A SEMI-EXPERIMENTAL NODAL SYNTHESIS METHOD FOR
THE ON-LINE RECONSTRUCTION OF THREE-DIMENSIONAL
NEUTRON FLUX-SHAPE S AND REACTIVITY

by

Robert P. Jacqmin

Ingénieur de l’Ecole Nationale Supérieure d’Ingénieurs Electriciens
de Grenoble, Grenoble, France (1985)

Ingénieur en Génie Atomique, Grenoble, France (1986)

Submitted to the Department of Nuclear Engineering
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1991

© Massachusetts Institute of Technology 1991
All Rights Reserved

Signature of Author

Department of Nuclear Engineering
August 30, 1991

Certified by
Allan F. Henry
Professor of Nuclear Engineering
Thesis Supervisor

Certified by
David D. Lanning
Professor of Nuclear Engineering
Thesis Reader

Accepted by
Allan F. Henry
Chairman, Department Committee on Graduate Students
A SEMI-EXPERIMENTAL NODAL SYNTHESIS METHOD FOR 
THE ON-LINE RECONSTRUCTION OF THREE-DIMENSIONAL 
NEUTRON FLUX-SHAPES AND REACTIVITY

by

Robert P. Jacqmin

Submitted to the Department of Nuclear Engineering on August 30, 1991 
in partial fulfillment of the requirements for the degree of Doctor of 
Philosophy in Nuclear Engineering.

ABSTRACT

The safety and optimal performance of large, commercial, light-
water reactors require the knowledge at all time of the neutron-flux 
distribution in the core. In principle, this information can be obtained by 
solving the time-dependent neutron diffusion equations. However, this 
approach is complicated and very expensive. Sufficiently accurate, real-
time calculations (time scale of approximately one second) are not yet 
possible on desktop computers, even with fast-running, nodal kinetics 
codes.

A semi-experimental, nodal synthesis method which avoids the 
solution of the time-dependent, neutron diffusion equations is described. 
The essential idea of this method is to approximate instantaneous nodal 
group-fluxes by a linear combination of K, precomputed, three-
dimensional, static expansion-functions. The time-dependent coefficients 
of the combination are found from the requirement that the reconstructed 
flux-distribution agree in a least-squares sense with the readings of J (\geq K) 
fixed, prompt-responding neutron-detectors. Possible numerical 
difficulties with the least-squares solution of the ill-conditioned, J-by-K 
system of equations are brought under complete control by the use of a 
singular-value-decomposition technique. This procedure amounts to the 
rearrangement of the original, linear combination of K expansion functions 
into an equivalent, more convenient, linear combination of R (\leq K) 
orthogonalized "modes" of decreasing magnitude. Exceedingly small 
modes are zeroed to eliminate any risk of roundoff-error amplification, 
and to assure consistency with the limited accuracy of the data. Additional 
modes are zeroed when it is desirable to limit the sensitivity of the results 
to measurement noise.
Numerical tests of this fitted-synthesis method demonstrate that best results are obtained when the synthesis is allowed to be discontinuous in time. In this discontinuous mode of operation, only a small subset of the library of expansion functions is retained at any instant in the synthesis formula. These expansion functions are flux shapes obtained for reactor conditions (control-rod positions, power level, ...) closely bracketing the actual, instantaneous state of the core.

A model of a commercial pressurized-water reactor assumed to be equipped with fixed neutron-flux detectors is used to test the method. Comparisons with reference solutions show that, if the neutron detectors are sufficiently numerous and reasonably distributed throughout the core, and if their characteristics are perfectly known, nodal group-fluxes can be reconstructed in real-time with maximum errors of only a few percent, even for very severe transients involving control-rod motions and thermal-hydraulic feedback effects. Reactivity can be inferred by inverse kinetics with an accuracy of better than one millihertz.

Thesis Supervisor: Allan F. Henry
Title: Professor of Nuclear Engineering
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis supervisor, Professor Allan F. Henry, for his guidance and trust. This research has largely benefitted from his original ideas and great expertise. Working under Professor Henry’s direction has been not only an honor and a privilege, but also a very enriching experience for me, both personally and professionally.

Many thanks go to Professor David D. Lanning for reviewing this thesis and for making numerous constructive comments during group meetings.

Special thanks are also due to Professor John E. Meyer for many helpful suggestions.

I would like to extend my appreciation to the following people: Drs. John Bernard and Kwan Kwok (MIT Nuclear Reactor Laboratory), Dr. Philippe Finck (Argonne National Laboratory), Bruce Hagemeier (Boston Edison Company), Professors Alar Toomre, Nick Trefethen, and Nathan Siu, Dr. Brian Aviles, Dr. Michael Houts, Kenneth Brooks, Christopher Steele, John Chun, Stan DiVittorio, Rachel Morton, and Gloria Rook (MIT).

The financial support received from the U.S. Department of Energy under the Research Grant Program (Nuclear Engineering Research) is gratefully acknowledged.

Last, but not least, I owe eternal gratitude to my family and my fiancee, Laurence, for their continuous encouragements. This thesis is dedicated to them.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................ 2

ACKNOWLEDGEMENTS .................................................................................... 4

TABLE OF CONTENTS ...................................................................................... 5

LIST OF FIGURES ............................................................................................ 9

LIST OF TABLES ............................................................................................... 29

Chapter 1. INTRODUCTION ........................................................................... 33

1.1. Introduction and Overview ....................................................................... 33
1.2. The Neutron Diffusion Equations ......................................................... 37
1.3. Numerical Solution of the Group Diffusion Equations ...................... 43
1.4. Nodal Diffusion Methods ....................................................................... 46
1.5. Recent Developments in Core Power-Distribution
    Monitoring .................................................................................................. 49
1.6. Objective .................................................................................................. 52
1.7. Organization ............................................................................................... 53

Chapter 2. GENERAL DESCRIPTION OF THE METHOD .......................... 55

2.1. Introduction ............................................................................................... 55
2.2. Flux Reconstruction by Nodal Synthesis ............................................. 56
2.3. Characteristics of the Synthesis Method and Anticipated
    Difficulties .................................................................................................. 63
    2.3.1. A Simple Interpretation of the Method ......................................... 63
    2.3.2. Anticipated Difficulties ................................................................. 66
2.4. Reactivity Calculation by Inverse Kinetics ....................................... 69
2.5. Neutron-Flux Measurements in Light-Water Reactors .................. 74
    2.5.1. Neutron-Flux Detectors ................................................................. 74
        a. Fission Chambers .......................................................................... 76
        b. Self-Powered Neutron Detectors ................................................ 77
    2.5.2. Reactor Nuclear Instrumentation ................................................ 79
2.6. Summary ................................................................................................... 83
Chapter 3. MINIMUM-NORM LEAST-SQUARES SOLUTION
OF ILL-CONDITIONED RECTANGULAR SYSTEMS
OF LINEAR ALGEBRAIC EQUATIONS .................... 84

3.1. Introduction ......................................................... 84
3.2. Rectangular Systems of Linear Algebraic Equations ....... 85
  3.2.1. The Geometrical Approach ................................. 85
  3.2.2. The Algebraic Approach ..................................... 91
3.3. Practical Methods for Solving Rectangular Systems of
  Linear Algebraic Equations ....................................... 97
  3.3.1. Normal Equations ............................................ 97
  3.3.2. QR Factorization ........................................... 98
  3.3.3. Singular Value Decomposition ............................ 101
    a. Definition and Geometrical Interpretation .............. 101
    b. Existence and Construction ................................. 105
    c. SVD and Rectangular Systems of Equations ............ 109
    d. SVD and Ill-Conditioning ................................. 113
    e. Numerical Considerations .................................. 126
3.4. Application of SVD to the Fitted-Synthesis Equations ..... 129
  3.4.1. Least-Squares Residual as an Error Estimate ........ 129
  3.4.2. Orthogonalization of the Expansion Functions
    by SVD ............................................................. 133
3.5. Summary .................................................................. 140

Chapter 4. FIRST APPLICATIONS OF THE METHOD AND
REFINEMENT ............................................................. 143

4.1. Introduction ........................................................... 143
4.2. Testing Methodology ................................................ 144
  4.2.1. The QUANDRY Code ........................................... 144
    a. Static Module ............................................... 146
    b. Transient Module .............................................. 147
  4.2.2. Testing Procedure ............................................ 149
  4.2.3. Computer Equipment and Programs .................... 153
4.3. Reactor Model and Nuclear Instrumentation ............... 154
4.4. Description of the Transient and Reference Calculation .. 157
4.5. First Application of the Fitted-Synthesis Method .......... 165
  4.5.1. Expansion Functions ........................................ 165
  4.5.2. SVD and Zeroing of Small Singular Values .......... 167
  4.5.3. Results .......................................................... 171
4.6. Refinement of the Method: Time-Discontinuous Synthesis . 190
Chapter 5. NOISE CONSIDERATIONS AND FILTERING TECHNIQUES ........................................... 263

5.1. Introduction ............................................................................................................. 263
5.2. Fitted Synthesis in the Presence of Measurement Noise ....................................... 265
  5.2.1. Deterministic Analysis .................................................................................... 265
  5.2.2. Statistical Analysis ......................................................................................... 270
    a. Noise Characterization ..................................................................................... 271
    b. Expected Values ............................................................................................... 272
    c. Covariance Matrices ....................................................................................... 274
    d. The ’R-over-J’ Rule ......................................................................................... 281
5.3. Simulations with Noisy, Unfiltered Detector-Readouts ........................................ 286
  5.3.1. Artificial Noise ............................................................................................... 286
  5.3.2. Tests with Eighty In-Core Detectors .............................................................. 288
  5.3.3. Tests with Four In-Core Detectors ............................................................... 318
  5.3.4. Tests with Forty-Eight Out-of-Core Detectors ............................................. 332
5.4. Noise-Filtering Algorithms .................................................................................... 349
  5.4.1. Trade-Offs in Filtering ................................................................................. 351
  5.4.2. Filter Design ................................................................................................. 352
    a. FIR Versus IIR Filters .................................................................................... 352
    b. Direct Method Versus Indirect Method ............................................................ 356
  5.4.3. Fitted Synthesis With Filtered Noisy Signals ............................................... 363
    a. Fixed-Memory FIR Filters ............................................................................. 363
    b. Fading-memory IIR Filters ............................................................................. 385
    c. Modified Fixed-memory FIR Filters ............................................................. 400
5.5. Summary ................................................................................................................ 410

Chapter 6. SECOND AND THIRD APPLICATIONS OF THE METHOD ........................................... 412
6.1. Introduction .............................................................................. 412
6.2. Application to a Coolant Inlet-Temperature Transient .... 413
   6.2.1. Description of the Transient ........................................... 413
   6.2.2. Synthesis Calculations ...................................................... 418
      a. Expansion Functions ....................................................... 419
      b. Tests With Noise-Free Signals ....................................... 421
      c. Tests With Noisy Signals ............................................... 429
   6.2.3. Conclusion ....................................................................... 441
6.3. Motivations for the Startup Transient .............................. 442
6.4. Application to a Reactor-Startup Transient ...................... 443
   6.4.1. Description of the Transient ........................................... 443
   6.4.2. Synthesis Calculations ...................................................... 452
      a. Expansion Functions ....................................................... 452
      b. Results ............................................................................. 455
   6.4.3. Conclusion ....................................................................... 481
6.5. Summary ................................................................................ 482

Chapter 7. CONCLUSIONS AND RECOMMENDATIONS .............. 483

    7.1. Summary of the Investigation .............................................. 483
    7.2. Conclusions ........................................................................ 489
    7.3. Recommendations for Future Research ............................. 491

REFERENCES .................................................................................. 494

APPENDIX A1. DESCRIPTION OF THE PWR MODEL .................. 517

APPENDIX A2. POLYNOMIAL FILTERS ..................................... 522

   A.2.1. Fixed-Memory Polynomial Filters .................................. 522
   A.2.2. Fading-Memory Polynomial Filters .............................. 525
LIST OF FIGURES

2.1. Geometrical Interpretation of Eq. 2.16 ............................. 65
2.2. Ranges of In-Core Fission Chambers in BWRs ................... 82
3.1. Minimum-Norm Solution of a One-by-Two System of Equations ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ ................................................................. 86
3.2. Unique Solution of a Two-by-Two System of Equations ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ ........................................................................ 86
3.3. Least-Squares Solution of a Three-by-Two System of Equations ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ ................................................................. 87
3.4. \( \mathbf{A} \) as a Linear Map from \( \mathbb{R}^K \) to \( \mathbb{R}^J \) ......................... 89
3.5. Least-Squares Solution and Residual Vector ......................... 94
3.6. Orthonormal Bases for the Four Fundamental Subspaces Associated with \( \mathbf{A} \) ................................................................. 103
3.7. The Image by \( \mathbf{A} \) of an Arbitrary Vector \( \mathbf{T} \) of \( \mathbb{R}^K \) ............. 105
3.8. Minimum-Norm Least-Squares Solution of ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ by Use of the SVD of \( \mathbf{A} \) ................................................................. 110
3.9a. Image of a Unit Circle of \( \mathbb{R}^K \) by an Ill-Conditioned Matrix \( \mathbf{A} \) of Rank \( R = 2 \) ................................................................. 114
3.9b. Image of a Unit Circle of \( \mathbb{R}^J \) by an Ill-Conditioned Matrix \( \mathbf{A}^\dagger \) or Rank \( R = 2 \) ................................................................. 115
3.10. Effect of Ill-Conditioning on the Minimum-Norm Least-Squares Solution of ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ ................................................................. 119
4.1. Synoptic Diagram of the Numerical Simulations ................... 150
4.2a. X-Y Cross-Sectional View of the Reactor Showing the 20 Strings of In-Core Neutron-Detectors ........................................ 155

4.2b. X-Z Cross-Sectional View of the Reactor and Its Nuclear Instrumentation ........................................................................ 156

4.3. Simulated Withdrawal of Rod-Banks C and D ....................... 159

4.4a. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (Operational Transient, Cusping Correction Applied) ............................................................................. 162

4.4b. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (Operational Transient, No Cusping Correction Applied) .............................................................................. 163

4.5. QUANDRY-Computed Fuel Temperature and Coolant Temperature in Node # 893 Versus Time (Operational Transient) ...................................................................................... 164

4.6. Eigenvalue $k_{eff}$ Versus Control-Rod Pattern at Time $t$ for Two Different Power-Levels ................................................................. 167

4.7. Maximum Relative Error in Group-Two Flux Versus Number of Singular Values Retained in the Synthesis (Operational Transient, 26 Expansion Functions) .................... 170

4.8. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................. 171

4.9a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) ........................................... 173

4.9b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) ........................................... 174

4.10a. Relative Error in Group-One Flux at $t = 52$ s Versus Node Number (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 175
4.10b. Relative Error in Group-Two Flux at $t = 52$ s Versus Node Number (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 176

4.11. Cutaway View of the Quarter-Core Reactor Model Showing the Positions of the Representative Nodes .................. 178

4.12. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 180

4.13. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 181


4.15. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 183

4.16. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 184

4.17. Reference and Reconstructed Amplitude Function Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 186

4.18. Reference and Reconstructed Total Power Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 187

4.19. Reference and Reconstructed Prompt Neutron Lifetime Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 188

4.20. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed) .................................................. 189
4.21. Graphical Interpretation of the Time-Discontinuous Synthesis Method ........................................... 193

4.22. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................. 198

4.23a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) ...................... 201

4.23b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) ......................... 202

4.24a. Relative Error in Group-One Flux at t = 52 s Versus Node Number (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................. 203

4.24b. Relative Error in Group-Two Flux at t = 52 s Versus Node Number (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................... 204

4.25. Coefficients of the Linear Combination of Orthogonalized Expansion Functions (Operational Transient, 80 In-Core Detectors) ................................................................. 206

4.26. Modal Contributions to the Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors) ............................................................ 209

4.27. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Detectors) ............................................................ 210

4.28. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors) ............................................................ 211

4.29. Modal Contributions to the Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Detectors) ............................................................ 212
4.30. Modal Contributions to the Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Detectors) .................................................................................................................. 213

4.31. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 214

4.32. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 215

4.33. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 216

4.34. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 217

4.35. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 218

4.36. Reference and Reconstructed Amplitude Function Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 220

4.37. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 221

4.38. Reference and Reconstructed Prompt Neutron Lifetime Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 222

4.39. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed) .................................................................................................................. 223

4.40. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 80 In-Core Detectors, One Expansion Function) ................................. 224
4.41. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 80 In-Core Detectors, One Expansion Function) ........................................... 225

4.42. Relative Error in Group-Two Flux at $t = 52$ s Versus Node Number (Operational Transient, 80 In-Core Detectors, One Expansion Function) ........................................... 226

4.43. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors, One Expansion Function) ........................................... 227

4.44a. Euclidean Norm of Solution-Vector and Relative Residual Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) ......................................... 234

4.44b. Euclidean Norm of Minimum-Norm Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, Four In-Core Detectors, One Singular Value Zeroed) .... 235

4.45a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) ............................... 236

4.45b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) .................................... 237

4.46a. Relative Error in Group-One Flux at $t = 52$ s Versus Node Number (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) .......................................... 238

4.46b. Relative Error in Group-Two Flux at $t = 52$ s Versus Node Number (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) .......................................... 239

4.47. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Detectors) ................................................................. 240

4.48. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Detectors) ................................................................. 241
4.49. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) ................................................................. 242

4.50. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) ................................................................. 243

4.51. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed) ................................................................. 244

4.52. X-Y Cross-Sectional View of the Reactor Model Showing the 12 Strings of Out-of-Core Neutron-Detectors ...................... 246

4.53. Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ................................................................. 248

4.54a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ....................................................... 249

4.54b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ....................................................... 250

4.55. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Detectors) ................................................................. 251

4.56. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Detectors) ................................................................. 252

4.57. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ....................................................... 254

4.58. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ....................................................... 255
4.59. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ......................................................... 256

4.60. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ......................................................... 257

4.61. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed) ......................................................... 258

5.1. Least-Squares Solution of ‘$$A \mathbf{T}(t) = \mathbf{C}(t) + \delta \mathbf{C}(t)$$’ .................. 267

5.2. Noise-Free and Noisy Outputs of the Neutron Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors) ......................................................... 290

5.3. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed) ....... 291

5.4. Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed) ......................................................... 293

5.5. First Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors) ......................................................... 294

5.6. Second Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors) ......................................................... 295

5.7. Third Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors) ......................................................... 296

5.8. Fourth Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors) ......................................................... 297

5.9. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed) ......................................................... 298
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.10.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed)</td>
<td>299</td>
</tr>
<tr>
<td>5.11.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed)</td>
<td>300</td>
</tr>
<tr>
<td>5.12.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed)</td>
<td>301</td>
</tr>
<tr>
<td>5.13.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed)</td>
<td>302</td>
</tr>
<tr>
<td>5.14.</td>
<td>Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>305</td>
</tr>
<tr>
<td>5.15.</td>
<td>Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>306</td>
</tr>
<tr>
<td>5.16.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>307</td>
</tr>
<tr>
<td>5.17.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>308</td>
</tr>
<tr>
<td>5.18.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>309</td>
</tr>
<tr>
<td>5.19.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>310</td>
</tr>
<tr>
<td>5.20.</td>
<td>Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed)</td>
<td>311</td>
</tr>
</tbody>
</table>
5.21. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, Two Singular Values Zeroed) ........................................... 314

5.22. Reference and Reconstructed Amplitude Function Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed) ........................................... 315

5.23. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed) ........................................... 316

5.24. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed) ........................................... 317

5.25. Relative Residual Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed) ............... 319

5.26. Euclidean Norm of Solution-Vector Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed) ........................................... 320

5.27. First Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors) ........................................... 321

5.28. Second Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors) ........................................... 322

5.29. Third Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors) ........................................... 323

5.30. Fourth Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors) ........................................... 324

5.31. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed) ........................................... 325

5.32. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed) ........................................... 326
5.33. Reference and Reconstructed Group-Two Flux in Node # 649
Versus Time (Operational Transient, Four In-Core Noisy
Detectors, No Singular Value Zeroed) ........................................... 327

5.34. Reference and Reconstructed Group-Two Flux in Node # 893
Versus Time (Operational Transient, Four In-Core Noisy
Detectors, One Singular Value Zeroed) ........................................... 328

5.35. Reference and Reconstructed Group-Two Flux in Node # 651
Versus Time (Operational Transient, Four In-Core Noisy
Detectors, One Singular Value Zeroed) ........................................... 329

5.36. Reference and Reconstructed Group-Two Flux in Node # 649
Versus Time (Operational Transient, Four In-Core Noisy
Detectors, One Singular Value Zeroed) ........................................... 330

5.37. Reference and Reconstructed Group-Two Flux in Node # 649
Versus Time (Operational Transient, Four In-Core Noisy
Detectors, Two Singular Values Zeroed) ......................................... 331

5.38. Noise-Free and Noisy Outputs of the Neutron Detector
Located in Node # 495 Versus Time (Operational Transient,
48 Out-of-Core Noisy Detectors) .................................................... 333

5.39. Relative Residual Versus Time (Operational Transient, 48
Out-of-Core Noisy Detectors, No Singular Value Zeroed) ...... 334

5.40. Euclidean Norm of Least-Squares Solution-Vector Versus
Time (Operational Transient, 48 Out-of-Core Noisy
Detectors, No Singular Value Zeroed) ............................................. 335

5.41. First Modal Coefficient Versus Time (Operational Transient,
48 Out-of-Core Noisy Detectors) .................................................... 336

5.42. Second Modal Coefficient Versus Time (Operational Transient,
48 Out-of-Core Noisy Detectors) .................................................... 337

5.43. Third Modal Coefficient Versus Time (Operational Transient,
48 Out-of-Core Noisy Detectors) .................................................... 338

5.44. Fourth Modal Coefficient Versus Time (Operational Transient,
48 Out-of-Core Noisy Detectors) .................................................... 339
5.45. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, No Singular Value Zeroed) ........................................ 340

5.46. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, No Singular Value Zeroed) ........................................ 341

5.47. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, No Singular Value Zeroed) ........................................ 342

5.48. Relative Residual Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed) ..... 343

5.49. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed) ........................................ 345

5.50. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed) ........................................ 346

5.51. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed) ........................................ 347

5.52. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, Two Singular Values Zeroed) ........................................ 348

5.53. Filtered Output $y_n$ Evaluated from a Sequence of Measurements $x_k$ ........................................................................... 349

5.54. FIR Filter of Eq. 5.79 .............................................................. 354

5.55. FIR Filter of Eq. 5.82 .............................................................. 355

5.56. Response Function of Ideal Low-Pass Filter ......................... 359

5.57. Response Function of Approximate Low-Pass Filter .......... 362
5.58. Notational Conventions for Fixed-Memory Polynomial Filter ................................................................. 363

5.59a. Variance Reduction Factor for Fixed-Memory Polynomial Filters of Various Degrees M Versus Length L of the Filtering Interval (Estimation at the Center of the Interval) .... 368

5.59b. Variance Reduction Factor for Fixed-Memory Polynomial Filters of Various Degrees M Versus Length L of the Filtering Interval (Estimation at the End of the Interval) ........ 369

5.60. Noise-Free and Filtered Noisy Outputs of the Neutron Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval) ............................................ 372

5.61. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) .......... 373

5.62. First Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval) ......................................................... 375

5.63. Second Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval) ......................................................... 376

5.64. Third Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval) ......................................................... 377

5.65. Fourth Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval) ......................................................... 378

5.66. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) ............. 379
5.67. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) .................................. 380

5.68. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) .................................. 381

5.69. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) ..................................................... 383

5.70. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) ..................................................... 384

5.71. Notational Conventions for Fading-Memory Polynomial Filter .................................................................................................................. 385

5.72a. Time Response of a Fading-Memory Polynomial Filter of Order One to a Step Change in Input Signal for Different Weighting-Factors Theta ........................................................................ 388

5.72b. Time Response of a Fading-Memory Polynomial Filter of Order Two to a Step Change in Input Signal for Different Weighting-Factors Theta ........................................................................ 389

5.72c. Time Response of a Fading-Memory Polynomial Filter of Order Three to a Step Change in Input Signal for Different Weighting-Factors Theta ........................................................................ 390

5.73. Variance Reduction Factor for Fading-Memory Polynomial Filters of Various Degrees M Versus Weighting-Factors Theta (0-Step Estimation) .................................................................................. 391

5.75. Noise-Free and Filtered Noisy Outputs of the Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8) ........................................................................ 394
5.76. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed) .... 395

5.77. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed) ............... 397

5.78. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed) ............... 398

5.79. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed) ............... 399

5.80a. Response Function of Second-Order Fixed-Memory Polynomial Filter Versus Normalized Frequency for Different Values of the Filtering-Interval Length, L .................. 403

5.80b. Response Function of Second-Order, Modified, Fixed-Memory Polynomial Filter Versus Normalized Frequency for Different Values of the Filtering-Interval Length, L .................. 405

5.81. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Modified Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) ............... 407

5.82. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Modified Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) ............... 408

5.83. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Modified Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed) .................................. 409
6.1. Coolant Inlet-Temperature and QUANDRY-Computed Core-Averaged Fuel Temperature Versus Time (Coolant Inlet-Temperature Transient) ................................................................. 415

6.2. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (Coolant Inlet-Temperature Transient) ............... 417

6.3. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (Coolant Inlet-Temperature Transient) ..... 418

6.4. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 422

6.5a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 423

6.5b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 424

6.6. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 425

6.7. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 426

6.8. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 427

6.9. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed) ............... 428

6.10. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed) ............... 430

24
6.11a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed) .................. 431

6.11b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed) .................. 432

6.12. Noise-Free and Noisy Outputs of the Detector Located in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors) ................................................. 433

6.13. Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed) ................................................................. 434

6.14. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed) ...... 435

6.15. Reference and Reconstructed Total Power Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed) ........................................ 437

6.16. Reference and Reconstructed Reactivity Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed) ........................................... 438

6.17. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed) ......................... 439

6.18. Reference and Reconstructed Total Power Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed) ................................................................. 440

6.19. Reference and Reconstructed Reactivity Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed) ................................................................. 441
6.20. The Five Phases of the Startup Transient ........................................ 445

6.21a. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (First Half of Startup Transient) ......................... 447

6.21b. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (Second Half of Startup Transient) ......................... 448

6.22a. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (First Half of Startup Transient) ...... 450

6.22b. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (Second Half of Startup Transient) ................................................................. 451

6.23a. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ........................................ 457

6.23b. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ........................................ 458

6.24a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ........................................ 459

6.24b. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ........................................ 460

6.25a. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ........................................ 461

6.25b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ........................................ 462

6.26a. Reference and Reconstructed Group-One Flux in Node # 330 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ........................................ 463
6.26b. Reference and Reconstructed Group-One Flux in Node # 330 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 464

6.27a. Reference and Reconstructed Group-Two Flux in Node # 330 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ................................................................. 465

6.27b. Reference and Reconstructed Group-Two Flux in Node # 330 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 466

6.28a. Reference and Reconstructed Group-Two Flux in Node # 1060 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ................................................................. 467

6.28b. Reference and Reconstructed Group-Two Flux in Node # 1060 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 468

6.29a. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ................................................................. 469

6.29b. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 470

6.30a. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ................................................................. 471

6.30b. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 472

6.31a. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (First Half of Startup Transient, 80 In-Core Detectors) ................................................................. 473

6.31b. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) ................................................................. 474
6.32a. Reference and Reconstructed Amplitude Function Versus Time (First Half of Startup Transient, 80 In-Core Detectors) .......... 475

6.32b. Reference and Reconstructed Amplitude Function Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) .......... 476

6.33a. Reference and Reconstructed Total Power Versus Time (First Half of Startup Transient, 80 In-Core Detectors) .......... 477

6.33b. Reference and Reconstructed Total Power Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) .......... 478

6.34a. Reference and Reconstructed Reactivity Versus Time (First Half of Startup Transient, 80 In-Core Detectors) .......... 479

6.34b. Reference and Reconstructed Reactivity Versus Time (Second Half of Startup Transient, 80 In-Core Detectors) .......... 480

6.35. Reference and Reconstructed Reactivity Versus Time With Effective External Source Neglected (Startup Transient, [40 s, 120 s] Interval, 80 In-Core Detectors) ......................... 481

A1.1. X-Y Rodded Cross-Section of the PWR Model (1/4 Core) Showing Boundary Conditions and Composition Numbers ...... 517
# LIST OF TABLES

3.1. Effect of Ill-Conditioning on the Solution of a Simple, Two-by-Two System, $\Delta \ T = \ C$, for Four Particular Right-Hand-Side Vectors $\ C$ ......................................................... 125

3.2. Compared Operation-Counts of Three Different Algorithms for Solving Rectangular, J-by-K Systems of Equations (J $\geq$ K) .................................................................................. 127


4.1. Numerical Parameters Used in the QUANDRY Reference Calculation (Operational Transient) ................................................................. 160

4.2. Eigenvalues $k_{\text{eff}}$ Corresponding to Two Power-Levels and 13 Control-Rod Patterns ................................................................. 166

4.3. Singular-Value Spectrum of the 80-by-26 Matrix $\Delta$ ............... 168


4.5. Figure Numbering for the Results of the Time-Continuous and Time-Discontinuous Syntheses .......................................................... 200

4.6. Comparisons Between Discontinuous-Synthesis, Continuous-Synthesis, and Point-Kinetics Results (80 In-Core Detectors) ......................................................... 228

4.7a. Maximum Errors in Group-Fluxes Reconstructed by a Four-Mode Time-Discontinuous Synthesis ................................. 259

4.7b. Maximum Errors in Group-Fluxes Reconstructed by a Three-Mode Time-Discontinuous Synthesis .................................................. 259

4.7c. Maximum Errors in Group-Fluxes Reconstructed by a Two-Mode Time-Discontinuous Synthesis .................................................. 260
5.1. Figure Numbering for the Results of the Tests With 80 In-Core Detectors ......................................................... 289

5.2. Variance Reduction Factor for the Reconstructed Group-Two Flux in Node # 893 (Four-Mode Synthesis) ......................... 303

5.3. Variance Reduction Factor for the Reconstructed Group-Two Flux in Node # 893 (Three-Mode Synthesis) ......................... 312

5.4. Figure Numbering for the Results of the Tests With 4 In-Core Detectors ................................................................. 318

5.5. Figure Numbering for the Results of the Tests With 48 Out-of-Core Detectors ............................................................... 332

5.6a. Coefficients $w_r(-L/2)$ and Variance Reduction Factors (VRF) for Fixed-Memory Polynomial Filters of Order $M = 0$ and 1 ........................................................................................................ 366

5.6b. Coefficients $w_r(-L/2)$ and Variance Reduction Factors (VRF) for Fixed-Memory Polynomial Filters of Order $M = 2$ and 3 ........................................................................................................ 367

5.7. Influence of the Order of the Polynomial, $M$, Length of Filtering Interval, $L$, and Sampling Rate, $1/\tau$, on the Performance of a Fixed-Memory Polynomial Filter ........................................ 370

5.8. Length $L_{eq}$ of the Filtering Interval of a Fictitious Fixed-Memory Polynomial Filter Equivalent to a Fading-Memory Polynomial Filter of Identical Order and Weighting Factor $\theta (\theta \geq 0.8)$ ........................................................................... 392

5.9. Influence of the Order of the Polynomial, $M$, Weighting Factor, $\theta$, and Sampling Rate, $1/\tau$, on the Performance of a Fading-Memory Polynomial Filter .................................................. 393

5.10a. Transfer Functions of Second- and Third-Order Fixed-Memory Polynomial Filters ...................................................... 402

5.10b. Transfer Functions of Second- and Third-Order, Modified, Fixed-Memory Polynomial Filters ........................................ 404
6.1. Numerical Parameters Used in the QUANDRY Reference Calculation (Coolant Inlet-Temperature Transient) .......................... 416

6.2. Eigenvalue $k_{\text{eff}}$ for Three Different Power-Levels and Three Different Values of Coolant Inlet-Temperature .......................... 419

6.3. Normalized Singular-Value Spectrum of the 80-by-9 Matrix $\Delta$ ............................................................................. 420

6.4. Numerical Parameters Used in the QUANDRY Reference Calculation (Startup Transient) .................................................. 446

6.5. Time-Intervals and Corresponding Expansion-Functions for the Discontinuous Synthesis (Startup Transient) .................. 453


A1.1. Two-Group Nodal Cross Sections and Diffusion Coefficients for the PWR Model ............................................................. 518

A1.2. Discontinuity Factors for all Nodes, Groups, and Directions ... 519

A1.3. Albedo Values for Radial and Axial Boundaries .................. 519

A1.4a. Coolant-Density Feedback Coefficients ............................... 520

A1.4b. Coolant-Temperature Feedback Coefficients .......................... 520

A1.4c. Fuel-Temperature Feedback Coefficients ............................... 521

A1.5. Delayed-Neutron Precursor Fractions and Decay Constants .... 521

A2.1. Normalization Constants of the First Discrete Legendre Polynomials ................................................................. 523

A2.2. Coefficients of Fixed-Memory Polynomial Filters of Order $M = 1, 2, 3,$ and $4$ ..................................................... 524

A2.3. Variance $\sigma_{\text{L}^2}$ for Fixed-Memory Filters of Order $M = 0, 1, 2,$ and $3$ (Filtering at the Center of the Interval) ............. 525
A2.4. Variance $\sigma_0^2$ for Fixed-Memory Filters of Order $M = 0, 1, 2,$ and $3$ (Filtering at the End of the Interval) ......................... 525

A2.5. Weights of Error Estimate in Recursive Formulation of Fading-Memory Filters of Order $M = 0, 1, 2,$ and $3$ ................. 528

A2.6. Variance $\sigma_0^2$ for Fading-Memory Filters of Order $M = 0, 1,$ 2, and $3$ (Filtering at the End of the Interval) ......................... 529
Chapter 1

INTRODUCTION

1.1. Introduction and Overview.

The safety and good economical performance of large, commercial, nuclear plants require the knowledge and control at any instant of the energy stored in the reactor core and its spatial distribution. Energy in a reactor is generated in fuel rods (approximately 1 cm in diameter) by neutron-induced fission reactions. It is then transferred to the surface of the cladded fuel-rods by conduction and carried away by a cooling fluid by convection. The means by which plant operators exercise control upon the energy stored in a reactor is by altering the delicate balance between the rate of energy addition and the rate of energy removal. During operation, this can be achieved by motions of control rods made of neutron-absorbing material for instance.

The rates of energy addition and removal, although fundamental to the safe and optimal operation of the plant, are not directly measurable. However, they can be related to other variables which are measurable, in particular local neutron flux, coolant temperature, pressure and flow-rate. In practice, restrictions are imposed directly on these observable variables and other quantities. These restrictions constitute a safety envelope which guarantees that, during normal, transient- or steady-state operation, the electricity-production objective is achieved without challenge to the integrity of any component, especially without overheating and failure of the fuel cladding material, which would cause release of radioactive fission products into the core cooling system.
In large, commercial, light-water reactors, the core thermal-power distribution is controlled both globally or in level, and locally or in shape. In pressurized-water reactors (PWRs), global control is achieved by motion of selected banks of control rods and by addition or dilution of soluble neutron-absorber (boric acid) in the coolant. Limited variations in the core-inlet coolant temperature are also allowed. In boiling-water reactors (BWRs), bulk control is accomplished either by rod motions or by action on flow control valves. Changes in the coolant recirculation flow through the core indirectly affect the fission rate because the latter is very sensitive to water temperature and density.

In both types of reactors, local control or shaping of the power distribution is partially achieved through the use of fuel pins of different enrichment in $^{235}$U, periodic rearrangement of the fuel assemblies, and incorporation of burnable neutron-absorbing material at predetermined locations in the core. During operation, “fine tuning” of the power shape is possible by insertion or withdrawal of control rods.

Ideally, one would like to be able to perform the following tasks in real time:

(i) Compute the instantaneous, three-dimensional distributions of neutron flux, fission power, xenon concentration, coolant temperature, and coolant density, and infer from them actual values of total power, reactivity, axial and radial peaking factors, axial offset, as well as margins to crucial thermal limits, in particular maximum linear heat-generation rate, departure from nucleate boiling limit, minimum critical power ratio;

(ii) Compare the theoretical predictions of observable quantities (e.g. neutron flux) with local measurements provided by a sufficiently rich and fast-responding plant instrumentation;

(iii) Act upon some control variable (e.g. control-rod position) so as to achieve a desired, “optimal” plant response while respecting all safety, technical, and operating constraints [E2].

Such a structure could be the heart of a closed-loop, on-line control system in a largely automated plant in which the operators would have only a supervisory role.
The main component of such an advanced control system is its on-line core power-distribution monitoring function. This function entails the determination of the instantaneous neutron-flux density throughout the reactor. Indeed, from this flux-density map, values of local heat-generation rates (i.e. pin power-density map) are readily computed. Other important variables such as fuel temperature, coolant temperature, and coolant density can subsequently be inferred from this distributed heat-source.

One major difficulty of reactor analysis stems from the fact that, when a nuclear reactor operates at nominal power, local heat-generation, heat transfer, and cooling mechanisms influence each other in a complicated manner. Realistic predictions of power-density maps require that important thermal-hydraulic phenomena be modeled, and that any change in the geometry or in local material properties be properly reflected in the neutronic parameters (cross sections). Mathematically, this means that one has to face a set of nonlinear, multidimensional, time-dependent, coupled, partial (integro-)differential equations. The geometrical complexity and high degree of heterogeneity of commercial reactors make the numerical solution of these equations a formidable endeavor, even for present-day powerful computers. Simplifying assumptions are necessary. They are made either in the modeling equations themselves, or in the discretization and numerical procedures used to solve them [H8, D11, N1, W7, V1].

In reactor static as well as transient analyses, the standard approach consists in solving the neutronic and thermal-hydraulic problems in tandem (using "coupled codes"). First, approximate values of fuel temperature, coolant temperature, and coolant density distributions are specified. Material properties are then updated to reflect these thermal-hydraulic conditions. The changes in material properties are "fed back" onto the neutronic parameters (feedback model), and neutron-flux and power distributions are computed. Using this known heat source, more accurate temperature distributions can be calculated. This iterative process is continued until convergence.
In transient studies, these iterative calculations must be repeated at every time step. In long-lasting transients (not considered in the subsequent discussion), effects caused by variations in the concentration of some nuclides (e.g. xenon, uranium) must also be taken into consideration, further complicating the calculations.

Until very recently, performing such three-dimensional iterative calculations was extremely expensive for steady-state problems, and prohibitively so for time-dependent problems. Instead, one had to rely on more approximate methods, on the reactor fixed instrumentation, and on periodic measurements provided by movable sensors. It is only in the past ten years or so, with the developments of efficient methods of reactor analysis and with the new generation of digital computers [J4, J5, L11, V2], that such an objective has appeared feasible.

In most reactors in operation today, because of the difficulty of performing accurate on-line (real time) calculations to simulate the dynamic behavior of the plant, and also because of a limited, inaccurate, or unreliable instrumentation, reactor operation is restricted to certain allowable regions well within the safety envelope. This conservative approach is intended to guarantee the safe operation of the plant under all circumstances, even in "worst case" scenarios. Operating limits are expressed in terms of restrictions on global quantities such as total power, axial offset, control-rod positions, quadrant tilt. The margins to safety limits which are imposed on these quantities are derived from the conservative assumption that the power distribution takes its most limiting shape.

The price to pay for introducing conservative operating-margins is that plant performance is usually far from optimal, which ultimately translates into an economic penalty. It is generally recognized that improved performance would result if accurate and reliable measurements and predictions of the neutron-flux distribution were available to the plant operators and control systems in real time. Under such circumstances, the needlessly restrictive operating-band could be relaxed. Potential benefits which would result from these margin gains include maximum core-power
output with minimum peaking, increased operating flexibility, higher plant availability and capacity factors, higher output power, better fuel utilization, reduction in fuel failure-rate, early detection and correction of anomalies, and improved confidence during plant operation [C2, D6, W11, K10].

Significant advances have been made in the past fifteen years in the development of methods for reactor-transient analyses [H7, S17, W7, A3, H9]. This progress, when combined with elaborate algorithms and numerical solution techniques, should make it possible to perform real-time calculations of three-dimensional neutron flux shapes on desktop (rather than mini-) computers in the near future, even for moderately fast transients. Central to these new developments are two-group neutron diffusion theory and nodal methods, which are discussed in the next sections.

1.2. The Neutron Diffusion Equations.

A rigorous description of the spatial and temporal variations of the fission power in a reactor requires the determination of the directional neutron-flux density, \( \Psi \), as a function of neutron position, \( r \), relative energy, \( E \), relative direction of travel, \( \Omega \), and time, \( t \) [A1, H8, G17, D11]. In principle, this information can be obtained by solving the governing time-dependent neutron transport equation for \( \Psi(r, E, \Omega, t) \). However, this is a very difficult task in general. Fortunately, for commercial light-water reactors, the difficulty can be circumvented since the neutron diffusion equation, a low-order approximation to the transport equation, has been found to yield sufficiently accurate results. Refinements even exist to correct the results of diffusion theory in those regions (near strong heterogeneities) where it breaks down [S7].

In diffusion theory, the dependent variable is the scalar neutron flux density, \( \Phi(r, E, t) \), defined as the average value of \( \Psi(r, E, \Omega, t) \) over all directions \( \Omega \). Solving the time-dependent neutron diffusion equation for
\( \Phi(r, E, t) \), although considerably easier than solving the transport equation, still remains a very challenging task in general.

Obtaining a numerical solution to the diffusion equation requires that the space, energy, and time variables be discretized. Usually, the energy variable is discretized first. This is done by dividing the neutron continuous energy-spectrum into intervals or "groups", \( \Delta E_g, \ g = 1, 2, \ldots, G \). Group fluxes are defined as

\[
\phi_g(r, t) \equiv \int_{\Delta E_g} \phi(r, E, t) \, dE.
\]  

The energy-group diffusion equations for the \( \phi_g(r, t) \)'s are obtained by integration of the continuous-energy diffusion equations over each energy interval \( \Delta E_g \). No additional approximation is necessary to obtain these equations provided all coefficients (group cross sections, diffusion coefficients, and speeds) are defined consistently. The result may be written in compact matrix form as

\[
- \mathbf{L} \Phi + (1-\beta) \chi_F E^T \Phi + \sum_{i=1}^{I} \chi_i \lambda_i C_i + \mathbf{Q} = \mathbf{V}^{-1} \frac{\partial}{\partial t} \Phi, 
\]  

\[
\beta_i E^T \Phi - \lambda_i C_i = \frac{\partial}{\partial t} C_i, \quad i = 1, 2, \ldots, I
\]

where

\( \Phi = \) column vector of group fluxes (neutrons cm\(^{-2}\) s\(^{-1}\)) defined as

\[
\Phi = \phi(r, t) = [\phi_1(r, t), \phi_2(r, t), \ldots, \phi_G(r, t)]^T,
\]  

\( G = \) total number of energy groups (-),

\( \mathbf{L} = \) net-loss (leakage plus absorption) operator (cm\(^{-1}\)) defined as
\[ L \phi = - \nabla \cdot D \nabla \phi + A \phi \quad \text{for all } \phi, \quad (1.5) \]

\[ D = \text{diag}\{D_1(\tau, t), D_2(\tau, t), \ldots, D_G(\tau, t)\}, \]

\[ D_g(\tau, t) = \text{diffusion coefficient for energy group } g \, (\text{cm}), \]

\[ A = \text{absorption operator (cm}^{-1}) \text{ defined as} \]

\[ A_{gg'} \equiv \Sigma_{tg}(\tau, t) \delta_{gg'} - \Sigma_{gg}(\tau, t), \quad \begin{cases} g = 1, 2, \ldots, G, \\ g' = 1, 2, \ldots, G, \end{cases} \quad (1.6) \]

\[ \Sigma_{tg}(\tau, t) = \text{total interaction cross section for group } g \, (\text{cm}^{-1}), \]

\[ \Sigma_{gg'}(\tau, t) = \text{scattering cross section from group } g' \text{ to group } g \, (\text{cm}^{-1}), \]

\[ F = [\nu \Sigma_{f_1}(\tau, t), \nu \Sigma_{f_2}(\tau, t), \ldots, \nu \Sigma_{f_G}(\tau, t)]^T, \]

\[ \Sigma_{fg}(\tau, t) = \text{fission cross section for group } g \, (\text{cm}^{-1}), \]

\[ \nu = \text{mean number of neutrons emitted per fission} \, (-), \]

\[ \chi_p = [\chi_{p_1}, \chi_{p_2}, \ldots, \chi_{p_G}]^T, \]

\[ \chi_{pg} = \text{fraction of prompt neutrons born in group } g \, (-), \]

\[ \chi_i = [\chi_{i_1}, \chi_{i_2}, \ldots, \chi_{i_G}]^T, \]

\[ \chi_{ig} = \text{fraction of delayed neutrons born in group } g \text{ from the } i\text{-th} \]

\[ Q = [Q_1(\tau, t), Q_2(\tau, t), \ldots, Q_G(\tau, t)]^T, \]
\( Q_g(\tau, t) = \text{external (nonfission) source of group-g neutrons (neutrons cm}^{-3} \text{s}^{-1}) \),

\[ \mathbf{v} = \text{diag} \{ v_1, v_2, \ldots, v_G \} , \]

\( v_g = \text{relative speed of group-g neutrons (cm s}^{-1}) \),

\( C_i = C_i(\tau, t), \; i = 1, 2, \ldots, I, \; \text{(atoms cm}^{-3}) \),

\( I = \text{total number of delayed-neutron precursor families (} \cdot) \),

\( \beta_i = \text{fraction of delayed neutrons from the i-th precursor family (} \cdot) \),

\[ \beta = \sum_{i=1}^{I} \beta_i = \text{total delayed-neutron fraction}, \]

\( \lambda_i = \text{decay constant of the i-th precursor family (s}^{-1}) \).

