_w + t_s \), should satisfy the Nyquist criterion relative to the bandwidth of \( \phi_i \), and the system should be compensating most of the time (i.e., \( t_e + t_w < t_s \)). The convergence time \( t_w \) can in principle be made arbitrarily fast by increasing \(|\dot{V}_b|\), (and using the MCP gain to control \( I_+ \)). In practice \(|\dot{V}_b|\) is constrained by the secondary emission characteristics of the system and the maximum possible MCP current.

This MSLM-based IPL tracks \( \phi_i \) with a constant compensated phase error of \( \phi_{e_u} \approx (3-4n)\pi/2 \). More specifically, the error will be

\[
\phi_{e_1} = \sin^{-1} \left( \frac{I_0 - I_1}{I_1} \right) - \pi
\]

(4.47)

for most spatial elements; \( \phi_{e_u} \), \( \phi_{e_1} \), for \( I_0, I_1 \). \( \phi_m \) of spatial elements which happen to start in a dead region, where \( \dot{\phi}_m = 0 \), will remain stationary. This is inconsequential as long as the dead region is narrow \( (I_0, I_1) \), since these elements are already compensated with \( \phi_{e_u} \). Using Eq. (4.47), the width of the dead regions is \( 2\left|\phi_{e_1} + \frac{\pi}{2}\right| \). Figure 4.20 shows that small
Fig. 4.19 State-space characteristic for the inverted-secondary framed mode MSLM-based IPL.
The upper plot is $\phi_m$ versus $\phi_m$ for $\phi_m = \frac{\sqrt{V}}{V} (I_o - I_1 \sin(\phi_1 - \phi_m)) - I_t$ with $\phi_1 = 0$. The lower plot is $I$ versus $\phi_m$ for $I = I_o - I_1 \sin(\phi_1 - \phi_m)$. 
fluctuations in $|E_1|$ do not perturb $\phi_e$, much; however this system would probably benefit from an agc which adjusts the MCP gain to make $I_0$ or $|E_0|$ responsive to changes in the average signal amplitude. Note that $\phi_e$ as given by Eq. (4.47) is independent of $\phi_i'$ or $\phi_m$; this is the desired behavior for phase estimation and compensation applications.

Due to the operational physics of the inverted-secondary MSLM, the "dead" regions ($\phi_m=0$) have no deleterious effect on the stability of the phase estimate. Once $I<I_t$ occurs in a resolution element, that element is physically "locked-out" and its modulator phase remains fixed. Subsequent changes in $\phi_i'$, or increases in $I$ (even background light $\gg I_t$) will not perturb $\phi_m'$. An internal field prevents any electrons from reaching that region of the crystal until the whole modulator is reset to $\phi_m=\phi_s$ at the start of the next frame. The surface and bulk resistivities are usually too large for the crystal charge to change by conduction, $(R_x/L_x)C_x=\text{weeks}$ in the visible-phocathode MSLM). These effects suggest against using the inverted-secondary mode in a heterodyne MSLM, because the drift of $\phi_i'$ would not be tracked. As mentioned previously, the inverted-secondary IPL should be operated with short frame times over which $\phi_i'$ is approximately constant.

An interesting hybrid operating scheme employs both the primary and inverted-secondary modes. A frame is started by uniformly initializing the crystal to $-\phi_n$. Then a short timed-integration (e^-deposition) mode is executed. Finally the inverted-secondary mode is employed to correct the fraction of input phases (usually <10%) which the primary mode failed to compensate.

In the future, the MSLM may be modified to operate with threshold in both the electron-deposition and inverted-secondary
Fig. 4.20 The active phase error, $\phi_{e1}$, as a function of signal amplitude, $|E_i|$, in the inverted-secondary MSLM-based IPL. Increasing $|E_t|$ decreases the sensitivity to amplitude fluctuations, however the width of the dead region increases with $|E_t|$. 

$|E_t| = 0.1|E_x|$ 

$|E_t| = 0.25|E_x|$ 

$(I_t = \frac{|E_t|^2}{2|Z_0|})$
framed modes. This may be achievable by coating the electro-optic plate with an amorphous semiconductor or using a PLZT electro-optic plate. The PLZT (see eq. (4.37)) is particularly attractive since its high resistivity, which was a handicap in the continuous mode, allows long storage times in the framed modes.

4.3.4 Heterodyne MSLM-based IPL

In section 3.2.1 it was seen that introducing a frequency offset $\Delta w = w_\phi - w_i$ between the optical frequency of the local oscillator (L.O.) beam ($w_\phi$) and that of the signal beam ($w_i$) can effectively reduce $G_0$ relative to the gain $G_1$, in a manner similar to the beneficial effect of detector intensity threshold. This suggests that heterodyne operation may improve the performance of an IPL employing an SLM lacking intensity threshold. (This section makes use of results from section 3.2.1 and it will be assumed that the reader is familiar with that material.) Based on Eq. (3.89), the input phase will be taken as $\phi'_0 = \phi'_0 + \Delta w t$, where it is assumed that $\phi'_0$ is a function of space and varies only slowly with time ($\dot{\phi}'_0 = 0$).

Continuous-mode MSLM

The low-pass heterodyne IPL discussion of section 3.2.1.2 is applicable to a heterodyne IPL employing a continuous-mode MSLM. Using Eqs. (3.94b) and (3.89) with the parameters of Eqs. (4.27)-(4.30), the error is characterized by

$$\dot{\phi}_e = -w_0 (\phi_e - \phi'_i - (G_0 \frac{\Delta w}{w_0}) - G_1 \sin \phi_e)$$

Comparison with $\dot{\phi}_e$ for the homodyne IPL (with $\phi'_i = 0$) reveals that the offset is effectively shifted from $G_0$ to $G_0 - \Delta w/w_0$. 
As illustrated in Figs. 3.14 and 3.17 this IPL (with 0<Δw<w_0(1+G_1)) exhibits three successive regimes of operation. There is a rapid convergence with time constant τ_0=1/(w_0(1+G_1)) to a small error quasi steady-state; this state is held for a tracking interval T_t (given by Eq. (3.98) while the modulator tracks ϕ_i_0 and Δwt; and after T_t the modulator begins a limit-cycle oscillation with period T_c=2π/Δw.

A practical mode of operation appears to be framing of the IPL at the rate 1/T_t. This maintains the IPL in its small error tracking interval, corresponding to the RC segment of Fig. 3.14. Since the error is small, the system is well characterized by the linearized steady-state of Eqs.(3.76), which can be rewritten as:

\[ \phi_m = \frac{G_1}{1+G_1}(\phi_i_0 + \frac{G_0}{G_1} \frac{\Delta w}{1+G_1} w_a + 2n\pi) + \frac{G_1}{1+G_1} \Delta wt \]

and

\[ \phi_e = \frac{\phi_i_0}{1+G_1} - \frac{G_1}{1+G_1} \left( \frac{G_0}{G_1} \frac{\Delta w}{1+G_1} w_a + 2n\pi \right) + \frac{\Delta wt}{1+G_1} \]

Note that relative to a homodyne IPL (Eq.(2.24) G_0 is reduced to about G_0-Δw/w_0, which can: reduce the static component of the error, decrease the convergence time, and improve the suppression of amplitude fluctuations. The growing error term is spatially uniform and hence does not degrade phase compensation or spatially-relative phase estimation performance. Eventually the linear approximation fails, but Figs. 3.14 and 3.17 show that this mode performs well during the full T_t interval.

T_t can be made arbitrarily large by increasing G_1, but in practice the modulator dynamic range ϕ_s will limit the frame time. A more practical frame time than (3.98) is

\[ T'_t = \frac{\phi_s - \phi_m - \pi}{G_1 \Delta w / (1+G_1)} = \frac{\phi_s - \phi_m - \pi}{\Delta w} \]  
(4.48)
It is advantageous to utilize the longest available tracking time $T'_t$ in order to minimize the fraction of dead time $f_d = (5\tau_0 + t_e)/(5\tau_0 + t_e + T'_t)$, where $t_e$ is the erase time between frames.

The modulator phase at the start of a frame, $\phi_{mo}$, should satisfy

$$\phi_{mo} \geq G_0 - G_1 + \pi$$

in order that initial phase acquisition can occur without skipping cycles or reaching the $\phi_m = 0$ limit of the modulator. (The low-pass homodyne IPL should also satisfy this condition. This condition need not be too wasteful of modulator dynamic range, e.g. with $G_1 = 10\pi$ and $|E_1|/|E_\omega| = 0.8$, $G_0 - G_1 = 0.25\pi$. ) The frequency shift should be chosen to obtain an effective system offset, $G_{oe}$, of zero during initial convergence, i.e.

$$G_{oe} \equiv G_0 - \frac{\Delta w}{\omega_0} - \phi_{mo} = 0$$

The system should also satisfy $|BW_i + \Delta w| < \omega_0 (1 + G_1)$, where $BW_i$ is the bandwidth of fluctuations in $\phi'_i$. (With $\phi_{mo} > G_0 - G_1 + \pi$, Eq. (4.50) implies that $\Delta w < \omega_0 (G_1 - \pi)$.)

The low-pass integrating MSLM-based IPL could also be operated continuously in the limit-cycle regime; however, as Fig. 3.17 illustrates, the spatially-relative error is only small for a fraction of the time (25% to 100% depending on $\phi'_i$).

A third possible operating mode is a simple and hold mode, where the IPL is turned on for a time interval of about $5\tau$ and then $\phi_m$ is held for the rest of the frame interval of length $1/(2BW_i)$ (Nyquist criterion). This mode offers the advantage of requiring a modulator tracking range of only about $3\pi$ radians, in order to acquire any $\phi'_i$ and allow for tracking $\Delta w t$ during acquisition. In practice, due to
the condition in Eq. (4.49), a modulator dynamic range of at least
\[ \phi_s = G_0 - G_1 + 3\pi \] is required. This is considerably less than the dynamic range required by the other two operating modes described above. The offset during initial convergence, \( G_{oe} \), should be zeroed by choosing \( \Delta w \) to satisfy Eq. (4.50). For a given \( G_0, G_1 \), and dynamic range, this sample and hold heterodyne mode performs better in terms of error, acquisition time, and immunity to amplitude fluctuations than the homodyne implementation. It may be possible to further enhance the immunity to amplitude fluctuations by controlling \( \Delta w \) with feedback from the average amplitude to maintain \( G_{oe} \approx 0 \). (For example, the speed of rotation of a half-wave plate or grating, or the frequency of a bragg AO modulator could be controlled).

A sample and hold process may be implementable in the continuous mode by employing a photoconductive crystal, such as \( \text{Bi}_{12}\text{SiO}_{20} \). During the sample time the crystal could be put into its low-pass filter regime by a conductivity-biasing light beam. During the hold time, the bias beam and the MCP would be turned off. (Of course the hold time should be much less than the dark charge-relaxation time \( \rho_e \) of the crystal.)

In summary, the sample and hold mode requires less dynamic range than, but lacks the continuous \( \phi_{io} \) tracking capability of, the first heterodyne framed mode described. Ultimately, hardware considerations, such as the properties of the modulator crystal and the maximum available dynamic range, will determine which mode is most desirable.

**Electron-deposition framed mode MSLM**

A heterodyne IPL implemented with the electron-deposition framed mode of the MSLM corresponds to the integrating heterodyne IPL of section 3.3.1.1. The dynamic behavior of the error is described by
\[ \dot{\phi}_e = (\Delta w - G_0') - G_1' \sin \phi_e \]

which follows from \( \phi_e = \phi_i' - \phi_m \) and Eqs. (3.90b) and (4.38).

The effective offset is thus \( G_{oe} = \Delta w - G_0' \); and by choosing \( \Delta w = G_0' \), the IPL can be made to operate with zero phase tracking error, minimum acquisition time constant \( \tau = 1/G_1' \), and maximum bandwidth, \( |\dot{\phi}_i'| < BW_i = G_1' \).

However, when immunity to amplitude fluctuations is of concern, a better choice for \( \Delta w \) is

\[ \Delta w = \frac{\pi}{S} \frac{|E_2|}{2z_0} = G_0' (|E_1| \to 0) \tag{4.51} \]

(from Eqs. (4.93b) with \( I_i' = 0 \), (4.28), and (4.38)). The L.O. intensity should be chosen to be larger than the greatest expected input intensity, i.e. \( |E_2|^2 > |E_1|^2 \). Both \( \Delta w \) and \( S \) are generally adjustable, and when Eq. (4.51) is satisfied, the system operates with diminishing error as \( |E_1| \) decreases. The ultimate minimum usable \( |E_1| \) is determined by such considerations as convergence time or noise.

As in the case of the homodyne electron-deposition MSLM-based IPL, this heterodyne system must also be framed to prevent the modulator from locking up at its maximum phase, \( \phi_s' \). Unlike in the homodyne case, however, the phase estimate in this heterodyne system is stable providing that \( \Delta w \) is within the range \( G_0' - G_1' < \Delta w < G_0' + G_1' \). (If MSLM saturation is encountered \( G_0' \) of Eq. (4.38c) replaces \( G_0' + G_1' \) as the upper limit.)

Two major types of framed operation are possible: the IPL can be operated continuously for time \( T_c \leq (\phi_s' - \phi_m' - \pi) / \Delta w \) and then reset to \( \phi_m' \), or operated for a maximum convergence time of a few time constants \( (\tau = 1/G_1') \) and then frozen (sample and hold mode) for a frame time \( T_f \) determined by the bandwidth of \( \phi_i' \), \( (T_f \approx 0.5/BW_i) \) where \( BW_i \ll 1/\tau \) is
assumed). The former mode offers the advantage of continuous tracking of \( \phi_{i0} \) during \( \Delta t \); but the sample and hold mode is less demanding of modulator dynamic range, requiring

\[
\phi_m + \phi_s \geq 2\pi + 5\Delta \omega t = \pi (2 + \frac{2.5}{\tau_m}) \geq 2.8\pi
\]

in order to acquire any initial value of \( \phi_{i0} \) and allow for drift with \( \Delta \omega t \). (\( r = |E_i|/|E_2| \) enters because it was assumed that \( \Delta \omega \) was chosen to satisfy Eq. (4.51)). This mode is particularly attractive because the electron-deposition MSLM is a natural sample and hold device; when the MCP is turned off, \( \phi_m \) is stored for long periods (e.g. weeks) depending on the crystal employed and fabrication details.

Inverted-secondary mode MSLM

The dynamic behavior of the error in an IPL employing the inverted secondary mode of the MSLM is approximately characterized by

\[
\dot{\phi}_e = \Delta \omega + G_s U(I_o + I_1 \sin \phi_e - I_t)
\]  

(4.52)

which follows from Eq. (4.46) with \( G_s = \pi V_\pi / |V_b| \).

At first glance it would seem that operation with \(-G_s \Delta \omega < 0\) will remove the "dead regions" which degraded the performance of the homodyne system of section 4.3.3. This would be the case if Eq. (4.52) was a complete description of the IPL operation. However, it is a property of this MSLM mode that after \( I_o + I_1 \sin \phi_e \) drops below \( I_t \) for the first time, \( \phi_m \) holds its value and \( \dot{\phi}_m = 0 \) for the rest of the frame time. Thus resolution cells, for which \( |\phi_e| = |\dot{\phi}_e| = \phi_m (at t = \tau) \) is less than \( \phi_{i1} \) of Eq. (4.47), are below threshold and remain inactive for the whole frame. As \( \phi_{i1} = \phi_{i0} + \Delta \omega t \) varies, the whole modulator eventually becomes inactive.

Nevertheless, this behavior does not preclude heterodyne operation.
Study of Figs. 4.19 and Eq. (4.46) reveals that resolution cells which are above threshold at \( t=0 \) will lock out after an interval

\[
T = \frac{\phi_1 + \phi_m - \phi_{10}}{\Delta w + G_S} \quad (4.53a)
\]

The multistability of \( \phi_1 \) is always such that \( T > 0 \). The phase at lock out is

\[
\phi_m = \frac{G_S}{\Delta w + G_S} \phi_{10} + \frac{\Delta w - G_S \phi_1}{\Delta w + G_S} \quad (4.53b)
\]

The second term of \( \phi_m \) is spatially uniform and the first term is a phase estimate of \( \phi_{10} \) when \( G_S \gg \Delta w \). This is essentially the sample and hold operating mode of the integrating heterodyne IPL. Inverted-secondary operation, however, offers the advantage of a faster convergence time; \( T \) in Equation \( (4.53a) \) can in principle be made very short by increasing the system parameter \( G_S \) (\( \approx 1/|V_b|/V_\pi \)).

It can be concluded that heterodyne operation in the inverted-secondary mode offers no particular advantage over homodyne operation, but is functionally useful if required in a particular application, e.g. with a drifting L.O. Both homodyne and heterodyne operation suffer from the same dead region limitations and offer the same time response advantages.

4.3.5 MSLM-based IPL summary

The primary challenge of incorporating the MSLM into an adaptive IPL system was seen to be the lack of intensity threshold or offset in most of the MSLM's operating modes. This results in an MSLM-based IPL characterized by \( G_0 > G_1 \), which is undesirable.

In the continuous MSLM mode some of the modulator's usable dynamic range can be sacrificed to overcome the lack of intensity threshold, since the effective offset of a low-pass homodyne IPL is \( G_0 - \phi_m \). Best compensated-
wave flatness is obtained by operating with $G_i$ as large as possible, in which case nulling $G_o - \phi_m$ is impractical; although $|G_o - \phi_m| \leq G_i$ must hold. The minimum required modulator dynamic range is generally at least $3\pi$ radians.

The electron-deposition framed MSLM mode results in an unstable integrating IPL. Although continuous operation for long time intervals is feasible under certain circumstances, "timed-integration" framed operation is superior. The timed-integration mode requires only $2\pi$ radians of dynamic range and is stabilized by operating for a precisely controlled write-time interval (which is a little less than $T_c$, the period of the $\dot{\phi}_m$ limit-cycle).

The inverted-secondary mode of the MSLM results in an integrating IPL with adjustable threshold. This system requires a dynamic range of at least $2\pi$ radians and can offer fast convergence times. The inverted-secondary IPL, however, suffers from state-space "dead regions"; which, with proper design, do not seriously degrade performance.

Heterodyne operation considerably enhances the performance of IPLs employing the electron-deposition framed modes or continuous mode of the MSLM by subtracting from $G_o$ a term proportional to the L.O. frequency shift, $\Delta w$. Although continuous phase tracking for long periods is possible, sample and hold operation is preferable in order to restrict the required modulator dynamic range to about $3\pi$ radians. The heterodyne electron-deposition framed mode is particularly attractive because it inherently implements the sample and hold operation and, with $\Delta w$ fixed according to Eq. (3.93h), it exhibits better immunity to amplitude fluctuations than the other modes.

There are a variety of additional techniques for potentially
enhancing the performance of the MSLM-based IPL. It may be possible to introduce an adjustable threshold into any MSLM operating mode by employing a PLZT electro-optic plate, an amorphous semiconductor coating on the dielectric mirror, or preceding the write-beam side of the MSLM with a 2-D platelet laser. Another technique which looks promising, but has not been thoroughly analyzed, is to operate with a heterodyne frequency offset $\Delta w$ between the L.O. and input beams and to modulate the MCP voltage at $\Delta w$ (or possibly modulate another voltage such as the gap voltage). Then, employing the intrinsic frequency response of the MSLM (e.g. low-pass filter in the continuous mode), it may be possible to implement electronic heterodyne detection, with the $G_1 \sin e_e$ term passing through the system at baseband frequency, but with the $G_0$ term at $\Delta w$ (and all signal components at harmonics of $\Delta w$) being filtered out or at least attenuated.

Additional techniques for reducing $G_0$ relative to $G_1$ become available if an MSLM possessing a bandpass spatial or temporal frequency response characteristic can be constructed. Then, as mentioned in section 3.1.5 (see Eqs.(3.31d,e)), if $G_0$ is made spatially (or temporally) uniform (e.g. with $|E_x| >> |E_y|$ and $|E_z|$ uniform and static) the modulator's spatial (or temporal) passband will block $G_0$. Low spatial frequency or dc content in $\phi_1$ will not be lost if a tilted L.O. beam is employed to shift the spectrum of $\phi_1$ into the modulator's spatial bandpass. Similarly, heterodyne operation can be employed to shift the time variation of $\phi_1$ into the modulator's temporal bandpass. If a choice is possible, the spatial method is preferrable, because in the temporal method heterodyne operation will eventually drive the modulator to its physical limits. Employing a deformable membrane instead of an electrooptic crystal may provide a means for obtaining a spatial bandpass characteristic.
Many of the techniques discussed in this section are easily generalized to other SLMs, should a new SLM become available which meets the minimal performance criteria outlined in section 4.1.

4.4 Conclusions

In the absence of a suitable alternative among the existing spatial light modulators, the MSLM was developed for use in implementing high-resolution, all-optical adaptive systems. The MSLM is potentially faster, more sensitive, and capable of greater dynamic range than most other optically-addressed spatial phase modulators.

The discussions of the previous section have suggested that there are a variety of practical options for employing the MSLM to implement high-resolution adaptive systems. The choice of a particular operating mode is largely dictated by MSLM hardware considerations and the particular application. For example, an electro-optical crystal with low bulk resistivity and an MSLM dynamic range of at least 3\(\pi\) radians is required for the continuous mode. Sample and hold operation of the heterodyne electron-deposition MSLM-based IPL appears to be the most desirable operating mode. This mode requires a dynamic range of between 2\(\pi\) and 3\(\pi\) radians and can provide ideal IPL performance, characterized by very good immunity to amplitude fluctuations and \(G_1 \gg G_{oe} = 0\). Where the additional complexity of a frequency shifted L.O. beam is not compatible with the specific application and/or the modulator dynamic range is 2\(\pi\) radians or less, the homodyne timed-integration or homodyne inverted-secondary modes are more desirable. The latter offers threshold and a potentially faster convergence time, but in practice requires an MSLM with a grid between the MCP and crystal to enhance secondary emission.
All-optical IPL systems employing the homodyne timed-integration and inverted-secondary MSLM framed modes have actually been demonstrated in the laboratory and are described in chapter 6. Unfortunately the fastest available MSLMs (e.g. with a grid between the MCP and crystal) could not be utilized in these experimental closed-loop adaptive optical systems because those MSLMs were demountable and hence restricted to only UV-sensitive (λ<260nm) photocathodes. All of the available electro-optical crystals were opaque to this radiation. Even with a suitable crystal, a coherent short-wave UV source is not readily available; although UV lasers do exist.\textsuperscript{49,230}

Even worse, the only visible-photocathode MSLM available at the time of these experiments could be used only in simple low-resolution adaptive demonstrations, because it employed a parallel crystal without a dielectric mirror. As discussed in Appendix B.3, a parallel-crystal Fabry-Perot produces undesirable intensity variations with \(V \chi\) (Fig. B.15) in an external interferometer. This device's limited dynamic range of \(\pi\) radians was also a handicap.

On the positive side, the MSLM performance defects are well understood and the required improvements are currently in the process of being implemented by other researchers.\textsuperscript{219,228} A facility for in-house visible photocathode transfer is also under construction.\textsuperscript{217} It is conceivable that within two years the MSLM will realize the performance objectives of ≤msec cycle times, 20 \(\lambda_p/\text{mm}\) resolution, ≥5 \(\pi\) dynamic range, and a half-wave sensitivity of better than 1 pJ in the visible.

Another visible-photocathode MSLM with a wedge crystal and larger expected dynamic range (> 3\(\pi\) radians) has very recently been assembled by Hamamatsu. Laboratory demonstrations of high-resolution adaptive optical systems incorporating this device will be described in future publications.
V. EXPERIMENTAL DISCRETE-CHANNEL OPTICALLY-ADDRESS PHASE MODULATORS

A variety of optically-driven phase modulating systems were constructed in order to implement and investigate the proposed adaptive optical techniques. Even in high-resolution applications, many variants of the interference phase loop (e.g., homodyne or heterodyne) involve essentially independent spatial elements operating in parallel. Thus many features of IPL operation can be studied with a single channel optical loop. Multiple spatial resolution elements are required to implement such systems as the Zernike phase-contrast IPL and to investigate overall IPL compensation performance with spatially corrupted wavefronts.

Discrete channel systems bring additional flexibility to laboratory adaptive optical investigations. A limited number of spatial channels affords an opportunity to study more carefully the effects of such defects as spatial coupling and spatially nonuniform gain and dynamic response on high-resolution system performance. Channels can be individually monitored and controlled; and the dynamic response characteristics of a large variety of SLMs, and even more ideal responses which would be difficult or impossible to obtain with an SLM, can be simulated.

Truly high-resolution applications with all their attendant special features (e.g., spatial resolution effects) are really only practical with a monolithic spatial light modulator. The MSLM which was described in the last chapter shows promise for future high-resolution applications.

5.1. Single-channel system

The single-channel optically-driven phase modulator, represented schematically in Fig. 5.1, employed an Si photodiode (Hamamatsu #S876-1010BQ, 0.4 A/W sensitivity) to drive a LiNbO$_3$ phase modulator. The phase modulator,
Fig. 5.1 Single-channel optically-addressed phase modulator.
\[ \frac{\lambda}{n} \frac{t}{e} = 231 \text{ volts; measured } 200 \text{ volts} \]

\[ w_\text{L} q(I_o - I_f) = w_\text{L} G_0 \]

\[ \frac{w_\text{L} \pi R + R}{V} (-e R e C V_1 - V_H) \]

Fig. 5.2 Typical Modulator state-space \((\phi_m \text{ versus } \phi_a)\) of single-channel IPL.
constructed by M. Smith, had a half-wave voltage of $V_m = 200$ volts. The versatile high voltage amplifier, shown schematically in Fig. C.1 of Appendix C, featured: variable gain, adjustable intensity threshold and offset, a controllable dominant single pole, an adjustable initial value of modulator phase, and diagnostic output signals proportional to the input intensity, the modulator phase, and the derivative of modulator phase. The frequency response and other characteristics of this amplifier are discussed in more detail in Appendix C. The frequency response relations were used to insure that the adjustable pole dominated over most of its range, and to understand the circuit behavior and limitations when this pole is at too large a frequency to dominate.

A type-zero response was chosen because it, or in particular a low-pass filter, is a good approximation to some high resolution modulators, such as the continuous-mode MSLM described in the last chapter. At the maximum bandwidth used ($f_m = 1/2\pi R_i C \approx 1 k\text{Hz}$) the effects of a second pole can be seen. It is also easy to produce laboratory phase turbulence with bandwidth $\ll f_m$, and hence simulate an infinite bandwidth system. The minimum bandwidth ($\sim 3$ Hz) was chosen to be less than many of the frequency components of laboratory turbulence. Actually, the bandwidth can be decreased even further in order to approximate an integrating IPL. (A better type-one integrating response can be obtained by replacing $R_f$ with $C$ in Fig. C.1).

For the component values employed, the frequency response of the modulator voltage, $\dot{V}_x$, to the control irradiance $\dot{I}$ is well approximated by Eq. (C.10) in Appendix C. Transforming Eq. (C.10) back to the time domain results in a differential equation for $v_x(t)$; and with
\[ \phi_m = \pi v_x / v_\pi \], this becomes a differential equation for the modulator phase,

\[ \dot{\phi}_m = w_L \left[ -\phi_m + \frac{\pi}{\pi} (v_H - \frac{R_c + R_e}{R_e} (v_R - R_f S A I(t))) \right] \quad (5.1) \]

This is only valid for (from Eq. (C.7))

\[ \frac{v_R - v_1}{R_f S A} U(v_R - v_1) = I_t < I_s = \frac{v_r}{R_f S A} \quad (5.2a) \]

and

\[ 0 < \phi_m < \phi_s = \pi v_H / v_\pi \quad (5.2b) \]

Here S and A are respectively the photodiode sensitivity in Amps/Watt and photodiode area. \( U(\cdot) \) is the unit step function. The cutoff frequency is

\[ w_L = \frac{1}{(R + R_f) C} \quad , \quad (5.3) \]

where \( R_f \) is the intrinsic output resistance (= 60\( \Omega \)) of the operational amplifier ("op amp"). Voltage \( v_R \) is given by \( v_R = v_+ + v_2 \), where \( v_2 \) is the op amp input offset voltage which is adjustable by \( R_2 \). The parameters \( v_+ \), \( v_H \), \( v_1 \), \( R \), \( C \), \( R_f \), \( R_c \), \( R_e \), and \( R_2 \) are all defined in Fig. C.1.

With the interferometric intensity \( I = I_o + I_1 \sin \phi_e \), from Eqs. (2.1) and (2.2), \( \dot{\phi}_m \) is given by

\[ \dot{\phi}_m = w_L (-\phi_m + G_o + G_1 \sin \phi_e) \quad (5.4) \]
where

\[ G_1 \equiv g I_1 = \frac{\pi}{vH} \frac{R_c + Re}{Re} R_S A I_1 \]  \hspace{1cm} (5.5a)

and

\[ G_0 = \frac{\pi}{vH} (v_H - \frac{R_c + Re}{Re} v_r) + g I_0 \]  \hspace{1cm} (5.5b)

According to Eqs. (5.2) this expression is valid for \( I_t < I < I_s \) and \( 0 < \phi_m < \phi_s \). These operational limits are discussed in more detail in Appendix C. Note that there is intensity saturation at \( I_s \) and when \( v_r > v_1 \). This flexible circuit offers both a separately adjustable threshold intensity \( I_t \) (Eq. (5.2a)) and a phase offset intensity \( I_f \) defined by:

\[ G_0 \equiv g(I_0 - I_f) \]  \hspace{1cm} (5.6)

From Eq. (5.5b)

\[ \frac{v_r - \frac{Re}{Re + R_c} v_H}{I_f} = \frac{1}{R_S} \]  \hspace{1cm} (5.7)

Alternatively \( G_0 \) can be rewritten in terms of the threshold intensity defined in Eq. (5.2a), i.e.,

\[ G_0 = \frac{\pi}{vH} (v_H - \frac{R_c + Re}{Re} v_1) + g(I_0 - I_t) \]  \hspace{1cm} (5.8)

(It is an actual threshold only when \( v_r > v_1 \).)

The offset, and operational limits on \( I \) and \( \phi_m \) can be better understood by examining the sample closed-loop state-space plot (\( \phi'_m \) vs \( \phi_m \) of
Fig. 5.2. (Remember that when the loop is closed, $\phi_e=\phi_i^t - \phi_m$.) Note that $\phi_m$ in Eq. (5.1) is proportional to $I_t$; hence the $I_t$ and $I_s$ limits bound $\phi_m$. These limits can also bound the steady-state ($\phi_m=0$) range of $\phi_m$ to between where they intercept the $\phi_m=0$ axis, unless the $\phi_m$ bounded range is shorter ($0<\phi_m<\phi_s$). This is the case in Fig. 5.2 at the lower limit of $\phi_m$ at $\phi_m=0$. The upper limit at $\phi_m=\phi_s$ coincides with the $I_t=I_s$ intercept.

This amplifier simulates a device with threshold but no offset, which corresponds to the system discussed in Chapter 2, when $v_r>v_1$ and (from Eq. (5.7))

$$v_r = \frac{R_e}{R_e + R_C} v_H$$

(5.9)

Then the $I_t=I_s$ intercept passes through the origin of the $\phi^t_m$ vs $\phi_m$ state-space plot.

**Component Values**

The MJ205 transistor breakdown voltage of 800 volts limited $v_H$ to about 700 volts and $\phi_s$ to less than $3.5\pi$ radians. As mentioned in Appendix C, $R_e$ (2.2k), $R(\leq 1k)$, and $C(\leq 50\mu F)$ were chosen to make $\omega_L$ independent of gain, offset, and threshold. For typical intensities ($I_t=20\mu W$), the maximum gain of this circuit is $G=55\pi$ ($R_f\leq 2M\Omega$, $R_c\leq 1.5M\Omega$, $S=0.4 A/W$).

There are a variety of constraints which narrow the choice of component values. For example in Chapter 3, it was suggested that some aspects of IPL operation are optimized when $G=\phi_s/2$, or from Eq. (5.5b)

$$\frac{v_H}{2} = \frac{R_c + R_e}{R_e} (v_r - R_f S A I_0)$$

(5.10)
Interferograms of modulator array

(b) $V_x = 0$  
(c) $V_x = 0.8V_\pi$  
(d) $V_x = 0$

on all elements  
on elements marked  
"x"  
on all elements

Fig. 5.3 Nineteen element spatial phase modulator array  
(b and c have different reference beam tilts and cannot be directly compared)
More specifically, in the first IPL system operated, \( I_0 \) and \( I_1 \) were measured; and \( v_1, v_r (<v_1), v_H, G_0, \) and \( G_1 \) were specified. Given these constraints, the values of the variable resistors \( R_f \) and \( R_c \) were explicitly determined by Eqs. (5.12) below. From Eqs. (5.5a) and (5.6) with \( f = G_0/\phi_s \) and \( R_c + R_e = R_c \)

\[
I_f = I_0 - \frac{G_0}{G_1} I_1 \quad (5.11a)
\]

and

\[
f v_H = \frac{R_c}{R_e} R_f S A (I_0 - I_f) \quad (5.11b)
\]

The above equations together with Eq. (3.9b) specify \( R_f \) and \( R_c \)

\[
R_f = \frac{f}{S A} \frac{v_r}{I_0 - (1-f) I_f} \quad (5.12a)
\]

\[
R_c = \frac{R_e v_H}{v_r} \frac{I_0 - (1-f) I_f}{I_0 - I_f} \quad (5.12b)
\]

The actual circuit parameter values used and resulting \( \phi_m \) vs \( \phi_m \) characteristics are presented in the experimental discussion of Chapter 6. Satisfyingly, the first IPL system locked immediately upon turn-on!

5.2 Nineteen-channel System

A nineteen channel optically driven spatial phase modulator was constructed. The phase modulating element, shown in Fig. 5.3a, consisted of a slab of z-cut LiNbO\(_3\) (38.1 mm dia x 3mm thick) with an indium-tin oxide (ITO) transparent electrode on one side and a hexagonal array of 19 discrete aluminum electrodes on the other face. The detector array in Figs. 5.4 was fabricated by dicing up a silicon solar cell into 19
Fig. 5.4 Nineteen element detector array
(The illustrated channel numbers were chosen to simplify detector wiring.)
discrete detectors (using a wafer scribe), remounting their cathodes to a conducting surface, and attaching thin anode leads. The twenty-channel high voltage amplifier constructed for this system is shown in Fig. C.7 of Appendix C. A circuit diagram of a single channel is given in Fig. C.5 along with a more detailed discussion of the circuit.

In operation, a phasefront passes through the modulating crystal, reflects from the Al electrodes and passes back out. The expected half-wave voltage for this reflection geometry, with beam propagation and the modulating field both along the c-axis is

\[ v_\pi = \frac{\lambda}{2n_0^3 r_{13}} = 3100 \text{ volts (LiNbO}_3, \lambda=633 \text{ nm)} \] (5.13)

However, the measured half-wave voltage was significantly larger and appeared to vary over the surface of the crystal; ranging from 6 to 7 kV, or nominally \( v_\pi \approx 6500 \) volts. As mentioned in Appendix 3.7, discrepancies between published and measured half-wave voltages are not unprecedented; poor electrode contact does not seem to account for the larger measured half-wave voltage, because the observed crystal surface conductivity provided a uniform intimate contact with the crystal.

M. Smith polished the crystal and applied the electrodes. The crystal had a 3° wedge angle between the two surfaces to allow the first surface reflection to be spatially filtered from the modulated rear reflection.

Figures 5.3b-d are interferograms of the modulator array, made with the interferometer diagrammed in Fig. 6.12. In b & c the reference beam is tilted. Figures b and d, with \( v_x=0 \), provide an indication of the overall crystal flatness (\( \sim \lambda \) from left to right, \( 2\lambda \) from top to
bottom, and $\lambda/2$ over the center 7 elements). Notice in b, that the metal pad reflections are shifted by $\pi$ relative to the reflection around the pads. In c, the fringes from the four modulated pads ($v_{x} \sim 0.8v_{\pi}$) almost coincide with the surrounding reflection.

The modulator was limited to a dynamic range of about $2\pi$ radians by arcing on the surface of the crystal between the modulator elements. Arcing occurred rarely with $v_{x}=13$ kV and every few seconds with $v_{x}=15$ kV. The utility of this system in adaptive optical applications could be greatly improved if another modulator were fabricated more carefully from a LiNbO$_3$ crystal with $v_{\pi}=3.1$ kV; then, with $v_{x}<16$ kV, more than $5\pi$ radians of modulation would be available.

Based on Eqs. (C.26) - (C.28) in Appendix C and $\phi_{m}=v_{x}/v_{\pi}$, a differential equation can be written for the modulator phase as a function of the interferometric driving irradiance, $I=I_{0}+I_{1}\sin \phi_{e}$

$$\dot{\phi}_{m} = w_{L} (-\phi_{m} + G_{0} + G_{1}\sin \phi_{e}) \quad (5.14)$$

where

$$G_{1} = gI_{1} = \frac{\pi}{v_{\pi}} \frac{u}{R_{2}} R_{2}^{1} R_{1}^{2} S_{A}I_{1} \quad (5.15)$$

$$G_{0} = \frac{\pi}{v_{\pi}} (v_{x0} - \frac{u}{R_{2}} \frac{R_{1}^{1} + R_{2}^{2}}{R_{2}^{1}} v_{1}) + g(I_{0} - I_{1}) \equiv g(I_{0} - I_{f}) \quad (5.16)$$

$$I_{1} = \frac{R_{1}^{1}}{R_{f}^{1} S_{A}} (R_{f}^{1} v_{f} - v_{1}) \quad (5.17)$$

and

$$w_{L} = \frac{1}{(R+R_{1})C} \quad (5.18)$$
This is valid for approximately $0 \leq \phi_m \leq \phi_s = \pi v_H/v_\pi$ and $I_t' \leq I_s$. The actual threshold $I_t'$ is equal to $I_t$ only when $I_t > 0$, i.e., $I_t' = I_t U(I_t)$, where $U(\cdot)$ is the unit step function. The saturation intensity, $I_s'$, is given by Eq. (C.28). In Appendix C, it is also shown that $\mu = 1500$ are properties of the 6BK4C vacuum tube, $R_f' = R_f || 30 \Omega$, $R_f' = 3$ volts, and $S$ is the sensitivity and $A$ is the area of the photodetector. In addition $R_1$, $R_2$, $R_f$, $R$, $C$, $v_1$, and $V_f$ ($-v_1 \leq V_f - v_1$) are defined in Fig. C.5; and $r_2$ is the fractional displacement ($0 \leq r_2 \leq 1$), from the grounded end, of the variable tap on $R_2$. Equation (5.16) defines the offset intensity

$$I_f = \frac{-1}{R_f' \cdot SA} \left[ \frac{r_2}{R_1 + R_2} v_{\pi 0} - \frac{R_f'}{R_r} v_f \right] = \frac{1}{R_f' \cdot SA} \left( R_f' - \frac{R_f'}{R_r} V_f \right)$$

Notice in Fig. C.5 that each channel has three potentiometers which independently control bandwidth ($R$), gain ($R_2$), and offset/threshold ($R_0$). $R_0$ is a ten-turn potentiometer mounted on the front panel, which controls $V_f$ and allows accurate offset adjustment, or manual control of $\phi_m$ (when $I = 0$), in each channel. The $R$ and $R_2$ potentiometers are trimpots, mounted to be easily accessible with a screwdriver. Additional construction and operational details of the twenty-channel amplifier are presented in Appendix C.

5.3. Twenty-channel large dynamic range system

Many IPL applications, such as the Zernike phase-contrast IPL, visualization of thick (multiple $\pi$ radian) phase objects, and continuous tracking of real-time turbulence, would benefit from a large ($>4\pi$ radian) dynamic range system.
By employing a transverse modulator with $v_x$ applied along the z-axis and light propagation along the y axis (e.g., single-channel system of Section 5.1), much lower half-wave voltage can be obtained. The 30 element prototype modulator illustrated in Fig. 5.5 was constructed from a block of LiNbO$_3$ (polished by R. Dillon) with a dielectric mirror ($\text{MgO} + \text{SiO}_2$ layers for 6328Å) on one face. The mirror surface was glued to a glass block cut from an optical flat; and the whole assembly was sliced (by J. Thackara) with a diamond saw into five slabs, as shown in Fig. 5.5a. Six modulator elements were defined per slab by evaporating aluminum stripe electrodes on the top of each slab and a ground plane on the bottom. In a scheme suggested by C. Warde, the glass section of each slab was cut to different lengths to allow direct electrode attachment to each element. The crystal had a wedge angle between the mirror face and front face order to eliminate the front reflection.