The appearance of the precursor concentrations, \( C_i \), as extra unknowns is a consequence of the fact that not all neutrons are born as "prompt" neutrons, i.e. at the instant of fission. A small fraction of them (\( \beta_i \)) are withheld in precursor nuclei and do not appear until these nuclei have undergone a beta-decay reaction which may take several seconds or even several tens of seconds. Although the \( C_i \)'s themselves are not observable and of little practical interest, the existence of the "delayed" neutrons is fundamental to the control of nuclear reactors.

Neutron group-fluxes and precursor concentrations are required to be continuous and nonnegative everywhere within the reactor, and to vanish on its outer boundary. The neutron current, approximated in diffusion theory by Fick's law as

\[ I \approx -D \nabla \phi , \] (1.7)
is required to be continuous across any internal surface. Initial conditions must also be specified.

Eqs. 1.2 and 1.3 are mere instantaneous balance equations for each neutron group and precursor family. The time scale of interest here is a few minutes or less. Therefore, long-term dynamic effects such as xenon-samarium poisoning or fuel depletion are not considered.

The group cross sections and diffusion coefficients are function of position because of reactor heterogeneity. They are function of position and time, explicitly because of external perturbations, and implicitly because of thermal-hydraulic feedback effects. These effects, which may be prompt ($\leq 10^{-3}$ s) or delayed ($\sim 0.1$ s or more), can lead to significant spatial and spectral distortions in flux shape and therefore must be taken into consideration. Mathematically, this implies that the neutronic model formed by Eqs. 1.2 and 1.3 must be complemented by equations and variables describing the thermal-hydraulic and feedback processes. The different time-scales involved in the full set of equations cause an inherent stiffness which requires the use of implicit numerical solution-techniques [V1, W7].

In steady-state conditions, Eqs. 1.2 and 1.3 reduce to

\[ - L_0 \phi_0 + \chi F_0^T \phi_0 + Q_0 = 0, \quad (1.8) \]

\[ \beta_i F_0^T \phi_0 = \lambda_i C_{i0}, \quad i = 1, 2, \ldots, I, \quad (1.9) \]

where

\[ \chi \equiv (1 - \beta) \chi_p + \sum_{i=1}^{I} \beta_i \chi_i \quad (1.10) \]

is a vector of total, group fission-spectra.
Above a certain power-level, external-source contributions are negligible with respect to fission sources, and Eq. 1.8 reduces to the following homogeneous system:

\[ L_0 \phi_0 = \frac{1}{\lambda} \chi F_0^T \phi_0. \]  

(1.11)

Reactors are designed in such a way that Eq. 1.11 has a solution other than the trivial one \( \phi_0 = 0 \). Mathematically, this can always be achieved by introducing an eigenvalue, \( 1/\lambda \), which is determined together with the solution vector \( \phi_0 \).

Generally, Eq. 1.11 has more than one eigenvalue-eigenvector solution-pair \( \{ \lambda, \phi_0^{(\lambda)} \} \). These multiple solutions are called "\( \lambda \)-modes". It can be shown that there is only one eigenvector, \( \phi_0 \), which is everywhere nonnegative. This eigenvector corresponds to the largest eigenvalue \( \lambda = k_{\text{eff}} \). This particular eigenvalue-eigenvector pair is called the "fundamental mode". It is the physical solution of Eq. 1.11. When \( k_{\text{eff}} = 1 \), the reactor is "critical".

All quantities with subscript '0' in Eqs. 1.8, 1.9 and 1.11 are independent of time. However, nonlinearities remain since \( L_0 \) and \( F_0 \) depend implicitly on the neutron distribution through thermal-hydraulic feedback effects. As already pointed out, this dependence must be known and accounted for at high power. This means that steady-state analogs of the time-dependent thermal-hydraulic and feedback models complementing Eqs. 1.2 and 1.3 must be adjoined to Eq. 1.11. This also implies that a tandem iterative procedure must be employed to obtain a solution to Eq. 1.11.

For commercial light-water-moderated reactors, the numerical solution of the group diffusion equations is greatly facilitated by the remarkable fact that accurate predictions of criticality and power distributions can be obtained by using only two energy groups (G = 2) [H8, T2]. Usually, numerical accuracies of ± 0.0005 in eigenvalue and a few percent in local power density are judged acceptable and are consistent with modeling and parametric uncertainties.

Most of the difficulties of light-water reactor analyses come from the spatial (rather than spectral) dependence of the neutron flux. The physical reason for this is that neutron diffusion lengths are fairly small (~ 1 cm) with respect to the size of the reactor core (~ a few meters), which can lead to significant spatial distortions in flux shape.

In these conditions, conventional numerical methods such as finite-difference techniques require a very small mesh-spacing to yield sufficiently accurate results [H8, W7]. These very fine meshes translate into huge matrix systems with many unknowns. Although advantage can be taken of the sparseness and simple structure of these systems, and in spite of efficient solution-techniques, three-dimensional computations remain very expensive. Nevertheless, much confidence has been established in the use of finite-difference codes [C1, V6]. Today, they remain a standard means of generating reference solutions which are useful to assess the performance of other methods. However, such codes are rarely used for full-core, 3-D, transient analyses [F1].

The finite-element method has also been applied to the neutron diffusion equations [K1, L1, R5, W7, H8, A3]. Finite-element methods have a number of advantages over finite-difference methods. In particular, they do not require as many unknowns to achieve comparable accuracy. In addition, they give numerical results at every location and not just at discrete points. Furthermore, any local mesh refining does not propagate needlessly throughout the grid. However, these advantages are completely offset by difficulties with the numerical solution of the final system of
equations. Even with simple elements and low-order polynomial trial-functions, this system does not have a structure as simple as that of finite-difference methods.

Modal and synthesis methods are alternatives to finite-difference and finite-element methods. In both methods, the scalar flux-density is expanded in terms of precomputed "expansion-functions" (or "basis-functions"). The result can be a drastic reduction in the number of unknowns. However, the cost of generating the required expansion-functions may limit the applicability of these methods.

In modal methods, the expansion functions are the eigenvectors of some operator [S13, D5, W7]. These eigenvectors or "modes" should form a complete set. The method is of practical interest when the formally-exact infinite sum can be well approximated by an expansion limited to only a few modes. There are situations where the generation of high-order modes is difficult and expensive, but the major difficulty with modal methods is rather in deciding a priori how many modes should be retained in the truncated expansion to achieve a given accuracy.

In synthesis methods [H8], much more flexibility is allowed in the choice of the basis functions. In particular, they are not restricted to the eigenmodes of a single operator. This great amount of flexibility is the key to the success of the method. However, this success comes at a price: a loss of mathematical rigor in the sense that no theoretical error bound can be derived for the solution. Nonetheless, in spite of this and other deficiencies [A2, C3, Y2], synthesis methods have been applied in various forms with considerable success since the early sixties [K3, K4, Y1, Y3, Y4, O1, O2].

One particular form of synthesis approximation is called "time synthesis" [K5, K7, B9, W7, Y3, Y4, M2]. It is briefly described here since the idea will be at the heart of the semi-experimental synthesis method to be described in Chapter 2. In the time synthesis, the space and time dependences of the group fluxes are separated by writing
\[ \phi_g(L, t) = \sum_{k=1}^{K} \psi_g^{(k)}(\tau) T_g^{(k)}(t), \quad g = 1, 2, ..., G. \] (1.12)

The \( \psi_g^{(k)} \)’s in Eq. 1.12 are known expansion-functions. They may be generated by performing, for instance, a series of \( \lambda \)-mode calculations (Eq. 1.11) for different reactor-conditions expected during transients. The \( T_g^{(k)} \)’s are unknown, scalar, group “mixing coefficients” which depend only on time. Various procedures exist to determine them. One method consists in substituting Eq. 1.12 into the time-dependent group diffusion equations (Eqs. 1.2 and 1.3) and in requiring the resulting formulas to be true in a weighted integral sense [S13, L5]. This leads to a system of \( G \times K \) ordinary differential equations for the \( T_g^{(k)} \)’s. Another method is to derive them from a variational principle [K5, L10, S13, K7]. A similar set of ordinary differential equations results and must be solved.

Numerical tests have demonstrated that the time-synthesis method can yield extremely good results provided an “appropriate” set of expansion functions is used. The physical reason for this success is that, for transients of practical interest, the prompt-neutron population readjusts itself very rapidly (in less than a few milliseconds) to any change in reactor conditions. This rapid readjustment accounts for most of the changes in flux shape and amplitude. In the subsequent evolution, the flux distribution remains close to a shape that can be well approximated by a linear combination of steady-state flux-shapes as in Eq. 1.12. Deviations caused by delayed precursor and feedback effects are usually fairly “homogeneous” and therefore do not lead to significant errors.

One unfortunate drawback of the time-synthesis method is that there is no recommended “best” procedure for generating the required expansion-functions. In addition, once a method has been chosen, there is no general prescription other than physical intuition for selecting the reactor conditions for which expansion functions should be generated. Moreover, there is no guarantee that increasing the number of basis functions will systematically lead to better results [K5].
A more detailed description of each of the above methods can be found in the listed references. Today, however, it can almost be said that, to a large extent, they have all been supplanted by modern nodal methods.

1.4. Nodal Diffusion Methods.

Over the past ten years or so, modern nodal methods have emerged as the method of choice for solving the three-dimensional neutron diffusion equations [S6, D9, D3, L3, K17, S11, H9]. A great deal of evidence has now been accumulated in favor of these methods [N2, D7].

For a given accuracy, modern nodal methods give criticality predictions (i.e. solve the nodal form of Eq. 1.11) with approximately two orders of magnitude less computational effort than finite-difference methods. These computing-time savings are achieved through a substantial reduction in the size of the system of equations to be solved. Specifically, nodal methods make it possible to use much larger meshes than finite-difference methods for the same accuracy (i.e. spatial truncation error). Even though nodal methods usually involve additional unknowns per mesh (surface currents), the net result is a reduction in the size of the final matrix system. In addition, the structure of this system is such that it is amenable to the same efficient numerical solution-techniques as those used in finite-difference codes. Typically, a two-group, three-dimensional, full-core, nodal criticality calculation involves on the order of 25,000 unknowns and requires a few minutes of CPU time on a workstation.

The quantities of interest in nodal methods are the group fluxes averaged over coarse meshes (say 20 cm x 20 cm x 20 cm) or "nodes":

$$\phi_{g_n} = \frac{1}{V_n} \iiint_{V_n} \phi_g(r) \, dr$$  \hspace{1cm} (1.13)
where $V_n$ is the volume of node $n$.

Steady-state calculations based on nodal methods proceed in several steps [S9, S11].

(1) In the first step, a set of few-group, homogenized, equivalent, diffusion-theory macroscopic cross sections and diffusion coefficients are determined for representative nodes in the reactor. These parameters are found from the requirements that overall criticality and individual reaction-rates be preserved. Different sets of values are computed to account for differences in fuel enrichment, boron concentration, fuel temperature, moderator temperature and density, past operating history (e.g. exposure, influence of control rods), ...etc. In each of these 2-D, 6-to-20-group, “assembly” calculations, the geometrical and material details of the interior of the node are explicitly represented. Different “cell” parameters are used to represent fuel pins, burnable-absorber pins, control-rods, water holes, ...etc. Approximate node boundary-conditions are used. The final result of each of these “lattice calculations” is a set of equivalent, few-group (typically two-group) cross sections and diffusion coefficients which are spatially constant within each node [H8, K15, S9].

(2) In the second step, a set of 3-D nodal equations involving these equivalent few-group parameters is solved. These equations consist of a “nodal balance equation” and “coupling equations”. The unknowns are the eigenvalue $\lambda = k_{\text{eff}}$, node averaged fluxes, $\phi_{g\alpha}$, and surface-averaged net currents. The nodal balance equation is obtained without approximation by simple integration of the group diffusion equations (Eq. 1.11) over the volume of an arbitrary node. Coupling equations assure closure of the system by providing additional (approximate) relationships between node-averaged fluxes and surface-averaged currents [W2, B8, S6, L3].

In modern nodal methods, these coupling equations are derived in a systematic and consistent way. For example, in 3-D Cartesian geometry, for a given node, they are obtained by first integrating the group diffusion equations over two directions, say $y$ and $z$, to yield one-dimensional equations for the twice-integrated nodal group-fluxes, $\phi_{g\alpha}(x)$. An
assumption is then made about the shape of the unknown transverse-leakage terms as a function of x. (Usually, a quadratic profile is assumed.) Next, the 1-D equations are solved either analytically (as in the Analytic Nodal Method [S6]) or by approximating the surface fluxes by a polynomial expansion (as in the Nodal Expansion Method [F6]). Integration of the result, \( \phi_{gn}(x) \), between the node boundaries yields a relationship between node-averaged flux, \( \phi_{gn} \), surface fluxes and surface currents. By repeating this procedure for neighbor nodes in the x-direction and by combining the results to eliminate all surface fluxes, one set of group coupling-equations is obtained. The two other sets are obtained similarly by exchanging the roles of x, y, and z.

The importance of the homogenization calculations (Step 1) cannot be overemphasized. To a large extent, the quality of the equivalent, homogenized reactor-parameters conditions the accuracy of the subsequent nodal calculations (Step 2). However, it must be pointed out that correction factors can be incorporated into the nodal calculations to reduce the errors caused by the homogenization procedure [K15, S7, S9].

(3) In an optional third step, the node-averaged group-fluxes on the one hand, and surface-averaged group-fluxes and currents on the other hand, can be used to reconstruct local, intranodal group-fluxes and pin power-density maps, if such a detailed, pin-by-pin, ‘finite-difference’-type information is desired (for accurate fuel-depletion calculations for instance). Various methods have been developed to perform such dehomogenization calculations [K16, F4, K9, F5, P3, G15, S9, S10, R2, R3].

Frequently, however, the quantities of immediate interest are precisely the eigenvalue \( k_{\text{eff}} \), node-averaged group-fluxes and power-density maps which result directly from the nodal calculations (Step 2). In these conditions, Step 3 is not needed.

The considerable success of nodal methods for solving the steady-state neutron diffusion equations (Eq. 1.11) has naturally lead to their application to time-dependent problems (Eqs. 1.2 and 1.3) [S6, S8, G2, C4,
C5, K2, A7, E3]. Nodal diffusion kinetics codes have been developed and are now routinely used for the analysis of reactor transients. Fast-running versions of these programs are being incorporated into complex, on-line core power-distribution monitoring and control systems.

1.5. Recent Developments in Core Power-Distribution Monitoring.

The importance of the continuous availability of the full-core, three-dimensional power-distribution during reactor operation has already been stressed, and some of the expected benefits have been enumerated. Significant developments have taken place in the past fifteen years and have resulted in advanced ‘fuel-management optimization’ and ‘power-shape monitoring’ systems. Most of these systems are tailored to a particular type of reactor and operates in conjunction with its instrumentation. Some are being tested in existing plants. Advantages offered by these systems include, in addition to those already mentioned, the reduction in the need for human intervention during operation, and the concomitant reduction in the risk of human error.

The following partial list attests to the interest that advanced systems of core-power monitoring and control have stimulated in recent years [C2]:
- Electric Power Research Institute’s Power Shape Monitoring System (PSMS) for BWRs [F9, B11] and PWRs [P1, P2, E7, K10], and Advanced Core Monitoring Framework [I1, I2],
- Exxon Nuclear Company’s POWERPLEX Core Monitoring Software System (CMSS) for BWRs [F2, F3],
- Westinghouse PWRs’ Core On-Line Surveillance Monitoring, and Diagnostics System (COSMOS), and BEACON core monitoring system [H4, E1, B5, B6],
- Siemens/KWU PWRs’ Power Density Detector (PDD) and Aeroball Systems [E5],
- General Electric BWRs’ 3-D MONICORE on-line core monitoring system [C6],
- Combustion Engineering PWRs' MINI-CECOR system, Core Operating Limit Supervisory System (COLSS) [H16, E6], and Better Axial Shape Selection System (BASSS) [H6, T3, T4, V5, T5],
- Babcock and Wilcox PWRs' Prompt Incore Detector System (PIDS) [R4, H5, R6, R7].

Similar systems have been developed or are under development in other countries [L6, W11, B12, S2, F8].

At the Massachusetts Institute of Technology (MIT), research efforts have lead to the development of a 'Power Plant Simulator and Controller' (POPSICLE) for PWR transient studies [A7].

An increasing number of these core-power monitoring and control systems are constructed around 3-D, two-group, nodal kinetics codes coupled with thermal-hydraulic codes [K10]. A few still rely on simpler models [N4, H12, F2], on modal or space-synthesis programs [T3, T5], or on measurements. Many systems possess some means of adapting the theoretical predictions to the available measurements. Some even have predictive capabilities [H11, F2, G3].

Even though the heart of all these simulators is the evaluation of the core power-distribution, other functions related to plant safety, control, and fuel-management are often integral part of these systems.

For control purposes, it is sometimes argued that the "best" of these systems are those that rely on a physical model (such as a 3-D nodal model). Although model-based calculation and control systems have indeed numerous well-known advantages [K11], they also have drawbacks which are sometimes overlooked.

A first drawback is the disproportionate burden which is often placed on the computational facilities as a result of the large memory and real-time capability required by model-based systems. In fact, many of the physics codes embodied in such systems are not yet able to perform accurate, three-dimensional, full-core power-distribution calculations in real time for moderately fast transients (time scale of ~ 1 s), even when
implemented on powerful minicomputers or workstations [N4, F2, L11, S2]. (The qualifier "on-line" usually implies that the power-distribution calculations are carried out in parallel to the operation of the plant, but not necessarily in real time.) Advantage can sometimes be taken of symmetries in the core composition pattern and control actions. However, this is not always possible. In spite of the speed-up brought by quasi-static-type approximations [S12, O6, T1, K2], much computing effort is still being spent in evaluating the 3-D neutron-flux shape iteratively at every time-step with thermal-hydraulic feedback effects taken into account. In addition, in these calculations, memory requirements are often substantial and, in some cases, more limiting than computer speed itself.

Notwithstanding these limitations, one can reasonably anticipate, that, with the advent of more powerful computers, it will eventually be possible to achieve the objective of faster-than-real-time full-core calculations. At present, however, this remains quite difficult with desktop machines, except for fairly slow transients.

Independently of these severe hardware requirements, there are numerous complications which results from the applications of these codes to plant-control. Any model-based controller should

(i) lead to a stable behavior of the closed-loop system,
(ii) take into account modeling and parametric uncertainties,
(iii) produce estimates of the errors in the calculated variables,
(iv) be robust, i.e. in particular fairly insensitive to (high-frequency) disturbances, and
(v) have some adaptive capabilities.

In purely numerical studies, it is often assumed that the model is exact, that the values of all parameters are known exactly, and that external perturbations (e.g. control-rod motion) and their effects on plant parameters (cross sections) are perfectly known. In practice, however, this is never the case. In an actual implementation, all modeling and parametric uncertainties should be quantified, periodically reevaluated, and factored into the design of the model-based controller. In addition, provision should be made to avoid excessive sensitivity of the model predictions to disturbances. Ideally, there should also be a means of adjusting some parameters in the model to achieve better agreement between computed and
measured quantities. Finally, model-based control systems should, as any control system, be able to accommodate a certain amount of noise in the measurements and be fault-tolerant [K11].

In view of the above enumeration, one may very well conclude that the most serious drawback of model-based control systems is the excessive complexity that they introduce, especially when they are intended to cover a wide range of plant operating conditions. It will probably take some time before these systems are fully accepted, implemented, licensed, and before the actual benefits (and drawbacks) resulting from their use can be better appreciated. Although there is little doubt that model-based methods will ultimately prevail, in the meantime, there is still room for innovative alternatives to the direct numerical solution of the full set of coupled, three-dimensional, time-dependent neutronic and thermal-hydraulic equations. These alternate methods should be simpler and should operate in real time without requiring intensive computational resources nor large software developments.

1.6. Objective.

The objective of this research is to develop a simple, approximate yet acceptably accurate method for estimating three-dimensional, instantaneous neutron-flux distributions without solving the governing time-dependent neutron diffusion equations (Eqs. 1.2 and 1.3). The essential idea of the method is to approximate instantaneous, nodal, group fluxes by a linear combination of precomputed expansion-functions as in the time-synthesis method. These expansion functions will be obtained from three-dimensional, steady-state, nodal calculations. The coefficients of the combination will not be determined in a theoretical manner, but instead directly from local flux measurements provided by fixed neutron detectors. Such a procedure should be computationally very efficient.

Important issues to be addressed, questions to be answered, and problems to be remedied include:
- Feasibility, i.e. accuracy, in particular for severe transients,
- Relative efficiency (cost and memory requirements),
- Selection of the expansion functions and their number,
- Possible near linear-dependences between expansion functions,
- Effect of numerical roundoff errors, measurement uncertainties
and noise,
- Sensitivity of the results to the number of detectors and their
locations,
- Identification of advantages, possible limitations, and shortcomings
of the method.

The investigation will rely entirely on numerical simulations. The
synthesis method will be implemented in the form of a computer program
and, for each transient studied, the results will be compared with a
reference numerical solution provided by a 3-D nodal kinetics code. The
synthesis program will receive as inputs the precomputed expansion-
functions and the flux measurements. The latter will not come from actual
neutron-detectors but will be simulated, i.e., in fact, inferred from the
reference solution itself.

1.7. Organization.

In Chapter 2, a general description of the proposed, semi-
experimental, nodal synthesis method is given. It is shown in particular
how the flux-reconstruction procedure leads to a rectangular system of
linear algebraic equations. Potential sources of difficulties or errors with
this system and with the method in general are identified.

Various numerical techniques for solving rectangular systems of
equations are reviewed in Chapter 3. The choice of the particular
technique of ‘Singular Value Decomposition’ is justified. This method is
described in detail with emphasis on its unique capability to detect and
remedy ill-conditioning in matrices.

A first application of the method to a severe operational transient in
a PWR is described in Chapter 4. The problem of the selection of the
expansion functions is discussed, and the importance of the number and location of the detectors is studied.

The effect of measurement noise is investigated in Chapter 5. Remedies in the form of simple filters are proposed.

In Chapter 6, results of applications of the method to other transients are presented.

Finally, Chapter 7 contains a summary of the investigation, conclusions, and recommendations for future research.
Chapter 2

GENERAL DESCRIPTION OF THE METHOD

2.1. Introduction.

The on-line monitoring of the three-dimensional neutron-flux distribution is essential to the safety, control, and optimum performance of light-water reactors. In Chapter 1, reasons were given for the difficulty of obtaining this detailed information in real time with sufficient accuracy and at an acceptable cost. With modern nodal diffusion methods, a direct approach does not appear unreasonable anymore and current developments concentrate on elaborate monitoring and control systems using fast-running, 3-D, nodal kinetics codes. These model-based systems are very promising, but their complexity is often such that they cannot operate in real time yet, except for slow transients. The need for simpler and more efficient methods still exists, in particular for alternatives to the direct, expensive numerical solution of the time-dependent, few-group, neutron diffusion equations. The general characteristics of a semi-experimental nodal synthesis method which would provide such an alternative were outlined.

A detailed description of this semi-experimental synthesis method is given in this Chapter. It is shown in particular how the method leads to a rectangular system of linear algebraic equations to be solved. The general characteristics of this system of equations are discussed, but the algebraic detail associated with its least-squares solution is deferred to Chapter 3. Most of the notation is introduced here and the definitions will not be repeated in subsequent chapters. The formulas required to compute point-kinetics quantities, in particular reactivity, from the reconstructed nodal
group fluxes are given. Since the synthesis method relies on neutron-flux measurements, a brief review of neutron detectors currently in use in reactors is also included. Results of numerical tests of the method will be presented in Chapters 4, 5, and 6.

2.2. Flux Reconstruction by Nodal Synthesis.

The proposed, semi-experimental, nodal synthesis method is based on a collapsed-group, time-synthesis approximation in which the instantaneous neutron flux, \( \phi_{gn}(t) \), in energy-group \( g \) and node \( n \) (Eq. 1.13), is written as a linear combination of a set of \( K \) precalculated expansion-functions, \( \psi_{gn}^{(k)} \).

\[
\phi_{gn}(t) = \sum_{k=1}^{K} \psi_{gn}^{(k)} T^{(k)}(t), \quad \begin{cases} g = 1,2,\ldots,G, \\ n = 1,2,\ldots,N. \end{cases}
\]  
(2.1)

The basis functions, \( \psi_{gn}^{(k)} \), are time independent. They are computed beforehand for selected steady-state reactor-conditions approximating and bracketing the reactor transient-conditions. (Admittedly this requires some \emph{a priori} knowledge about the perturbation initiating the transient and about the subsequent evolution of the reactor.)

The \( T^{(k)} \)'s are scalar “mixing coefficients” to be determined. In Eq. 2.1, there are the same for all energy groups \( g \) and depend only on time. Once the expansion functions have been computed, these coefficients are the only unknowns in Eq. 2.1.

Synthesis methods have been extensively and successfully used despite recognized drawbacks (see Section 1.3). However, until recently, applications were limited to 1-D and 2-D geometries because of the cost of generating three-dimensional expansion-functions directly from finite-difference calculations [Y4]. Today, modern nodal diffusion codes provide a means of generating these 3-D shapes in a reasonably accurate and fairly
inexpensive manner (see Section 1.4). Therefore, it is now possible to apply the synthesis idea more realistically to 3-D flux distributions as in Eq. 2.1. Whether this type of synthesis can be as accurate as simpler syntheses of 1-D and 2-D flux-shapes remains to be verified though.

The nodal synthesis approximation, Eq. 2.1, permits achieving a substantial reduction in the number of unknowns, from \( G \times N \) (\( \sim 50,000 \)) to \( K \) (\( \sim 10 \)). The effect of Eq. 2.1 is, in fact, to replace the original space-, energy-, and time-dependent problem by a problem containing only time-dependent unknowns. This simplification can be expected to result in appreciable computing-time savings. However, it may also lead to a loss of accuracy. The savings are justified insofar as the simultaneous deterioration in accuracy remains acceptably small. One may intuitively expect the quality of the reconstructed nodal group-fluxes to depend to a large extent upon the expansion functions since all spatial and spectral variations are "lumped" into these basis functions [K5].

In this study, the expansion functions, \( \psi_{g_n}^{(k)} \), will be obtained either from nodal, criticality calculations (Eq. 1.11) or from subcritical, steady-state calculations with known external sources (Eq. 1.8). In the case of basis functions generated from criticality calculations, only fundamental \( \lambda \)-modes, i.e. those eigenvectors of Eq. 1.11 that are everywhere nonnegative (see Section 1.2), will be retained in the synthesis. One may argue that improved results could be obtained by incorporating higher-order \( \lambda \)-modes in the synthesis (as in a modal expansion). However, this will not be done for at least two reasons. The first one is that generating such harmonics would complicate the synthesis and increase the overall cost of the method. The second reason is an \textit{a posteriori} justification: The applications of Chapter 4 and 6 will demonstrate that sufficiently accurate results can be obtained \textit{without} the use of high-order \( \lambda \)-modes in Eq. 2.1.

For a given "best" choice of expansion functions, the success of Eq. 2.1 still depends on the accurate determination of the \( K \) mixing coefficients, \( T^{(k)} \). One way to find these coefficients is to substitute Eq. 2.1 into the group, nodal, neutron diffusion equations and to require the result
to be true in some weighted-integral sense [D4, L5]. The drawback of this approach is that it requires the evaluation of a number of integrals or inner products which involve space-, energy-, and time-dependent arguments (cross sections). All functional dependences must be known for a proper evaluation of these inner products. In particular, the instantaneous effects of external perturbations and thermal-hydraulic feedback must be known and taken into consideration. Ultimately, the determination of the $T^{(k)}$'s amounts to the solution of a set of coupled, partial and ordinary differential equations, an undesirable complication.

In the present, semi-experimental method, a simpler approach is adopted. It is based on the observation that there are only a small number of unknowns to be determined, namely the $K$ coefficients $T^{(k)}(t)$. The idea is then to use actual flux-measurements provided by the reactor instrumentation to find numerical values for these $K$ unknowns. The procedure is detailed below.

Assuming that the reactor is equipped with a total of $J$ fixed neutron-detectors, the signals returned by these detectors may be formally written in the following, general, linear form:

$$
C^{(j)}(t) = \int_{V^{(j)}} d\tau \int_{E_{\text{min}}^{(j)}}^{E_{\text{max}}^{(j)}} \Sigma^{(j)}(\tau, E, t) \phi(\tau, E, t) \, dE, \quad j = 1, 2, \ldots, J, \quad (2.2)
$$

where continuous space- and energy-variables have been used, and where

- $V^{(j)}$ is the volume of the $j$-th detector,
- $E_{\text{min}}^{(j)}$ and $E_{\text{max}}^{(j)}$ are respectively the lower and upper energy-sensitivity limits of the $j$-th detector ($E_{\text{min}}^{(j)} \geq 0$, $E_{\text{max}}^{(j)} \leq +\infty$),
- $\phi(\tau, E, t)$ is the instantaneous flux density of neutrons of relative energy $E$ at point $\tau$ within $V^{(j)}$, and
- $\Sigma^{(j)}(\tau, E, t)$ is the response function of the $j$-th detector.
\[ \Sigma^{(i)}(r, E, t) \] depends on the particular type of detector (material composition, geometry, physical interactions involved). In practice, \( \Sigma^{(i)}(r, E, t) \) incorporates corrections for such effects as burnup of emitting material, undesirable contributions (such as those from gamma radiations), self-shielding, and neutron-flux depression \[ J3].

For operation in “pulse mode” (a mode in which individual impulsions can be resolved), a detector is termed “counter” and its output has units of \( \text{counts s}^{-1} / (\text{neutrons cm}^{-2} \text{s}^{-1}) \). For operation in “current mode”, the units are usually \( \text{Ampere} / (\text{neutrons cm}^{-2} \text{s}^{-1}) \). In the following, for reasons of simplicity, this distinction between units will not be made. It will be assumed instead that the \( C^{(i)} \)'s are all expressed in the same arbitrary units since the proper conversion or normalization factors can always be included in the \( \Sigma^{(i)} \)'s.

Eq. 2.2 is an observation equation which formally relates detector “output”, \( C^{(i)}(t) \), to detector “input”, \( \phi(r, E, t) \). A more useful “equivalent” formula involving the discrete unknowns, \( \phi_{gn}(t) \), can be derived from Eq. 2.2:

\[
C^{(i)}(t) = \sum_{g=1}^{G} \sum_{n=1}^{N} V_n \Sigma_{gn}^{(i)}(t) \phi_{gn}(t), \quad j = 1, 2, ..., J. \tag{2.3}
\]

where \( V_n \) is the volume of the node \( n \), and the \( \Sigma_{gn}^{(i)}(t) \)'s are equivalent, homogenized, nodal, group cross sections representative of the \( j \)-th detector and surrounding materials at time \( t \).

The procedure leading from Eq. 2.2 to Eq. 2.3 is in many aspects similar to the technique used in nodal methods to obtain equivalent, homogenized cross sections and diffusion coefficients (Section 1.4). This homogenization calculation must be carried out over an appropriate number of nodes and energy groups. In Eq. 2.3, for each detector \( j \), the summations have been extended from the homogenization region to the entire reactor volume and neutron energy-spectrum, with the convention
that $\Sigma_{g_n}^{(j)}(t)$ is zero outside the homogenization region. Note that, if the detectors are sensitive only to a single group of neutrons, the summation over energy in Eq. 2.3 can be dropped. In addition, if the detector is located entirely within one node, and if the homogenization region is chosen to include that node only, then the summation over space can also be dropped.

In general, the factors $\Sigma_{g_n}^{(j)}(t)$ in Eq. 2.3 may be interpreted as node- and group-dependent weight functions giving the appropriate instantaneous contributions $V_n \Sigma_{g_n}^{(j)}(t) \phi_{g_n}(t)$ of all nodal group fluxes to the output $C^{(j)}(t)$ of the j-th detector. In addition to being dependent on a particular detector (and on the "history" of that detector), these weight functions also vary with local reactor-conditions such as fuel exposure, assembly history, moderator density, proximity of control rods, ... etc, which makes their accurate evaluation quite difficult in general [C8, C9, T6, V4, R1]. Although no detail is given here about the two-step calculation of these $\Sigma_{g_n}^{(j)}(t)$'s, it should be realized that the practical value of Eq. 2.3 is highly dependent upon an accurate determination of these weights.

Although the $\Sigma_{g_n}^{(j)}(t)$'s depend on time, this time-dependence is slow with respect to the time scale of interest ($\sim 1$ s). (Note also that any known, positive, time-dependent factor, $f^{(j)}(t)$, can always be extracted from the $\Sigma_{g_n}^{(j)}(t)$'s in Eq. 2.3 provided a factor $1/f^{(j)}(t)$ is simultaneously incorporated into the $C^{(j)}(t)$'s.) In the sequel, no time-dependence will be shown explicitly, but one should bear in mind that the $\Sigma_{g_n}^{(j)}(t)$'s will inevitably have to be updated from time to time.

The use of a matrix notation for Eqs. 2.1 and 2.3 will greatly simplify the rest of the presentation. Define:

- a G×N-element column-vector of nodal group-fluxes,

$$\underline{\phi}(t) \equiv [\phi_{11}(t), \phi_{21}(t), ..., \phi_{G1}(t), \phi_{12}(t), ..., \phi_{GN}(t)]^T,$$

(2.4a)
- a similar, $G \times N$-element column-vector for each expansion function,

$$
\psi^{(k)} = \left[ \psi_{11}^{(k)}, \psi_{21}^{(k)}, ..., \psi_{G1}^{(k)}, \psi_{12}^{(k)}, ..., \psi_{GN}^{(k)} \right]^T, \quad k = 1, 2, ..., K, \quad (2.4b)
$$

- a $G \times N$-by-$K$ matrix whose columns are the expansion-function vectors $\psi^{(k)}$,

$$
\Psi = \left[ \psi^{(1)}, \psi^{(2)}, ..., \psi^{(K)} \right], \quad (2.4c)
$$

- a $K$-element column-vector of mixing coefficients,

$$
\Upsilon(t) = \left[ T^{(1)}(t), T^{(2)}(t), ..., T^{(K)}(t) \right]^T, \quad (2.4d)
$$

- a $G \times N$-element column-vector of weights for each detector,

$$
\Sigma^{(j)} = \left[ V_1 \Sigma_{11}^{(j)}, V_1 \Sigma_{21}^{(j)}, ..., V_1 \Sigma_{G1}^{(j)}, V_2 \Sigma_{12}^{(j)}, ..., V_N \Sigma_{GN}^{(j)} \right]^T, \quad j = 1, 2, ..., J, \quad (2.4e)
$$

- a $G \times N$-by-$J$ matrix whose columns are the weight vectors $\Sigma^{(j)}$,

$$
\Sigma = \left[ \Sigma^{(1)}, \Sigma^{(2)}, ..., \Sigma^{(J)} \right], \quad (2.4f)
$$

- and a $J$-element column-vector of measurements,

$$
\mathbf{C}(t) = \left[ C^{(1)}(t), C^{(2)}(t), ..., C^{(J)}(t) \right]^T. \quad (2.4g)
$$

With these definitions, Eqs. 2.1 and 2.3 may be rewritten simply as

$$
\phi(t) = \sum_{k=1}^{K} \psi^{(k)} T^{(k)}(t), \quad (2.5)
$$
\[ C^{(j)}(t) = \sum^{(j)^T} \phi(t), \quad j = 1,2,...,J, \quad (2.6) \]

or, in an even more compact form,

\[ \phi(t) = \Psi T(t), \quad (2.7) \]

\[ C(t) = \Sigma^T \phi(t). \quad (2.8) \]

For the forthcoming discussion, it will be convenient to define a vector of reconstructed nodal group-fluxes as

\[ \hat{\phi}(t) = \sum_{k=1}^{K} \psi^{(k)} T^{(k)}(t) = \Psi T(t), \quad (2.9) \]

and to introduce formally an **error vector**

\[ \delta\phi(t) = \hat{\phi}(t) - \phi(t). \quad (2.10) \]

Combination of Eqs. 2.8, 2.9 and 2.10 yields

\[ \Sigma^T \Psi T(t) - \Sigma^T \delta\phi(t) = C(t), \quad (2.11a) \]

or equivalently

\[ \Delta T(t) + E(t) = C(t), \quad (2.11b) \]

with

\[ \Delta = \Sigma^T \Psi \quad (2.12) \]

and

\[ E(t) = -\Sigma^T \delta\phi(t). \quad (2.13) \]
If the error vector \( \delta \phi(t) \) is small with respect to \( \hat{\phi}(t) \) (in the sense of some arbitrary vector-norm in \( \mathbb{R}^{GN} \)), then the vector \( E(t) \) is also small with respect to \( A \ T(t) \). In these favorable conditions, Eq. 2.11b may be written as

\[
A \ T(t) \approx C(t). \tag{2.14}
\]

Eq. 2.14 constitutes a system of linear algebraic equations for the unknown vector \( T(t) \) of mixing coefficients. If a solution, \( T(t) \), can be found to those equations, it can be substituted into Eq. 2.9 to give an estimate of the instantaneous flux-distribution, \( \hat{\phi}(t) \).

This procedure appears simple and straightforward. However, serious doubts remain concerning its accuracy. In fact, this question is so primordial that it is appropriate to pause here to identify possible sources of errors before even attempting to solve Eq. 2.14.

2.3. Characteristics of the Synthesis Method and Anticipated Difficulties.

2.3.1. A Simple Interpretation of the Method.

One might think that obtaining a solution to Eq. 2.14 is a trivial matter: All that is needed to determine \( T(t) \) is to invert the \( A \) matrix. In addition, so long as \( A \) does not change, the same inverse matrix can be used repeatedly to process successive right-hand-side vectors \( C(t) \) received from the instrumentation. Unfortunately, this simple idea is not directly applicable to Eq. 2.14 because \( A \) is not necessarily a square matrix and therefore cannot be “inverted”.

63
In general, the coefficient matrix $\mathbf{A}$ in Eq. 2.14 is rectangular, with $J$ rows and $K$ columns. Every entry of $\mathbf{A} = \Sigma^T \Psi$ is an inner product and is positive. Each row corresponds to a particular detector, whereas each column corresponds to a particular expansion function. This can be seen clearly if Eq. 2.12 is rewritten using Eq. 2.4f as

$$\mathbf{A} = \left[ \Sigma^{(1)}^T \Psi, \Sigma^{(2)}^T \Psi, \ldots, \Sigma^{(J)}^T \Psi \right]^T,$$

(2.15a)

or alternatively using Eq. 2.4c

$$\mathbf{A} = \left[ \Sigma^T \psi^{(1)}, \Sigma^T \psi^{(2)}, \ldots, \Sigma^T \psi^{(K)} \right].$$

(2.15b)

Eq. 2.15b shows that each column of $\mathbf{A}$ is the image through the instrumentation (i.e. by the flux-to-measurement operator $\Sigma^T$) of an expansion function.

As an illustration, consider the simple case where $J = 3$ and $K = 2$. In this particular situation, Eq. 2.14 takes on the following form:

$$\begin{bmatrix} \Sigma^{(1)}^T \psi^{(1)} & \Sigma^{(1)}^T \psi^{(2)} \\ \Sigma^{(2)}^T \psi^{(1)} & \Sigma^{(2)}^T \psi^{(2)} \\ \Sigma^{(3)}^T \psi^{(1)} & \Sigma^{(3)}^T \psi^{(2)} \end{bmatrix} \begin{bmatrix} T^{(1)}(t) \\ T^{(2)}(t) \end{bmatrix} = \begin{bmatrix} C^{(1)}(t) \\ C^{(2)}(t) \\ C^{(3)}(t) \end{bmatrix}.$$ 

(2.16)

Subsequent chapters will show that a desirable situation is $J \gg K$, i.e. many more detectors than basis functions. In these conditions, Eq. 2.14 is an overdetermined system and has no solution in general. Nonetheless, in principle, it is always possible to solve it in a least-squares sense.

Assuming temporarily that such a least-squares solution has been found, the reconstructed flux-vector, $\hat{\phi}(t)$, can be immediately computed using Eq. 2.9. $\hat{\phi}(t)$ may be interpreted as that particular, instantaneous, linear combination of $K$, three-dimensional, nodal expansion-functions which is in best least-squares agreement with $J$ local flux measurements.
representative of the unknown vector $\phi(t)$. Figure 2.1 is a simple geometrical illustration of this statement for $K = 2$ and $J = 3$, as in Eq. 2.16. The figure assumes that (i) the reactor is one-dimensional, (ii) there is only one neutron energy-group, and (iii) the position variable is continuous.

![Figure 2.1. Geometrical Interpretation of Eq. 2.16.](image)

This simple idea of approximating the actual flux-distribution by directly fitting a combination of precomputed flux-shapes to detector readouts is actually not new [R8, L9, K19, K13, B17, H15, L8]. A number of analogous methods which consist in adjusting an approximate formula (often a truncated modal expansion) to measurements have been suggested in the past and, in some cases, have even been applied to actual reactors [K19]. For synthesis approximations, however, it seems that no result
involving three-dimensional, nodal flux-shapes has ever been reported. In addition, these hybrid methods suffer from a number of drawbacks and limitations which are not always recognized.

Some potential difficulties with the present, fitted-synthesis method have already been mentioned. The following analysis is a more systematic identification of possible sources of difficulties and errors. This analysis will prove extremely useful in subsequent chapters by highlighting areas where attention is needed. This in turn will suggest ways of improving the method and even remedies to some problems.

2.3.2. Anticipated Difficulties.

A first important observation is that the synthesis method does not belong to the category of model-based techniques (Section 1.5). No system of time-dependent equations is solved to determine \( \hat{\phi}(t) \). The present approach is also different from, and simpler than, methods which combine a model with measurements [B4, B17, M1], in some cases to accelerate the convergence of iterative criticality calculations [M4], or to force agreement with a model by adjustment of free parameters ("leakage correction terms") [C6]. In those methods, a large system of equations must still be solved.

The fitted-synthesis method does not rely on any dynamic model, but rather on a postulated, approximate formula and on measurements. The elimination of the model greatly alleviates the computational burden. However, this simplification comes at a price: Without a model, it becomes very difficult to predict the future evolution of the plant, i.e. to make "projections" which could be useful to the plant operators. In addition, because of this absence of a model and because of the great flexibility allowed in the generation of the expansion functions, no theoretical estimate is available for the error, \( \delta\hat{\phi}(t) \), in the estimate of the state vector,
\( \hat{\phi}(t) \). In fact, as already pointed out in Section 1.4, it is not even possible to derive a theoretical upper bound for \( \| \delta \hat{\phi}(t) \| \).

In addition to these inherent limitations, the fitted-synthesis method apparently suffers from a lack of robustness to multiple sources of errors. These errors may arise from either the synthesis approximation itself, the expansion functions, \( \psi^{(k)} \), or the mixing coefficients, \( T^{(k)}(t) \). They may have either theoretical, numerical, or experimental origin.

As discussed in Section 1.3, the time-synthesis formula relies on the assumption that the instantaneous neutron-flux distribution can be well approximated by a linear combination of static flux-shapes even during moderately fast transients. An inappropriate mode of generation, a poor selection of reactor configurations, or an insufficient number (K) of expansion functions in Eq. 2.9 are all potential causes of systematic errors in \( \hat{\phi}(t) \). Finding a proper set of basis functions is a necessary condition for a successful flux-synthesis.

Another type of error in the procedure is the energy-group collapse: In Eq. 2.9, the mixing coefficients, \( T^{(k)} \), are assumed to be the same for all energy groups. This approximation could be relaxed if energy-selective neutron-detectors were used for measurements of fast and thermal neutrons separately.

Once reactor conditions have been chosen for the generation of expansion functions \( \psi^{(k)} \), errors may arise from the calculation of these \( \psi^{(k)} \)'s. Sources of errors are:

(i) The static, nodal, neutronic, thermal-hydraulic, and feedback models which contain approximations,

(ii) The various parameters used in those models which are not known exactly, and

(iii) The iterative procedure used to solve the equations, which does not yield a perfectly-converged numerical solution. Normally, this last
type of error can be controlled and is negligibly small (0.1 % or less) with respect to the others (~ a few percent).

Independently of the $\psi^{(k)}$’s, the quality of the reconstructed fluxes still depends on the vector of mixing coefficients, $T(t)$. Errors may arise from either $A$, $C(t)$, or the least-squares procedure used to solve Eq. 2.14.

Errors in the $\Sigma^{(j)}$’s (detector response-function and subsequent homogenization calculations), and in the expansion functions, $\psi^{(k)}$, are responsible for uncertainties in the $A$ matrix.

Measurement noise introduces random errors in $C(t)$. Systematic errors also enter $C(t)$ if one decides to consider the $\Sigma^{(j)}$’s as exact.

Finally, numerical difficulties associated with the least-squares solution of Eq. 2.14 may amplify roundoff errors and seriously affect $T(t)$. Such difficulties can be expected because the $A$ matrix in Eq. 2.14 is likely to be very ill-conditioned, i.e. almost rank-deficient, because of near linear-dependences between some of its columns or some of its rows. Linear dependences between columns of $A$ occur when some of the expansion functions are linear combinations of other expansion functions (Eq. 2.15b). Linear dependences between rows may happen because of symmetries in both the detector locations and core composition-pattern (Eq. 2.15a). Unless special precautions (to be specified in Chapter 3) are taken, ill-conditioning can make the solution of Eq. 2.14 completely meaningless [K13].

In view of all these difficulties, one may legitimately wonder whether the proposed method has any chance of success at all. Remarkably, subsequent chapters will show that it can indeed be successful, and that, more importantly, necessary conditions for this success can be specified.

Priority will be given to the most serious difficulties, i.e.:
(i) Possible very large errors caused by numerical roundoff in the least-squares solution of Eq. 2.14,
(ii) Systematic errors caused by the synthesis approximation, and
(iii) Random errors caused by measurement noise.
It must first be shown that these difficulties can be overcome before considering other types of problems such as systematic errors resulting from an imperfect instrumentation.

Accordingly, the $\Sigma^{(i)}$'s are assumed to be perfectly known in the following. This amounts to postulating that the instrumentation is totally reliable and provides unbiased measurements. Such ideal conditions can be approached by an accurate evaluation of the detector response-functions (including appropriate corrections), careful homogenization calculations, frequent updates, and the combined use of fault-tolerant sensors and an efficient signal-validation algorithm [E6, B17].

Part of the challenge is to select the expansion functions (number K and suitable reactor-conditions) and the detectors (number J and positions) so as to minimize the errors in the reconstructed fluxes. Since this challenge is problem dependent, one cannot expect to find an ideal choice of these free parameters for all situations. In addition, there is little flexibility in the detector number and positions once the instrumentation is in place. However, one may be able to find a choice which gives satisfactory results for practical purposes.

2.4. Reactivity Calculation by Inverse Kinetics.

From the vector $\hat{\phi}(t)$ of nodal group-fluxes reconstructed from Eq. 2.9, a number of other, important, time-dependent quantities can be readily evaluated. These include the three-dimensional, nodal, core power-distribution, the value and location of the maximum heat-generation rate, axial flux difference, axial offset, quadrant tilts, power peaking factors, total power, and point-kinetics parameters [G17, A1, H10, H8].
In this investigation, only two integral quantities will be of interest: total fission-power,

\[ \tilde{P}(t) = 3.204 \times 10^{-11} \sum_{g=1}^{G} \sum_{n=1}^{N} \Sigma_{f_{gn}}(t) \hat{\phi}_{gn}(t) V_n, \quad (2.17) \]

and reactivity.

The latter can be determined by "inverse kinetics". The procedure consists in first evaluating the amplitude function,

\[ \tilde{n}_{\text{eff}}(t) \equiv \frac{1}{K_0} \sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{gn}^* \frac{1}{V_{gn}} \hat{\phi}_{gn}(t) V_n, \quad (2.18) \]

effective delayed-neutron fractions,

\[ \tilde{\beta}_{\text{eff}}(t) \equiv \frac{\sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{gn}^* \chi_{ig} \Sigma_{f_{g'_{gn}}}(t) \hat{\phi}_{g'_{gn}}(t) V_n}{\sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{gn}^* \chi_{ig} \sum_{g'=1}^{G} \Sigma_{f_{g'_{gn}}}(t) \hat{\phi}_{g'_{gn}}(t) V_n}, \quad i = 1,2,...,I \quad (2.19) \]

and prompt neutron lifetime,

\[ \tilde{\Lambda}(t) \equiv \frac{\sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{gn}^* \frac{1}{V_{gn}} \hat{\phi}_{gn}(t) V_n}{\sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{gn}^* \chi_{gn} \sum_{g'=1}^{G} \Sigma_{f_{g'_{gn}}}(t) \hat{\phi}_{g'_{gn}}(t) V_n}, \quad (2.20) \]

using the reconstructed flux-vector \( \hat{\phi}(t) \).
In the above formulas, \( \phi_{a gn}^* \) are preselected weights and \( K_0 \) is an arbitrary, positive, normalization constant. The neutron speeds, \( v_{gn} \), delayed and total fission spectra, \( \chi_{i gn} \) and \( \chi_{g n} \), and fission cross sections, \( \Sigma_{fgn}(t) \), are equivalent, homogenized, nodal, few-group parameters.

The weights \( \phi_{a gn}^* \) may be chosen as adjoint fluxes corresponding to a particular, critical state of the reactor. However, other choices are equally acceptable [H8].

From the values of amplitude function, prompt neutron lifetime, and effective delayed-neutron fractions, effective precursor concentrations, \( \hat{c}_{i eff}(t) \), can be computed by solving the precursor point-kinetics equation (a weighted integral form of Eq. 1.3):

\[
\frac{d}{dt} \hat{c}_{i eff}(t) + \lambda_i \hat{c}_{i eff}(t) = \frac{\hat{\beta}_{i eff}(t)}{\Lambda(t)} \hat{n}_{eff}(t), \quad i = 1,2,\ldots,I, \tag{2.21}
\]

subject to known initial conditions. In practice, this can be done by first integrating Eq. 2.21 formally between \( t - \Delta t \) and \( t \) to obtain

\[
\hat{c}_{i eff}(t) = \exp(-\lambda_i \Delta t) \hat{c}_{i eff}(t - \Delta t)
\]

\[
+ \int_{t - \Delta t}^{t} \exp\left[-\lambda_i (t - \tau)\right] \frac{\hat{\beta}_{i eff}(\tau)}{\Lambda(\tau)} \hat{n}_{eff}(\tau) \, d\tau, \quad i = 1,2,\ldots,I \tag{2.22}
\]

Successive integrations by parts of the convolution integral lead to
\[ \tilde{c}_{\text{eff}}(t) = \exp(-\lambda_i \Delta t) \tilde{c}_{\text{eff}}(t - \Delta t) \]

\[ + \frac{1}{\lambda_i} \left[ \left( \frac{\tilde{\beta}_{\text{eff}}}{\tilde{\Lambda}} \tilde{n}_{\text{eff}} \right)_t - \exp(-\lambda_i \Delta t) \left( \frac{\tilde{\beta}_{\text{eff}}}{\tilde{\Lambda}} \tilde{n}_{\text{eff}} \right)_{t - \Delta t} \right] \]

\[ - \frac{1}{\lambda_i^2} \left[ \frac{d}{dt} \left( \frac{\tilde{\beta}_{\text{eff}}}{\tilde{\Lambda}} \tilde{n}_{\text{eff}} \right)_t - \exp(-\lambda_i \Delta t) \frac{d}{dt} \left( \frac{\tilde{\beta}_{\text{eff}}}{\tilde{\Lambda}} \tilde{n}_{\text{eff}} \right)_{t - \Delta t} \right] \]

\[ + ... \quad (2.23) \]

Formulas of arbitrary accuracy for \( \tilde{c}_{\text{eff}}(t) \) can be easily derived from Eq. 2.23 by Taylor expanding \( \frac{\tilde{\beta}_{\text{eff}}}{\tilde{\Lambda}} \tilde{n}_{\text{eff}} / \tilde{\Lambda} \) and its derivatives about \( t \) or \( t - \Delta t \). A first-order formula is obtained by neglecting the terms of order \( \Delta t^2 \) and higher. The result may be expressed as

\[ \tilde{c}_{\text{eff}}(t) = \exp(-\lambda_i \Delta t) \tilde{c}_{\text{eff}}(t - \Delta t) \]

\[ + \frac{1}{\lambda_i} \left[ 1 - \exp(-\lambda_i \Delta t) \right] \frac{\tilde{\beta}_{\text{eff}}(t - \Delta t)}{\tilde{\Lambda}(t - \Delta t)} \tilde{n}_{\text{eff}}(t - \Delta t) + o(\Delta t^2), \quad (2.24a) \]

or alternatively as

\[ \tilde{c}_{\text{eff}}(t) = \exp(-\lambda_i \Delta t) \tilde{c}_{\text{eff}}(t - \Delta t) \]

\[ + \frac{1}{\lambda_i} \left[ 1 - \exp(-\lambda_i \Delta t) \right] \frac{\tilde{\beta}_{\text{eff}}(t)}{\tilde{\Lambda}(t)} \tilde{n}_{\text{eff}}(t) + o(\Delta t^2). \quad (2.24b) \]

Finally, the solution, \( \tilde{c}_{\text{eff}}(t) \), obtained from Eq. 2.24a or 2.24b can be substituted into the neutron point-kinetics equation to yield an estimate for the instantaneous reactivity, \( \tilde{\rho}(t) \):

\[ \tilde{\rho}(t) = \tilde{\beta}_{\text{eff}}(t) + \frac{\tilde{\Lambda}(t)}{\tilde{n}_{\text{eff}}(t)} \left[ \frac{d}{dt} \tilde{n}_{\text{eff}}(t) - \sum_{i=1}^{1} \lambda_i \tilde{c}_{\text{eff}}(t) - q_{\text{eff}}(t) \right] \quad (2.25) \]
where

\[
\tilde{\beta}_{\text{eff}}(t) = \sum_{i=1}^{I} \tilde{\beta}_{i\text{eff}}(t).
\]  

(2.26)

Although, strictly speaking, \(\tilde{\beta}_{i\text{eff}}(t)\) and \(\tilde{\Lambda}(t)\) are time-dependent, numerical experience shows that the effects of this dependence on \(\tilde{\rho}(t)\) are so small that they can be legitimately ignored in practice. This is an important observation since it implies that, for the purpose of evaluating reactivity, variations in the fission cross sections in Eqs. 2.19 and 2.20 may be neglected, i.e. that the thermal-hydraulic feedback effects responsible for these variations need not be taken into account, a tremendous simplification. (Note that feedback effects are reflected in the \(\tilde{\phi}(t)\) vector from the synthesis.)

The term \(\frac{d}{dt} \tilde{n}_{\text{eff}}(t)\) in Eq. 2.25 may be approximated by

\[
\frac{d}{dt} \tilde{n}_{\text{eff}}(t) \approx \frac{\tilde{n}_{\text{eff}}(t) - \tilde{n}_{\text{eff}}(t - \Delta t)}{\Delta t},
\]  

(2.27)

a first-order formula which is consistent with the approximation made in evaluating the precursor concentrations (Eq. 2.24).

The effective external source, \(q_{\text{eff}}(t)\), when not negligible, must be estimated. If the distribution of nonfission sources is known, then \(q_{\text{eff}}(t)\) can be calculated from its definition:

\[
q_{\text{eff}}(t) = \frac{1}{K_0} \sum_{g=1}^{G} \sum_{n=1}^{N} \phi_{g \text{eff}}^{*} Q_{g \text{n}}(t) V_{n}.
\]  

(2.28)
During reactor operation at high power, for not too fast transients, the terms \( \frac{d}{dt} \hat{n}_{\text{eff}}(t) \) and \( q_{\text{eff}}(t) \) in the right-hand side of Eq. 2.25 are negligibly small with respect to the precursor term, \( \Sigma \lambda_i \hat{c}_{i\text{eff}}(t) \).

From the standpoint of estimating reactivity, the overall procedure leading to Eq. 2.25 may be viewed as a sophisticated reactivity-meter which accounts for changes in both flux level and flux shape.

Before turning to Chapter 3 and to the least-squares solution of Eq. 2.14, a brief review of neutron detectors used in commercial light-water reactors is given, mainly for completeness, but also because it is important not to lose sight of practical considerations. Additional details can be found in the references, in particular in References K14, S14, and M3.


2.5.1. Neutron-Flux Detectors.

Various types of neutron-detection devices are used in light-water reactors to monitor the (thermal) neutron-flux distribution inside [S14] and outside [M3] the core. Generally, these detectors have the form of thin cylinders with lengths ranging from centimeters to meters. Some in-core detectors are fixed for continuous measurements while others are movable for periodic measurements or calibrations of fixed detectors. Fixed in-core detectors are usually arranged in vertical strings of four or more, equally-spaced and positioned either between fuel bundles (BWRs) or in thimbles within fuel assemblies (some PWRs). Coaxial cables assure the electrical connection between detector and signal-processing equipment.

Measuring the local neutron-flux in the core of reactors is a very challenging task in general because of very hostile ambient conditions (high...
level of neutron and gamma radiations, high temperature and pressure, corrosive environment, ...) [C7] and because of geometrical constraints (small clearances).

The following is a nonexhaustive list of requirements or desirable characteristics for neutron sensors:
- Small size, i.e. reduced length for good spatial resolution, and small diameter for compatibility with the space limitations;
- For a given detector-size, high signal-to-background ratio, i.e. a relatively high neutron-sensitivity. Sensitivity is a very important, practical characteristic of any detector. The absolute neutron-sensitivity is defined as the ratio of the detector output signal to a unit incident neutron flux [B14, J3]. Many factors influence sensitivity, in particular the neutron energy-spectrum. Calculations (by electron transport codes) must take into account all these factors [J3, H2, N3]. For comparisons with measurements in reactors, gamma-sensitivity calculations may have to be performed [G8];
- Long lifetime (≥ 1 year), especially for fixed in-core detectors.

This implies slow burnup-rate of the neutron sensitive material so as to maintain a neutron relative sensitivity above an acceptable minimum value;
- Small or well-defined changes in characteristics (sensitivity, burnup, spectral response,...) with time to eliminate or reduce the need for corrections or calibrations;
- Large range of (linear) operation;
- High reliability despite the severity of the environment;
- Fast response for applications to real-time control and protection;
- Ease to manufacture and operate;
- Low cost;
- Insensitivity to vibrations.

Note that there is no "ideal" detector since some of the above requirements are conflicting, for instance high sensitivity and slow burnup rate or small size.

On the basis of these requirements, only two types of detectors are acceptable in practice: ionization chambers [S14, F10, B15, K14] and self-powered neutron detectors, also called SPNDs or collectrons [W3, S4, D10, T8, J9, K18, V3, B18, G9].
a. Fission Chambers.

Ionization chambers are the most common type of neutron sensors. They are almost always fission chambers. In PWRs, boron-lined ion chambers (BF$_3$ proportional counters) are also used, but their rapid burnout limits their use to the startup range.

Fission chambers are small ion chambers, typically 6 mm in case diameter and 2.5 cm in active length. Neutron-induced fissions in the U$_3$O$_8$-coated electrode (highly enriched in uranium 235) cause gas ionization. A pulse or current results from the collected charges. The collection time is very short, typically less than 1 microsecond. The gas is usually argon at high pressure for increased sensitivity. The polarization voltage must be high enough, usually between 50 V and 300 V, to maintain saturated operation.

The sensitivity of a miniature fission-chamber is typically $10^{-17}$ A / (neutrons cm$^{-2}$ s$^{-1}$). This relatively high value is the main advantage of fission chambers over self-powered neutron detectors. Another advantage is that fission chambers obviate the need for any spectral correction as their uranium content depletes because their spectral response follows that of the fuel. Disadvantages include manufacturing difficulties, insufficient reliability, and high cost.

With variations in their design, fission chambers can operate in different modes depending on the power range to be covered:

- Pulse mode in the source range;
- Mean-current (DC) mode, or ‘Mean-Square Variance’ (MSV) mode (chamber current squared for enhanced neutron-to-gamma response) [S14, O5, B10] in the intermediate range;
- Mean-current mode in the power range.

Gamma compensation is necessary in the source and intermediate ranges.

The time constants of signal-conditioning circuits and processing equipment (amplification, multiplexing, analog-to-digital conversion,
filtering, validation [M7]) are longer than the charge collection-times. Therefore, for sufficiently high incident fluxes, these time constants determine the maximum sampling-rate or minimum response-time of the instrumentation. In current mode, response times of 0.01 s are quoted in the literature [P5]. In pulse mode, the acquisition time may be relatively long when the local flux-level is very low. As a consequence, response times of several seconds are not unusual when a minimum counting-rate is imposed for statistical reasons.

b. Self-Powered Neutron Detectors.

Self-powered neutron detectors are simpler than, and not as expensive as, fission chambers. However, their sensitivity is much lower, typically $10^{-20}$ A / (neutrons cm$^{-2}$ s$^{-1}$) or less, and parasitic effects such as gamma-ray interactions in the connecting coaxial cable may be a significant source of background noise. In addition, spectral corrections must generally be performed.

SPNDs are used only in current mode. They are of solid construction. There is no gas-filled region nor applied external voltage. The outer diameter is typically 1 mm and the active length 20 cm or more. SPNDs are generally nonmovable detectors located in the core. They were developed mainly for PWRs [W4, W5]. Tests in several commercial reactors have shown that some SPNDs can operate satisfactorily for several years.

SPNDs can be of two different types: delayed-responding SPNDs or prompt-responding SPNDs.

Delayed-responding SPNDs use the beta-decay current following neutron capture in their emitter (central wire). The current is proportional to the net beta-escape rate from the emitter. The metallic outer shell is the collector. It is made of stainless steel or Inconel. Magnesium oxide or aluminium oxide forms the insulator between emitter
and collector. The emitter material should have a large neutron-capture cross section and the half-life for beta decay of the excited nucleus should be short. The beta rays from the \((n, \beta^-)\) reaction should also have enough energy to be able to reach the collector without being absorbed in the emitter or in the insulator.

In practice, only two materials are used as emitters in delayed-responding SPNDs:

- Rhodium 103, \(^{103}\)Rh, which gives either one of two radioisotopes of \(^{104}\)Rh upon absorption of a neutron. The 2200 m/s absorption cross section, \(\sigma_a\), is 139 b for one of them (ground state) and 11 b for the other (metastable). The ground state of \(^{104}\)Rh is characterized by a half-life \(T_{1/2}\) of 42 s, whereas the metastable state has \(T_{1/2} = 264\) s [A6];

- Vanadium 51, \(^{51}\)V, which gives \(^{52}\)V with only one beta-decay mode characterized by \(\sigma_a = 4.9\) b, \(T_{1/2} = 226\) s.

Both types of detectors are extensively used is spite of their delayed response. Typical dimensions are:

- Emitter diameter \(\sim 0.5\) mm;
- Outer shell diameter \(\sim 1.5\) mm;
- Active length \(\sim 40\) cm for \(^{103}\)Rh, 100 cm for \(^{51}\)V.

Calibrations can be performed automatically (by computer) because of well-established sensitivity depletion curves [D2].

Prompt-responding SPNDs use the \((n,\gamma)(\gamma,\epsilon)\) reactions in their emitter material. A fraction of the secondary (photo- or Compton) electrons penetrate the surrounding insulator and reach the collector, giving rise to a current.

In general, the neutron sensitivity of such detectors is considerably less than that of delayed-responding SPNDs of identical size because of a much lower neutron-to-electron conversion efficiency. However, their response is essentially instantaneous, which makes them suitable for safety-related functions. The output current of prompt-responding SPNDs is a composite signal because of contributions to the \((\gamma,\epsilon)\) reactions from (i) external gamma rays (prompt gamma-rays from fission reactions and delayed gamma-rays from radioactive fission products), and possibly (ii)
delayed betas following the neutron-capture reaction. (In fact, rhodium and vanadium delayed SPNDs have a small prompt component.) Special procedures must be employed to recover the neutron flux from the composite signal [M5, L12, E4, Y5, Y6]. Frequent, if not continuous, calibrations are necessary when the variations in neutron sensitivity with exposure are not well known. (Long-term exposure can give rise to delayed currents from long-lived radioisotopes.)

The following materials have been considered as possible emitters for fast-responding SPNDs:
- cobalt 59 (\(\sigma_a = 37\) b) [B13, J6, B14],
- cadmium 112 [J3],
- platinum 195 [S3, L13, G8],
- molybdenum 95 (\(\sigma_a = 14\) b) [K18],
- hafnium 181 (\(\sigma_a = 106\) b) [W4, W5, T7],
- and even uranium 235 surrounded by a very thin insulator [B2, B16].

As with fission chambers, the response time is limited by the electronics associated with the detectors. Values of 0.1 s have been reported [T7].

2.5.2. Reactor Nuclear Instrumentation.

In Westinghouse PWRs, on-line neutron-flux monitoring relies on detectors located outside the reactor vessel where the ambient conditions are less severe [G1, G13, P4, G16]. These detectors are fairly long (length > 1 m), and therefore have poor spatial resolution. Their primary function is the detection of any gross, radial or axial unbalance in flux shape, such as would occur during xenon oscillations. Three sets of such detectors are used, with overlapping ranges:
- For the source (startup) range, a set of two gamma-compensated fission chambers operating in pulse mode, or BF₃ proportional counters;
For the intermediate range, a set of two fission chambers operating in DC-current or MSV mode;

For the power range, a set of four monitors, each composed of two long, uncompensated, ion fission-chambers operating in DC-current mode.

At present, in most plants, there are no fixed in-core detectors. In-core power monitoring is accomplished by periodic flux-measurements by means of small, movable fission-chambers [B1, P4, G16]. Calibrations rely on heat-balance calculations.

A notable exception is the Seabrook nuclear power plant in which an experimental, in-core power-monitoring system is being tested [G12, D6]. This system relies on 58 strings of fixed, in-core, platinum SPNDs. Each string contains five detectors. The originality of this detection system comes from the fact that the platinum detectors are not used as neutron detectors, but instead as gamma detectors. Although most materials used as emitters in SPNDs are primarily neutron sensitive, platinum is primarily (~80%) gamma sensitive [G10, O3]. Gamma-flux measurements are converted into neutron-flux values by a "numerical correlation" involving a 2-D assembly lattice code and a 3-D nodal code. Specifically, the (prompt and delayed) gamma flux at each detector location is calculated by the lattice code (CASMO-3) using 18 gamma energy-groups. This calculation makes use of detector-sensitivity values obtained from Monte-Carlo calculations [N3]. The lattice code also gives neutron-flux values at the detector location. Total gamma-flux, gamma-sensitivity, and neutron-flux values are then combined to determine equivalent, two-group, neutron cross sections in the detector region as a function of typical core parameters. These equivalent neutron cross sections are subsequently used by the nodal code (SIMULATE-3) to reconstruct local neutron group-fluxes at the detector locations. Predictions can be made for the detector outputs and compared with the actual signals. The time-response of the system is on the order of one minute. Early results seem to indicate that the system can give very accurate results in spite of a lack of spatial resolution caused by the gamma detectors [O3].

Babcock & Wilcox PWRs are equipped with 364, fixed, in-core, rhodium, self-powered neutron detectors distributed in 52 strings, each
composed of seven detectors [R4, H5, R6, R7]. Scanning is performed every 30 seconds. These delayed-responding detectors are soon to be replaced by prompt-responding hafnium detectors, or, more precisely, by pairs of prompt-responding hafnium detectors (used for neutron detection) and delayed-responding rhodium detectors (used for on-line calibration of the hafnium detectors) [W4, W5].

In Combustion Engineering PWRs, there are 33 instrumented assemblies, each containing four or five, fixed, 40-cm long, rhodium detectors [H6, O1, O2, T4, T5].

German, Siemens/Kraftwerk-Union PWRs are equipped with 48 strings of six, fixed, cobalt prompt SPNDs (‘Power Density Detection System’), calibrated at regular intervals by the ‘Aerobal’ flux-mapping system [E5]. At least one other German PWR uses fixed in-core rhodium detectors [W11].

In Canadian, 600 MWe, pressurized, heavy-water-moderated reactors, there are two sets of fixed in-core detectors mounted horizontally and vertically [A5, L7]:
- A set of 85, small, prompt-responding, platinum SPNDs, some used for local flux control, others for protection;
- And a set of 102, long, vanadium SPNDs used both for flux mapping and calibration purposes.

In General Electric BWRs, all detectors are fixed fission-chambers located in the core [M6, F7, G5]. There are (see Fig. 2.2):
- Four chambers operating in pulse mode to cover the source range (SRM);
- Eight chambers operating in MSV mode to cover the intermediate range (IRM);
- Between 144 and 172 chambers positioned in four equally-spaced planes and operating in DC-current mode to cover the power range (LPRM). Periodic calibration of the LPRM system is performed by means of a traversing in-core probe (TIP).
Figure 2.2. Ranges of In-Core Fission Chambers in BWRs.  
(Adapted From References G5 and K14.)

The above presentation, although succinct, should make it clear that:  
(i) Small, reliable, fast-responding neutron detectors for in-core use  
are available commercially, either in the form of fission chambers or self-  
powered neutron detectors.  
(ii) These detectors have a limited range of operation and must  
generally be coupled with some means of on-line or periodic calibration;  
(iii) Many commercial light-water reactors are equipped with some  
type of fixed, in-core neutron sensors [S15].

Some plants still lack such sensors. However, there seems to be an  
increasing awareness of the need for a fixed in-core instrumentation,  
preferably of the prompt-responding type [R7, W6]. Advanced, real-time,  
optimum-control systems envisioned for future plants will require such a  
prompt instrumentation [K11, U1].
2.6. Summary.

In this chapter, the general characteristics of a semi-experimental nodal synthesis method have been presented. The terminology "semi-experimental" reflects the fact that the method combines (off-line) theoretical calculations with (on-line) experimental measurements from flux detectors. The method may also be called "fitted synthesis" since it consists in a fit of a linear combination of precomputed, three-dimensional expansion functions to local flux measurements. It was shown how reactivity can be inferred from the reconstructed flux distribution with essentially no need for any extra information concerning the thermal-hydraulic conditions of the plant. Finally, a brief review of neutron instrumentation used in commercial light-water reactors was given.

A number of potential sources of difficulties and errors with the fitted-synthesis method were identified. Of particular concern are systematic errors caused by the synthesis approximation, numerical difficulties associated with the solution of the least-squares problem, and random errors caused by noise in the instrumentation. The first two of these problems are addressed in Chapter 3 and 4, whereas the third one is discussed in detail in Chapter 5.
Chapter 3

MINIMUM-NORM LEAST-SQUARES SOLUTION
OF ILL-CONDITIONED RECTANGULAR SYSTEMS
OF LINEAR ALGEBRAIC EQUATIONS

3.1. Introduction.

The general description of a semi-experimental synthesis method for reconstructing the instantaneous neutron-flux distribution in light-water reactors was given in Chapter 2. The essence of the method is to approximate the actual neutron-flux by a linear combination of three-dimensional expansion-functions obtained from steady-state nodal calculations. By requiring that the K coefficients of the combination agree in a least-squares sense with J local flux-measurements, a rectangular system of linear algebraic equations, \( \mathbf{A} \mathbf{T} = \mathbf{C} \), is obtained. This system must be solved for the vector of mixing-coefficients, \( \mathbf{T} \), given a measurement vector, \( \mathbf{C} \). A potentially serious numerical difficulty is that the J-by-K matrix \( \mathbf{A} \) may be ill-conditioned as a result of near linear-dependences between expansion functions. This ill-conditioning can make the solution-vector meaningless by amplifying roundoff-errors or measurement noise.

This chapter contains a review of the general characteristics and properties of rectangular systems of equations. Possible solution-techniques are then presented. The choice of an orthogonalization method using the 'Singular Value Decomposition' of \( \mathbf{A} \) is justified. This method is described in great detail and is then specialized to the particular system of equations obtained in Chapter 2. A physical interpretation is given of the
underlying orthogonalization procedure and of the technique which consists in zeroing small singular values to remedy the ill-conditioning problem.

Simple geometrical arguments are used extensively throughout the presentation to facilitate the comprehension. Familiarity with basic matrix-theory is assumed.

3.2. Rectangular Systems of Linear Algebraic Equations.

3.2.1. The Geometrical Approach.

Any rectangular, J-by-K, real matrix $\mathbf{A}$ can be interpreted as a linear map from $\mathbb{R}^K$ to $\mathbb{R}^J$: Any vector $\mathbf{T} \in \mathbb{R}^K$ is transformed by $\mathbf{A}$ into an image vector $\mathbf{A} \mathbf{T} \in \mathbb{R}^J$. Solving $\mathbf{A} \mathbf{T} = \mathbf{C}$ is equivalent to finding a vector $\mathbf{T} \in \mathbb{R}^K$ whose image $\mathbf{A} \mathbf{T}$ by the $\mathbf{A}$ mapping coincides with the given vector $\mathbf{C} \in \mathbb{R}^J$, or, if not possible, lies as close as possible to $\mathbf{C}$. This interpretation is central to the forthcoming developments.

The nature of a rectangular system of linear algebraic equations, $\mathbf{A} \mathbf{T} = \mathbf{C}$, depends on the number of rows, J, number of columns, K, and $\text{ran}^\perp$, R, of the coefficient matrix $\mathbf{A}$. Three general cases can be distinguished [L4]:

(i) If $J < K$, $\mathbf{A} \mathbf{T} = \mathbf{C}$ is a system with more unknowns than equations. In general, $R = J$ and such an underdetermined system is consistent, i.e. it has an infinite number of solutions. Among all these solutions and in the absence of any constraint, the "best compromise" is obtained for that particular vector, $\mathbf{T}$, which has minimum Euclidean norm. Figure 3.1 is an illustration of such a situation for $J = 1$ and $K = 2$. Several solution-vectors are shown. They all have the same image, $\mathbf{A} \mathbf{T} = \mathbf{C}$, by the linear map $\mathbf{A}$. The selected solution-vector is the one which has the shortest length.
Figure 3.1. Minimum-Norm Solution of a One-by-Two System of Equations \( \mathbf{A} \mathbf{T} = \mathbf{C} \).

(ii) If \( J = K \), then \( \mathbf{A} \) is square. In general, \( R = J = K \) and \( \mathbf{A} \mathbf{T} = \mathbf{C} \) has a unique solution, \( \mathbf{T} = \mathbf{A}^{-1} \mathbf{C} \). Figure 3.2 is a geometrical interpretation of this case for \( J = K = 2 \).

Figure 3.2. Unique Solution of a Two-by-Two System of Equations \( \mathbf{A} \mathbf{T} = \mathbf{C} \).
(iii) If \( J > K \), \( \mathbf{A} \mathbf{T} = \mathbf{C} \) is a system with more equations than unknowns. In general, \( R = K \) and such an overdetermined system is inconsistent: It has no solution. In these conditions, one is usually interested in the least-squares solution-vector, i.e. in that particular vector \( \mathbf{T}_{LS} \) which is such that \( \mathbf{A} \mathbf{T}_{LS} \) gives "the best approximation of \( \mathbf{C} \)" in \( \mathbb{R}^J \). A more precise definition of \( \mathbf{T}_{LS} \) is "the vector of \( \mathbb{R}^K \) which minimizes the Euclidean norm of the vector \( \mathbf{C} - \mathbf{A} \mathbf{T} \). Figure 3.3 is an illustration of this case for \( J = 3 \) and \( K = 2 \).

![Diagram](image)

Figure 3.3. Least-Squares Solution of a Three-by-Two System of Equations \( \mathbf{A} \mathbf{T} = \mathbf{C} \).

The three, "nice", general situations described above correspond to a matrix \( \mathbf{A} \) that has full rank: \( R = \min \{ J, K \} \). In the case of the actual system of equations to be solved (Eq. 2.14), however, it was pointed out in Section 2.3.2 that linear-dependences between columns are to be expected, i.e. that \( \mathbf{A} \) is likely to be rank deficient: \( R < \min \{ J, K \} \).
Rank degeneracies in a J-by-K, real matrix $\mathbf{A}$ can be best described by using the mathematical notion of *null-space* (or kernel) of $\mathbf{A}$. In the case of square matrices, null-space and range are familiar concepts. For rectangular matrices, there are not two, but four fundamental subspaces to consider, two in $\mathbb{R}^K$ and two in $\mathbb{R}^J$ [S19]. These four vector-spaces are:

(i) $\mathbf{R}(\mathbf{A}^T) = \text{row space of } \mathbf{A} = \text{space spanned by the rows of } \mathbf{A}$;
(ii) $\mathbf{N}(\mathbf{A}) = \text{null-space of } \mathbf{A} = \text{orthogonal complement of } \mathbf{R}(\mathbf{A}^T)$ in $\mathbb{R}^K$;
(iii) $\mathbf{R}(\mathbf{A}) = \text{column space or range of } \mathbf{A} = \text{space spanned by the columns of } \mathbf{A}$;
(iv) $\mathbf{N}(\mathbf{A}^T) = \text{left null-space of } \mathbf{A} = \text{orthogonal complement of } \mathbf{R}(\mathbf{A})$ in $\mathbb{R}^J$,

and they verify:

$$\mathbf{R}(\mathbf{A}^T) \oplus \mathbf{N}(\mathbf{A}) = \mathbb{R}^K, \quad (3.1a)$$

$$\mathbf{R}(\mathbf{A}) \oplus \mathbf{N}(\mathbf{A}^T) = \mathbb{R}^J, \quad (3.1b)$$

$$\dim \mathbf{R}(\mathbf{A}^T) = \dim \mathbf{R}(\mathbf{A}) = R = \text{rank}(\mathbf{A}), \quad (3.1c)$$

$$\dim \mathbf{N}(\mathbf{A}^T) = J - R, \quad (3.1d)$$

$$\dim \mathbf{N}(\mathbf{A}) = K - R. \quad (3.1e)$$

This decomposition of $\mathbb{R}^K$ and $\mathbb{R}^J$ into orthogonal subspaces is illustrated in Figure 3.4.
Figure 3.4. $\mathbf{A}$ as a Linear Map from $\mathbb{R}^K$ to $\mathbb{R}^J$.
(Adapted from References P6 and S19.)

Note that, when $\mathbf{A}$ is square, one needs to distinguish only two subspaces because $\mathbb{R}^K = \mathbb{R}^J$, $\mathbb{R}(\mathbf{A}^T) = \mathbb{R}(\mathbf{A})$, and $\mathbb{N}(\mathbf{A}) = \mathbb{N}(\mathbf{A}^T)$. For rectangular matrices, this picture unfolds into four subspaces as shown in Figure 3.4.

When $\mathbf{A}$ has full-rank as in the examples of Figures 3.1, 3.2, and 3.3, at least one null-space, $\mathbb{N}(\mathbf{A})$ or $\mathbb{N}(\mathbf{A}^T)$, reduces to $\{\Omega\}$. On the other hand, when $\mathbf{A}$ is rank-deficient, both $\mathbb{N}(\mathbf{A})$ and $\mathbb{N}(\mathbf{A}^T)$ are of dimension at least one. Rank degeneracies reduce the common dimension of both the row and column spaces of $\mathbf{A}$, and, at the same time, increase the dimensions of the null-spaces.

89
Since \( \mathbf{R}(\mathbf{A}^T) \) and \( \mathbf{N}(\mathbf{A}) \) are orthogonal complements in \( \mathbf{R}^K \) (Eq. 3.1a), any vector \( \mathbf{T} \) of \( \mathbf{R}^K \) can be written in a unique manner as the sum of a vector \( \mathbf{T}_R \in \mathbf{R}(\mathbf{A}^T) \) and a vector \( \mathbf{T}_N \in \mathbf{N}(\mathbf{A}): \\
\mathbf{T} = \mathbf{T}_R + \mathbf{T}_N, \\
\begin{cases} 
\mathbf{A}^T \mathbf{T}_R = \mathbf{A}^T \mathbf{T} \\
\mathbf{A}^T \mathbf{T}_N = \mathbf{0}.
\end{cases} \tag{3.2} \\

It can be seen from Eq. 3.2 and Figure 3.4 that:
(i) The null-space component of any vector \( \mathbf{T} \) of \( \mathbf{R}^K \) is mapped onto \( \mathbf{0} \) in \( \mathbf{R}^J \);
(ii) The addition to \( \mathbf{T} \) of an arbitrary vector belonging to \( \mathbf{N}(\mathbf{A}) \) does not affect the image vector \( \mathbf{A}^T \mathbf{T} \) in \( \mathbf{R}^J \);
(iii) The left null-space of \( \mathbf{A}, \mathbf{N}(\mathbf{A}^T) \), is not "reachable" by the \( \mathbf{A} \) mapping.

From these elementary observations, it follows that:
(i) '\( \mathbf{A}^T \mathbf{T} = \mathbf{C} \)' has no solution unless \( \mathbf{C} \in \mathbf{R}(\mathbf{A}) \);
(ii) an infinite number of vectors \( \mathbf{T} \) have the same image by \( \mathbf{A} \) unless \( \mathbf{N}(\mathbf{A}) = \{ \mathbf{0} \} \).

In the case of the system of equations '\( \mathbf{A}^T \mathbf{T} = \mathbf{C} \)' derived by the synthesis method of Chapter 2 (Eq. 2.14):
(i) In general, \( J > K \). (For reasons given later in this Chapter, it is desirable to have more detectors than expansion functions.) Therefore, \( \mathbf{C} \) is unlikely to lie entirely in \( \mathbf{R}(\mathbf{A}) \) (Figure 3.4). In these conditions, '\( \mathbf{A}^T \mathbf{T} = \mathbf{C} \)' should be expected to have no solution, and one will have to content oneself with a least-squares solution.
(ii) Because of the high probability of rank degeneracy, the null-space \( \mathbf{N}(\mathbf{A}) \) will not reduce to \( \{ \mathbf{0} \} \), and, as a consequence, the least-squares solution-vector will not be unique. One way to remove this ambiguity is to impose upon the least-squares solution to have minimal Euclidean norm, i.e. to have a null component, \( \mathbf{T}_N \), in the direction of \( \mathbf{N}(\mathbf{A}) \). This vector is
the minimum-norm, least-squares solution to ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’. Among all vectors minimizing \( \| \mathbf{C} - \mathbf{A} \mathbf{T} \| \), it is the one which has the shortest length.

In fact, the solution of any rectangular system of linear algebraic equations, underdetermined or overdetermined, degenerate or not, may be regarded as a ‘minimum-norm, least-squares solution’ in a broad sense. In the following, it will be convenient to use this terminology when referring to a general rectangular system without further specification.

3.2.2. The Algebraic Approach.

The preceding section showed that the concepts of column space, row space, and null-spaces shed considerable light upon the nature of any rectangular system of equations. However, the insight gained was mainly geometrical. In practice, one would also like to have formulas to compute the desired, minimum-norm, least-squares solution-vector.

To obtain such formulas rigorously, one must first prove a geometrically intuitive result (see Figure 3.3), namely that a least-squares solution is such that the vector \( \mathbf{C} - \mathbf{A} \mathbf{T} \) of \( \mathbb{R}^j \), when not zero, is perpendicular to \( \mathbb{R}(\mathbf{A}) \).

The proof makes use of the following, real, positive-definite function of \( \mathbf{T} \) [G11]:

\[
\varphi(\mathbf{T}) = \frac{1}{2} \| \mathbf{C} - \mathbf{A} \mathbf{T} \|^2.
\]  

(3.3)

with the convention that ‘\( \| \cdot \| \)’ represents the Euclidean norm.

Solving ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ is equivalent to minimizing \( \varphi(\mathbf{T}) \). Indeed, \( \varphi(\mathbf{T}) \) is zero for an exact solution-vector, \( \mathbf{T} \), and is otherwise minimum for a least-squares solution, \( \mathbf{T}_{LS} \) (which, by definition, minimizes \( \| \mathbf{C} - \mathbf{A} \mathbf{T} \| \)).
Although other norms could be considered in Eq. 3.3, the choice of the Euclidean norm makes the algebra simple since \(\varphi(T)\) is then a differentiable function of \(T\).

With the Euclidean norm, Eq. 3.3 may be written as

\[
\varphi(T) = \frac{1}{2} (C - A T)^T (C - A T)
= \frac{1}{2} T^T (A^T A) T - T^T A^T C + \frac{1}{2} C^T C. \tag{3.4}
\]

The following (easily proven) lemma:

\[
P(T) = \frac{1}{2} T^T M T - T^T D \Rightarrow \nabla P(T) = \frac{1}{2} (M + M^T) T - D,
\]

may be applied to Eq. 3.4 with \(M = A^T A = M^T\) and \(D = A^T C\). The result is:

\[
\varphi(T) = P(T) + \frac{1}{2} C^T C
\]

and

\[
\nabla \varphi(T) = \nabla P(T) = A^T A T - A^T C = - A^T (C - A T). \tag{3.5}
\]

Setting \(\nabla \varphi(T)\) to zero then yields the minimization condition (a maximum being excluded since \(\varphi\) has no upper bound):

\[
A^T (C - A T) = 0. \tag{3.6}
\]

The solution of Eq. 3.6 is the set of vectors \(T\) satisfying

\[
C - A T \in N(A^T). \tag{3.7}
\]
Eq. 3.7 implies that $C - AT$ is either $Q$ or is perpendicular to $R(A)$ when $T = T_{LS}$, as one set out to prove.

Because of its importance, the vector

$$E(T) = C - AT$$

has been given a name: It is called the \textit{residual vector}. When $T = T_{LS}$, it is denoted $E_{LS}$. Its Euclidean norm, $\|E_{LS}\|$, is then minimum and gives an \textit{absolute} measure of the distance of $C$ from the projection "plane", $R(A)$, i.e. of the quality of the least-squares fit (see Figure 3.5). In a good fit, the components of the $E_{LS}$ vector should be small compared with those of the $C$ vector, and $\|E_{LS}\|$ should be small with respect to $\|C\|$.

The \textit{relative residual}, defined as

$$\frac{\|E(T)\|}{\|C\|} = \frac{\|C - AT\|}{\|C\|},$$

is also minimum, equal to $\|E_{LS}\| / \|C\| = \sin \theta$ in Figure 3.5, when $T = T_{LS}$. It is a \textit{relative} measure of the quality of the least-squares solution. It should be small with respect to unity.
Eqs. 3.6 and 3.7 can be used to establish well-known formulas for the minimum-norm and least-squares problems. This is done below using a systematic approach which includes all possible cases. The pivotal notions are again the null-spaces, $N(\Delta^T)$ and $N(\Delta)$.

There are only two possibilities with Eq. 3.7: Either
- (1) $N(\Delta^T) = \{0\}$, or else
- (2) $N(\Delta^T)$ is not reduced to $\{0\}$.

**Case 1:** $N(\Delta^T) = \{0\}$ requires both $J \leq K$ (underdetermination) and no rank degeneracy in $\Delta$, i.e. $R = J$. In this (unlikely) situation, one still has two alternatives: either $N(\Delta) = \{0\}$ or not.

- $N(\Delta)$ can be $\{0\}$ only if $J \geq K$, and since one already has $R = J \leq K$, one concludes that $J = K = R$, i.e. $\Delta$ is square and has full rank. There is a unique solution:

$$ T = \Delta^{-1} C. $$

(3.10)
Alternatively, if $N(A)$ is not reduced to $\{0\}$, then, of necessity, $R = J < K$ and \( \mathbf{A}^T \mathbf{C} = \mathbf{C} \) is an underdetermined, full-rank system. Among all its solutions, the one which has minimum norm is $\mathbf{T} = \mathbf{T}_R + \mathbf{T}_N$ with $\mathbf{T}_N = 0$, i.e. $\mathbf{T} = \mathbf{T}_R$. Since this vector belongs to $\mathbf{R}(\mathbf{A}^T)$, one can find a vector $\mathbf{C'} \in \mathbf{R}^J$ such that $\mathbf{T} = \mathbf{A}^T \mathbf{C'}$. From the fact that $\mathbf{A}^T \mathbf{C} = \mathbf{C}$, it follows that $\mathbf{A}^T \mathbf{C'} = \mathbf{C}$. The matrix $\mathbf{A}^T \mathbf{A}$ is a square, $J$-by-$J$, symmetric matrix. Its rank is equal to the rank of $\mathbf{A}$, $R = J$. Therefore, it is invertible, and one may write $\mathbf{C'} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{C}$. The final, minimum-norm solution may be written as

$$\mathbf{T} = \mathbf{A}^T (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{C}. \quad (3.11)$$

**Case 2:** $N(\mathbf{A}^T)$ is not reduced to $\{0\}$. This occurs whenever $J > K$ (overdetermined system) or when there are rank-degeneracies in $\mathbf{A}$. In the particular situation where there are no rank deficiencies, then $R = K < J$. This is the full-rank, overdetermined case. Using Eq. 3.5, one may write $$(\mathbf{A}^T \mathbf{A}) \mathbf{T} = \mathbf{A}^T \mathbf{C}. \quad (3.10)$$ The matrix $\mathbf{A}^T \mathbf{A}$ is a square, $K$-by-$K$, symmetric matrix. Its rank is equal to the rank of $\mathbf{A}$, $R = K$. It is therefore invertible, and the final, least-squares solution-vector may be written as

$$\mathbf{T}_{LS} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}. \quad (3.12)$$

Note that $\mathbf{A}^T (\mathbf{A}^T \mathbf{A})^{-1}$ and $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ are respectively right and left inverses of $\mathbf{A}$. Eq. 3.10 may be viewed as a limiting case of Eqs. 3.11 and 3.12 since, when $\mathbf{A}$ is square and invertible, Eqs. 3.11 and 3.12 reduce to Eq. 3.10.

No mention has been made thus far of any weighting in the rectangular system of equations to be solved. In fact, it has been implicitly assumed that \( \mathbf{A}^T \mathbf{C} = \mathbf{C} \) represents a "properly-weighted" system, i.e. that
any weighting matrix has been absorbed in the notation. In the light of Eqs. 3.11 and 3.12, however, a few comments are appropriate.

A least-squares solution to ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)' is best justified when the components of the right-hand-side vector, \( \mathbf{C} \), are equally reliable. (In these conditions, it can be shown that the least-squares solution has the desirable statistical property of being a “minimum-variance estimate”.) When this is not the case, the system ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)' should be *left*-multiplied or “row-weighted” by a \( J \)-by-\( J \) weighting matrix, \( \mathbf{W} \) (ideally such that \( \mathbf{W}^T \mathbf{W} \) is the covariance matrix associated with the measurement vector \( \mathbf{C} \)). This is equivalent to replacing \( \mathbf{A} \) by \( \mathbf{W} \mathbf{A} \), and \( \mathbf{C} \) by \( \mathbf{W} \mathbf{C} \).

On the other hand, it may be desirable in some cases to redefine or rescale the components of the unknown vector, \( \mathbf{T} \), i.e. to replace \( \mathbf{T} \) by \( \mathbf{W}' \mathbf{T} \) where \( \mathbf{W}' \) is a \( K \)-by-\( K \) matrix. This redefinition of the unknowns requires *right*-multiplication or “column-weighting” of the matrix \( \mathbf{A} \) by \( \mathbf{W}'^{-1} \) in ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’.

One can easily convince oneself from Eq. 3.11 and 3.12 that left-multiplication of ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ does not affect the minimum-norm solution of an underdetermined system. However, right-multiplication does. The converse is true for the least-squares solution of an overdetermined system. Therefore, any left- or right-multiplication of \( \mathbf{A} \) in ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’ requires care [L4, D8, G11].

Eqs. 3.11 and 3.12 are simple, well-known formulas for the minimum-norm and least-squares problems respectively. Unfortunately, neither of them is applicable when rank degeneracy occurs because the matrices \( \mathbf{A} \mathbf{A}^T \) and \( \mathbf{A}^T \mathbf{A} \) then become singular. However, the procedure by which Eqs. 3.11 and 3.12 have been obtained has evidenced the key role played by these square matrices, as well as by \( \mathbf{R} \), the rank of \( \mathbf{A} \).
3.3. Practical Methods for Solving Rectangular Systems of Linear Algebraic Equations.