For light polarized along the z axis, the expected reflection half-wave voltage,

$$v_x = \frac{\lambda}{2n_0^3 r_{33}} \frac{t}{\ell} \quad (5.20)$$

is 350 volts; $v_x = 380$ volts was measured. The dynamic range, $\phi_s$, was limited by arcing between electrodes on the crystal and glass surfaces. Of all the insulators tried, Krylon corona dope (50 kV/mm breakdown) on a well cleaned surface was the most effective. It withstood 3.5 kV; however arcing at the crystal-mirror-glass interface occurred at 2 to 2.5 kV, which limited $\phi_s$ to $5\pi$ radians.

Due to the square aspect ratio ($w = h = 4.75$ mm), it was expected that
Fig. 5.5 Thirty-element spatial phase modulator array
fringing fields would have a strong influence, necessitating the cutting of each slab into individual modulators. However, the modulator has not yet been diced into 30 elements because experiments with the 5 slabs revealed that it is extremely difficult to optically align individual elements to the accuracy demanded by adaptive wavefront correction. (If the elements are aligned within $\lambda/2$ of a common plane (piston aberration) and with less than $\lambda/2$ variation across an element (tilt aberration), the IPL can compensate the modulator to produce flat wavefronts.)

A variety of promising techniques were developed for optically aligning the modulator elements. In one technique the slabs were stood on an optical flat with as little tilt as possible and then glued around the edge with a fast setting rigid glue (Elmers "wonder bond"). Figure 5.6, which is an interferogram of mounted test slabs cut from an optical flat, shows approximately the minimum tilt obtainable with this procedure. In a slightly modified technique, the five modulator slabs were glued with a flexible glue (Duco "Weldit"), which allowed the tilt to change. To fine-tune tilt, the assembly illustrated in Fig.5.5b,c was constructed by J. Thackara. Fine-pitch screws were incident on the left, right, and protruding step of each slab, gravity and the electrical leads provided a restoring force in the vertical direction. Slow creep over time proved to be a major difficulty.

Figures 5.7(a), (b), (c) show the slabs before fine tilt alignment; the reference beam is adjusted approximately parallel to the center slab. In(a) and (b) the third element from the right in the center slab is modulated ($v_x = 1370V$ and $400V$) while the other elements are clamped at $v_x = 0$. Fringing of the fields is evident. (The right side of the top slab is damaged).
Fig. 5.6 Alignment of glass slabs

(a) \( v_x = 130 \text{ volts} \) on center element
(b) \( v_x = 400 \text{ volts} \) on center element
(c) \( v_x = 0 \text{ volts} \) on center element

Fig. 5.7 Interferograms of thirty element modulator array
Due to the alignment difficulties, it is proposed to only partially cut each slab into individual elements; starting from the glass side and stopping about 3 mm short of the front face. By inserting teflon spacers between slices and etching the electrodes from the uncut 3mm, $v_x$ in excess of 5kV and $\phi_m > 10\pi$ should be achievable with this modulator. Because the slices are vertically aligned, only five teflon sheets are required.

Actually the uncut assembly shown in Fig. 5.7a can be employed as a 15 element modulator, by operating the elements in pairs and blocking out the interelectrode areas with opaque strips. If each slab is operated as a unit, this device becomes a very large dynamic range 5 element linear array. The 0.1 mm teflon insulation between slabs should easily withstand 8 kV, corresponding to $\phi_m > 20\pi$ radians.

The square aspect ratio simplifies modulator construction, particularly when the elements are individually diced out. In addition arrays of square detector elements are commercially available. A 5x5 array of 25 square diodes was purchased from Centronic (# MD25-0, 3mm x 3mm, 0.1mm separation).

As a suggestion for future work, a superior discrete modulator could be built by starting all over and dicing an LiNbO$_3$ block in both directions, but stopping short of the front face to expedite alignment. Electrode contact could be made by inserting aluminized teflon strips between elements in the z direction; external contact could be effected by using strips with various amounts of overhang out of the cut. As above, teflon sheets can provide insulation in the x direction. Finally, the whole assembly can be encapsulated in epoxy for rigidity and durability.
This assembly should be able to withstand voltages in excess of 10 kV and corresponding modulation >25π radians! (A typical diamond saw cut width is ≈0.6 mm and teflon breaks down at 79 kV/mm.) As discussed in Appendix B.7 nonuniform contact between the aluminized teflon strips and the crystal may be problematical. Crystal surface conductivity, however, can compensate for nonuniform contact, depending on how the crystal surface is prepared.

5.4. A proposed medium-resolution discrete-channel system

Employing the proposed modulator described above, it may be possible to construct a medium-resolution (hundreds of elements) discrete channel system. Large numbers of elements are easily fabricated by dicing the crystal in both directions; and the simplicity of the IPL allows the channel electronics to be simplified to merely a resistor, photodiode, and modulator element, as illustrated in Fig. 5.8a.

It is apparent from the circuit model in Fig. 5.8b that this system has two poles, however \( R_d << R_f || R_L \) is generally true, allowing the frequency response to be modelled by the single-pole expression:

\[
\tilde{V}_x = \frac{w_o}{j\omega + w_o} \left( \frac{R_f}{R_L + R_f} \tilde{V}_H + R_i || R_f \right) S A I
\]  

(5.21a)

where

\[
w_o = (R_L || R_f)(C_d + C_x)
\]  

(5.21b)

A typical value of \( C_d \) is \( \approx 10^{-11} F \) (at \( V_H - V_x = 0 \)) for a 2 mm x 2 mm diode. (The change in \( C_d \) with \( V_x \) will not affect low frequency operation.) A typical value of \( C_x \) is \( 3.5 \times 10^{-12} F \) (LiNbO\(_3\) with \( V_H = 150 V, \lambda = 13 \) mm and
Fig. 5.8  Simple discrete-channel system
In order to keep the dark signal small, the maximum gain is limited by the constraint \( R_f \ll R_L \). Diodes with \( R_L > 5 \times 10^6 \) are available. With \( R_f \ll R_L \), the response to the intensity \( I = I_0 + I_1 \sin \phi_e \) can be approximated by

\[
\dot{\phi}_m = w_0 (-\phi_m + G_0 - G_1 \sin(\phi_m - \phi'_1))
\]

(5.22)

where

\[
G_0 = g_0 I_0
\]

(5.23a)

\[
G_1 = g_0 I_1
\]

(5.23b)

\[
g_0 = R_f S_A \pi / v_\pi
\]

(5.23c)

and

\[
w_0 = R_f (C_d + C_x)
\]

(5.23d)

This is valid for: \( 0 < I_S = v_H / (R_f S_A) \) and \( 0 < \phi_m < \phi_S = \pi v_H / v_\pi \).

The saturation intensity level, \( I_S \), and maximum modulator phase, \( \phi_S \), are limited by the breakdown voltage of the photodiode; values as large as 600 volts are available. (These high-voltage photodiodes are fabricated by a planar diffused technology, which may make integrated fabrication of the whole array on a single silicon wafer feasible.)

Similar IPL systems, with no intensity offset or threshold, were discussed in Section 4.3, in the context of systems employing monolithic spatial light modulators. In fact, Fig. 4.14 can be regarded as the state-space plot (\( \dot{\phi}_m \) vs \( \phi_m \)) for this system with \( v_H = 600 \) volts, \( v_\pi = 150 \) volts,
and \( |E_i| = 80\% |E_o| \) (i.e. \( \phi_s = 4\pi \), \( G_o = 3.5\pi \), and \( G_i = 3.42\pi \)).

This chapter described discrete-channel, optically-addressed phase modulators constructed in order to investigate IPL adaptive optical systems. Only the characteristics of the open-loop detector-to-modulator hardware have been presented here. Chapter 6 will report the results of incorporating the single-channel and nineteen-channel devices into actual closed-loop optical systems which demonstrate most of the essential features of IPL performance.
VI EXPERIMENTAL ADAPTIVE OPTICAL SYSTEMS

This chapter will describe a variety of experimental adaptive optical systems which were constructed in order to demonstrate the practicality of the Interference Phase Loop technique in phase estimation and compensation applications. Much of the basic IPL theory of chapters 2 and 3 was experimentally verified with a single-channel system. Wavefront shaping and phase compensation were demonstrated with a nineteen-element system. Although no truly high-resolution system has been constructed, some all-optical phase compensation results have been obtained with a monolithic SLM, the Microchannel Spatial Light Modulator discussed in Chapter 4.

6.1. Single-channel system

A single-channel homodyne IPL was constructed by using the optically-driven phase modulator of Section 5.1 in conjunction with a Mach-Zehnder interferometer, as diagrammed in Fig. 6.1. This system was employed to experimentally demonstrate that the IPL can perform real-time estimation and compensation of phase fluctuations over multiple \[ \pi \] radians of dynamic range, even in the presence of amplitude fluctuations. Other aspects of the theory of Chapters 2 and 3, such as recovery from modulator cutoff and saturation and the dynamic modulator state-space behavior \((\delta_m, \psi_m)\), were also experimentally verified.

6.1.1. Aberrators for performance evaluation

The system was evaluated with independent amplitude and phase aberrating media. Controlled static phase shifts were produced by laterally moving a warped transparent plate (e.g., a microscope slide) through the incident beam with a micrometer stage. Larger phase shifts were produced by tilting a glass plate in the beam. (The beam was flat
Fig. 6.1 Single-channel homodyne interference phase loop
over the distance it was laterally shifted as the plate tilted.) The open-loop interferometer was used to measure the phase shift as a function of the tilt and position. The tilt was controlled by monitoring the position, on a distant screen, of a beam reflected from the glass plate.

Amplitude-only modulation was produced with a rotating polarizer in the polarized laser beam or a commercial KD*P electrooptic modulator (Lasermetrics model # 3078M2). The rotating polarizer was mounted in an angularly-calibrated rotating ring, and was cantilevered from off the optical table to allow manipulation without vibrating the interferometer. The KD*P modulator was operated between parallel polarizers; producing the amplitude modulation \( \cos(\pi v/v_\pi) \), and no phase modulation (with applied voltage, \( v \)).

Simultaneous amplitude and phase aberration was produced by laterally shifting objects, such as partially aluminized slides or photographic transparencies, through the beam.

Low frequency dynamic phase fluctuations (=0 to 15 Hz) of variable magnitude (<\( \pi /4 \) to >\( 8\pi \)) were produced by placing a hot-plate or variable heat soldering iron in the signal beam. Allowing the free circulation of room air through the interferometer also produced similar turbulence; cardboard shields were employed when air turbulence was undesirable.

Wideband (approximate 0 to 200 Hz) phase fluctuations were produced by allowing mechanical vibrations into the interferometer. Their magnitude, (\( \pi /25 \) to >\( \pi \) radians rms), could be controlled by the construction of the interferometer, time of operation, and flotation of the optical table. Surprisingly, a floating table during the day had significantly more vibration than a nonfloating table at night (12 am - 5 am).
6.1.2 Experimental performance of single-channel system

Initially, all the parameters of the amplifier in Fig. C.1 were carefully chosen to result in a particular $\phi_m$ versus $\phi_n$ characteristic. The voltages in Fig. C.1 were adjusted to $v_1=30$ volts, $v_r=15$ volts, and $v_H=600$ volts ($\phi_S = v_H/v = 3\pi$). According to Eq. (5.2a), since $v_r < v_1$ there was no intensity threshold; but there was an offset in the detector signal. The measured intensities $A_{I_0}$ ($\sim 40 \mu W$) and $A_{I_1}$ ($\sim 10 \mu W$) were used with the desired values of $G_0$ and $G_1$ to specify the adjustable system parameters $R_c$ and $R_f$ through Eqs. (5.12). As expected the parameter values turned out not to be critical; and real-time procedures were developed to characterize and optimize the system performance while in operation. Typical values of $G_1$ ranged form $2\pi$ to $40\pi$ and $G_0$ from $-3\pi$ to $3\pi$. A representative modulator state-space characteristic ($\phi_m$ versus $\phi_n$) for the single-channel IPL was given in Fig. 5.2. (For most of the results reported here $I_1 < I_0 - I_I$ held)

Figure 6.2 shows the closed-loop performance of the single-channel IPL (with $G_1=11.4\pi$ and $G_0=2.9\pi$) in response to calibrated phase shift induced by tilting a microscope slide. The upper trace is $-v_0$ from the circuit of Fig. C.1 and is proportional to the interferometer intensity, $I$. The lower trace is $v_x$, which is proportional to $\phi_m$. In Figure 6.3c the loop was opened, (by disconnecting the modulator), in the presence of the same phase variation; $v_x$ can be viewed here as a low-pass filtered and amplified version of $I$.

In order to interpret the behavior to the IPL it is helpful to recall from Eqs. (2.1)-(2.5) and (2.18a) that $I = I_0 + I_1 \sin \phi_e$, where:
Fig. 6.2 Basic phase estimation and compensation behavior of the IPL.
\(G_0 = 2.9\pi, G_1 = 11.4\pi, \phi_s = 3\pi\)

Fig. 6.3 Behavior of the open-loop interferometer.
(to approximately the same phase variation as in Fig. 6.2)
\[ \phi_e = \phi_i - (\phi_2 + \phi_0) - \phi_1 - \phi_2 = \sin^{-1} \left( \frac{\phi_m G_o}{G_1} \right) + 2n\pi = 2n\pi \] (6.1a)

and
\[ \phi_e \equiv \phi_i^\prime - \phi_m \equiv \phi_c - \phi_1 - \phi_2 \] (6.1b)

Here \( \phi_0 \) is the intrinsic phase of the modulator, \( \phi_1 \) is the phase incurred in transit from the modulator to L.O. beam combiner, \( \phi_2 + \pi/2 \) is the L.O. phase, \( \phi_c = \phi_i - \phi_m - \phi_0 \) is the compensated phase, and \( n = \text{INT}(\phi_e + \pi)/2\pi \).

Despite significant variation in \( \phi_i^\prime \) (up to 2\pi radians), the expected stability of the phase error \( \phi_e = 2n\pi \) is evident in the upper trace of Fig. 6.2. Since \( \phi_e \) is approximately constant, the lower trace, \( \phi_m = \phi_i^\prime - \phi_e \), is essentially an estimate of the aberrator phase, \( \phi_i \). (This estimate is relative to the system reference phase \( \phi_0 = \phi_0 + \phi_1 + \phi_2 \), i.e., \( \phi_{\text{est}} = 0 \) when \( \phi_i = \phi_b \).) In fact, this single channel IPI is a useful real-time phase measurement device\(^7\), by scanning a transparent object through the beam, spatial phase variations can also be mapped. Since \( \phi_m \) is proportional to \( v_x \), \( v_x \) is an electrical representation of the phase estimate.

Figure 6.2 is also a good demonstration of adaptive phase compensation. In response to the \( \phi_i \) variation in the lower trace, the upper trace, which is essentially an interferometric representation of the compensated phase (\( \phi_c \) in Fig. 6.1), is only slightly perturbed.

Figure 6.3 illustrates the open-loop (modulator disconnected) response to approximately the same phase variation as in Fig. 6.2. The upper trace is essentially an ordinary interferometric phase representation, truncated to the range \(-\pi/2 < \phi_e < 0\). Note that this elementary interferometer cannot continuously track in excess of \( \pi \) radians and does not distinguish increasing from decreasing phase.
IPL operation can generally be easily measured and optimized in real-time. The basic measurement and adjustment procedures are presented in Appen. C.1.2, where it is shown that the voltage at \( v_o \) in Fig.C.1 can be written as \( v_o = v_{oo} + 0.5 v_{o\pi} \sin \phi_e \). By simultaneously monitoring \( v_o \) and \( v_x \), as in Figs.6.2 and 6.3, \( v_{oo}, v_{o\pi}, \Delta v_x/\Delta v_o \), and \( v_{x0} \) (the value of \( v_x \) when \( v_o = v_{oo} \)) can be measured and used to infer \( G_o \) and \( G_1 \) through Eqs. (C.16).

The basic optimization procedure is to adjust \( R_e \) to make \( \Delta v_x/\Delta v_o \) (and \( G_1 \)) large, and then adjust \( R_f \) to make \( v_{x0} = v_H/2 \), (i.e., \( G_o = \phi_s/2 \)).

The left half of Fig. 6.4a shows closed-loop IPL operation in the presence of turbulence created by a hot plate. The lower trace \( (v_x) \) is an estimate of the turbulence phase and the upper trace \( (-v_o) \) is a representation of the compensated phase. The uncompensated phase \( (-\sin \phi_i) \) is illustrated on the right side of Fig. 6.4a, where the loop was opened (with switch \( S_2 \) in Fig. C.1). Compensation of more severe (higher temperature) hot plate turbulence is depicted in Fig. 6.4b.

Figure 6.5 displays the closed-loop and open-loop \( (v_H=0) \) response to "wideband" (0-150 Hz) phase fluctuations with two different loop-filter bandwidths. Notice that with large bandwidth \( (\sim 1600 \text{ Hz}) \), all the fluctuations are tracked by \( \phi_m \) and compensated out of \( \phi_c \). With a narrower filter bandwidth \( (\sim 10 \text{ Hz}) \), the out-of-bandwidth spectral components are tracked by \( \phi_m \) with greatly diminished gain and remain uncompensated in \( \phi_e \). This behavior is in agreement with the linearized theory of Chapter 2, where from Eqs. (2.36) and (2.37),
Fig. 6.4 IPL performance with air turbulence from a hot-plate
Fig. 6.5 IPL performance with wideband phase fluctuations
($G_1=20\pi$, two different loop-filter bandwidths: $f_1=1600\text{Hz}$, $f_2=10\text{Hz}$)
\[
\tilde{\phi}_m(w) = \frac{w_L G_1}{jw + w_L G_1} \tilde{\phi}_i'(w)
\]
(6.2a)

\[
\tilde{\phi}_e(w) = \frac{jw}{jw + w_L G_1} \tilde{\phi}_i'(w)
\]
(6.2b)

The sum of these signals is independent of frequency, i.e., \(\tilde{\phi}_m + \tilde{\phi}_e = \tilde{\phi}_i'\).

At low frequencies, \(\tilde{\phi}_m = \tilde{\phi}_i'\) and \(\tilde{\phi}_e = 0\); whereas at high frequencies, \(\tilde{\phi}_m = 0\) and \(\tilde{\phi}_e = \tilde{\phi}_i'\).

The closed-loop and open-loop (modulator disconnected) response to a large increase in phase followed by a large decrease (approximately \(\pm 7\pi\)) is illustrated in Fig. 6.6. Jumping to a new trajectory after \(\phi_m\) saturation or cutoff, in accordance with Fig. 2.5, is clearly evident.

The expected restart phases, \(\phi_{mc}\) and \(\phi_{ms}\) after cutoff and saturation respectively, are given by Eqs. (3.22) & (3.23) as a function of \(G_0, G_1\), and \(\phi_s\).

In Fig. 6.6a \(\phi_{ms} = 2.0\pi\) and \(\phi_{mc} = 1.1\pi\), which agrees fairly well with the calculated values, \((\phi_{ms} = 2.1\pi\) and \(\phi_{mc} = 1.1\pi\)) for \(G_0 = 2\pi\) and \(G_1 = 11\pi\).

(Notice that the measured values of \(\phi_{ms}\) and \(\phi_{mc}\) from a \(v_x\) trace provide an alternative means for determining \(G_0\) and \(G_1\).) The variation of \(\phi_{mc}\) and \(\phi_{ms}\) in Fig. 6.6a is due to \(\phi_i\) varying rapidly enough to change while the IPL is relocking. The asymmetry in the upper trace is due to intensity saturation.

As expected, when \(\phi_m\) is active \((0 < \phi_m < \phi_s)\), \(\phi_e\) is compensated and changes very little. Also note that, in contrast to the open-loop system in Fig. 6.6b, in the IPL there is no ambiguity as to whether \(\phi_i\) is increasing or decreasing. Figure 6.6a also illustrates bistability in that the same values of \(\phi_i\) occur during the phase decrease as during the increase, but the \(\phi_m\) values are about 2\(\pi\) lower.
Fig. 6.6 IPL performance with large phase variations.
($G_0 = 2.1\pi$, $G_1 = 11.2\pi$, $\phi_s = 3.12\pi$)
Figures 6.7 demonstrate the insensitivity of the homodyne IPL to signal amplitude variations. In (a) the IPL phase estimate remained unperturbed while the intensity of the signal beam was varied by more than an order of magnitude with an electrooptic amplitude-only modulator. The last segment on the right in (a) and (b) corresponds to no signal intensity. Notice that the error in (a) (Δφ_e = Δφ_m, since \( \phi_1 = \text{const} \)) suddenly begins to grow at very small intensities, which is in agreement with the discussions in section 3.1.1.

Figure 6.7b illustrates that the interferometer intensity also remains unperturbed as the signal amplitude varies. This can be understood by using Eq. (2.23) with Eqs. (5.5a) and (5.7) to write: \( I = I_f + \phi_m/g \). Since \( I_f \) and \( g \) are fixed system constants, \( I \) remains constant as long as \( \phi_m \) does. At small intensity, the error grew less in Fig. 6.7b than 6.7a, because in 6.7b \( \phi_1 \) was closer to \( G_0(0) \) of Eq. (3.6c). In Figs. 6.7 \( |E_1| < 0.5 |E_x| \) always held; according to section 3.1.1 the maximum value of \( |E_1| \) at which this IPL can function is about \( 2 |E_x| \).

IPL state-space trajectories of \( \dot{\phi}_m \) vs \( \phi_m \) (actually \( \dot{v}_x \) vs \( v_x \)) are depicted in Figs. 6.8a and b. Fig. 6.8a is the low-pass IPL response to a step decrease in \( \phi_1 \); \( \phi_m \) moves from right to left along a sinusoidal loop, similar to those in Fig. 6.2a. The structure on the right is due to rapid fluctuations of \( \phi_1 \) before the step. Figure 6.8b illustrates the turn-on transient (left to right) with a frequency response characterized by more than one pole. The upper intensity saturation limit on \( \phi_m \) is clearly visible.
Fig. 6.7 IPL immunity to signal amplitude fluctuations
Fig. 6.8 Experimental IPL state-space trajectories

Fig. 6.9 Modulator state-space ($\dot{m}$ versus $\phi_m$) with inverted modulation
A variety of additional experiments were carried out. The IPL system of Fig. 5.1 was employed to measure both the attenuation and phase shift of transparent objects shifted through the signal beam.

In another experiment, the electrical connections to the modulator were reversed. As expected from the applicable $\phi_m$ vs $\phi_m$ state-space plot of Fig. 6.9, compensation still occurred, but with $\phi_e = (2\pi - 1)\pi$ and $\nu_x$ estimating the conjugate phase, $-\phi_i$. In response to an increasing phase, e.g., tilting a glass plate away from normal incidence, $\nu_x$ decreased.

The IPL performed well with a wide range of gains. At lower gains, the convergence time constant $\omega G_1$ was longer and the error phase larger; for example with $G_1 = 2\pi$, $G_o = 1.5\pi$, and $\phi_s = 3\pi$, the error ranged between about $-\pi/4$ and $+\pi/4$ radians. At larger gains, in excess of $G_1 = 40\pi$, electrical noise at 60Hz and other frequencies became a significant component of the detector-to-modulator signal. Despite careful shielding, this noise was attributed to stray RF coupling, and AC powerline noise passed through the power supply. This noise prevented operation down to thermal or shot noise limits. Constant sources of background light did not degrade performance because they could be electronically offset in $G_o$.

In the future, a superior single-channel system could be constructed by employing one channel of the twenty-channel amplifier of Section 5.2 to obtain larger gain, $G_1$, and dynamic range ($\phi_s = 75\pi$ for $\nu_x = 15kV$). As additional future work, a phase measurement device could be constructed to accurately profile large phase fluctuations in thick objects, such as biologic cells. In addition, the configuration in Fig. 6.1,
with mirror $M_1$ replaced by a beamsplitter rotated by $90^\circ$, provides a 
noncontacting alternative to a stylus for accurately profiling reflective 
rough surfaces located at $P_1$ (about $20\mu m$ dynamic range, $\lambda/20$ precision).

6.2. **Nineteen-channel system**

A homodyne IPL with nineteen spatial resolution elements was 
constructed by incorporating the optically driven modulator of Section 5.2 
into an interferometer. Basic adaptive phase compensation of static and 
dynamic wavefront aberrations was demonstrated.

**Experimental IPL configurations**

The first system, which bears a resemblance to a Mach-Zehnder inter-
ferometer, is depicted in Figs. 6.10. This configuration featured 
distinct signal beam, compensated beam, and LO beam paths. Lens $L_2$ was
Fig. 6.10a Nineteen-element laboratory IPL (Version l-l).
optional, since the modulator and detector scale sizes were directly compatible. However, the spatial filter improved the interferogram contrast by eliminating undesired beams. The inversion of the modulator image at the detector was corrected by the modulator wiring. Additional optical alignment details are presented in Appendix C.2.3.

Figure 6.10c illustrates that, in the interferogram at the detector, the L.O. beam and compensated beam from the modulator are flipped versions of the collimated beam. This puts excessive demands on the collimated beam at $L_1$, and is responsible for the fringes around the edges of Fig. 6.11 (which is an interferogram taken in plane $P_2$ of Fig. 6.10a). Phase compensation experiments performed with the interferometer in Fig. 6.10a will be described later.

In order to obtain a flatter wavefront over more modulator elements, the improved Twyman-Green based IPL of Fig. 6.12 was also constructed. The advantages of this configuration include: a reduction in the number of potentially aberrating beamsplitters, and no relative flipping of the collimated beam between the L.O. and compensated beams. The overlap of the aberrated and compensated beam paths slightly diminishes the versatility of this system; however all the essential features of IPL operation remain unchanged. With $\phi_i$ as the aberrator phase, the error phase is still driven to a constant (from Eq. 6.1, $\phi_e=2\phi_i-\phi_m-\phi_0-\phi_x+2\pi n$). If $\phi_1+\phi_x$ is spatially uniform, so is the compensated phase $\phi_c$, i.e., $\phi_c=2\phi_i-\phi_m-\phi_0+\phi_1+\phi_x+\text{const}(x,y)$. If a beam was split out at point A in Fig. 6.12, its phase would be the conjugate of $\phi_i$, i.e., $\phi_A=\phi_i-\phi_m-\phi_0+\phi_c-\phi_x=\text{const}-\phi_i$. Note that the internal aberrations of the modulator, $\phi_0(x,y)$, are removed from $\phi_c$ and $\phi_A$. It is also easy to show that $\phi_c$ is still spatially uniform when different paths are
Fig. 6.10 Nineteen-element laboratory IPL (version I-1)
Fig. 6.11 Interferogram from nineteen-element IPL (Version I-1).

Fig. 6.12 Improved nineteen-element laboratory IPL (I-2).
(M₂ allows the same DET to MOD wiring as for version I-1; BS₂ attenuates the L.O. beam; P₁, P₂, P₃, P₄, P₅ are planes referred to in the text).
followed in each of the two passes through the aberrator. The interferograms in Figs. 5.3b, c, d were photographed in plane $P_2$ of this system.

6.2.1. Techniques for performance evaluation.

Individual IPL channels were evaluated by simultaneously monitoring the voltages $v'_g$ (proportional to $-\phi_m$) and $v_0$ (proportional to $-1$) in the circuit of Fig. C.5. This is analogous to the $v_x$ and $v_3$ measurements made on the single-channel system of the last section; and, as described in Appendix C, the real-time parameter measurement and optimization procedures employed were similar to those of the single-channel system. However shortcut procedures were also developed to allow rapid adjustment of all nineteen channels.

Monitoring the Fourier transform (or far field) of the compensated beam provides an indication of the performance of an adaptive optical system; particularly when beam steering or plane wave compensation is desired. In the former case the whole transform shifts; and in the latter case the zero-order power (plane wave component) increases, which could for example improve the performance of the communications receiver in Fig. 1.1.

Since the modulator samples the input wave periodically in space, the Fourier transform is an infinite two dimensional array of repeating transforms; as illustrated in Figs. 6.13. The reciprocal lattice vector of solid state physics (e.g., p. 61 in Ref. 235) is useful for finding the basis repetition pattern corresponding to an arbitrary regular modulator array. The Fourier transform photograph in Fig. 6.13e was taken in plane $P_1$ of the system in Fig. 6.10a with no aberrator.
**SPATIAL DOMAIN**

(a) 2-D view

(b) 1-D profile

\[ P = \text{modulator period (6.55mm)} \]
\[ d = \text{width of modulator element (5.79mm)} \]
\[ a = \text{diameter of modulator aperture (31.75mm)} \]
\[ f = \text{truncated compensated wave} \]
\[ g = \text{small-scale detail} \]

**FOURIER DOMAIN**

(c) 2-D view

(d) 1-D profile

\[ w_o = \text{sampling frequency } \left( \frac{2}{\sqrt{3}} \right) \]
\[ w_1 = \text{minimum spot diameter } (2.44 \cdot 2\pi/a) \]
\[ F = \text{transform spots } (F(f)) \]
\[ G = \text{envelope encompassing most of } F(g) \]
\[ G_1 = F(\text{uniform modulator element}) \]

(e) photograph

---

Fig. 6.13 Fourier transform of the compensated beam.
Fourier transform data was taken in the form of both photographs of the magnified transform and the time response of the intensity received by a small (sub diffraction-limited) detector centered on the zero order (or a higher order) transform spot. Film negatives were also made for many of the transform photographs; in the future these can be scanned to obtain one-dimensional intensity profiles through the transform.

Spatial phase compensation performance was also evaluated by comparing interferograms of aberrated and compensated wavefronts. Based on Eqs. (2.22) it should also be possible to obtain continuous grey-tone visualization of a large dynamic range ($\gg \pi$ radian) phase object inserted into the $\phi_i$ position of the operating IPL. Unfortunately the limited dynamic range of the 19 element system restricted the phase object to $< 2\pi$ radians, which was so much less than $G$, that the object phase had negligible contrast and was "compensated" rather than visualized. Some image compensation experiments were also attempted.

**Spatial Phase Aberrators**

Finding suitable spatial phase aberrators is difficult due to the low resolution of this discrete element system. As mentioned previously, an adaptive system must possess high enough resolution to impress a representation of the spatial phase distortion, to be estimated and compensated, back on the original corrupted wave. In a discrete element system, this essentially means that the Nyquist criterion must be satisfied, i.e., the spatial sampling frequency $1/P$, where $P$ is the distance between modulator elements, must exceed twice the maximum spatial frequency of the aberration to be compensated.
Each detector element then sees a shade of grey, rather than fringes (assuming no uniform tilt); and a correctable aberration has up to \( N \) uniform resolution cells, where \( N \) is the number of IPL channels (e.g., 19).

This restrictive condition illustrates the utility of high resolution adaptive systems. The nineteen-element system can only tolerate about 2.5\( \lambda \) phase variation across the aperture. However, to compensate even such mild aberrators as a microscope slide (\( \sim 5\lambda/\)inch) in the same 1.25" diameter aperture would require \( \sim 120 \) elements. Alternatively, a lens could be inserted between the aberrator and modulator to reduce the effective compensated aperture size.

The most readily available static phase aberrator is the intrinsic distortion of the phase modulator, \( \phi_o(x,y) \). As mentioned previously, the IPL removes \( \phi_o(x,y) \) from the compensated beam. Beamsplitter, \( \phi_1 \), and local oscillator, \( \phi_2 \), distortions can also provide aberrations for evaluation of the IPL. According to Eqs. (6.1), the IPL replaces \( \phi_o(x,y) \) by \( \phi_1(x,y) + \phi_2(x,y) \) in the compensated phase, \( \phi_c \), but removes all aberrations from \( \phi_e \). For example, if the local oscillator beam is tilted while the IPL is running, no change should be observed in \( \phi_e \); but the modulator should impress the tilt on \( \phi_c \) and cause the Fourier transform of \( \phi_c \) to shift. This corresponds to using the IPL to perform beam steering or shaping of the compensated wavefront.

The major external spatial phase aberrators employed were slightly distorted (<3\( \lambda/\)inch) optical windows and sheets of plexiglass warped over a hot-plate. (Surprisingly, aside from a uniform wedge angle between the two surfaces, plexiglass flat to about \( \lambda/\)inch was
obtainable). The major dynamic aberrator was a variable temperature soldering iron below the beam path. Phase fluctuations were produced with a bandwidth of approximately 0.5 Hz to 5 Hz, and with a dynamic range variable from $<\pi$ radians to $>2\pi$ radians. The heat source was positioned to produce a spatial variation of about 2.5 waves across the modulator.

6.2.2. **Experimental Performance of nineteen-channel system**

The first experiments involved an open-loop, non-IPL demonstration of phase compensation in order to gain insight into the basic capabilities and limitations of these laboratory systems. Phase compensation was accomplished by employing a hill-climbing procedure to optimize a sharpness function in a manner similar to the "image sharpening" approach $^{11,12,14,15,46}$. More specifically, the loop was opened by blocking the array detector in Figs. 6.10 and 6.12. Then, with minimum gain in each channel, the front panel offset controls were employed to manually vary $\phi_m$ while monitoring the zero order intensity $I_s$ at a diffraction limited detector in the Fourier plane ($P_1$). From initial conditions of $\phi_m=0$ in each cell, the channels were sequentially adjusted to maximize $I_s$. Multiple iterations over all the channels were performed until no further improvement was evident. Theoretically, $^{11,46}$ this procedure should find the global maximum $I_s$ of which the system is capable, corresponding to plane wave compensation.

These experiments provided a benchmark for evaluating the compensation behavior of the IPL. This hill-climbing procedure was also employed to investigate the suboptimal performance which resulted when the aberration scale size was too small to be properly compensated, or when the total modulator dynamic range was $<2\pi$ radians.
In addition, these experiments provided an indication of L.O. (local oscillator) beam quality. With a flat L.O. beam, an improvement should be evident in the interferogram at $P_2$ (in Figs. 6.10 or 6.12) when $I_s$ is maximized; L.O. beam aberrations will appear as new distortions (other than tilt).

Once $I_s$ has been optimized, the compensated and uncompensated waves can be compared by turning $V_h$ on and off. Figure 6.14a illustrates a fivefold improvement in $I_s$, the intensity seen at a diffraction-limited detector in the Fourier plane. The lower trace is the $I_s=0$ level, the upper trace corresponds to the sequence $V_h=0$ (uncompensated), $V_h=9$ kV, and $V_h=0$. Figure 6.14b shows marked sharpening of the Fourier image of the compensated wave relative to the aberrated wave. Figure 6.14c demonstrates beam steering; $I_s$ was optimized with the pinhole detector slightly below the lower spot of the uncompensated transform, resulting in a shift of the compensated transform in that direction.

Figs. 6.14a, b, c were obtained with the system in Fig. 6.10. Figs. 6.14 do not illustrate a global maximum of $I_s$ because the amperage limitation (1.25 ma) of the $V_h$ supply limited $\phi_s$ to less than $2\pi$ radians. (13 kV at 2.5 ma was required.) Due to this dynamic range inadequacy, the maximum value of $I_s$ was obtained with a few modulator elements at their limits (0 or $\phi_s$). Later experiments were performed with a 20 kV, 25 ma supply.

Compensation of static distorted wavefronts by the closed-loop IPL is illustrated in Figures 6.15. In Fig. 6.15a, the irradiance $I_s$ at a
diffraction-limited detector in the Fourier plane is seven-times more intense for the IPL-compensated wave than the corrupted wave. The sequence of events in that photograph is: uncompensated (IPL off, with $V_H=0$); IPL operating with $V_H=12.5$ kV; $V_H=0$; and dark level. The phase aberration was distorted plexiglass plus dynamic fluctuations due to room air turbulence. The maximum compensated irradiance level was about equal to the level with no distorter. (Achieving the full theoretical irradiance improvement of 19 times, would require a more severe aberration.) Figure 6.15b compares Fourier transforms of the corrupted and IPL-compensated wavefronts.

In the presence of temporal phase fluctuations, induced by a soldering iron, IPL adaptive compensation and stabilization of both the $\phi_e$ interferogram and the $\phi_c$ Fourier transform were demonstrated. The IPL-compensated and uncompensated ($V_H=0$) behavior of $I_S$, the irradiance at a diffraction-limited detector in the Fourier plane, are compared in Fig. 6.16a. The IPL results in a marked stabilization of $I_S$. The short trace in the bottom is the dark level ($I_S=0$). In Fig. 6.16b, both static and dynamic aberrations were present. The upper trace in Fig. 6.16b is the IPL enhanced and stabilized $I_S$ irradiance. The smaller magnitude trace in the middle is the uncompensated irradiance, and the lower segment is the $I_S=0$ level. Fig. 6.17a shows timed exposures of the Fourier plane in the presence of IPL compensated and uncompensated dynamic spatial phase fluctuations.

Fig.6.17b compares the closed-loop and open-loop ($V_H=0$) simultaneous behavior of $V_o$ (proportional to $-I$, lower trace) and $V_g'$ (proportional to $-\phi_m$, upper trace) in a single channel (#19). Notice that the interferometric
Fig. 6.14 Manual phase compensation by image sharpening.
(a) Intensity, $I_s$, at a diffraction limited detector in the Fourier plane. (This is essentially the Strehl ratio)
(b), (c) Photographs of Fourier transform of aberrated and compensated waves.
(a) Intensity at a diffraction limited detector in the Fourier plane, (Strehl ratio).

(b) Photographs of Fourier transform of aberrated and compensated waves. (This enlargement was made at the expense of grey-scale dynamic range. In the original, the center spot of the compensated image was much brighter than the surrounding areas and any of the spots in the aberrated image.)

Fig. 6.15 IPL phase compensation of mostly static wavefront aberrations.
(a) Intensity, $I_s$, at a diffraction limited detector in the Fourier plane.

(b) $I_s$ with static and dynamic aberrations

Fig. 6.16 IPL phase compensation of dynamic wavefront aberrations. (Air turbulence induced by a soldering iron.)
(a) Photographs of Fourier transform of aberrated and compensated waves. (The dynamic range in this enlargement is saturated. In the original, the center spot of the compensated image was brighter than the surrounding areas and any of the spots in the aberrated image.)

(b) Response of a single channel (#19)

Fig. 6.17 IPL phase compensation of dynamic wavefront distortions
intensity I is essentially constant when the loop is closed. This figure also demonstrates that the performance of each individual channel is essentially the single-channel homodyne behavior of section 6.1. The nineteen-channel IPL was operated with channel gains ranging over $5\pi \leq G_1 \leq 50\pi$ and with offsets of $-2\pi \leq G_0 \leq 2\pi$. Typical values for the optical power averaged over a detector element were $A_d I_0 = 0.2 \mu W$ and $0.04 \mu W < A_d I_1 < 0.4 \mu W$, where $A_d$ is the area of a detector element (0.32 cm$^2$).

Photographs were also taken of corrupted and compensated $\phi_e$ interferograms. Unfortunately the changes are not represented well in the photographs, due to the low contrast of the interferograms and low resolution of the distortions. Generally, the grey levels of the central elements tend to become more uniform, fringes are swept toward the edges, and broken fringes become more continuous.

On rare occasions spatial oscillations could be observed, with groups of modulator elements "shimmering." This is thought to be due to spatial overlap of the modulation due to each pad at the phase modulator. In particular, the large phase change ($> \pi$ radians) when one element recovers from saturation ($\phi_m = \phi_s$) or cutoff ($\phi_m = 0$) may perturb the phase in adjacent modulator elements enough to cause the IPL to enter or to leave saturation or cutoff. These oscillations became more severe when the dynamic range of the modulator was reduced to $\pi$ radians or less. In some instances beamsplitter resonances induced higher frequency ($> 50$ Hz) vibrations into the interferometer; however these could be excluded from the IPL with the low-pass filter.