In practice, there are three possible methods for solving rectangular systems of linear algebraic equations:
- The Normal Equations,
- QR Factorization, and
- Singular Value Decomposition.
Each of these methods is presented below.

3.3.1. The Normal Equations.

A standard technique for solving rectangular systems of linear algebraic equations consists in forming the so-called Normal Equations [L4, H1, S19, G11], i.e.:

\begin{align}
\text{• For } J < K, \quad & \begin{cases}
(i) \quad (A A^T) C' = C, \\
(ii) \quad T = A^T C';
\end{cases} \quad (3.13) \\
\text{• For } J \geq K, \quad & (A^T A) T = A^T C. \quad (3.14)
\end{align}

One immediately recognizes in these equations Eqs. 3.11 and 3.12 written in a “noninverted” form. From the discussion of the previous section, one concludes that the method of the Normal Equations is applicable only when there is no rank degeneracy in \(A\). Only then can the matrix \(A^T A\) (or \(A A^T\)) be inverted. (In practice, a Cholesky factorization is an efficient solution method.) The method fails in the rank-degenerate case.

As one might intuitively expect, it also fails in practice when \(A\) is “almost” rank-deficient, i.e. ill-conditioned. The reason for this is that
forming $\mathbf{A}^T \mathbf{A}$ or $\mathbf{A} \mathbf{A}^T$ numerically may lead to an irreparable loss of information. The following simple example illustrates this point:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ \varepsilon & 0 \end{bmatrix}, \quad 0 < \varepsilon \ll 1; \quad \mathbf{A}^T \mathbf{A} = \begin{bmatrix} 1 + \varepsilon^2 & 1 \\ 1 & 1 + \varepsilon^2 \end{bmatrix}.$$  

If $\varepsilon^2 < \varepsilon_m$, the machine accuracy, then the computer equates $1+\varepsilon^2$ to 1. As a result, $\mathbf{A}^T \mathbf{A}$ becomes numerically singular.

In practice, ill-conditioning is expected in the $\mathbf{A}$ matrices arising from the synthesis method (Eq. 2.9). This ill-conditioning cannot be detected by the method of the Normal Equations. Another method is therefore called for, preferably an “orthogonalization” method such as ‘QR Factorization’ or ‘Singular Value Decomposition’.

3.3.2. QR Factorization.

A least-squares solution to an overdetermined system of equations, ‘$\mathbf{A}^T \mathbf{A} = \mathbf{C}$’, can be obtained from a QR factorization of the J-by-K matrix $\mathbf{A}$ [L4, S16, D8, H1, S19, G11, G7]:

$$\mathbf{A} = \mathbf{Q} \mathbf{R} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \quad (3.15)$$

where $\mathbf{Q}$ is a J-by-J, orthogonal matrix, while $\mathbf{R}_1$ is a K-by-K, upper-triangular matrix. The column of the $\mathbf{Q}$ matrix are orthonormal vectors forming a basis of $\mathbf{R}^J$.

Such a factorization is not unique.
When the rank of $\mathbf{A}$, $\mathbf{R}$, is equal to $\mathbf{K}$, the matrix $\mathbf{R}_1$ is invertible and the least-squares solution of the overdetermined system $\mathbf{A} \mathbf{T} = \mathbf{C}$ is (without proof)

$$
\mathbf{T}_{LS} = \begin{bmatrix} \mathbf{R}_1^{-1} & \mathbf{0} \end{bmatrix} \mathbf{Q}^T \mathbf{C}.
$$

(3.16)

For underdetermined systems, a QR factorization of $\mathbf{A}^T$, instead of $\mathbf{A}$, is performed:

$$
\mathbf{A}^T = \mathbf{Q} \mathbf{R} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}
$$

(3.17)

where $\mathbf{Q}$ is now a $\mathbf{K}$-by-$\mathbf{K}$, orthogonal matrix, and $\mathbf{R}_1$ a $\mathbf{J}$-by-$\mathbf{J}$, upper triangular matrix. The column of the $\mathbf{Q}$ matrix form an orthonormal basis of $\mathbf{R}^\mathbf{K}$.

When $\mathbf{R} = \mathbf{J}$, $\mathbf{R}_1$ is invertible and the minimum-norm solution of the underdetermined system $\mathbf{A} \mathbf{T} = \mathbf{C}$ is (without proof)

$$
\mathbf{T} = \mathbf{Q} \begin{bmatrix} (\mathbf{R}_1^{-1})^T \mathbf{C} \\ \mathbf{0} \end{bmatrix}.
$$

(3.18)

When $\mathbf{A}$ is rank-deficient, $\mathbf{R}_1$ is singular and cannot be inverted. However, it is always possible, before factorizing $\mathbf{A}$ (or $\mathbf{A}^T$), to permute its columns so as to isolate a subset of $\mathbf{R}$ columns which are linearly independent. For example, in the overdetermined case, if the matrix with pivoted columns is denoted $\mathbf{A} \mathbf{\Pi}$, the $\mathbf{K}$-by-$\mathbf{K}$, upper-triangular matrix $\mathbf{R}_1$ in its QR factorization may be partitioned as

$$
\mathbf{R}_1 = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \end{bmatrix}
$$

(3.19)

where $\mathbf{R}_{11}$ is $\mathbf{R}$-by-$\mathbf{R}$, upper triangular, and non-singular.
A least-squares solution is then

\[ T_{LS} = \Pi \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^T C. \] (3.20)

In general, however, this least-squares solution is \textit{not} the one of minimal Euclidean norm, unless \( R_{12} = 0 \) [G11].

In addition to being able to handle rank-deficiencies, the QR-Factorization method can also detect and remedy ill-conditioning. The procedure relies on the analysis of the relative magnitude of the diagonal entries of \( R_1 \) (or \( R_{11} \)) [G11].

Furthermore, a QR factorization can be "updated" when a row (detector) or a column (expansion function) is appended to the original system. This means that a new factorization does not need to be recomputed from scratch, but can be found directly from the old one at a reduced computing cost [D8, G11].

Moreover, there exist well-tested algorithms (such as Householder’s Method or the Modified Gram-Schmidt Method) which have been developed to perform QR factorizations very efficiently [W10, D1, D8, H1, G11].

In spite of all these advantages, however, the QR Factorization method was \textit{not} chosen to solve Eq. 2.14 for at least two reasons:

- The first reason is that there exist cases (admittedly rather pathological though) where the method fails to detect near rank-deficiency [G11];

- The second and most compelling reason is that a standard QR factorization provides only a "partial picture" of a matrix. The next section shows that a much more complete and useful picture can be obtained by another orthogonalization method: ‘Singular Value Decomposition’.
3.3.3. Singular Value Decomposition.

a. Definition and Geometrical Interpretation.

Any J-by-K, real matrix $\mathbf{A}$ of rank $R$ can be written as the product of three real matrices,

$$
\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T
$$

(3.21)

where

$$
\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K] \text{ is a K-by-K, orthogonal matrix,}
$$

$$
\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_J] \text{ is a J-by-J, orthogonal matrix, and}
$$

$\mathbf{S}$ is a $R$-by-$R$, diagonal matrix: $\mathbf{S} = \text{diag} \{s_1, s_2, \ldots, s_R\}$

with $s_1 \geq s_2 \geq \ldots \geq s_R > 0$. The $s_i$'s are called the singular values of $\mathbf{A}$.

The columns of $\mathbf{V}$, $\mathbf{v}_i$, are called the right singular vectors of $\mathbf{A}$. They form an orthonormal basis of $\mathbb{R}^K$.

The columns of $\mathbf{U}$, $\mathbf{u}_i$, are called the left singular vectors of $\mathbf{A}$. They form an orthonormal basis of $\mathbb{R}^J$.

Eq. 3.21 is called a singular value decomposition (SVD) of $\mathbf{A}$ [L4, S16, G4, D8, S18, K12, L2, H14, P6, H1, S19, G11, G7]. It may also be written in an expanded form as

$$
\mathbf{A} = \sum_{i=1}^{R} s_i \mathbf{u}_i \mathbf{v}_i^T.
$$

(3.22)

which shows that, in fact, only the first $R$ columns of the $\mathbf{V}$ and $\mathbf{U}$ matrices in Eq. 3.21 are important.
It can be seen from Eq. 3.21 or Eq. 3.22 that, since \( u_i^T u_j = \delta_{ij} \) and \( v_i^T v_j = \delta_{ij} \),

\[
\mathbf{A} v_i = \begin{cases} 
  s_i \, u_i, & i = 1,2,...,R, \\
  0, & i = R+1,R+2,...,K,
\end{cases}
\]  

(3.23)

and

\[
\mathbf{A}^T u_i = \begin{cases} 
  s_i \, v_i, & i = 1,2,...,R, \\
  0, & i = R+1,R+2,...,J.
\end{cases}
\]  

(3.24)

Eqs. 3.23 and 3.24 show that the \( v_i \)'s and \( u_i \)'s are not just any orthonormal bases of \( \mathbf{R}^K \) and \( \mathbf{R}^J \) respectively, but that:

(i) the first \( R \) of the \( v_i \) vectors form an orthonormal basis for the row space of \( \mathbf{A} \), \( \mathbf{R}(\mathbf{A}^T) \);

(ii) the last \( K-R \) of the \( v_i \) vectors form an orthonormal basis for the null-space of \( \mathbf{A} \), \( \mathbf{N}(\mathbf{A}) \);

(iii) the first \( R \) of the \( u_i \) vectors form an orthonormal basis for the column space of \( \mathbf{A} \), \( \mathbf{R}(\mathbf{A}) \);

(iv) the last \( J-R \) of the \( u_i \) vectors form an orthonormal basis for the left null-space of \( \mathbf{A} \), \( \mathbf{N}(\mathbf{A}^T) \);

(v) the image by \( \mathbf{A} \) of any of the right singular vectors, \( v_i \), \( i = 1,2,...,R \), is in the direction of one of the left singular vectors, \( u_i \), \( i = 1,2,...,R \), and has length \( s_i = \| A \, v_i \| = \| A^T \, u_i \|, \) \( i = 1,2,...,R \).

Thus, the \( \mathbf{V} \) and \( \mathbf{U} \) orthogonal matrices in the SVD of \( \mathbf{A} \) represent four orthonormal bases, one for each of the four fundamental subspaces associated with \( \mathbf{A} \) (see Section 3.2). These four bases are shown in Figure 3.6. The basis of the row space, \( \{ v_1, v_2, ..., v_R \} \), is mapped by \( \mathbf{A} \) onto the basis of the column space, \( \{ u_1, u_2, ..., u_R \} \), with multiplication by the singular values \( s_i \) (Eq. 3.23). Conversely, \( \mathbf{A}^T \) maps the basis of the column
space onto the basis of the row space, again with multiplications by the $s_i$'s (Eq. 3.24). The bases of both null-spaces are mapped to zero.

Even though the null-spaces $\mathbb{N}(A)$ and $\mathbb{N}(A^T)$ may have very different dimensions, there is still a pleasant symmetry in the SVD picture, which did not exist in the QR-Factorization method. (In fact, SVD is sometimes called a "complete" orthogonalization method [G11, G7].)

![Diagram](image)

Figure 3.6. Orthonormal Bases for the Four Fundamental Subspaces Associated with $A$ (Adapted from Reference S19).

Application of Eq. 3.22 to an arbitrary vector $T$ of $\mathbb{R}^K$ yields:
\[ \mathbf{AT} = \sum_{i=1}^{R} s_i \left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{u}_i. \]  \hspace{1cm} (3.25)

The right-hand side of Eq. 3.25 can be interpreted as the result of successive linear transformations (see Figure 3.7):

(i) In the first transformation, the vector \( \mathbf{T} \) is projected onto the basis vectors \( \mathbf{v}_i \):

\[
\mathbf{T} = \mathbf{T}_R + \mathbf{T}_N, \quad \begin{cases} 
\mathbf{T}_R = \sum_{i=1}^{R} \left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{v}_i \in \mathbf{R}(\mathbf{A}^T), \\
\mathbf{T}_N = \sum_{i=R+1}^{K} \left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{v}_i \in \mathbf{N}(\mathbf{A}).
\end{cases} \hspace{1cm} (3.26)

(ii) In the second transformation, the image in \( \mathbf{R}(\mathbf{A}) \) of each projection is formed:

\[
\begin{cases} 
\left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{v}_i \rightarrow s_i \left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{u}_i, \quad i = 1,2,\ldots,R, \\
\left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{v}_i \rightarrow 0, \quad i = R+1,R+2,\ldots,K.
\end{cases} \hspace{1cm} (3.27)

(iii) The full image-vector, \( \mathbf{AT} \), is finally obtained in a third linear operation by adding these various independent contributions:

\[
\mathbf{AT} = \mathbf{AT}_R + \mathbf{AT}_N, \quad \begin{cases} 
\mathbf{AT}_R = \sum_{i=1}^{R} s_i \left( \mathbf{v}_i^T \mathbf{T} \right) \mathbf{u}_i \in \mathbf{R}(\mathbf{A}), \\
\mathbf{AT}_N = \mathbf{0}.
\end{cases} \hspace{1cm} (3.28)
b. Existence and Construction.

There is unquestionably an "eigenvalue-eigenvector flavor" which emanates from the above description of the SVD process. One may have noticed that Eq. 3.21 resembles a diagonalization formula. This flavor cannot possibly come from \( \mathbf{A} \) or \( \mathbf{A}^T \) which are not square in general. In fact, it comes from two square matrices encountered earlier: \( \mathbf{A}^T \mathbf{A} \) and \( \mathbf{A} \mathbf{A}^T \).

The existence of the SVD of an arbitrary matrix \( \mathbf{A} \) can be readily inferred from the following elementary facts:

\[
\mathbf{R}(\mathbf{A}^T \mathbf{A}) = \mathbf{R}(\mathbf{A}^T), \quad (3.29a)
\]

\[
\mathbf{N}(\mathbf{A}^T \mathbf{A}) = \mathbf{N}(\mathbf{A}), \quad (3.29b)
\]
\[ \mathbf{R}(\mathbf{A}^T \mathbf{A}) \oplus \mathbf{N}(\mathbf{A}^T \mathbf{A}) = \mathbf{R}^K, \quad (3.29c) \]

and

\[ \mathbf{R}(\mathbf{A} \mathbf{A}^T) = \mathbf{R}(\mathbf{A}), \quad (3.30a) \]

\[ \mathbf{N}(\mathbf{A} \mathbf{A}^T) = \mathbf{N}(\mathbf{A}^T), \quad (3.30b) \]

\[ \mathbf{R}(\mathbf{A} \mathbf{A}^T) \oplus \mathbf{N}(\mathbf{A} \mathbf{A}^T) = \mathbf{R}^J. \quad (3.30c) \]

Eqs. 3.30a,b state that row space and null-space of \( \mathbf{A} \) are respectively identical to column space and null-space of \( \mathbf{A}^T \mathbf{A} \).

Eqs. 3.30a,b state that column space and left null-space of \( \mathbf{A} \) are respectively identical to column space and null-space of \( \mathbf{A} \mathbf{A}^T \).

The above identities justify a statement made earlier (in Section 3.2.2), namely that the ranks of the two square matrices, \( \mathbf{A}^T \mathbf{A} \) and \( \mathbf{A} \mathbf{A}^T \), is equal to the rank of \( \mathbf{A} \).

\( \mathbf{A}^T \mathbf{A} \) and \( \mathbf{A} \mathbf{A}^T \) are both square, symmetric, positive semi-definite matrices of rank \( R \). From an algebraic standpoint, these properties make them much "nicer" matrices than either \( \mathbf{A} \) or \( \mathbf{A}^T \).

These properties can be used to construct (and thus prove the existence of) the \( \mathbf{y}_i \) and \( \mathbf{u}_i \) basis-vectors, as well as the singular values, \( s_i \), appearing in the SVD of \( \mathbf{A} \). The mathematical construction described here, although not used in numerical calculations, is simple and sheds light on the SVD process. The method starts from the \( \mathbf{A}^T \mathbf{A} \) matrix, but it could equally well begin with \( \mathbf{A} \mathbf{A}^T \).

Since the \( K \)-by-\( K \) matrix \( \mathbf{A}^T \mathbf{A} \) is symmetric, positive semi-definite, and of rank \( R \), it has the following properties:
(i) Its eigenvalues are real;
(ii) Exactly R of them are positive (counting multiplicities);
(iii) A complete set of orthonormal eigenvectors can be found.

The positive eigenvalues are denoted \( s_i^2 \) (with \( s_i \) chosen positive), \( i = 1,2,...,R \), and the corresponding eigenvectors \( \mathbf{v}_i \). These \( R \) eigenvectors lie in \( \mathbb{R}(\mathbf{A}^T \mathbf{A}) = \mathbb{R}(\mathbf{A}^T) \). They are orthogonal to one another, and their length may be normalized to unity. They thus constitute an orthonormal basis of \( \mathbb{R}(\mathbf{A}^T) \) (Figure 3.6).

This basis may be completed by \( K-R \) additional, orthonormal vectors, \( \mathbf{v}_{R+1}, \mathbf{v}_{R+2}, ..., \mathbf{v}_K \), so that the full set \( \{ \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K \} \) forms an orthonormal basis of \( \mathbb{R}^K \). By construction, \( \{ \mathbf{v}_{R+1}, \mathbf{v}_{R+2}, ..., \mathbf{v}_K \} \) constitutes an orthonormal basis for the orthogonal complement of \( \mathbb{R}(\mathbf{A}^T \mathbf{A}) \) in \( \mathbb{R}^K \), i.e. for \( \mathbb{N}(\mathbf{A}^T \mathbf{A}) = \mathbb{N}(\mathbf{A}) \). Hence \( \mathbf{A} \mathbf{v}_i = 0, i = R+1,R+2,...,K \).

The eigenvalue-eigenvector relationships may be written as:

\[
\mathbf{A}^T \mathbf{A} \mathbf{v}_i = \begin{cases} 
  s_i^2 \mathbf{v}_i, & i = 1,2,...,R, \\
  \mathbf{0}, & i = R+1,R+2,...,K. 
\end{cases} \tag{3.31}
\]

From the known \( \mathbf{v}_i \)'s and \( s_i \)'s, \( R \) vectors, denoted \( \mathbf{u}_i \), can be defined as the scaled image in \( \mathbb{R}^J \) of the \( \mathbf{v}_i \) vectors:

\[
\mathbf{u}_i = \frac{1}{s_i} \mathbf{A} \mathbf{v}_i, \quad i = 1,2,...,R. \tag{3.32}
\]

One can easily verify, using Eqs. 3.31 and 3.32, that these vectors are orthonormal and that they satisfy:

\[
\mathbf{A}^T \mathbf{u}_i = s_i \mathbf{v}_i, \quad i = 1,2,...,R. \tag{3.33}
\]

In fact, the \( \mathbf{u}_i \)'s are eigenvectors of the J-by-J matrix \( \mathbf{A} \mathbf{A}^T \) since, from Eqs. 3.32 and 3.33,
\[ \mathbf{\Delta} \mathbf{\Delta}^T \mathbf{u}_i = s_i^2 \mathbf{u}_i, \quad i = 1, 2, \ldots, R. \] (3.34)

One concludes that the set \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_R\} forms an orthonormal basis of \( \mathbf{R}(\mathbf{A} \mathbf{A}^T) = \mathbf{R}(\mathbf{A}) \) (Figure 3.6). As with the \( \mathbf{v}_i \)'s, this basis can be completed by \( J-R \) orthonormal vectors, \( \mathbf{u}_{R+1}, \mathbf{u}_{R+2}, \ldots, \mathbf{u}_J \), so that the full set \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_J\} is an orthonormal basis of \( \mathbf{R}^J \). The extra vectors, \( \mathbf{u}_{R+1}, \mathbf{u}_{R+2}, \ldots, \mathbf{u}_J \), form an orthonormal basis for the orthogonal complement of \( \mathbf{R}(\mathbf{A} \mathbf{A}^T) \) in \( \mathbf{R}^J \), i.e. for \( \mathbf{N}(\mathbf{A} \mathbf{A}^T) = \mathbf{N}(\mathbf{A}^T) \). They are therefore such that \( \mathbf{A}^T \mathbf{u}_i = 0, \quad i = R+1, R+2, \ldots, J \).

The final result of the construction is Eqs. 3.21 and 3.22. Since the bases of \( \mathbf{N}(\mathbf{A}) \) and \( \mathbf{N}(\mathbf{A}^T) \) are somewhat arbitrary, a singular value decomposition such as Eq. 3.21 is not unique. However, the expanded form of Eq. 3.22 is unique (to within inconsequential, simultaneous sign changes in the \( \mathbf{v}_i \)'s and \( \mathbf{u}_i \)'s).

In summary, the SVD of a rectangular matrix \( \mathbf{A} \) of rank \( R \) amounts to the construction of orthonormal bases for each of the four fundamental subspaces associated with \( \mathbf{A} \). The bases for the row space of \( \mathbf{A} \), \( \mathbf{R}(\mathbf{A}) \), and column space of \( \mathbf{A} \), \( \mathbf{R}(\mathbf{A}^T) \), both of dimension \( R \), are in a direct, vector-to-vector correspondence, the singular values of \( \mathbf{A} \) acting as "amplifying factors" in the mapping between the two subspaces (Eqs. 3.23 and 3.24). This decomposition of the \( \mathbf{A} \) mapping into what may be called "singular modes" \( (s_i, \mathbf{v}_i, \mathbf{u}_i) \) makes it easy to go from \( \mathbf{R}^K \) to \( \mathbf{R}^J \) and, more importantly, from \( \mathbf{R}^J \) to \( \mathbf{R}^K \), which is what is needed to solve ‘\( \mathbf{A} \mathbf{T} = \mathbf{C} \)’.

In some sense, the singular value decomposition of rectangular matrices can be viewed as a generalization of the spectral decomposition of square matrices. However, the analogy is limited since, in general, the singular values of a square matrix are not equal to its eigenvalues.
c. SVD and Rectangular Systems of Equations.

It is manifest from the discussion of the previous two sections that the properties of SVD make it ideally suited for solving rectangular systems of linear algebraic equations. All that is required is to "invert" the \( \mathbf{A} \) mapping by reverting the simple correspondence between the bases of \( \mathbf{R}(\mathbf{A}) \) and \( \mathbf{R}(\mathbf{A}^T) \).

Specifically, the minimum-norm, least-squares solution of the rectangular system \( \mathbf{A} \mathbf{T} = \mathbf{C} \) is readily obtained by

(i) Projecting the \( \mathbf{C} \) vector onto the basis-vectors \( \mathbf{u}_i \):

\[
\mathbf{C} = \mathbf{C}_R + \mathbf{C}_N,
\]

\[
\begin{align*}
\mathbf{C}_R &= \sum_{i=1}^{R} (\mathbf{u}_i^T \mathbf{C}) \mathbf{u}_i \in \mathbf{R}(\mathbf{A}), \\
\mathbf{C}_N &= \sum_{i=R+1}^{J} (\mathbf{u}_i^T \mathbf{C}) \mathbf{u}_i \in \mathbf{N}(\mathbf{A}^T); \\
\end{align*}
\]  

(3.35)

(ii) Finding the antecedent in \( \mathbf{R}(\mathbf{A}^T) \) of each projection-vector, \( (\mathbf{u}_i^T \mathbf{C}) \mathbf{u}_i, i = 1,2,...,R \), making up \( \mathbf{C}_R \) by using the vector-to-vector correspondence of the SVD and the linearity properties of the mapping:

\[
\begin{align*}
\frac{1}{s_i} (\mathbf{u}_i^T \mathbf{C}) \mathbf{v}_i &\rightarrow (\mathbf{u}_i^T \mathbf{C}) \mathbf{u}_i, \quad i = 1,2,...,R, \\
Q &\rightarrow (\mathbf{u}_i^T \mathbf{C}) \mathbf{u}_i, \quad i = R+1,R+2,...,K; \\
\end{align*}
\]  

(3.36)

(iii) And, finally, adding the various independent contributions:
\[
T_{LS} = T_R + T_N,
\]

\[
\begin{cases}
T_R = \sum_{i=1}^{\text{R}} \frac{1}{s_i} (u_i^T C) v_i \in \mathbf{R}(\Delta^T), \\
T_N = 0.
\end{cases}
\] (3.37)

Figure 3.8 illustrates this procedure for a system of rank \( R = 2 \). This figure should be compared with Figure 3.7.

![Diagram showing least-squares solution](image)

---

**Figure 3.8.** Minimum-Norm Least-Squares Solution of \( \Delta^T C \) by Use of the SVD of \( \Delta \).

It will be useful in the following to keep in mind this geometrical interpretation of the minimum-norm, least-squares solution of \( \Delta^T C \), \( T_{LS} \), as being the result of an "orthogonal-projection step", of \( C \) onto the projection "plane" \( \mathbf{R}(\Delta) \), followed by an "inversion step", from \( \mathbf{R}(\Delta) \) to \( \mathbf{R}(\Delta^T) \).
Although Eq. 3.37 was obtained from geometrical arguments, it could have been established algebraically too. Indeed, for an arbitrary vector $\mathbf{T}$ of $\mathbb{R}^K$, the squared Euclidean norm of the residual vector, $E(T) = \mathbf{C} - \mathbf{A}^T \mathbf{T}$, of Eq. 3.8 is

$$
\|E(T)\|^2 = \sum_{i=1}^{J} [u_i^T E(T)]^2 = \sum_{i=1}^{J} [u_i^T (\mathbf{C} - \mathbf{A}^T \mathbf{T})]^2
$$

$$
= \sum_{i=1}^{J} [u_i^T \mathbf{C} - u_i^T (\mathbf{A}^T \mathbf{T})]^2
$$

$$
= \sum_{i=1}^{R} [u_i^T \mathbf{C} - s_i v_i^T \mathbf{T}]^2 + \sum_{i=R+1}^{J} [u_i^T \mathbf{C}]^2, \quad (3.38)
$$

which is minimum when $\mathbf{T}$ is such that

$$
u_i^T \mathbf{C} - s_i v_i^T \mathbf{T} = 0, \quad i = 1, 2, ..., R, \quad (3.39)$$

or

$$v_i^T \mathbf{T} = \frac{1}{s_i} u_i^T \mathbf{C}, \quad i = 1, 2, ..., R, \quad (3.40)$$

i.e. when $\mathbf{T}$ has the form given by Eq. 3.37.

In these conditions, Eq. 3.38 reduces to

$$
\|E(T_{LS})\|^2 = \|E_{LS}\|^2 = \sum_{i=R+1}^{J} [u_i^T \mathbf{C}]^2. \quad (3.41)
$$

Eq. 3.37 may also be written compactly as

$$T_{LS} = \mathbf{A}^\dagger \mathbf{C} \quad (3.42)$$
where
\[
\mathbf{A}^\dagger = \sum_{i=1}^{\infty} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^T.
\]  \hspace{1cm} (3.43)

\(\mathbf{A}^\dagger\) is called the pseudo-inverse (or “Moore-Penrose generalized inverse”) of \(\mathbf{A}\) \([B7]\). Eq. 3.43 may be recast in a form analogous to Eq. 3.21:
\[
\mathbf{A}^\dagger = \mathbf{V} \begin{bmatrix} \mathbf{S}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{U}^T.
\]  \hspace{1cm} (3.44)

Eqs. 3.43 and 3.44 show that \(\mathbf{A}^\dagger\) can be readily inferred from the SVD of \(\mathbf{A}\).

From Figure 3.8, it can be seen that the effect of the pseudo-inverse, \(\mathbf{A}^\dagger\), is to invert \(\mathbf{A}\) “wherever this is possible”, i.e. between \(\mathbf{R}(\mathbf{A}^T)\) and \(\mathbf{R}(\mathbf{A})\). \(\mathcal{C}_R\) is mapped onto \(\mathcal{T}_R\) by \(\mathbf{A}^\dagger\), whereas \(\mathcal{C}_N\) is mapped onto \(\mathcal{O}\):
\[
\begin{align*}
\mathbf{A}^\dagger \mathcal{C}_R &= \mathbf{A}^\dagger \mathcal{C} = \mathcal{T}_R = \mathcal{T}_{LS}, \\
\mathbf{A}^\dagger \mathcal{C}_N &= \mathcal{O}.
\end{align*}
\]  \hspace{1cm} (3.45)

Note that this irreducible, left null-space component, \(\mathcal{C}_N\), is equal to the residual vector, \(\mathbf{E}(\mathbf{T})\):
\[
\mathcal{C}_N = \mathbf{E}(\mathbf{T}) = \mathcal{C} - \mathbf{A}^T \mathbf{T}.
\]  \hspace{1cm} (3.46)

As one may intuitively expect, the pseudo-inverse, \(\mathbf{A}^\dagger\), reduces to the regular inverse, \(\mathbf{A}^{-1}\), for square invertible matrices, and to the left and right inverses of Eqs. 3.11 and 3.12 for full-rank, rectangular matrices:

- In the full-rank, underdetermined case, \(\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T\),
- In the full-rank, overdetermined case, \(\mathbf{A}^\dagger = \mathbf{A}^T (\mathbf{A}^T \mathbf{A})^{-1}\).
Contrary to the Normal-Equations or QR-Factorization methods, the SVD method (decomposition of $\mathbf{A}$ followed by pseudo-inversion) gives a single formula for $T_{LS}$, valid for any rectangular system, underdetermined or overdetermined.

By recognizing subspaces in both $\mathbf{R}^K$ and $\mathbf{R}^I$, and by constructing orthonormal bases for each of them separately, the SVD method gives a far more complete picture of a rectangular system of equations than the other two methods. Note that QR factorization (with pivoting) constructs only one basis, either for $\mathbf{R}(\mathbf{A}^T)$ or for $\mathbf{R}(\mathbf{A})$. As a consequence, finding the minimum-norm, least-squares solution of $'\mathbf{A}^T \mathbf{T} = \mathbf{C}'$ from the QR factorization of $\mathbf{A}$ is not as straightforward as from the singular value decomposition of $\mathbf{A}$.

In addition to these advantages, the SVD method has the distinctive characteristics of being able to handle rank-degeneracies and to detect, quantify, and even remedy ill-conditioning in an "optimal" fashion, as explained below.

d. SVD and Ill-Conditioning.

The construction of the SVD (Section b) has shown that the singular values of $\mathbf{A}$ are the square roots of the nonzero eigenvalues, $s_i^2$, of the square, symmetric, positive semi-definite matrices $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A} \mathbf{A}^T$. When one (or more) of these eigenvalues is zero, both $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A} \mathbf{A}^T$ are singular. In these conditions, $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{A}^T \mathbf{A}$, and $\mathbf{A} \mathbf{A}^T$ are all rank-deficient.

Ill-conditioning, or near rank-deficiency in $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A} \mathbf{A}^T$, occurs when the smallest, nonzero eigenvalue, $s_R^2$, is "small" or, more exactly, is much smaller than the largest eigenvalue, $s_I^2$. A measure of the degree of ill-conditioning in these matrices is provided by the ratio $s_I^2 / s_R^2$. A natural measure of ill-conditioning in $\mathbf{A}$ is therefore
\[
\sqrt{\frac{s_1^2}{s_R^2}} = \frac{s_1}{s_R} \equiv \kappa(\mathbf{A}),
\] (3.47)

ratio of the largest to the smallest singular value.

\(\kappa(\mathbf{A})\) is the condition number of the rectangular matrix \(\mathbf{A}\). It is minimum, equal to one, for perfectly conditioned matrices such as the orthogonal matrices \(\mathbf{V}\) and \(\mathbf{U}\) in the SVD of \(\mathbf{A}\). It is much greater than unity for ill-conditioned matrices.

Note that the conditioning of \(\mathbf{A}^T \mathbf{A}\) and \(\mathbf{A} \mathbf{A}^T\) may be considerably worse than that of \(\mathbf{A}\) since

\[
\kappa(\mathbf{A}^T \mathbf{A}) = \kappa(\mathbf{A} \mathbf{A}^T) = [\kappa(\mathbf{A})]^2 \geq \kappa(\mathbf{A}).
\] (3.48)

Figure 3.9a shows the geometrical consequences of ill-conditioning in a J-by-K matrix \(\mathbf{A}\) of rank \(R = 2\).

\[\begin{array}{c}
\mathbf{R}^K \\
\mathbf{R}(\mathbf{A}^T) \\
\mathbf{R}(\mathbf{A}) \\
\end{array}\]

Figure 3.9a. Image of a Unit Circle of \(\mathbf{R}^K\) by an Ill-Conditioned Matrix \(\mathbf{A}\) of Rank \(R = 2\).
For the pseudo-inverse, $\mathbf{A}^\dagger$, the ratio of the largest to the smallest singular value is

$$\frac{1}{s_R} = \frac{s_1}{s_R} = \kappa(\mathbf{A}).$$  \hfill (3.49)

Therefore, $\mathbf{A}$ and $\mathbf{A}^\dagger$ share the same condition number.

Figure 3.9b shows the effect of ill-conditioning on pseudo-inversion.

Figure 3.9b. Image of a Unit Circle of $\mathbf{R}^J$ by an Ill-Conditioned Matrix $\mathbf{A}^\dagger$ or Rank $R = 2$.

For a better understanding of Figures 3.9a,b, imagine that one has some means of perturbing the $\mathbf{A}$ matrix in such a way that the ratio $s_1 / s_R = \kappa(\mathbf{A})$ can be varied at will, but the $\{\mathbf{v}_i\}$ and $\{\mathbf{u}_i\}$ bases are not affected. If one perturbs $\mathbf{A}$ so that $s_1 / s_R$ increases, the conditioning of $\mathbf{A}$ and $\mathbf{A}^\dagger$
worsens: The ellipsoids in Figures 3.9a and 3.9b become thinner. (Their aspect ratio is $s_1 / s_R$.) In the limit $s_1 / s_R \to \infty$, the $v_R$ and $u_R$ directions "disappear" ($R = 2$ in the figures), or, more precisely, enter the null-spaces, $N(A)$ and $N(A^T)$, whose dimensions then increase by one.

The inverse ratio, $s_R / s_1 = 1 / \kappa(A)$, can be given a slightly different, more mathematical interpretation. This interpretation is based on the following properties and definitions:

$$s_1 = \max_{T \neq 0} \frac{\|A^T T\|}{\|T\|} = \max_{\|T\| = 1} \|A^T T\| \equiv \|A\|_2,$$  \hspace{1cm} (3.50)

$$\frac{1}{s_R} = \max_{T \neq 0} \frac{\|A^\dagger T\|}{\|T\|} = \max_{\|T\| = 1} \|A^\dagger T\| \equiv \|A^\dagger\|_2,$$  \hspace{1cm} (3.51)

$$s_R \equiv \|\delta A\|.$$

$$s_1$$ in Eq. 3.50 measures the maximum "amplifying power" of $A$. By definition, this is the spectral norm of $A$, $\|A\|_2$, i.e. the matrix norm induced by the Euclidean vector-norm [H14, G11]. It is an absolute measure of the "size" of $A$.

Similarly, $1 / s_R$ in Eq. 3.51 measures the maximum "amplifying power" of $A^\dagger$. It is equal to the spectral norm of $A^\dagger$, $\|A^\dagger\|_2$.

$s_R$ is also an absolute measure of the distance of $A$ from the set of all rank-deficient, J-by-K matrices. It is the amount of change, $\|\delta A\|$, which is needed in the entries of $A$ to make $A + \delta A$ of rank $R-1$ (Eq. 3.52).

The ratio

$$\frac{s_R}{s_1} = \frac{\|\delta A\|}{\|A\|} = \frac{1}{\|A\| \|A^\dagger\|} = \frac{1}{\kappa(A)}$$  \hspace{1cm} (3.53)
is therefore a measure of the relative distance of $A$ from a rank-deficient matrix.

Ill-conditioning in $A$ may have very detrimental consequences on the minimum-norm, least-squares solution of $\{A \ T = C\}$. In some extreme cases, it may even make this vector completely meaningless.

To see how this may happen, first observe that, with an ill-conditioned matrix $A$, the antecedent, $T_{LS} = A^\dagger C$, of any vector $C$ of $R^J$ having a nonzero component in the $u_R$ direction is “stretched” in the $v_R$ direction, as the ellipsoid of Figure 3.9b indicates (for $R = 2$). The components of $T_{LS}$ in the $v_R$ and $v_1$ directions are respectively:

\[ v_R^T T_{LS} = \frac{1}{s_R} u_R^T C \quad (3.54a) \]

and

\[ v_1^T T_{LS} = \frac{1}{s_1} u_1^T C. \quad (3.54b) \]

The relative “stretching ratio” for the $v_R$ direction is therefore

\[ \frac{|v_R^T T_{LS}|}{|v_1^T T_{LS}|} = \frac{s_1}{s_R} = \kappa(A), \quad (3.55) \]

which is very large with respect to unity since $A$ is ill-conditioned.

The fairly large component, $|v_R^T T_{LS}| = \frac{1}{s_R} |u_R^T C|$, is not a problem in itself. However, its combination with the other components in Eq. 3.37 may cause a dramatic loss of information in $T_{LS}$. 

117
To see this clearly, suppose for instance that the $C$ vector lies entirely in $R(A)$ and points predominantly in the $u_1$ direction, with only a relatively small component, $u_R^T C$, in the perpendicular direction $u_R$, as shown in Figure 3.10 for $R = 2$. This small component may not have any physical meaning if the ratio $|u_R^T C| / |u_1^T C|$ is smaller than the relative accuracy in the measurements. Alternatively, $u_R^T C$ may come from noise in the measurements. Upon pseudo-inversion by $A^\dagger$, the main component of $C$, $u_1^T C$, is multiplied by $1 / s_1$, whereas the smaller component, $u_R^T C$, is multiplied by the large factor $1 / s_R$. For sufficiently severe ill-conditioning in $A$, the term $\frac{1}{s_R} (u_R^T C) v_R$ will dominate the sum

$$\sum_{i=1}^{R} \frac{1}{s_i} (u_i^T C) v_i = T_{LS},$$

possibly even "swamping" all the other (meaningful) contributions.
This phenomenon may seem paradoxical in that, if $\mathbf{A}$ were made exactly rank-deficient by somehow forcing $s_R$ to zero, the troublesome component in $T_{LS}$, $\frac{1}{s_R} (\mathbf{u}_R^T \mathbf{C}) \mathbf{v}_R$, would not exist because then the summation in Eq. 3.41 would terminate at $i = R-1$. In Figure 3.10, for instance, since $R = 2$, $T_{LS}$ would simply be equal to $\frac{1}{s_1} (\mathbf{u}_1^T \mathbf{C}) \mathbf{v}_1$.

The explanation to this apparent “discontinuity” is that one cannot force $s_R$ to zero without simultaneously affecting the $\{\mathbf{v}_i\}$ and $\{\mathbf{u}_i\}$ bases. As $s_R \to 0$, the $\mathbf{v}_R$ and $\mathbf{u}_R$ directions leave $\mathbf{R}(\mathbf{A}^T)$ and $\mathbf{R}(\mathbf{A})$ respectively, and become part of the corresponding null-spaces. The vector

$\frac{1}{s_R} (\mathbf{u}_R^T \mathbf{C}) \mathbf{v}_R$ enters $\mathbf{N}(\mathbf{A})$ (from infinity!), while its antecedent by $\mathbf{A}^\dagger$,
\((u_R^T C) u_R\), enters \(N(A^T)\). Once the vectors are in their respective null-spaces, all correspondence between them ceases.

Mathematically, this transition occurs only when \(A\) becomes effectively rank-deficient. In practice, however, one can expect numerical difficulties to arise for sufficiently ill-conditioned matrices, i.e. for \(s_R / s_1\) small enough.

Fortunately, there is an obvious and natural remedy to these difficulties: It is to set any singular value, \(s_i\), in the SVD of \(A\) to zero if the ratio \(s_i / s_1\) is smaller than some predefined threshold value, \(\varepsilon\). Zeroing small singular values removes ill-conditioned modes, \((s_{R}, \nu_{R}, u_{R})\), \((s_{R-1}, \nu_{R-1}, u_{R-1})\), \((s_{R-2}, \nu_{R-2}, u_{R-2})\), ..., from the row space and column space of \(A\). Eqs. 3.22 and 3.43 then become "truncated sums".

The threshold value, \(\varepsilon\), should be sufficiently large so that enough small modes are eliminated. However, too large a value may lead to an increase in systematic errors because of the loss of meaningful information (signaled by a relative residual no longer small with respect to unity). A compromise must be found, and the selected value will not be optimal for all situations. How many modes should be eliminated depends on the particular application. In some cases, the decision may be complicated by the fact that several singular values are "clustered" around the selected threshold. (This clustering problem will be addressed in Chapter 4.)

Nevertheless, it is usually possible to find an appropriate trade-off value by considering the following three general factors which always influence the choice of \(\varepsilon\):

(i) The first factor is machine precision, \(\varepsilon_m\), defined as the smallest positive number which, when added to one on the computer, gives a result different from unity. This number sets a lower limit on \(s_R / s_1\). Below \(\varepsilon_m\), meaningless digits ("roundoff errors") will quasi-inevitably ruin the minimum-norm, least-squares solution of \(A^T = C\), \(T_{LS}\) [W9].

(ii) The second factor is accuracy in the data. Eq. 3.53 suggests that the smallest singular value, \(s_R\), may not have any physical meaning if it is
such that the ratio $s_R / s_1$ is less than the relative accuracy, $\varepsilon_A$, in the entries of $A$. This small singular value probably arises from nonsignificant digits in the entries of $A$. In fact, if $s_R / s_1 < \varepsilon_A$, one cannot even be certain that $s_R$ is not exactly zero, i.e. that $A$ is not actually of rank R-1 (or lower). In this ambiguous situation, it is legitimate to set $s_R$ to zero.

(iii) The third factor is measurement noise. In some cases, it may be desirable to impose a lower bound, $\varepsilon_C$, on $s_R / s_1$ so as to limit the sensitivity of $T_{LS}$ to noise in $C$, even if this leads to a small increase in systematic errors.

By taking all three factors into account, a general zeroing criterion may be formulated as

$$s_i \to 0 \text{ if } \frac{s_i}{s_1} < \varepsilon = \max \{\varepsilon_m, \varepsilon_A, \varepsilon_C\}, \quad i = 2, \ldots, R. \quad (3.56)$$

Whenever a singular value is zeroed, it is prudent to verify that the residual of the least-squares fit remains acceptably small.

To corroborate the above geometrical and mathematical arguments, and to demonstrate further the well-founded character of the zeroing procedure, consider the following, simple, two-by-two matrix:

$$A = \begin{bmatrix} 1 & 2+2\varepsilon \\ 1-\varepsilon & 2 \end{bmatrix}, \text{ with } 0 < \varepsilon << 1. \quad (3.57)$$

The degree of ill-conditioning in $A$ can be made arbitrarily high by decreasing the value of $\varepsilon$. Since $A$ is square, eigenvalues and eigenvectors exist. They are:

$$\begin{cases} 
\lambda_1 = 3, & w_1 \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\
\lambda_2 = \frac{2}{3} \varepsilon^2, & w_2 \approx \begin{bmatrix} 2 \\ -1 \end{bmatrix}.
\end{cases} \quad (3.58)$$
Note that, in this particular example, the ill-conditioning is entirely in the eigenvalues. (An example where it is entirely in the eigenvectors is

the matrix

\[
\begin{bmatrix}
3 & 2\varepsilon \\
1 / \varepsilon & 2
\end{bmatrix}
\].

The singular values of \( A \) are:

\[
\begin{align*}
\sigma_1 & = \sqrt{10} \left( 1 + \frac{3}{10} \varepsilon \right), \\
\sigma_2 & = \sqrt{10} \left( \frac{1}{5} \varepsilon^2 - \frac{3}{50} \varepsilon^3 \right).
\end{align*}
\]  \hspace{1cm} (3.59)

Its condition number is therefore

\[
\kappa(A) = \frac{\sigma_1}{\sigma_2} \approx \frac{5}{\varepsilon^2}.
\]  \hspace{1cm} (3.60)

The result of the singular value decomposition of \( A \) may be written in the form of Eq. 3.22, i.e.:

\[
A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T
\]  \hspace{1cm} (3.61)

with

\[
\sigma_1 u_1 v_1^T \approx \begin{bmatrix}
1 - \frac{2}{5} \varepsilon^2 & 2 + 2 \varepsilon + \frac{1}{5} \varepsilon^2 \\
1 - \varepsilon + \frac{2}{5} \varepsilon^2 & 2 - \frac{1}{5} \varepsilon^2
\end{bmatrix}
\]  \hspace{1cm} (3.62)

and

\[
\sigma_2 u_2 v_2^T \approx \begin{bmatrix}
\frac{2}{5} \varepsilon^2 & -\frac{1}{5} \varepsilon^2 \\
-\frac{2}{5} \varepsilon^2 & \frac{1}{5} \varepsilon^2
\end{bmatrix}.
\]  \hspace{1cm} (3.63)
The inverse (or pseudo-inverse) of $\mathbf{A}$ may be expressed in the form of Eq. 3.43 as

$$\mathbf{A}^{-1} = \mathbf{A}^\dagger = \frac{1}{s_1} \mathbf{v}_1 \mathbf{u}_1^T + \frac{1}{s_2} \mathbf{v}_2 \mathbf{u}_2^T$$

(3.64)

with

$$\frac{1}{s_1} \mathbf{v}_1 \mathbf{u}_1^T \approx \begin{bmatrix} \frac{1}{10} - \frac{3}{50} \varepsilon & \frac{1}{10} - \frac{8}{50} \varepsilon \\ \frac{2}{10} + \frac{4}{50} \varepsilon & \frac{2}{10} - \frac{6}{50} \varepsilon \end{bmatrix}$$

(3.65)

and

$$\frac{1}{s_2} \mathbf{v}_2 \mathbf{u}_2^T \approx \begin{bmatrix} \frac{1}{\varepsilon^2} \frac{-1}{\varepsilon^2} \\ \frac{-1}{2 \varepsilon^2} \frac{1}{2 \varepsilon^2} \end{bmatrix}.$$  

(3.66)

On the other hand, an exact calculation yields

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{\varepsilon^2} & \frac{-1 - \varepsilon}{\varepsilon^2} \\ \frac{-1 + \varepsilon}{2 \varepsilon^2} & \frac{1}{2 \varepsilon^2} \end{bmatrix}.$$  

(3.67)

It can be seen from Eqs. 3.61, 3.62, and 3.63 that, as $\varepsilon$ approaches zero, the term $s_1 \mathbf{u}_1 \mathbf{v}_1^T$ becomes the dominant term in $\mathbf{A}$. At the same time, however, Eqs. 3.64, 3.65, and 3.66, show that it is the term $\frac{1}{s_2} \mathbf{v}_2 \mathbf{u}_2^T$ which becomes preponderant in $\mathbf{A}^{-1} = \mathbf{A}^\dagger$, consistently with the expression for the exact inverse, Eq. 3.67.