**Image compensation**

Adaptive image compensation experiments were also attempted. Actually
the improved transforms of Figs. 6.14, 6.15, and 6.17 are elementary demonstrations of image compensation of an image whose far field is the modulator. Attempts were made to pass incoherent (color-filtered white light) and coherent images from P₅ to P₇ in Fig. 6.12. (Images incoherent with the IPL source introduce amplitude modulation which the IPL can be designed to tolerate. Polarization, spatial filter SF₂, or wavelength can also be employed to block the image from the control loop, with care that λ₁-λ₄ is approximately the same for the image and IPL.) Amplitude-only images could also be introduced at P₃ or P₄ in Fig. 6.12. A flat image is required at P₃ since phase distortions there will not be compensated from the output image at P₇.

In particular, attempts were made to introduce the far-field (Fourier transform) of an image at P₃ in Fig. 6.12, and reconstruct the compensated image at P₇. A Fourier hologram (Vander Lugt Filter) can be used for the far-field amplitude filter at P₃. The filter in Fig. 6.18a was made by photographing the Fourier transform of an image similar to the negative of Fig. 6.18c, which included a reference pinhole.

Due to the low resolution of the modulator, correctable aberrations (≤2.5λ/1.25") were too mild to corrupt any of the images tested. Since a correctable aberration only increases the modulator transform spots in Fig. 6.13c,d from about 2/a to 2/p in frequency diameter, it is not difficult to show that a transform which can be aberrated and compensated must satisfy the Nyquist criterion at the modulator. This corresponds to only about 19 resolution cells in both the image and transform.
Fig. 6.18 Far-field image (Fourier hologram) for image compensation experiments. (Recorded on Polaroid type 55 positive-negative film. Holograms with a variety of magnifications were made. The symmetry of (a) is due to the fact that the word "MIT" has thin lines orthogonal to 0°, 90°, and approximately ±45°.)
6.3 "All-optical" phase compensation with the MSLM

The visible-photon cathode MSLM was employed to build an "all-optical" IPL, with which basic phase compensation was successfully demonstrated. This closed-loop IPL system, pictured in fig. 6.19, employed the Fabry-Perot interferometer, resulting from reflections between the front and back crystal surfaces to compensate the nonuniform internal phase distortion, \( \phi_0(x,y) \), of the crystal. (The intensity as a function of \( \phi_m \) for interference between the surfaces of an uncoated LiTaO_3 crystal (\( R_1 = R_2 = 15\% \)) is illustrated in fig. B.13 of appendix B.3.)

IPL phase compensation was demonstrated with both the "timed-integration" and "inverted-secondary" MSLM framed modes of chapter 4.3. The continuous mode could not be employed due to the extremely long time constant (\( \tau_2 = \text{weeks} \)) of the visible-photon cathode MSLM.

The feedback path of fig. 6.19 was aligned by first storing a pattern on the crystal and then closing the loop in the timed-integration framed mode. When properly aligned the stored phase distribution is compensated; when misaligned the image is spatially propagated as a series of repetitions (for \( 2n\pi < \phi_0 + \phi_m < (2n+1)\pi \)) or alternating inverted repetitions (for \( (2n-1)\pi < \phi_0 + \phi_m < 2n\pi \)). The imaging lens, \( L_2 \) in figure 6.19, is required because the electron lens in the MSLM inverts and rescales (reduces by \( 0.89x \)) the write image. Fine registration was accomplished by moving \( L_2 \) and moving and tilting adjusting \( M_4 \).

6.3.1 Experimental performance of the timed-integration MSLM-based IPL

The operational theory of the timed-integration MSLM-based IPL with a Fabry-Perot interferometer is essentially the same as that presented
Fig. 6.19 "All-optical IPL employing the MSLM."
in section 4.3.2; except that the MSLM control-beam irradiance $I$, as a
function of $\phi_m$, is given by Eq. (B.57). Actually Eq. (B.57) and Fig. B.3.2b
reveal that, since $R$ is small ($R=0.15$), $I$ can be approximated by the
familiar form,

$$I = I_0 + I_1 \sin \phi_e$$  \hspace{1cm} (6.3)

where

$$I_0 = \frac{2R}{(1-R)^2}$$

and

$$\phi_e = \phi'_i - \phi_m$$  \hspace{1cm} (6.4a)

$$\phi'_i = \phi_0 - \frac{\pi}{2}$$  \hspace{1cm} (6.4b)

(Remember, $\phi_0(x,y)$ is the intrinsic phase of the modulator.)

Hence the state-space plots ($\dot{\phi}_m$ vs $\phi_m$ and $\dot{\phi}_e$ vs $\phi_e$) in Fig. 4.16 are
applicable to this experiment, ($I_0+I_1$ was small enough to avoid saturation). The more exact $\dot{\phi}_e$ vs $\phi_e$ characteristic from Eq. (B.57), with $\phi_e$ given by
Eqs. (6.4) ($\phi_e = -\frac{\pi}{2} - \delta$), is illustrated in Fig. 6.20. Due to attenuation in the crystal and background irradiance, the irradiance at $\phi_e = (3-4n)\pi/2$
is not zero.

The initial $\phi_e$ state-space, at the start of the frame with $\phi_m$ erased
to $\phi_n$, is depicted in Fig. 6.20. Since $\phi_0$ varies over $<2\pi$ radians,
$-\pi < \phi'_i \leq \pi$ is assumed. The Fabry-Perot interferogram corresponding to this
initial uncompensated spatial phase distribution is displayed in Fig.
6.21a. The compensated intensity after a minimum-time frame is in the
vicinity of the null at point B in Fig. 6.20. Minimum-time was obtained
by monitoring the Fabry-Perot interferogram for compensation of the
initial phase distribution without "escape" of resolution elements from
quasiequilibrium, (see section 4.3.2). In this experiment the limited
Fig. 6.20 Initial error state-space for the timed-integration framed IPL with a Fabry-Perot interferometer. (Initial condition: $\phi_m = -\phi_n$, $-\pi < \phi_e < \phi_n + \pi$)

Fig. 6.21 Phase compensation results with the timed-integration framed IPL. (c) shows the residual phase error, (d) shows the effect of stray background light)
modulator operating range, \((-0.44 \pi = \phi_n \leq \phi_m \leq \phi_s = 0.63 \pi, \Delta \phi_m = \pi\)\) restricted the compensatable phase error to the segment between A and B in Fig. 6.20. Fortunately the actual initial \(\phi_e\) distribution was encompassed by these bounds; the phase extremes at the crystal edge and center are denoted in Fig. 6.20.

The uncompensated and compensated interferograms in Figs. 6.21a, b were photographed with the same exposure. The longer exposure in Fig. 6.21c (twelve times longer) reveals that the residual phase error is very small indeed; the irradiance was barely visible by dark-adapted eye. The slight error intensity in the lower left of Fig. 6.21c is due to light reflected from the MCP. (This light also caused the fine fringes in Fig. 6.21a.) Neglecting that area, the compensated phase flatness was estimated (using Fig. B.2.2b) to be \(\phi_\gamma / 20\) (2\(\Delta \phi_e = \pi / 10\)). The bright ring around the outside of the image is an inactive region of the MSLM, which is shielded by the MCP support ring. Fig. 6.21d demonstrates operation in the presence of relatively high background irradiance (no light isolation chamber in Fig. 6.19); this irradiance caused the center region to escape from quasi-equilibrium.

6.3.2 Experimental performance of the inverted-secondary MSLM-based IPL

With the Fabry-Perot intensity approximation of Eqs. (6.3) and (6.4), the discussion of section 4.3.3 and the \(\phi_m\) vs \(\phi_m\) characteristic of Fig. 4.19 are applicable to the inverted-secondary, Fabry-Perot IPL. The corresponding initial \(\phi_e\) vs \(\phi_e\) characteristic is sketched in Fig. 6.22.

A Fabry-Perot interferogram of the initial error distribution, with \(\phi_m = \phi_s\), is shown in Fig. 6.23a. This is the same phase distribution as in Fig. 6.21a, but with a uniform phase shift of \(\Delta \phi_m = \phi_n + \phi_s = \pi\). The
inverted-secondary compensated distribution is illustrated in Fig. 6.23b. Figs. 6.23a and b were photographed with the same exposure as Figs.6.21a and b. The residual phase error, photographed with approximately 12x more exposure, is depicted in Fig. 6.23c. This residual error corresponds to \( \phi_e \) within the B to D dead region in Fig. 6.22. (The central ring resulted from a change in \( V_b \) and hence \( I_t \) for \( V_b < 0 \) \( (-\phi_n < \phi_m < 0) \).) Fig. 6.23d shows the compensated error with a lower threshold than in Figs. 6.23 b and c. The dead region is narrower and the flatness is improved; however the initial value of \( \phi_e \) at the edge was beyond D in Fig.6.22. This region could be compensated (at \( \phi_e = \frac{3\pi}{2} \)) by continuing the inverted-secondary mode until \( \phi_n + \phi_s \geq 2\pi \).

6.3.3 Attempted high-resolution MSLM-based IPL

Unfortunately this MSLM could not be directly employed for compensation of external high-resolution phase aberrations, due to the dominating Fabry-Perot interference pattern. While it is shown in appendix B.3 (see Fig. B.14) that a net usable phase modulation of approximately \( \phi_m/2 \) is available; appendix B.3 also reveals that the \( \phi_m \) vs \( \phi_m \) (i.e., I vs \( \phi_m \) in Fig. B.15) characteristic in an external interferometer (e.g., homodyne IPL) has undesirable \( \phi_m \) minima. These result in \( \phi_m \) solutions which are independent of \( \phi_i \) and thus do not compensate \( \phi_i \) in \( \phi_e \). In addition the dynamic range of this visible-phococathode MSLM is only about \( \pi/2 \) in this application.

However it was found that the Fabry-Perot crystal surface reflections could be removed by reflecting the readout beam off the MCP. A homodyne IPL was constructed by employing the reflection from BS\(_3\) in Fig. 6.19 as the local oscillator beam, and spatial filter SF\(_2\) to pass the MCP reflection while blocking the crystal reflections. (This can be accomplished due to an
Fig. 6.22 Initial error state-space for the inverted-secondary framed IPL (Initial conditions: $\phi_m = -\phi_s$, $-\phi_s - \pi < \phi_e \leq -\phi_s + \pi$)

Fig. 6.23 Phase compensation results with the inverted-secondary framed IPL. (c) shows the residual phase error, (d) shows compensation with a lower threshold)
overall tilt between the MCP and crystal. An interferogram taken at plane $P_3$ in Fig. 6.19 is displayed in Fig.6.24. Since the dynamic range of this MSLM is only about $\pi$ radians, the expected closed-loop behavior is merely a widening of the dark fringes at the expense of the light fringes, (or visa-versa if $\phi_x$ is changed by $\pi$ while the compensating $\phi_m$ is being stored). The likelihood of achieving this behavior is very low due to the extreme nature of the MCP distortion. This distortion exceeds the MSLM resolution over much of its area and requires ultra-precise alignment of the size and position of the feedback image. In addition, mechanical vibrations can destroy the interferogram while the loop is operating, particularly during the long cycle times required to obtain the desired threshold behavior of the inverted-secondary framed mode in the available visible-photocathode MSLM. (The four-fold symmetry of the MCP interferogram in Fig. 6.24 appears to be a property of the MCP manufacturing process rather than mechanical stress, since the same pattern was observed with ring and three-point support systems. Possible surfaces which could produce the observed fringe pattern include: a saddle, four narrow ridges with intervening valleys, or the inverse of the latter.)

These preliminary all-optical adaptive demonstrations were halted when it became apparent that new visible-photocathode MSLMs would soon be available with a wedged crystal, larger dynamic range, and higher MCP strip current (faster speed). In fact, a few of these MSLMs have recently been assembled; and truly high-resolution all-optical adaptive phase compensation is expected to be demonstrated in the near future.

This chapter has addressed some of the practical considerations involved with the implementation and evaluation of actual IPL systems. Besides placing the IPL theory of earlier chapters into a more concrete
Fig. 6.24 Interferogram of wavefront reflected from MCP in visible-photocathode MSLM.
perspective, the results of this chapter, in conjunction with chapter 5
and appendix C, are a source of useful experimental guidelines for future
workers in this area. Features of IPL operation which have been
demonstrated in this chapter include: phase estimation of static and
dynamic phase fluctuations over multiple $\pi$ radians of dynamic range,
insensitivity to amplitude fluctuations, modulator state-space behavior,
real-time adaptive compensation of spatially and temporally aberrated
wavefronts, real-time evaluation and optimization of IPL performance,
and monolithic "all-optical" operation with a spatial light modulator.
VII SUMMARY AND SUGGESTIONS FOR FUTURE WORK

Techniques and devices have been developed to allow the construction of high-resolution adaptive optical systems containing a million or more spatial resolution elements. This research was motivated by the possibility these systems offer for considerably enhanced performance in such applications as optical communications through low-visibility or turbulent atmospheric conditions, imaging through the atmosphere, and high energy laser systems. The previous technology, consisting of arrays of individual phase estimators hard-wired to modulator arrays, was found to be too cumbersome for achieving the required numbers of resolution elements. Instead, an "all-optical" parallel processing approach was pursued, which offers resolution as well as speed and sensitivity advantages over the previous approaches.

The major contributions of this work include: development of a new easily implemented phase measurement and compensation technique called an Interference Phase Loop (IPL); a detailed theoretical analysis of the homodyne, heterodyne, and Zernike phase-contrast interferometric implementations of the IPL, with the inclusion of nonlinear effects; development of the Microchannel Spatial Light Modulator (MSLM) from an early prototype stage to a usable device, along with an enhanced understanding of its operating theory; experimental demonstrations of the feasibility of IPL adaptive phase compensation, employing single-element and nineteen-element discrete-channel test systems; and successful demonstration of all-optical, although low resolution, phase compensation with an MSLM/IPL adaptive system. The major results and conclusions of this thesis will be reviewed in greater detail in the paragraphs which follow. The comments at the ends of most chapters
provide a more complete summary.

The applications overview in the introduction suggested that high-resolution adaptive optics will probably have the greatest utility for the compensation of atmospheric turbulence phase distortions in optical communications, imaging, and high-energy propagation applications, as well as a few other aberrations encountered in high energy systems and large optical systems. The demands of these applications, $10^2$ to $10^4$ resolution elements, $<k$Hz bandwidth, and quantum-limited sensitivity, are within the capabilities of the approaches of this thesis. These approaches can also achieve the spatial diversity, $10^6$ to $10^9$ resolution elements, required for phase-compensated low-visibility atmospheric communications. There is however some question as to whether the scattering-channel coherence time is typically long enough to allow an adaptive system to collect sufficient signal energy.

The basic Interference Phase Loop technique was shown to consist of detecting the output intensity from an elementary interferometric phase sensor, e.g. a homodyne interferometer, and feeding this signal back to drive a phase-only modulator in the path of the wavefront being measured and compensated. It was shown that this configuration can stably estimate and compensate phase in real-time over multiple $\pi$ radians of dynamic range without phase quadrant ambiguity, even in the presence of signal amplitude fluctuations. The IPL is generally much simpler to implement than alternative phase estimation systems offering comparable performance.

Chapter 3 provided an in-depth examination of IPL behavior, extending the analysis beyond the fundamental static and dynamic behavior introduced in Chapter 2 and considering three interferometric implementations:
homodyne, heterodyne, and Zernike phase-contrast. Topics discussed included: the effects of signal amplitude fluctuations, phase acquisition transient response, multistability and 2nπ degeneracy, spatial resolution requirements, IPL behavior with various detector and modulator frequency response characteristics, spatial phase compensation with the IPL, the effects of noise on IPL performance, and minimum signal requirements.

Such major operating characteristics as: steady-state phase estimation accuracy, phase acquisition time, phase tracking bandwidth, stability, and signal-to-noise ratio were considered for IPLs employing optically-addressed modulators with frequency responses characterized by one or two dominant time constants. It was seen that the bandwidth of the low-pass \(H(s)=\frac{w_0}{s+w_0}\) or integrating \(H(s)=\frac{w_0}{s}\) single-pole systems is approximately \(\frac{1}{w_0 G_1}\). Phase fluctuation components outside the IPL bandwidth (temporal or spatial) are not tracked by the phase modulator and contribute to the estimation and compensation phase error.

In many instances the loop filter is a fixed property of the optically-addressed phase modulator. Where there is a choice, however, the single-pole filters are preferable; since more complicated systems do not generally offer advantages which sufficiently offset their difficulty of analysis and adjustment.

It was shown that shot noise due to the signal, local oscillator, and/or background irradiance tends to introduce approximately Gaussian distributed fluctuations into \(\phi_m\) and \(\phi_e\) about their noiseless equilibrium values. These fluctuations have an rms magnitude of \(\sigma_{\phi}=(S/N)^{-1/2}\). A nonlinear analysis showed that the IPL signal-to-noise ratio, \(S/N\), must exceed a threshold value of approximately 0.5 in order that noise-induced jumping between adjacent system equilibria does not prevent phase acquisition.
The minimum required signal irradiance is generally a function of noise
and other performance constraints considered together. Ideally, at
least a few photons must be collected in a resolution cell during the
system integration time.

Analysis of the heterodyne interferometric implementation of the IPL
showed that, much like a conventional phase-locked loop, the IPL tends to
lock the compensated signal frequency to the local oscillator optical
frequency. This reduces the stability and frequency tolerance con-
straints on the local oscillator source. Heterodyne operation can also
improve performance by reducing the effects of the system offset, \(G_0\).
The heterodyne IPL often requires framed operation to prevent the
modulator phase from reaching its physical limits or entering into
limit-cycle oscillations.

The self-interference (Zernike phase-contrast) configuration of the
IPL, which requires no external reference beam, looks attractive for such
applications as astronomical image compensation. It was shown to be
operable in white light and to have an intrinsic "automatic gain control"
property which improves its insensitivity to amplitude fluctuations. It
also exhibits a "self-centering" feature which tends to maintain the
spatial average of the error at a preset value (e.g. zero), diminish
2\(\pi\) ambiguity, and reduce the dynamic range required by the phase modu-
lator. The principle disadvantage of this configuration is gain reduction
by approximately \(e^{-\sigma_1^2}\) during initial convergence, where \(\sigma_1\) is the uncom-
pensated rms spatial phase fluctuation.

Unlike the linearized investigations of many previous closed-loop adap-
tive optical systems\(^3,13,110,149,237\), the analysis in Chapters 2 and 3 inclu-
cluded such important IPL nonlinearities as the basic transcendental
nature of the operating equations, detector threshold and saturation, and phase modulator dynamic range constraints. Consideration of these nonlinearities is essential to the understanding of such fundamental IPL behavior as: initial phase acquisition, stability, phase-tracking over multiple \( \pi \) radian dynamic ranges, control and utilization of multistability, recovery after reaching the physical limits of the modulator, and the frequency-tracking range of the heterodyne IPL. In fact, a linearized analysis completely overlooks such important features as \( \phi_m \) limit-cycle oscillations in the heterodyne IPL and the fundamental S/N threshold due to noise-induced loss of phase-lock.

In some instances, an understanding of the nonlinear behavior allowed more meaningful linearized expressions to be written. For example, Eq. (2.26b) represents the multistable nature of the phase estimate, with the \( n \)-value being given by Eq. (2.25). Equation (2.26b) also shows that the phase error of the low-pass homodyne IPL depends on the multistability \( n \)-value. Equation (3.127) is a general linearized solution for the steady-state of the low-pass phase-contrast IPL, including the effects of spatial coupling and multistability.

In the absence of suitable alternatives among existing spatial light modulators, significant effort was devoted to development of the MSLM. The operating theory of Section 4.1 and Appendix B showed that the MSLM is capable of at least three distinct operating modes: continuous, electron-deposition framed, and inverted-secondary framed. Appendix B also presented a more detailed analysis of the secondary emission behavior of the MSLM and discussed a variety of image processing operations which the MSLM is inherently capable of performing. The performance of the MSLM was improved from a nonfunctional prototype stage to
usable devices with sensitive visible photocathodes (6% quantum efficiency at 655nm), framing rates of a few Hz, storage times of weeks, spatial resolutions up to 2 lp/mm, and dynamic ranges in excess of $\pi$ radians. A variety of performance improvements were suggested which should ultimately result in MSLMs with quantum-limited sensitivity, better than kHz framing rates, 20 lp/mm resolution, and multiple $2\pi$ radian dynamic range.

Section 4.3 showed that the major challenges of attempting to employ an MSLM in an all-optical IPL adaptive system include effectively utilizing the intrinsic MSLM operating modes and diminishing the detrimental effects of the $G_0$ phase-offset term. Possible techniques for reducing the effects of $G_0$ include: low-pass operation (continuous MSLM mode) with $\phi_m = G_0$, heterodyne IPL operation with $w_0 G_0 = \Delta \omega$, utilization of the intrinsic intensity threshold of the "inverted-secondary" MSLM operating mode or modifying other MSLM modes to exhibit threshold, and spatial or temporal filtering of the $G_0$ term.

The basic feasibility and much of the operational theory of the IPL were demonstrated with discrete-channel test systems. Although all-optical configurations are ultimately of greatest interest, discrete channels bring greater flexibility to laboratory investigations. Channels can be individually controlled and monitored, and a large variety of system time responses can be studied.

Since the resolution elements of a homodyne IPL are spatially decoupled, its basic operational theory could be verified with a single-channel optically-driven phase modulator. The experimentally demonstrated behavior included: real-time estimation and compensation of phase fluctuations over dynamic ranges in excess of $2\pi$ radians, insensitivity to
more than an order of magnitude of amplitude fluctuation, compensation of wideband phase fluctuations due to air turbulence and mechanical vibrations, recovery from modulator saturation and cutoff, multistability, and the $\phi_m$ versus $\phi_m$ state-space behavior.

A nineteen-element test system was employed to demonstrate IPL compensation of static and dynamic wavefront aberrations. Considerable improvement, corresponding to real-time "straightening out" of the wavefront distortions, was observed in the far-field (Fourier transform plane) of the IPL compensated wavefront. The intensity seen by a diffraction-limited detector in the Fourier plane, i.e. the Strehl ratio, was also enhanced and stabilized.

All-optical phase compensation employing an MSLM was successfully demonstrated. Unfortunately, only a low-resolution demonstration was possible, because the only available visible-photocathode MSLM employed a parallel crystal which resulted in undesirable intensity modulations in an external interferometer.

Future work

In the near future attempts will be made to employ recently improved MSLMs for high-resolution laboratory demonstrations of all-optical MSLM/IPL adaptive phase compensation. These new MSLMs include visible photocathode devices assembled in collaboration with Hamamatsu Corp. and demountable devices built by J. Thakara.

There is considerable opportunity for further improvement of the MSLM toward its ultimate resolution, speed, sensitivity, and dynamic range performance goals. A variety of areas for further MSLM research were suggested in Section 4.2.3. Adaptive IPL applications in particular could benefit from: employing an amorphous semiconductor coating or a
PLZT electrooptic plate to enhance the intensity thresholding capabilities of the MSLM, and extending the available phase modulation dynamic range by utilizing one of the optical techniques referenced in Appendix G.9.

Future high-performance MSLMs could be employed for impressive demonstrations of real-time high-resolution phase compensation both in the laboratory and in actual field applications, such as optical communications through turbulence (and possibly low-visibility conditions) and astronomical image compensation. A few potential laboratory image compensation experiments were suggested in Section 6.2.2. High-performance MSLMs could also be employed in a broad range of optical signal processing applications. Particularly significant results should be possible in areas where the intrinsic processing capabilities of the MSLM can be utilized, such as edge detection, hard-clipping thresholding, or optical binary logic.

Some important applications require modifications to the MSLM or that an alternative specialized optically-addressed modulator be developed. In astronomical image compensation, for example, a nondispersive phase modulator is desirable. One possibility is a modified MSLM which proximity focuses charge onto a deformable reflecting membrane instead of a crystal; this is related to Somers PEMLM\textsuperscript{200}. Optically-addressed modulators capable of functioning in high energy applications also need to be developed. One interesting compromise is to use conventional discrete-channel water-cooled modulators, but to greatly extend the number of realizable resolution elements by exploiting the simplicity of the IPL phase estimation technique. Reduction of the electronics in each channel to merely a high-voltage photodiode, resistor, and modulator element may be feasible, as suggested in Section 5.4.
There is also an opportunity for the development and demonstration of additional techniques for effectively utilizing an MSLM or other SLM in an all-optical IPL. A few potentially valuable schemes for reduction of the troublesome $G_0$ offset term were proposed in Section 4.3.5, such as employing a spatial or temporal carrier frequency (i.e. tilted L.O. beam or heterodyne operation respectively) in conjunction with an SLM having a spatial or temporal bandpass characteristic. It may also be worthwhile to investigate operating schemes which involve modulation, e.g. sinusoidal, of the MCP voltage or other MSLM bias voltage.

Another interesting area for further work is the theoretical and experimental investigation of the alternative high-resolution phase-compensation configurations proposed in Appendix F, which employ nonlinear optical processes. The four-wave mixing approach appears to be particularly promising, since it may potentially provide gain.

There are a variety of theoretical issues related to basic IPL performance which deserve further consideration. A more rigorous analysis of IPL behavior with very low photon fluxes, where low-density shot noise is important, would be useful. Closed-loop photon-photon correlations\textsuperscript{110} should also be considered. The linearized analyses of Robinson\textsuperscript{149} and Dyson\textsuperscript{110} and the nonlinear PLL noise discussion in Viterbi\textsuperscript{164} (particularly pages 85-87) provide useful insights into formulating this analysis.

It would also be informative to theoretically study phase compensation performance with $2n\pi$ ambiguities in the modulator phase. The finite size of the point spread function of a monolithic modulator can result in a gradual transition, and hence phase error, at the boundary between regions where the modulator phase differs by $2n\pi$. This can, for example, degrade the compensated Strehl ratio. Another applications oriented
issue is the optimum division of a low-level signal between the adaptive system detector and the compensated application (e.g., communications receiver).

Although the phase-contrast IPL is well understood, a more rigorous analysis of the performance degradations due to: partial spatial and temporal coherence of the input light, signal amplitude fluctuations, and white light operation with a dispersive electrooptical modulator would be of value. These considerations, in conjunction with the aberration statistics of a particular application, should allow a theoretical determination of the practical limitations of the phase-contrast IPL. For example, the minimum stellar magnitude required for initial convergence in astronomical image compensation is of interest. A study of IPLs implemented with other elementary interferometers, such as radial-shearing, may also be valuable.

An experimental investigation of some of the theoretical issues outlined above would also be worthwhile, as would experimental demonstrations of the intricacies and special features of the various IPL interferometric implementations. Some topics of particular interest are "self-centering", white light operation, performance with amplitude fluctuations, and initial phase acquisition in the phase-contrast IPL and frequency locking, $G_0$ offset control, and limit-cycle oscillations in the heterodyne IPL. The all-optical MSLM/IPL system could be employed for these experiments. Discrete-channel systems, however, would allow more comprehensive laboratory investigations, including performance studies with a variety of loop-filter frequency responses.

The above experiments generally require a modulator dynamic range in excess of 2- radians, which suggests the parallel activity of constructing
an improved nineteen or twenty element modulator to be driven by the existing high-voltage amplifier array. The modulator proposed in Section 5.3, involving the dicing of a slab of LiNbO₃ without cutting the front surface, offers the possibility of dynamic ranges in excess of 25π radians.

Large dynamic range discrete-channel IPL systems could also be useful for field investigations in actual applications. For example, a single-channel homodyne or heterodyne system could yield valuable information on the coherence time and dynamic range of phase fluctuations introduced by propagation through atmospheric turbulence or low-visibility conditions. Since obtaining a stable local oscillator could be difficult, it may be more practical to employ a self-interference twenty-channel IPL to collect similar data on spatially-relative aberrations. The existing twenty-channel amplifier is compact and rugged enough to be taken to the field, particularly if an enclosure is added to the high voltage section. Chapters 5 and 6 and Appendix C provide useful experimental guidelines for constructing and operating these discrete-channel systems. A discrete-channel IPL is also a general purpose precise phase measurement device applicable to a broad range of phase measurement applications, such as profiling surfaces or studying thick transparent objects.
A. INTERFERENCE PHASE LOOP DETAILS

A.1. Average value of a class of stochastic phasors

Phasors of the form $ae^{j\theta}$ are widely utilized to analyze systems in disciplines ranging from electronic circuit analysis to optics. In stochastic problems, the ensemble average of these phasors is often of interest. The ensemble average (overbar) can also be represented as a phasor:

$$\overline{ae^{j\theta}} = a \overline{\cos \theta} + j \overline{a \sin \theta} = ce^{j\phi}$$ \hspace{1cm} (A.1)

It is immediately apparent from Eq. (A.1) that

$$c = \sqrt{(a \cos \theta)^2 + (a \sin \theta)^2} \leq \bar{a}$$ \hspace{1cm} (A.2a)

$$\phi = \tan^{-1}\left(\frac{a \sin \theta}{a \cos \theta}\right)$$ \hspace{1cm} (A.2b)

$$\cos \phi = \frac{a \cos \theta}{c}$$ \hspace{1cm} (A.3a)

$$\sin \phi = \frac{a \sin \theta}{c}$$ \hspace{1cm} (A.3b)

The inequality in Eq. (A.2a) follows from the triangle inequality, i.e., with $z_k = a_k e^{j\theta_k}$ as a member of the phasor ensemble:

$$c = \frac{1}{N} \sum_{k=1}^{N} |z_k| \leq \frac{1}{N} \sum_{k=1}^{N} |z_k| = \bar{a}$$ \hspace{1cm} (A.4)

In many instances the real part of the phasor is of greatest interest:

$$\text{Re}(ae^{j\theta}) = a \cos \theta$$

Equations (A.1)-(A.3b) are valid for spatial or temporal averages as well as ensemble averages.
Making assumptions on the statistics of $a$ and $\theta$ results in interesting simplifications of the expressions for $c$ and $\phi$. If $a$ and $\theta$ are statistically independent, the average phasor can be factored into two parts, i.e., $\overline{a e^{-j\theta}}$. With $P(\theta)$ as the probability distribution function of $\theta$, it follows that

$$
e^{-j\theta} = \int_{-\infty}^{\infty} e^{-j\theta} P(\theta) d\theta$$

(A.5)

Introducing the constant $w=1$

$$e^{-j\theta} = \int_{-\infty}^{\infty} e^{-j\theta} P(\theta) d\theta = H(P(\theta)) \bigg|_{w=1} = C_x(P) \bigg|_{w=1}$$

(A.6)

Here $H$ denotes the Fourier transform operation and the last term on the right is the characteristic function of $P(\theta)$ evaluated at $w=1$.

Making a change of variables to a new variable $t$ with zero mean; $t=\theta - \bar{\theta}$ yields,

$$e^{-j\theta} = e^{-j\bar{\theta}} \int_{-\infty}^{\infty} e^{-jwt} P(t+\bar{\theta}) dt$$

(A.7)

The distribution $P$ can be rewritten as a real function with even and odd parts, $P=P_e+P_o$; which result in:

$$e^{-j\theta} = 2e^{-j\bar{\theta}} \left[ \int_{0}^{\infty} P_e(t+\bar{\theta}) \cos(t) dt - j \int_{0}^{\infty} P_o(t+\bar{\theta}) \sin(t) dt \right]$$

(A.8)

When $P$ is an even function, the quantity in brackets is real and hence does not contribute to $\phi$, the phase of $e^{-j\theta}$. 
An important conclusion follows from Eq. (A.8); for any probability function \( P(\theta) \) which is symmetric about the mean value of \( \theta \), the phase of the average phasor \( e^{j\theta} \) is just the average phase \( \bar{\theta} \), i.e., \( \phi = \bar{\theta} \) or:

\[
\frac{e^{-j\theta}}{e^{-j\bar{\theta}}} = c_p e^{-j\bar{\theta}} \tag{A.9}
\]

where

\[
c_p = 2 \int_0^\infty P(t+\bar{\theta}) \cos(t)dt = \mathcal{F}(P) \bigg|_{w=1} = \mathcal{C}_x(P) \bigg|_{w=1} \leq 1 \tag{A.10}
\]

The last inequality arises since \( \int_{-\infty}^\infty P(t)dt = 1 \) and \( |\cos(t)| \leq 1 \).

Thus if \( \Delta \) and \( \theta \) are statistically independent and \( P(\theta) \) is even about \( \bar{\theta} \):

\[
c = \bar{\Delta} \mathcal{F}(P) \bigg|_{w=1} \leq \bar{\Delta} \tag{A.11a}
\]

and

\[
\phi = \bar{\theta} \tag{A.11b}
\]

A specific example of a probability distribution which is even about its mean is the Normal or Gaussian distribution. If \( \Delta \) and \( \theta \) are independent and \( \theta \) is a Gaussian distributed random variable of mean \( \bar{\theta} \) and variance \( \sigma^2 \); it follows that:

\[
\frac{ae^{-j\theta}}{\sqrt{2\pi}\sigma} \mathcal{F}(e^{-t^2/2\sigma^2}) \bigg|_{w=1} \quad e^{-j\bar{\theta}} = \bar{\Delta} e^{-\sigma^2/2} e^{-j\bar{\theta}} \tag{A.12}
\]

This is a well known result.\(^{30,31,112,247}\)

Another easy case to evaluate is for \( \theta \) uniformly distributed over \( 2\pi \):

\[
\bar{\Delta} e^{-j\bar{\theta}} = \bar{\Delta} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-j\theta} d\theta = \bar{\Delta} \frac{j}{\pi} \sin \pi = 0 \tag{A.13}
\]
A.2. **Single lens optical convolution**

The single lens system of Fig. A.2 is adequate for many applications in which the dual lens system of Fig. A.1 is traditionally employed. This is surely common knowledge; but is briefly reviewed here, since it is not often seen in the literature. The elimination of unnecessary lenses from an optical processor is of practical importance; real lenses introduce distortion and high quality lenses are expensive. However, in some instances, the dual lens system can offer the advantage of having the second lens compensate aberrations introduced by the first lens.

Since the major results are well known, they are referenced rather than derived. In Figs. A.1 and A.2: a plane wave is incident from the left; \( g_1(x,y) \) is the complex input transmittance which produces \( G(x',y') \) at \( P_2 \); \( H(x',y') \) is the transmittance of a filter at \( P_2 \); \( g_0(r,t) \) is the output field; \( f_1 \) is the focal length of lens \( L_1 \) and \( f_2 \) is the focal length of lens \( L_2 \).

**Dual lens system**

Using the Rayleigh-Sommerfield integral with the Fresnel approximation, it is not difficult to show \(^{100,108}\) that a lens produces a Fourier transform multiplied by a quadratic phase factor at its rear focal plane. The quadratic phase factor disappears when \( d_2 = f_1 \) in Fig. A.1. The final output of the dual lens system is a convolution multiplied by a complex constant \(^{100,108}\)

\[
g_0(r,t) = -\frac{e^{i2k(f_1+f_2)}}{4\pi^2} \frac{f_1}{f_2} \iint_{P_1} dx dy \ g_1(-x-y) h\left(-\frac{f_1}{f_2}r-x\right), \ -\left(-\frac{f_1}{f_2}t-y\right) \] \quad (A.14a)

\[
= c_2[g_1(-x,-y) * h(-x,-y)] \quad (A.14b)
\]
Fig. A.1 Dual lens optical convolution.

\[ g_1(x,y) \quad \leftrightarrow \quad G(x',y') \quad \leftrightarrow \quad g_0(r,t) \]

\[ d_2 = d_3 = f_1 \quad \quad \quad d_2' = d_3' = f_2 \]

Fig. A.2 Single lens optical convolution

\[ d_3 = f_1 \]

\[ \frac{1}{d_2} = \frac{1}{d_5} = \frac{1}{f_1} \]
Here \( h(x,y) \) is the inverse Fourier transform of the filter \( H(x',y') \). When there is no filter at \( P_2 \), \( H(x',y') = 1 \) and an inverted and scaled image appears at \( P_3 \):

\[
g_0(r,t) = c_2 g_i\left(-\frac{f_1}{f_2}, -\frac{f_1}{f_2} t\right)
\]

(A.15)

**Single lens system**

The output of the single lens system can be found by using the well-known lens propagation result\(^{100,108}\) to go from plane \( P_1 \) to \( P_2 \) in Fig. A.2, and then using the Rayleigh-Sommerfield integral with the Fresnel approximation\(^{100,108}\) to propagate the field through free space from \( P_2 \) to \( P_3 \). When \( d_3 = f_1 \) and \( 1/d_5 + 1/d_2 = 1/f_1 \) in Fig. A.2, it can be shown that the output field is

\[
g_0(r,t) = c'_2 \exp\left(j \frac{k}{2d_4} (r^2 + t^2)\right) \iint_{P_1} dx dy \, g_i(-x,-y) \, h(-\frac{f_1}{d_4} r-x), -\frac{f_1}{d_4} t-y))
\]

(A.16a)

where

\[
c'_2 = e^{jk(d_2+d_5)} \frac{d_2}{4\pi^2 \frac{d_5}{d_2}}
\]

(A.16b)

Thus the only significant difference between the single lens convolution, Eqs. (A.14), and the dual lens result, Eqs. (A.16), is a quadratic phase factor in the former case. In applications where only the intensity is of concern, e.g., when there is a detector at \( P_3 \), the two systems are essentially identical. The only difference is that \( d_4 \) in the single lens case replaces \( f_2 \) of the dual lens result, (since \( d_5/d_2 = d_4/f_1 \)). When there is no filter at \( P_2 \), an inverted and scaled image appears at \( P_3 \); as one expects from the condition \( 1/d_5 + 1/d_2 = 1/f_1 \).
A.3. The Zernike phase-contrast filter

In the frequency plane of a Zernike phase-contrast system (e.g., Fig. A.3) is a filter of the form

$$H(u,v) = 1 + (be^{j\phi_h} - 1)P(u,v)$$  \hspace{1cm} (A.17)

where \(P(u,v)\) is a small aperture in the center of the \(u, v\) plane, and \(b\) and \(\phi_h\) are constants. This appendix attempts to provide insight into the relation between the pupil function, \(P(u,v)\), and Zernike phase-contrast behavior, as an aid to designing and predicting the performance of actual Zernike systems.

In the following discussion the input aperture, \(P_i(x,y)\) in Fig. A.3, is assumed to be square, which often allows the two-dimensional result to be expressed as the product of two 1-D results. With \(\Pi(x'/\Delta x)\) as a pulse of width \(\Delta x\) and unit magnitude, the field in the input aperture of Fig. A.3 is

$$g'(x', y') = g(x', y')P_i(x', y') = \Pi\left(\frac{x'}{\Delta x}\right) \Pi\left(\frac{y'}{\Delta x}\right) g(x', y')$$  \hspace{1cm} (A.18)

Here \(g\) is the input field just before the input aperture. The area of this aperture is \(A_m=\Delta x^2\). The 1-D Fourier transform of the field in the input aperture (assuming \(g(x', y')=g_x(x')g_y(y')\)) is

$$G'_x(f_x) = (\Delta x \text{ sinc}(f_x \Delta x))*G_x(f_x)$$  \hspace{1cm} (A.19)

Here \(G_x = \mathcal{F}(g_x)\), \(\mathcal{F}\) is a Fourier transform operator, and \(*\) is a convolution operator.

The filter aperture \(P(u,v)\) is also assumed to be expressible as the product of two 1-D apertures, i.e., \(P(u,v) = P_x(f_x)P_y(f_y)\), where
\[
\begin{align*}
&g(x',y') \\
&
\begin{cases}
&\downarrow \\
&P_1(x',y') \\
&T
\end{cases}
\\
&\quad \\
&L_1
\\
&H(u,v)
\\
&G(f_x,f_y)
\\
&\quad \\
&o(x,y)
\\
&\quad \\
&P_2 \\
&P_3
\end{align*}
\]

\[
H(u,v) = 1 + (be^{j\phi} - 1)P(u,v)
\]

\[
u = 2\pi f_x, \quad v = 2\pi f_y
\]

Fig. A.3 Zernike phase-contrast system.
\[ u = 2\pi f_x \quad \text{and} \quad v = 2\pi f_y \quad . \]

The 1-D Zernike phase-contrast output field, after passing through the Fourier filter of Eq. (A.17) and retransformation, is

\[ o(x) = \mathcal{F} [G_x'(f_x) + (b e^{-j\phi_h}) P_x(f_x) G_x'(f_x)] = \quad (A.20a) \]

or

\[ o(x) = g_x'(x) + E_r(x) \quad (A.20b) \]

where

\[ g_x'(x) = g_x(x) \quad \Pi \left( \frac{x}{\Delta x} \right) \quad (A.21a) \]

\[ E_r = (b e^{-j\phi_h}) h(x) \quad (A.21b) \]

and

\[ h(x) = \mathcal{F} [P_x(f_x) G_x'(f_x)] \quad (A.21c) \]

Note that \( E_r \) is an effective reference wave which interferes with \( g \) to make its phase visible; if \( E_r = |E_r| e^{j\phi_r} \) and \( g = |g| e^{j\phi_g} \), the Zernike output intensity is

\[ |o|^2 = |E_r|^2 + |g|^2 + 2|E_r||g|\cos(\phi_g - \phi_r) \quad (A.22) \]

In order to obtain a uniform interferogram, \( E_r \) and hence \( h(x) \) should be approximately constant over \(-\Delta x/2 \leq x \leq \Delta x/2\); which is the aperture of \( g(x) \).