If the small mode, $(s_2, \mathbf{v}_2, \mathbf{u}_2)$, is discarded, the following approximate inverse is obtained:
\[ \Delta^* = \frac{1}{s_1} \nu_1 u_1^T \approx \begin{bmatrix} 1 & 1 \\ 10 & 10 \\ 2 & 2 \\ 10 & 10 \end{bmatrix} \]

The disastrous consequences of ill-conditioning are illustrated in Table 3.1. The first row of the table contains four nearly-identical right-hand-side vectors \( \mathcal{C} \). The second row contains the corresponding, exact solution-vectors, \( \mathcal{T} \), of \( \Delta^* \mathcal{T} = \mathcal{C} \). It can be seen that these exact solution-vectors, although yielding a zero residual vector, \( \mathcal{C} - \Delta^* \mathcal{T} \) (third row of table), exhibit large variations from \([1/2, 1/2]^T\) to \([-\infty, +\infty]^T\) for \( \mathcal{C} \) near the point \([3/2, 3/2]^T\). Numerical difficulties can be expected for \( \varepsilon \) small enough.

The approximate solution, \( \mathcal{T}_{LS} \) (fourth row), in contrast, gives a nonzero residual vector, \( \mathcal{C} - \Delta \mathcal{T}_{LS} \) (fifth row). However, \( \mathcal{T}_{LS} \) remains consistently near the point \([3/10, 6/10]^T\), approaching it in a well-behaved fashion as \( \varepsilon \to 0 \).
Table 3.1. Effect of Ill-Conditioning on the Solution of a Simple, Two-by-Two System, ‘$\mathbf{A} \mathbf{T} = \mathbf{C}$’, for Four Particular Right-Hand-Side Vectors $\mathbf{C}$.

<table>
<thead>
<tr>
<th>$\mathbf{C}$</th>
<th>$\begin{bmatrix} 3 \ 2 \ 3 \ 2 \end{bmatrix}$</th>
<th>$\begin{bmatrix} 3 + \varepsilon \ 2 \ 3 \ 2 \end{bmatrix}$</th>
<th>$\begin{bmatrix} 3 \ 2 \ 3 - \frac{1}{2} \varepsilon \ 2 \end{bmatrix}$</th>
<th>$\begin{bmatrix} 3 + \varepsilon \ 2 \ 3 - \frac{1}{2} \varepsilon \ 2 \end{bmatrix}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{T} = \mathbf{A}^{-1} \mathbf{C}$</td>
<td>$\begin{bmatrix} -\frac{3}{2} \ 2 \varepsilon \ 3 \ 4 \varepsilon \end{bmatrix}$</td>
<td>$\begin{bmatrix} -\frac{1}{2} \ 2 \varepsilon \ 1 + \frac{1}{2} \ 2 \varepsilon \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 - \frac{1}{2} \ 2 \varepsilon \ \frac{1}{4} \varepsilon \end{bmatrix}$</td>
<td>$\begin{bmatrix} \frac{1}{2} \ \frac{1}{2} \end{bmatrix}$</td>
</tr>
<tr>
<td>$\mathbf{C} - \mathbf{A} \mathbf{T}$</td>
<td>$\begin{bmatrix} 0 \ 0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 0 \ 0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 0 \ 0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 0 \ 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\mathbf{T}_{LS} = \mathbf{A}^* \mathbf{C}$</td>
<td>$\begin{bmatrix} 3 - \frac{33}{10} \varepsilon \ 100 \ 6 - \frac{6}{10} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 3 - \frac{23}{10} \varepsilon \ 100 \ 6 - \frac{160}{10} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 3 - \frac{38}{10} \varepsilon \ 100 \ 6 - \frac{160}{10} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 3 - \frac{28}{10} \varepsilon \ 100 \ 6 + \frac{4}{10} \varepsilon \ 100 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\mathbf{C} - \mathbf{A} \mathbf{T}_{LS}$</td>
<td>$\begin{bmatrix} -\frac{150}{100} \varepsilon \ 100 \ \frac{75}{100} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} \frac{323}{100} \varepsilon \ 100 \ \frac{373}{100} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} \frac{238}{100} \varepsilon \ 100 \ \frac{338}{100} \varepsilon \ 100 \end{bmatrix}$</td>
<td>$\begin{bmatrix} o(\varepsilon^2) \ 100 \ 0 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

One concludes from Table 3.1 that, for $\varepsilon$ small enough, the vector $\mathbf{T}_{LS} = [3/10, 6/10]^T$ is as acceptable a solution as the exact vector $\mathbf{T}$.

In fact, $\mathbf{T}_{LS}$ is the minimum-norm solution of the rank-deficient system
\[
\begin{bmatrix}
1 & 2 \\
1 & 2
\end{bmatrix}^T = 
\begin{bmatrix}
2 \\
2 \\
3 \\
2
\end{bmatrix}.
\] (3.69)

In summary, the SVD technique makes it possible not only to detect and quantify ill-conditioning in a rectangular system of equations, 'A T = C', but also to remedy the problem in a consistent and optimal way. This is accomplished very easily by mere inspection of the singular-value spectrum of A, which gives a complete characterization of the conditioning of A. Upon judicious zeroing of the smallest singular values, a minimum-norm, least-squares solution which is consistent with the accuracy of the data and not too sensitive to perturbations in C can be found. This zeroing procedure will be given a physical justification in Section 3.4.2.

e. Numerical Considerations.

The advantages of the SVD method over the Normal-Equations and QR-Factorization methods appear to be overwhelming. However, the applicability of SVD is still subordinate to certain numerical considerations. Specifically, for SVD to be of any practical interest, an algorithm must be available to determine the singular values and singular vectors of a matrix A without explicitly forming the matrices A^T A and A A^T, and without requiring excessive computational efforts.

Fairly recently, elaborate algorithms capable of performing such a decomposition in an efficient and unconditionally stable fashion have been developed [W10, G11]. Highly-optimized computer programs based on these algorithms have also become available [S5, G4, D8, P6].

One widely-used SVD algorithm is the "Golub-Reinsch Algorithm" [G11]. In this method, the matrix to be decomposed is first reduced to bidiagonal form by successive multiplications by Householder reflection
matrices. The SVD of the bidiagonal matrix is subsequently computed by a variant of the QR algorithm. Details about this procedure can be found in References W10, D8 and G11.

Table 3.2 compares the operation counts of three algorithms capable of solving overdetermined systems of linear algebraic equations. Table 3.3 compares the costs of inversion algorithms for square systems.

Table 3.2. Compared Operation-Counts of Three Different Algorithms for Solving Rectangular, J-by-K Systems of Equations (J ≥ K). (From Reference G11.)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Operation Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Equations</td>
<td>JK² + K³/3</td>
</tr>
<tr>
<td>QR-Factorization (Householder Orthogonalization)</td>
<td>2JK² - 2K³/3</td>
</tr>
<tr>
<td>SVD (Golub-Reinsch)</td>
<td>4JK² + 8K³ (V and S only)</td>
</tr>
<tr>
<td></td>
<td>14JK² + 8K³ (V, S, and first K columns of U)</td>
</tr>
<tr>
<td></td>
<td>4J²K + 8JK² + 9K³ (V, S, and U)</td>
</tr>
</tbody>
</table>
Table 3.3. Compared Operation-Counts of Three Different Algorithms for Solving Square, K-by-K Systems of Equations. (From Reference G11.)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Operation Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Elimination</td>
<td>$2K^3/3$</td>
</tr>
<tr>
<td>QR-Factorization (Householder Orthogonalization)</td>
<td>$4K^3/3$</td>
</tr>
<tr>
<td>SVD (Golub-Reinsch)</td>
<td>$12K^3$</td>
</tr>
</tbody>
</table>

From these tables, SVD appears to be considerably more expensive than any other method. In point of fact, in many applications, the cost disadvantage of SVD is often the overriding reason for its rejection in favor of some other, cheaper method.

However, it is important to remember that the $A$ matrix of the fitted-synthesis method has a fairly small size (say $J \sim 100$ rows and $K \sim 10$ columns). In such circumstances, the cost of any algorithm is likely to be dictated more by overhead considerations than by the number of multiplications and divisions represented by the asymptotic counts of the above tables.

Note also from Table 3.2 that

(i) In any method, a large number of rows ($J$) is not too penalizing provided the number of columns ($K$) is small. The only exception is SVD when the entire $U$ matrix is explicitly needed;

(ii) The cost difference between the QR-Factorization and SVD algorithms is minimum (factor of two) for matrices having a small number of columns relative to the number of rows ($K \ll J$).
3.4. Application of the SVD Method to the Fitted-Synthesis Equations.

The previous discussion on rectangular systems of equations (Section 3.2) and on the various solution-techniques available (Section 3.3) has been kept fairly general. SVD was found to be the most attractive method mainly on the basis of its unique ability to handle ill-conditioning.

Before turning to numerical tests, it is useful to specialize further the application of SVD to the particular system of equations which results from the fitted-synthesis described in Chapter 2. This specialization will show that:

(i) Certain restrictions on the choice of the free parameters of the synthesis method (detector positions and number, number of expansion functions and reactor conditions for which they are generated) need to be imposed;

(ii) Considerable physical (rather than geometrical or mathematical) insight can also be gained into the double process of orthogonalization by SVD and zeroing of small singular values.

3.4.1. Least-Squares Residual as an Error Estimate.

One should recall from Chapter 2 that the actual system of equations to be solved is not \( \mathbf{A} \mathbf{T} = \mathbf{C} \), but more exactly

\[
\mathbf{A} \mathbf{T}(t) + \mathbf{E}(t) = \mathbf{C}(t)
\]  

(2.11b)

where

\[
\mathbf{E}(t) = - \mathbf{\Sigma}^T \mathbf{\Delta} \phi(t)
\]  

(2.13)
is a J-element vector representative of the G×N-element vector of systematic errors, \( \delta \Phi(t) = \hat{\Phi}(t) - \Phi(t) \), caused by the flux-synthesis approximation:

\[
\Phi(t) = \hat{\Phi}(t) = \sum_{k=1}^{K} \Psi^{(k)} T^{(k)}(t) = \Psi^T \mathbf{T}(t). \tag{2.0}
\]

The J-by-G×N, flux-to-measurement operator, \( \Sigma^T \), converts the vector of actual, nodal, group fluxes, \( \Phi(t) \), into the reduced-order observation-vector, \( \mathbf{C}(t) \):

\[
\Sigma^T \Phi(t) = \mathbf{C}(t). \tag{2.8}
\]

\( \Sigma^T \) also enters the matrix \( A = \Sigma^T \Psi \) of Eq. 2.11b.

One immediately recognizes in \( E(t) \) of Eq. 2.11b the residual vector of the least-squares problem (Section 3.2.2). From this observation and from Eq. 2.13, it follows that:

(i) Since the Euclidean norm of this vector, \( \|E(t)\| \), is minimized by the least-squares solution of \( A^T T(t) = C(t) \), \( T_{LS}(t) \), the full error-vector, \( \delta \Phi(t) \), is also minimized in some weighted-integral sense (or in the sense of some norm of \( R^{GN} \));

(ii) An experimental estimate of \( \delta \Phi(t) \) is available in the vector \( E_{LS}(t) \). Individual components of \( E_{LS}(t) \) give weighted estimates of local absolute errors, while the Euclidean norm of \( E_{LS}(t) \), \( \|E_{LS}(t)\| \), gives a weighted-integral measure of these same errors.

A relative integral measure of \( \delta \Phi(t) \) is provided by the relative residual (Eq. 3.9) equal to

\[
\frac{\|E_{LS}(t)\|}{\|C(t)\|} = \frac{\|\Sigma^T \delta \Phi(t)\|}{\|\Sigma^T \Phi(t)\|}. \tag{3.70}
\]
Note, however, that this interpretation of \( \| E_{LS}(t) \| \) and \( \| E_{LS}(t) \| / \| C(t) \| \) as measures of the absolute and relative errors in the synthesis approximation is conditional to the assumption that the measurements making up \( C(t) \) are noise free. Noise in the measurements will be considered in detail in Chapter 5.

This capability of estimating errors in \( \hat{\phi}(t) \) is an extremely important characteristic of the fitted-synthesis method because, as explained in Sections 1.3 and 2.3.2, no theoretical bound exists for \( \delta \hat{\phi}(t) \).

Unfortunately, a good least-squares fit (\( \| E_{LS}(t) \| / \| C(t) \| \ll 1 \)) is only a necessary condition for a successful synthesis. Strictly speaking, a good or even perfect fit (\( \| E_{LS}(t) \| = 0 \)) does not guarantee that, upon reconstruction of \( \hat{\phi}(t) \) by Eq. 2.9, large errors will not result at some unmonitored locations.

There can be two reasons for this. One is related to the neutron detectors, the other to the expansion functions, \( \Psi^{(k)} \), entering the columns of the \( G \times N \)-by-\( K \) matrix \( \Psi \).

For a given set of expansion functions, one can easily imagine situations where an insufficient, inoperable, or poorly-distributed instrumentation can cause failure of the synthesis method with no "warning" from the residual vector. For example, by using only a few neutron detectors and by positioning them all in some corner of the reactor, large errors will almost certainly result away from the instrumented region. Common sense and a reliable and "sufficiently rich" instrumentation should prevent such situations to occur.

For a given set of \( J \) detectors reasonably distributed throughout the core, the residual vector can still fail to be a reliable indicator of the errors in \( \hat{\phi}(t) \) if the expansion functions are too numerous and poorly chosen. One can easily convince oneself of this reality from the following geometrical arguments.
Recall from Eq. 2.15b and Figure 2.1 that each column of $\mathbf{A}$ is the
"image" through the instrumentation ($\Sigma^T$) of an expansion function. The
column space of $\mathbf{A}$, $\mathbf{R}(\mathbf{A})$, is therefore directly affected by the expansion
functions. A change in their number may change the dimension, $R$, of
$\mathbf{R}(\mathbf{A})$. The replacement of one expansion function by another may alter
both the dimension of $\mathbf{R}(\mathbf{A})$ and its position in $\mathbf{R}^J$ (see Figure 3.8). *The expansion functions must be chosen in such a way that the projection
"plane", $\mathbf{R}(\mathbf{A})$, remains close to $\mathbf{C}(t)$, i.e. in such a way that $\|\mathbf{E}(t)\| \text{remains small with respect to } \|\mathbf{C}(t)\|$ at all times. They are many different ways to
achieve this objective, including "poor strategies".*

An example of such a poor strategy is to include a systematically
large number of arbitrary expansion-functions in the matrix $\mathbf{A} = \Sigma^T \Psi$.
Eventually, by inclusion of a sufficient number ($K \geq J$) of linearly-
independent functions, $\mathbf{R}(\mathbf{A})$ becomes equal to the entire observation space,$\mathbf{R}^J$, leaving little space in $\mathbf{N}(\mathbf{A}^T)$ for a nonzero residual vector. In other
words, $\mathbf{A}^T \mathbf{T}(t) = \mathbf{C}(t)$’ is a full-rank, underdetermined system; it has a
unique, minimum-norm solution, which is such that the detector readings
are matched exactly. The residual vector, $\mathbf{E}(t)$, is constantly $\mathbf{0}$. However,
in spite of this nil residual, the synthesized flux-vector, $\hat{\phi}(t)$, will probably
be erroneous except at the detector locations.

Overdetermination is intuitively a much preferable situation. Not
only does it avoid the uselessness of a constantly-nil residual vector, but it
also provides $J-R$ extra degrees of freedom (the dimension of $\mathbf{N}(\mathbf{A}^T)$)
which can be advantageously used by the least-squares procedure to
discriminate between "good" and "bad" expansion-functions: Basis
functions which do not contribute to improving the fit significantly are
given relatively small weights, $\mathbf{T}^{(k)}_{\text{LS}}$. Since the success of the synthesis
method depends so much on an appropriate choice of expansion functions,
and since this choice may sometimes be a delicate one, it seems highly
desirable to let the synthesis method operate in this selective,
overdetermined mode.
Other arguments in favor of \( J > K \) (and even \( J \gg K \)) are computing-cost considerations and robustness to measurement noise.

In summary, great caution should be exercised in the interpretation of \( \mathbf{E}_{\text{LS}}(t) \) and \( \| \mathbf{E}_{\text{LS}}(t) \| \| \mathbf{C}(t) \| \) as indicators of the quality of the synthesis. One should always bear in mind that \( \mathbf{E}_{\text{LS}}(t) \) depends on the choice of the expansion functions and can be made arbitrarily small by increasing their number. As a general rule, a small relative residual is only a necessary condition for a good flux-synthesis.

For \( J > K \), a relative residual, \( \| \mathbf{E}_{\text{LS}}(t) \| / \| \mathbf{C}(t) \| \), which is not small with respect to unity is an unambiguous sign of a poor synthesis. On the other hand, if \( \| \mathbf{E}_{\text{LS}}(t) \| / \| \mathbf{C}(t) \| \ll 1 \), if the expansion functions have been properly selected (a problem discussed in Chapter 4), and if, in addition, measurement noise is negligible, one can have reasonable confidence that \( \mathbf{E}_{\text{LS}}(t) \) is a reliable indicator of the errors caused by the synthesis. This is particularly true when \( J \gg K \). In such a strongly overdetermined situation, \( \mathbf{R}(\mathbf{A}) \) represents only a thin projection “plane” in \( \mathbf{R}^J \), and it is extremely unlikely that \( \mathbf{C}(t) \) could lie near that plane if the latter had been constructed from inappropriate expansion-functions.

### 3.4.2. Orthogonalization of the Expansion Functions by SVD.

The specialization of the SVD method to the synthesis equations of Chapter 2 makes it possible to give a very interesting, physical interpretation of the orthogonalization process. This interpretation relies in fact on a simple notational change.

Since the \( \mathbf{V} \) matrix in the SVD of \( \mathbf{A} \) is an orthogonal matrix, it satisfies \( \mathbf{V} \mathbf{V}^T = \mathbf{I} \) where \( \mathbf{I} \) is the \( K \)-by-\( K \) identity matrix of \( \mathbf{R}^K \). Therefore, the fitted linear combination of expansion functions (Eq. 2.9) may be rewritten as:
\[ \hat{\varphi}(t) = \sum_{k=1}^{K} \psi^{(k)} T^{(k)}_{LS}(t) = \Psi T_{LS}(t) = \Psi V V^T T_{LS}(t) \]

\[ = \Psi' T_{LS}(t) = \sum_{i=1}^{R} \psi^{(i)} T^{(i)}_{LS}(t). \]  

(3.71)

where the following definitions have been introduced:

\[ T'_{LS}(t) = V^T T_{LS}(t), \]  

(3.72)

\[ \Psi' = \Psi V, \]  

(3.73)

or, equivalently,

\[ T^{(i)}_{LS}(t) = v_i^T T_{LS}(t) = \sum_{k=1}^{K} \nu^{(k)}_{i} T^{(k)}_{LS}(t), \quad i = 1,2,...,R, \]  

(3.74)

\[ \psi^{(i)} = \sum_{k=1}^{K} \psi^{(k)} \nu^{(k)}_{i}, \quad i = 1,2,...,K. \]  

(3.75)

Eq. 3.71 shows that, in effect, the initial linear combination of K expansion functions, \( \psi^{(k)} \), has been replaced by a new linear combination of expansion functions, \( \psi^{(i)} \). Note that, although there are K such new functions, only the first R ones appear in the synthesis. These new basis-functions are linear combinations of the original ones (Eq. 3.75).

Similarly, the coefficients, \( T^{(i)}_{LS}(t) \), of these new expansion-functions are linear combination of the original mixing-coefficients, \( T^{(k)}_{LS}(t) \) (Eq. 3.74). One recognizes in these coefficients the projections of \( T_{LS} \) onto the row space of \( \Delta, R(\Delta^T) \). In fact, \( T'_{LS} \) is simply \( T_{R} \), the row-space component of \( T_{LS} \) (Eq. 3.37).

With this new notation, Eqs. 2.9 and 2.11a become
\[
\phi(t) = \tilde{\phi}(t) - \delta\phi(t) = \sum_{i=1}^{R} \psi^{(i)} T_{\text{LS}}^{(k)}(t) - \delta\phi(t)
\] (3.76)

and

\[
\Sigma^T \Psi' T_{\text{LS}}(t) - \Sigma^T \delta\phi(t) = \mathbf{C}(t)
\] (3.77)

respectively.

Furthermore, observe that, from Eqs. 3.73, 2.12, and 3.21,

\[
\Sigma^T \Psi' = \Sigma^T \Psi V = AV = U \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}
\] (3.78)

or, in an expanded form, using Eq. 3.22,

\[
\Sigma^T \psi^{(i)} = A v_i = \begin{cases} s_i u_i, & i = 1,2,\ldots,R, \\ 0, & i = R+1,R+2,\ldots,J. \end{cases}
\] (3.79)

Recall that the effect of \(\Sigma^T\) is to convert any vector of \(\mathbb{R}^{GN}\) (such as \(\phi(t), \delta\phi(t), \psi^{(k)},\) or \(\psi^{(i)}\)) into a vector of the measurement space, \(\mathbb{R}^{J}\) (to which \(\mathbf{C}(t)\) belongs). The vector \(\Sigma^T \psi^{(i)}\) may therefore be thought of as that portion of \(\psi^{(i)}\) "visible through the instrumentation window".

On the other hand, Eq. 3.79 shows that \(\Sigma^T \psi^{(i)}\) is the image by \(A\) of the right singular vector, \(v_i\). \(\Sigma^T \psi^{(i)}\) belongs to \(\mathbb{R}(A)\) and is equal to either \(s_i u_i\) if \(i \leq R\), or to \(0\) if \(i > R\):

(i) \(\Sigma^T \psi^{(i)} = 0\), \(i = R+1,R+2,\ldots,J\), simply means that a linear combination (\(\psi^{(i)}\)) of the original expansion-functions, \(\psi^{(k)}\), is zero in \(\mathbb{R}(A)\), at least as "sensed" by the instrumentation (\(\Sigma^T\)). There are J-R such
zero linear combinations, i.e. linear dependences. J-K of them result from overdetermination, and K-R from rank degeneracies. These linear dependences are the reason why only the first R of the $\Psi^{(i)}$'s need to be retained in the synthesis (Eq. 3.71).

Linear dependences between the $\Psi^{(k)}$'s is generally the cause of rank-deficiencies in $A$. However, even when there is no actual linear-dependence in the expansion functions $\Psi^{(k)}$, rank deficiencies in $A$ (and therefore extra zero vectors $\Sigma^T \Psi^{(i)}$) may still occur if there are symmetries in both the detector locations and core-composition pattern. One can even imagine "unfortunate" situations where rank degeneracy occurs although the $\Psi^{(k)}$'s are linearly-independent and the detectors asymmetrically-positioned.

(ii) $\Sigma^T \Psi^{(i)} = s_i u_i$, $i = 1,2,\ldots,R$, on the other hand, means that the first R of the new basis-vectors, $\Psi^{(i)}$, are orthogonal (in $\mathbb{R}^{GN}$) in a weighted sense, since, from the orthogonality of the $u_i$'s,

$$\Psi^{(i)^T} \Sigma \Sigma^T \Psi^{(j)} = s_i^2 \delta_{ij} = \begin{cases} i = 1,2,\ldots,R, \\
 j = 1,2,\ldots,R. \end{cases}$$  \hspace{1cm} (3.80)

Moreover, the length of the vector $\Sigma^T \Psi^{(i)}$ is equal to the $i$-th singular value, $s_i$:

$$\|\Sigma^T \Psi^{(i)}\| = s_i, \hspace{0.5cm} i = 1,2,\ldots,R,$$  \hspace{1cm} (3.81)

and therefore

$$\frac{\|\Sigma^T \Psi^{(i)}\|}{\|\Sigma^T \Psi^{(1)}\|} = \frac{s_i}{s_1}, \hspace{0.5cm} i = 1,2,\ldots,R.$$  \hspace{1cm} (3.82)

Eq. 3.82 is consistent with the observations of item (i) above and with Eq. 3.53. It shows that the existence of a small singular value, $s_1$, is a
reliable indication that a linear combination \( \psi^{(i)} \) of the original expansion-functions is relatively small (again, as “sensed” by the instrumentation). Beyond a certain linear-dependence tolerance, \( \varepsilon \), for the ratio \( s_i / s_1 \), \( \sum^T \psi^{(i)} \) is zero to within physical or numerical uncertainties. Some of the original expansion-functions are perceived by the set of detectors as near linear-combinations of other expansion-functions, and, as a result, a linear combination of the columns of \( \mathbf{A} \) is almost zero.

This observation and Eq. 3.82 make it possible to give a physical (rather than numerical) justification for zeroing small \( s_i \)'s: Zeroing a small singular value, \( s_i \), amounts to eliminating \( \psi^{(i)} \) from the set of orthogonalized expansion-functions, i.e. to discard a linear combination of the original set of basis-functions on grounds that it appears too small to have an unambiguous physical meaning.

On the other hand, for \( s_i \)'s above the uncertainty threshold, \( \varepsilon s_1 \), the \( \psi^{(i)} \)'s are physically meaningful. Eqs. 3.80 and 3.81 make it natural to interpret these \( \psi^{(i)} \)'s as expansion modes for \( \hat{\phi}(t) \) in Eq. 3.71, \( \psi^{(1)} \) being the “fundamental mode”, and the other \( \psi^{(i)} \)'s being “harmonics” or smaller corrections to this fundamental. (Recall that \( s_1 \geq s_2 \geq \ldots \geq s_R > 0 \).) Possibly “missing” modes enter the error vector \( \delta \phi(t) \).

Observe that, contrary to modal expansions (Section 1.3), the \( \psi^{(i)} \) modes of the synthesis expansion are not generated as eigenmodes of a single operator, but are instead constructed by orthogonalization of various fundamental \( \lambda \)-modes (the expansion functions \( \psi^{(k)} \)) corresponding to different (static) operators.

The coefficients of the \( \psi^{(i)} \) modes in the synthesis are the new mixing or modal coefficients, \( T^{(i)}_{LS}(t) \), \( i = 1, 2, \ldots, K \). Since, as noted above, \( T_{LS} = T_R \), it follows from Eq. 3.37 that
\[ T_{LS}^{(i)}(t) = \begin{cases} \frac{1}{s_i} u_i^T C(t), & i = 1,2,...,R, \\ 0, & i = R+1,R+2,...,K. \end{cases} \] (3.83)

Eq. 3.83 shows that the proper least-squares “blending” of the orthogonalized expansion-functions, \( \Psi^{(i)} \), is determined by the projections of the measurement vector, \( C(t) \), onto the first \( R \) basis-vectors \( u_i \), divided by the corresponding singular values \( s_i \).\(^1\)

Although the scalars \( s_i \) are fundamental quantities, the importance of the \( u_i \) vectors should not be overlooked. The \( u_i \)'s give the directions in \( R^J \) of the orthogonal vectors \( \Psi^{(i)} \) (Eq. 3.79). For a given instrumentation, they depend exclusively on the choice of the basis functions \( \Psi^{(k)} \).

The importance of a “careful” selection of the \( \Psi^{(k)} \)'s has already been stressed (in particular in Sections 2.3.2 and 3.4.1). In the light of the interpretation of the synthesis formula as a modal expansion, the following, slightly more precise statement may be made: A necessary condition for an accurate flux-synthesis is that all relevant perturbation-modes be reflected in the basis \( \{ u_1, u_2, ..., u_R \} \) of \( R(\Delta) \), i.e. be present in the \( \Psi^{(k)} \)'s, and be large enough to “survive” the zeroing procedure. This implies that the

\(^1\) As one may suspect, this “column orthogonalization” in which \( \Psi \) and \( T_{LS} \) are replaced by \( \Psi' \) and \( T'_{LS} \) respectively has a row analog. Upon replacing the vector \( C \) by \( U U^T C \), the rectangular system ‘\( \Delta T = C \)’ where \( \Delta = \Sigma^T \Psi \) becomes ‘\( \Sigma' T \Psi 'T = C_R \)’ where \( C_R = U^T C \) is the orthogonal projection of \( C \) onto \( R(\Delta) \) (Eq. 3.35), and \( \Sigma' = \Sigma U \) is a matrix of “orthogonalized” detector weight-functions: \( \Sigma^{(i)} = \sum_{j=1}^{J} u_{1}^{(j)} \Sigma^{(j)} \), \( i = 1,2,...,J \). This point is not pursued here, however, since the physical interpretation of the \( \Sigma^{(i)} \)'s is not as straightforward as that of the \( \Psi^{(i)} \)'s.
number, locations, and characteristics, $\Sigma^{(i)}$, of the detectors should be such that these modes are “sufficiently” weighted.

In practice, one can expect the $C(t)$ vector to point predominantly in the $u_1$ direction most of the time, i.e. in the direction of $\psi^{(1)}$, the fundamental mode. In these conditions,

$$|u_i^T C(t)| << |u_1^T C(t)|, \quad i = R,R-1,... . \tag{3.84}$$

It follows from Eqs. 3.79, 3.83, and 3.84 that

$$\|\Sigma^T \psi^{(i)} T_{LS}^{(i)}(t)\| << \|\Sigma^T \psi^{(1)} T_{LS}^{(1)}(t)\|, \quad i = R,R-1,... . \tag{3.85}$$

Inequalities 3.85 show the expected relative importance of the vectors $\Sigma^T \psi^{(i)} T_{LS}^{(i)}(t)$ in the left-hand side of Eq. 3.77. Only the first few of these vectors, and therefore of the $\psi^{(i)}$’s, represent significant contributions.

The only situation where one can expect inequalities 3.84 and 3.85 to break down is when local perturbations occur and are strong enough to be detected. Then $C(t)$ will have a nonnegligible component in some direction $u_i \neq u_1$, and the corresponding vector $\Sigma^T \psi^{(i)} T_{LS}^{(i)}(t)$ will become significant in Eq. 3.77. At the same time the Euclidean norm of the solution-vector, $\|T_{LS}(t)\| = \|T_{LS}(t)\|$, will become substantially larger than when no perturbation mode is excited.

These last remarks have an immediate practical consequence: Small singular values which correspond to high-order perturbation modes may have to be set to zero if $C(t)$ contains a significant amount of noise. The reason for this is that the presence of noise in $C(t)$ may excite one of these high-order modes, thereby leading to an anomalously large component, $|u_i^T C|$, in some direction $u_i$ corresponding to a small $s_i$. This effect is amplified upon multiplication by the factor $1/s_i$ (Eq. 3.83), as already
discussed in Section 3.3.3.d. As a consequence, the corresponding mode, $\Psi^{(i)}$, may be given too large a weight, $T_{L_S}^{(i)}(t)$, in the synthesis, and fairly large errors may result in $\hat{\phi}(t)$. The effect should be maximum for high-order harmonics $\Psi^{(i)}$ such as those arising from local perturbations (rod motions). At the same time, the impact should be only local since $\Psi^{(i)}$ will be small away from the perturbed regions. This has been confirmed by numerical experiments, some of which will be described in Chapter 5.

To summarize, the effect of orthogonalization on the solution of the fitted-synthesis equations, $\mathbf{A} \mathbf{T} = \mathbf{C}$, by SVD is a rearrangement of the K basis-functions into a new set of R (≤ K) orthogonalized functions or modes. The relative importance of these modes is measured by the singular values of $\mathbf{A}$. Ill-conditioning in $\mathbf{A}$ means that the flux sensors have detected the presence of high-order modes in the synthesis. Deliberate elimination of the highest-order modes is justified if their physical meaning cannot be ascertained, or if excessive sensitivity of the results to measurement noise is feared. This elimination increases systematic errors, but it is always possible to verify that the effect is only small and/or local.

3.5. Summary.

In this Chapter, the general properties of rectangular systems of linear algebraic equations, $\mathbf{A} \mathbf{T} = \mathbf{C}$, have been discussed using both geometrical and algebraic arguments. Three candidate solution-methods, the Normal Equations, QR Factorization, and Singular Value Decomposition, were compared. The method based on orthogonalization by SVD, in spite of its fairly high cost, was found to be the most suitable one.

The SVD of $\mathbf{A}$ consists in constructing orthonormal bases for each of the four fundamental subspaces associated with $\mathbf{A}$ (row space, null-space,
column space, and left null-space) in such a way that $\mathbf{A}$ is "diagonal" when expressed in these orthonormal bases. The minimum-norm, least-squares solution of $\mathbf{A} \mathbf{T} = \mathbf{C}$ is readily obtained by pseudo-inversion of the diagonalized mapping. The diagonalized form of $\mathbf{A}$ contains its singular values which give a complete characterization of the conditioning of $\mathbf{A}$. The ratio of the smallest to the largest singular value measures the degree of ill-conditioning in $\mathbf{A}$. Below a certain threshold-value for this ratio, small singular values should be set to zero for consistency with the limited accuracy of $\mathbf{A}$, and to avoid numerical difficulties.

The SVD method was then more specifically applied to the rectangular system of equations obtained in Chapter 2. Overdetermination, i.e. neutron detectors outnumbering expansion functions, was found to be a necessary condition for reliable performance of the synthesis method. The residual vector of the least-squares fit was interpreted as a measure of the systematic errors caused by the synthesis approximation under the assumption that all other errors (in particular random errors arising from measurement noise) were negligible.

Finally, it was shown that solving the fitted-synthesis equations by SVD also permits a natural interpretation of the reconstructed flux-vector, $\hat{\phi}(t)$, as a linear combination of orthogonal modes, a fundamental mode and harmonics. An accurate synthesis requires that all the modes relevant to the transient under study be present and be detectable. These modes are formed from linear combinations of the expansion functions included in the synthesis. They contain the same spatial and spectral information rearranged in a more convenient form. Their ordering and relative importance follow those of the singular values of $\mathbf{A}$. The procedure which consists in zeroing small singular values was therefore interpreted as the elimination of high-order harmonics, i.e. of a near-zero linear combination of expansion functions.

In Chapter 4, the methodology for testing the method and a first application are presented. Evidence of the fact that ill-conditioning and
roundoff errors are under complete control is given. The important question of how to generate and select an appropriate set of expansion-functions in a systematic way is addressed. The influence of the number of detectors and of their positions is also studied.

In Chapter 5, the effects of noise in the measurements are investigated in detail since these effects are suspected to be a potential source of difficulties. The results of complementary tests of the method are presented in Chapter 6.
Chapter 4

FIRST APPLICATIONS OF THE METHOD
AND REFINEMENT

4.1. Introduction.

In Chapter 2, a semi-experimental, nodal synthesis method has been presented. The essential idea of this method is a direct reconstruction of the instantaneous neutron-flux distribution from a linear combination of \( K \) precomputed, three-dimensional, nodal expansion-functions fitted to \( J \) local detector readouts. The solution of the \( J \)-by-\( K \) linear least-squares problem, \( \Delta \mathbf{T}(t) = \mathbf{C}(t) \), relies on the singular value decomposition of \( \Delta \). This procedure, which has been described in detail in Chapter 3, amounts to the reorganization of the expansion functions into a new set of “orthogonal modes” of well-defined magnitude. Near linear-dependences between the expansion functions lead to very small or “high-order” modes. When judged excessively small, these modes are zeroed to prevent any numerical difficulties from occurring.

In this chapter, the results of a first numerical evaluation of the fitted-synthesis method are reported. “Ideal” experimental conditions are assumed since the primary objective of the tests is to demonstrate the feasibility of the method, i.e. to assess its theoretical accuracy independently of other considerations.

The general methodology of the tests is first presented. An application of the method to a severe “operational transient” in a PWR is then described. In analyzing the results, advantage is taken of the physical insight gained into the method in Section 3.4. A natural refinement of the
method, including a procedure for selecting an appropriate set of expansion functions, is proposed and tested. Results of additional numerical experiments involving a reduced number of detectors on the one hand, and out-of-core detectors on the other hand, are also reported and analyzed.

Throughout this chapter, extensive use is made of graphical illustrations. In several instances, the analysis of, and comparisons between plots will provide invaluable information.

4.2. Testing Methodology.

4.2.1. The QUANDRY Code.

The main vehicle for the numerical simulations is the nodal diffusion code ‘QUANDRY’ developed at MIT in the late seventies for the analysis of light-water reactors [S6]. The code, written in FORTRAN, provides a numerical solution to both the steady-state and time-dependent, two-group, neutron-diffusion equations in 3-D Cartesian geometry. The theoretical basis of the program is the ‘Analytic Nodal Method’ [S6]. The distinctive characteristic of this nodal method is that the “coupling equations” (see Section 1.4) are obtained by an exact, analytical integration of the 1-D equations for the group surface-fluxes, rather than by an approximate expansion method. The only assumption made in the derivation is that of a quadratic spatial shape for the transverse-leakage terms.

Since its creation, the QUANDRY code has been extensively tested. Over the years, considerable experience has been accumulated with the use of the program at MIT [S6, S7, T1, G2, K2, A7, J1] and elsewhere [H13, E3]. Confidence has been established in its ability to predict nodal flux and power distributions accurately for both steady-state and transient conditions. Several options have been added to the original program, including an “adjoint-calculation” option, an “external-source” option, and a “quasi-static” option [O6, J1].
The unknowns of interest in QUANDRY are the node-averaged group fluxes, \( \phi_{gn} \), \( g = 1,2; \) \( n = 1,2,...,N \). Auxiliary unknowns are the net leakages defined as the difference between net currents on each face of a node in a given direction and in a given energy-group. The size of the matrix system of node-averaged group fluxes and net leakages is \( 8N \) by \( 8N \).

For steady-state problems, the Analytic Nodal Method leads to a system of equations having the form of either Eq. 1.11 (homogeneous system) or Eq. 1.8 (inhomogeneous system). For time-dependent problems, the system has the form of Eqs. 1.2 and 1.3.

The matrix elements, or “coupling coefficients”, are functions of the nodal, group cross sections and diffusion coefficients. They are influenced by the neutron flux through changes in the thermal-hydraulic conditions of the core. These nonlinearities and the large size of the system make the use of an iterative solution-technique imperative.

The version of the program used in this investigation still contains the original, simple, 1-D, ‘WIGL’ thermal-hydraulic model [S6]. In this model, fuel and moderator temperature distributions are computed from a given fission-source distribution and from a known inlet coolant-temperature (assumed uniform). From the moderator-temperature distribution and from the user-specified pressure (assumed constant and uniform), a moderator-density distribution is inferred by a routine which uses fits of water-property curves. Subsequent feedback on cross sections relies on a simple linear formula with node- and group-dependent coefficients.

The following is a brief description of the main variables and input parameters of the QUANDRY code. The algebraic detail of the Analytic Nodal Method is omitted since it is not relevant to this investigation.
a. Static Module.

The static module of the QUANDRY code receives as inputs:
- The core geometry;
- Nodal, group, homogenized diffusion-theory parameters for each material composition;
- Thermal-hydraulic data for the WIGL model, including coolant pressure, \( p \), and total mass flow-rate, \( \dot{m} \);
- Fuel-temperature, moderator-temperature, and moderator-density coefficients, as well as reference cross sections for the feedback model;
- Neutronic boundary-conditions on all sides;
- Thermal-hydraulic boundary-condition \((T_{\text{inlet}})\) on the bottom plane.

For homogeneous, criticality calculations, a normalization power-level must also be specified.
For nonhomogeneous problems, the external-source distribution must be supplied.

The heart of the QUANDRY static module is a three-level iterative calculation:

(i) Fission-source-term or "outer" iterations accelerated by Wielandt's method of eigenvalue shifting. Each outer iteration includes
- Modified, block, Gauss-Seidel or "inner" iterations,
- Evaluation of new thermal-hydraulic conditions,
- Cross-section feedback and coupling-coefficient updates;

(ii) Each inner iteration consists of
- A flux-matrix inversion by "flux" iterations, followed by
- Leakage updates (from the new fluxes and old leakages);

(iii) In the flux iterations, nodal, group fluxes are computed by a Cyclic Chebyshev Semi-Iterative method. The optimum relaxation parameter is computed from an estimate of the spectral radius of the iteration matrix.

The rate of convergence of the overall procedure is governed by the choice of a number of numerical parameters, including:
- The number of outer iterations per coupling-coefficient updates (typically four);
- The minimum number of inner iterations per outer iteration (usually one or two);
- The minimum and maximum numbers of flux iterations per inner iteration (\(\sim 1\) and 5 respectively);
- The required degree of (partial) convergence of the fluxes before the leakages are updated in any inner iteration (\(\sim 15\) to 30%);
- The maximum relative change in total fission-source for convergence of outer iterations (\(\sim 10^{-6}\));
- The maximum relative change in local power for convergence of the flux iterations (\(\sim 10^{-4}\)).

A balance must be found between some of these parameters to minimize computing cost.

Outputs of static calculations include:
- Node-averaged group fluxes, \(\phi_{gn}\), and directional net leakages;
- Node-averaged power-densities;
- Node-averaged fuel temperature, moderator temperature, and moderator density.

In criticality calculations, the eigenvalue, \(k_{eff}\), is also returned.

b. Transient Module.

The transient module of the code requires, in addition to the above data, a number of dynamic parameters for the neutronic and thermal-hydraulic models:
- Homogenized neutron-speeds, \(v_{gn}\);
- Number of delayed-neutron precursor families, \(I\);
- Delayed-neutron fractions, \(\beta_i\), and time-constants, \(\lambda_i\);
In QUANDRY, all neutrons (prompt and delayed) are assumed to be born in the fast group ($\chi_{P1} = \chi_{d1} = 1$, $\chi_{P2} = \chi_{d2} = 0$). As a consequence, $\beta_{i\text{eff}} = \beta_i$, $i = 1, 2, ..., I$. (See Eqs. 1.10 and 2.19.)

The nature of the transient, i.e. the external perturbations, must be fully specified. There are three possibilities:
- Cross section perturbations, $\Delta \Sigma_{g_n}(t)$, to simulate control-rod motions for instance;
- Change in inlet coolant temperature, $\Delta T_{\text{inlet}}(t)$;
- Change in total coolant flow-rate, $\Delta m(t)$.

Any transient calculation is preceded by a static calculation. The result of this static calculation determines the initial (equilibrium) conditions for the transient.

In time-dependent calculations, the same inner/flux iterative strategy as in the static module is used at every time-step to determine instantaneous, node-averaged, group fluxes and net leakages.

"Time intervals" are defined by the user. A new spectral radius for the flux-iteration matrix is estimated at the beginning of each time-interval. An appropriate splitting of the time-axis may therefore result in computing-time savings.

For each time-interval, the following parameters must be specified:
- Number of time-steps;
- Time-step size;
- Coupling-coefficient or matrix updating-frequency;
- Minimum number of inner iterations per time-step;
- Convergence threshold-value, $\varepsilon_\phi$, for the flux iterations. At any given time-step, convergence is achieved at the r-th flux iteration if $r \geq 4$ (minimum of four flux iterations per inner iteration) and

$$\varepsilon_\phi^{(r)}(t) < \varepsilon_\phi$$

where

$$\varepsilon_\phi^{(r)}(t) = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\phi_{2n}^{(r)}(t) - \phi_{2n}^{(r-1)}(t)}{\phi_{2n}^{(r-1)}(t)} \right|.$$  \hspace{1cm} (4.1)

N is the total number of nodes, including reflector nodes.
The choice of the above parameters strongly influences the overall computing-time. However, in this investigation, computing-cost will not be an overriding concern; instead, priority will be given to numerical accuracy.

There is no automatic time-step-adjustment algorithm in the version of QUANDRY currently in use at MIT [P6]. Considerations of accuracy require small values for both the time-step size (small temporal truncation-errors) and \( \varepsilon_\phi \) (good spatial convergence). These two values should not be chosen independently of each other. If a fairly small value is specified for \( \varepsilon_\phi \), the code may perform many more inner iterations per time-step than the user-specified minimum value. This is especially true when relatively large time-steps are used. The result is an increase in computing-time. In some cases, the final cost can be larger than if a smaller time-step had been used.

The code returns as outputs instantaneous values of node-averaged, group fluxes, \( \phi_{gn}(t) \), power densities, fuel temperatures, moderator temperatures, and moderator densities. Additional outputs include

(i) Total fission power, \( P(t) \), and
(ii) Values for point-kinetics quantities: amplitude function, \( n_{\text{eff}}(t) \), prompt neutron lifetime, \( \Lambda(t) \), and reactivity, \( \rho(t) \). The evaluation of these integral quantities require that weight functions be specified [H8, J1]. These weight functions may come from a preliminary, steady-state, adjoint calculation corresponding to some selected, reference reactor-condition.