"Exactly" diffraction limited case

The case where \( P(u, v) \) is exactly diffraction limited will be discussed first, i.e.,

\[ P_x(f_x) = \text{sinc}(f_x \Delta x) \quad (A.23) \]

Such a filter could be fabricated, for example, by recording the transform of the unobscred aperture \( P_i(x, y) \) on film.
The Fourier transform of $G'_x(f_x)P(u_x)f_x$ in Eq. (A.21c) becomes

$$h(x) = \mathcal{F}\{[(\Delta x \text{sinc}(f_x \Delta x)) \ast G'_x(f_x)] \text{sinc}(f_x \Delta x)\}$$

$$= \left[\pi\left(\frac{x}{\Delta x}\right)g_x(x)\right] \ast \frac{1}{\Delta x} \pi\left(\frac{x}{\Delta x}\right)$$

$$= \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} \pi\left(\frac{x-x'}{\Delta x}\right) g_x(x-x') dx'$$

$$= \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} \pi\left(\frac{x}{\Delta x}\right) g_x(x-x') dx' \quad (x''=x-x')$$

Finally

$$h(x) = \begin{cases} 
\frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} g_x(x') dx' & 0 < x < \Delta x \\
\frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} g_x(x') dx' & -\Delta x < x < 0 \\
0 & x < -\Delta x, x > \Delta x
\end{cases} \quad (A.24)$$

Using brackets to denote the spatial average, note that $h(0) = \langle g_x \rangle$ and if $g_x$ is symmetric about zero $h(\Delta x/2) = h(-\Delta x/2) = \langle g_x \rangle / 2$. In many applications this diffraction-limited $P(u,v)$ is thus unsatisfactory, since the reference wave amplitude varies by about a factor of two over the phase visualization aperture.
Sub diffraction limited case

The case where \( P(u,v) \) is a square aperture, smaller than the
diffraction limited spot, will be considered next, i.e.

\[
P_x(f_x) = \Pi \left( \frac{X}{\Delta f} \right)
\]

The second transform of \( G_x'(f_x)P_x(f_x) \) is

\[
h(x) = \mathcal{F}^{-1} \left\{ [\Delta x \text{sinc}(f_x \Delta x) \ast G_x(f_x)] \Pi \left( \frac{X}{\Delta f} \right) \right\} = \]

\[
[\Pi \left( \frac{X}{\Delta x} \right) g_x(x)] \ast \Delta f \text{sinc}(x \Delta f) = \]

\[
h(x) = \Delta f \int_{-\Delta x / 2}^{\Delta x / 2} g_x(x') \text{sinc}((x-x') \Delta f) dx' \quad (A.27)
\]

Requiring a uniform reference wave over the output aperture imposes the
constraint that \( \text{sinc}(\Delta f \Delta x) = 1 \). Since \( \text{sinc}(1/4) = .9 \), a good rule of
thumb is \( \Delta f \Delta x < 1/4 \). Then Eq. (A.27) becomes approximately uniform for
\( |x| < \Delta x / 2 \)

\[
h(x) = \Delta f \Delta x < g_x > \quad \text{with} \ \Delta f \Delta x < 1/4 \quad (A.28)
\]

Actually this result could have been obtained directly from (A.26) by
making a stationary phase approximation, i.e. \( \Delta f \Delta x < 1/4 \) implies
\( 2\pi f \Delta x < \pi / 4 \) over the limits of the Fourier Transform.
Returning to two-dimensions;

\[ h(x,y) = A_F A_m <g(x,y)> \]  \hspace{1cm} (A.29)

Here \( A_F = \Delta f^2 \) is the area of the frequency domain filter \( P(u,v) \); and \( A_m = \Delta x^2 \) is the area of the input aperture \( P_i(x',y') \).

When the transform is performed by a lens of focal length \( f \), the physical width of the \( P(u,v) \) aperture is \( \Delta x_f = \lambda f / A_f \) (e.g., \( \Delta x_f \approx 6 \mu m \) for \( \lambda = 633 \text{nm}, f = 1 \text{m}, \Delta f = 1/4 \Delta x, \Delta x = 1 \text{-inch} \)). This is small, but probably realizable; an alternative is use additional lenses to magnify the Fourier plane. The physical area of \( P(u,v) \) is

\[ A_p = (\lambda f)^2 A_f \] \hspace{1cm} (A.30)

A photographic means for producing these sub-diffraction limited filters would be to underexpose a photograph of the transform of the unobscured input aperture \( P_i(x',y') \). A reflecting filter with \( \phi_h = \pi \) could possibly be produced by placing an alluminized dot on a dielectric beamsplitter.

Using the single lens optical system of Fig. A.2, the phase-contrast output field follows from Eqs. (A.16), (A.20), (A.21) and (A.29) as

\[ o(x,y) = c_0 [g(-x,-y) + A_m A_f <g(x,y)> (e^{j \phi_h} - 1)] \] \hspace{1cm} (A.31)

Here \( c_0 \) is a complex constant from Eqs.(A.16); \( x = x'd_5/d_2 \), and \( y = y'd_5/d_2 \).

The results of this section reveal that there is a trade-off between the spatial uniformity and amplitude of the reference wave,
$E_r$ of Eq. (A.21b). If a 50% reduction at the edges of the interferogram is tolerable, $|E_r|$ can be as strong as $|E_r|=2|g|$. On the other hand, approximate spatial uniformity reduces $|E_r|$ to $|E_r|<\frac{1}{8}|g|$. These values assume $b<1$.

Actually, effective values of $b>1$ can be obtained by attenuating the region surrounding $P(u,v)$ in the Fourier plane. With an attenuation factor of $\alpha$, Eq. (A.17) becomes

$$H(u,v) = \alpha + (be^{-ih} - \alpha)P(u,v) = \alpha(1+(b' e^{-ih} - 1)P(u,v))$$  \hspace{1cm} (A.32)

Here $b'=b/\alpha$ can be greater than unity. Such a filter could be produced photographically, or by drilling a pinhole in a beamsplitter or attenuating filter. The same basic output result, Eq. (A.31), is still valid if it is multiplied by $\alpha$ and $b$ is replaced by $b/\alpha$. 
A.4. The discrete, linearized, Zernike phase-contrast IPL

With small phase error and \( N \) discrete resolution element, the behavior of the phase-contrast IPL near equilibrium is describable by a system of \( N \) coupled linear equations in the Laplace domain. The equation, Eq. (3.135), for the \( k^{\text{th}} \) modulator element was presented in the main text and is repeated here:

\[
\hat{\phi}_{m_k} = \frac{\sum_{j=1}^{N} G_{1,j} \hat{\phi}_{i_k}^{m_j}}{N<\bar{G}_{1}>} + \hat{\phi}_{1} + \frac{1}{N<\bar{G}_{1}>} \sum_{j=1}^{N} G_{1,j} \hat{\phi}_{m_j}^{m_j} \tag{A.33}
\]

Here the lowest subscript indexes the modulator element, the tilde denotes Laplace domain variables and averages are used interchangeably with summations (e.g., \( \frac{1}{N} \sum_{j=1}^{N} G_{1,j} = <G_{1}> \)). This system of \( N \) equations will be solved for an uncoupled expression for \( \hat{\phi}_{m_k}^{m_j} \) in one modulator element, which is not a function of \( \hat{\phi}_{m_j}^{m_j} \) in other modulator elements. It is convenient to rewrite Eq. (A.33) as

\[
\frac{\sum_{j=1}^{N} G_{1,j} \hat{\phi}_{m_k}^{m_j}}{N<\bar{G}_{1}>} + \frac{1}{N<\bar{G}_{1}>} \sum_{j=1}^{N} G_{1,j} \hat{\phi}_{m_j}^{m_j} = N<\bar{G}_{1}>\hat{\phi}_{1}^{m_k} - \sum_{j=1}^{N} G_{1,j} \hat{\phi}_{i_k}^{m_j} + N<\bar{G}_{1}>(c \hat{\phi}_{1}^{m_k} + \hat{\phi}_{i_k}^{m_j} - \frac{2\pi}{S}n_{1}) \tag{A.34}
\]

Matrix techniques provide a straightforward approach for solving this system of equations. Although the main text utilizes a much simpler technique, the matrix solution is presented to show that the two approaches are equivalent.
The N resolution elemental equations can be written as the following matrix equation:

\[
\begin{align*}
\mathbf{C}_m &= \mathbf{D} \mathbf{\phi}_i + \mathbf{\bar{V}}_b \quad \text{with } \mathbf{\bar{V}}_m = \begin{pmatrix}
\phi_{m1} \\
\phi_{m2} \\
\phi_{m3} \\
\vdots \\
\phi_{mn}
\end{pmatrix} \quad \text{and } \mathbf{\bar{\phi}}_i = \begin{pmatrix}
\phi_{i1} \\
\phi_{i2} \\
\phi_{i3} \\
\vdots \\
\phi_{in}
\end{pmatrix}
\end{align*}
\] (A.35)

Here:

\[
\mathbf{C} = \mathbf{A} \mathbf{G} = \begin{pmatrix}
a_1 & -1 & -1 & \ldots \\
-1 & a_2 & -1 & \ldots \\
-1 & -1 & a_3 & \ldots \\
\vdots & \vdots & \vdots & \ddots \\
\vdots & \vdots & \vdots & \vdots & \ddots \\
\vdots & \vdots & \vdots & \ddots & a_N
\end{pmatrix} \begin{pmatrix}
G_{i1} & 0 & 0 & \ldots \\
0 & G_{i2} & 0 & \ldots \\
0 & 0 & G_{i3} & \ldots \\
\vdots & \vdots & \vdots & \ddots \\
\vdots & \vdots & \vdots & \ddots & G_{iN}
\end{pmatrix}
\] (A.36a)

\[
\text{where } a_i = \frac{N G_{i1}}{G_{i1} f_i} - 1 \quad \text{(A.36b)}
\]

\[
\mathbf{D} = \begin{pmatrix}
d_1 & -1 & -1 & \ldots \\
-1 & d_2 & -1 & \ldots \\
-1 & -1 & d_3 & \ldots \\
\vdots & \vdots & \vdots & \ddots \\
\vdots & \vdots & \vdots & \ddots & \ddots
\end{pmatrix} \mathbf{\bar{G}}
\] (A.37a)

\[
\text{where } d_i = \frac{N G_{i1}}{G_{i1}} - 1 \quad \text{(A.37b)}
\]
\[
\begin{pmatrix}
\tilde{e}_1 - \frac{2\pi}{s} n_1 + c\phi' & \\
\tilde{e}_2 - \frac{2\pi}{s} n_2 + c\phi' & \\
\tilde{e}_3 - \frac{2\pi}{s} n_3 + c\phi' & \\
\tilde{e}_N - \frac{2\pi}{s} n_N + c\phi'
\end{pmatrix}
\]

\[
\phi_b = N < G_1 >
\]

(A.38)

The solution of Eq. (A.35) is

\[
\phi_m = \phi^{-1} \phi_1 + \phi^{-1} \phi_b
\]

(A.39)

The problem is thus reduced to finding the inverse of \( \phi \), which is \( \phi^{-1} \). The inverse of \( \phi \) is trivial \( (G^{-1})_{ii} = 1/G_i \). In his paper on the stability effects of channel cross-coupling in the COAT adaptive optical system, O'Meara gives hints for inverting a matrix with the form of \( \phi \), but with all \( a_i \) equal. These hints will be used to find \( \phi^{-1} \).

First, Det \( \phi \) will be evaluated. Adding all the other rows to row one and subtracting column one from the other columns does not change the value of the determinant of \( \phi \):

\[
\text{Det} \phi = \left| \begin{array}{cccc}
a_{1-N+1} & a_2 - a_1 & a_3 - a_1 & \ldots & a_N - a_1 \\
-1 & a_2 + 1 & 0 & 0 \\
-1 & 0 & a_3 + 1 & 0 \\
& \vdots & & \vdots \\
-1 & 0 & 0 & a_N + 1
\end{array} \right|
\]

(A.40)

Expanding Det \( \phi \) down the first column:

\[
\text{Det} \phi = \frac{B}{a_{1+1}} [a_{1-(N-1)} + \sum_{i=1}^{N} \frac{a_i - a_1}{a_i + 1}]
\]

(A.41)
Here \( B = \prod_{i=1}^{N} (a_i + 1) \) and \( a_1 - a_1 = 0 \) were used.

Employing the identity

\[
N = \sum_{i=1}^{N} \frac{a_i}{a_i + 1} + \sum_{i=1}^{N} \frac{1}{a_i + 1}
\]  

(A.42)

and using Eq. (A.36b) allows Eq. (A.41) to be simplified to:

\[
\text{Det} \bar{A} = B \left[ 1 - \sum_{i=1}^{N} \frac{1}{a_i + 1} \right] = B \left[ 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{G_i} \right] = B \left[ 1 - \langle G_1 \rangle \right]
\]

(A.43)

It will simplify the following algebra to define

\[
P = 1 - \frac{\langle G_1 F \rangle}{\langle G_1 \rangle} = \frac{\langle (1-F)G_1 \rangle}{\langle G_1 \rangle} \quad (\text{Det} \bar{A} = BP)
\]

(A.44)

In order to evaluate \( \bar{A}^{-1} \), it is sufficient to determine the form of two general elements: \( (\bar{A}^{-1})_{ij} \) and \( (\bar{A}^{-1})_{ij} \). The cofactor of \( \bar{A}_{11} \) can be evaluated by excluding \( a_1 \) from the previous result for \( \text{Det} \bar{A} \):

\[
\text{Cof}(\bar{A})_{11} = \begin{vmatrix}
  a_2 & -1 & -1 & \ldots & -1 \\
  -1 & a_3 & -1 & \ldots & -1 \\
  -1 & -1 & a_4 & \ldots & -1 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  -1 & -1 & -1 & \ldots & a_N \\
\end{vmatrix} = B \frac{1}{a_1 + 1} \left( P + \frac{1}{a_1 + 1} \right)
\]

(A.45)
Since the channel number are arbitrary, this generalizes to any i:

\[
(\bar{A}^{-1})_{ii} = \frac{\text{Cof}(\bar{A})_{ii}}{\text{Det}(\bar{A})} = \frac{p + \frac{1}{a_i + 1}}{(a_i + 1)^p} = \frac{G_{1_i}^{\bar{r}_i}}{N^{G_{1_i}}}(1 + \frac{G_{1_i}^{\bar{r}_i}}{N^{(1-F)G_{1_i}}})
\]  
(A.46)

Similarly,

\[
\text{Cof}(\bar{A})_{12} = \begin{vmatrix}
-1 & -1 & -1 & \ldots & -1 \\
-1 & a_3 & -1 & -1 \\
-1 & -1 & a_4 & -1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-1 & -1 & -1 & a_N & -1
\end{vmatrix} = - \begin{vmatrix}
-1 & -1 & -1 & \ldots & -1 \\
0 & a_3 + 1 & 0 & 0 \\
0 & 0 & a_4 + 1 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & a_N + 1
\end{vmatrix} = \frac{B}{(a_1 + 1)(a_2 + 1)}
\]  
(A.47)

and

\[
(\bar{A}^{-1})_{ij} = \frac{1}{(a_i + 1)(a_j + 1)^p} = \frac{G_{1_i} G_{1_j}^{\bar{r}_i \bar{r}_j}}{N^{2G_{1_i}}<(1-F)G_{1_i}>}
\]  
(A.48)

This inverse has been checked by evaluating \( \bar{A}^{-1} \bar{A} = \bar{A} \bar{A}^{-1} = I \).
Using Eqs. (A.46) and (A.48), the in-channel transfer function is:

\[
(\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{B}})_{ii} = (\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{B}})_{ii} = \mathcal{P}_i - \frac{\mathcal{F}_i G_1 (1 - \mathcal{F}_i)}{N <(1-F)G_1>}
\]

(A.49)

and the cross-coupling transfer function is:

\[
(\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{B}})_{ij} = \frac{\mathcal{F}_i G_1 (1 - \mathcal{F}_j)}{N <(1-F)G_1>}
\]

(A.50)

The resulting complete solution for the \(k\)th modulator element follows from Eq. (A.39) as

\[
\tilde{\varphi}_{in} = \mathcal{F}_k \left( \phi_{in}^k_1 - \frac{< (1-F) G_1 \phi_1 >}{< (1-F) G_1 >} \right) + \tilde{\varepsilon}_k + \frac{< G_1 \mathcal{F}_1 > + < G_1 c' \phi_1 >}{< (1-F) G_1 >} - \frac{2\pi}{\lambda n_k}
\]

(A.51)

Aside from being in discrete spatial channel form, this result is identical to the decoupled Equation (3.137) derived by other means in the main text.

As shown in the main text, this formal approach is not required for the Zernike phase-contrast IPL. The techniques introduced here, however, can be readily applied to systems with more complicated spatial coupling, e.g. due to the spatial overlap of the influence functions from each element of a discrete element modulator.
A.5. Stability of the linearized Zernike phase-contrast IPL

Equation 3.137 of the main text describes the response of the phase-contrast IPL after a perturbation from an equilibrium where \( \phi_e \) is small. There is a stable return to equilibrium if the Laplace response has no poles in the right half of the complex frequency plane.

In order to evaluate the stability of Eq. 3.137 it is necessary to define the open-loop and closed-loop transfer functions more precisely. It is also helpful to break the system into \( N \) discrete resolution elements. Eq. 3.137 then becomes identical to Eq. A.51 of Appendix A.4. Assuming that the open-loop transfer function can be expressed as a ratio of polynomials:

\[
G_k \cdot H_k \equiv \frac{M_k}{D_k}
\]  

(A.52a)

where

\[
M_k \equiv \prod_{i=1}^{Nz} (s + z_{ik})
\]

and

\[
D_k \equiv \prod_{i=1}^{Nd} (s + d_{ik})
\]

(A.52b)

(A.52c)

The corresponding closed-loop transfer function is

\[
P_k = \frac{\hat{H}_k G_k}{1 + \hat{H}_k G_k} = \frac{M_k}{P_k}
\]

(A.53a)

where

\[
P_k \equiv D_k + M_k \equiv \prod_{i=1}^{Np} (s + b_{ik})
\]

(A.53b)

Although the values of the poles and zeros are allowed to vary from element to element, for simplicity it is assumed that the number of poles and zeros is not spatially variant.
The average in Eq. (3.137) can now be rewritten in terms of summations over $M_k$, $D_k$, and $P_k$, for example:

$$\langle (1-\Phi) G_1 \rangle = \frac{1}{N} \sum_{i=1}^{N} G_1 \frac{D_i}{P_i} = \frac{1}{NB} \sum_{j=1}^{N} [G_1 D_j \prod_{\ell \neq j} P_\ell]$$

where $B = \prod_{i=1}^{N} P_i$ (i.e., the product of the $N$ values of $P_i$).

Eq. (3.137) now becomes:

$$\phi_k = \frac{M_k}{P_k} \left[ \tilde{\phi}_k + \sum_{j=1}^{N} \left( \sum_{\ell \neq j} (M_j \tilde{c}_j - D_j \Phi_c) \right) + NB <G_1 \phi_1^c> \right]$$

Examination of the definitions of $M_k$, $D_k$, and $P_k$ reveals that the characteristic equation of the first two terms is $P_k$. Thus the "in-channel" response is stable if the closed-loop transfer function in that channel is stable.

The characteristic equation for the remaining "cross-coupling" terms is

$$\sum_{j=1}^{N} G_1 D_j \prod_{\ell \neq j} P_\ell = NB \langle (1-\Phi) G_1 \rangle = 0$$

If $\hat{H}_k$ and $\hat{F}_k$ (at the current values of $G_1$) are stable in every resolution cell, it will be shown that the "cross-coupling" terms are also stable. With stable $\hat{H}_k$ and $\hat{F}_k$, it is assumed for all $i$ and $k$ that $\text{Re}(b_{ik}) > 0$ and $\text{Re}(d_{ik}) > 0$ holds in Eqs. (A.52) and (A.53). Then, from Eqs. (A.52c), (A.53b), and (A.54), each term in the summation of Eq. (A.52) is a polynomial in $s$ of degree $n_t + n_d + n_p - 1$, with all coefficients positive.
The summation of Eq. (A.52) can be rewritten as a single polynomial of degree $N_t + N_d + N_p - 1$:

$$
\sum_{j=1}^{N} G_j \prod_{k \neq j} P_k = s \sum_{l=0}^{N_t} \sum_{i=0}^{N_d + N_p - 2} A_l s_i = s \sum_{i=1}^{N_t} (s + r_i) \quad (A.57)
$$

This equation defines $A_i$ and the poles of the "cross-coupling" terms, $r_i$. Since all $A_i$ are greater than zero, it follows from Descartes Rule of Signs that $\text{Re}(r_i) > 0$ holds for all $r_i$. The system thus has only stable poles, at zero or in the left half of the $s$ plane.

It can be concluded that the spatial coupling of the phase-contrast IPL does not seriously degrade the perturbational IPL stability. If the closed-loop and open-loop responses at each spatial location are stable when each loop acts independently, then the phase-contrast IPL is also stable.
B. MICROCHANNEL SPATIAL LIGHT MODULATOR DETAILS

B.1 MSLM Circuit Modelling Details

The Laplace frequency responses and temporal behavior corresponding to the simple MSLM circuit model of Figure 4.2 and more complete model of Figure 4.1 are presented here. These results provide a framework for interpreting experimentally measured operational parameters, guide device improvement research, and lead to accurate simplified expressions for the time and frequency domain behavior of the modulator.

B.1.1 Two-Pole Model

Laplace Solution

The equations of the two-pole circuit model of Figure 4.2 were solved in the Laplace domain to yield the Laplace frequency response:

\[ \tilde{V}_E(s) = \frac{1}{\frac{1}{C_{E}} + \frac{1}{R_C}} \tilde{V}_b(s) + \frac{1}{C_{E}} \tilde{V}(s) + \frac{1}{R_C} \frac{1}{s} \tilde{V}_b(s) + \frac{1}{s} \tilde{V}_E(s) - \frac{1}{s} V_0 - \frac{1}{s} V_0 \]

\[ \tilde{V}_E(s) = \frac{1}{\frac{1}{R_C} + \frac{1}{s}} \tilde{V}_b(s) - \frac{1}{C_{E}} \frac{1}{s} \tilde{V}(s) - \frac{1}{R_C} V_0 + \frac{1}{R_C} V_0 \]

(B.1a)

Here \[ \| \] is an operation, i.e. \[ \|X||Y||Z = (X^{-1}Y^{-1}Z^{-1})^{-1} \], \[ V_0 \] and \[ V_0 \] are the initial conditions at \[ t=0 \]; and the solutions of the system characteristic equation are:

\[ \lambda_1, \lambda_2 = \frac{1}{2 C_g(R_S | R_L)} C_g(R_S | R_L) + \frac{1}{2} \sqrt{\frac{C_E(R_S | R_L)}{C_g(R_S | R_L)}} + \frac{2}{C_g(R_S | R_L)} \]

(B.2a)

Note that the roots are always real. It is also easy to prove that they are positive by using:

\[ \frac{1}{2 C_g(R_S | R_L)} C_g(R_S | R_L) + \frac{4}{C_g(R_S | R_L)} \leq \frac{C_E(R_S | R_L)}{C_g(R_S | R_L)} \]

\[ \frac{C_E C_g(R_S | R_L)}{C_g(R_S | R_L)^2} + \frac{2}{C_g(R_S | R_L)} \leq \frac{C_E C_g(R_S | R_L)}{C_g(R_S | R_L)^2} \]
which follows from the identity $R_S^2 \geq (R_S || R_E)(R_S || R_L)$.

In most cases the MSLM is characterized by $R_S < R_E || R_C$, allowing the approximations:

$$\lambda_1 \approx \frac{1}{R_S(C_g || C_E)} \quad (B.2b) \quad \lambda_2 \approx \frac{1}{(R_E || R_C)(C_g + C_E)} \quad (B.2c)$$

**Time Response**

The temporal response can be found by inverse transforming the Laplace response. With zero initial conditions and steps in $V_b$ and $I_e$,

($\tilde{V}_b = V_b/s$ and $\tilde{I}_e = I_e/s$); the step response is:

$$V_E(t) = V_{ES}\left[\frac{\lambda_1(1-e^{-\lambda_2 t}) - \lambda_2(1-e^{-\lambda_1 t})}{\lambda_1 - \lambda_2}\right] + \frac{V_b}{C_E R_S}\left[\frac{e^{-\lambda_2 t} - e^{-\lambda_1 t}}{\lambda_1 - \lambda_2}\right] \quad (B.3a)$$

$$V_g(t) = V_{GS}\left[\frac{\lambda_1(1-e^{-\lambda_2 t}) - \lambda_2(1-e^{-\lambda_1 t})}{\lambda_1 - \lambda_2}\right] + \frac{V_b}{C_g R_S - I}\left[\frac{e^{-\lambda_2 t} - e^{-\lambda_1 t}}{\lambda_1 - \lambda_2}\right] \quad (B.3b)$$

$$I_x(t) = \frac{V_b}{R_S} + \frac{IR_L}{R_S + R_E + R_L} V_b \left[\frac{\lambda_1(1-e^{-\lambda_2 t}) - \lambda_2(1-e^{-\lambda_1 t})}{\lambda_1 - \lambda_2}\right]$$

$$+ \frac{1}{R_s}\left[\frac{I - V_b}{C_g + \frac{C_E}{R_S}}\right]\left[\frac{e^{-\lambda_2 t} - e^{-\lambda_1 t}}{\lambda_1 - \lambda_2}\right] \quad (B.3c)$$

where $V_{ES}$, $V_{GS}$ and $I_{XS}$ are the steady-state values reached

$$V_{ES} = \frac{R_E(V_b + R_L I)}{R_E + R_L + R_S} \quad (B.4a) \quad V_{GS} = \frac{R_L(V_b - (R_S + R_E)I)}{R_S + R_E + R_L} \quad (B.4b)$$

$$I_{XS} = \frac{V_b + IR_L}{R_E + R_L + R_S} \quad (B.4c)$$

**B.1.2 Four-Pole Model**

**Laplace Solution**

The circuit equation of the four-pole circuit model of Figure 4.1 are considerably more difficult to solve in the Laplace domain. The
The transfer function for the crystal voltage with zero initial condition is:

\[
\hat{V}_x(s) = \frac{1}{R_S C_x C_g} \left( s + \frac{1}{R_Q R_Q} \right) \left( s + \frac{1}{R_T C_T} \right) \hat{V}_x C_g \left( s + \frac{1}{R_L C_g} \right) \left( s + \lambda_1 \right) \left( s + \lambda_2 \right) \left( s + \lambda_3 \right) \left( s + \lambda_4 \right)
\]

(B.5a)

The characteristic equation is:

\[
(s + \lambda_1)(s + \lambda_2)(s + \lambda_3)(s + \lambda_4) = \frac{(R_p C_x C_c Q X_T + D)(1 + R_C G) + R_C G D}{R_S R_C C_c C_T C_G} \left( s + \frac{1}{R_X X_T} \right) \left( s + \frac{1}{R_Q Q_T} \right) = 0
\]

where

\[
D = C_x C_T X_T + C_x C_Q X_Q + C_Q C_T Q_T
\]

and \( X = s + \frac{1}{R_X X_T} \); \( Q = s + \frac{1}{R_Q Q_T} \); \( T = s + \frac{1}{R_T C_T} \); \( G = s + \frac{1}{R_L C_g} \)

Time Response

There is no advantage to deriving the time response from the Laplace solution since the algebra is very cumbersome and numerical methods are required to solve the fourth-order characteristic equation. It is much easier to employ matrix techniques; the circuit equations then become:

\[
\hat{V} = \bar{A} \hat{V} + B
\]

where

\[
\bar{A} = \begin{pmatrix}
\frac{1}{C} \left( \frac{1}{R_L R_S} \right) & \frac{1}{R_C C_g} & \frac{1}{R_S C} & \frac{1}{R_S C} \\
\frac{1}{R_S C_X} & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_X} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_T} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \right) \\
\frac{1}{R_S C_T} & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_T} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_T} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \right) \\
\frac{1}{R_S C_Q} & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_Q} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_Q} \right) & \frac{1}{C} \left( \frac{1}{R_S R_p} + \frac{1}{R_Q} \right)
\end{pmatrix}
\]

(B.6a)
\[ \bar{V} = \begin{pmatrix} V_g \\ V_X \\ V_T \\ V_Q \end{pmatrix} \quad (B.6b) \] and
\[ \bar{b} = \begin{pmatrix} \frac{V_b}{R_S C_g} - \frac{I}{C_g} \\ \frac{V_b}{R_S C_X} \\ \frac{V_b}{R_S C_T} \\ \frac{V_b}{R_S C_Q} \end{pmatrix} \quad (B.6c) \]

\[ (\dot{\bar{V}} = \frac{d\bar{V}}{dt}) \]

The homogeneous solution (\( \bar{b} = \bar{0} \)) can be found by substituting \( \bar{V} = \bar{c} e^{\lambda t} \) into \( \dot{\bar{V}} = \bar{A}\bar{V} \) to yield the eigenequation \( (\bar{A} - \lambda \bar{I})\bar{c} = 0 \) and the characteristic equation \( \text{Det} |\bar{A} - \bar{I}| = 0 \).

Superimposing solutions, a general solution is:
\[ \bar{V} = \bar{L}_1 \bar{c}_1 e^{\lambda_1 t} + \bar{L}_2 \bar{c}_2 e^{\lambda_2 t} + \ldots = \bar{C}\bar{E}\bar{L} \]

where: \( \bar{c} = (\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n) = \) Modal matrix of eigenvectors
\[ \bar{E} = \text{diag}(e^{\lambda_1 t}) \]
\[ \bar{L} = \text{vector of coefficients determined by initial conditions}. \]

The method of undetermined coefficients can be used for the particular solution. For the step response, \( \bar{V} = \bar{V}_p \neq f(\text{time}) \) is substituted into equation (B.6a); resulting in \( \bar{V}_p = -\bar{A}^{-1}\bar{b} \).

The full solution is \( \bar{V} = \bar{C}\bar{E}\bar{L} - \bar{A}^{-1}\bar{b} \), where \( \bar{L} \) can be found by using the initial conditions vector \( \bar{V}_0 = \bar{V}(t=0) \). Since \( \bar{E}(t=0) = \bar{I} \):
\[ \bar{L} = \bar{E}^{-1}(\bar{V}_0 + \bar{A}^{-1}\bar{b}) \]

The full step response solution is
\[ \bar{V} = \bar{C}\bar{E}\bar{L}\bar{V}_0 + (\bar{C}\bar{E}\bar{L} - \bar{I})\bar{A}^{-1}\bar{b} \quad (B.7) \]
Note that according to Frobenius's theorem of matrix theory \( F(\bar{A}) = \bar{X} \text{diag}F(\bar{\lambda})\bar{\lambda}^{-1} \), \( \bar{\lambda} \in \bar{E}^{-1} \) is identical to \( e^{At} \) which is known as the state transition matrix.

The response to a step in \( V_b \) and/or I was computed numerically from equation (B.7) on an IBM 370 for a wide range of initial conditions and parameter values. Accurate EISPACK algorithms were utilized to find the eigenvalues and eigenvectors; their accuracy was verified by substituting them in the identities:

\[
\bar{\lambda}^{-1}\bar{AC} = \text{diag}\lambda \equiv \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots \\ 0 & \lambda_2 & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix} \quad \text{and} \quad \bar{\lambda} = \bar{AC} - \text{diag}\lambda = 0 \equiv \begin{pmatrix} 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix}
\]

Based on numerical experiments with typical ranges of parameters (\( C_g, C_X \ll C_{Q}, C_{T}, R_p \rightarrow \infty \)) approximate (~1% accuracy) expressions were obtained for the four eigenvalues. The MSCM is then characterized by the four time constants:

\[
\begin{align*}
\tau_1 &= |\lambda_1|^{-1} = (R_L || R_S)(C_T || C_Q || C_G || C_X) \\
\tau_2 &= |\lambda_2|^{-1} = (R_S + R_L) || (R_X + R_Q + R_T)[C_g + C_X || C_Q || C_T] \\
\tau_3 &= |\lambda_3|^{-1} = R_Q C_Q = \rho_Q e_Q = 400 \text{ sec (for SiO}_2) \\
\tau_4 &= |\lambda_4|^{-1} = R_T C_T = \rho_T e_T = 10^{-3} \text{ sec (for TiO}_2)
\end{align*}
\]

B.1.3 Proof that four-pole model is always overdamped

For any choice of parameter values the four-pole system is overdamped; characterized by real negative eigenvalues. Reviewing basic matrix theory, two matrices \( \bar{A} \) and \( \bar{Q} \) related by a similarity transform \( Q = T \bar{A} T^{-1} \), where \( T \) is an arbitrary nonsingular matrix) have
identical eigenvalues. It should also be remembered that a real symmetric matrix has real eigenvalues.

Continuing, it is convenient to rewrite $\bar{\bar{A}}$ in equation (B.6) as the product of a diagonal matrix $\bar{\bar{D}}$ and symmetric matrix $\bar{\bar{S}}$:

$$\bar{\bar{A}} = \bar{\bar{D}} \bar{\bar{S}} = \begin{pmatrix} \frac{-1}{C_g} & 0 & 0 & 0 \\ 0 & \frac{-1}{C_x} & 0 & 0 \\ 0 & 0 & \frac{-1}{C_q} & 0 \\ 0 & 0 & 0 & \frac{-1}{C_t} \end{pmatrix} \begin{pmatrix} \frac{1}{R_L} + Y_s & Y_s & Y_s & Y_s \\ Y_s & \frac{1}{R_x} + Y_1 & Y_1 & Y_1 \\ Y_s & Y_1 & \frac{1}{R_q} + Y_1 & Y_1 \\ Y_s & Y_1 & Y_1 & \frac{1}{R_T} + Y_1 \end{pmatrix}$$  (B.9)

Here: $Y_1 = \frac{1}{R_S} + \frac{1}{R_p}$ and $Y_s = \frac{1}{R_S}$

$\bar{\bar{A}}$ can be similarity transformed to a real symmetric matrix $\bar{\bar{Q}}$ by employing $T = (\bar{\bar{D}}^{1/2})^{-1}$; i.e.,

$$\bar{\bar{Q}} = (\bar{\bar{D}}^{1/2})^{-1} \bar{\bar{A}} (\bar{\bar{D}}^{1/2}) = \bar{\bar{D}}^{1/2} \bar{\bar{S}} \bar{\bar{D}}^{1/2}$$  (B.10)

Since $\bar{\bar{Q}}$ and $\bar{\bar{A}}$ are related by a similarity transform and $\bar{\bar{Q}}$ is symmetric with real eigenvalues, $\bar{\bar{A}}$ must have real eigenvalues.

The characteristic equation $\text{DET}(\bar{\bar{A}} - \lambda I) = 0$ was explicitly written in terms of the system parameters in equation (B.5b) of the Laplace solution. Since all physical resistances and capacitances have real positive values, equation (B.5b) is a polynomial with all positive coefficients.
According to Descartes's rule of signs, a polynomial with no variations of sign in the coefficients can have no positive real roots. Since all the eigenvalues are real, they must be negative; corresponding to a decaying time response.

B.2 Secondary Emission in the MSLM

Secondary emission assumes a role in the behavior of any device in which moderate energy electrons ($qV > 20\text{eV}$) bombard a target. Unlike the majority of such applications, in which monoenergetic electrons are produced by an electron gun; the MCP in the MSLM generates electrons with a broad distribution of energies. All operating modes of the MSLM, and in particular the inverted-secondary mode and special functions of appendix B.4, are influenced by secondary emission; as expressed through load-line analysis on the $J$ versus $V_g$ characteristic of Figures 4.3 and 4.8. Secondary emission is preferable to photoconductivity or other processes for electron removal; it does not degrade image storage characteristics and can often be utilized without additional complexity in MSLM fabrication or operation. A thorough
understanding of secondary emission behavior in the MSLM is essential to analyzing and improving the MSLM's operating characteristics.

There are three major components to secondary emission behavior: the energy distribution of MCP electrons, the secondary emission properties of the target (e.g. crystal or dielectric mirror), and the dynamics of electrons in the MCP-to-target field. The basic terms and concepts involved will be introduced through a discussion of the simpler monoenergetic electron case.

B.2.1 Monoenergetic electrons

In most target materials the ratio of secondary electrons to primary electrons, \( \delta' = \frac{i_s}{i_p} \), varies with the energy \( qV_e \) of the primary electrons as shown in Fig.B.1.\textsuperscript{208,240,241} Between the first and second crossover energies, \( qV_I \) and \( qV_{II} \), each primary electron generates more than one secondary; the secondary emission coefficient \( \delta' \) peaks at \( \delta_p \), which occurs at an intermediate energy \( qV_p \). The details of this relationship depend on target material and surface condition and the electron angle of incidence; typical values for an insulating target are:\textsuperscript{208,240} \( 20 < V_I < 100 \) volts, \( 1.5 < \delta_p < 20 \), \( 200 < V_p < 600 \) volts and \( 2kV < V_{II} < 20kV \).

A monoenergetic approximation to the MSLM is sketched in Figure B.2. Only the most fundamental monoenergetic secondary emission dynamics will be discussed here; a more thorough discussion for the same basic geometry appears in Kazan and Knoll.\textsuperscript{240}

The gap current \( i_g \) as a function of gap voltage \( V_g \), which is illustrated in Figure B.3, is the behavior of greatest interest. The electron energy at the target, \( qV_e \), is the sum of the gap acceleration energy, \( qV_g \); and initial electron ejection energy, \( qV_a \). When
Fig. B.1 Secondary emission coefficient, $\delta'$, as a function of primary electron energy, $qV_e$.

Fig. B.2 MSLM Secondary emission geometry. The electron energy is $qV_e = q(V_a + V_g)$. The crystal voltage is $V_x = V_b - V_g$. The gap current is $i_g = i_p - i_s$. The MCP is modelled as an emissive cathode and an acceleration grid.
Fig. B.3 Gap current, $i_g$, as a function of gap voltage, $V_g$, with monoenergetic electrons. ($J_{Pr} = neGA_I/hv =$total irradiance-induced MCP ejection current. $V_g=V_e-V_a$. A stable equilibrium occurs at $V_g=V_0$ when $R_x \to \infty$).

Fig. B.4 Special cases of $i_g$ versus $V_g$. (□=stable equilibria for finite $R_x$, ○= stable equilibria for $R_x \to \infty$)
$V_e = V_g + V_a < V_I$ in Figure B.1, there is a net flow of primary electrons to the target. With an insulating target ($R_x \to \infty$) these electrons accumulate and decrease $V_g$ until $V_g = -V_a$, at which point all additional electrons are repelled. This is a condition termed "lock-out" in which charge can be neither added nor removed. The decrease of $i_g$ as $V_g$ approaches $-V_a$ is gradual rather than abrupt due to deflection of primary electrons, by any negative gap field, to the back of the MCP or other grounded surfaces. This effect depends on the details of MSLM geometry.

With $V_e = V_a + V_g > V_I$, according to Figure B.1, there is a net removal of secondaries from the target. The trajectory in Figure B.3 should follow the dotted line ($i_g = i_p - i_s = i_p(1-\delta')$); however the positive gap field when $V_g > 0$ attracts the secondaries back to the target crystal. The effective secondary current $i_{se} = i_p \delta_e$ is reduced to zero at $V_I$, where no secondaries possess enough energy to escape from the target. Equilibrium ($i_g = 0$, $\delta_e = 1$) occurs at $V_0$, where all but one secondary returns to the target for each incident primary electron. If another electrode at a potential $V_2 > 0$ was placed near the target, this electrode would collect secondaries until an equilibrium occurred at $V_g = V_2 + V_0$.