4.2.2. Testing Procedure.

Figure 4.1 is a synoptic diagram of the testing procedure.
A reference solution to the transient under study, $\phi(t)$, is generated by the transient module of the QUANDRY code. (The directional net leakages are of no interest to the fitted synthesis.) Instantaneous, local group-fluxes are extracted from that reference solution and combined with user-specified detector weight-functions, $\Sigma^{(j)}$, to form the J detector signals:
\[ C^{(j)}(t) = \Sigma^{(j)T} \phi(t), \quad j = 1,2,...,J. \] (2.6)

Until Chapter 5, these signals are assumed to be noise-free. (The "Noise" and "Filters" boxes of Figure 4.1 are bypassed.)

The static module of QUANDRY is used to generate \( K \) expansion-functions, \( \psi^{(k)} \), corresponding to either critical conditions (fundamental \( \lambda \)-modes), or subcritical conditions with external sources.

The synthesis program receives as input the \( \Sigma^{(j)} \)'s and \( \psi^{(k)} \)'s and combines them to form the \( J \)-by-\( K \) matrix \( \mathbf{A} \) of Eq. 2.12. The general inputs of the synthesis code are:
- Number of energy groups (equal to two here);
- Geometry;
- Number of neutron detectors, locations, and characteristics (\( \Sigma^{(j)} \));
- Library of expansion functions, \( \psi^{(k)} \);
- Linear-dependence tolerance, \( \varepsilon \), for zeroing small singular values;
- Number of delayed-neutron precursor families, \( I \);
- Neutron, prompt and delayed energy-spectra (\( \chi_{p1} = \chi_{d1} = 1, \chi_{p2} = \chi_{d2} = 0 \), as in QUANDRY);
- Effective external source, \( q_{\text{eff}} \), if any, for computing the reactivity;
- Parameters for the noise-filtering routines;
- Detector output-signals, \( C^{(j)}(t) \).

In addition, if total power, \( \hat{P}(t) \), and prompt neutron lifetime, \( \hat{\Lambda}(t) \), are to be computed according to Eqs. 2.17 and 2.20, values for the homogenized, nodal, group, fission cross-sections, \( \Sigma_{fn}(t) \), must be supplied. In principle, these cross sections depend on time implicitly through feedback effects. However, this time-dependence will be neglected in the forthcoming applications.
As with the reference QUANDRY calculations, a vector of weight functions, \( \phi_0^* \), is also necessary for evaluating the point-kinetics parameters: \( \tilde{n}_{\text{eff}}(t), \tilde{\Lambda}(t), \) and \( \tilde{\rho}(t) \) (Eqs. 2.18, 2.20, and 2.25).

The main output of the synthesis code is the reconstructed flux-vector, \( \hat{\phi}(t) \), from which \( \tilde{n}_{\text{eff}}(t), \tilde{\Lambda}(t), \) and \( \tilde{\rho}(t) \) are derived.

Comparisons can be made between reference values, \( \phi(t), n_{\text{eff}}(t), \Lambda(t), \rho(t) \) on the one hand, and reconstructed values, \( \hat{\phi}(t), \tilde{n}_{\text{eff}}(t), \tilde{\Lambda}(t), \) and \( \tilde{\rho}(t) \) on the other hand.

In particular, the vector of absolute errors, \( \delta\phi(t) = \hat{\phi}(t) - \phi(t) \), can be calculated. This vector is the ultimate indicator of the performance of the fitted-synthesis method.

For an unambiguous evaluation of the individual components of \( \delta\phi(t) \), it is necessary that the elements of the reference vector, \( \phi(t) \), be computed with sufficient (spatial) numerical accuracy themselves. To this end, the following convergence-criterion is substituted to the “average-error” test of Eq. 4.1:

\[
\varepsilon^{(r)}_\phi(t) < \varepsilon_\phi \quad \text{where} \quad \varepsilon^{(r)}_\phi(t) = \max_{\text{All nodes}} \frac{\left| \phi^{(r)}_n(t) - \phi^{(r-1)}_n(t) \right|}{\phi^{(r-1)}_n(t)} .
\] (4.2)

Eq. 4.2 guarantees that convergence in the group-two neutron-flux at time \( t \) is achieved everywhere to within a certain tolerance, \( \varepsilon_\phi \). Such a criterion will presumably require more computational effort than Eq. 4.1 for the same value of \( \varepsilon_\phi \), but, again, minimizing the computing-cost of the reference solution is not the objective of this investigation.

Consistency of the overall procedure is guaranteed by the fact that the neutronic, thermal-hydraulic, and feedback models used for generating
the expansion functions are the steady-state forms of the corresponding, 
time-dependent models used for generating the reference solution.

4.2.3. Computing Equipment and Programs.

Although QUANDRY and the synthesis code could in principle run 
in parallel on two different computers, a mode of operation in two 
successive steps with a single computer was adopted for simplicity. In a 
first step, the QUANDRY reference-calculation is performed and 
instantaneous values for the node-averaged, group fluxes, $\phi_{gn}(t)$, total 
power, $P(t)$, and point-kinetics quantities, $n_{eff}(t)$, $\Lambda(t)$, $\rho(t)$, are saved on 
disk. In a second step, the synthesis program is executed. At every 
"sampling time" (sampling period = $\tau$), the necessary reference 
information is read from disk.

The computer used for all the simulations is a DEC MicroVAX III 
workstation.

The synthesis code is written entirely in FORTRAN 77. It uses 
double-precision arithmetic (8-bit exponent, 55-bit mantissa for floating-
point numbers), as opposed to QUANDRY which, for the most part, uses 
single-precision arithmetic (8-bit exponent, 23-bit mantissa). No special 
consideration had to be given to the dynamic storage of variables with the 
synthesis program since its memory requirements are very modest.

The singular value decomposition of the matrix $A$ in the synthesis 
code can be performed by SVD routines from either the LINPACK or 
EISPACK packages [D8, G4]. Both routines were tested and performed 
equally well. The LINPACK routine was found to be marginally slower, 
but less susceptible of overflows than the EISPACK routine. Scaling 
routines have since been incorporated to avoid any risk of overflow.
In solving the rectangular system of equations, \( \mathbf{A} \mathbf{T}(t) = \mathbf{C}(t) \), the pseudo-inverse of \( \mathbf{A} \), \( \mathbf{A}^\dagger \), is never formed explicitly (Eq. 3.43). Instead, a more efficient, back-substitution-type calculation is carried out based on Eqs. 3.35, 3.36, and 3.37.

4.3. Reactor Model and Nuclear Instrumentation.

A single reactor model is used throughout this investigation. The selected model is that of a typical, Westinghouse, 1000-MWe PWR.

It has already been mentioned (Section 2.5.2) that few PWRs are presently equipped with fixed, prompt-responding, in-core flux-sensors. This may make it difficult, if not impossible, to apply the fitted-synthesis method to these reactors. In this respect, the choice of a BWR model would have been preferable. However, the version of QUANDRY currently in use at MIT lacks the capability to simulate BWR thermal-hydraulic conditions properly. On the other hand, there is evidence that the thermal-hydraulic and feedback models embodied in QUANDRY, although simple, still permit a realistic simulation of PWR static and transient conditions, at least within a limited range of variations for the fuel and moderator temperatures. For these reasons, a PWR model was selected.

The core model contains 3,474 nodes. The radial reflector is explicitly represented by 1,728 additional nodes. When the quarter-core radial symmetry is taken into account, these numbers reduce to 1,008 and 450 respectively. There are 18 axial subdivisions or "planes", each 20-cm thick, numbered from one (bottom) to 18 (top). Complete description of the reactor model is given in Appendix A1.

For the purpose of testing the fitted-synthesis method, the reactor is assumed to be equipped with a total of \( J = 80 \), identical, fixed, fast-responding, power-range neutron-detectors (of the fission-chamber of
SPND type) distributed in four equally-spaced planes (# 2, 7, 12, 17), as shown in Figures 4.2a and 4.2b.

Figure 4.2a. X-Y Cross-Sectional View of the Reactor Showing the 20 Strings of In-Core Neutron-Detectors.
Figure 4.2b. X-Z Cross-Sectional View of the Reactor and its Nuclear Instrumentation.
There are 20 vertical strings. For simplicity, no variation in the axial spacing of the detectors is considered. Symmetrical positions across the X and Y axes, as well as across the main diagonals, are deliberately avoided to maximize the information returned by the instrumentation.

Assemblies containing part- or full-length control-rods are not equipped with detectors. However, observe in Figure 4.2a that each rodded node (rod-banks C and D) has at least one instrumented neighbor.

The detectors are assumed to be sensitive to thermal neutrons only, and to be contained within a single node, which requires a length of less than 20 cm. The homogenization volume is chosen to be that node only. The characteristics weight-functions, $\Sigma^{(j)}$, of Eq. 2.4e are given by

$$
\Sigma_{gn}^{(j)} = \delta_{g2} \delta_{nnj} \begin{cases}
g = 1,2; \\
j = 1,2,\ldots,80; \\
n = 1,2,\ldots,5202;
\end{cases}
$$

(4.3)

where $n_j$ is the number of the node containing the j-th detector. In these conditions, $C^{(j)}(t)$ in Eq. 2.6 is a direct measure of the group-two flux in node $n_j$.

In subsequent tests (Sections 4.7), the positions of the detectors will be varied. However, their characteristics will remain unchanged, i.e. Eq. 4.3 will always hold. The results of the tests are essentially unchanged when the detectors have a mixed spectral response.

4.4. Description of the Transient and Reference Solution.

The transient studied in this chapter is referred to as "operational transient" to distinguish it from two other transients which will be analyzed in Chapter 6, and which will be referred to as "coolant inlet-temperature transient" and "startup transient".
The reactor is initially \( t < 0 \) critical and in equilibrium at \( P = 20 \% \) \( P_{\text{nominal}} \), with \( P_{\text{nominal}} = 3,338 \text{ MWth} \). All fission cross sections have been previously divided by the eigenvalue \( k_{\text{eff}} = 1.045361 \).

A rather extreme transient is initiated at \( t = 0 \) by the simulated withdrawal of control-rod banks C and D of Figures 4.2a,b at the constant speed of \( v_{\text{rod}} = 2 \text{ cm s}^{-1} \). In QUANDRY, these rod motions are simulated by linear perturbations of cross sections and diffusion coefficients in successive nodes in the \( Z \) direction. (Values for the \( \Delta \Sigma_{g_n} \)'s and \( \Delta D_{g_n} \)'s can be found in Appendix A1.)

At \( t = 60 \text{ s} \), rod bank C reaches the top of the core, while bank D continues its upward motion. All rod motions cease at \( t = 120 \text{ s} \). Rod bank D remains partially inserted. The transient is followed until \( t = 180 \text{ s} \). Figure 4.3 illustrates this sequence of rod motions.
The coolant inlet-temperature and total mass flow-rate are maintained constant at 555 K and 3.87 \times 10^3 \text{ kg s}^{-1} respectively. Six families of delayed-neutron precursors are used in the simulation. The total delayed-neutron fraction is $\beta = \beta_{\text{eff}} = 0.00650$. 

Figure 4.3. Simulated Withdrawal of Rod Banks C and D.
Since quadrant symmetry is preserved during the transient, only a quarter of the core needs to be modeled for both the reference and synthesis calculations. Apart from considerations of computing cost, there are two important, beneficial consequences to this size reduction. The first consequence is that memory requirements for the QUANDRY calculation are kept within reasonable limits. The second consequence is that the ratio of instrumented nodes to power-producing nodes is increased from 80/3,474 to 80/1,008. For each of the four instrumented planes, the increase is from 20/193 to 20/56. The influence of the number of detectors and their locations on the performance of the fitted-synthesis method will be studied in Section 4.7.

The numerical parameters used in the QUANDRY calculation are listed in Table 4.1.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Number of Time-Steps</th>
<th>Time-Step Size (s)</th>
<th>$\varepsilon_\phi$ in Eq. 4.2</th>
<th>Number of Time-Steps per Matrix Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 40 s]</td>
<td>320</td>
<td>0.125</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[40 s, 120 s]</td>
<td>640</td>
<td>0.125</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[120 s, 130 s]</td>
<td>40</td>
<td>0.25</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[130 s, 180 s]</td>
<td>200</td>
<td>0.25</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
</tbody>
</table>

The value $\varepsilon_\phi = 10^{-3}$ essentially guarantees that the first three digits of $\phi_{2n}(t)$ are numerically correct. (It is recognized that the actual, physical, relative uncertainty is certainly larger than $10^{-3}$.)
With an imposed, minimum number of two inner iterations per time-step, the total CPU time is approximately 10 hrs 52 min. (This number reduces to 6 hrs 25 min with $\varepsilon_\phi = 10^{-2}$, and to 4 hrs 30 min when the original convergence test, Eq. 4.1, is used with $\varepsilon_\phi = 10^{-3}$.)

The reference flux values are stored every $\tau = 0.25$ s for subsequent use by the synthesis program.

Adjoint fluxes and net leakages corresponding to the initial reactor conditions are chosen as weight functions in the evaluation of the point-kinetics quantities. (Other weight-functions could have been used.) These adjoint fluxes and net leakages were computed by the static module of QUANDRY.

Figure 4.4a shows plots of QUANDRY-computed total fission-power and reactivity versus time. Thermal-hydraulic feedback effects limit the reactivity insertion, and the reactor ultimately reaches an equilibrium state at a power level approximately equal to 120 % $P_{\text{nominal}}$. 

161
A correction scheme more elaborate than the one in the original code was developed and incorporated into QUANDRY to reduce as much as possible control-rod "cusping effects" [S6, J7]. These nonphysical effects are caused by the imperfect simulation of rod motion within partially-rodved nodes. Despite the correction, some residual effects remain and small cusps are still visible in the reactivity curve in Figure 4.4a. However, the effect was considerably reduced, as evidenced by Figure 4.4b, the analog of Figure 4.4a when no cusping-correction is applied. In Figure 4.4b, cusps are even visible in the power curve.
Figure 4.4b. QUANDRY-Computed Reactivity and Total Fission-Power Versus Time (Operational Transient, No Cusping-Correction Applied).

The cusping-correction scheme does not require any preparatory work such as precomputed tables of correction factors. It operates "online" and, in addition, does not cause any appreciable increase in total computing-time.

It was initially feared that these nonphysical cusping effects could hamper a proper evaluation of the synthesis procedure by causing artificially large, local distortions in flux shape. However, numerical tests subsequently proved that this was not the case.
In fact, it should be pointed out that cusping effects are considerably more pronounced in the reactivity curve than in any local-flux profile, as some of the graphs below will demonstrate.

Plots of local values of fuel and moderator temperature are shown in Figure 4.6. Node # 893 is located in Plane # 12, next to the central axis.

![Graph showing fuel and coolant temperature over time](image)

Figure 4.5. QUANDRY-Computed Fuel Temperature and Coolant Temperature in Node # 893 Versus Time (Operational Transient).

For this particular node, the increase in fuel temperature is \( \sim 350 \text{ K} \), whereas the increase in coolant temperature is \( \sim 46 \text{ K} \).
In view of these fairly large temperature-variations, one may legitimately question the validity of the linear feedback model with constant coefficients. However, possible errors in the feedback model equally affect reference solution, expansion functions, and detector signals. Therefore, the consistency of the tests is preserved.

4.5. First Application of the Fitted-Synthesis Method.

4.5.1. Expansion Functions.

A total of 26 expansion functions $\Psi^{(k)}$ are generated from criticality calculations corresponding to
- two bracketing power-levels: $P = 20\% \ P_{\text{nominal}}$ and $P = 120\% \ P_{\text{nominal}}$;
- and 13 different control-rod patterns: $t = 0$ s, $t = 10$ s, $t = 20$ s, ..., $t = 120$ s. (See Figure 4.3.)

No expansion function is computed for configurations in which nodes are partially rodded.

These calculations are performed by the static module of QUANDRY. Thermal-hydraulic feedback effects are included in all calculations. The numerical parameters used guarantee that at least the first four digits of any local-flux value are numerically correct. Each calculation requires approximately 3 minutes of CPU time.

Table 4.2 shows the 26 computed eigenvalues, $k_{\text{eff}}$. These eigenvalues are plotted as a function of control-rod position in Figure 4.6.
Table 4.2. Eigenvalues $k_{\text{eff}}$ Corresponding to Two Power-Levels and 13 Control-Rod Patterns.

<table>
<thead>
<tr>
<th>Rod Pattern at Time (s)</th>
<th>$k_{\text{eff}}$ at $P = 20% P_{\text{nominal}}$</th>
<th>$k_{\text{eff}}$ at $P = 120% P_{\text{nominal}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.045361</td>
<td>1.033623</td>
</tr>
<tr>
<td>10.0</td>
<td>1.046660</td>
<td>1.035048</td>
</tr>
<tr>
<td>20.0</td>
<td>1.048083</td>
<td>1.036502</td>
</tr>
<tr>
<td>30.0</td>
<td>1.049531</td>
<td>1.037905</td>
</tr>
<tr>
<td>40.0</td>
<td>1.050905</td>
<td>1.039206</td>
</tr>
<tr>
<td>50.0</td>
<td>1.052142</td>
<td>1.040383</td>
</tr>
<tr>
<td>60.0</td>
<td>1.053218</td>
<td>1.041430</td>
</tr>
<tr>
<td>70.0</td>
<td>1.054142</td>
<td>1.042368</td>
</tr>
<tr>
<td>80.0</td>
<td>1.054928</td>
<td>1.043215</td>
</tr>
<tr>
<td>90.0</td>
<td>1.055589</td>
<td>1.043963</td>
</tr>
<tr>
<td>100.0</td>
<td>1.056135</td>
<td>1.044608</td>
</tr>
<tr>
<td>110.0</td>
<td>1.056572</td>
<td>1.045146</td>
</tr>
<tr>
<td>120.0</td>
<td>1.056903</td>
<td>1.045571</td>
</tr>
</tbody>
</table>
Figure 4.6. Eigenvalue $k_{eff}$ Versus Control-Rod Pattern at Time $t$ for Two Different Power-Levles.

4.5.2. SVD and Zeroing of Small Singular Values.

The 80-by-26 matrix $\Delta$ of Eq. 2.12 is formed by combining the $J = 80$ weight functions, $\Sigma^{(j)}$, of Eq. 4.3 with the $K = 26$ expansion-functions, $\Psi^{(k)}$. No normalization of the $\Psi^{(k)}$'s, i.e. column weighting of $\Delta$ is performed.
The SVD of that matrix is then computed. The 26 singular values are listed in Table 4.3. They have been normalized to \( s_1 \), the largest singular value.

Table 4.3. Normalized Singular-Value Spectrum of the 80-by-26 Matrix \( A \).

<table>
<thead>
<tr>
<th>i</th>
<th>( s_i / s_1 )</th>
<th>i</th>
<th>( s_i / s_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000 ( 10^0 )</td>
<td>14</td>
<td>7.232 ( 10^{-5} )</td>
</tr>
<tr>
<td>2</td>
<td>8.863 ( 10^{-2} )</td>
<td>15</td>
<td>6.658 ( 10^{-5} )</td>
</tr>
<tr>
<td>3</td>
<td>4.769 ( 10^{-2} )</td>
<td>16</td>
<td>3.346 ( 10^{-5} )</td>
</tr>
<tr>
<td>4</td>
<td>1.385 ( 10^{-2} )</td>
<td>17</td>
<td>2.656 ( 10^{-5} )</td>
</tr>
<tr>
<td>5</td>
<td>5.859 ( 10^{-3} )</td>
<td>18</td>
<td>1.212 ( 10^{-5} )</td>
</tr>
<tr>
<td>6</td>
<td>4.839 ( 10^{-3} )</td>
<td>19</td>
<td>9.196 ( 10^{-6} )</td>
</tr>
<tr>
<td>7</td>
<td>2.066 ( 10^{-3} )</td>
<td>20</td>
<td>6.586 ( 10^{-6} )</td>
</tr>
<tr>
<td>8</td>
<td>1.245 ( 10^{-3} )</td>
<td>21</td>
<td>2.648 ( 10^{-6} )</td>
</tr>
<tr>
<td>9</td>
<td>5.524 ( 10^{-4} )</td>
<td>22</td>
<td>1.761 ( 10^{-6} )</td>
</tr>
<tr>
<td>10</td>
<td>4.727 ( 10^{-4} )</td>
<td>23</td>
<td>1.198 ( 10^{-6} )</td>
</tr>
<tr>
<td>11</td>
<td>3.602 ( 10^{-4} )</td>
<td>24</td>
<td>5.327 ( 10^{-7} )</td>
</tr>
<tr>
<td>12</td>
<td>2.202 ( 10^{-4} )</td>
<td>25</td>
<td>1.940 ( 10^{-7} )</td>
</tr>
<tr>
<td>13</td>
<td>1.270 ( 10^{-4} )</td>
<td>26</td>
<td>1.513 ( 10^{-7} )</td>
</tr>
</tbody>
</table>

The condition number of \( A \) is \( \kappa(A) = s_1 / s_{26} = 1 / 1.513 \times 10^{-7} \gg 1 \), an indication of very severe ill-conditioning. Some of the expansion functions are "sensed" by the detectors as near linear combinations of others.
Since the relative numerical accuracy in the elements of the $\Psi^{(k)}$'s is approximately $5 \times 10^{-5}$, a linear-dependence tolerance of $\varepsilon = 5 \times 10^{-5}$ should be imposed for consistency reasons, independently of other considerations (See Sections 3.3.3.d and 3.4.2). This would lead to the zeroing of the 11 smallest singular values, i.e. to the elimination of the 11 smallest orthogonal modes, $(s_i, \chi_i, m_i, \Psi^{(i)})$, from the synthesis formula:

$$\hat{\phi}(t) = \sum_{i=1}^{R} \Psi^{(i)} \, T_{LS}^{(i)}(t).$$  \hspace{1cm} (3.71)

However, these small modes are temporarily retained to verify some of the theoretical predictions of Chapter 3.

Twenty-six synthesis-calculation are performed with 1, 2, 3, ..., and 26 modes included. In each case the results are compared with the reference solution.

Figure 4.7 summarizes these results. The maximum relative error over all nodes and all time-steps is shown as a function of the number of modes retained in the synthesis.
Figure 4.7. Maximum Relative Error in Group-Two Flux Versus Number of Singular Values Retained in the Synthesis (Operational Transient, 26 Expansion Functions).

Any local- or average-error chart would have a ‘U’ shape similar to that of Figure 4.7. This general profile merely indicates that large errors occur when either too few or too many modes are retained. The cause of the errors is loss of information in the first case, and amplification of roundoff errors in the second case. Errors appear to be smallest when from 10 to 18 modes are used.

With $\varepsilon = 5 \times 10^{-5}$, 15 modes are retained. A “reduced” matrix $\Delta'$ is obtained, having condition number $\kappa(\Delta') = 1 / 6.658 \times 10^{-5}$ and rank $R' = 15$. 
4.5.3. Results.

Application of the fifteen-mode synthesis formula yields a relative residual \( \| E_{LS}(t) \| / \| C(t) \| \) in Eq. 3.70 smaller than \( 3.5 \times 10^{-4} \) throughout the transient, as shown in Figure 4.8. This small value with respect to unity is a sign that the least-squares fit of the 15 orthogonal modes to the 80 detector readouts is fairly successful.

Figure 4.8. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
The Euclidean norm of the least-squares solution vector, \( \| \mathbf{T}_{LS}(t) \| \), is also plotted in Figure 4.8. Note that \( \| \mathbf{T}_{LS}(t) \| \) does not exhibit any “wild” variations characteristic of the solution of an ill-conditioned system, but instead remains in the fairly narrow interval [0.6, 1.3].

Evaluation of the error vector, \( \Delta \phi(t) = \hat{\phi}(t) - \phi(t) \), makes it possible to compute a node-averaged relative error in group-g flux:

\[
\varepsilon^{(\text{ave})}_g(t) = \frac{1}{N_p} \sum_{n=1}^{N_p} \left| \frac{\Delta \phi_{g_n}(t)}{\phi_{g_n}(t)} \right| .
\]  

(4.4)

The summation in Eq. 4.4 is over the \( N_p \) power-producing nodes only. The absolute values prevent any error cancellation.

The maximum relative error in group-g flux, defined as

\[
\varepsilon^{(\text{max})}_g(t) = \frac{\Delta \phi_{g_{n_0}}(t)}{\phi_{g_{n_0}}(t)} ,
\]

(4.5a)

where

\[
\left| \frac{\Delta \phi_{g_{n_0}}(t)}{\phi_{g_{n_0}}(t)} \right| = \max_{\text{(Power Producing Nodes Only)}} \left| \frac{\Delta \phi_{g_n}(t)}{\phi_{g_n}(t)} \right| ,
\]

(4.5b)

can also be determined.

Plots of \( \varepsilon^{(\text{ave})}_g(t) \) and \( \varepsilon^{(\text{max})}_g(t) \) versus time are shown in Figures 4.9a and 4.9b for \( g = 1 \) and \( g = 2 \) respectively.
Figure 4.9a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
The maximum values of $\varepsilon_g^{(ave)}(t)$ over all time-steps are less than 0.5% and 0.8% in group one and two respectively. Note the larger errors in the first 60 seconds of the transient, when both rod-banks C and D are simultaneously moving.

A detailed analysis of the error vector, $\delta \phi(t)$, reveals that most errors are actually confined within a $\pm 2\%$ band except for nodes in which perturbations occur. This is illustrated by Figures 4.10a and 4.10b which show the relative-error distributions at $t = 52$ s, $\delta \phi_{gn}(52)/\phi_{gn}(52)$, as a function of node number, n, for group one and group two respectively.
Reflector nodes are not included. Node numbering is such that the left-hand side of the figures corresponds to the bottom of the reactor, whereas the right-hand side corresponds to the top of the reactor.

Figure 4.10a. Relative Error in Group-One Flux at t=52s Versus Node Number (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.10b. Relative Error in Group-Two Flux at t=52s Versus Node Number (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).

Observe that the nodes in which the largest relative errors occur are “clustered” in groups of three or four. (This is equally true of the absolute errors.) It is easily verified that each cluster of three nodes belongs to rod-bank C, whereas each cluster of four nodes belongs to rod-bank D. Fairly large errors tend also to occur in the immediate neighborhood of some of these rodded nodes.

In Figures 4.10a,b, time is “frozen” and the focus is on the spatial distribution of the errors. Alternatively, one may focus on a particular node and plot errors as a function of time.
The following nodes were selected as “representative” nodes:

- Node # 893, in Plane # 12, next to the central axis. This node contains a neutron detector;
- Node # 651, in Plane # 9, two nodes away from the central axis;
- Node # 649, in Plane # 9 also, on the central axis. This node is perturbed (i.e. becomes partially rodded) between $t = 50$ s and $t = 60$ s;
- Node # 495, in Plane # 7, in the reflector;
- Node # 244, in Plane # 4, on the central axis. The tip of a bank-D rod moves across this node between $t = 0$ and $t = 10$ s.

The positions of these nodes are indicated in the cutaway view of Figure 4.11. A few other nodes have also been numbered for future reference.
Figure 4.11. Cutaway View of the Quarter-Core Reactor Model Showing the Positions of the Representative Nodes.
Also indicated in Figure 4.11 is $t_D$ defined as

$$t_D = \frac{1}{v_{\text{rod}}} Z_D,$$  \hspace{1cm} (4.6)

where $Z_D$ is the elevation of the tip of rod-bank D measured from its initial, inserted position.

Figures 4.12 through 4.16 show plots of reference and reconstructed group-two fluxes, $\phi_{2n}(t)$ and $\hat{\phi}_{2n}(t)$ respectively, as a function of time for the five representative nodes. Relative errors, $\delta\phi_{2n}(t) / \phi_{2n}(t)$, are also plotted for a better appreciation of the quality of the flux synthesis.
Figure 4.12. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.13. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.14. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.15. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.16. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).

Corresponding plots could be given for the fast group, but, as one may intuitively expect, it is in the thermal group that the largest errors occur.

Fairly large errors (~ 20 %) occur in Node # 649 as well as in other perturbed nodes during the first half of the transient, when the two rod-banks are simultaneously moving. However, for the rest of the nodes, errors are less than a few percent in both groups throughout the transient.
The largest error in group one is 9.6%; it occurs in Node # 973 at t = 35 s.

The largest error in group two is 24.9%; it occurs in Node # 325 at t = 9 s (Figure 4.11).

The total CPU time for the 720 time-steps of the synthesis calculation is 312 s (excluding the point-kinetics calculations and the comparisons with reference values, but including some I/O time). This is considerably less than the time required for generating the reference solution, even when a relaxed spatial-convergence criterion is used in QUANDRY.

Computation of the SVD of the 80-by-26 matrix $\Delta$ requires 1.9 s.

Reference and reconstructed total power, amplitude function, prompt neutron lifetime, and reactivity are displayed in Figures 4.17 through 4.20.
Figure 4.17. Reference and Reconstructed Amplitude-Function Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.18. Reference and Reconstructed Total Power Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.19. Reference and Reconstructed Prompt Neutron Lifetime Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).
Figure 4.20. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 26 Expansion Functions, 11 Singular Values Zeroed).

In all cases, the agreement between the QUANDRY and synthesis calculations is excellent. In particular, the error in reactivity is less than one millibeta throughout the transient.

Note the slight drift in \( \hat{P}(t) \) and \( \hat{\Lambda}(t) \) caused by the neglect of the time-variations in the fission cross sections. A similar drift appears in nodal power-profiles. If variations in \( \Sigma_f \) were taken into account, no drift would appear, and the relative error in \( \hat{P}(t) \) would be essentially identical to that in the amplitude function, \( \hat{n}_{\text{eff}}(t) \), i.e. less than 0.1 %. However, updating \( \Sigma_f \) would require coupling of the synthesis code with both a
thermal-hydraulic model and a feedback model. In view of the fairly small errors in \( \hat{P}(t) \) and \( \hat{\Lambda}(t) \) (0.8 \% maximum deviation in \( \hat{P}(t) \)), there is little incentive in introducing such models (and complications) for the sole purpose of correcting these errors.

Observe that there is no apparent drift in the reactivity curve, \( \hat{\rho}(t) \).

The reconstructed prompt-neutron lifetime remains essentially constant, equal to 2.973 \( 10^{-5} \) s.

In summary, this first, straightforward application of the fitted-synthesis method to a PWR model equipped with 80 in-core detectors demonstrates that instantaneous, nodal, group fluxes can be reconstructed with very good accuracy, except possibly for a few nodes in which local perturbations are introduced.

Although these results may be considered satisfactory, a slight modification of the method makes it possible to improve them substantially, and, at the same time, to explain the larger errors in a few isolated nodes. This modification is described in the next section.


4.6.1. Description.

Repeated synthesis calculations show that retaining between 10 and 18 singular values \( s_i \) does not lead to any significant improvement nor deterioration in the results. This observation is not too surprising if one notices that, in Table 4.3, a number of singular values \( s_i \) are such that \( s_i / s_1 \) is approximately equal to the relative accuracy in the expansion functions \( \psi^{(k)} \), i.e. \( 10^{-4} \) to \( 10^{-5} \). There is no clear cutoff-value or obvious zeroing-threshold in the \( \{ s_i \} \) spectrum.
In fact, the SVD technique is sometimes criticized for producing such “clusters” of small singular values with the result that it is not easy to decide exactly how many of these clustered $s_i$’s should be eliminated and how many should be retained [D8].

The detailed analysis of the synthesis method performed in Chapter 3 can be invoked to understand why SVD is not to blame for this difficulty.

In this particular problem, the SVD of $A$ converts the 26 expansion functions, $\psi^{(k)}$, into 26 orthogonal basis-functions, $\psi^{(i)}$, of well-defined importance, $s_i$.

For $i > 18$, the $\psi^{(i)}$ modes are physically meaningless because $s_i / s_1$ is less than the numerical accuracy in the $\psi^{(k)}$. These small modes are therefore legitimately discarded;

For $i \leq 18$, on the other hand, the $\psi^{(i)}$'s may contain some physical information, and certainly do so for $i < 10$. However, because of the fairly large number of expansion functions included in the synthesis (relative to the number of detectors), this information is not “clearly” reflected in the $A$ matrix: Too many high-order modes are present. As a result, it is difficult for the neutron sensors to detect distinctively small flux-perturbations in the direction ($\mathbf{u}_i$) of these high-order modes. (The effect would be even worse if measurement noise were present.)

In a certain sense, too much detailed information has been “lumped” into $A$, and part of that information cannot be easily recovered by SVD and by the limited number of neutron detectors. The problem is with the way the matrix $A$ is formed, not with its SVD. The reason why consequences on $\hat{\Phi}(t)$ are not too serious is simply that the clustered modes, $\psi^{(i)}$, $i = 10, 11, \ldots, 18$, correspond either to relatively small, extended perturbations, or to local distortions in flux shape affecting only a limited number of nodes.

The above interpretation suggests a remedy to the “confusing” situation in $A$ and to the related singular-value clustering problem: It is to limit the information content of the $A$ matrix by retaining at any instant
only a few appropriate expansion-functions, namely those $\Psi^{(k)}$'s that bracket the instantaneous reactor-conditions "the most closely".

For instance, in the particular transient under study, for $t$ in the interval [20 s, 30 s], only the four expansion-functions computed for the two control-rod patterns at $t = 20$ s and $t = 30$ s and for the two different power-levels $P = 20\% P_{\text{nominal}}$ and $P = 120\% P_{\text{nominal}}$ should be retained.

This procedure is equivalent to subdividing the time-axis into "time-intervals" or "time-domains". In each of these time-domains, only a selected subset of the library of expansion-functions is used. A new matrix $A$ is formed at the beginning of each time-interval. (Note that, at this point, updated values for the $\Sigma^{(i)}$'s can be used.) The SVD of $A$ is computed and used until the end of that interval.

In effect, by this modification, the original, continuous synthesis which uses a single "thick" matrix containing all the information is replaced by a discontinuous synthesis [Y1, K7] which uses instead a series of "thin" matrices, each containing only a small fraction of that information.

Since each matrix $A$ in this time-discontinuous synthesis has only a few columns, its SVD yields only a few modes $\Psi^{(i)}$. The consequences of neglecting one of these modes are now much easier to predict. The situation is also much less "confusing" for the detectors since they only need to find the proper combination of a fundamental mode and a few harmonics.

Figure 4.21 is a graphical interpretation of this discontinuous synthesis which uses selected subsets of expansion-functions (for four different power-levels and six distinct control-rod patterns). It is assumed that rod motion is the only external perturbation.
Each intersection point on the grid corresponds to a precomputed, static expansion-function, $\psi^{(k)}$. Safety considerations, operating restrictions, or mere physical limits result in a bounded domain containing all acceptable dynamic trajectories. One such trajectory is sketched. (The fact that a trajectory may encounter an intersection point on the grid at some instant $t_1$ does not imply that $\hat{\phi}(t_1) = \phi(t_1)$ in general.)

From Figure 4.21, the notion of “bracketing” expansion-functions may be given a more precise meaning: Bracketing expansion-functions are those static flux-shapes which correspond to the four corners of the mesh containing the instantaneous state of the reactor represented by the point $(Z(t), P(t))$.  

Figure 4.21. Graphical Interpretation of the Time-Discontinuous Synthesis Method.
In the continuous synthesis, all points on the grid are retained independently of the actual, instantaneous state of the reactor. In the discontinuous synthesis, on the other hand, only the four corner-points of the mesh which contains the instantaneous state are retained, which is intuitively more satisfactory. In some sense, the information is “presorted” before being presented to the synthesis code.

It is sometimes argued that synthesis methods are of limited practical interest because the exact solution of the problem must be known beforehand to identify bracketing conditions. Figure 4.21 suggests a way around this difficulty for the present method: The reactor instrumentation can be used to determine the location of the instantaneous state of the reactor on the grid.

In the simple case of Figure 4.21, this determination requires only estimates of the instantaneous power-level, $P(t)$, (computed from neutron-flux measurements), and control-rod positions, $Z(t)$. From this information, the four nearest intersection-points on the grid, $(P_1, Z_1)$, $(P_1, Z_2)$, $(P_2, Z_1)$, and $(P_2, Z_2)$ can be identified. The corresponding expansion-functions can then be extracted from the library and input to the synthesis program.

By this method of “on-line subset selection”, one does not need to know the solution of the problem beforehand to identify bracketing functions. All that is needed is to identify the proper mesh on the grid, and this can be done using (real-time) information from the plant. Note that the experimental values of $P(t)$ and $Z(t)$ need not be very accurate. At the same time, the mesh size should be consistent with measurement uncertainties. In case of doubt about the actual position along a particular axis, it is always possible to retain an extra expansion-function in the synthesis.

In practice, this shape-selection procedure can be easily automated without requiring repeated searches through the entire library of expansion-functions. Assume for instance that, at some time $t$, power-level, $P(t)$, and control-rod pattern, $Z(t)$, are within known intervals
denoted \([P_1, P_2]\) and \([Z_1, Z_2]\) respectively (Figure 4.21). At later times, new values of \(P\) and \(Z\) are received from the reactor instrumentation. \(P\) is compared with \(P_1\) and \(P_2\), while \(Z\) is compared with \(Z_1\) and \(Z_2\). If, for any reason, it is detected that \(Z\), for example, leaves the interval \([Z_1, Z_2]\), and enters the neighboring interval \([Z_2, Z_3]\), the two expansion-functions corresponding to the points \((P_1, Z_1)\) and \((P_2, Z_1)\) are simply replaced by those corresponding to the points \((P_1, Z_3)\) and \((P_2, Z_3)\). Subsequent values of \(Z\) are compared with \(Z_2\) and \(Z_3\) instead of \(Z_1\) and \(Z_2\). An analogous, independent logic applies to \(P\).

This idea of a discontinuous fitted-synthesis using only carefully-selected expansion-functions at any time-step has been applied to the operational transient of the previous section. Results and comparisons with the continuous synthesis are presented below. The generalization of the above discussion and of Figure 4.21 to more than two variables is deferred to Section 4.6.3.

4.6.2. Application.

The synthesis program was modified to allow the use of multiple subsets of expansion functions in a given transient. Since the code is purely experimental and primarily aimed at assessing the performance of the time-discontinuous-synthesis idea, no extensive changes were performed. In particular, no automatic shape-selection algorithm was implemented. Instead, provision was made for defining “time intervals” in which the appropriate expansion-functions can be directly specified, as explained in the previous section. The code switches automatically from one subset of basis-functions to another whenever a new time-interval is reached.

For the reactor model and operational transient described in Sections 4.3 and 4.4, twelve time-intervals are defined, as shown in Table 4.4. At the start of each interval, a new matrix \(A\) is formed from the four “nearest” expansion-functions, and its SVD is computed. (In the following,
even though $A$ changes with time, no time-dependence will be shown explicitly either in $A$ or in its SVD for notational simplicity.) In this particular example, there is no difficulty in identifying these nearest expansion-functions since no intermediate power-level exists. The singular values of the twelve, 80-by-4 matrices are reported in Table 4.4.

**Table 4.4. Time-Intervals for the Discontinuous Synthesis and Normalized Singular-Value Spectra of the 80-by-4 Matrices.**

| Time-Interval | $s_i / s_1$
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i = 1, 2, 3, 4$</td>
</tr>
<tr>
<td>[0 s, 10 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.296 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>1.037 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>3.277 $10^{-4}$</td>
</tr>
<tr>
<td>[10 s, 20 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>1.901 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>5.492 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.023 $10^{-4}$</td>
</tr>
<tr>
<td>[20 s, 30 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>1.755 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>1.440 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>2.929 $10^{-4}$</td>
</tr>
<tr>
<td>[30 s, 40 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.703 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>1.250 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>3.181 $10^{-4}$</td>
</tr>
<tr>
<td>[40 s, 50 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.036 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>9.507 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>1.923 $10^{-4}$</td>
</tr>
<tr>
<td>[50 s, 60 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>1.742 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>4.730 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.408 $10^{-4}$</td>
</tr>
</tbody>
</table>

| Time-Interval | $s_i / s_1$
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i = 1, 2, 3, 4$</td>
</tr>
<tr>
<td>[60 s, 70 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>1.922 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>6.159 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.334 $10^{-4}$</td>
</tr>
<tr>
<td>[70 s, 80 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.457 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>9.089 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.792 $10^{-4}$</td>
</tr>
<tr>
<td>[80 s, 90 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>3.012 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>1.031 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>3.485 $10^{-4}$</td>
</tr>
<tr>
<td>[90 s, 100 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.646 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>8.721 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.509 $10^{-4}$</td>
</tr>
<tr>
<td>[100 s, 110 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.373 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>7.009 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.007 $10^{-4}$</td>
</tr>
<tr>
<td>[110 s, 180 s]</td>
<td>1.000 $10^{+0}$</td>
</tr>
<tr>
<td></td>
<td>2.289 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>7.062 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>2.585 $10^{-4}$</td>
</tr>
</tbody>
</table>
Observe that the condition numbers, $\kappa = s_1 / s_4$, of the successive matrices remain such that

$$1.923 \times 10^{-4} \leq \frac{1}{\kappa} \leq 3.485 \times 10^{-4}$$

throughout the transient. Therefore, $1 / \kappa$ is always larger than the numerical accuracy in the expansion functions, $5 \times 10^{-5}$, and, contrary to the case of the continuous synthesis, none of the modes $(s_i, \psi^{(i)})$ needs to be eliminated on numerical grounds.

The relative residual, $\|E_{LS}(t)\| / \|C(t)\|$, obtained by application of the discontinuous-synthesis method is shown in Figure 4.22. Also plotted in that figure is the Euclidean norm of the least-squares solution vector, $\|T'_{LS}(t)\|$.
Figure 4.22. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).

The relative residual never exceeds 0.0018, a sign of a good least-squares fit. However, comparison with Figure 4.8 reveals that this value is larger than that obtained earlier with the continuous synthesis. One might therefore be tempted to conclude that the reconstructed flux-vector, \( \hat{\phi}(t) \), will be worse than before.

Actually, this is not the case. As explained in Section 3.4.1, this larger residual simply means that the least-squares fit of the measurement vector, \( \mathbf{C}(t) \), to the projection "plane", \( \mathbf{R}(A) \), is "not as good" (Figure 3.8).
In fact, this deterioration in the residual of the least-squares fit could have been predicted. The dimension, \( R \), of the column space of \( \mathbf{A} \), \( \mathbf{R}(\mathbf{A}) \), is smaller than before: Each matrix \( \mathbf{A} \) is of rank \( R = 4 \), whereas the single matrix of the continuous synthesis had rank \( R = 15 \) (after zeroing of 11 of its singular values). The reason for the larger residual is that, loosely speaking, it is easier for the 80-element vector \( \mathbf{C}(t) \) to be close to a space which has dimension 15 than to a space which has dimension only 4.

One should also recall from Section 3.4.1 that:

(i) A relative residual small with respect to unity is only a necessary condition for a good synthesis;

(ii) The relative residual is a reliable measure of the systematic errors, \( \| \delta \phi(t) \| / \| \phi(t) \| \), caused by the synthesis approximation only when an “appropriate” set of expansion functions is used.

The continuous synthesis satisfies the necessary condition of Item i better than the discontinuous synthesis.

On the other hand, the condition of Item ii is better fulfilled by the discontinuous synthesis than by the continuous synthesis. This gives the discontinuous synthesis a decisive advantage over the continuous synthesis, as proved by the results presented below.

To facilitate comparisons between the results of the two types of syntheses, figure numbers have been listed in Table 4.5.
Table 4.5. Figure Numbering for the Results of the Time-Continuous and Time-Discontinuous Syntheses.

<table>
<thead>
<tr>
<th>Description</th>
<th>Time-Continuous Synthesis</th>
<th>Time-Discontinuous Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average and Maximum Relative Errors in Group-One Flux Versus Time</td>
<td>Figure 4.9a</td>
<td>Figure 4.23a</td>
</tr>
<tr>
<td>Average and Maximum Relative Errors in Group-Two Flux Versus Time</td>
<td>Figure 4.9b</td>
<td>Figure 4.23b</td>
</tr>
<tr>
<td>Relative Errors in Group-One Flux at t = 52 s Versus Node Number</td>
<td>Figure 4.10a</td>
<td>Figure 4.24a</td>
</tr>
<tr>
<td>Relative Errors in Group-Two Flux at t = 52 s Versus Node Number</td>
<td>Figure 4.10b</td>
<td>Figure 4.24b</td>
</tr>
<tr>
<td>Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time</td>
<td>Figure 4.12</td>
<td>Figures 4.26, 4.31</td>
</tr>
<tr>
<td>Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time</td>
<td>Figure 4.13</td>
<td>Figures 4.27, 4.32</td>
</tr>
<tr>
<td>Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time</td>
<td>Figure 4.14</td>
<td>Figures 4.28, 4.33</td>
</tr>
<tr>
<td>Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time</td>
<td>Figure 4.15</td>
<td>Figures 4.29, 4.34</td>
</tr>
<tr>
<td>Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time</td>
<td>Figure 4.16</td>
<td>Figures 4.30, 4.35</td>
</tr>
<tr>
<td>Reference and Reconstructed Amplitude Function Versus Time</td>
<td>Figure 4.17</td>
<td>Figure 4.36</td>
</tr>
<tr>
<td>Reference and Reconstructed Total Power Versus Time</td>
<td>Figure 4.18</td>
<td>Figure 4.37</td>
</tr>
<tr>
<td>Reference and Reconstructed Prompt Neutron Lifetime Function Versus Time</td>
<td>Figure 4.19</td>
<td>Figure 4.38</td>
</tr>
<tr>
<td>Reference and Reconstructed Reactivity Versus Time</td>
<td>Figure 4.20</td>
<td>Figure 4.39</td>
</tr>
</tbody>
</table>
Node-averaged relative errors ($\varepsilon_g^{(\text{ave})}(t)$ of Eq. 4.4) in group-one and group-two neutron-fluxes are plotted in Figure 4.23a and 4.23b respectively. Maximum errors ($\varepsilon_g^{(\text{max})}(t)$ of Eq. 4.5) are also shown.

Figure 4.23a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.23b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).

$\varepsilon_{g}^{(\text{ave})}(t)$ is less than 0.3 % in both groups. Note the abrupt variations when the synthesis programs switches time interval.