When the target is insulating, $i_g$ vs $V_g$ in Figure B.3 is effectively a state-space plot ($V_g \sim i_g / C_x$ vs $V_g$). Collection of primaries ($i_g > 0$) decreases $V_g$, secondary removal ($i_g < 0$) increases $V_g$, and equilibria ($i_g = 0$) occur at $V_g = -V_a$ and $V_g = +V_0$. Depending on the ejection energy $qV_a$, which is a fixed parameter in this discussion, there are three major types of behavior. In Figure B.3 $V_I - V_a < V_0$ held; Figure B.4 illustrates $i_g$ vs $V_g$ for $V_I - V_a > V_0$ and $V_II - V_a < V_0$. Notice that Figure B.3 and case C in Figure B.4 are bistable.
As mentioned in section 4.2.1 of the main text and illustrated in Figure B.4, with finite target conductivity, equilibrium points lie on the system load-line, \( V_g \approx \frac{V_b}{R_X i_g} \), rather than at \( i_g = 0 \).

The monoenergetic \( i_g \) vs \( V_g \) relations in cases a, b, c of Figure B.3 and B.4, cannot adequately explain the observed \( i_g \) vs \( V_g \) behavior, e.g. Figure 4.7b. Case c is the most similar curve; however, \( V_a > V_{II} = V_0 \) is not reasonable since \( V_{II} > 1500 \) volts and \( V_a \) cannot exceed the MCP bias \( V_m \) (generally < 1500 volts). Case a in Figure B.3 can match the observed behavior with \( V_0 < 0 \); but \( V_0 < 0 \) is not reasonable because \( V_g = V_0 < 0 \) causes all the secondaries to escape from the target \( (i_g < 0, \text{ not } i_g = 0) \).

**B.2.2 Multienergetic electron distribution**

The energy \( qV_a \) of an MCP electron is a function of from where, on the MCP pore wall, it was emitted before ejection. Since most MCP's have slanted or curved pores to prevent ion feedback, virtually all incident electrons impact a pore wall and loose memory of their MCP entrance energy. The MCP exit electron current, \( i_p(V_a) \), is thus expected to be distributed in energy, \( qV_a \), from zero to a maximum of \( qV_m \), the MCP bias voltage, as illustrated in Figure B.5. The shape of this distribution was estimated by fitting theoretically predicted \( i_g \) vs \( V_g \) curves, resulting from a variety of assumed \( i_p(V_a) \) distributions, to actual experimental measurements of \( i_g \) vs \( V_g \). Theoretical \( i_g \) vs \( V_g \) relations, resulting from a couple of assumed simple \( i_p(V_a) \) energy distributions, will be developed in this section.
Fig. B.5 Idealized MCP electron energy distribution.

\[ i'_p = \frac{3i'_{po}}{3V_a} \] is the fraction of MCP electrons with energy \( qV_a \).

Fig. B.6 Energy distribution of secondary electrons. \( \delta'' = \frac{3\delta'}{3V_s} - \frac{3^2\delta''}{3V_s 3V_e} \)

\( qV_e \) is the ejection energy of the secondary electrons. Curves are shown for two values of primary electron energy, \( qV_e \).
Negative Gap Voltage

With negative gap voltage, \( V_g < 0 \), no secondaries are attracted back to the target, but MCP primaries with \( V_a < |V_g| \) are repelled back to the MCP and do not reach the target. The resulting gap current \( I_g \) can be computed from the \( \delta' \) versus \( V_e \) distribution of Figure B.1 \((\delta' = \frac{3\delta}{2V_e})\), and the \( I_p' \) versus \( V_a \) distribution of Figure B.5 \((I_p' = \frac{3I_p}{2V_e})\).

Using \( V_g' = |V_g| = -V_g \) and \( V_e' = V_a + V_g \); the effective primary current is:

\[
I_p = \int_{V_g'}^{V_m} I_p'(V_a) \, dV_a = \int_{0}^{V_m-V_g'} I_p'(V_e + V_g') \, dV_e \quad \text{(B.11)}
\]

The gap current is

\[
I_g(V_g') = I_p - I_s = \int_{V_g'}^{V_m} I_p'(V_a)[1-\delta'(V_a-V_g')] \, dV_a \quad \text{(B.12a)}
\]

\[
= \int_{0}^{V_m-V_g'} I_p'(V_e+V_g)[1-\delta'(V_e)] \, dV_e \quad \text{(B.12b)}
\]

Note that \( I_p \) in Eq.(B.11) is only a portion of the total MCP ejection current, \( I_{po} \)

\[
I_{po} = \int_{0}^{V_m} I_p'(V_a) \, dV_a \quad \text{(B.13)}
\]

The second term in equations(B.12) is effectively a convolution between \( I_p' \) and \( \delta' \). Useful insights can be gained by graphically translating these functions relative to each other and observing the prescribed overlap integrals. This formulation neglects such higher order effects as: secondary emission at the back of the MCP and other grounded surfaces, MSLM geometry dependent deflection of MCP electrons away from the crystal, and additional extraction of electrons from the MCP when \( V_g > 0 \). Experiments suggest that these effects have a negligible influence.
Positive Gap Voltage

With $V_g > 0$, all MCP electrons are assumed to contribute to the primary current at the target; but many secondaries are attracted back to the target. The secondaries are emitted with an energy distribution $i_s' = \frac{d^i_s}{dV_s}$, which is depicted in Figure B.6 for two values of incident energy $qV_e$. The secondary energy cannot exceed the primary energy $qV_e$, except by the small thermal energy of target electrons. Secondaries with energy $qV_s$ less than $qV_g$ do not escape from the target to contribute to the secondary current. This results in a reduction of the secondary emission coefficient from $\delta'(V_e)$ of Figure B.1 to the effective value $\delta'(V_e, V_g)$ which is defined through:

$$i_s = \int_{V_g}^{V_e} i_s'(V_e, V_s) \, dV_s \equiv \int_{V_g}^{V_e} i_s'(V_e, V_s) \, dV_s$$

Here $\delta''(V_e, V_s) = \frac{\partial \delta'}{\partial V_s} = \frac{\partial^2 \delta}{\partial V_s \partial V_e}$

($\delta''$ versus $V_s$ appears in Fig. B.6, and $\delta'$ versus $V_e$ appears in Fig. B.1)

From Eq. (B.14a)

$$\delta' = \frac{\partial \delta}{\partial V_e} = \int_{V_g}^{V_e} \delta''(V_e, V_s) \, dV_s$$

In terms of $\delta'$ the total gap current can be expressed as:

$$i_g(V_g) = \int_0^{V_g} i_e'(V_a)[1-\delta'(V_a+V_g, V_g)] \, dV_a \quad \text{for } V_g > 0$$

$$= \int_{V_g}^{V_g+V_m} i_e'(V_e-V_g)[1-\delta'(V_e, V_g)] \, dV_e \quad \text{for } V_g > 0$$

In order to evaluate the integrals in equation (B.11) through (B.15), specific expressions are required for the relations represented in
Figures B.1, B.5, and B.6. Kazan and Knoll suggest that the secondary emission coefficient $\delta'(V_e)$ of Figure B.1 can be approximated in many materials by an expression of the form:

$$\delta'(V_e) = a V_e \exp(-b\sqrt{V_e})$$  \hspace{1cm} (B.16)

The coefficients $a$ and $b$ can be expressed in terms of the peak emission coefficient $\delta_p$, which occurs at energy $qV_p$. Evaluating $\frac{3\delta'}{\delta V_e}$ to find the maximum reveals:

$$a = \frac{\delta_p e^2}{V_p} \quad \text{and} \quad b = \frac{2}{\sqrt{V_p}}$$

Eq. (B.16) becomes

$$\delta' = \delta_p e^2 V_e \cdot 2\sqrt{V}$$

where

$$V = V_e/V_p$$  \hspace{1cm} (B.17)

The associated crossover voltages $V_I$ and $V_{II}$ where $\delta' = 1$ can be found by graphical techniques; such as the intersections of $Y_1 = 2\sqrt{V}$ and $Y_2 = 2 + \ln(\delta_p) + \ln(V)$ plotted as a function of $V$. For example with $\delta_p = 3$ and $V_p = 500$ volts, $V_I = 40$ volts and $V_{II} = 3000$ volts follows.

Kazan and Knoll also suggest that the secondary energy distribution $\delta''(V_s)$ of Figure B.6 can be approximated by an infinite ($V_s \rightarrow \infty$) Maxwellian distribution:

$$\delta''(V_e, V_s) = a e^{-bV_s} = \frac{\delta'}{V_s}$$ \hspace{1cm} (B.18)

Here the distribution was normalized to $\delta'(V_e)$ of Figure B.1, i.e.,

$$\delta' = \int_0^\infty a e^{-bV_s} = a/b; \quad \text{and} \quad V_s$$

is the average secondary energy:

$$\bar{V_s} = \int_0^\infty V_s \cdot \frac{\delta''}{\delta'} \cdot dV_s = 1/b$$
Using Eq. (B.18) in the integral of Eq. (B.14),
\[
\delta' = \frac{\delta'(V_e)}{V_g} \int_0^\infty e^{-V_s/V_g} dV_s = \frac{\delta'(V_e) e^{-V_g/V_s}}{V_g} \tag{B.19}
\]

\[i_p'(V_a)\]

Finding a single simple expression for the MCP energy distribution \(i_p'(V_a)\) of Figure B.5 is more difficult. However, retardation \(V_g < 0\) experiments with an aquadag (graphite) electrode with \(\delta_p < 1\) to suppress secondary emission, show a linear fall-off at low energies \((-20eV < V_g < 0)\), and an exponential fall off at greater retardation, see Figure B.7a.

This suggests that a composite pulse and exponential distribution may fit \(i_p'(V_a)\) well:
\[
i_p'(V_a) = \frac{i_{PL}(U(V_a) - U(V_a - V_L))}{V_L} + \frac{i_{PH}}{V_a} e^{-V_a/V_a} = i_{PL} + i_{PH} \tag{B.20}
\]

Here \(U(x)\) is the unit-step function, i.e. \(U(x < 0) = 0, U(x > 0) = 1\). Note that \(q\tilde{V}_a\) is the average energy of the exponential component and \(qV_L\) is the maximum energy of the low-energy pulse. This expression is normalized to the total MCP electron current \(i_{po} = \int_{V_a}^{V_m} i_p'(V_a) dV_a\):
\[
i_{po} = \int_0^V i_p'(V_a) dV_a = \int_0^V (i_{PL} + i_{PH}) dV_a = i_{PL} + i_{PH} \tag{B.21}
\]

(This normalization and taking \(q\tilde{V}_a\) as the average energy of the exponential assumes that \(\tilde{V}_a << V_m;\) typically \(\tilde{V}_a < 100\) volts and \(V_m > 1000\) volts.)

The primary current with a retarding field \(V_g = -V_g' < 0\), can be determined by evaluating Eq. (B.11) using \(i_p'\) from Eq. (B.20),
\[ i_p = \int_{V_g}^{\infty} i'_p(V_a) \, dV_a \]

For \(-V_L < V_g < 0\):
\[ i_p = i_p L (1 - \frac{V_g}{V_L}) + i_p H e^{-V_g/V_a} \]  
(B.22a)

For \(V_g < -V_L\):
\[ i_p = i_p H e^{-V_g/V_a} \]  
(B.22b)

The average energy of the MCP energy distribution of equation (B.20) is:
\[ V_{avg} = \frac{1}{i_{po}} \int_0^{V_a} V_a i'_p(V_a) \, dV_a = \frac{i_p H V_a + i_p L V_L/2}{i_p H + i_p L} \]  
(B.23)

The distribution of \(i'_p(V_a)\) in equation (B.20) agrees fairly well with distributions recently reported in the literature.224

\[ i_g \text{ vs } V_g \]

The desired \(i_g\) as a function of \(V_g\) relation can be obtained by employing the energy distribution of equations (B.17), (B.19) and (B.20) in the integrals of equations (B.12) and (B.15). For \(V_g = -V'_g < -V_L\), using equations (B.12a), (B.17), and (B.20):

\[ i_{gH} = \frac{i_p H}{V_a} \int_{V'_g}^{\infty} V_a e^{-V_a/V_a} \left[ 1 - \delta_p \left( \frac{V_a - V'_g}{V_p} \right) \exp(2 - 2 \sqrt{\frac{V_a - V'_g}{V_p}}) \right] \, dV_a \]

Using equation (B.22b) and the change of variables \(u = \sqrt{\frac{V_a - V'_g}{V_p}}\), results in

\[ i_{gH} = i_{po} e^{-V'_g/V_a} (1 - \delta_H) \]  
(B.24a)

where

\[ \delta_H = \frac{2e^{2\delta_p} \int_{0}^{u_m} u^3 \exp(-2u) \, du}{V_a \int_{0}^{u_m} u^3 \exp(-2u) \, du} \]  
(B.24b)  
(for \(V_g < -V_L\))
and \[ u_m = \sqrt{(V_m - V'_g)/V_p} \].

Here \( \delta_H \) is an effective secondary emission coefficient.

The \( \delta_H(V_g) \) integral was evaluated by recognizing the relationship:

\[
\int_{u_1}^{\infty} u^3 \exp(-au^2) du = -\frac{3}{a^3} \left[ \int_{u_1}^{\infty} \exp(-au^2) du \right] = -\frac{3}{a^3} F
\]

Using the change of variables \( z = \sqrt{b}(u+a/b) \), \( F \) evaluates to

\[ F = \sqrt{\frac{\pi}{2b}} \, e^{-a^2/2b} \, \operatorname{erfc}(u \sqrt{\frac{b}{2}} + \frac{a}{\sqrt{2b}}) \]

The final fully evaluated result is given in Eq. (B.30); two significantly different methods were employed to assure its correctness.

For the case of \(-V_L < -V'_g < 0\), Eqs. (B.12a), (B.17), (B.20), and (B.24b) yield:

\[ i_g = i_{gH} + i_{PL} \int_{V_g}^{V_L} \left[ 1 - \delta_p \left( \frac{V_a - V'_g}{V_p} \right) e^{2 \sqrt{(V_a - V'_g)/V_p}} \right] dV_a \]

Evaluating the first term of the integral and making the change of variables \( u = \sqrt{(V_a - V'_g)/V_p} \) results in:

\[ i_g = i_{gH} + i_{PL} (1 - \frac{V'_L}{V_L})(1 - \delta_L) \]  \hspace{1cm} (B.25a)

where

\[ \delta_L = \frac{2 \delta_p e^2}{u_L^2} \int_0^{u_L} u^3 e^{-2u} du \]  \hspace{1cm} (for \(-V_L < V'_g < 0\))  \hspace{1cm} (B.25b)

and \[ u_L = \sqrt{(V_L - V'_g)/V_p} \]

Eq. (B.25b) is easily integrated by parts; the resulting expression appears in Eq. (B.29).

For the case of \( V_g > 0 \), Eqs. (B.14a), (B.17), (B.19), and (B.20) result in:
\[ i_g = i_p H(1 - \delta_{He}) + i_p L(1 - \delta_{Le}) \]  \hspace{1cm} (B.26a)

where
\[ \delta_{He} = \int^v_{v_a} \exp \left( \frac{v_g}{v_a} - \frac{v_q}{v_s} \right) u^3 \exp \left( -2u - \frac{v_p u^2}{v_a} \right) \, du \]  \hspace{1cm} (B.26b)

and
\[ \delta_{Le} = \int^v_{v_L} \exp \left( -v_g/v_s \right) u^{-2} \, du \]  \hspace{1cm} (B.26c)

Here \( u = \sqrt{(v_a + v_g)/v_p} \) and
\[ u_0 = \sqrt{v_a/v_p}, \quad u_m = \sqrt{v_m/v_p}, \quad \text{and} \quad u_L = \sqrt{v_L/v_p} \]

The fully evaluated secondary emission behavior as expressed through \( i_g(v_g) \) is summarized in Eqs. (B.28) through (B.32).
Parameterization

The electron dynamics of the theoretical MSLM secondary emission model are fully characterized by the following parameters:

\( i_{po} \) is the total irradiance induced MCP electron current in a resolution cell of area \( A_r \); note

\[
i_{po} = J_p A_r = \frac{n q G A_r I}{h \nu r}
\]  
(B.27)

Also,

\[
i_{po} = i_{pL} + i_{pH}
\]

\( i_{pL} \) is the low-energy component of the MCP electron current.

\( i_{pH} \) is the current in the high-energy tail of the MCP electron current (exponential energy distribution).

\( q \bar{V}_a \) is the average MCP exit energy of the high-energy electrons in the exponential component of the MCP energy distribution.

\( qV_L \) is the maximum MCP exit energy of the low-energy component of the MCP energy distribution.

\( qV_m \) is the maximum energy of the high-energy MCP electrons.

\( \delta_p \) is the peak of the secondary emission coefficient, \( \delta'(V_e) \).

Typical values are: 1.8 to 2.9 for \( \text{LiNbO}_3 \), 2 to 2.6 for \( \text{Al} \), \( <1 \) for graphite, and 7.5 for \( \text{KD} \).

\( qV_p \) is the electron energy \( qV_e = qV_p \) which causes peak secondary emission, \( i.e. \delta_p = \delta'(V_p) \). Typical values are: 200 to 500 volts for \( \text{LiNbO}_3 \), 500 volts for \( \text{BaO} \), 300 to 400 volts for \( \text{Al} \), and 220 volts for \( \text{KD} \).

\( q \bar{V}_s \) is the average energy of secondary electrons.

\( V_g \) is the voltage across the MCP-to-target gap.
The behavior of the model does not depend on all the absolute parameter values, but rather on the dimensionless relative parameters

\[ r = \frac{V_p}{V_a} \quad \text{and} \quad u = \sqrt{\frac{V_e}{V_p}} = \sqrt{\frac{V_a + V_g}{V_p}}. \]

Important values of \( u \) are:

\[ u_0 = \sqrt{\frac{V_g}{V_p}} \quad \text{for} \quad V_a = 0 \]
\[ u_m = \sqrt{\frac{(V_m + V_g)}{V_p}} \quad \text{for} \quad V_a = V_m \]

and

\[ u_L = \sqrt{\frac{(V_L + V_g)}{V_p}} \quad \text{for} \quad V_a = V_L \]

other important relative parameters are \( V_g/V_L \) and \( V_g/V_a \).

The model also makes use of the Normal Distribution Function, \( \phi(z) \);

where

\[ \phi(z) = \int_{-\infty}^{z} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} = \frac{1}{2}(1+erf(z/\sqrt{2})) \]

Tabulations of \( \phi(z) \) or the error function, \( erf(z) \), are readily available.
Basic Result: Theoretical Expression for MSLM Secondary Emission Behavior

The general solution for \( i_g \) as a function of \( V_g \) is:

for \( V_g = -V'_g < 0 \)

\[
\begin{align*}
  i_g &= i_{p_L} \left( 1 - \frac{V_g}{V_L} \right)(1-\delta_L) + i_{p_H} e^{-V_g/V_L} a(1-\delta_H) \\
  \text{ (B.28a)}
\end{align*}
\]

for \( V_g > 0 \)

\[
\begin{align*}
  i_g &= i_{p_L}(1-\delta_{L_e}) + i_{p_H}(1-\delta_{H_e}) \\
  \text{ (B.28b)}
\end{align*}
\]

The effective partial secondary emission coefficients are:

\[
\delta_L = \left\{ \begin{array}{ll}
  \frac{\delta_p e^2}{4u_L} \left[ 3 - e^{-2u_L(4u_o^3+6u_o^2+6u_o+3)} \right] & \text{(-}\!V_L < V_g < 0) \\
  1 & \text{(to null (1-}\delta_L)) \\
  \text{(V_g < -V_L)}
\end{array} \right.
\]

\[
\delta_H = \frac{\delta_p e^2}{r} \left[ \alpha(1-e^{-P_m}) - \gamma_m e^{-P_m} - \beta(\phi_m - \phi_r) \right] 
\]

\[
\delta_{L_e} = \frac{\delta_p e^2 V_p}{4V_L} e^{-V_g/V_L} s[e^{-2u_o(4u_o^3+6u_o^2+6u_o+3)} - e^{-2u_L(4u_L^3+6u_L^2+6u_L+3)}] 
\]

\[
\delta_{H_e} = \frac{\delta_p e^2}{r} \exp \left( \frac{V_g}{V_a} - \frac{V_g}{V_s} \right) \left[ \alpha e^{-P_o - P_m} + \gamma_o e^{-P_o} - \gamma_m e^{-P_m} - \beta(\phi_m - \phi_r) \right] 
\]

Here \( r = \frac{V_p}{V_a} \) \quad \beta = \sqrt{\frac{2}{r}} (r + 3) e^{1/r} \quad \alpha = 1 + \frac{1}{r}

\[
\begin{align*}
  u_o &= \sqrt{\frac{V_g}{V_p}} \\
  u_L &= \sqrt{\frac{V_L+V_g}{V_p}} \\
  u_m &= \sqrt{\frac{V_m+V_g}{V_p}} \\
  \gamma_x &= u_x (ru_x - 1) \quad \text{where } x = o, m, \text{ or } L \\
  p_x &= \gamma_x + 3u_x = ru_x^2 + 2u_x \\
  \phi_r &= \phi \left( \sqrt{\frac{2}{r}} \right) \\
  \phi_m &= \phi \left( \sqrt{2r} u_m + \sqrt{\frac{2}{r}} \right)
\end{align*}
\]
Theoretical secondary emission behavior

The theoretical \( i_g \) versus \( V_g \) behavior of equations (B.28) is graphed in figures B.7 a and b. The parameters were chosen to fit actual experimental data which are shown as points on these graphs. Note the excellent agreement between theory and experiment. The graph of \( \delta_H \) as a function of \( r = \frac{V_p}{V_a} \) in fig.B.8 (from Eq.(B.30)) illustrates the dependence of secondary emission on the average energy of the high-energy component of MCP electrons; typically \( r > 3 \) in the MSLM.

An overall effective secondary emission coefficient can be defined by

\[
\delta_{\text{eff}} = 1 - \frac{i_g}{i_{\text{PL}} + i_{\text{PH}}} \tag{B.33}
\]

Consideration of typical parameter values reveals that the expressions for \( \delta_L, \delta_H, \delta_{\text{Le}}, \) and \( \delta_{\text{He}} \) in equations (B.29) through (B.32) can be simplified. Typically: \( V_p > 300 \ \text{volts}, \ V_a < 100 \ \text{volts}, \ r > 3, \) and \( V_m > 600 \ \text{volts}. \) When \( V_g > 0; \ u_m = \sqrt{V_m / V_p} > \sqrt{2}; \) and hence \( \phi_m = \phi(\sqrt{2}u_m + \sqrt{2}/r) \) \( > \phi(4) \approx 1 \) and \( e^{-p_m} = \exp(-ru_m^2 - u_m) < e^{-8}; \) (Note 0.977 < \( \phi(z > 2) < 1 \)) These same approximations are valid for \( V_g < 0 \) as long as \( -V_g = V_g' < V_m - u_m V_p, \) (where \( u_m' = (\sqrt{1 + 3r} - 1)/r \) insures \( e^{-p_m} < e^{-3} \) and \( \phi_m > \phi(2) \)). Equations (B.30) and (B.32) are then accurately approximated by:

\[
\delta_L = \frac{\delta_p e^2}{r} [\alpha - \beta(1 - \phi_r)] \quad \text{(for } V_m - u_m V_p < V_g < 0) \tag{B.34}
\]

\[
\delta_H = \frac{\delta_p e^2}{r} \exp\left(-\frac{V_g}{V_a}\right) \left[\frac{V_g}{V_s}\right] \left[(\alpha + \gamma_o) e^{-P_o - \beta(1 - \phi_r)}\right] \quad \text{(for } V_g > 0) \tag{B.35}
\]

where \( P_o = \frac{V_g}{V_a} + 2\sqrt{V_g / V_p} \); and \( \alpha, \gamma_o, \beta, \) and \( \phi_r \) are independent of \( V_g \).
Comparison of theoretical and experimental $i_g$ versus $V_g$ secondary emission characteristic. The solid line is a theoretical curve for: $V_m=1000\ \text{v}$, $V_a=47\ \text{v}$, $V_L=9\ \text{v}$, $V_p=300\ \text{v}$, $V_s=5.5\ \text{v}$, $\delta_p=1.3$, $i_{pH}=1.46\times10^{-8}\ \text{A}$, and $i_{PL}=0.94\times10^{-8}\ \text{A}$. The dots are actual experimental data points taken with an aquadag (graphite) coated stainless steel electrode and Galileo MCP.

Fig. B.7a Comparison of theoretical and experimental $i_g$ versus $V_g$ secondary emission characteristic.
The solid line is a theoretical curve for: \( V_m = 600 \text{v}, \ V_a = 90 \text{v}, \ V_L = 4.8 \text{v}, \ V_p = 300 \text{v}, \ V_S = 3 \text{v}, \ \delta_p = 2.9, \ i_pH = 1.46 \times 10^{-8} \text{A}, \) and \( i_pL = 2.04 \times 10^{-8} \text{A}. \) The dots are experimental data points taken by A. Weiss with an Al electrode and Varian MCP.

**Fig. B.7b** Comparison of theoretical and experimental \( i_g \) versus \( V_g \) secondary emission characteristic.
Fig. B.8 Dependence of the high-energy secondary emission coefficient, $\delta_H$, on the average energy $q\bar{v}_a$ of high energy MCP electrons. ($\delta_H$ versus $r = v_p/\bar{v}_a$, with $\delta_p = 1$ and $V_m >> \bar{v}_a$)
Note that $\delta_H$ is independent of $V_g$. This implies that if the MCP energy distribution is indeed exponential, experimental retardation measurements ($V_g = -V_g' < 0$) should reveal a large range of voltages,

$V_L < V_g^' < V_m V_p$ (typically $15 < V_g^' < 800$ volts for $V_m = 1000$ volts),

for which (from equation (B.28a)):

$$i_g = \text{const} \cdot e^{-V_g/V_a} \quad (B.36)$$

Hence $V_a$ can be experimentally determined independently of the secondary emission properties of the target! Actually experimental log $i_g$ vs $V_g$ plots show a slight change of slope (decrease) between -150 and -250 volts, which suggests that the sum of two exponential distributions would be a more exact model.

For $V_g < 0$, $u_L < \sqrt{V_L/V_p} < 0.2$ (typically $V_L < 15$ volts, $V_p > 300$ volts) and $e^{-2u_L}$ can be approximated by its Taylor series. Equation (B.29) then becomes:

$$\delta_L = \frac{\delta_D e^2}{2} u_L^2 (1 - \frac{4}{5} u_L + \frac{4}{3} u_L^2 - \cdots) = \delta_L(0) (1 - \frac{V_g^'}{V_L}) \quad (B.37a)$$

where

$$\delta_L(0) = \delta_L(V = 0) = \frac{\delta_D e^2 V_L}{2 V_p} \quad (B.37b)$$

Because the average secondary energy is small, i.e. $\bar{V}_S < 5$ volts, the $e^{-V_g/\bar{V}_S}$ factor dominates $\delta_{Le}$ and $\delta_{He}$ for $V_g > 0$. Since $e^{-V_p/\bar{V}_S} < e^{-60}$, the terms in square brackets in equation (B.32) really only matter for $V_g << V_p$ and Equation (B.31) can be approximated by

$$\delta_{Le} = \frac{\delta_D e^2}{2 V_L} e^{-V_g/\bar{V}_S} [2(u_L^4 - u_0^4) + 16(u_0^5 - u_L^5)] \quad (for \; 0 < V_g < V_p) \quad (B.38)$$

With small negative gap voltages ($-V_L < V_g < 0$), using equation (B.37b) and $e^{-V_g/\bar{V}_a} \approx 1 - \frac{V_g}{\bar{V}_a}$, the $V_g^'$ dependence of $i_g$ is approximately:
\[ i_g = i_{PL}(1-\frac{V_g}{V_L})^2 + i_{PH}(1-\delta_H)(1-\frac{V_g}{V_L}) \] 

(B.39) 

(for \(-V_L < V_g < 0\))

This quadratic equation in \(V_g\) can be solved for the equilibrium, \(i_g = 0\), gap voltage of an insulating target. Since \(\delta_L(0) \ll 1\), a less accurate expression for \(i_g\) is

\[ i_g = i_{PL}(1-\frac{V_g}{V_L}) + i_{PH}(1-\delta_H(0)) \] 

(B.40) 

(for \(-V_L < V_g < 0\))

With small positive gap voltages, in equations (B.28b), (B.35), and (B.38), \(i_g\) is approximately

\[ i_g = (i_{PL}+i_{PH})e^{-V_g/V_s} \] 

(B.41) 

(for \(0 < V_g < 3\bar{V}_s\))

Estimation of MSLM secondary emission parameters.

No pretense is made that the distributions \(\delta'(V_e), i'(V_a)\), and \(\delta''\)
\((V_e, V_g)\) which went into the theoretical model are exact; however, the good agreement between theory and experiment in figures B.7 suggests that the model is an adequate approximation. Some of the fitting parameters may thus be good estimates of actual physical MSLM characteristics. The simplified expressions of equations (B.34) to (B.41) are useful for finding these parameters; suggesting the generalized \(i_g\) vs \(V_g\) characteristic of figure B.9. The parameters \(V_L, \bar{V}_a\), and \(\bar{V}_s\) can be found from the slope of the appropriate region of experimental \(i_g\) versus \(V_g\) or log \(i_g\) versus \(V_g\) characteristics. Extrapolation of the decaying exponential (equation (B.36)) to \(V_g = 0\), yields \(i_{PH}(1-\delta_H) = i_H(0)\); and \(i_{PL}(1-\delta_L(0)) = i_L(0)\). The measured value of \(i_g\) at \(V_g = 0\) is \(i_g(0) = i_H(0) + i_L(0)\).

Using the inferred values of \(V_L, V_a, \bar{V}_g, i_H(0), \) and \(i_L(0)\); a variety of techniques can be employed to determine the target parameters \(V_p\) and
Fig. B.9 Estimation of MSLM secondary emission parameters from its $i_g$ versus $V_g$ characteristic.
\( \delta_p \). With \( \delta_p = 1 \), and \( \delta_L(V_p) \) and \( \delta_H(V_p) \) scaled by \( k \), equation (B.28b) becomes

\[
i_g = i_{pL} \left( 1 - \frac{V_i}{V_L} \right) (1 - k \delta_L) + i_{pHe} - \frac{V_i}{V_a} (1 - k \delta_H)
\]

(B.42)

where

\[
i_{pH} = \frac{i_H(0)}{1 - k \delta_H}
\]

(B.43a)

and

\[
i_{pL} = \frac{i_L(0)}{1 - k \delta_L(0)}
\]

(B.43b)

Using Eqa. (B.42) and (B.43):

\[
i_g(\infty) \equiv i_g(V_g >> 0) = i_{pL} + i_{pH}
\]

(B.44)

Eqs. (B.43) and (B.44) can be solved for \( \delta_p \) as a function of \( V_p \) and \( i_g(\infty) \),

\[
\delta_p = k = \frac{-b \pm \sqrt{b^2 - 4d}}{2}
\]

(B.45)

where

\[
b = \frac{i_H(0) \delta_L + i_L(0) \delta_L(0) - i_g(\infty)(\delta_H + \delta_L(0))}{i_g(\infty) \delta_H \delta_L(0)}
\]

and

\[
d = \frac{i_g(\infty) - i_g(0)}{i_g(\infty) \delta_H \delta_L(0)}
\]

This equation overestimates \( \delta_p \) because it only considers recollection of secondaries as the cause of the increase of \( i_g \) from \( i_g(0) \) to \( i_g(\infty) \); in actuality field aided extraction of MCP electrons and collection of electrons which missed the target at \( V_g = 0 \) are also responsible. Deviations between the calculated and experimental \( i_g(V_g > 0) \) curves and measurements with the same MCP configuration, but different targets, can be utilized to correct for these effects.

Using equation (B.45) for \( \delta_p \), the remaining parameter, \( V_p \), can be chosen for the best fit to the experimental \( i_g \) vs \( V_g \) curves. Alternatively, another expression for \( \delta_p = f(V_p) \) can be derived at such points as \( (V_g, 0) \)
or \((V_{gp}, i_p)\) in figure B.8, resulting in two equations in the two variables, \(\delta_p\) and \(V_p\). On multiple experiments with the same MCP \(V_L, \bar{V}_a, i_p,\) and \(i_p\) are known constants, and \(\delta_p\) can be determined from \(V_p\) through the relations below, which follow from equations (B.42) and (B.43).

\[
\delta_p = k = \frac{1}{\delta_H} \left( \frac{i_H(0)}{i_{PH}} \right) = \frac{1}{\delta_L(0)} \left( \frac{i_L(0)}{i_{PL}} \right) = \frac{i_g(\infty) - i_g(0)}{i_p L(0) + i_p H} \tag{B.46}
\]

As will be seen, \(\delta_p\) and \(V_p\) are more easily determined when an acceleration grid is inserted between the MCP and crystal.

The process of using this theoretical model to interpret MSLM experimental measurements is continuing. Of particular interest are: the variation of the MCP electron distribution with \(V_m\) and with the ejected current magnitude (e.g. with saturated current levels), and the dependence of secondary emission properties on crystal surface history and coatings.
Secondary emission with a grid between the MCP and crystal

Secondary emission performance of the MSLM can be greatly improved by the inclusion of a grid at potential \( V_c \), as schematically illustrated in figure B.10. With \( V_g \) in the range of \( 0 \leq V_g \leq V_c \), all the primary electrons arrive at the target and the reverse field, \( V_t < 0 \), causes all secondaries to be collected at the grid. This behavior is modeled by the secondary emission integral for \( V_g > 0 \), equation (B.26a), with removal of the \( e^{-V_g/\bar{V}_s} \) factors. The final expression for \( 0 \leq V_g \leq V_c \) is equation (B.28b), with \( \delta_L e \) and \( \delta_H e \) given by equations (B.31) and (B.32) in the limit \( \bar{V}_s \to \infty \). For \( V_g > V_c \), multiply equations (B.31) and (B.32) by \( e^{V_c/\bar{V}_s} \), and for \( V_g < 0 \) equations (B.28a), (B.29) and (B.30) do not need to be modified.

A typical calculated \( i_g \) vs \( V_g \) characteristic for the MSLM with a grid is presented in Fig.B.11. The value of \( V_g \) during electron removal is determined by the \( \bar{V}_b < 0 \) loadline, \( i_g = \c X \bar{V}_b + \frac{V_b - V_g}{R_x} \). The continuous MSLM mode operates by primary electron deposition when \( V_b > V_c \), and by secondary emission (\( i_g < 0 \)) when \( V_b < V_c \). These two cases possess reverse contrast relative to each other. The operational details can be determined by load-line analysis on figure B.11.

A grid greatly increases the secondary current by accelerating the majority of MCP electrons to voltages beyond the first cross-over voltage, \( V_1 \) in Fig. B.1. As illustrated in Fig. B.11, secondary currents in excess of the primary current may even be possible. The increased electron energy should also enhance the proximity focus between the MCP and crystal. Without a grid, proximity focussing fails when an electron deposition framed mode is written long enough for the equilibrium at \( V_g = V_1 = 0 \) in Fig. 4.8 to be reached. With a grid, equilibrium occurs at \( V_g = V_c \) (for \( R_x \to \infty \) and \( \bar{V}_b = 0 \)) in Fig. B.11; although \( V_t = 0 \), \( V_g \) still provides significant proximity focussing acceleration.
Fig. B.10  MSLM Secondary emission geometry with a grid.

Fig. B.11  Theoretical $i_g$ versus $V_g$ characteristic with a grid.

$V_g$ is the grid voltage. The parameter values are the same as those used for Fig. B.7b.
B.2.4 The Inverted-Secondary Framed Mode Integral

As derived in equation (4.15) of the main text, the crystal voltage reached during the inverted secondary framed model depends on the write beam irradiance through

$$V_x = V_{x0} - \int_{V_{g1}'}^{V_m} \left( \frac{C_{vb}}{I_\lambda F(V_g')} \right)^{-1} dV_g'$$  \hspace{1cm} (B.47)

Here $V_{g1}' = -V_{g1}$ and $V_g' = -V_g$. $F(V_g)$ is the irradiance, $I$, to current density, $J$, proportionality coefficient, defined by:

$$J = I \cdot F(V_g)$$  \hspace{1cm} (B.48)

From Eqs. (B.27) and (B.28a)

$$F(V_g) = \frac{neG}{h\nu} \left[ \frac{i_{pL}}{i_{po}} (1 - \frac{V_g'}{V_L}) (1 - \delta_L) + \frac{i_{PH}}{i_{po}} e^{-V_g'/\bar{V}_a (1 - \delta_H)} \right]$$  \hspace{1cm} (B.49)

Referring back to figure 4.8, $V_{g1}$ (where the large $R_x$ load-line and $i_g(V_g)$ intersect) is near $i_g = 0$. Analysis of equation (B.28a) reveals that $i_g = 0$ occurs very near $V_g = -V_L$. Since most of the range of the integral in (B.47) is at $V_g < -V_L$, the $i_{pL}$ current does not have enough energy to appreciably contribute to secondary emission and $F(V_g)$ can be approximated by:

$$F(V_g) = \frac{neG}{h\nu} f_H (1 - \delta_H(0)) e^{-V_g'/\bar{V}_a} \equiv F_0 e^{-V_g'/\bar{V}_a}$$  \hspace{1cm} (B.50)

Here $f_H = i_{PH}/i_{po}$; and $\delta_H(0) \equiv \delta_H(V_g = 0)$ is used since $\delta_H$ is constant over most of the range $-V_m < V_g < 0$. 


With the lumped constant \( b = \frac{C_X V_b}{F_0 A r} \), Eq. (B.47) becomes

\[
V_x = V_{xo} - \int_{V_m}^{V_m + V'_{gp}} \frac{dV'}{V' - l - \frac{b}{l} e^{V'/V_a}}
\]

Since neglect of the \( i_{pl} \) term, which is mostly a primary flow, causes an overestimation of the secondary current, it would be more accurate to integrate from \( V'_{gp} = -V_{gp} \), where \( V_{gp} \) (\( = -V_L \)) is defined as the voltage of maximal secondary current.

This integral is easily evaluated with the aid of common integral tables, resulting in.

\[
V_x = V_{xo} - (V_{m} - V'_{gp}) + \bar{V}_a \ln\left( \frac{1 - \frac{b}{l} e^{V_{m}/V_a}}{1 - \frac{b}{l} e^{V'_{gp}/V_a}} \right)
\]

\[
= V_{xo} - \bar{V}_a \ln\left( \frac{1 - \frac{b}{l} e^{-V_{m}/V_a}}{1 - \frac{b}{l} e^{-V'_{gp}/V_a}} \right) \quad (B.51)
\]

This expression can be simplified by introducing the new lumped constant:

\[
a = \frac{1}{b} e^{-V_{gp}/V_a} = \frac{F(V'_{gp})A_b}{C_X V_b} \quad (B.52)
\]

Here \( aI < 1 \) always holds when (B.47) is applicable, and, from Eq. (B.50), \( F(V'_{gp}) = F_0 \exp(-V'_{gp}/\bar{V}_a) \). During secondary emission charge removal \( V'_{b} < 0 \), \( F(V_{gp}) < 0 \), and hence \( a > 0 \). Employing \( a \):

\[
V_x = V_{xo} - \bar{V}_a \ln\left( \frac{1 - aI \exp((V'_{gp} - V_{m})/\bar{V}_a)}{1 - aI} \right) \quad (B.53)
\]

Since \( V_{m} \gg V_{a} \approx V'_{a} \) \( (V_{m} = 1kV, \bar{V}_a = 80 \text{ volts}, V'_{a} = 10 \text{ volts}) \); the desired expression for \( V_x(I) \) can be written:

\[
V_x = V_{xo} + \bar{V}_a \ln(1 - aI) \quad (B.54)
\]
B.3 Parallel crystal Fabry-Perot interferometer

When the electro-optic crystal in the MSLM is polished with parallel faces, it becomes a Fabry-Perot interferometer. Although this offers a stable intensity representation of modulator phase, insensitive to vibration and air turbulence; the amplitude, phase and intensity of the readout beam are not linear functions of the modulation phase.

The behavior of a Fabry-Perot interferometer is well known; some of the results applicable to the MSLM will be reviewed here. The parameters which characterize a Fabry-Perot are defined in figure B.12. There \( r \) and \( t \) are field reflection and transmission coefficients; \( R \) is the intensity reflection coefficient at interface 1, \( T \) is the intensity transmission coefficient into the crystal at interface 1; \( \alpha \) corresponds to roundtrip attenuation and/or unequal interface reflectivities; and \( \delta = \phi_m + \phi_0 \) is the phase retardation of a roundtrip path inside the crystal, where \( \phi_m \) is the electrically controlled component of phase shift.

The ratio of the total reflected field \( E_m \) to the incident field \( E_i \) can be expressed as:

\[
\frac{E_m}{E_i} = r_1 \left( \frac{1 - \alpha e^{j\delta}}{1 - \alpha R e^{j\delta}} \right)
\]

The intensity ratio is

\[
\left| \frac{E_m}{E_i} \right|^2 = \frac{R[(1-\alpha)^2 + 4\alpha \sin^2(\delta/2)]}{(1 - R)^2 + 4\alpha R \sin^2(\delta/2)}
\]

The overall phase-shift of the Fabry-Perot is:

\[
\phi_f \equiv \tan^{-1} \left( \frac{\text{Im}(E_m/E_i)}{\text{Re}(E_m/E_i)} \right) = \tan^{-1} \left( \frac{\alpha(R-1)\sin \delta}{1 + \alpha^2 R - \alpha(R+1) \cos \delta} \right)
\]
Notes:

\( r_2 = -\alpha r_1 \)

\( \delta = 2\left(\frac{2\pi}{\lambda} n_x\right)L = \phi_0 + \phi_m \)

\( R = r_1^2 \)

\( T = t t' = 1 - R \)

Fig. B.12 Geometry and definition of terms for parallel-crystal Fabry-Perot
When a crystal with negligible absorption and no dielectric mirror is employed, $|r_2|=|r_1|$, $\alpha=0$, and

$$\left| \frac{E_m}{E_i} \right|^2 = \frac{4R\sin^2(\delta/2)}{(1-R)^2+4R\sin^2(\delta/2)} \quad (B.57)$$

$$\tan \phi_f = \frac{R^{-1}}{R+1} \tan \left( \frac{\pi - \phi_m - \phi_0}{2} \right) \quad (B.58)$$

A useful implicit expression for $\phi_f$ is

$$\phi_f = \frac{1}{2}(\phi_0 + \phi_m - \pi) + \sin^{-1}(R\sin(\frac{\pi}{2} - \frac{\delta}{2} - \phi_f))$$

Unlike the case of unequal reflectivities, when $|r_1|=|r_2|$ the intensity goes to zero at $\delta = \phi_m + \phi_0 = 2n\pi$. The phase at an intensity minima $(\delta = 2n\pi)$ is $\phi_f = (2n+1)\frac{\pi}{2}$ for any $R < 1$. With low finesse (small $R$), $\phi_f \approx \frac{1}{2}(\phi_0 + \phi_m - \pi)$; and two cycles of $\phi_m$ are required to change $\phi_f$ by $2\pi$. As one would expect, when $R = 1$ no light can enter the Fabry-Perot and $\phi_f = 0$.

Figures B.13 and B.14 illustrate the dependence of the intensity and phase of the reflected light from the Fabry-Perot on the electrooptic phase, $\phi_m$. The small $R$ curves correspond to LiNbO$_3$ with $n_o=2.29$ (0.633nm) and $R=0.15$. In the symmetric case, $r_1^2 = r_2^2$, $\phi_f$ is fit very well by the line segments:

$$\phi_f = (n-1)\frac{\pi}{2} + b(\phi_m + \phi_0 - n\pi) \quad (B.60a)$$

where

$$n = \text{INT}\left(\frac{\phi_m + \phi_0 + \pi/2}{\pi}\right) \quad (B.60b)$$

and

$$b = \begin{cases} 0.6 & \text{for even } n \\ 0.4 & \text{for odd } n \end{cases} \quad (B.60c)$$

Notice that $\phi_m$ must vary over $4\pi$ radians to cycle $\phi_f$ through $2\pi$ radians. The small $R$ curves are also applicable to LiTaO$_3$ $(R=0.14)$. In the
Fig. B.13 Output intensity modulation, $I_{\text{out}}$, of a parallel-crystal Fabry-Perot. ($I_{\text{out}} = |E_m/E_i|^2$, with $|E_i|^2/2Z_0 = 1$, $R_1 = r_1^2$, and $R_2 = r_2^2$)

Fig. B.14 Output phase modulation, $\phi_f$, of a parallel crystal Fabry-Perot.
asymmetric case, which corresponds to LiNbO\(_3\) or LiTaO\(_3\) with a dielectric mirror on one surface \((R_1 = 0.15, R_2 = 0.99)\), the Fabry-Perot interference only slightly perturbs the intensity and phase from the ideals of \(I_m = I_i\) and \(\phi_f = \phi_o + \phi_m\).

Because both phase and amplitude modulation are produced by the parallel crystal MSLM, it is not suitable for many phase modulation applications. For example, attempts to visualize \(\phi_f\) with a reference beam\(\sqrt{I_2} e^{i\phi_2}\) in an external interferometer (e.g. Twyman-Green) result in the intensity

\[
I = I_2 + G(\phi_m)I_1 + 2\sqrt{I_2 I_1} G(\phi_m) \cos(\phi_1 - \phi_2 + \phi_f(\phi_m)) \quad (B.61)
\]

Here \(E_1 = \sqrt{I_1} e^{i\phi_1}\), \(G(\phi_m) = |E_m/E_1|^2\) of Eq. (B.55), \(\phi_f(\phi_m)\) is given by Eq. (B.56). Equation (B.61) can be simplified by employing such approximations as Eqs. (B.59) for \(\phi_f\). Figure B.15 illustrates the behavior of \(I\) as a function of \(\phi_m\) for various values of \(\phi_1 - \phi_2\) in the case of LiTaO\(_3\) or LiNbO\(_3\) \((R = 0.15)\).

The parallel crystal MSLM does have interesting applications as an intensity modulator; particularly in a high finesse configuration with mirrors on both crystal surfaces. For example according to the \(R_1 = R_2 = 0.9\) curve of Fig. B.13, when \(\phi_o + \phi_m = n\pi\) only a small change in \(\phi_m\) is required to produce full intensity modulation. This effect can be employed to increase the charge sensitivity and decrease the write and erase times of the MSLM in intensity modulation applications. The Fabry-Perot intensity modulator also exhibits other interesting properties, such as threshold by employing the sharp break from maximum intensity into an intensity null. Some additional applications are mentioned in appendix B.4.
Fig. B.15 Intensity, I, produced by a parallel-crystal Fabry-Perot in an external interferometer. A readout beam $\sqrt{I_1}e^{i\phi_1}$ is reflected from a parallel-crystal MSLM and then interfered with the reference beam $\sqrt{I_2}e^{i\phi_2}$ to produce the intensity, I, as a function of the internal crystal phase, $\delta$. (Here $R_1=R_2=0.15$ corresponds to LiTaO$_3$ or LiNbO$_3$, $I_1=1$, and $I_2=0.45$)
B.4. **Special functions achievable with the MSLM**

The fundamental MSLM operation is to modulate the phase of a readout beam in proportion to the irradiance of a control beam. By employing an external interferometer or Fabry-Perot crystal, with reflections from both faces, the readout beam becomes a grey-scale intensity representation of the control-image. With a suitable electrooptic crystal, (e.g. KDP), between polarizers, the amplitude of the readout beam is modulated. In the framed modes images can be stored for long periods (e.g., weeks), depending on the bulk and surface resistivity of the electrooptic crystal.

A variety of more sophisticated operations are also possible with the MSLM. These include: thresholding of the control image (including level-detection or hardclipping), edge enhancement, contrast reversal, image addition and subtraction, and binary logic operations (NOT, AND, OR, EXCLUSIVE OR, and NOR).

**Threshold operations**

Section 4.2.1.3 shows that in the inverted-secondary framed mode, charge is only removed from regions in which the control-image irradiance exceeds a distinct threshold (see Fig. 4.10). This threshold can be precisely controlled by adjusting $V_b$, (the rate at which $V_b$ is reduced), or $V_m$ (which controls the MCP gain). The slope of the $V_x$ vs $I$ characteristic is usually so large that a real-time level-detection or hardclipping operation results, where all image areas below the threshold level remain unmodified at $V_x = V_{b0}$ while those above saturate at the uniform level $V_x = V_{bN}$.

Besides real-time thresholding, stored-image thresholding operations can also be performed. After an image has been written by electron deposition (e.g., primary framed mode), a charge distribution, $C_{X,Y}(x,y)$ corresponding to the write image irradiance, is stored on the crystal
surface. This image can be thresholded by a two-step process. First
\( V_b \) is reduced to a voltage \( V_{b1} \) with the MCP electron flux turned off;
capacitive coupling results in the gap voltage distribution
\( V_g(x,y)=V_{b1}-V_x(x,y) \). Next a uniform erase-beam of electrons is applied
and \( V_b \) is ramped down still further to \( V_{b2} \). According to Fig. 4.8, regions
where \( V_g \) was less than \( V_{g3} \) after the first step maintain their previous
level, \( V_x \). However regions below the threshold level, \( V_{xt}=V_{b1}-V_{g3} \) are
uniformly erased to \( V_x=V_{b2} \) in Fig. B.16. This process can be repeated at
successively higher threshold levels, \( V_{xt} \), on the same captured image to
achieve multiple thresholding. Fig. B.16 is a sketch of the \( V_x \) vs. previously
stored \( V_{xin} \) transfer function for stored image thresholding.

**Contrast Enhancement**

If \( V_b \) in the inverted-secondary framed mode is chosen to place the
small dynamic irradiance range of a low contrast image just below \( I_t \) of
Fig. 4.10, the very large \( V_x \) versus \( I \) slope will result in large real-time
contrast enhancement. The crystal voltage will be extended over the full
dynamic range \( \Delta V_x=V_{b0}-V_{bN} \).

The contrast of a weak stored image \( V_x=V_{x1}+\Delta V_x(x,y) \) (\( \Delta V_x<100 \) volts)
can be enhanced relative to its surroundings (\( V_{x1} \)) by exploiting the
previously discussed stored-image thresholding technique. With no
electron flux, \( V_b \) is reduced to \( V_{x1} \); resulting in \( V_g<V_{g3} \) in the weak
image (note: \( V_g=V_b-V_x=-\Delta V_x<V_{g3} \)), but \( V_g=0>V_{g3} \) in the surrounding
regions. When \( V_b \) ramps down with uniform illumination, the contrast between
the image and its surroundings is greatly improved. This effect is
particularly impressive when the initial image lacked adequate contrast
to be visible in an interferometric readout system.
Fig. B.16 Transfer function for stored image thresholding.
(The crystal charge is $\sigma_x = C \frac{V_x}{A}$, and the corresponding phase modulation is $\pi \frac{V_x}{V_\pi}$.)
When intensity modulation rather than phase modulation is desired, C. Warde has suggested that intensity contrast can be enhanced by putting mirrors on both crystal surfaces. The resulting Fabry-Perot interferometer has a readout intensity \( I_{\text{out}} \) versus modulator phase characteristic similar to \( R_1 = R_2 = 0.9 \) in Fig. B.13. A low-contrast stored image can be made visible by writing additional spatially uniform charge to move it to one of the high-gain regions, e.g., bb' or b'b" in Fig. B.13. Actually, as depicted in the \( R_1 = R_2 = 0.9 \) curve of Fig. B.14, the phase also has high-gain regions; but the accompanying null in \( I_{\text{out}} \) makes these difficult to employ.

**Edge Enhancement**

Edge enhancement of images has been observed and is theorized to be produced by the processes outlined below.

Real-time edge enhancement can occur in the inverted-secondary framed mode. Consider an object in the image with illumination \( I_{\text{obj}} \) on a background with illumination \( I_{\text{back}} < I_{\text{obj}} \). Edge enhancement occurs when \( V_b \) is chosen to place \( I_{\text{obj}} \) slightly below \( I_t \) of Eq. (4.18). As \( V_b \) begins to ramp down, the weak \( I_{\text{back}} \) region immediately "locks out". Most of the repelled electrons return to the MCP. However some are deflected into the object region, which is not yet locked out, and contribute to the MCP electron flux \( J_{\text{edge}} \) at the edge of the object. (According to Eq. B.48a \( J \) is proportional to \( I \), e.g., \( J_t = (F/Ar)I_t \).) As \( V_b \) continues to decrease the object interior will lock out before the edge since \( |J_{\text{edge}}| > |J_{\text{obj}}| \). Some of these interior electrons make an additional contribution to the edge current, raising \( J_{\text{edge}} \) above the threshold level \( J_t \). At this point both the interior of the object and the background are locked out, but \( (\text{since } |J_{\text{edge}}| > |J_t|) \) the contrast of the edge can be arbitrarily increased by decreasing \( V_b \).
Stored-image edge enhancement is hypothesized to make use of stored-image thresholding on one side of the edge and the above real-time process on the other side. Consider as object region charged to \(V_{x1}\) surrounded by a background of lesser charge \(V_{x2}\). If \(V_b\) is adjusted to \(V_{b1}\) \((V_{x2} < V_{b1} < V_{x1})\) with no e\(^-\) flux; when a uniform control beam is applied, the object region will have a negative gap field \((V_g = V_{b1} - V_{x1} < V_g < 0)\) which repels electrons. Some of these repelled electrons will be attracted by the positive gap field \((V_g = V_{b1} - V_{x2} > 0)\) in the background, contributing to the to edge current \(J_{edge}\). If \(V_b\) is now ramped down at a rate such that the MCP current is sufficient to remove charge only in the edge regions \(|J_{back}| < |J_t| < |J_{edge}|\), only the edge will not lock out. Again, the edge contrast increases as \(V_b\) decreases.

It has been suggested that edge enhancement may occur when an object is written with \(s = 2n\pi\) on a background with \(s = 2(n \pm 1)\pi\) in Fig. B.13. The object and background are both at intensity nulls, but the edge includes a maxima. 

Contrast Reversal

The inverted-secondary framed mode exhibits reversed contrast phase modulation relative to the continuous mode or e\(^-\) deposition framed modes. In the former case, the crystal voltage \(V_x\) decreases as irradiance \(I\) increases; in the latter cases \(V_x\) increases with \(I\). When a large dynamic range \((V_x > V_a)\) is required, the inverted-secondary framed mode performs best with halftone (two irradiance level) images.

In applications where the readout phase is employed modulo \(2n\pi\), writing a uniform offset phase can be equivalent to reversing the contrast. For example, the irradiance in an external interferometer is \(I_0 + I_1 \sin(\phi - \phi_m)\); after a uniform \(V_p\) is added to \(V_x\), this becomes \(I_0 - I_1 \sin(\phi' - \phi_m)\). (\(\phi' - \phi_m\) is the optical path difference in the interferometer and \(\phi_m\) is the controllable MSLM component; i.e., \(\phi_m = \nu V_x / V_p\)). Alternatively, the same effect can be
obtained by adding \( \pi \) radians to \( \phi_1 \). As pointed out by C Warde, with a dual-mirrored Fabry-Perot crystal the intensity contrast of a stored image can be reversed by adding uniform charge to move the image from \( \text{bb}' \text{ to } \text{b}'' \) (or \( \text{b}' \text{ to } \text{c}'' \), etc.) in Fig. B.13. Real-time or stored-image amplitude contrast reversal can be accomplished by merely reorienting the output polarizer of a birefringent MSLM.

**Image addition and subtraction**

When a second image \( B \) is written in the primary framed mode over a previously stored image \( A \), the resulting charge distribution, and hence phase modulation (Pockels effect), is the sum of the two images, \( A + B \). If the first image, \( A \), was written with reverse contrast by the inverted-secondary mode, the charge and phase modulation represent the difference \( B - A \). In writing the second image \( B \), care should be taken to avoid depositing so much charge that \( V_g \) is reduced enough to impair proximity focusing.

**Logic Operations**

The MSLM has properties which make it well suited as an optical parallel digital processing element. Binary or two irradiance-level operation is easily implemented by operating between the crystal voltage saturation limits (i.e., \( V_{bo} \) and \( V_{bn} \) in Fig. 4.10. These saturation logic levels are determined by the \( V_b \) power supply, and hence should be reproducible and spatially uniform. The threshold capability of the inverted-secondary framed mode in Fig.4.10 provides a "schmitt trigger" action with a sharp logic threshold to regenerate logic levels. The intrinsic storage capability of the framed modes provides for data latching and memory functions. Secondary emission also exhibits bistability (e.g., \( V_{g2} \) and \( V_m \) in Fig. 4.8) which may be useful for implementing
"flip-flop" functions. Bistability can also be produced in closed-loop configurations based on the IPL of Section 2.2. A variety of schemes have been proposed to implement many logic functions, such as NOT, AND, OR, NOR, and EXCLUSIVE OR in a single MSLM. One set of logic conventions and functional implementations will be outlined below. Elsewhere [84,22] C. Warde describes a different system based on the properties of the high finesse dual-mirrored Fabry-Perot MSLM.

A hypothetical system employing the inverted-secondary framed mode as the basic building block is illustrated in Fig. B.17. An internal interferometer (Fabry-Perot) MSLM is employed for compactness. The supply voltages are adjusted to place $V_{b_0}$ at an intensity null ($\pi V_{b_0}/V_\pi + \delta = \delta = 2n\pi$ in Fig. B.13 ) for logical '0' and $V_{bN}$ at an intensity maxima $(V_{b_0} - V_{bN} = (2n-1)V_\pi)$ for logical '1'. Although less efficient, the lower finesse $I$ vs $\delta$ characteristic of Fig. B.13 $(R_1 = R_2 = .15)$ is preferable in order to suppress the effects of spatial variations in $V_\pi$ (the backround phase of the electro-optic crystal). The proposed input images consist of arrays of spots to simplify alignment, (e.g., a vertical array of horizontal binary numbers). By operating at a nominal logical '1' light level of $I_1 > 2I_t$ and logical '0' level of $I_0 < .1I_t$, the logic noise margin is about $I_t$. ($I_t$ is the logic threshold irradiance, from Eq. (4.18)

$I_t = c_\lambda \dot{V}_b / A_t F$. ) The major fan-out limitation is beam-splitter losses.

The OR operation is performed by the basic inverted-secondary framed mode buffer when two logic images are simultaneously incident. Since $I_1 + I_1 > I_t$ and $I_1 + I_0 > I_t$.

An AND gate can be implemented by employing an inverted-secondary MSLM with $\dot{V}_b$ increased to move the threshold up to $I_t'' = 3I_t$. Then
Fig. B.17 Possible example of MSLM optical digital logic.

D = Spatial detector (Write-beam side of MSLM)
M = Spatial reflective modulator (Read-beam side of MSLM)
= Optical power distribution
= Optical digital-image paths
= Electrical control signals
$I_1 + I_2 > 4I_t > I_t$ is the only combination which produces a logical '1' output. An even simpler AND implementation is to employ one logic image to write on the MSLM and the other as the readout beam.

The NOT operation can be produced by an inverted-secondary MSLM with $V_{bo}$ and $V_{BN}$ readjusted to result in an intensity maximum and null respectively. This device will also perform a NOR operation on two incident write beams; and with a modified threshold of $3I_t$, it performs the NAND operation.

Alternatively, the NOT operation can be implemented with the e-deposition framed mode if $t_w = t = C V_x / (A r_1)$ (from Eq. (4.12)). With $\delta = 2n\pi$ initially in Fig. B.13, $I_1$ adds $\pi$ to $\delta$ and moves an intensity maximum to a minimum. The noise margin about $I_1$ can be improved by writing for a longer time to saturate $V_x$ at $V_{bo}$, where $V_{bo}$ has been set to a null $(\delta = 2n\pi)$.

Additional functions can be implemented by employing more involved framed modes. For example, an EXCLUSIVE OR (Half adder) cycle begins with the MSLM erased to $V_{BN}$ which is set to an intensity null $(\delta = 2n\pi)$. Then a primary write with $t_w = t_d / 2$ is executed, producing an output which is the logical OR of the two input beams. This is followed by a secondary erase with $V_{bo}$ adjusted for threshold $I_t^1 = 3I_t \leq 2I_1$, which brings the input combination $I_1 + I_1$ back down to a logical '0' at $V_{BN}$.

Most of the special functions described in this appendix have been demonstrated in the laboratory. (The initial demonstrations by this researcher were not photographed; however another researcher J. Thackara, recently made the demonstration photographs in Figs. B.18 through B.23.)

The visible photocathode MSLM sealed by Hamamatsu Corp. was employed.)
Contrast reversal is illustrated in Fig. B.18. The square in (a) was written by the electron deposition mode; and (b) was written by the inverted-secondary mode. Image addition of horizontal and vertical bars is demonstrated in Fig. B.19. Fig. B.20 illustrates subtraction of a horizontal bar from a disk; in (c) a vertical bar is added. In Fig. B.21b the spot on the lower left was removed by analogue thresholding. Fig. B.21c shows digital thresholding (or hard-clipping); all areas above the threshold have the same readout intensity. Fig. B.22 demonstrates stored-image contrast enhancement of the horizontal bars. Fig. B.23b exhibits the logical AND operation between the horizontal and vertical bar images in (a).
Fig. B.18  (A), (B) CONTRAST REVERSAL

Fig. B.19  (A), (B) IMAGE ADDITION

Fig. B.20  (A)-(C) IMAGE SUBTRACTION

Fig. B.21  (A)-(C) ANALOG AND DIGITAL THRESHOLDING

Fig. B.22  (A)-(B) CONTRAST ENHANCEMENT

Fig. B.23  (A), (B) AND OPERATION
B.5 MSLM sensitometry

A large number of measures are available to evaluate the sensitometry of optical devices; in order to avoid confusion, the measures used are defined here. The measures of light flux employed in this thesis are radiant flux density, \( I \) measured in \( \text{watts/cm}^2 \), and exposure, \( I \cdot t \) measured in Joules/cm². In the literature \( \text{W/cm}^2 \) is often termed "radiant exitance," "radiant emittance," or "irradiance." No use is made of photometric units based on the "lumen", which is weighted by the human eye response.

Since the MSLM converts light into an electron current density \( J \) incident on the electro-optic crystal, current responsivity, \( r_I \), is an important characteristic:

\[
r_I = \frac{J}{I} = \frac{\text{Amps/cm}^2}{\text{Watts/cm}^2} = \frac{\text{Amps}}{\text{Watts}}
\]  

(B.63a)

The conditions of measurement should be stated since \( r_I \) depends on \( \nu = \text{frequency}, \ V_{\text{mcp}}, \ V_g, \) and \( J \). Unless otherwise specified, it will be assumed that \( V_g \) is positive enough that all secondary electrons return to the crystal and \( J \) is in the linear region below MCP saturation; then

\[
r_I = nG \frac{e}{h\nu}
\]  

(B.63b)

Here \( n \) is the photocathode quantum efficiency, \( G \) is the MCP gain, \( e \) is the electron charge, and \( h \) is Planck's constant. The expected value with one MCP is \( r_I = 500 \text{ A/W} \) (\( G = 10^4, n = 10\%, h\nu = 2 \text{ eV} \)); and with two MCP's \( r_I = 5 \times 10^5 \text{ A/w} \) (\( G = 10^7 \)).
In the electron-deposition framed modes, the crystal voltage $V_x$ and phase modulation are proportional to the exposure or integrated irradiance. The exposure to voltage transfer function is the speed responsivity, $r_s$

$$
\frac{dV_x}{dt} = \frac{volts/sec}{watts/cm^2} = \frac{V_x}{I_t} = \frac{volts}{Joules/cm^2} = \frac{A_r}{C_x r_I}
$$

(B.64a)

Here

$$
C_x = \frac{\varepsilon_x \varepsilon_0 A_r}{\lambda_x}
$$

(B.64b)

is the crystal resolution cell capacitance, $A_r$ is the area of a resolution cell, $\lambda_x$ is the crystal thickness, and $t_w$ is the write time. As the speed responsivity increases, the MSLM writes faster with a given irradiance $I$. Both $r_s$ and $I$ must be specified in order to infer the time response.

A more conventional measure is the half-wave exposure sensitivity, $S_{\pi}$, which is the exposure required for a phase shift of $\pi$ radians:

$$
S_{\pi} = \frac{V_{\pi}}{r_s} = \frac{C_x V_{\pi}}{A_r r_I} = \frac{\sigma_{\pi}}{nGe} = \frac{Joules}{cm^2}
$$

(B.65)

Here $V_{\pi}$ is the half-wave voltage and $\sigma_{\pi} = C_x V_{\pi} / A_r$ is the half-wave surface charge density. The expected half-wave sensitivity with one MCP is on the order of $S_{\pi} = 0.3 \text{nJ/cm}^2$ ($r_I = 500 \text{ A/w}$ and $\sigma_{\pi} = 1.6 \times 10^{-7} \text{Coul/cm}^2$ for a 0.5 mm thick LiNbO$_3$ crystal) and $S_{\pi} = 0.3 \text{ pJ}$ for two MCP’s ($r_I = 5 \times 10^5 \text{ A/w}$).

If not measured directly, $S_{\pi}$ can be calculated from other measured
quantities via Eq. (B.65); as long as \( V_\pi \) is physically realizable and \( I \) is not large enough to cause MCP saturation.

At a given irradiance \( I \), the time \( t_\pi \) to write \( \pi \) radians of phase shift is

\[
t_\pi = S_\pi (I) / I = \frac{\hbar \nu \sigma}{\eta e G(I) I} \tag{B.66}
\]

Here \( S_\pi \) is written as a function of \( I \) because the shortest rise times occur for large irradiances, which places the MCP in a quasisaturation regime where \( G(I) \) decreases with increasing \( I \).

In the continuous mode, the steady-state voltage and phase modulation are proportional to \( I \). Useful measures are voltage responsivity \( r_v \), which relates \( V_x \) to irradiance, and half-wave irradiance \( I_\pi \) which is the irradiance required for \( \pi \) radians of modulation

\[
r_v = \frac{V_x}{I} = \rho_x \frac{\ell_x}{R_x} I = \frac{\text{volts}}{\text{watts/cm}^2} \tag{B.67}
\]

\[
I_\pi = \frac{V_\pi}{r_v} = \text{watts/cm}^2 = \frac{\sigma \hbar \nu}{R_x C \eta \text{Ge}} = \frac{S_\pi}{R_x C} \tag{B.68}
\]

(Here \( R_x = \rho_x \ell_x / A_r \) is the resolution cell bulk resistance.) Eq. (B.68) is only really useful when \( V_\pi \) is physically realizable (e.g., below breakdown) and \( I_\pi \) is not in the saturation regime where \( r_v \) and \( r_l \) change with \( I \).

The quantities \( r_l \), \( r_v \), and \( I_\pi \) convey no direct information on the continuous-mode time response, which is characterized by the time constant \( \tau \propto R_x C_x \).
Half-wave sensitivity is a less meaningful measure for the continuous mode. However an output step of $\pi$ radians requires that an irradiance of $I_\pi$ be present for the rise time period $t_{\pi}^\text{R} 2.2 \tau_2$. This corresponds to an effective half-wave sensitivity of

$$S'_\pi = I_\pi t_{\pi}^\text{R} = \frac{V_\pi}{r_\pi} = \frac{V_\pi}{r_\pi} (2.2R_x C_x) = 2.2 \frac{C_x V_\pi}{A_\text{r} r_I} = 2.2 \frac{h \nu r_\pi}{n \text{Ge}} = 2.2 S_\pi$$

(B.69)

Here $S_\pi$ is given by Eq. (B.65). Using this measure the primary framed mode is about twice as sensitive as the continuous mode.

The measures of MSLM sensitivity generally improve, and the MSLM can operate with greater speed and dynamic range for a given write-beam irradiance, as the MCP gain $G$ is increased. However sections 3.1.9 and 4.2.3.1 showed that noise (e.g. shot noise) gets more serious as $G$ is increased (and $S_\pi$ is decreased), especially with a high-resolution MSLM. In a particular application, noise performance can be traded off against speed, dynamic range, and other performance considerations to find an optimal value of $G$ and associated minimum usable signal energy. The discussions of sections 3.1.9 and 4.2.3.1 reveal that when signal shot noise is the dominant noise source, the minimum required irradiance $I_m$ is generally that which produces a few photons in a resolution element during the write-time interval. Two MCPs are often sufficient to obtain the optimal gain, although three MCPs may be required when the resolution is low. When dark current is important, the number of photons produced by $I_m$ should exceed the square root of the number of dark current counts, which can allow operation with $I_m$ significantly below the effective input irradiance $I_d$ corresponding to the dark current (typically $I_d=10^{-15} \text{W/cm}^2$).
B.6. Proximity focussing in the MSLM

Theoretical models which agree well with experimental data are presented in the literature, and reviewed here, for proximity focusing from a photocathode\textsuperscript{244} and from an MCP\textsuperscript{224}. The basic concept underlying these models is that electrons are emitted with an angular distribution of velocities, which causes a lateral spread in position as they are accelerated across the proximity focus gap of width \( \lambda \). From elementary classical mechanics

\[ \lambda = \frac{1}{2} at^2 + v_\parallel t = \frac{1}{2} \frac{eV}{m} t^2 + v_\parallel t \quad (B.71) \]

Here \( e \) is the electron charge and \( m \) is the electron mass. If the axial component of emission velocity \( v_\parallel \) is small, i.e. \( v_\parallel^2 \ll 2V_g/m \), the transit time is approximately

\[ t = \sqrt{\frac{2m}{eV}} \lambda \quad (B.72) \]

The corresponding radial displacement \( r \), due to the radial emission velocity \( v_\perp \) or radial energy \( eV_r = 1/2 v_\perp^2 \) is

\[ r = v_\perp t = 2\sqrt{\frac{V_r}{V_g}} \lambda \quad (B.73) \]

Given the statistical probability distribution \( P(V_r) \) of the initial radial energy, it follows from elementary probability theory that the radial probability distribution is
\[ p(r) = p(V_r) \left[ \frac{dV}{dr} \right] = \frac{V}{\frac{V_r}{2}} P(V_r) = \frac{\sqrt{V}}{2\pi \sigma^2} e^{\frac{-r^2}{2\sigma^2}} P(V_r) \] (B.74)

Eq. (B.73) was used to evaluate the derivative. The point spread function \( psf(r) \) is the fraction of electron current density, \( J(r) \), from a pencil beam (of current \( I_p \)), which falls between \( r \) and \( r+dr \), i.e.

\[ psf(r) = \frac{J(r)}{I_p} = \frac{P(r)dr}{2\pi \sigma^2} = \frac{Vg}{4\pi \sigma^2} P(V_r) \] (B.75)

The MTF, \( M(u) \), for proximity focus is then the Fourier transform of Eq. (B.75).

\[ M(u) = \mathcal{F}(P(V_r)) = \frac{Vg}{2\pi \sigma^2} e^{\frac{-r^2}{2\sigma^2}} \] (B.76)

The radial emission energy distribution \( P(V_r) \) from a photocathode can be approximated by a Maxwellian distribution:

\[ P(V_r) = \frac{1}{\bar{V}_r} e^{-V_r/\bar{V}_r} \] (B.77)

Here \( \bar{V}_r \) is the average radial emission energy, and using Eqs. (B.73) and (B.75)

\[ psf(r) = \frac{Vg}{4\pi \sigma^2} e^{\frac{-r^2}{2\sigma^2}} \]

The Fourier transform of Eq. (B.75) is the MTF

\[ M_p(u) = e^{-u/\bar{V}_r} = e^{(u/N_p)^2} \] (B.79)
The resolution is conveniently taken as that frequency, \( N_p \), where the contrast decreases to \( e^{-1} = 37\% \)

\[
N_p = \frac{1}{2\pi \lambda} \sqrt{\frac{V_g}{V_r}} \quad \text{lp/mm}
\]  

(B.80)

where \( \lambda \) is the gap width specified in mm. Using a slightly different approach, the Russians present an MTF formula for proximity focus from an MCP

\[
M_m(u) = \exp \left( -\frac{12 \lambda^2 V_r'}{V_g u^2} \right) = e^{-(u/N_p)^2}
\]  

(B.81)

The corresponding \( e^{-1} \) resolution is

\[
N_m = \frac{1}{\sqrt{12\lambda}} \sqrt{\frac{V_g}{V_r}} \quad \text{lp/mm}
\]  

(B.82)

Approximately 80% of the electrons have radial energy less than \( eV_r' \), Eq. (B.82) matches the experimental MTF well when

\[
V_r' = V_{a.9} \sin^2 \theta_{.9}
\]

Here 90% of the electrons are emitted with energy less than \( eV_{a.9} \) and 90% have directions within a cone angle of \( \theta_{.9} \) from the channel axis. According to the numerical data in reference it appears that \( V_{a.9} \) is well approximated by \( \bar{V}_a \) of the high energy portion of the MCP energy distribution of Eq. B.20 in Appendix B.2. For the MCP in Ref. 224, as \( V_m \) is increased from 600 volts to 1350 volts, \( \theta_{.9} \) decreases from 7.75\(^\circ\) to 6.5\(^\circ\); and \( V_{a.9} \) grows from 60 volts to 100 volts as \( V_m \) varies from 1000 V to 1350 V.
Equation (B.82) can be utilized to estimate the minimum gap voltage $V_{gm}$ which must be maintained in the electron-deposition framed or continuous modes in order that proximity focusing from the MCP does not degrade the MSLM resolution below $N_m$.

$$V_{gm} = 12z^2 V'_r N_m^2$$  \hspace{1cm} (B.83)

As an aid to MSLM operation, table B.1 gives values of $V_{gm}$ for typical MSLM specifications. Ref. was used to estimate $V'_r$.

Table B.1

Minimum MCP-to-crystal voltage, $V_{gm}$, for proximity focus

<table>
<thead>
<tr>
<th>$\lambda$ (gap width) (mm)</th>
<th>$N_m$ (1p/mm)</th>
<th>$V_m$ (volts)</th>
<th>$V'_r$ (volts)</th>
<th>$V_{gm}$ (volts)</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>5</td>
<td>1000</td>
<td>1.0</td>
<td>75</td>
<td>sealed system</td>
</tr>
<tr>
<td>0.5</td>
<td>5</td>
<td>1200</td>
<td>1.5</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>20</td>
<td>1000</td>
<td>1.0</td>
<td>1200</td>
<td>usual sealed system operation</td>
</tr>
<tr>
<td>0.5</td>
<td>20</td>
<td>1200</td>
<td>1.5</td>
<td>1800</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>5</td>
<td>1200</td>
<td>1.5</td>
<td>4.5</td>
<td>demountable system</td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td>1200</td>
<td>1.5</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>5</td>
<td>1200</td>
<td>1.5</td>
<td>450</td>
<td>large gap</td>
</tr>
<tr>
<td>1.0</td>
<td>20</td>
<td>1200</td>
<td>1.5</td>
<td>7200</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>30</td>
<td>1200</td>
<td>1.5</td>
<td>4050</td>
<td>high resolution</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>1200</td>
<td>1.5</td>
<td>11250</td>
<td></td>
</tr>
</tbody>
</table>
B.7 Experimental development of the MSLM

This appendix briefly reviews some of the experimental results obtained in the course of developing the MSLM. The basic experimental configuration employed to evaluate demountable MSLMs is sketched in Fig. B.24. The setup used to study the sealed visible-photocathode MSLM is depicted in Fig. B.25, including the electronic bias circuits.

The photographs in Fig. B.26a were made in the image plane of the demountable MSLM configuration of Fig. B.24, with a vertical bar ultraviolet pattern being written onto the photocathode (Gold). The image, which was produced by interference between light reflected from the front and back crystal surfaces, is the first readily discernable image read out from an MSLM. Figures 4.13 and B.18 to B.23 illustrate improved images which were subsequently obtained with the visible-photocathode MSLM.

In some instances the crystal in the demountable MSLM was replaced by a phosphor screen in order to more carefully study such resolution degrading effects as: improper uv focus, poor proximity focus, undesired writing by indirect uv light, and charge deposition due to such extraneous effects as point emission. Figure B.26b shows the word "TIM" and other patterns written onto a phosphor screen. The two bright spots in the upper right region are due to point emission between the MCP and screen. A large number of such spots appeared when the screen-to-MCP voltage exceeded 5 kV. In the usual case of a crystal instead of a phosphor screen, the invisible uv write-beam was focused by passing it through a 2537Å interference filter to a flourescent target painted around the photocathode.
Fig. B.24 Experimental configuration for evaluating demountable MSLMs.
Notes:

1) Scope #1 monitors changes in the MCP electron flux, since the strip current is static and when the crystal is neither charging nor discharging all the electrons return to the MCP output electrode.

2) The meter labeled $V_f$ monitors $V_{G2}$, the electrostatic lens voltage.

3) Many of the resistors serve the function of protecting the device against internal arcing or shorts.

4) $M = 10^6$ ohm resistance, $K = 10^3$ ohms.

5) $S_1$ initiates a sequence of:
   a) Trigger scope #2.
   b) Turn on LED for fixed time interval $T_c$.
   c) After a short delay (~40 msec, turn on MCP for time interval $T_m$.

6) Scope #2 or the DVM monitors the charge on a small region of the crystal, by way of the interferometric intensity changes caused by charge induced phase modulation in the crystal.

Fig. B.25 Experimental configuration for evaluating the sealed visible-photocathode MSLM.
a) First discernable MSLM image. (A vertical bar written to 2 charge levels, $V_x = 5$ and $V_x = 7$ kV).

b) Image written onto a phosphor screen in the place of the MSLM crystal.

Fig. B.26 Demountable MSLM imaging experiments.
The earliest MSLMs employed very distorted electrooptic crystals which resulted in ten or more fringes per inch in the readout interferograms created by light reflected from the front and back crystal surfaces. As a guide to future researchers, a double exposure photographic technique will be described which was found useful for detecting image-induced shifts in these complicated readout fringe patterns. With the film placed in the image plane of Fig. B.24, the first exposure was of the readout interferogram from the uniformly discharged \((V_x(x,y)=0)\) or charged \((V_x(x,y)=V_b)\) crystal. The second exposure was of the readout interferogram after writing an image charge distribution onto the crystal. The fringes tended to move and hence were broader in the photographs of the image-charged areas. Part of the readout interferogram was often masked out during the second exposure in order to leave an area of unwritten fringe pattern to be used as a reference in quantitative measurements of the image-induced fringe shifts.

The MSLM write and erase times were measured by observing the output of a detector placed behind a narrow slit in the readout-beam image plane. The noise immunity in the demountable configuration of Fig. B.24 was increased by employing a chopped read beam and a tuned amplifier following the photodiode. Figure B.27 shows the time response of a demountable MSLM; in Fig. B.27a the electron-deposition 0 to 2.5 kV rise time (10% to 90%) is about 270 msec. Fig. B.27b depicts a fall time of about 840 msec \((e^{-1}\) in 340 msec) obtained by leakage through a grounded semiconducting coating on the crystal. Rise times as short as 100 msec for 2.5 kV were observed in demountable systems, the primary limitation being the maximum available uv source intensity.
a) Electron-deposition rise time following turn on of the write-light beam. (Demountable MSLM)

b) Resistive erasure fall time following a step decrease in $V_b$. (Demountable MSLM)

c) Electron-deposition rise time following turn on of the write-light beam. (Sealed visible photocathode MSLM)

Fig. B.27 Time response measurements. (The time response of the output current from a photodiode behind a narrow slit in the readout interferogram is displayed. The photodiode gives an interferometric indication of the modulator phase and thereby $V_x$.)
Figure B.27c shows an electron-deposition 0 to 4 kV rise time of about 200 msec obtained with the sealed visible photocathode configuration of Fig. B.25. This time was near the strip current limited value, with the MCP output current density being approximately equal to the strip current density.