The peak-values for $\varepsilon_{g}^{(\text{max})}(t)$ are only - 2.8 % in group one ($t = 54 \text{ s}$, Node # 1398), and - 5.5 % in group two ($t = 10 \text{ s}$, Node # 325) respectively.

Comparisons with Figures 4.9a,b indicate that mean and maximum errors have been reduced by an average factor of approximately three with respect to the continuous synthesis.
Figures 4.24a and 4.24b show the distributions of relative errors at $t = 52$ s. These figures are the analogs of Figure 4.10a and 4.10b.

![Graph showing relative error vs. node number](image)

**Figure 4.24a.** Relative Error in Group-One Flux at $t=52$ s Versus Node Number (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.24b. Relative Error in Group-Two Flux at $t=52s$
Versus Node Number (Operational Transient, 80
In-Core Detectors, No Singular Value Zeroed).

Observe that the dispersion of the points in Figures 4.24a,b is very
different from that obtained with the continuous synthesis. Not only has
the magnitude of the errors been reduced, but the number of isolated
"clusters" too.

This last point is extremely interesting. In fact, close observation of
Figures 4.24a,b reveals that there are only two clusters of markedly
isolated points: One is made of three points, the other of four points.
These two clusters correspond to rodded nodes in which perturbations
occur at $t = 52s$. The cluster of four points includes Node # 649 from
rod-bank D (Plane # 9), whereas the cluster of three points contains Node #
1398 from rod-bank C (Top Plane). Apart from these two clusters and, to a lesser degree, their immediate neighbors, no node is significantly affected at that particular instant. In particular, no anomalously-large error occurs in the rodded nodes which are perturbed at some other time during the transient.

This is in sharp contrast with the continuous-synthesis results (Figures 4.10a,b) in which relatively large errors were observed in essentially all perturbed nodes at \( t = 52 \text{ s} \) (many isolated clusters of three and four points), whether a perturbation was actually occurring in those nodes at \( t = 52 \text{ s} \) or not.

Part of the explanation for this difference in error-distribution patterns has already been given.

Recall that, in the continuous synthesis, all the information provided by the 26 expansion-functions, \( \psi^{(k)} \), is contained in a single matrix \( A \). This information is rearranged by SVD into orthogonal modes, \( \psi^{(i)} \), some of them corresponding to local perturbations occurring at various times during the transient. In the synthesis process (Eq. 3.71), the neutron detectors are not sufficiently numerous and/or adequately positioned to determine accurately the proper weights, \( T_{L5}^{(i)}(t) \), to be assigned to these high-order modes. Errors in the coefficients translate into errors in \( \hat{\phi}(t) \). This can happen not only at any time \( t \), but also in any node perturbed during the transient.

In the discontinuous synthesis, on the other hand, only a few appropriate modes, \( \psi^{(i)} \), are present in the matrix \( A \) at any instant. Not only is it easier for the detectors to find the proper "blending" of these few modes, but it is also impossible for other "inappropriate" modes to be excited since they are absent from the set of expansion functions. The synthesis is "much cleaner", and spurious errors do not risk to occur.

In this particular problem, it is interesting to examine the evolution of each of the four terms, \( \psi^{(i)} T_{L5}^{(i)}(t) \), \( i = 1,2,3,4 \), in the modal expansion, Eq. 3.71.
The four modal coefficients, $T_{LS}^{(i)}(t)$, are plotted together in Figure 4.25.

![Modal Coefficients Graph](image)

**Figure 4.25.** Modal Coefficients Versus Time (Operational Transient, 80 In-Core Detectors).

Observe that the coefficient $T_{LS}^{(1)}$ of the fundamental mode, $\psi^{(1)}$, behaves in a much smoother fashion than the other $T_{LS}^{(i)}$’s. No discontinuity in slope is visible for that coefficient.
The negative sign of $T_{LS}^{(i)}$ should not be interpreted as a physical inconsistency. In fact, this negative sign can be changed into a positive one by reversing the signs of all the elements of $\Psi^{(1)}$. The signs of the other $T_{LS}^{(i)}$'s are equally arbitrary.

This sign arbitrariness can be traced back to the $v_i$ and $u_i$ vectors from the SVD of $A$ (Section 3.3.3.b). Indeed, recall from Section 3.4.2 that:

(i) Any pair $(v_i, u_i)$ may be replaced by $(-v_i, -u_i)$ without affecting the SVD of $A$;

(ii) The components of the $v_i$ vectors enter the calculation of the $\Psi^{(i)}$ mode:

$$\Psi^{(i)} = \sum_{k=1}^{K} \Psi^{(k)} v_i^{(k)}, \quad i = 1,2,...,K, \quad (3.75)$$

while the components of the $u_i$ vectors enter the evaluation of the corresponding coefficient, $T_{LS}^{(i)}(t)$:

$$T_{LS}^{(i)}(t) = \begin{cases} \frac{1}{s_i} u_i^T C(t), & i = 1,2,...,R, \\ 0, & i = R+1,R+2,...,K. \end{cases} \quad (3.83)$$

The smooth behavior of $T_{LS}^{(i)}(t) = \frac{1}{s_1} u_1^T C(t)$ is indirect evidence that the measurement vector, $C(t)$, has a smoothly- and slowly-varying component in the $u_1$ direction of the projection "plane" $R(A)$.

This confirms that the mode $\Psi^{(1)}$, or, more completely, $(s_1, v_1, u_1, \Psi^{(1)})$, has all the attributes of a fundamental mode (except possibly for its sign). In the synthesis (Eq. 3.71), the principal contribution comes from this fundamental term. The terms involving the other modes, $\Psi^{(i)}$,
i = 2, 3, 4, may be viewed as corrections of decreasing magnitude (s_i). The coefficients, T^{(i)}_{LS}(t), of these higher-order modes (or perturbation modes) vary more rapidly than T^{(i)}_{LS}(t).

To appreciate better the contribution of each term, \psi^{(i)} T^{(i)}_{LS}(t), in Eq. 3.71, three types of syntheses were performed, the first one involving all four modes (four-mode synthesis), the second one only the three largest modes (three-mode synthesis), and the third one only the two largest modes (two-mode synthesis). The results of these three flux-syntheses are presented in Figures 4.26 through 4.30 for the five representative nodes # 893, 651, 649, 495, and 244 (Figure 4.11).

In comparing the results of the four-mode, three-mode, and two-mode syntheses with one another, one may be tempted to interpret the magnitude of the first neglected term as representing a bound on, or even an estimate of, the magnitude of the error, \delta \phi(t), in \hat{\phi}(t). There is indeed some legitimacy in this interpretation. However, one should remember that, as emphasized in Chapter 2, no rigorous, theoretical statement can be made concerning the error vector, \delta \phi(t), even when the synthesis formula is conveniently rearranged in the form of a modal expansion such as Eq. 3.71. Nevertheless, comparisons between "truncated" expansions (R = 2, 3, 4 in Eq. 3.71) are always of interest and would be extremely valuable in the absence of a reference solution.
Figure 4.26. Modal Contributions to the Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors).
Figure 4.27. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Detectors).
Figure 4.28. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors).
Figure 4.29. Modal Contributions to the Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Detectors).
Figure 4.30. Modal Contributions to the Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Detectors).

Two conclusions can be drawn from these figures:

(i) As expected, the agreement between the QUANDRY and synthesis solutions improves systematically as a larger number of modes is included in the synthesis;

(ii) As predicted, high-order modes do not contribute significantly to the sum (Eq. 3.71), except in or near rodded nodes (e.g. Node # 649 or Node # 244), and only when perturbations occur in or near these nodes.

The agreement between the four-mode synthesis and the reference solution is so good that no discrepancy is visible in the above figures. The three-mode reconstruction appears to be extremely accurate too.
Sharp transitions are nevertheless expected every time the synthesis code switches from one set of expansion functions to another. These transitions or discontinuities can be seen in the relative-error curves of Figures 4.31 through 4.35.

Figure 4.31. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.32. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.33. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.34. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.35. Reference and Reconstructed Group-Two Flux in Node \# 244 Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).

Comparisons with Figures 4.12 through 4.17 confirm that the overall reduction in the errors mainly comes from a significant improvement in the perturbed nodes such as Node \# 649.

Note that, for this particular node, the relative error takes on larger values (not exceeding 1.7 \% in magnitude though) for \( t \) in the interval [50 s, 60 s], i.e. when the node is partially rodded. These errors would have been only marginally larger (~2.0 \% maximum) if the highest-order term, \( \psi^{(4)} T_{LS}^{(4)}(t) \), had been ignored, but substantially bigger (66 \% peak-value) had the term \( \psi^{(3)} T_{LS}^{(3)}(t) \) been also neglected (Figure 4.28).
Similar comments can be made for Node # 244 which is perturbed between t = 0 s and t = 10 s.

The results have deteriorated slightly in instrumented nodes such as Node # 893. This is consistent with the earlier discussion concerning the degradation in the least-squares fit and the larger relative residual. Remarkably, errors in Node # 893 remain quite small between t = 80 s and t = 90 s, even though a perturbation occurs during that time interval in neighbor node # 892.

Note the appreciable variations in group-two flux experienced by the reflector node # 495.

The total CPU time for the entire 180-second transient is 80 s. This is significantly less than the 312 s required by the continuous synthesis. These computing-time savings are a direct consequence of the reduced number of expansion-functions (four instead of 26). The $A$ matrices have fewer columns; their construction, the solution of the least-squares-fit problem by SVD, and the subsequent flux-reconstruction all require less computing-time. (Chapter 5 will show that a reduced number of columns in $A$ is also desirable when noise is present in the measurements.)

Each of the twelve SVD computations requires only a fraction of a second. The total time for these calculations is actually less than the 1.9 s needed to compute the SVD of the 80-by-26 matrix of the continuous synthesis.

To complete the comparisons, total fission-power and point-kinetics quantities as resulting from the discontinuous flux-synthesis are presented in Figure 4.36 through 4.39.
Figure 4.36. Reference and Reconstructed Amplitude Function Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.37. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.38. Reference and Reconstructed Prompt Neutron Lifetime Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).
Figure 4.39. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Detectors, No Singular Value Zeroed).

As with the continuous synthesis, there is again a slight drift in reconstructed power caused by the assumption of constant fission cross sections.

To appreciate better the quality of the results and, at the same time, the severity of the transient, a third type of synthesis was performed using the same reactor model and instrumentation as before, but with a single expansion-function for the entire transient (i.e. K = 1). The initial flux-shape was chosen for this “single-mode” synthesis or point-kinetics-type calculation.
Figure 4.40 shows the resulting relative residual and Euclidean norm of $T'_{LS}(t)$ plotted as a function of time.

Figure 4.40. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 80 In-core Detectors, One Expansion Function).

Observe that the relative residual, which previously did not exceed 0.0018, now takes on values as large as 0.18. Large errors are therefore expected in the synthesized flux-vector, $\hat{\phi}(t)$.

Three figures, Figures 4.41, 4.42, and 4.43, suffice to confirm this anticipated result.
Figure 4.41. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 80 In-Core Detectors, One Expansion Function).
Figure 4.42. Relative Error in Group-Two Flux at $t=52s$
Versus Node Number (Operational Transient, 80
In-Core Detectors, One Expansion Function).

Figure 4.41 shows that the node-averaged relative error increases rapidly from 0 to approximately 21% for $t > 120$ s. A similar rise is visible in the magnitude of the maximum error which ultimately reaches 70%.

Figure 4.42 gives a “snapshot” of the error distribution in group-two flux at $t = 52$ s. Relative errors greater than 20% in magnitude occur in many nodes.
Figure 4.43. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Detectors, One Expansion Function).

For Node # 649 of Figure 4.43, the group-two flux is systematically underestimated by approximately 13 % during the second half of the transient.

These large errors in flux-shape are indicative of the severity of the flux tilt that accompanies the rod withdrawal.

Table 4.6 compares the results of the three types of syntheses.
Table 4.6. Comparisons Between Discontinuous-Synthesis, Continuous-Synthesis, and Point-Kinetics Results (80 In-Core Detectors).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group-One Flux</td>
<td>- 2.8 % (#1398, ( t=54 )s)</td>
<td>- 9.6 % (#973, ( t=35 )s)</td>
<td>- 61.3 % (#1398, ( t=180 )s)</td>
</tr>
<tr>
<td>Group-Two Flux</td>
<td>- 5.5 % (#325, ( t=10 )s)</td>
<td>- 24.9 % (#325, ( t=9 )s)</td>
<td>- 70.4 % (#1398, ( t=180 )s)</td>
</tr>
<tr>
<td>Amplitude Function</td>
<td>- 0.052 %</td>
<td>0.058 %</td>
<td>3.3 %</td>
</tr>
<tr>
<td>Total Power</td>
<td>0.80 %</td>
<td>0.79 %</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Prompt Neutron Lifetime</td>
<td>- 1.1 %</td>
<td>- 1.1 %</td>
<td>- 1.1 %</td>
</tr>
<tr>
<td>Reactivity</td>
<td>0.72 m( \beta )</td>
<td>0.84 m( \beta )</td>
<td>7.0 m( \beta )</td>
</tr>
<tr>
<td>Total CPU Time (( \tau = 0.25 )s)</td>
<td>80 s</td>
<td>312 s</td>
<td>44 s</td>
</tr>
</tbody>
</table>

In all calculations, integral quantities benefit from error cancellations. This effect is particularly helpful to the single-mode synthesis (point-kinetics calculation).

In summary, comparisons between discontinuous and continuous syntheses have evidenced the unquestionable superiority of the time-discontinuous synthesis. The results obtained by this method are remarkably accurate considering the severity of the transient studied. One may view this as additional numerical evidence in support of the basic
assumption of time-synthesis methods, namely that an appropriate linear combination of carefully selected, static, three-dimensional, neutron flux-shapes \textit{can} be a good approximation to the true instantaneous flux-shape.

The comparisons have also confirmed that the key to the success of the fitted synthesis is in the choice of the basis-functions. To obtain good results, it is necessary that the proper basis-functions be included in the synthesis. The results of the tests show that it is preferable to \textit{exclude} from the synthesis any "irrelevant" flux-shape, i.e. shapes corresponding to reactor conditions which are very distant from the instantaneous state, and that one should strive to limit the set of basis-functions used at any instant to an appropriate minimum.

This is best achieved by the time-discontinuous synthesis because only expansion-functions closely bracketing the instantaneous state of the reactor are retained at any time-step. This procedure "helps" the least-squares fit by automatically eliminating irrelevant modes (or directions $u_i$ in measurement space), particularly high-order perturbation-modes which may otherwise be given nonzero weights in the synthesis, and hence cause nonnegligible errors locally. At the same time, limiting the number of expansion functions makes the interpretation of the results easier. Additional benefits include reduced computing-cost and less sensitivity to measurement noise (See Chapter 5).

One may think that the results could be further improved by refining the splitting of the time-axis, i.e. by computing expansion-functions for reactor conditions with partially-rodded nodes, and by having the discontinuous-synthesis program switch more frequently from one subset of basis-functions to another. However, this will not be done for at least three reasons.

The first reason is that the magnitude of the errors is already acceptably small. In fact, it is within the uncertainties caused by modeling and parametric errors (nodal model and homogenized diffusion-theory parameters).

The second reason is that a discontinuous synthesis using expansion-functions computed for closer reactor-conditions may not yield improved results. The higher-order modes which may appear will have small
magnitude \( (s_a) \), and examination of Table 4.4 suggests that they will probably have to be zeroed for numerical reasons.

The third reason is that it is desirable to minimize the preparatory work and bookkeeping associated with the expansion functions, as explained in the next section.

4.6.3. Generalization.

In practice, application of the time-discontinuous fitted-synthesis method described in Section 4.6.1 is likely to be complicated by the fact that there will be more independent variables to be monitored than just total power and control-rod positions. Boron concentration (PWR), total coolant flow-rate, inlet coolant-temperature, are examples of other important variables which influence the neutron flux shape. On a longer time-scale, xenon concentration and even fuel exposure would have to be taken into consideration too.

Expansion functions should be computed for different values of these variables in the range of physically-realizable or allowed variations. In general, if the number of independent variables is \( D \), Figure 4.21 becomes a \( D \)-dimensional picture.

This generalization immediately brings up a number of difficulties with the method.

A first difficulty is with the generation of the library of expansion-functions. If two bracketing expansion-functions are computed per dimension, one obtains a library of \( 2^D \) basis-functions. In fact, more than \( 2^D \) expansion-functions will probably be needed since having only two extreme flux-shapes available would certainly not be sufficient for certain variables (e.g. control-rod pattern). These calculations may represent a substantial preparatory work. Regular updates of the library will also be necessary. However, part of this work could certainly be automated.
A second difficulty is with the number of expansion functions, $K$, needed at any instant in the time-discontinuous synthesis. This number should remain as small as possible since performance is expected to deteriorate as $K$ approaches $J$, the number of detectors. There are many situations where $K$ need not be as large as $2^D$, for instance when some variables do not change or vary very slowly. Unfortunately, one cannot exclude transients in which several perturbations occur simultaneously, causing changes in many variables, and forcing the incorporation of a large number of expansion-functions in the synthesis, at least temporarily.

A third difficulty with the method is that it may not be possible to measure certain physical quantities in real-time (~ 1 s). For example, estimating the boron concentration in a PWR requires several minutes [B3, M9, J2]. In such circumstances, if real-time performance is essential, one has no other alternative than to include extra expansion-functions in the synthesis to be certain that the instantaneous state of the reactor is bracketed.

Although none of these difficulties appears insurmountable, it is important to bear in mind their implications, which come in addition to other independent limitations arising from an imperfect or noisy instrumentation (Chapter 2).

4.7. Influence of the Number and Positions of the Neutron Detectors.

It may be argued that the results of the numerical tests of the time-discontinuous synthesis-method were favorably influenced not only by the fact that $D$ was equal to only two, but also by the presence of a large number of neutron detectors well-distributed throughout the reactor core, and by helpful symmetries.

The objective of this section is to demonstrate that the number and positions of the detectors are not as crucial parameters as the expansion
functions. They simply need to be "reasonably" chosen. This is fortunate in the sense that there is considerably less flexibility in the detectors than in the expansion functions.

Several synthesis-calculation are performed for the same reactor model and operational transient as before. In one series of calculations, a reduced in-core instrumentation is used. In another series, out-of-core detectors only are used.

In all the tests, the same twelve subsets of expansion functions as those used previously are supplied to the discontinuous-synthesis code. It was proved in Section 4.6 that accurate results can be obtained with these twelve subsets and with 80 in-core detectors. Any simulation using a different set of detectors but the same subsets of expansion-functions should therefore permit appreciating any negative impact of the instrumentation on the results.

4.7.1. Discontinuous Fitted-Synthesis Using Four In-Core Detectors.

The reactor model described in Section 4.3 is no longer assumed to be instrumented with 80 in-core neutron-sensors, but with only \( J = 4 \) such sensors. These four detectors are those of the string positioned next to the central axis, in the X-Z plane, in Figure 4.2b. This string contains Node # 893 (Figure 4.11).

This assumption of a reactor equipped with only one string of detectors, or, equivalently, having only one operable string left, is admittedly not realistic. However, the objective of the present test is not to simulate realistic experimental conditions, but instead to demonstrate that, even in extreme situations, the time-discontinuous fitted-synthesis method is still capable of performing adequately when accurate detector-signals are available.
It is again assumed that the detectors are prompt-responding, ideal, identical, and yield reliable, unbiased signals. Their characteristic weight-functions, $\Sigma^{(j)}$, are given by Eq. 4.3 as before, i.e. $C^{(j)}(t)$ is a direct measure of the local, node-averaged, group-two neutron-flux.

The operational transient of Section 4.4 constitutes a very severe test for the discontinuous fitted-synthesis method when only four detectors are available. In fact, since the method uses $K = 4$ expansion-functions in every time-interval, the system $\hat{A} \ T(t) = C(t)$ (Eq. 2.14) is no longer overdetermined but square: $\hat{A}$ is a 4-by-4 matrix.

It has been pointed out in Section 3.4.1 that such a situation where $J \leq K$ is generally undesirable because the fitted-synthesis method has no extra degree of freedom to discriminate between "good" and "bad" expansion-functions. However, the procedure of on-line selection of the basis-functions reduces the need for such a discriminative function. As already mentioned, the expansion functions are known to be "appropriate".

The singular value decompositions of the twelve successive matrices $\hat{A}$ are computed by the synthesis code. The condition numbers, $\kappa = s_1 / s_4$, were all such that:

$$4.453 \times 10^{-5} \leq \frac{1}{\kappa} \leq 2.549 \times 10^{-4}.$$  

The lower bound is approximately equal to the relative accuracy in the expansion functions. No singular value was therefore zeroed.

In these conditions, all 4-by-4 matrices, $\hat{A}$, have full rank: $R = 4$. The system $\hat{A} \ T(t) = C(t)$ has an exact solution, $T(t) = \hat{A}^\dagger C(t) = \hat{A}^{-1} C(t)$, and the residual vector, $E(t)$, is nil.

This is confirmed by Figure 4.44a which shows that the relative residual, $\|E(t)\| / \|C(t)\|$, is zero to within machine accuracy ($\epsilon_m = 2^{-55} \approx 3 \times 10^{-17}$).
Figure 4.44a. Euclidean Norm of Solution-Vector and Relative Residual Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).

Although a perfect synthesis would give such a zero residual, one would rather have $E(t) \neq 0$ here so as to have some indication about the performance of the present synthesis (which is known to be imperfect). This would be especially important if no reference solution were available. One natural way to obtain a feeling for the systematic errors caused by the synthesis approximation is to eliminate the highest-order mode, $\psi^{(4)}$, in Eq. 3.71 by setting $s_4$ to zero in the SVD of $A$. As explained in Chapter 3, the effect of this zeroing is to make $A$ of rank three instead of four, i.e. to reduce the dimension of the projection “plane”, $R(A)$. In these conditions, the system $A T(t) = C(t)$ does not have an exact solution anymore.
However, its minimum-norm least-squares solution, $T_{LS}(t)$, can be found. If the relative residual of the new, fitted, three-mode expansion is not small with respect to unity, then poor performance can be expected for the four-mode synthesis.

![Graph showing Relative Residual and Euclidean Norm vs Time](image)

**Figure 4.44b.** Euclidean Norm of Minimum-Norm Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, Four In-Core Detectors, One Singular Value Zeroed).

It is reassuring to find out that the relative residual of the three-mode synthesis remains very small ($\leq 0.0013$) with respect to unity throughout the transient.
From the reference solution, node-averaged and maximum relative errors, $\varepsilon^{(\text{ave})}(t)$ and $\varepsilon^{(\text{max})}(t)$, can be computed. They appear in Figures 4.45a and 4.45b for group one and group two respectively.

Figure 4.45a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).
Figure 4.45b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).

The fairly small magnitude of the average and maximum errors in both figures is remarkable. Node-averaged errors are less than 0.8 % in both groups throughout the transient. Maximum errors never exceed 8.4 % (group one) and 17.6 % (group two) in magnitude.

Equally remarkable is the very limited dispersion of the errors in reconstructed nodal group-fluxes, as illustrated by Figures 4.46a and 4.46b for $t = 52$ s.
Figure 4.46a. Relative Error in Group-One Flux at t=52s Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).
Notice again the presence of the same pair of isolated clusters of three and four points as in Figures 4.24a and 4.24b. Note also the detrimental effect that these larger errors have on neighbor nodes.

As in the simulations using 80 detectors, it is enlightening to compare the results of four-mode, three-mode, and two-mode fitted-syntheses with one another and with the reference solution. Comparing such truncated modal expansions (Eq. 3.71 with \( K < 4 \)) with one another, and examining the corresponding relative residuals, would provide extremely useful information in the absence of a reference solution.
Comparisons were made for representative Nodes # 651 and 649. The results appear in Figures 4.47 and 4.48 respectively.

Figure 4.47. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Detectors).
Figure 4.48. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Detectors).

The general profiles are similar to those obtained with 80 detectors (Figures 4.27 and 4.28), the term \( \psi^{(4)} T_{LS}^{(4)}(t) \) representing again a small correction to an already very accurate, three-mode synthesis.

Relative errors in group-two flux in three representative nodes, Nodes # 893, 651, and 649, are shown in Figures 4.49, 4.50, and 4.51.
Figure 4.49. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeros).
Figure 4.50. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).
Figure 4.51. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Detectors, No Singular Value Zeroed).

As expected from the exact nature of the fit, the group-two flux in Node # 893 is reconstructed without error. It is identical to the detector signal in that node.

The relative error in Node # 651 is always less than 1 % in magnitude, and takes on its maximum value when the tip of rod-bank D reaches the plane containing that node. This occurs between $t = 50 \text{ s}$ and $t = 60 \text{ s}$. Note that there was no similar error-peak with the 80 detectors (Figure 4.32).
For Node # 649, perturbed between \( t = 50 \text{ s} \) and \( t = 60 \text{ s} \), the error curve is similar to that obtained with 80 detectors (Figure 4.33), except for a higher peak of \(-14\%\) instead of \(-1.7\%\) previously.

Additional tests were performed with a varying number of in-core detectors (between 4 and 80). Similarly good results were obtained with only marginal improvements or deteriorations. The only exception to this general observation was in extreme, "unreasonable" situations, for example when all the detectors were positioned in a single plane.


One may argue that placing the neutron detectors inside the core, next to regions where rod-motions occur, favors the detection and synthesis of high-order modes, and hence leads to a better synthesis than if the instrumentation were placed far away from the perturbed regions.

The results of the following test prove that this is only partially true.

The PWR model of Section 4.3 is now assumed to be deprived of any in-core flux-sensors, and to be equipped instead with a "side instrumentation" consisting of 48 neutron-detectors located in reflector regions. These detectors are arranged in 12 identical strings of four, with the same axial spacing as before (Figure 4.2b). The locations of these 12 strings is shown in Figure 4.52.
Figure 4.52. X-Y Cross-Sectional View of the Reactor Model Showing the 12 Strings of Out-of-Core Neutron-Detectors.

As in the positioning of the 80 in-core detectors of Figure 4.2a, symmetries across the X and Y axes and across the main diagonals are deliberately avoided.
The condition numbers, $\kappa = \frac{s_1}{s_4}$, of the twelve 48-by-4 matrices $A$ formed by the synthesis program are such that

$$9.467 \times 10^{-5} \leq \frac{1}{\kappa} \leq 3.297 \times 10^{-4}.$$ 

The value $9.467 \times 10^{-5}$ is above the roundoff threshold of $\sim 5 \times 10^{-5}$. Therefore, all four modes can be retained in the synthesis.

In this overdetermined problem with appropriately-selected expansion-functions, the relative residual of the least-squares fit, $\frac{\|E_{LS}(t)\|}{\|C(t)\|}$, gives a reliable measure of the systematic errors caused by the synthesis approximation, $\frac{\|\delta \phi(t)\|}{\|\phi(t)\|}$. (See Section 3.4.1.)

Figure 4.53 is a plot of this relative residual as a function of time.
Figure 4.53. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).

The maximum value of \( \|E_{LS}(t)\| / \|\mathbf{C}(t)\| \) is 0.0012 << 1, indicative of a very good fit and suggestive of a successful synthesis.

This is confirmed by the analysis of the error vector, \( \delta \phi(t) \). As can be seen in Figures 4.54a,b, the node-averaged error, \( e^{(ave)}_g(t) \), is always less than 0.4 % in both energy-groups. At the same time, peak-values for the maximum error, \( e^{(max)}_g(t) \), are only - 3 % in group one and - 6 % in group two.
Figure 4.54a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).
Figure 4.54b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).

The dispersion of the errors is analogous to that obtained in the tests using 80 and 4 in-core detectors (Figure 4.24a,b and Figure 4.46a,b). Most points fall within a narrow band about the zero axis, except for isolated clusters of three and four points corresponding to nodes in which perturbations occur.

Plots of four-, three- and two-mode syntheses are compared in Figures 4.55 and 4.56 for two representative nodes, Node # 651 and Node # 649.
Figure 4.55. Modal Contributions to the Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Detectors).
Figure 4.56. Modal Contributions to the Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Detectors).

The high-order term, $\psi^{(4)} T_{LS}^{(4)}(t)$, appears again as a small, barely-visible correction to the three-mode synthesis. On the other hand, neglecting the term $\psi^{(3)} T_{LS}^{(3)}(t)$ causes appreciable errors in Node # 649 and other rodded nodes.

These results demonstrate that the synthesis method is capable of recovering from the side instrumentation small perturbation-modes ($\psi^{(4)}$) originating in the core.
Part of the explanation for this remarkable performance is in the fact that some of the detectors are positioned not too far away from moving rods, as can be seen in Figure 4.52. They are therefore able to sense perturbations, at least partially.

The other reason is in the use of only four, carefully-selected expansion-functions at any time-step, with the consequence that the space \( \mathbf{R}(\Delta) \) is "tailored" to the instantaneous conditions of the reactor, as explained in Section 3.4.2: \( \mathbf{R}(\Delta) \) contains the perturbation direction \( \mathbf{u}_4 \), but no other "irrelevant" perturbation-direction. If the projection of the measurement vector, \( \mathbf{C}(t) \), onto \( \mathbf{R}(\Delta) \) has some nonzero component, \( \mathbf{u}_4^T \mathbf{C}(t) \), in the direction \( \mathbf{u}_4 \), the mode \( \psi^{(4)} \) will be present in the synthesis with weight \( T^{(4)}_{LS}(t) = \frac{1}{s_4} \mathbf{u}_4^T \mathbf{C}(t) \) (Eq. 3.83).

Figures 4.57 through 4.61 show the group-two flux reconstructed by the four-mode synthesis for the five representative nodes of Figure 4.11. Corresponding reference-curves and relative errors are also plotted. Node # 893 is no longer instrumented, but reflector Node # 495 now contains a detector.
Figure 4.57. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).
Figure 4.58. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).
Figure 4.59. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).
Figure 4.60. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).
Figure 4.61. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 48 Out-of-Core Detectors, No Singular Value Zeroed).

The error curves are very similar to those obtained with 80 in-core detectors (Figures 4.31 through 4.35). In particular, error peaks occur again between $t = 50$ s and $t = 60$ s for Node # 649, and between $t = 0$ s and $t = 10$ s for Node # 244.

The magnitude of the errors in rodded nodes # 649 and 244 is remarkably small considering the remote locations of the flux sensors.

Table 4.7a compares the results obtained by the four-mode discontinuous synthesis method using (i) 80 in-core detectors, (ii) 4 in-core detectors, and (iii) 48 out-of-core detectors.
Table 4.7a. Maximum Errors in Group-Fluxes Reconstructed by a Four-Mode Time-Discontinuous Synthesis.

<table>
<thead>
<tr>
<th></th>
<th>80 In-Core Detectors</th>
<th>4 In-Core Detectors</th>
<th>48 Out-of-Core Detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Error in Group-One Flux</td>
<td>- 2.8 % (#1398, t=54s)</td>
<td>- 8.4 % (#1398, t=54s)</td>
<td>- 2.9 % (#1060, t=15s)</td>
</tr>
<tr>
<td>Maximum Error in Group-Two Flux</td>
<td>- 5.5 % (#325, t=10s)</td>
<td>- 17.6 % (#1398, t=53s)</td>
<td>- 6.0 % (#1058, t=100s)</td>
</tr>
</tbody>
</table>

Tables 4.7b and 4.7c are the analogs of Table 4.7a for the three-mode and two-mode syntheses respectively.

In analyzing these tables, one should not lose sight of the fact that the errors displayed are maximum values over all nodes and all time-steps. These numbers should not overshadow an important result: In all tests, very small errors were observed in the vast majority of the nodes at any instant.

Table 4.7b. Maximum Errors in Group-Fluxes Reconstructed by a Three-Mode Time-Discontinuous Synthesis.

<table>
<thead>
<tr>
<th></th>
<th>80 In-Core Detectors</th>
<th>4 In-Core Detectors</th>
<th>48 Out-of-Core Detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Error in Group-One Flux</td>
<td>- 4.7 % (#1094, t=100s)</td>
<td>- 9.8 % (#1398, t=54s)</td>
<td>- 5.8 % (#244, t=0s)</td>
</tr>
<tr>
<td>Maximum Error in Group-Two Flux</td>
<td>- 10.8 % (#325, t=10s)</td>
<td>- 20.4 % (#1398, t=53s)</td>
<td>- 13.4 % (#244, t=0s)</td>
</tr>
</tbody>
</table>
Table 4.7c. Maximum Errors in Group-Fluxes Reconstructed by a Two-Mode Time-Discontinuous Synthesis.

<table>
<thead>
<tr>
<th></th>
<th>80 In-Core Detectors</th>
<th>4 In-Core Detectors</th>
<th>48 Out-of-Core Detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximum Error in</strong></td>
<td>- 58.4 % (#244, t=1s)</td>
<td>- 44.7 % (#979, t=6s)</td>
<td>41.1 % (#487, t=30s)</td>
</tr>
<tr>
<td><strong>Group-One Flux</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Maximum Error in</strong></td>
<td>- 128. % (#244, t=1s)</td>
<td>- 98.5 % (#244, t=0s)</td>
<td>87.4 % (#487, t=30s)</td>
</tr>
<tr>
<td><strong>Group-Two Flux</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the three-mode synthesis, although not as accurate as those of the full, four-mode synthesis, are reasonably good. A two-mode synthesis, on the other hand, leads to unacceptably large errors.

From the above tables and from the results of the tests, one concludes that the time-discontinuous, fitted-synthesis method is capable of reconstructing node-averaged group fluxes with great accuracy, even with a reduced in-core instrumentation or with out-of-core detectors only. Best results are obtained when a large number of in-core flux-sensors are used, but the number of detectors and their positions are not as crucial parameters as the number and the mode of selection of the expansion functions, at least when measurement noise is negligible.

4.8. Summary.

In this chapter, numerical simulations were conducted to evaluate the theoretical accuracy of the fitted-synthesis method described in Chapter 2 and 3. A PWR model was used in these simulations.
The tests consisted in reconstructing instantaneous neutron-flux distributions during a severe transient involving control-rod motions and thermal-hydraulic feedback effects. A reference solution to this transient was computed by the QUANDRY code. Reference local-flux values were used to form the neutron-detector signals required by the synthesis method.

Calculations involving 80 in-core detectors and a single set of 26 QUANDRY-computed, static expansion-functions demonstrated that straightforward application of the fitted-synthesis method could give good results (errors of only a few percent in local group fluxes), except in rodded nodes. CPU time was on the order of minutes as opposed to hours for the reference solution.

The modal form of the synthesis approximation derived in Chapter 3 was used to analyze these results, in particular to explain the anomalously large errors (~ 20 %) in perturbed nodes. The cause of these errors was traced back to the lumping of all the information provided by the expansion functions into a single matrix, which resulted in "clustered perturbation-modes".

A natural remedy to the problem was to allow the synthesis to be discontinuous in time. In this modified form, only a small subset of the library of expansion functions was retained at any instant (four in this particular case). This subset was made of flux shapes determined for steady-state critical conditions which closely bracket the instantaneous state of the reactor. Unnecessary expansion-functions were excluded from the synthesis. A procedure which uses on-line information from the nuclear and nonnuclear instrumentation of the plant was suggested as a practical means to identify these "nearest" shapes, without the need for any a priori information about the future evolution of the reactor.

The time-discontinuous, fitted-synthesis method was found to yield better results than the time-continuous method, and to be less expensive. For the transient studied, the time-discontinuous method was capable of reproducing reference flux-values to within a few percent in both groups, in all nodes, and at any time-step. The maximum error in reactivity was
less than one millibeta. In addition, there was no singular-value- or mode-clustering phenomenon, and the physical interpretation of the reconstructed fluxes in terms of superimposed modes was easier.

The simulations were repeated with two different sets of detectors: a single string of 4 in-core detectors in the first case, 48 out-of-core detectors in the second case. In both cases, the discontinuous synthesis method performed remarkably well, demonstrating a lower-than-expected sensitivity to the number and locations of the detectors.

All simulations performed thus far assumed ideal experimental conditions, in particular noise-free measurements. The results can be expected to deteriorate when noise is present in the detector readouts, unless there is some means of eliminating or reducing that noise. This important question of the sensitivity of the fitted-synthesis method to measurement noise is investigated in the next chapter.

In Chapter 6, the time-discontinuous synthesis method is applied to two different transients: a “coolant inlet-temperature transient”, and a “startup transient”.
Chapter 5

NOISE CONSIDERATIONS AND FILTERING TECHNIQUES

5.1. Introduction.

A semi-experimental, nodal synthesis method for reconstructing instantaneous neutron-flux distributions in real-time has been developed and tested in the previous chapters. This method consists in fitting a linear combination of \( K \), precomputed, 3-D, static basis-functions to \( J \) local flux-measurements.

Several potential difficulties with this procedure were identified in Chapter 2, and were addressed in Chapters 3 and 4. In particular, the solution of the least-squares-fit problem by singular value decomposition of the coefficient matrix was shown to eliminate any risk of uncontrolled growth in roundoff errors. The SVD orthogonalization technique also made it possible to rewrite the synthesis formula in the convenient form of a modal expansion, a fundamental term representing the main contribution to the reconstructed flux, and harmonics supplying local corrections of decreasing importance.

The important question of the accuracy of the method was addressed in Chapter 4. It was shown through numerical tests that the method is able to reconstruct instantaneous, nodal group-fluxes with maximum errors of only a few percent, even for severe transients. These small systematic errors were achieved by letting the synthesis be discontinuous in time: At any instant, only a small subset of selected basis-functions is used; these few flux-shapes closely bracket the instantaneous state of the reactor. In this discontinuous form of synthesis, the number and positions of the neutron detectors were found to be of lesser importance than the choice of the basis functions.
An additional, independent problem with the method was mentioned in several occasions in the previous chapters, but left unsolved: Noise in the detector readouts may have a detrimental effect on the results. In all the tests performed in Chapter 4, it was assumed that the signals returned by the instrumentation were noise-free, an unlikely situation in practice.

The influence of measurement noise on the time-discontinuous fitted-synthesis method is examined in detail in this chapter. The study will rely on simple simulations and comparisons with reference, noise-free values. The reactor model and transient will be the same as those described and analyzed in Chapter 4. The instrumentation will be varied in number and position. The three sets of detectors (80 in-core detectors, 4 in-core detectors, and 48 out-of-core detectors) of Chapter 4 will be used again.

In practice, noise in the instrumentation can have various origins. The exact cause of measurement noise is of no interest to this investigation. The basic assumptions are that (i) there exists some amount of noise, i.e. fluctuations of random sign in the measurements, and that (ii) this noise is additive. The main objectives are (i) to understand how these fluctuations can affect the results of the synthesis method, (ii) to quantify the effects relative to the magnitude of the noise, and (iii) to find remedies to a possible excessive sensitivity of the method to that noise.

In the following, the basic equations obtained in Chapter 2 and 3 are modified to account explicitly for the presence of noise in the detector signals. The effect of measurement noise on the results of the synthesis are then studied by combining elementary statistical notions and the orthogonality properties of SVD. The theoretical predictions are subsequently verified by numerical tests. The testing methodology is the same as in Chapter 4. Finally, several filtering algorithms aimed at moderating the consequences of noise are presented and tested.

As in Chapter 4, extensive use is made of graphical illustrations throughout this chapter.

5.2.1. Deterministic Analysis.

The impact of measurement noise onto the synthesized flux-vector, \( \hat{\phi}(t) \), can be most easily studied by replacing the measurement vector,

\[
\mathbf{C}(t) = \left[ C^{(1)}(t), C^{(2)}(t), \ldots, C^{(j)}(t) \right]^T,
\]

by \( \mathbf{C}(t) + \delta\mathbf{C}(t) \) where \( \mathbf{C}(t) \) represents the "exact", noise-free signals, while \( \delta\mathbf{C}(t) \) represents noise contributions or "random errors":

\[
\delta\mathbf{C}(t) = \left[ \delta C^{(1)}(t), \delta C^{(2)}(t), \ldots, \delta C^{(j)}(t) \right]^T.
\]

Any systematic error is excluded from the \( \delta C^{(j)} \)'s. As explained in Chapter 2, possible systematic errors such as those resulting from erroneous detector weight-functions, \( \Sigma^{(j)} \), are not considered here.

The output of the j-th detector is \( C^{(j)}(t) + \delta C^{(j)}(t) \). Neither \( C^{(j)}(t) \) nor \( \delta C^{(j)}(t) \) are known individually; only their sum is available.

With this formal separation, the observation equation, Eq. 2.8, becomes:

\[
\Sigma^T \phi(t) = \mathbf{C}(t) + \delta\mathbf{C}(t).
\]

Substitution of the synthesis approximation,

\[
\phi(t) = \hat{\phi}(t) - \delta\phi(t) = \Psi \mathbf{T}(t) - \delta\phi(t),
\]

into Eq. 5.3 yields
\[ \Sigma^T \Psi \mathbf{T}(t) - \Sigma^T \delta \phi(t) = \mathbf{C}(t) + \delta \mathbf{C}(t), \] 

(5.5a)

which is identical to Eq. 2.11a apart from the explicit representation of the noise by a separate term in the right-hand side.

Eq. 5.5a may be rewritten in a form similar to Eq. 2.11b:

\[ \mathbf{A} \mathbf{T}(t) + \mathbf{E}(t) = \mathbf{C}(t) + \delta \mathbf{C}(t), \] 

(5.5b)

with

\[ \mathbf{A} \equiv \Sigma^T \Psi \] 

(2.12)

and

\[ \mathbf{E}(t) \equiv - \Sigma^T \delta \phi(t) \] 

(2.13)

as in Chapter 2.

Even though the J-by-K matrix \( \mathbf{A} \) changes with time in the discontinuous synthesis-method, no time-dependence is explicitly shown in \( \mathbf{A} \). The analysis below applies to a particular, arbitrary time-interval in which \( \mathbf{A} \) is fixed. It is also assumed that \( J \geq K \geq R = \text{rank} \mathbf{A} \).

The new rectangular system of linear algebraic equations to be solved at every time-step is

\[ \mathbf{A} \mathbf{T}(t) = \mathbf{C}(t) + \delta \mathbf{C}(t). \] 

(5.6)

Its minimum-norm, least-squares solution is the K-element vector

\[ \mathbf{T}_{LS}(t) = \mathbf{A}^\dagger [\mathbf{C}(t) + \delta \mathbf{C}(t)]. \] 

(5.7)

where the pseudo-inverse of \( \mathbf{A} \), \( \mathbf{A}^\dagger \), is given by equation 3.44:
\[ A^\dagger = \mathcal{V} \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T. \] (3.44)

Figure 5.1 illustrates this least-squares fit to a noisy measurement-vector.

\[ \begin{array}{c}
\mathbf{R^K} \\
N(\Delta) \\
T_{LS}(t) \\
\mathbf{R(\Delta^T)} \\
0 \\
\end{array} \quad \begin{array}{c}
\Delta \\
N(\Delta^T) \\
E_{LS}(t) \\
C(t) + \delta C(t) \\
R(\Delta) \\
\mathbf{R^J} \end{array} \]

\[ \begin{array}{c}
\mathbf{R^K} \\
0 \\
T_{LS}(t) \\
\mathcal{C}(t) \\
\end{array} \quad \begin{array}{c}
\Delta^\dagger \\
E_{LS}(t) \\
C(t) + \delta C(t) \\
\mathbf{R^J} \\
\end{array} \]

Figure 5.1. Least-Squares Solution of \( A \cdot T(t) = \mathcal{C}(t) + \delta \mathcal{C}(t) \).

The residual vector corresponding to \( T_{LS}(t) \) is

\[
E_{LS}(t) = \mathcal{C}(t) + \delta \mathcal{C}(t) - A \cdot T_{LS}(t)
= N [\mathcal{C}(t) + \delta \mathcal{C}(t)]
\] (5.8)
where
\[
N = \mathbf{I} - \mathbf{A} \mathbf{A}^\dagger = \mathbf{U} \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} \mathbf{U}^T
\]  
(5.9)

is a J-by-J orthogonal-projection matrix. Any vector of \(\mathbf{R}^J\) is projected by \(N\) onto the left null-space of \(\mathbf{A}\), \(\mathbf{N}(\mathbf{A}^T)\).

In the absence of noise (\(\delta\mathbf{C}(t) = 0\)), a different least-squares solution-vector, and a different residual vector would be obtained, as indicated in Figure 5.1.

When noise is present, the residual vector, \(\mathbf{E}_{LS}(t)\), is no longer a measure of systematic errors only. Instead, it is a combined measure of both the systematic errors caused by the synthesis approximation, and random errors caused by noisy measurements. The two types of errors are indistinguishably mixed.

If the magnitude of the noise, \(\|\delta\mathbf{C}(t)\|\), is small with respect to \(\|\mathbf{C}(t)\|\), the vector \(\mathbf{A} \mathbf{T}_{LS}(t)\) in \(\mathbf{R}(\mathbf{A})\) will be near the corresponding vector obtained in the absence of noise. However, because of possible ill-conditioning in \(\mathbf{A}\), the point \(\mathbf{T}_{LS}(t) = \mathbf{A}^\dagger [\mathbf{C}(t) + \delta\mathbf{C}(t)]\) of \(\mathbf{R}^K\) may very well be located far away from its noise-free counterpart. In some cases, the result can be a meaningless reconstructed flux-vector, \(\hat{\phi}(t)\), unless corrective action is taken.