Figure B.28a is from a series of experiments in which secondary emission was first demonstrated in the MSLM. The upper trace is the bias voltage $V_b$ of Figs. 4.1 and 4.2. The lower trace is the image-beam intensity seen by the image-plane photodetector in Fig. B.24, which is proportional to $V_x$ and increases as electrons are deposited on the crystal. With no write-beam, the decrease in $V_b$ at $t_1$ produces only a small change in $V_x$ due to capacitive voltage division between the crystal and gap ($\Delta V_x = C_g \Delta V_b / (C_g + C_x)$). At $t_2$ the write-beam is driving an MCP electron current which causes $V_x$ to decrease with $V_b$ by secondary emission. At $t_3$, $V_b$ is decreased faster then the maximum erasure rate for the given MCP current level. Then, as discussed in Section 4.2.1.3, the system "locks out" with $V_x$ decreasing only slightly and $V_b$ becoming so negative that no electrons can reach the crystal surface. Fig. B.28b shows similar image-beam behavior with a thicker crystal ($x = 4\text{mm}$) so that $C_x = C_g$.

Figure B.28c is a more quantitative representation of this behavior, obtained by carefully calibrating the photodiode output, as read on a DVM, against $V_x$. The AB and DE segments are locked-out and have a slope of $C_g / (C_x + C_g) = 0.15$ as expected from independent measurements of $C_x$ and $C_g$. The B to B' jump occurs because the increase in $V_b$ from A to B capacitively couples sufficient voltage into the gap to decrease
a) Response of $V_X$ (readout interferometric intensity) to $V_b$ with write-light on and off. Capacitive voltage division occurs at $t_1$, secondary emission at $t_2$, and lockout at $t_3$.

b) Response of $V_X$ to the following sequence:
- $t_0$: $V_b=2.9$ kV (capacitive voltage division)
- $t_1$: write-light on (electron deposition)
- $t_2$: write-light off
- $t_3$: $V_b=0$ volts (capacitive voltage division)
- $t_4$: write-light on (locked out)
- $t_5$: $V_b=2.9$ kV (electron deposition)
- $t_6$: $V_b=0$ volts (secondary emission)

Fig. B.28 Secondary emission behavior.
c) Response of $V_x$ to $V_b$. The write-light is continuously on, but segments DE and AB are locked out due to a rapid decrease of $V_b$. The slope of the DE and AB segments is $C_g/(C_g+C_x)$. BB' is an example of recovery from lock out (and bistability).

d) Response of $V_x$ to the instantaneous switching of $V_b$ from 0 to -500 volts at $t_0$. Secondary emission eventually reduces $V_x$ to -500 volts. Before $t_0$, $V_b$ was increased in 100 volt increments to calibrate the photodiode current in terms of $V_x$.

Fig. B.28 Secondary emission behavior.
the negative locking-out \( V_g \) voltage enough to allow some MCP electrons to reach the crystal. These electrons then cause secondary emission, which increases \( V_g \) to approximately zero \( (V_{gL} \) in Fig. 4.8) and results in a corresponding decrease in \( V_x \) \( (V_x = V_b - V_g) \).

In Fig. B.28d, \( V_b \) was first increased in 100 volt increments to calibrate \( V_x \), and then at \( t_o \), \( V_b \) was rapidly decreased by 500 volts. (The photodiode happened to be at a nonlinear part of the interferogram). The small fraction of MCP electrons with more than 500 volts of energy caused \( V_x \) to eventually decrease by secondary emission. Data of the type in Fig. B.28d can be used to experimentally determine the MSLM's \( i_g \) versus \( V_g \) characteristic. Since \( V_b \) is fixed, the gap voltage is given by \( V_g = V_b - \Delta V_x = -500 - \Delta V_x \), where \( \Delta V_x \) is the decrease in \( V_x \) after \( t_o \), as read directly from the calibrated photograph. The gap current \( i_g \) can be determined from the instantaneous rate of change of \( V_x \) in the photograph, i.e. \( i_g = (C_g + C_x) dV_x / dt \). As discussed in Section B.2.2, the \( i_g \) versus \( V_g \) characteristic can be employed to make valuable inferences on the MCP energy distribution and the crystal secondary emission properties. In other experiments, \( i_g \) versus \( V_g \) characteristics were measured directly with the crystal replaced by conducting targets of graphite, aluminum, or stainless steel.

A variety of measurements of MCP current \( i_p \) as a function of write beam irradiance \( I \) were made in order to evaluate the performance of the photocathode-MCP combination. The MCP current was collected with a positively biased electrode replacing the crystal, or a small electrode between the crystal and MCP, or directly at the MCP output electrode (anode). The latter method is feasible because when the crystal is
fully charged, most of the MCP output electrons return to the MCP anode; if this component is time varying, it is easily separated from the MCP strip current. The increase of MCP anode current with illumination, however, underestimates the actual electron flux (~50% of total), because when the MCP is emitting electrons the actual strip current flowing in the walls decreases. Beside chopper modulation and attenuating filters, the write-beam intensity was modulated by varying the power into the write-light source. Fig. B.29a shows MCP current for various MCP bias voltages \( V_m = 0, 500, 970, 1070, 1190, 1280, 1400, 1510 \) with a uv lamp powered by 60 Hz ac. Simultaneous quantitative measurements of the uv intensity were made through a 2537Å interference filter with a calibrated photodiode, allowing Fig. B.29a to be converted to a family of curves of \( i_p \) versus \( I \) with \( V_m \) as a parameter. The expected \( a e^{b V_m} \) dependence of \( i_p \) on \( V_m \) was observed in this and other experiments, but with a decrease of \( b \) by about a factor of 2 for \( V_m > 1050 \) volts. Ion feedback noise was observable above \( V_m = 1300 \) volts. Direct measurements were also made of the photocathode emission current in order to determine the quantum efficiency of the various photocathodes employed.

Figure B.29b shows the response of MCP current to a step in illumination. The initial overshoot and decay in Fig. B.29b is due to the high initial MCP flux \( (i_p = 1.9\mu A, i_{strip} = 5.1\mu A) \) depleting charge from the pore walls faster than it can be replenished. When \( i_p \) is less than about 20% of the strip current there is no initial overshoot.

A variety of additional measurements were made, some of which allowed the model parameter values in Fig. 4.1 and 4.2 to be experimentally determined. For example, a crystal with its MCP surface grounded through a partially conducting coating was connected through a known
a) The variation of MCP current with time varying illumination, for various MCP bias voltages, $V_m=0, 500, 970, 1070, 1150, 1280, 1400, 1510$. The illumination $I(t)$ was also measured, allowing $i_p$ versus $I$ to be determined for various values of $V_m$.

b) Response of MCP current to a step in illumination. The initial pulse is due to saturated operation, with $i_p=40\% i_{strip}$. The time integral of the pulse gives the amount of excess charge available in high current pulsed MCP operation.

Fig. B.29 Microchannel plate (MCP) measurements.
capacitance, $C_0$, to $V_b$. Detecting the read-beam with the photodiode in Fig. B.24 provided a noncontacting means for reading $V_x$, which allowed $C_x$ to be calculated ($C_x = (V_b - V_x)C_0/V_x$). Similar charging and discharging experiments also allowed $C_g$ to be measured; the associated rise and fall transient responses provided values for $R_x$ and $R_L$.

A few specialized circuits were constructed to aid in MSLM development. In order to facilitate the measurement of the weak photocathode and MCP currents, a well-shielded, sensitive, low-noise, current amplifier was built. (Commercial tuned amplifiers, lock-in amplifiers, and sensitive electrometers were also employed). A sensitive photodiode amplifier was assembled to make accurate measurements on the readout-beam interferograms. An electronic shutter and two coupled high-voltage pulsers were constructed in order to implement the sequences required for effective utilization of the MSLM. For example a sequence employed to write in the visible-phocathode MSLM was: turn $V_b$ on, then pulse $V_m$ and write-light on, end write-light pulse after $t_w$ (must be short enough to prevent $V_x = V_b$ and hence loss of proximity focus), end MCP pulse after $t_1 > t_w$ (to prevent stray light from reducing contrast of the image), turn off $V_b$. In an MSLM with a crystal having a low surface resistivity a sequence used was: turn write-beam and MCP on, pulse $V_b$ on for $t_w$, immediately pulse shutter to record the image before it can decay, turn write-beam and MCP off.
Practical considerations concerning the fabrication and use of LiTaO$_3$ and LiNbO$_3$ phase modulators.

Many important aspects of the behavior of LiNbO$_3$ and LiTaO$_3$ were observed which can be critical to the success or failure of a particular application. The electrooptic coefficients, as inferred from measured half-wave voltages, have been observed to vary significantly from manufacturer to manufacturer and even over a single boule. This may be due to crystal impurity variations or improper poling. (Most of the crystals studied were accurately aligned by x-ray diffraction). For example, the z-cut LiNbO$_3$ crystal employed in the visible-phocathode MSLM had an expected reflection half-wave voltage of 4360 volts$^{252}$ $\left( V_{\pi} = \lambda/(2n_0^3r_{13}) \right)$ and measured $V_{\pi}$ of $\sim$8000 volts. The z-cut LiNbO$_3$ crystal in the nineteen-element modulator had an expected $V_{\pi}$ of $\sim$3100 volts$^{252}$, and an actual $V_{\pi}$ of $\sim$6500 volts. (Nonuniform electrode contact cannot account for these large values of $V_{\pi}$; because in the MSLM electrons were deposited directly on one crystal surface and an ITO electrode was sputtered on the other face; and in the nineteen-element modulator, surface conductivity spread charge from points of intimate contact.

In applications where a spatial modulator is constructed by applying a potential between reflecting discrete electrodes on the rear surface of a z-cut crystal and a transparent front electrode (e.g. nineteen-element system), crystal surface conductivity can be a problem. Due to the surface preparation and contamination, the surface of the crystal is generally much more conductive than the highly resistive bulk. As a result, if there is a single electrode or all-but-one electrodes are floating, the modulation is not confined to under the active electrode, but spreads over the whole crystal surface. With multiple electrodes,
the individual electrodes are not resolved, but there are gradual
gradients of modulation between the electrodes. This effect is also
evident in the work of other researchers, e.g. Refs. 210,245,246.
The use of assorted solvents and acid etches did not offer much improve-
ment once a surface became conductive. However, discrete steps in surface
potential could be produced across scratches or scribe marks on the
crystal surface.

Surface conductivity was present but did not have an adverse
effect on the nineteen-element system, because: all the electrodes
were actively driven, the interelectrode gaps were very small, and the
interelectrode regions reflected no light into the modulated beam.
Surface conductivity was a limitation in many implementations of the MSLM,
but not the visible-phocathode MSLM (where the support ring was clearly
imaged).

Since the relative permittivity of these materials is very large
($e_r=50$), it is essential that the gap between the crystal and any
modulating electrodes be small and uniform. For example, with a
crystal thickness of $x=0.5\text{mm}$ half of the applied voltage is lost in a
gap thickness of $g = 0.5\text{mm} / e_r = 10\mu\text{m} < 0.5\text{mil}$; with large applied voltages
even a 10% variation (1 $\mu\text{m}$) can be observable in the modulation! This
effect caused the nonuniform support ring image, evident on Fig. 4.13c
in the visible-phocathode MSLM. In the nineteen-element system,
some of the pads lifted slightly, but surface conductivity had the
beneficial effect of smoothing the potential under the whole pad.

$\text{LiNbO}_3$ and $\text{LiTaO}_3$ are strongly pyroelectric, and must be heated and
cooled very slowly to prevent fracture. The actual rate depends on the
size and shape of the crystal; a conservative rate is 2°C/minute. Although this researcher destroyed no crystals, others have had the misfortune of fracturing crystals by such seemingly innocuous procedures as boiling them in solvents, running cold tap water over them, or applying a volatile solvent (e.g. acetone) to a warm crystal.
C. DESIGN AND OPERATIONAL DETAILS OF THE DISCRETE CHANNEL SYSTEMS

C.1. Single-channel System

C.1.1. Response of the single-channel high-voltage amplifier

Figure C.1 is a schematic diagram of the high-voltage amplifier employed to implement the single-channel IPL discussed in Sections 5.1 and 6.1; and Fig. C.2 is an equivalent circuit for the frequency response of this amplifier. The model subdivides at node $V_o$ into two decoupled circuits.

The full two-pole frequency response of $\tilde{V}_o$ was solved for, but in this application the photodiode ($w_d \approx 2\pi \times 10kH_Z$, Hamamatsu # S876-1010BQ) dominates, and the operational amplifier ($w_f = w_o / A_o = 2\pi \times 4.5MHz$, RCA CA3140) has little effect on the frequency response. In the limit $w_f \to \infty$, the full frequency response simplifies to

$$\tilde{V}_o(w) = \frac{\tilde{V}_r - R_f \frac{R_L}{R_L + R_d} \tilde{I}_d(w)}{1 + jw/w_d} \quad (C.1)$$

where

$$w_d = \frac{1}{(R_L || R_d)C_d} \quad (C.2)$$

The reverse bias arrangement in Fig. C.1, employing $R_s$, increases $w_d$, but at the expense of a higher dark current and noise pickup. Usually $R_L >> R_d$; and when $w << w_d$, $\tilde{V}_o$ simplifies to:

$$\tilde{V}_o = \tilde{V}_r - R_f \tilde{I}_d \quad (C.3)$$
Notes:
1) $R_f1$ and $R_f2$ control gain.
2) $R_2$ controls offset and intensity threshold.
3) $R_c1$ and $R_c2$ control gain and effect offset and threshold.
4) $R$ controls bandwidth.
5) $R_4$ controls the initial value of $V_x$ when the IPL is turned on by closing $S_2$.
6) $S_{2a}$ reduces battery current drain when the loop is open.
7) $I_{C1}$ is a CA3140T op amp.

Fig. C.1 Schematic diagram of the single-channel high voltage amplifier circuit.
Notes:

a) \( R_C = R_{C1} + R_{C2} \)

b) \( R_i = 600 \Omega \) = output resistance of op amp. (Should be in series with \( v_o \) source, but has been moved since \( R_f \gg R_i \) is assumed.)

c) The crystal is modelled as being purely capacitive \( (C_x = 2 \times 10^{-11} F) \), since \( R_x > 10^{13} \Omega \). When frequencies near a piezoelectric resonance are of interest, a series inductance and modified value of \( C_x \) corresponding to the specific resonance should be employed.

d) Since the op amp (CA3140T) has CMOS inputs, its input resistance \( (\approx 10^{12} \Omega) \) and input offset current \( (\approx 10^{-11} A) \) have been neglected.

e) The adjustable op amp input offset voltage is included in \( v_r \).

f) \( A_0 w_o = w_r \) is the unity gain frequency (or gain-bandwidth product) of the op amp. \( (A_0 = 10^5, w_r = 4.5 \times 10^6 \text{ Hz}, w_0 = 45 \text{ Hz}) \)

g) This model ignores the loading effects (10M\( \Omega \) resistance) of oscilloscopes monitoring \( v_o \), \( v_c \), and \( v_t \).

Fig. C.2 Equivalent circuit for the frequency response of the single-channel HV amplifier
Since $i_d$ is often small, the defects of real op-amps and photodiodes can have a significant effect on Eq. (C.3). The generalized op-amp of Fig. C.3a includes input resistance, input offset current, offset voltage, and a bias voltage $V_f$ (which will be used in Section C.2.1). The photodiode is characterized by

$$i_d = i_L - i_s = i_{LO}(\exp(q(v_- - v_+)/k_B T) - 1) - SAI$$  \hspace{1cm} (C.3a)

where

$$R_L = \frac{k_B T}{q i_{LO}} = 0.025/i_{LO}$$  \hspace{1cm} (C.3b)

Here $q$ is the electron charge, $k_B$ is the Boltzmann constant, $v_- - v_+$ is the voltage across the photodiode as defined in Fig. C.3a, $i_s = SAI$ is the photocurrent generated by the irradiance $I$ over area $A$ in a photodiode with sensitivity $S$ (amps/watt); and $R_L$ is the effective dark-current leakage resistance. The steady-state time response corresponding to Fig. C.3a is

$$v_o = \frac{v_r - R_f i_s}{1 + \frac{1}{A D R_f || R_f || R_- || R_L}}$$  \hspace{1cm} (C.4)

The offset voltage, $v_r$, now becomes

$$v_r = \frac{R_f v_+}{R_r || R_f || R_-} + \frac{R_f v_2}{R_r || R_f || R_- || R_L} - \frac{R_f v_f + R_f i_-}{R_r}$$  \hspace{1cm} (C.5)

When $R_f$ is not too large, the denominator of Eq. (C.4) becomes unity and Eq. (C.4) corresponds to Eq. (C.3). On the other hand, the maximum possible gain is limited by the op-amp and photodiode defects to Eq. (C.4) in the limit of $R_f \rightarrow \infty$; which corresponds to the circuit of Fig. C.3b. In the actual single channel amplifier, $R_f \rightarrow \infty$, $i_- = 10^{-11}$amps and $R_- = 10^{12} \Omega$. 
Fig. C.3a Generalized photodiode op-amp model

\[ V_o = A_o (V_2 + V_+ - V_-) \]

\[ V_2 = \text{input offset voltage} \]

Fig. C.3b Maximum gain model

\[ V_o = A_o (V_2 + V_+ - V_-) \]
held. The offset voltage,

$$v_r = v_+ + \frac{R_f}{R_f || R_L} \cdot v_2$$  \hspace{1cm} (C.6)

is the controllable via the op-amp offset voltage ($v_2$) adjustment ($R_2$ in Fig. C.1).

The fact that $v_o$ cannot exceed the supply limit, $0 < v_o < v_1$, adds irradiance thresholding if $v_r > v_1$ and saturation to the response, i.e., Eq. (C.3) is only valid for irradiance in the range

$$I_t = \frac{v_r - v_1}{R_f S A} \cdot U(v_r - v_1) \leq I \leq \frac{v_r}{R_f S A} = I_s$$  \hspace{1cm} (C.7)

Here $U(\ )$ is the unit step function.

The frequency response of $\tilde{V}_x$ as a function of $\tilde{V}_o$ in Fig. C.2 is too tedious to repeat here. At frequencies where the low-pass filter employing $C$ dominates; $C_x(\approx 17 \text{ pf})$, $C_\mu$, and $C_\pi$ can be neglected. In order to make gain and bandwidth adjustments independent of each other, a large value of $C$ was employed along with the constraints: $R << R_e$, $R_b << R_e$ and $\beta > 1$. Within these constraints the frequency response simplifies to:

$$\tilde{V}_x(w) = \tilde{V}_H - \frac{w_L}{w_L + j_w} \cdot \frac{R + R_e}{R_e} \cdot (\tilde{V}_o(w) - \tilde{V}_j)$$  \hspace{1cm} (C.8)

where

$$w_L \approx \frac{1}{(R_1 + R) || \left(R_b + (\beta + 1)R_e\right) C} \approx \frac{1}{(R + R_1) C}$$  \hspace{1cm} (C.9)

The crystal was connected between $v_c$ and the emitter to obtain a larger
dynamic range, since the minimum value of $v_c$ is not zero ($v_{c_{min}} \approx 0.3 + R_e v_h/R_c$). If independent gain ($R_c$) and bandwidth ($R$) adjustment was not desired, the low-pass filter could have been implemented by placing $C$ in parallel with $R_f$. From step-response measurements, the 3 db frequency of $V_o/I_d$ was found to be 3.5 kHz and with $C=0$, the 3 db frequency of $V_c(w)/V_o(w)$ was 17 kHz.

The approximate overall response of the single-channel amplifier follows from Eqs. (C.3) and (C.8) as

$$V_x(w) = \frac{w_L}{w_L + jw} \frac{R_c + R_e}{R_e} (V_H - R_f SAV_H(w))$$

(C.10)

This is only valid if

$$I_t \leq I \leq I_s$$

and

$$0.3 \text{ volts} < v_x < v_H$$

(C.11)

hold for all time. In order to obtain Eq. (C.10) $V_j$ was lumped into $v_r$ and $jwV_H = 0$ (since $V_H(w) = v_H \delta(w)$) was used.

The monitored output voltage in Fig. C.1 is

$$V_c = \frac{R_e}{R_c + R_e} V_H + \frac{R_c}{R_c + R_e} V_x$$

(C.12)

This is approximately $V_x$, since $R_c >> R_e$ usually holds. The output at $V_t$ is

$$V_t = \frac{jwR_c o}{1+jwR_c o} V_c \approx jwR_c o V_x$$

(C.13)

where the second expression on the right is valid for $w << 1/R_c o$ and $R_c >> R_e$. In the time domain:
\[ v_t \approx R_e \frac{v_o}{v_x} \]  

(C14)

C.1.2. **Real-time parameter measurement and optimization in the single-channel system.**

By simultaneously monitoring \( v_o \) and \( v_x \), the system parameters of the single-channel IPL were measured and adjusted in real-time. The relations between \( \phi_e \), \( \beta \), \( v_o \), \( v_x \), and \( \phi_m \) and their operating limits, based on the discussions in Section 5.1 and the previous section, are reviewed in Fig. C.4. A variety of additional useful parameters are also defined by their usage in this figure; for example \( v_o \) can be expressed as:

\[ v_o = v_{oo} - \frac{v_{om}}{2} \sin \phi_e \]  

(C.15a)

\( v_x \) and \( v_o \) can be simultaneously measured with a dual trace oscilloscope (e.g., Fig. 6.2) to determine \( \Delta v_x/\Delta v_o \) (= \( R_c/R_e = v_H/v_{om} \)) and \( v_{xo} \), which is the value of \( v_x \) when \( v_o = v_{oo} \) (i.e., \( \phi_e = 0 \)). If the loop is opened by disconnecting the modulator and an external aberrator is employed to vary \( \phi_e \), the cyclic variation of \( v_o \) can often be used to infer \( v_{oo} \) and \( v_{om} \) in Eq. (C.15a). Once \( v_{oo} \) and \( v_{om} \) are known, the aberrator itself can be calibrated by using

\[ \phi = \sin^{-1} \left( \frac{2(v_{oo}-v_o)}{v_{om}} \right) \]  

(C.15b)

In fact \( v_m \) of the modulator crystal can be measured by connecting the modulator crystal and opening the loop with switch \( S_2 \) in Fig. C.1. Then \( R_4 \) in Fig. C.1 can be used to vary \( v_x \); and \( \phi_m \) can be inferred from the measured values of \( v_o \) by using Eq. (C.15b).
RELATIONS:

$I = I_0 + I_1 \sin \phi_e$

$I_i$ - signal irradiance

$I_\pi$ is local oscillator irradiance

$I_0 = I_1 + I_\pi$

$I_1 = 2I_1 I_\pi$

$-V_o = -V_r + R_f SAI$

$V_o = \frac{\imath V_0 \pi \sin \phi_e}{2}$

$V_{om} = \frac{\Delta V_0}{\Delta V_i}$

$V_x = \frac{\Delta V_x}{\Delta V_o} V_0$

($\phi_m = \frac{\pi V_x}{V_o}$

$G_0 = g(I_0 - I_f)$

$V_{o\pi} = V_{om} \Delta V_0 = 2R_f SAI = 4 \Delta V_i \Delta V_i$

$V_{oo} = \frac{1}{2}(V_{om} + V_{om}) = V_o(\phi_e = 0) = V_r - R_f SAI = V_r - V_x - V_i$

$V_{xo} = V_x(V_o = V_{oo}) = V_x(\phi_e = 0)$

$\Delta V_x = \frac{R_C}{R_e} \frac{V_H}{V_{om}}$

Fig. C.4 Relations between $\phi_e$, $I$, $V_o$, $V_x$, and $\phi_m$.

(Steady-state relations for case of $V_{om} = V_r$).
(Note: \( v_\pi = \frac{\Delta \phi_{\text{in}}}{\Delta v_X} \)). These measurements provide an experimental measure of \( G_0 \) and \( G_1 \):

\[
G_1 = \frac{v_{0\pi}}{2} \frac{\Delta v_X}{\Delta v_0} \frac{\pi}{v_\pi} \quad (C.16a)
\]

\[
G_0 = v_{x_0} \frac{\pi}{v_\pi} \quad (C.16b)
\]

There are many alternative measurements, such as:

\[
G_1 = \frac{\Delta v_{0\pi}}{2} \frac{\phi_S}{v_{om}} \quad \text{and} \quad G_0 = (1 - \frac{v_{00}}{v_{om}}) \phi_S
\]

Here \( v_{om} \) can be measured as the value of \( v_O \) when \( v_X \) just equals zero (see Fig. C.4).

A variety of procedures can be used to optimize IPL performance. The basic objective is to make \( \Delta v_X/\Delta v_0 \) (i.e., \( G_1 \)) large, and \( v_{x_0} = \frac{v_H}{2} \) (i.e., \( G_0 = \frac{\phi_S}{2} \)). In practice, with the modulator disconnected, \( R_c \) is increased to make \( G_1 \) large while \( v_X \) and \( v_O \) are monitored. Then \( R_f \) is adjusted to make \( v_{00} = v_{om}/2 \); and \( R_2 \) in Fig. C.1 can be used for fine adjustment of \( v_r \) and hence \( v_{00} \). A phase aberrator is used to vary \( \phi_e \), and hence \( v_O \) and \( v_X \), while these adjustments are being made. In closed-loop operation \( v_{om}/2 \) is easily measured as the center of the tracking range of \( v_O \). In cases where \( v_{om} < v_{om} \), adjusting \( v_{00} = 0 \) results in good IPL performance.

As illustrated in Fig.C.4, the range limitation on \( v_O \) often makes direct measurement of \( v_{00} \) and \( v_{0\pi} \) difficult; but \( v_O \) and \( v_{0\pi} \) can be measured by a variety of indirect means. For example, the change in \( v_O \) due to the signal, \( \Delta v_{0i} \), and due to the L.O. irradiance, \( \Delta v_{0k} \), can be measured by alternatively blocking each beam. One then expects:
\[ v_{oo} = v_r - \Delta v_{oi} - \Delta v_{ol} \quad \text{and} \quad v_{o\pi} = 2\sqrt{\Delta v_{oi} \Delta v_{ol}} \]

Alternatively, \( R_F \) can be reduced so that \( v_{omax} \) and \( v_{omin} \) in Fig. C.4 and hence \( v_{oo} \) and \( v_{o\pi} \) can be measured. At any larger gain the new value \( v'_{omax} \) is always observable and the new values, \( v'_{o\pi} \) and \( v'_{oo} \), are

\[ v'_{o\pi} = a v_{o\pi} \quad \text{and} \quad v'_{oo} = v_r - a(v_r - v_{oo}) \quad \text{where} \quad a = \frac{v_r - v'_{omax}}{v_r - v_{omax}} \]

During IPL alignment the gain changes are small and \( v'_{oo} = v'_{omax} - v_{o\pi}/2 \) can be used to adjust \( v'_{oo} = v_{om}/2 \). Still another alternative is to frequency shift the signal beam (i.e., \( \Phi_e = \omega \)); then \( v_{oo} \) is the level crossed by \( v_o \) at equally spaced intervals.

C.2. Nineteen-channel system

C.2.1. Response of the twenty-channel HV amplifier

Figure C.5 is a schematic circuit diagram for one channel of the twenty-channel amplifier. Although simplicity was a major design criterion, two stages of op amp gain were required to amplify the weak irradiance of the expanded and aberrated beams. Based on the discussion of Fig. C.3b, the maximum possible transconductance of a single 741-type op amp (e.g., LM324, LM1458, \( A_0 R_1 \sim 10^7 \Omega \)) with the photodetector was found to be \( A_0 R_1 || R_L \sim 3 \times 10^7 \Omega \).

The frequency response of the first stage is very similar to that of the first stage of the single-channel system in Fig. C.2 (e.g., Eqs. C.1 and C.2). Based on the response of \( v_o \) to a step in irradiance the 3db frequency of this stage is about 1.2 kHz. Under most circumstances the system frequency response is dominated by the RC filter which follows this stage. Using Eqs. (C.4) and (C.5), \( \tilde{v}_c \) is thus well approximated by:
Fig. C.5. Schematic circuit diagram of one channel of the twenty-channel TV amplifier.

Notes:
1) 12 channels employ 1/2 of an LM324 Quad 741-type op amp.
2) 8 channels employ an LM458 Dual 741-type op amp.
3) \( R_0 \) controls offset gain.
4) \( R_2 \) controls bandwidth (4msec < \( 1/\tau \) < 200msec).
5) \( R_g = 4.7K \) with LM458 and \( 15K \) with LM324.
\[ V_c(w) = \frac{w_L}{w_L + j_w} (V_r - R'_f SAI(w)) \] (C.17)  
(for \(-v_1 < V_r - R'_f SAI < v_1\))

where

\[ w_L = (R_i + R)C \quad (R_i = \text{output resistance of op amp}, \ 2k\Omega) \] (C.18)

\[ \hat{V}_r = (R'_f i_ - - \frac{R'_f}{R_f} v_f) \delta(w) \] (C.19a)

and

\[ R'_f = R_f || [A_o (R_r || R_f || R_ - || R_L)] = R_f || A_o R_L = R_f || 30M\Omega \] (C.19b)

Here \( v_2(\sim 0.5mv) \) is the op amp input offset voltage, \( i_- (\sim 0.5\mu A) \) is the input offset current, \( R_ - (= 1M\Omega ) \) is the input resistance, and \( R_1 \) is the output resistance of the op amp (\( <2k\Omega \)). \( R_L \) is given by Eq. C.3b. With the fixed component values \( R_r = 3.9 \ M\Omega \) and \( R_f = 10M\Omega \); it follows that: \( R'_f = 7.5M\Omega \), \( R'_f i_- \approx 3 \text{ volts}, \) and \( v_r = 3 - 2.6 v_f \)

The noninverting amplifier in Fig. C.5 provides the required additional gain, without loading the adjustable low-pass filter (\( w_L \)). In order to discuss this stage, it is convenient to define the resistance

\[ R_A = R_1 + (1-r_2)R_2 \] and \[ R_B = r_2 R_2; \] where \( R_A + R_B = R_1 + R_2 \) and \( r_2 \) is the relative displacement of the tap on \( R_2 \) from the grounded end. With the inclusion of the op amp defects \( i_-, v_2, \) and \( R_-, v_G \) as a function of \( v_c \) is:

\[ v_G = v'_r + g_1 v_c \] (C.20)

Here
\[ v_r' = R_A i - + g_1 v_2 = g_1 v_2 \]  

(C.21a)

and

\[ g_1 = \left( \frac{1}{A_o} + \frac{R_B |R_A| |R_-|^{-1}}{R_A} \right) = \frac{R_A + R_B}{R_B} = \frac{1}{r_2} \frac{R_1 + R_2}{R_2} \]  

(C.21b)

\((R_B < R_A < R_- < 1 M\Omega)\) and \(R_A i \approx 0.03\) volts held for the chosen component values. Usually \(A_o = 10^5 >> g_1\); however the maximum possible gain for \(r_2 \rightarrow 0\) is \(g_1 = A_o\).

If the op amp is modelled as having a dominant pole (i.e., \(A_o(w) = w_T/(w_0 + jw)\)), the closed-loop bandwidth of this stage is approximately \(w_T/g_1\). Since \(g_1 < 500\) is usually employed and \(w_T = 2\pi \times 10^6\), this stage does not have a significant effect on the overall frequency response of the system.

An equivalent circuit for the frequency response, Eq. (C.22), of the high voltage vacuum tube is presented in Fig. C.6. Usually the adjustable RC filter prevents high frequencies from reaching this stage, and \(v_x\) is given by the steady-state relation of Eq. (C.23), i.e.

\[ v_x = v_{xo} - \mu v_G' \]  

(C.24)

Here \(\mu = R_p g_m\) and

\[ v_{xo} = \frac{R_p V_H}{R_p + R_H} \]  

(C.25)

This is valid for \(v_{xm} \leq v_x \leq v_H\).
Fig. C.6 Response of the high-voltage vacuum tube amplifier.

a) Equivalent Circuit

![Equivalent Circuit Diagram]

6BK4C (or 6EL4A) vacuum tube:

\[ C_g = 2.6 \times 10^{-12} \text{F}, \quad C_L = 10^{-12} \text{F}, \quad C_p = 3 \times 10^{-14} \text{F} \]

\[ 0.3 \Omega < R_p < 10 \Omega, \quad 1000 < R_p \cdot q_m < 2000, \quad R_g = 10^3 \Omega \text{ (nonlinear)} \]

LiNbO_3 Crystal:

\[ C_x = 3 \times 10^{-12} \text{F}, \quad R_x > 10^{14} \Omega \]

b) Frequency Response (Neglecting \( C_p \), which is very small; and assuming \( V_G < 0 \))

\[
\tilde{V}_x(w) = \left( \frac{w_g}{w + jw} \right) \left( \frac{w_p}{w + jw} \right) \frac{R_p \ || \ |R_x| \ |R_H| \ V_H}{R_p \ || \ |R_x| \ |R_H|} (R_p \ || \ |R_x| \ |R_H|) \ g_m V_G
\]

(C.22a)

where

\[
w_g = \frac{1}{C_g R_g} > 2\pi \times 4 \text{MHz}
\]

(C.22b)

and

\[
w_p = \frac{1}{(C_L + C_x) R_p \ || \ |R_x| \ |R_H|} > 2\pi \times 4 \text{KHz}
\]

(C.22c)

c) Steady-state (For \( R_x >> R_H >> R_p \) and \( w << w_p << w_g \))

\[
V_x = \frac{R_p}{R_p + R_H} V_H - g_m R_p V_G' \quad (V_{xm} \leq V_x \leq V_H)
\]

(C.23)

(Here \( V_{xm} \) is the minimum value of \( V_x \) reached for large positive \( V_G' \);
\( V_{xm} = 40 \text{ volts for } V_H = 10 \text{Kv, } R_H = 10^8 \Omega \))
The operating parameters of the tube were experimentally measured. Good linearity was observed for \( v_{xm} \leq v_x \leq v_H \). However \( \mu \) and \( R_p \) were found to vary from tube to tube and to be slightly dependent on \( v_H \).

For the operating conditions in Fig. C.5 \((R_H = 10^8 \Omega, V_H = 12kV)\), the parameters of the tube were measured to be: \( \mu = 1500 \), \( v_{xm} = 40 \) volts, and \( v_{xo} = 900 \) volt. Good linearity was observed in the active region \( v_{xm} \leq v_x \leq v_H \); (corresponding to approximately \(-7\) volts \(< v_G < +8 \) volts). However, \( \mu \) and \( R_p \) were found to vary from tube to tube and to be slightly dependent on \( v_H \).

Using Eqs. (C.17) - (C.24), an overall differential equation can be written for the crystal voltage in Fig. C.5:

\[
\begin{align*}
\frac{v_x}{v_L} + v_x &= \begin{cases} 
  v_{xo} - \frac{\mu}{r_2} \frac{R_1 + R_2}{R_2} v_1 & I_s < I_t \\
  v_{xo} - \frac{\mu}{r_2} \frac{R_1 + R_2}{R_2} \left( R_f i_i - \frac{R_f}{R_r} v_f - R_f S A I \right) & I_t < I < I_s \\
  v_{xo} + \frac{\mu}{r_2} \frac{R_1 + R_2}{R_2} v_1 & I_s < I_t
\end{cases} \\
\text{(C.26a)}
\end{align*}
\]

\[
I_t' = \frac{v_{ro} - \frac{R_f}{R_r} v_f - v_1}{R_f S A} \quad U(v_{ro} - \frac{R_f}{R_r} v_f - v_1) \quad \text{(C.27)}
\]

and

\[
I_s = \frac{v_{ro} + \frac{R_f}{R_r} v_f + v_1}{R_f S A} \quad \text{(C.28)}
\]

\( U(\ ) \) is the unit step function and \( v_2 (< 7.5 \text{ mV}) \) was omitted from Eqs. (C.25).
In the steady-state, the additional constraint $v_{xm} < v_x < v_H$ must also be satisfied. When $v_G$ is positive, $\mu$ becomes $\mu_+ = \mu \frac{R_G}{(R_G + R_g)}$.

C.2.2. Construction and Operational Comments

Since a dynamic range of at least $2\pi$ radians was desired, care was taken to construct the high voltage amplifier and modulator to withstand at least 13kV ($v_\pi = 6.5$ kW). A variety of high voltage cables were evaluated for leakage and breakdown by applying $\pm 13$ kV to the center conductor and measuring the leakage current back to the supply through a bare wire wound tightly around the outer insulation. Surprisingly 5 kV probe, wire outperformed thick insulated 10kV wire, double insulated 12 kV wire, and double insulated 25 kV wire. The 6BK4C triodes are rated at 26 kV, and special high voltage resistors (Vicoreen 208F, 15kV, 100MΩ, 4 watts) were employed for $R_H$.

Other basic high voltage construction practices were observed including: well cleaned insulators, wide spacing, and avoidance of sharp points; and no conductors touched the plexiglass rear panel, which could carry a water film. Although the modulator elements were potted in Krylon corona dope (breakdown 50kV/mm), arcing between the modulator elements on the crystal syrface (> .75 mm separation) was the overall limiting factor. Arcing rarely occurred at 13kV, but at 15kV an arc occurred every few seconds. With continued operation arcing became more frequent at $v_H > 13$ kV, probably due to carbonized pathways.

Since twenty channels were required, careful breadboard evaluation was carried out. A variety of integrated circuit op amps were evaluated, (including LM741, CA3140, LM3900 (QUAD), LM324 (QUAD741, 14 pin DIP) and LM1458 (DUAL741, 8 pin DIP)). The LM324 and LM1458 both performed
well, occupied comparable space, and had similar costs per op amp. Because vendor shortages were encountered for both at different times, some of each were employed.

In operation the LM1458s proved to be superior. The first stage offset voltage ($V_r$ in Eq. C.19a) was found to increase slightly (~.02 volts) in LM324 channels when $V_G$ increased above ~1 volt and the op amp was required to source >1 ma into the grid. With large second stage gain ($v_G/v_C$), $V_G$ would then increase by a volt or more. This has no immediate effect since the tube is already fully on ($\phi_x=0$); however in the closed-loop IPL, recovery from $\phi_m=0$ is now different. This deficiency was compensated by using a large grid resistor ($R_g=15k$) for the LM324, ($R_g=4.7k$ for the LM1458); this allowed $v_G/v_C$ to be increased to 500 with no noticeable offset shift. Some reduction of gain is suffered for $V_G>0$. ($G_1$ is reduced on the $\phi_m$ vs $\phi_m$ IPL characteristic for small $\phi_m$; but this is insignificant as long as $G_1$ remains large).

Knowledge of the physical layout of the circuit is invaluable for its effective use. Figure C.7a shows the layout of the amplifier board; each channel used half of a 16 pin DIP socket. The channel numbers in Figure C.7 correspond to the photodiode numbers in Fig. 5.4b. The diodes are connected by shielded ribbon cable to a 22 pin DIP socket on the board. The grid, $R_b$, and power connections also plug in, allowing complete removal of the board. The component layout of a single channel is illustrated in Fig. C.7b. Testpoints for attaching clips to monitor $V_o, V_C$, and $V_G$ are shown; and $V_G$ and $v_G'$ can also be monitored at $R_g$ which is located at the tube socket. The channel assignments for $R_o$ (offset) on the front panel are shown in Fig. C.7c. When the amplifier board is mounted in the chassis, $R$ (bandwidth) and $R_2$ (gain) are readily accessible with a screwdriver. The high voltage jacks on the
a) Twenty-channel amplifier layout. (See Fig. C.7b)

b) Typical channel layout. (Uses 1/2 of a 16 pin DIP socket)

c) Front panel. ($R_o$ offset adjustment)  
(CW = + bias on $v_o,v_1,v_g$)

d) Rear panel. (HV xtal jacks)

Fig. C.7 Operating details of the twenty-channel high-voltage amplifier.  
(Location of channels, and channel connections and controls.)
Fig. C.8 Twenty-channel high voltage amplifier
rear panel allow arbitrary reconfiguration of the modulator elements; the channel assignments in Fig. C.7d look disorganized because the primary concern was to minimize high voltage lead lengths and entanglements from the tubes. Fig. C.8 contains photographs of the twenty-channel HV amplifier.

C.2.3. Optical alignment of the nineteen element IPL.

Some comments based on hours spent in the dark will be made on the optical alignment of the laboratory IPL, which will hopefully illuminate the labors of future researchers in this area. Although the beamsplitters and lenses referred to in this discussion are those of the system in Fig. 6.12; these suggestions are applicable to other configurations.

To begin, the laser and \( M_1 \) are adjusted to place the beam at the proper height and in the center of the modulator (with \( BS_1 \) and \( BS_2 \) in place, and lenses \( L_0 \) and \( L_1 \) removed). Then \( L_0 \) and \( SF_1 \) are adjusted for a clean expanded beam; the center of which is marked by a pointer (e.g., a pencil).

Accurate beam collimation is essential to obtain a plane-wave \( L_0 \) beam and hence planar compensated wave. Lens \( L_1 \) is mounted horizontally from a rod clamp on a vertical support, in an arrangement which allows \( x \) and \( y \) tilt adjustment as well as \( x, y, z \) position adjustment. Collimation is monitored by placing a high quality, fairly thin (<1/2"), optical flat in the position of \( BS_1 \), with its normal at an angle of \( \sim 20^\circ \) to the optical axis. The front and back BS reflections can be observed on a screen to produce a shearing interferogram of the collimated beam. The center of \( L_1 \) is made coincident with the beam center pointer; and \( x \) and \( y \) tilt and \( z \) are adjusted to minimize the number (<1) of fringes in the shearing interferogram. If \( BS_1 \) is truly parallel, the interferogram will be invariant as \( BS_1 \) is shifted along the optical axis or rotates.
Next BS₃ and BS₅ are adjusted for unobstructed images of the modulator at L₃ and L₄. The modulator is centered on the beam center pointer and modulator tilt is adjusted to propagate the rear surface reflection back into the pinhole at SF₁. Aperture A₁ is adjusted to just delimit the modulator.

Components L₃ and SF₂ are adjusted for a clear image of the modulator at P₂. BS₂ is position and tilted to register the LO and MOD apertures on a face of BS₁ and to make the LO and MOD beams parallel. A sheet of thin white paper held against the beam combiner face of BS₁ aids in beam registration. If BS₁ is employed to fine-tune beam parallelism; the steps in the previous paragraph should be repeated.

A spatial filter provides an invaluable tool in adjusting beam parallelism. The beams are almost parallel when both pass through SF₂, which is a pinhole in the Fourier plane of L₃. For the following steps, BS₁ is blocked.

With L₄, L₅, L₃, and SF₂ removed, the center of the modulator image is marked on screens at P₁ and P₂. L₃ is reinserted at a distance of twice its focal-length from the modulator (unity magnification), and L₄ at one focal-length. L₃ and L₄ are positioned so that the image centers coincide with the previously marked positions in planes P₁ and P₂; which assures that the less aberrated center regions of the lenses are being employed. SF₂ is adjusted to remove all undesirable images. BS₄, M₂, and the detector are positioned for proper magnification and alignment; the detector should be rotated if necessary. The modulator image should be visible on the detector; and the tilt controls on M₂ can be used for fine registration.
Lens $L_5$ and plane $P_1$ are positioned to obtain the desired magnification of the Fourier transform. Lens $L_4$ is mounted in a manner similar to $L_1$; and $L_4$, $L_5$, and $P_1$ should be fine-tuned to optimize the transform. Aperture $A_2$ should be adjusted to partially block the undesired reflection from $BS_3$, which is a wedge. (The wedge angle should clearly separate the transforms at $P_1$).

Finally, $BS_2$ is unblocked and fine-tuned to minimize the fringes in the interferogram at $P_2$. Many of the adjustments were critical and interdependent. Starting all over and carefully readjusting the system will generally improve the optical alignment and result in a flatter interferogram at $P_2$ with more uniform irradiance. (The system can be rapidly aligned in five minutes, but more than 16 hours has been spent without achieving optimal alignment.)

The quality of lenses $L_1$ and $L_4$ is critical; better lenses (e.g., Melles Griot) should be employed. Beamsplitter $BS_2$ should be flat to $\lambda/10$ and be either very parallel ($\leq 1$ arcsec) or a wedge of greater than a few arcminutes. When $BS_2$ is a wedge, $SF_2$ removes the reflection from the undesired surface. In the intermediate almost-parallel case, both reflections pass through $SF_2$ and modulate the interferogram at the detector with troublesome, high-contrast interference fringes.

Aerotech (Model # AOM110-3) beam-splitter holders are excellent for precision beam alignment ($<1$ arcsec); but they are underdamped and tend to resonate at harmonics between $5\mu H_2$ and 300 $H_2$. Oriel (Model # 1750 and 1758) and NRC (Model # GM-2) holders are more steady, but less precise. It is usually desirable to include one and only one Aerotech holder for fine beam alignment (e.g., at $BS_2$).
C.2.4. Real-time parameter measurement and adjustment

The basic IPL parameter measurement and adjustment procedures for the nineteen-element system were very similar to those discussed in Section C.1.2. for the single-channel system. Since $v_x$ was too large to easily monitor; $v_G^i$ and $v_o$ or $v_c$ in Fig.C.5 were monitored. The modulator phase $\phi_m$ was calibrated directly in terms of $v_G^i$ by cross-wiring a modulator element to a channel with no light incident on the detector. Resistor $R_o$ in that channel varied $\phi_m$, while the resulting changes in the interferogram intensity were monitored at $v_c$ of another channel. Fig. C.9 is an example of a $\phi_m$ calibration; the average value over the modulator was $\Delta \phi_m/\Delta v_G^i \approx 0.25\pi/V$.

(Alternatively: $\frac{\Delta \phi_m}{\Delta v_G^i} = \frac{\pi}{V} \frac{\Delta v_x}{V} = \frac{\mu}{V} \frac{\pi}{V} = 0.23\pi/V$)

The intensity is generally small enough that $I_t < I_s$ (in Eqs. (C.26)-(C.28)). This simplifies parameter measurement because

$$v_o = v_{oo} + \frac{v_o\pi}{2} \sin \phi_e \quad \text{from Eq. (C.15a)}$$

(C.29)

can be directly observed. By simultaneously monitoring $v_G^i$ and $v_o$; $v_{Go}$ (value of $v_G^i$ when $v_o = v_{oo}$), $v_{Gc}$ (value of $v_G^i$ when $v_x = 0$); and $\Delta v_G^i/\Delta v_o$ can also be measured; then:

$$G_o \approx (v_{Gc} - v_{Go}) \frac{\Delta \phi_m}{\Delta v_G^i} \quad \text{(C.30a)}$$

and

$$G_1 = \frac{v_o\pi}{2} \frac{\Delta v_G^i}{\Delta v_o} \frac{\Delta \phi_m}{\Delta v_G^i} \quad \text{(C.30b)}$$
Fig. C.9 Interferometric readout of modulator phase, $\phi_m$, as grid voltage, $V'_G$, is varied.
There are other possibilities such as solving Eqs. (3.22) and (3.23) for $G_0$ and $G_1$.

Real-time system alignment is simplified by the fact that bandpass, gain, and offset are independently adjustable. First gain ($R_2$) and bandwidth ($R$) are set. Then $v_{02}$ (the value of $v_0$ for which $\phi_m=\phi_s/2$) is found by monitoring $v_0$ and $v_{01}$. Next $v_H$ is turned off to open the loop, and $\phi_1$ is varied to find $v_{00}$, which is the center of the $v_0$ range (see Eq. (C.29)). Finally $R_0$, the offset resistor on the front panel, is adjusted to obtain $v_{00}=v_{02}$ (i.e., $G_0=\phi_s/2$). ($v_{00}=0$ is sufficient when $\Delta v_0/\Delta v_0$ is large.)

Since there are 19 channels a shortcut procedure was also developed.

The interferometer is slightly misaligned so that, the detector is averaging the intensity over a few fringes, i.e., $I=I_0+I_1\sin\phi_e$=$I_0$ and $\phi_m=G_0$. With the loop closed, the current from the HV supply is monitored while the offset ($R_0$) is adjusted in each channel. The current is adjusted to the center of its variation range ($\Delta i=v_H/R_H=.15\text{ma}$) for that channel. Finally the interferometer is realigned. (This procedure has difficulties when system defects such as nonparallel beam-splitters cause $I_0(x,y)$ to change with the interferometer alignment.)

A less optimal, but simpler shortcut alignment procedure is to adjust the offset while the optically aligned IPL is operating. As the offset is varied in a given channel, a range of $R_0$ dial positions will be found for which the HV supply current varies. Over most of that range, the current remains at an intermediate value corresponding to $\phi_e=0$, (Or $v_x=v_0\phi_1'/\pi$). The objective is to move $R_0$ into the center of its active range for that channel. The major difficulty with this technique is that wavefront fluctuations in $\phi_1'$ across other elements will also cause the
HV supply current to vary during the adjustment process. Another variant of the procedure is to view the actual interferogram (plane $P_2$ in Figs. 6.10 and 6.12) instead of supply current. Three states should be apparent; $\phi_m=0$, $\phi_m=\phi_s$, and $\phi_m=\phi_1$.

When multiple $\phi_m$ equilibria exist, the operation of the IPL is influenced by the initial value of $\phi_m$ before the loop is closed, $\phi_{m0}$. The easiest initial value of $\phi_m$ to implement is $\phi_{m0}=0$, which occurs when the IPL is started by turning $V_H$ on, from its off condition of zero volts. Superior operation is obtained with the initial condition $\phi_{m0}=\phi_s/2$; then with a dynamic range $\geq 2\pi$ radians, all elements should lock in their active region, $0<\phi_m<\phi_s$. Arbitrary reinitialization to $\phi_{m0}$ can be implemented by illuminating the detector with a saturating light source ($I>I_s$ results in $v_x=v_H$), and setting $v_H$ to $\phi_{m0}v_\pi/\pi$. A suitable source is a flashlight bulb near the detector. The system is restarted by turning off the source and then bringing $v_H$ back to $\phi_s v_\pi/\pi$. According to Eqs. (C.24) and (C.25) with $R_p<<R_H$, adjusting $v_H$ does not perturb $v_X$ much.
D. BEHAVIOR OF THE STREHL IN ADAPTIVE OPTICAL SYSTEMS

In the adaptive optical system of Fig. 1.1, there is a diffraction limited pinhole detector, D, in the Fourier transform plane of lens $L_1$. This detector examines the zero-order of the Fourier transform of the compensated wavefront. The Strehl ratio $^{50,125}$ is the ratio of: the intensity seen by this detector to the intensity which would be seen with perfect compensation, i.e., a totally uniform wavefront. As such, the Strehl ratio is a commonly used measure of the performance of adaptive optical systems, and is discussed to some extent in the literature $^{30,31,46,50,122}$. More practically, the Strehl ratio describes the intensity seen by the compensated direct detection communications receiver in Fig. 1.1. As discussed in Section 3.2.1 of the main text, the Strehl ratio also assumes a major role in the performance of the Zernike phase-contrast interference phase loop. The properties of the Strehl ratio discussed here will be used to motivate discussions in the main text.

With the outgoing wavefront $|E_i(x,y,t)|e^{j\theta(x,y,t)}e^{-jwt}$ at point A in Fig. 1.1, the intensity at the detector, D, is

$$I_d = c_1 \left| \mathcal{F}_{w=0}(|E_i|e^{j\theta}) \right|^2 \quad (D.1)$$

where $c_1$ is a real constant less than unity. (The optical carrier $e^{-jwt}$ will be suppressed in the following discussion.) The zero-order of the Fourier transform is essentially the spatial average

$$\mathcal{F}(\cdot) = A_m \int_{A_m} \int_{A_m} |E_i| e^{j\theta} dx dy = A_m <|E_i|e^{j\theta}> \quad (D.2)$$
The brackets here denote spatial averaging and $A_m$ is the area of the exit aperture of the modulator. This complex field can be represented as a new phasor

$$ce^{j\theta}c = |E_i|e^{j\theta}$$

(D.3)

The basic properties of $c$ and $\phi_c$ are discussed in Appendix A.1. (Eqs. (A.1)-(A.4) are valid for spatial averages.) It is also useful to define

$$c' = \frac{c}{|E_i|} < 1$$

(D.4)

The inequality on the right follows from Eq. (A.2a). The detector intensity can now be written:

$$I_d = c_1A_m c'^2 = c_1A_m|E_i|^2 c'^2$$

(D.5)

With perfect phase compensation $\theta$ is spatially uniform and $c'=1$.

Since the adaptive systems of concern in this thesis compensate only phase, this discussion will stress the modified Strehl ratio, $Q$, in which the phase of the perfectly compensated wave is flat, but the amplitude can be spatially variant, i.e.

$$Q = \frac{I_d}{c_1A_m|E_i|^2} = c'^2$$

(D.6)

If the performance of each spatial resolution element of the adaptive system is known, Eqs. (D.6) and (A.2a) allow $Q$ to be calculated. With discrete spatial resolution elements, the averages become summations. Alternatively, various assumptions can be introduced which allow general observations to be made on the behavior of $Q$. The latter route will be taken.
In an adaptive optical system it is useful to write $\theta$ as

$$\theta = \phi_i(x,y,t) - \phi_m(x,y,t) - \phi_o(x,y) \quad (D.7)$$

where $\phi_i$ is the incident corrupted phase, $\phi_m$ is the compensating modulator phase and $\phi_o$ is a static aberration introduced by the adaptive system.

The following assumptions will be made: (1) $\phi_i$ and $|E_i|$ are statistically independent; (2) $\phi_i$ is spatially ergodic, with equal ensemble average (denoted by overbar) and spatial average (brackets); (3) $\phi_i$ is characterized by a probability distribution, $P(\phi_i)$, which is symmetric about the mean value of $\phi_i$, such as Gaussian; and (4) $\langle |E_i| \rangle$ and $\phi_i$ are deterministic variables in the time domain. These conditions could be satisfied by a static aberrator, such as "shower glass", or possibly a foggy or low visibility atmosphere frozen in time. This model is also somewhat applicable to a turbulent atmosphere with cells of differing phase shift drifting through the signal beam. The fourth condition makes the stochastic dynamic behavior of the adaptive system a separate problem and allows use of its deterministic dynamic equations (e.g., Eq. (2.12) and (2.13) of the IPL). In conjunction with this constraint, it is also assumed that the temporal variations in $\phi_i$ are slow enough to be tracked by the adaptive system.

**Uncompensated Strehl ratio**

With the adaptive system turned off, it is assumed that $\phi_m$ is spatially uniform. It will also be assumed that $\phi_o$ is spatially uniform. (An alternative is to assume that $\phi_o$ is ergodic and include it in $\phi_i$).

Employing the additional assumptions stated above, in conjunction
with the results of Appendix A.1, allows the simplification

$$
\langle |E_i| e^{3\theta} \rangle = \langle |E_i| \rangle c' e^{j(\phi_i - \phi_m - \phi_0)}
$$

where

$$
c' = \frac{\mathcal{C}_i}{w = 1 P(\phi_i)}
$$

is real. When $\phi_i$ is a Gaussian distributed random variable with mean $\bar{\phi}_i$ and variance $\sigma_i^2 = \phi_i^2 - \bar{\phi}_i^2$, Appendix A.1 shows

$$
c' = e^{-\frac{\sigma_i^2}{2}}
$$

From Eq. (D.6), the uncompensated Strehl ratio is thus

$$
Q = e^{-\sigma_i^2}
$$

and from Eq. (D.5)

$$
I_d = c_1 A_m |E_i|^2 e^{-\sigma_i^2}
$$

This result, Eq. (D.12), explicitly shows that, subject to the stated assumptions, the detector intensity degradation is almost wholly due to the rms phase fluctuations, $\sigma_i$. Conversely, any attempt to compensate $|E_i|$ with an attenuating amplitude modulator usually reduces $\langle |E_i| \rangle$ and make matters worse! Since $\sigma_i^2$ was not constrained to be cyclic, this result should be applicable to both strong fluctuations $\sigma_i^2(2\pi)^2$, and weak fluctuations, $\sigma_i^2<(2\pi)^2$.

Strehl ratio with imperfect phase compensation

When the adaptive system is operating, since $\phi_m$ is trying to conjugate $\phi_i$, $\phi_m$ will be a function of $\phi_i$. Given the probability distribution of $\phi_i$, that of $\theta = \phi_i - \phi_m - \phi_0$ can be determined from elementary
probability theory as

\[ P_\theta(\phi) = P(\phi_i) \left| \frac{\phi_0}{\phi_i} \right| \]  

(D.13)

The compensated strehl ratio follows from Eqs. (D.6) and (D.9):

\[ Q = \left[ \sum_{\omega=1}^{N} (P_\theta(\phi)) \right]^2 \]  

(D.14)

The specific example of the homodyne IPL with a type-zero loop filter will be developed in more detail. Since the IPL is usually operated with enough gain to keep the error phase, \( \phi_e \), small, the linear approximation of Eq. (2.24) can be used for \( \phi_m \). Also employing Eqs. (2.3) and (2.4),

\[ \bar{\phi} = \phi_i - \phi_m - \phi_0 = \frac{\phi_i}{1+G_1} - \frac{G_o - 2n\pi}{1+G_1} \frac{G_0 - 2n\pi}{G_1} - \frac{\phi_0}{G_1} - \frac{\phi_0 - \phi_2}{2} - 2n\pi \]  

(D.15)

Given \( G_o \) and \( G_1 \), and using Eq. (D.13)

\[ P_\theta(\phi|G_o,G_1) = \frac{P(\phi_i)}{1+G_1} \]  

(D.16)

If \( G_1, G_o, \phi_o, \phi_1, \) and \( \phi_2 \) are all approximately spatially uniform, and \( P(\phi_i) \) is symmetric about \( \phi_i \); it is easy to show that \( P(\phi) \) is also symmetric about the mean, \( \bar{\phi} = \phi_i - \phi_m - \phi_0 \). The final \( 2n\pi \) in Eq. (D.15), due to the \( 2n\pi \) multistability of \( \phi_m \), can be neglected since \( \theta \) only appears as \( \cos \theta \) or \( \sin \theta \). The \( 2n\pi \) in the brackets can also be neglected if it is the same in each channel due to the initial condition of \( \phi_m \), or if \( 2\pi(n-\bar{n})\ll G_1 \) holds.
In particular if \( P(\phi) \) is Gaussian, then \( P_\theta \) is also Gaussian with variance:

\[
\sigma^2 = (\langle \phi_i - \phi_m \rangle - (\phi_i - \phi_m))^2 = \frac{\sigma_i^2}{(1+G)^2} \tag{D.17}
\]

The corresponding Strehl ratio with imperfect interference phase loop compensation is:

\[
Q = \exp(-\frac{\sigma_i^2}{(1+G)^2}) \tag{D.18}
\]

The compensated intensity seen by the direct detection communications in Fig.1.1 is

\[
I_d = c_i A_0 \langle |E_i| \rangle^2 Q \tag{D.19}
\]

Under these assumptions, the homodyne IPL with a type-zero loop filter performs essentially perfect phase compensation (\( Q \approx 1 \)), as long as the gain is much larger than the RMS phase variation, i.e., \( G \gg \sigma_i \).

**Strehl ratio in atmospheric turbulence**

Fried\(^{37,97,248}\) has shown that the spatial phase fluctuations due to propagation through atmospheric turbulence are approximately Gaussian distributed, and thus Eq. (D.11) is a good approximation for the uncompensated Strehl ratio. The corresponding phase variance is given by\(^{30}\)

\[
\sigma_i^2 = 1.031 (r_R/\lambda)^{5/3} \tag{D.20}
\]

Here \( r_R \) is the receiver aperture radius and \( \lambda \) is the wavelength in \( \text{nm} \), propagation path \( L \) in meters, and atmospheric structure constant
\( C_n^2 \) in meters\(^{-1/3} \), the coherence radius is \(^9^5\)

\[
C_o = 6 \times 10^{-9} \left( \frac{n^2}{C_n^2} \right)^{3/5} \quad \text{(in meters)} \quad (D.21)
\]

\( C_n^2 = n^2 / L_0^{2/3} \), where \( n^2 = \bar{n}^2 - \bar{n}^2 \) is the rms refractive index fluctuation and \( L_0 \) is the largest scale size of the turbulence \(^9^7,^1^0^6\). Equation (D.20) shows that in atmospheric turbulence the phase variance \( \sigma_i^2 \) increases with the aperture size.

The compensated Strehl ratio result of Eq. (D.18) assumed that the system resolved the coherence cells. When the radius, \( r_o \), of the smallest system resolvable spot size is greater than the coherence cell radius, even with perfect phase estimation and compensation there is a residual phase variance of

\[
\sigma^2 = 1.031 \left( r_o / r_o \right)^{5/3} = \sigma_i^2 \left( r_R / r_o \right)^{-5/3} = \sigma_i^2 N^{-5/6} \quad (D.22)
\]

Here \( N = (r_R / r_o)^2 \) is the effective number of adaptively corrected subapertures in the receiver aperture.
half-core FOV $\theta_R$ from a transmitted pulse of duration $T$, power $P_T$, and initial beam divergence $\theta_T$ (Gaussian half-cone angle) is approximately

$$n_s = n_d(L) + n_w(L) \quad (E.1a)$$

$$= \left[ \frac{n_d^0 T}{2h\nu} \right] \left[ \frac{r_R^2}{\theta_T^2} \right] \left( e^{-k_e' L} + \frac{e^{-k_e' L}}{(1 + \frac{\theta_T^2}{\theta_R^2})(1 + \frac{\theta_T^2}{\theta_R^2})} \right) \quad (E.1b)$$

Here $n$ is the quantum efficiency, $h\nu$ is the photon energy, $k_e = k_s'$, $k_e' = (1-\alpha)k_s$ and $\theta_o(L) = \lambda/(2\pi D_o)$ is the rms beam divergence (half-angle) of the scattered light's angular spectrum. The term in brackets is the average photon count for free-space propagation. The first term, $n_d$, is the average number of photons collected from the direct un-scattered beam (essentially a plane wave), and $n_w$ is the average number of scattered photons collected by a receiver with FOV $\theta_R$. (The FOV of a direct-detection receiver is controlled by the radius $r_d$ of the Fourier-plane detector, i.e., $\theta_R \approx r_d/f$, where $f$ is the focal length of the lens. When diffraction-limited, $\theta_R = 0.61\lambda/r_R$). The coherence length in the receiver input aperture (coherence cell radius) is approximately

$$\rho_o = \frac{\lambda}{2\pi \theta_F} \sqrt{\frac{3}{2\alpha k_s L}} \quad (E.2)$$

Typical values are $\theta_F = 11.3\text{mr}$, $\alpha = 0.57$, and $\rho_o = 6.1\mu\text{m}$ (for $k_s L = 4$ and $\lambda = 0.53\mu\text{m}$).
E. THE MULTIPLE FORWARD SCATTER MODEL FOR LOW-VISIBILITY OPTICAL COMMUNICATIONS

This Appendix briefly reviews some of the important features of applying the multiple forward scatter (MFS) propagation model to a low visibility communications link. These issues are discussed in greater detail in References 8, 89, 92, and 93. This discussion assumes familiarity with Section 1.4 of the main text.

A key assumption of the MFS model is that the single-scatter phase function, $P(\theta)$, can be divided into two components. These are: a strong forward-peaked component with rms scattering angle $\theta_f$ and scattering coefficient $\alpha_k_s$, and a diffuse broad and backscattering component with scattering coefficient $(1-\alpha)k_s$. \(\alpha_k_s = 2\pi \int_0^{\theta'} k_s \sin \theta \, d\theta\), where $\theta'$ is an angle beyond which $d(\alpha_k_s)/d\theta'$ becomes very small, and $\theta_f$ is approximately $\theta_f = (2\pi P(\theta))^{-1/2}$. Typically, $\alpha = 0.57$. Most atmospheric aerosols are larger than $\lambda$, and hence possess scattering phase functions consistent with this model.\(^88\text{-}^90\)

At visible wavelengths the absorption coefficient, $k_a$, is usually small\(^88\) and most of the forward scattered light eventually arrives at the receiver. The diffusely scattered light, however, is assumed to be lost. The effective extinction coefficient seen by a wide field-of-view (FOV) receiver, which collects the scattered light, is thus $k_e' = (1-\alpha)k_s$. The direct line-of-sight beam seen by a narrow-FOV receiver is attenuated by the full scattering coefficient, i.e., \(k_e = k_s\).

More specifically, the average number of signal photons $n_s$ collected by a direct-detection receiver of aperture radius $r_R$ and
Over path lengths longer than a few times $k_{e}^{-1}$, a wide-FOV receiver, with $\theta_R = \theta_0$, generally sees only the $n_w$ component, because $n_d$ is greatly attenuated. A diffraction-limited receiver over the same path, however, sees only the much weaker $n_d$ component, since Eq. D.1b shows that $n_w$ is attenuated by a factor of $(\theta_R^\theta_0)^2 = (\theta_0^\theta_R)^2 = 1/M$. Note that $M$ is the number of coherence cells in the input aperture ($\approx 2 \times 10^7$ for $r_R^1$ inch). Wide-FOV operation thus increases the signal by a factor of about $e^{\alpha k_0}$ over diffraction-limited operation. (This assumes $\theta_T = \theta_0$; actually $\theta_T < \theta_0$ is more desirable). Over longer pathlengths $n_d$ may be attenuated so much that $n_w$ dominates even the diffraction-limited signal. The wide FOV improvement then follows from Eq. (E.1b) as approximately $(\theta_0/\theta_R)^2 = M$.

The average number of background photons collected by a direct detection receiver is approximately $1.886 n_{b} = (\pi r_R^2) \Omega N_{\lambda} \Delta \lambda$, where $\Delta \lambda$ is the spectral bandwidth of the optical filter which precedes the receiver, $\Omega$ is the solid angle of the receiver FOV ($\Omega = \pi \theta_R^2$), and $N_{\lambda}$ is the background radiance due to reflected and scattered sunlight and other sources. For diffraction-limited operation $n_{b} = n_{b_{d}} = \lambda^2 N_{\lambda} \Delta \lambda$; and with a wide FOV of $\theta_R = \theta_0$, $n_{b} = n_{b_{w}} = M n_{b_{d}}$. A diffraction-limited FOV thus attenuates background light by a factor of $1/M$ relative to a wide-FOV direct detection receiver.

Adaptive phase compensation flattens out the wide-FOV scattered wavefront so that it can be collected by a diffraction-limited receiver, thus offering both wide-FOV scattered signal reception and diffraction-limited background rejection. In order for it to be
both desirable and feasible to employ the additional complexity of adaptive phase compensation, as opposed to simple wide-FOV direct-detection (i.e., a "photon bucket"), a variety of constraints must be met. For desirability: 1) the wide-FOV signal must exceed the diffraction-limited signal, (i.e., \( n_w(L_1) > n_d(L_1) \)). 2) The diffraction-limited signal must be insufficient for the desired performance (path-length, bandwidth, error rate). Performance can be specified by demanding that the signal-to-noise ratio (SNR) exceed a threshold, \( \text{SNR}_t \). For shot-noise limited performance, this condition becomes: \( \text{SNR}_d = n^2_d / (n_{bd} + n_d^2) = n_d(L_1) < \text{SNR}_t \). 3) The uncompensated wide-FOV "photon bucket" receiver must be background-limited, i.e., \( n_w(L_1) < n_{bw}(L_1) \). (More rigorously: \( \text{SNR}_w = n_w^2 / n_{bw} < \text{SNR}_t \)). In many instances, \( L_1 \) in these constraints can be interpreted as the minimum link length over which scattering is severe enough to justify adaptive phase compensation. Feasibility considerations generally determine the maximum possible link length, \( L_2 \); for successful adaptive communications: 4) The compensated wide-FOV signal must be sufficient for communications, (i.e., \( \text{SNR}_w(n_w(L_2)) > \text{SNR}_t \)). 5) There must be enough signal to drive the phase compensation system, (i.e., \( \text{SNR}_c(L_2) = n_c^2 / (n_c + n_{bc}) > \text{SNR}_{tc} \)). Here \( n_c \) and \( n_{bc} \) are the average number of signal and background photons respectively seen by the phase compensation system. Adaptive phase compensation is not worthwhile in situations where \( L_2 < L_1 \).

In many cases, the dominant desirability constraint is (3) and the dominant feasibility constraint is (5). Using Eq.(E.1b).
condition (3) can be rewritten as

$$\frac{n_w}{n_{bw}} = \frac{I_s}{N_\lambda \Delta \lambda} < 1$$  \hspace{1cm} (E.3a)

where

$$I_s = \frac{9P_T e^{-k'e'L}}{32\pi^2 k_f^2 \theta_f^4 L}$$  \hspace{1cm} (E.3b)

The compensator drive signal can also be expressed in terms of $I_s$:

$$n_c = \frac{n_d}{\hbar \nu} I_s T_c \lambda^2$$  \hspace{1cm} (E.4)

Here $T_c$ is the phase distortion coherence time, and $d$ is the product of the signaling duty cycle and fraction of signal diverted to the compensator, (an optimistic $d=1$ will be assumed). In the range of useful link lengths, $L_1 < L < L_2$, $I_s$ can be eliminated between Eqs. (E.3a) and (E.4) resulting in

$$n_c < \frac{n_N \lambda^2 \Delta \lambda T_c}{(\hbar \nu)}$$  \hspace{1cm} (E.5)

Section 3.1.7 and Ref. 8 shows that the right side of this Equation is the average background photon count, $n_{bc}$, of a self-interference phase compensation system.

Demanding that a simple wide-FOV direct-detection be background limited (condition (3)) in order to justify phase compensation thus implies that the phase compensator is also background limited, i.e., $n_c < n_{bc}$. This rather severe constraint does not preclude phase compensation, but simultaneous satisfaction of condition (5) demands
\[ n_{bc} > n_c \cdot \frac{\beta}{2} + \sqrt{\left( \frac{\beta}{2} \right)^2 + n_{bc}} \quad \text{(here } \beta = \text{SNR}_{ct} = 4). \]

The existence of a useful range \((L_1 < L < L_2)\) for phase-compensated communications thus demands

\[ n_{bc} = \frac{n}{\hbar \nu} \lambda^2 N_\lambda \Delta \lambda \tau_c > \beta + 1 = 5 \quad \text{(E.6)} \]

This important result reveals that low-visibility communications with phase compensation for background suppression is only viable when the background radiance is very large, narrow-band spectral filters are unavailable, and/or the scattering channel varies only slightly with time.
Many nonlinear optical processes produce a wave proportional to $E^*_i E_r$, where $E_r$ is a pump or reference wave. This has an important implication: if $E_r = E_i$, the phase-compensated wave $E_i E^*_i$ is produced. For example, Fig. F.1 illustrates high-resolution adaptive phase compensation with a single SLM. Note that the output waves $E_a$ and $E_c$ are phase compensated. Figure F.2 is a closed-loop variation, where the compensated wave is used to create the hologram. That configuration, which is mentioned in the main text, is unworkable; because when initially turned on, there is no $E_c$ wave to create the hologram which will diffract $E_c$.

An even simpler phase-compensation geometry can be obtained by simultaneously writing and reading out a real-time hologram in a photorefractive material \( \text{e.g. Bi}_{12}\text{SiO}_{20} \), as sketched in Fig. F.3. There $E_i$ and $E_r$ are polarized at 45° relative to each other; and an analyzer is employed to block $E_r$, since the compensated output is in the same direction as $E_r$.

In Fig. F.4, a backward-propagating (reflected) version of $E_i$ is employed for readout. (The readout wave does not wash out the hologram because it creates a grating with planes approximately parallel to the xy plane. However, due to the bias voltage $V_b$ along x, the crystal can only resolve grating planes approximately parallel to the yz plane.

A major disadvantage of the real-time holographic approach in Figs. F.1 - F.4 is the large amount of signal energy lost, and resulting attenuation of the compensated beam, due to diffraction efficiencies.
AMPLITUDE MODULATION OF:
\[ t_M = |E_i|^2 + |E_r|^2 + E_i^* E_r e^{-j(k_i - k_r) \cdot \vec{r}} \cdot E_i^* E_r e^{-j(k_r - k_i) \cdot \vec{r}} \]

Fig. F.1 Real-time holographic adaptive phase compensation with an SLM.

There are four angularly separated output waves. Two waves \((E_a\) and \(E_b)\) are reflected from mirror \(m_0\) after being diffracted from \(E_i^*\) on its way into the modulator. The other two \((E_c \text{ and } E_d)\) are diffracted from the reflected on-axis wave on its way out of the modulator. Two of the four outputs are phase compensated:

\[ E_a = \alpha r |F_1|^2 E_r e^{-j(k_r - 2k_i) \cdot \vec{r}} \]
\[ E_c = \alpha' r (|E_i|^2 + |E_r|^2) |E_i|^2 E_r e^{-jk_r \cdot \vec{r}} \]

Here \(\alpha\) and \(\alpha'\) are real constants less than unity (attenuation), and \(r=1\) is the reflectivity of the mirror on the back of the modulator.

Fig. F.2 Closed-loop real-time holographic phase compensation.
(Does not work.)
Fig. F.3 Real-time holographic adaptive phase compensation with a photorefractive material. (P = linear polarizer. Six waves are reconstructed in four angular directions.)

\[ E'_i = E_i e^{-j\mathbf{k}_i \cdot \mathbf{r}}, \quad E'_r = E_r e^{-j\mathbf{k}_r \cdot \mathbf{r}}, \text{ and} \]

\[ E'_c = \alpha(|E_i|^2 + |E_r|^2(1+\alpha'))|E_i|^2 E_r e^{-j(\mathbf{k}_r - 2\mathbf{k}_i) \cdot \mathbf{r}} \]

Fig. F.4 Real-time holographic phase compensation with a photorefractive material (\(\alpha\) and \(\alpha'\) are diffraction efficiencies).

Fig. F.5 Phase compensation by four-wave mixing.
much less than unity. One potential solution to the difficulty may be to employ the four-wave mixing geometry of Fig. F.5. With a suitable choice of nonlinear mixing material, it may be possible to produce an amplified compensated $E_c$ wave.

The interference phase loop is usually more efficient in its use of signal light than the systems of Figs. F.1 - F.4, since the IPL modulator often reflects (or transmits) $\sim$100% of its input light, (efficiency can be further enhanced by polarizing beamsplitter techniques, as in Fig. 2.2c). The compensation configurations in Figs. F.1 - F.5 are generally limited to monochromatic light, coherent with a local reference or pump beam; whereas the self-interference implementations of the IPL can function in polychromatic light. It should also be noted that the IPL is less demanding of SLM resolution than the holographic approach of Fig. F.1, because there is usually no tilted-reference-beam spatial carrier in the IPL. The nonlinear mixing processes of Figs. F.3 - F.5, however, generally have no difficulty achieving the required resolutions. They are also capable of fast response times.

Although nonlinear processes have been employed in many phase conjugation applications, (see appendix G.16), the real-time nonlinear phase-compensation techniques mentioned here do not seem to have been previously reported. These techniques, and in particular four-wave mixing, are worthy of further study. Important issues, which need to be addressed to determine their feasibility, include: satisfaction of phase matching or Bragg conditions, availability of suitable materials, effects of other material properties (e.g. optical activity in $\text{Bi}_{12}\text{SiO}_{20}$), maximum possible diffraction efficiencies and/or gain, sensitivity to low-level signals, and noise considerations.
G. BIBLIOGRAPHIC COMMENTS ON PHASE MEASUREMENT AND RELATED TOPICS

This appendix contains bibliographic material pertaining to a wide variety of phase measurement techniques. These references are not all inclusive, but should provide a good starting point for the reader interested in pursuing a given topic in greater detail.

(The entries under each subheading are in chronological order)

G.1 Basic two-beam homodyne interferometry:


Malacara, D., Editor, Optical Shop Testing, New York: John Wiley and Sons, 1978. (Also applicable to Sections G.4 to G.10)

G.2 Two-beam heterodyne interferometers enhanced by electronic analogue and/or digital post-processing of the interferogram:


The following are "scanning interferometers," which can be viewed as a homodyne interferometer with a slowly changing phase in one arm or a heterodyne interferometer with a small frequency offset:


G.3 Two-beam homodyne interferometers (e.g. a hologram) with optical and electronic post processing.


G.4 Phase visualization techniques which pass the spatial Fourier transform of the input waveform (whose phase is to be measured) through a specialized spatial frequency-plane filter. The source distribution, or degree of partial coherence of the input waveform, is often constrained.

G.4.1 Variants of Zernike phase-contrast and Schlieren techniques, including systems with electronic post processing:


G.4.2 General frequency-plane filtering techniques for visualization of $\phi$ and/or $d\phi/dt$:


Frequency-plane filtering techniques which visualize the spatial derivative, \( d\phi/dx \):


G.5 Lateral-shearing interferometers implemented with gratings.


G.6 Radial-shearing interferometers.


G.7 Polarization interferometers (lateral-shearing: Total doubling, differential, compensated; radial-shearing).


Dyson, J. op. cit., 1957.


G.8 Scatterplate interferometers.


G.9 Variable sensitivity interferometers which modify the dynamic range of the interferogram so that a full cycle (i.e. a fringe) corresponds to an optical path length of either a fraction of a wavelength or many wavelengths (as opposed to one fringe per wavelength in an elementary interferometer).


G.10 Systems employing noninterferometric phase-slope (dφ/dx) visualization techniques which are useful with incoherent white light. (See also Sec. G.4.3.)

Hartmann sensor:


Moire effect:

G.11 Determination of phase from spatial intensity (and/or Fourier transform intensity) measurements, using the analytic functional and propagational Fourier transform properties of optical fields. This is known as "the phase problem" of optics.


G.12 Numeric algorithms (mostly iterative) for computing the phase of a wavefront given its intensity and the intensity or other information concerning its spatial Fourier transform.


Dallas, W. J. "Digital computation of image complex amplitude from image and diffraction intensity: An alternative to holo-


G.13 Optimum phase estimators which minimize the mean-squared error of the estimate.


G.14 Closed-loop systems in which a function of the phase estimate is fed back to modify the phase estimation system. Many of these are adaptive optical systems, which only implicitly estimate phase.

G.14.1 Two-beam interferometric systems which modulate the wavefront being measured with a function of the phase estimate.


Dyson, F. J., 1975, op. cit.

Dicke, R. H., 1975, op. cit.


G.14.2 Two-beam interferometric systems which modulate the reference beam with a function of the phase estimate.


Johnson, G. W., et. al., 1979, op. cit.

G.14.3 Frequency multidither adaptive systems.

O'Meara, T. R.  "the multidither principle in adaptive optics,"  

Angelbeck, et. al. "Multidither adaptive optics system operation 
using a separate dither and corrector mirror,"  Proc. of 

G.14.4 "Image sharpening" adaptive systems.

Muller, R. A. and A. Buffington.  "Real-time correction of atmos-
pherically degraded telescope images through image sharpen-

stellar image using a real-time phase correction system: 

G.15 Variants of holographic interferometry.

Dandliker, R. E., and F. M. Mottier.  "Two-reference-beam holographic 

Marom, E. and J. Katz.  "Unconventional interferometric realizations 
based on holographic nonlinear effects,"  Appl. Opt. 16, 5, 


G.16 Nonlinear optical processes

G.16.1 Self-reference interferometry by conjugate wavefront generation.

Agrawal, G. P.  "Phase determination by conjugate wavefront genera-


G.16.2 Conjugate wave generation by stimulated Brillouin scattering.

Zel'dovich, B. Ya., et. al.  "Connection between the wavefronts of 
the reflected and excited light in stimulated Mandel'shtam-

Nosach, O. Yu., et. al.  "Cancellation of phase distortions in an 
amplifying medium with a Brillouin mirror,"  Sov. Phys. JETP. 

G.16.3 Conjugate wavefront generation by stimulated Raman scattering.


G.16.4 Conjugate wavefront generation by photon echos.


G.16.5 Applications of three-wave mixing (optical parametric amplification).


G.16.6 Application of degenerate four-wave mixing.


G.16.7 Real-time holography in photorefractive materials


G.16.8 Review papers on application of nonlinear optical processes.


REFERENCES


44. SPIE Proc. on "Imaging through the atmosphere," 75, 1976.


57. SPIE Proc. on Active optical devices and applications, 228, 1980.


205. __________, Preliminary data sheet, PLZT shutter/modulator, Motorola Opto-Ceramic Products, 3434 Vassar St., Albuquerque, N.M.


213. Applications for Microchannel Plates, Varian Assoc., Palo Alto, California.


225. Correspondence with Epoxy Technology Inc. P.O. Box 567, 14 Fortune Dr., Billerica, Mass.


252. Optical Crystals, Lithium Niobate and Lithium Tantalate, Mountain View California: Crystal technology Inc.