This reason for these difficulties can be easily seen from the modal form of the synthesis approximation:
\[
\hat{\phi}(t) = \Psi^T \mathbf{T}_{LS}(t) = \sum_{i=1}^{R} \psi^{(i)} T_{LS}^{(i)}(t)
\]  
(3.71)

where the vector of modal coefficients,
\[ T'_{LS}(t) = \mathbf{V}^T T_{LS}(t), \]  

(3.72)

is simply \( T_{LS}(t) \) expressed in the \( \{v_i\} \) orthonormal basis of \( \mathbf{R}^K \). The components, \( T'^{(i)}_{LS}(t) \), of this vector are obtained from Eq. 3.83 with \( \mathbf{C}(t) \) replaced by \( \mathbf{C}(t) + \delta\mathbf{C}(t) \):

\[
T'^{(i)}_{LS}(t) = \begin{cases} 
\frac{1}{s_i} \mathbf{u}_i^T \mathbf{C}(t) + \frac{1}{s_i} \mathbf{u}_i^T \delta\mathbf{C}(t), & i = 1,2,\ldots,R, \\
0, & i = R+1,R+2,\ldots,K.
\end{cases}
\]  

(5.10)

Eq. 5.10 shows that, for \( i \leq R \), the new modal coefficient, \( T'^{(i)}_{LS}(t) \), is the sum of a noise-free component, \( \mathbf{u}_i^T \mathbf{C}(t) / s_i \), and a noisy contribution, \( \mathbf{u}_i^T \delta\mathbf{C}(t) / s_i \).

To understand how noise can adversely affect \( \mathbf{\hat{f}}(t) \), assume, for instance, that the noise-free vector \( \mathbf{C}(t) \) is predominantly in the fundamental direction, \( \mathbf{u}_1 \), and has no component in the \( \mathbf{u}_i \) direction, i.e. \( \mathbf{u}_i^T \mathbf{C}(t) = 0 \). In general, the noise projection along \( \mathbf{u}_i \), \( \mathbf{u}_i^T \delta\mathbf{C}(t) \), will not be zero. If the singular value \( s_i \) is a small number with respect to \( s_1 \), then \( T'^{(i)}_{LS}(t) \), which is zero when no noise is present, may instead be very large in magnitude compared with the other modal coefficients.

Whether this relatively large coefficient will have an impact on the reconstructed flux-vector, \( \mathbf{\hat{f}}(t) \), of Eq. 3.71 depends on the corresponding mode \( \psi'^{(i)} \). On the average, this mode has small magnitude because, as shown in Chapter 3:

\[ \| \Sigma^T \psi'^{(i)} \| = s_i, \quad i = 1,2,\ldots,R. \]  

(3.81)

However, Eq. 3.81 does not prevent \textit{some component} \( \psi'^{(i)} \) from being relatively large in magnitude with respect to the other modes in some energy-group \( g \) and node \( n \). This may happen when the following two
conditions are simultaneously met: (i) $\psi^{(i)}_{\text{sn}}$ is a high-order perturbation-mode reflecting local perturbations incorporated in the original set of $K$ expansion-functions $\psi^{(k)}$, and (ii) no detector is located near those regions where $|\psi^{(i)}_{\text{sn}}|$ is large. In such circumstances, large errors may occur in $\hat{\phi}_{\text{sn}}(t)$ because $\psi^{(i)}_{\text{sn}}$ is multiplied by a large, physically-meaningless coefficient, and the term $\psi^{(i)}_{\text{sn}} T^{(i)}_{\text{LS}}(t)$ may dominate all the other contributions in the synthesis (Eq. 3.71).

From these simple considerations, one concludes that the sensitivity of the synthesis to noise depends directly on the conditioning of $A$, i.e. on its condition number $\kappa(A) = s_1 / s_R$. For an ill-conditioned matrix, $\kappa(A)$ is much larger than unity, and the sensitivity to noise may be excessive. This sensitivity can be reduced by the same procedure used in Chapters 3 and 4 to neutralize roundoff errors, i.e. by eliminating the smallest modes from the synthesis. The difference here is that these modes contain some small but valuable physical information, and zeroing them implies accepting a slight increase in systematic errors.

5.2.2. Statistical Analysis.

The value of the deterministic analysis of the previous section is mainly qualitative because the noise vector, $\delta C(t)$, formally introduced in Eq. 5.3 is unknown. Because of this noise, the values of $T_{\text{LS}}(t)$, $T'_{\text{LS}}(t)$, $E_{\text{LS}}(t)$, $\hat{\phi}(t)$, and other reconstructed quantities cannot be predicted exactly.

A more useful, quantitative characterization of the results can be obtained from a simple statistical analysis. Indeed, if measurement noise can be characterized statistically, predictions can be made for the expected values of $T_{\text{LS}}(t)$, $T'_{\text{LS}}(t)$, $E_{\text{LS}}(t)$, $\hat{\phi}(t)$, etc..., and for the corresponding expected deviations about these mean values, i.e. variances.

To facilitate the statistical analysis, it is assumed that the noise-vector, \( \delta \mathbf{C} \), (i) is \textit{stationary}, (ii) has zero expected-value

\[
\mathbb{E} \{ \delta \mathbf{C} \} = 0 ,
\]  \hspace{1cm} (5.11)

and (iii) is characterized by a known, \( J \)-by-\( J \) covariance matrix

\[
\text{cov} \{ \delta \mathbf{C} \} = \mathbb{E} \left \{ \left [ \delta \mathbf{C} - \mathbb{E} \{ \delta \mathbf{C} \} \right ] \left [ \delta \mathbf{C} - \mathbb{E} \{ \delta \mathbf{C} \} \right ]^T \right \} = \mathbb{E} \left \{ \delta \mathbf{C} \delta \mathbf{C}^T \right \} . \hspace{1cm} (5.12)
\]

It is reasonable to restrict this covariance matrix to a diagonal form, i.e. to assume that the various detector-signals are \textit{uncorrelated}:

\[
\mathbb{E} \left \{ \delta \mathbf{C}^{(j)} \delta \mathbf{C}^{(m)} \right \} = 0 , \quad j \neq m , \quad \begin{cases} j = 1,2,\ldots,J, \\ m = 1,2,\ldots,J. \end{cases} \hspace{1cm} (5.13)
\]

For consistency with the requirement of a properly-weighted least-squares system (Chapter 3), Eq. 5.3 is left-multiplied by the \( J \)-by-\( J \) diagonal matrix \( \text{diag} \{ 1 / \sigma^{(j)^2} \} \) where the \( \sigma^{(j)^2} \)'s are the variances of the \( \delta \mathbf{C}^{(j)} \)'s,

\[
\sigma^{(j)^2} = \mathbb{E} \left \{ \delta \mathbf{C}^{(j)^2} \right \} , \quad j = 1,2,\ldots,J, \hspace{1cm} (5.14)
\]

or estimates of the latter. The result is a "balanced" observation equation with scaled detector weight-functions, \( \Sigma^{(j)} \), and a measurement vector, \( \mathbf{C}(t) + \delta \mathbf{C} \), in which all components have equal variance. This common variance is denoted \( \sigma^2 \).

In these conditions, the covariance matrix of the \( \delta \mathbf{C}^{(j)} \)'s becomes

\[
\text{cov} \{ \delta \mathbf{C} \} \equiv \sigma^2 \mathbf{I}. \hspace{1cm} (5.15)
\]
In the following, no other assumption needs to be made about the distribution of the $\delta C^{(i)}$'s, i.e. no statistical model (probability density function) needs to be postulated. In other words, any distribution of errors characterized by the above first two moments is acceptable as “noise”.

b. Expected Values.

Expected values for $T_{LS}(t)$, $T'_{LS}(t)$, $E_{LS}(t)$, and $\hat{\phi}(t)$, are readily inferred from the linear relationships between this various quantities and the noisy measurement-vector $C(t) + \delta C$.

The expected value of the least-squares solution vector (Eq. 5.7) is

$$
\mathbb{E} \{ T_{LS}(t) \} = \mathbb{E} \{ \Lambda^\dagger [ C(t) + \delta C ] \} = \Lambda^\dagger \mathbb{E} \{ C(t) + \delta C \}
$$

$$
= \Lambda^\dagger [ \mathbb{E} \{ C(t) \} + \mathbb{E} \{ \delta C \} ]
$$

$$
= \Lambda^\dagger [ C(t) + 0 ] = \Lambda^\dagger C(t), \quad (5.16)
$$

i.e. the noise-free value. $T_{LS}(t) = \Lambda^\dagger [ C(t) + \delta C ]$ is therefore an unbiased estimate of the unknown, noise-free, least-squares solution-vector.

Similarly, from Eqs. 3.72, 5.10, 5.8, and 3.71:

$$
\mathbb{E} \{ T'_{LS}(t) \} = \mathbb{E} \{ V^T T_{LS}(t) \} = V^T \Lambda^\dagger C(t)
$$

$$
= \begin{bmatrix}
S^{-1} & 0 \\
0 & 0
\end{bmatrix}
U^T C(t), \quad (5.17)
$$
\[ E \{ T_{LS}^{(i)}(t) \} = \begin{cases} \frac{1}{s_i} u_i^T C(t), & i = 1, 2, \ldots, R, \\ 0, & i = R+1, R+2, \ldots, K, \end{cases} \]  
(5.18)

\[ E \{ E_{LS}(t) \} = N C(t), \]  
(5.19)

\[ E \{ E_{LS}^{(j)}(t) \} = \begin{cases} 0, & j = 1, 2, \ldots, R, \\ u_j^T C(t), & j = R+1, R+2, \ldots, J, \end{cases} \]  
(5.20)

\[ E \{ \hat{\phi}(t) \} = \Psi' E \{ T_{LS}'(t) \} = \sum_{i=1}^{R} \psi^{(i)} E \{ T_{LS}^{(i)}(t) \} \]
\[ = \sum_{i=1}^{R} \psi^{(i)} \frac{1}{s_i} [u_i^T C(t)], \]  
(5.21)

The right-hand sides of these equations are the values obtained when no noise is present. In fact, the above relationships are identical to formulas obtained in Chapter 3, except that \( T_{LS}(t), T_{LS}'(t), T_{LS}^{(i)}(t), E_{LS}(t), E_{LS}^{(j)}(t), \) and \( \hat{\phi}(t) \) have been replaced by their expected values.

Note that, even though the vector \( E \{ \hat{\phi}(t) \} \) is an unbiased estimate of the unknown, noise-free, reconstructed flux-vector of the synthesis, it remains a biased estimate of the unknown, physical, flux-vector, \( \phi(t) \), because of systematic errors caused by the synthesis approximation. The above analysis has shown that, if the noise vector satisfies \( E \{ \delta C \} = \Omega \), measurement noise will not add any additional bias to \( \hat{\phi}(t) \).
c. Covariance Matrices.

The covariance matrix of the mixing coefficients, $\mathbf{T}_L^k(t)$, is the $K$-by-$K$ matrix

$$\text{cov} \{ \mathbf{T}_L(t) \} = \mathbb{E} \left\{ \left[ \mathbf{T}_L(t) - \mathbb{E}\{ \mathbf{T}_L(t) \} \right] \left[ \mathbf{T}_L(t) - \mathbb{E}\{ \mathbf{T}_L(t) \} \right]^T \right\}. \quad (5.22)$$

Substitution of Eqs. 5.7 and 5.17 into Eq. 5.22 yields

$$\text{cov} \{ \mathbf{T}_L(t) \} = \mathbf{A}^+ \text{cov} \{ \mathbf{S} \} \mathbf{A}^+^T. \quad (5.23)$$

Eqs. 5.15 and 3.44 may be used to rewrite Eq. 5.23 in the following form:

$$\text{cov} \{ \mathbf{T}_L(t) \} = \sigma^2 \mathbf{A}^+ \mathbf{A}^+^T$$

$$= \sigma^2 \mathbf{V} \begin{bmatrix} (\mathbf{S}^{-1})^2 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}^T = \sigma^2 \sum_{i=1}^{R} \frac{1}{s_i^2} \mathbf{v}_i \mathbf{v}_i^T. \quad (5.24)$$

or

$$\mathbf{V} \text{cov} \{ \mathbf{T}_L(t) \} \mathbf{V}^T = \sigma^2 \begin{bmatrix} (\mathbf{S}^{-1})^2 & 0 \\ 0 & 0 \end{bmatrix} \quad (5.25)$$

Eq. 5.25 proves that the covariance matrix of the $\mathbf{T}_L^k(t)$'s is diagonal\(^2\) when expressed in the orthonormal basis $\{ \mathbf{v}_i \}$ of $\mathbb{R}^K$.

\(^2\) It can be shown that the least-squares vector, $\mathbf{T}_L(t)$, of the properly-weighted system, $\mathbf{A} \mathbf{T}(t) = \mathbf{C}(t) + \mathbf{S} \mathbf{C}$, is the "best", linear, unbiased estimate in the sense that it minimizes the variances of the $\mathbf{T}_L^k(t)$'s [M8]. Any other linear unbiased estimate would have larger
It follows immediately from Eq. 5.25 and from the definition
\( T_{LS}^{(t)}(t) = V^T T_{LS}(t) \) that the covariance matrix of the modal coefficients,
\( T_{LS}^{(t)}(t) \), is simply

\[
\text{cov} \{ T_{LS}^{(t)}(t) \} = V \text{cov} \{ T_{LS}(t) \} V^T = \sigma^2 \begin{bmatrix}
(S^{-1})^2 & 0 \\
0 & 0
\end{bmatrix},
\tag{5.26}
\]
i.e. is a diagonal matrix.

This result could have been expected from Eq. 5.18, and demonstrates one more time that it is more convenient to work with the modal coefficients, \( T_{LS}^{(0)}(t) \), than with the mixing coefficients, \( T_{LS}^{(k)}(t) \).

Eq. 5.26 proves that the \( T_{LS}^{(0)}(t) \)’s are uncorrelated. This is a direct consequence of (i) the assumption of uncorrelated \( \delta C^{(j)} \)’s, and (ii) the orthogonalization process by SVD.

The diagonal elements of \( \text{cov} \{ T_{LS}^{(t)} \} \) are the variances of the \( T_{LS}^{(0)}(t) \)’s:

\[
\text{var} \{ T_{LS}^{(0)}(t) \} = \begin{cases}
\frac{1}{s_i^2} \sigma^2, & i = 1,2,\ldots,R, \\
0, & i = R+1,R+2,\ldots,K.
\end{cases}
\tag{5.27}
\]

Eq. 5.27 is an interesting result which shows that, for \( i \leq R \), the dispersion in the values of \( T_{LS}^{(0)}(t) \) about \( E \{ T_{LS}^{(0)}(t) \} \) is constant (for a given matrix \( A \) and for stationary noise), and is governed by the reciprocal of the corresponding singular value \( s_i \). A large dispersion can therefore be

variances. It can also be shown that, if the random errors, \( \delta C^{(j)} \), have a Gaussian distribution, then \( T_{LS}(t) \) is also the maximum-likelihood estimator.
expected for a small $s_i$. This result is consistent with the qualitative
description of noise sensitivity given at the end of Section 5.2.1.

On the other hand, there is no dispersion for $i > R$ since $T'_{LS}(t)$ has
no component in the directions $v_{R+1}, v_{R+2}, \ldots, v_K$ of $N(A)$.

Application of the identity
\[
\text{var} \{X\} = \mathbb{E} \{X^2\} - [\mathbb{E} \{X\}]^2,
\]
(5.28)
to $X = T'_{LS}(t)$ in Eqs. 5.18 and 5.27 yields
\[
\mathbb{E} \left\{ \left[ T'_{LS}(t) \right]^2 \right\} = \begin{cases} 
\frac{1}{s_i^2} \left[ u_i^T C(t) \right]^2 + \frac{1}{s_i^2} \sigma^2, & i = 1,2,\ldots,R, \\
0, & i = R+1,R+2,\ldots,K.
\end{cases}
\]
(5.29)

For $i \leq R$, the right-hand side of Eq. 5.29 is the sum of two terms: a
contribution from the noise-free vector, $\left[ u_i^T C(t) \right]^2 / s_i^2$, and a contribution
from noise, $\sigma^2 / s_i^2$. The term $\left[ u_i^T C(t) \right]^2 / s_i^2$ is the i-th component of the
least-squares solution-vector of the noise-free problem. The term, $\sigma^2 / s_i^2$,
on the other hand, is a bias caused by the noise. This bias does not lead to
any particular difficulty. It should simply be expected in plots of the $T'_{LS}(t)$'s.

A formula for the expected value of the square of the Euclidean
norm of $T'_{LS}(t)$ can be derived by summing both sides of Eq. 5.29 from
$i = 1$ to $i = K$. The result is:
\[
\mathbb{E} \left\{ \| T'_{LS}(t) \|^2 \right\} = \sum_{i=1}^{R} \frac{1}{s_i^2} \left[ u_i^T C(t) \right]^2 + \sigma^2 \sum_{i=1}^{R} \frac{1}{s_i^2}.
\]
(5.30)

The various terms simply add in root-mean-square fashion because $\{v_i\}$ is
an orthonormal basis.
The J-by-J covariance matrix of the $E_{LS}(t)$'s is

$$\text{cov} \{E_{LS}(t)\} = N \text{cov} \{\delta C\} N^T = \sigma^2 N N^T$$

$$= \sigma^2 U \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} U^T = \sigma^2 \sum_{i=R+1}^{J} u_i u_i^T,$$  \hspace{1cm} (5.31)

or, in diagonal form,

$$\text{cov} \{U^T E_{LS}(t)\} = U^T \text{cov} \{E_{LS}(t)\} U = \sigma^2 \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}.$$  \hspace{1cm} (5.32)

Eq. 5.32 shows that, in the same way $\text{cov} \{T_{LS}(t)\}$ was diagonal in the \{v_i\} basis of $R^K$, $\text{cov} \{E_{LS}(t)\}$ is diagonal in the \{u_i\} basis of $R^J$.

The variances of the orthogonal projections of $E_{LS}(t)$ on the $u_i$ directions are

$$\text{var} \{u_i^T E_{LS}(t)\} = \begin{cases} 0, & i = 1,2,\ldots,R, \\ \sigma^2, & i = R+1,R+2,\ldots,J. \end{cases}$$  \hspace{1cm} (5.33)

This is another expected result. There is no dispersion in the first $R$ projections $u_i^T E_{LS}(t)$ since the residual vector of the least-squares fit is, by definition, always perpendicular to $u_1, u_2, \ldots, u_R$ (i.e. to $R(A)$), whether there is noise in the measurements or not.

In the last $J-R$ $u_i$-directions, on the other hand, the components of $E_{LS}(t)$ are identical to, and therefore as noisy as, the null-space components of $C(t) + \delta C$, $C_N(t) + \delta C_N$.

The analog of Eq. 5.29 can be obtained from Eq. 5.33 and from the identity 5.28 applied to $X = u_i^T E_{LS}(t)$:
\[
E \left\{ [u_i^T E_{LS}(t)]^2 \right\} = \begin{cases} 
0, & i = 1, 2, \ldots, R, \\
[u_i^T C(t)]^2 + \sigma^2, & i = R+1, R+2, \ldots, J.
\end{cases} \tag{5.34}
\]

Addition of the \(J\) terms on both sides gives a formula similar to Eq. 5.30:

\[
E \left\{ \|E_{LS}(t)\|^2 \right\} = \sum_{i=R+1}^{J} [u_i^T C(t)]^2 + (J - R) \sigma^2 \\
= \|C_N(t)\|^2 + (J - R) \sigma^2. \tag{5.35a}
\]

Eq. 5.35a shows that the square Euclidean norm of the residual vector is no longer a measure of the systematic errors caused by the synthesis approximation. The expected value of \(\|E_{LS}(t)\|^2\) is now the sum of (i) a systematic-error term, \(\|C_N(t)\|^2\), equal to the value of \(\|E_{LS}(t)\|^2\) when no noise is present, and (ii) a random-error term, \((J-R) \sigma^2\). As with \(\|T'_{LS}(t)\|^2\), this random-error term causes a bias in \(\|E_{LS}(t)\|^2\).

When systematic errors are nil or negligible, Eq. 5.35a reduces to a well-known formula:

\[
E \left\{ \|E_{LS}(t)\|^2 \right\} = (J - R) \sigma^2, \tag{5.35b}
\]

which is sometimes used in statistical analyses to estimate \(\sigma^2\) (when \(J \gg R\)).

Note that no singular value appears in the formulas involving \(E_{LS}(t)\). This is natural because \(E_{LS}(t)\) does not depend on the \(A\) mapping itself, but only on the measurement vector, \(C(t) + \delta C\), and on the \(\{u_i\}\) basis of \(R^J\).

Eqs. 5.30 and 5.35a suggest that, from a statistical standpoint, it is the square of \(\|T'_{LS}(t)\|^2\), rather than \(\|T'_{LS}(t)\|^2\) itself, and the square of \(\|E_{LS}(t)\|^2\), rather than \(\|E_{LS}(t)\|^2\) itself, which are the most convenient variables to use.
The last covariance matrix of interest is the GN-by-GN covariance matrix of the reconstructed, nodal group-fluxes, $\hat{\phi}_{gn}(t)$:

$$\text{cov} \left\{ \hat{\phi}(t) \right\} = \Psi \text{cov} \left\{ T_{LS}(t) \right\} \Psi^T = \Psi' \text{cov} \left\{ T'_{LS}(t) \right\} \Psi'^T. \quad (5.36)$$

Substitution of Eq. 5.26 into Eq. 5.36 yields

$$\text{cov} \left\{ \hat{\phi}(t) \right\} = \sigma^2 \Psi' \begin{bmatrix} (S^{-1})^2 & 0 \\ 0 & 0 \end{bmatrix} \Psi'^T, \quad (5.37)$$

This matrix is not diagonal in general. This is no surprise since the GN-element vector, $\hat{\phi}(t)$, is inferred from only $J$ measurements ($J << GN$). Strong correlations must therefore exist between the various $\hat{\phi}_{gn}(t)$’s.

It can be verified, however, that the smaller, J-by-J covariance matrix of the $\hat{C}^{(i)}(t)$’s, the elements of the vector

$$\hat{C}(t) = \Sigma^T \hat{\phi}(t) = A T_{LS}(t) = C(t) + \delta C(t) - E_{LS}(t), \quad (5.38)$$

the “image” of $\hat{\phi}(t)$ through the instrumentation, is diagonal when expressed in the orthonormal basis $\{u_i\}$ of $R^J$:

$$\text{cov} \left\{ \hat{C}(t) \right\} = \text{cov} \left\{ \Sigma^T \hat{\phi}(t) \right\} = \Sigma^T \text{cov} \left\{ \hat{\phi}(t) \right\} \Sigma. \quad (5.39)$$

The vector $\hat{C}(t)$ is shown in Figure 5.1. It is the orthogonal projection of $C(t) + \delta C$ onto $R(A)$.

Using Eqs. 5.36, 5.23, 3.21, and 3.44, Eq. 5.39 may rewritten as

$$\text{cov} \left\{ \hat{C}(t) \right\} = A \text{cov} \left\{ T_{LS}(t) \right\} A^T$$
\[
\begin{align*}
&= \mathbf{A} \mathbf{A}^\dagger \text{cov} \{\delta \mathbf{C}(t)\} \left[\mathbf{A} \mathbf{A}^\dagger\right]^T \\
&= \mathbf{U} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T \text{cov} \{\delta \mathbf{C}(t)\} \mathbf{U} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T, \tag{5.40}
\end{align*}
\]

or, equivalently,
\[
\begin{align*}
\mathbf{U}^T \text{cov} \{\tilde{\mathbf{C}}(t)\} \mathbf{U} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T \text{cov} \{\delta \mathbf{C}(t)\} \mathbf{U} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{5.41}
\end{align*}
\]

In general, \(\text{cov} \{\delta \mathbf{C}(t)\}\) is not diagonal in the \(\{u_i\}\) basis. The effect of the left and right multiplications by \(\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\) in Eq. 5.41 is to extract the first \(R\) diagonal elements from \(\mathbf{U}^T \text{cov} \{\delta \mathbf{C}(t)\} \mathbf{U}\), thus making \(\text{cov} \{\tilde{\mathbf{C}}(t)\}\) diagonal in the \(\{u_i\}\) basis.

In this particular application, \(\text{cov} \{\delta \mathbf{C}(t)\}\) is just \(\sigma^2 \mathbf{I}\). Therefore, Eq. 5.40 reduces to
\[
\text{cov} \{\tilde{\mathbf{C}}(t)\} = \sigma^2 \mathbf{U} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T = \sigma^2 \sum_{i=1}^{R} u_i u_i^T. \tag{5.42}
\]

Combination of Eqs. 5.15, 5.31, and 5.42 gives the unifying formula:
\[
\text{cov} \{\tilde{\mathbf{C}}(t)\} = \text{cov} \{\delta \mathbf{C}\} - \text{cov} \{\mathbf{E}_{LS}(t)\}. \tag{5.43}
\]

All three matrices are diagonal in the \(\{u_i\}\) basis.

Eq. 5.43 shows that, independently of the basis chosen to express the matrices, the variances of the \(\tilde{\mathbf{C}}^{(j)}(t)\)'s are smaller than the variances of the
\( \delta C^{(j)} \)'s by an amount equal to the variances of the \( E_{LS}^{(j)} \)'s. The reason for this is that the variances of the elements of \( \delta C_N \) (which lies in the space \( N(A^T) \) defined by \( u_{R+1}, u_{R+2}, ..., u_J \)) do not affect \( \tilde{C} \), consistently with the fact that \( C_N(t) + \delta C_N \) (the component of \( C(t) + \delta C \) in \( N(A^T) \)) does not affect the solution of the least-squares fit.

d. The 'R-over-J Rule'.

The formula obtained for \( \text{cov} \{ \tilde{C}(t) \} \) (Eq. 5.42) makes it possible to infer an approximate yet very useful expression for the variances in the reconstructed group-two fluxes. This expression is derived below by two different methods: The first one makes use of mathematical arguments only; the second one incorporates physical arguments.

From Eq. 5.43, the variances in the \( \tilde{C}^{(j)}(t) \)'s are

\[
\text{var} \left\{ \tilde{C}^{(j)}(t) \right\} = \sigma^2 \sum_{i=1}^{R} u_{i}^{(j)} \cdot u_{i}^{(j)}, \quad j = 1, 2, ..., J. \tag{5.44}
\]

Since the \( u_i \) vectors are orthonormal vectors of \( R^J \), they satisfy:

\[
\sum_{i=1}^{J} \left[ u_{i}^{(j)} \right]^2 = 1, \quad j = 1, 2, ..., J. \tag{5.45}
\]

Therefore, Eq. 5.44 may be rewritten as

\[
\text{var} \left\{ \tilde{C}^{(j)}(t) \right\} = \sigma^2 \frac{\sum_{i=1}^{R} \left[ u_{i}^{(j)} \right]^2}{\sum_{i=1}^{J} \left[ u_{i}^{(j)} \right]^2}. \tag{5.46}
\]
The ratio in the right-hand side of Eq. 5.46 is always a number in the interval [0, 1], confirming that the variances in the $\hat{C}^{(j)}(t)$'s are indeed smaller than (or, at worst, equal to) the variances in the $[C^{(j)}(t) + \delta C^{(j)}]$'s. Since there is no a priori reason for the j-th components of the $\mu_i$ vectors to be completely different from one another, an approximate value for this ratio is $R / J$, giving for Eq. 5.46:

$$\text{var} \left\{ \hat{C}^{(j)}(t) \right\} = \sigma^2 \frac{R}{J}, \quad j = 1,2,\ldots,J.$$  \hspace{1cm} (5.47)

Note from Eq. 5.46 that the above formula is exact on the average, i.e.:

$$\frac{1}{J} \sum_{j=1}^{J} \text{var} \left\{ \hat{C}^{(j)}(t) \right\} = \sigma^2 \frac{R}{J}.$$  \hspace{1cm} (5.48)

Eq. 5.48 is the rigorous mathematical translation of a geometrically intuitive result (Figure 5.1): Projecting the J-element vector $C(t) + \delta C$ onto the R-dimensional "plane" $R(\hat{A})$ has the effect of reducing the variance by an average factor of $R / J$.

If the detector weight-functions are given by Eq. 4.3, i.e. by

$$\Sigma_{g\ell}^{(j)} = \delta_{g2} \delta_{\ell n_j} \quad \text{where} \quad n_j \quad \text{is the number of the node containing the j-th detector,}$$

then $\hat{C}^{(j)}(t) = \hat{\phi}_{2n_j}(t), \ j = 1,2,\ldots,J$. In these conditions, Eqs. 5.47 and 5.48 become

$$\text{var} \left\{ \hat{\phi}_{2n_j}(t) \right\} = \sigma^2 \frac{R}{J}, \quad j = 1,2,\ldots,J,$$  \hspace{1cm} (5.49)

and

$$\frac{1}{J} \sum_{j=1}^{J} \text{var} \left\{ \hat{\phi}_{2n_j}(t) \right\} = \sigma^2 \frac{R}{J}.$$  \hspace{1cm} (5.50)
respectively. Simple extension of Eq. 5.49 to all the nodes \( n \) gives

\[
\text{var} \{ \hat{\phi}_{2n}(t) \} = \sigma^2 \sum_{j} R_j, \quad n = 1,2,...,N. \tag{5.51}
\]

This approximate relationship can also be obtained by the following method in which physical arguments are used to justify the necessary approximations.

Although \( \text{cov} \{ \hat{\phi}(t) \} \) in Eq. 5.37 is not a diagonal matrix, its diagonal entries still give relationships for the variances of the \( \hat{\phi}_{gn}(t) \)'s:

\[
\text{var} \{ \hat{\phi}_{gn}(t) \} = \sigma^2 \sum_{i=1}^{R} \frac{1}{s_i^2} [\psi_{gn}^{(i)}]^2, \quad \left\{ \begin{array}{l}
g = 1,2,...,G, \\
n = 1,2,...,N. \end{array} \right. \tag{5.52}
\]

Two preliminary comments can be made about Eq. 5.52:

(i) The variance of \( \hat{\phi}_{gn}(t) \), i.e. the dispersion of the reconstructed flux, \( \hat{\phi}_{gn}(t) \), in group \( g \) and node \( n \), about its expected value is constant for a given matrix \( \Delta \) and for stationary noise.

(ii) The variance of \( \hat{\phi}_{gn}(t) \) is the sum of \( R \) contributions,

\[\sigma^2 [\psi_{gn}^{(i)}]^2 / s_i^2, \quad \text{one for each of the modes} \ (s_i, \psi_i, \mu_i, \psi^{(i)}). \]

A small singular value \( s_i \) may give a large contribution unless the corresponding mode \( \psi^{(i)} \) is also very small in magnitude in energy-group \( g \) and node \( n \). Eq. 5.52 expresses quantitatively a result obtained qualitatively at the end of Section 5.2.1.

Recall from Eq. 3.81 that \( s_i \) is a measure of the magnitude of \( \psi^{(i)} \):

\[
s_i^2 = \| \sum_{i} \psi^{(i)} \|^2 = \psi^{(i)T} \sum \sum \psi^{(i)}, \quad i = 1,2,...,R, \tag{5.53a}
\]

or, in expanded form,
\[ s_i^2 = \sum_{j=1}^{J} \left[ \Sigma_{g}^{(j)} \psi_{n}^{(i)} \right]^2, \quad i = 1, 2, \ldots, R. \]  (5.53b)

With \( \Sigma_{g}^{(j)} = \delta_{g2} \delta_{nnj} \) as in Eq. 4.3, Eq. 5.53b becomes

\[ s_i^2 = \sum_{j=1}^{J} \left[ \psi_{2n}^{(i)} \right]^2, \quad i = 1, 2, \ldots, R, \]  (5.54)

which may also be written as

\[ s_i^2 = J \left[ \psi_{2n}^{(i)} \right]^2, \quad i = 1, 2, \ldots, R, \]  (5.55)

where

\[ \left[ \psi_{2n}^{(i)} \right]^2 = \frac{1}{J} \sum_{j=1}^{J} \left[ \psi_{2nj}^{(i)} \right]^2, \quad i = 1, 2, \ldots, R \]  (5.56)

is the “average” value of the square group-two mode \( \psi_{2n}^{(i)} \). The average is over the \( J \) instrumented nodes \( n_j \). Admittedly, such an average has no rigorous physical meaning; however, \( \bar{n} \) in Eq. 5.56 may be thought of as a fictitious “average instrumented-node” representative of all the nodes \( n_j \).

For fundamental \( (i = 1) \) and low-order modes, \( \left[ \psi_{2n}^{(i)} \right]^2 \) does not depend appreciably on the detector positions because these modes are “felt” everywhere throughout the reactor core. For high-order perturbation-modes, on the other hand, this average may be relatively small if the detectors are all located far away from the perturbed regions.

Substitution of Eq. 5.55 into Eq. 5.52 yields
\[
\text{var} \{ \hat{\phi}_{g_n}(t) \} = \sigma^2 \frac{R}{J} \sum_{i=1}^{R} \frac{\left[ \psi_{g_n}^{(i)} \psi_{g_n}^{(i)} \right]^2}{\left[ \psi_{2_n}^{(i)} \psi_{2_n}^{(i)} \right]^2}, \quad \begin{cases} g = 1,2,\ldots,G, \\ n = 1,2,\ldots,N. \end{cases} \tag{5.57}
\]

Each term in the right-hand side of Eq. 5.57 depends on the node number, \( n \), and energy group, \( g \). For \( g = 2 \), unless the instrumentation is concentrated in some corner of the reactor, and, at the same time, node \( n \) is far away from that corner, the ratio \( \left[ \psi_{g_n}^{(i)} \right]^2 / \left[ \psi_{2_n}^{(i)} \right]^2 \) should be near unity. A possible exception is with high-order modes for which this ratio can be significantly larger than one for some nodes \( n \). In all other conditions, the sum of the \( R \) ratios in Eq. 5.57 should be approximately equal to \( R \), giving the desired formula:

\[
\text{var} \{ \hat{\phi}_{2_n}(t) \} \approx \sigma^2 \frac{R}{J}, \quad n = 1,2,\ldots,N. \tag{5.51}
\]

Eq. 5.51 shows that the variance in the reconstructed group-two flux in any node \( n \) (with the possible exception of some perturbed nodes \( n \)) is approximately equal to the noise variance, \( \sigma^2 \), multiplied by the ratio \( R / J \). This ratio, or variance-reduction factor, is a good estimate of the inherent noise-reduction capability of the fitted-synthesis method. Note that, as with Eqs. 5.48 and 5.50, Eq. 5.51 is exact when averaged over all instrumented nodes:

\[
\frac{1}{J} \sum_{j=1}^{J} \text{var} \{ \hat{\phi}_{2_n}(t) \} = \sigma^2 \frac{R}{J}. \tag{5.58}
\]

Eq. 5.51 makes sense in that, for a fixed value of \( R \):

(i) If \( J \) increases (more detectors), the dispersion in the \( \hat{\phi}_{2_n}(t) \)'s is reduced;

(ii) If \( J = R \), there is no noise reduction.

Eqs. 5.51 and 5.58 also suggest that, from a noise-reduction standpoint, it is desirable to have as \( J >> R \), i.e. many detectors and few expansion-functions (\( K \geq R \)).

285
Since the $\Sigma_{g_n}^{(i)}$'s are zero for $g = 1$, there is no analogous formula for the $\hat{\phi}_{1n}(t)$'s. Their variances "follow" those of the $\hat{\phi}_{2n}(t)$'s since both are reconstructed from the same matrix $\mathbf{A}$ and mixing-coefficients $T_{kLS}(t)$.

In the following, Eq. 5.51 will be referred to as the 'R-over-J rule'. In deriving this relationship from the more rigorous form of Eq. 5.52, a number of assumptions have been made, some of them echoing important statements already made about potential difficulties. In general, it should be remembered that the R-over-J rule, albeit exact on the average, is only approximate for a particular node. It is valid locally insofar as

(i) The J neutron-detectors are reasonably distributed so that $\left[ \psi_{2_n}^{(i)} \right]^2$ is a representative average of the $\left[ \psi_{2n_j}^{(i)} \right]^2$'s in Eq. 5.56;

(ii) The characteristic weight-functions are given by Eq. 4.3, which implies in particular that the detectors are primarily sensitive to thermal neutrons, and are located within a single node $n_j$ each (otherwise, Eqs. 5.47 and 5.48 apply);

(iii) No high-order perturbation mode affects the node of interest, $n$, or else there are detectors in or near that node (so that $\left[ \psi_{2n}^{(i)} \right]^2 / \left[ \psi_{2n_j}^{(i)} \right]^2 \approx 1$ in Eq. 5.57).

When Item iii is not satisfied, the variance of $\hat{\phi}_{2n}(t)$ is larger than $\sigma^2 R / J$.

5.3. Simulations with Noisy, Unfiltered Detector-Readouts.

5.3.1. Artificial Noise.
In Chapter 4, the discontinuous synthesis-method was tested using a PWR model. A severe operational transient was analyzed with three different sets of detectors: 80 in-core detectors (Section 4.6.2), 4 in-core detectors (Section 4.7.1), and 48 out-of-core detectors (Section 4.7.2). All detector signals were assumed to be noise-free. Very accurate results were obtained, in particular with 80 detectors.

These simulations have been repeated with artificial noise added to all signals. The random component, \( \delta C^{(j)}(t) \), added to each signal, \( C^{(j)}(t) \), was generated in the following manner:

\[
\delta C^{(j)}(t) = \alpha C^{(j)}(t), \quad j = 1,2,\ldots,J,
\]  

(5.59)

where \( \alpha \) is a random variable having a Gaussian distribution characterized by

\[
E\{\alpha\} = 0,
\]

(5.60)

and

\[
E\{\alpha^2\} = (0.05)^2 = 2.5 \times 10^{-3}.
\]

(5.61)

It follows from this choice and from Eq. 5.59 that

\[
E\{\delta C^{(j)}\} = 0, \quad j = 1,2,\ldots,J,
\]

(5.62)

\[
E\{\delta C^{(j)}\}^2 = \sigma^{(j)^2} = C^{(j)}(t) E\{\alpha^2\}
\]

\[
= (0.05)^2 C^{(j)}(t), \quad j = 1,2,\ldots,J,
\]

(5.63)

\[
E\{\delta C^{(j)} \delta C^{(m)}\} = 0, \quad j \neq m, \quad \begin{cases} j = 1,2,\ldots,J, \\
m = 1,2,\ldots,J. \end{cases}
\]

(5.64)
The assumptions of unbiased and uncorrelated signals are therefore verified. However, the assumptions of equal-variance measurements and stationary noise are not satisfied. Eq. 5.63 indicates that, instead, each detector-readout is characterized by its own variance, \( \sigma^{(j)^2} \), which depends on time. This dependence is a simple proportionally to the noise-free signal, \( C_0(t) \).

In the tests, however, it is assumed that this time-dependence is not known and is difficult to estimate. Accordingly, no row-weighting is applied to the \( \Delta \) matrices. In such conditions, the least-squares problem is not properly weighted, and its solution is not "the best one" in the sense that it does not have minimum-variance.

The choice of a dynamic, rather than static, noise is deliberately made for two reasons. The first reason is that the presence of dynamic noise gives rise to interesting phenomena which would remain hidden by considering only stationary noise. The second reason is to demonstrate that, even when the conditions of Section 5.2.2.a are not satisfied, the statistical analysis performed, although not applicable in principle, still retains some value and makes it possible to explain the results of the tests.

The choice of a fairly high value of 5 % for the expected relative noise-level is also deliberate. With such a value, random errors should completely dominate the systematic errors caused by the synthesis approximation. Indeed, it was found in the tests of Chapter 4 that, with a four-mode, time-discontinuous synthesis, systematic errors in reconstructed group fluxes were, on the average, less than 0.1 %, and never exceeded a few percent except in a handful of isolated nodes, and for a limited time only.

5.3.2. Tests with Eighty In-Core Detectors.

This first series of tests involves the instrumentation described in Section 4.3 and shown in Figures 4.2a and 4.2b. There are \( J = 80 \) neutron
detectors, equally distributed in four planes. Section 4.6.2 contains a set of graphs obtained from simulations using noise-free detector-signals. Some of these graphs will serve as references for the present tests.

To facilitate comparisons between graphs, figure numbers are summarized in Table 5.1.

<table>
<thead>
<tr>
<th></th>
<th>Four-Mode Synthesis</th>
<th>Three-Mode Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Residual</td>
<td>Figure 5.3</td>
<td>Figure 5.14</td>
</tr>
<tr>
<td>Euclidean Norm of Solution Vector</td>
<td>Figure 5.4</td>
<td>Figure 5.15</td>
</tr>
<tr>
<td>Modal Coefficients</td>
<td>Figures 5.5, 5.6, 5.7, 5.8</td>
<td>Figures 5.5, 5.6, 5.7</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 893</td>
<td>Figure 5.9</td>
<td>Figure 5.16</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 651</td>
<td>Figure 5.10</td>
<td>Figure 5.17</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 649</td>
<td>Figure 5.11</td>
<td>Figure 5.18</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 495</td>
<td>Figure 5.12</td>
<td>Figure 5.19</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 244</td>
<td>Figure 5.13</td>
<td>Figure 5.20</td>
</tr>
</tbody>
</table>

Table 5.1. Figure Numbering for the Results of the Tests With 80 In-Core Detectors.
Figure 5.2 shows the noisy output of the detector located in Node # 893 (Figure 4.11).

![Graph showing flux versus time with and without noise](image)

**Figure 5.2. Noise-Free and Noisy Outputs of the Neutron Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors).**

This noise has manifestly zero mean, and the dispersion of the points increases as the local group-two flux rises. (Recall that the detector noise-free output is equal to the actual group-two flux in Node # 893.)

The time-discontinuous fitted-synthesis method is applied with the same subsets of four expansion-functions per time-interval as in Section
4.6.2, and with all four modes included. The resulting relative residual, \( \| E_{LS}(t) \| / \| C(t) \| \), is shown in Figure 5.3. The relative residual obtained when no noise is present is also plotted for comparison.

![Relative Residual Versus Time](image.png)

**Figure 5.3.** Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).

Figure 5.3 confirms that random errors completely dominate systematic errors.

To explain the figure, one needs the help of Eq. 5.35b divided on both sides by \( \| C(t) \|^2 \):

291
\[
\mathbb{E} \left\{ \frac{\|E_{LS}(t)\|^2}{\|C(t)\|^2} \right\} = (J - R) \frac{\sigma^2}{\|C(t)\|^2},
\] (5.65)

In Eq. 5.65, contrary to the assumptions of Section 5.2.2, \( \sigma^2 \) is not the variance of all detectors, and it is not constant. However, it may be interpreted as an "average" detector-variance, and may be approximated from Eq. 5.63 as

\[
\sigma^2 \approx \frac{1}{J} \sum_{j=1}^{J} \sigma^{(j)}^2 = \frac{1}{J} (0.05)^2 \|C(t)\|^2.
\] (5.66)

Substitution of this approximation into Eq. 5.65 gives

\[
\mathbb{E} \left\{ \left[ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right]^2 \right\} \approx \frac{J - R}{J} (0.05)^2.
\] (5.67)

This equation shows that the expected value of the square of the relative residual is a constant approximately equal to

\[
\mathbb{E} \left\{ \left[ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right]^2 \right\} \approx \frac{80 - 4}{80} (0.05)^2 = 2.375 \times 10^{-3} \approx (0.0487)^2.
\]

If one makes the additional approximation:

\[
\mathbb{E} \left\{ \left[ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right]^2 \right\} \approx \left[ \mathbb{E} \left\{ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right\} \right]^2,
\] (5.68)

(i.e. if one neglects the variance in \( \|E_{LS}(t)\| / \|C(t)\| \)), then the following estimate of the expected value of the relative residual is obtained:

\[
\mathbb{E} \left\{ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right\} = 0.0487.
\] (5.69)
This result is in good agreement with Figure 5.3. The value computed from the 721 points is 0.0478 (variance $\approx 3.19 \times 10^{-5}$).

Figure 5.4 is a plot of the Euclidean norm of the least-squares solution-vector, $\|T'_{LS}(t)\|$, as a function of time.

Figure 5.4. Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).

Observe that, in conformity with Eqs. 5.18, 5.30, and 5.60, both expected value and dispersion increase with time. As with the relative residual, $\|T'_{LS}(t)\|$ is now biased.
The fairly large values for $\| T'_{LS}(t) \|$ are a sign that some of the components of $T'_{LS}(t)$ have large magnitude themselves. This is confirmed by plots of these individual components, $T'^{(i)}_{LS}(t)$, as a function of time (Figures 5.5 through 5.8).

![Graph showing modal coefficient versus time with markers for 'Without Noise' and 'With Noise'.](image)

**Figure 5.5.** First Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors).
Figure 5.6. Second Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors).
Figure 5.7. Third Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors).
Figure 5.8. Fourth Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors).

As predicted by Eq. 5.18, the modal coefficients are unbiased: the points are equally distributed on both sides of the noise-free values.

The dispersion of the points with respect to the expected values increases with time, which is in agreement with Eq. 5.27. In this particular simulation, the variances of the $T_{LS}^{(i)}(t)$'s are equal to

$$\text{var} \left\{ T_{LS}^{(i)}(t) \right\} = \frac{\sigma^2}{s_i^2} \approx \frac{1}{80} (0.05)^2 \frac{\|C(t)\|^2}{s_i^2}, \quad i = 1,2,3,4.$$  \hspace{1cm} (5.70)
The dispersion increases with $i$: The smallest variance corresponds to $i = 1$ (largest singular value $s_1$), and the largest variance to $i = R = 4$ (smallest singular value $s_4$). Recall from Section 4.6.2 that, in this particular simulation, $s_1 / s_4 = 5000$ throughout the transient.

Reconstructed node-averaged group-two fluxes, $\hat{\phi}_{2n}(t)$, appear in Figures 5.9 through 5.13 for the five representative nodes # 893, 651, 649, 495, and 244 of Figure 4.11. The reference, QUANDRY-computed solutions are also shown for comparison.

![Figure 5.9](image)

Figure 5.9. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.10. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.11. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.12. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.13. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, No Singular Value Zeroed).

Comparison of Figures 5.2 and 5.9 for the group-two flux in instrumented node # 893 permits a first verification of the R-over-J rule (Eq. 5.51). The rule predicts an average variance of

$$\text{var} \left\{ \hat{\phi}_{2n}(t) \right\} = \frac{4}{80} \sigma^2 = 0.05 \sigma^2 = \frac{1}{(4.5)^2} \sigma^2 \quad (5.71)$$

where $\sigma^2$ increases approximately as the average detector-output, $\|C(t)\|^2 / J$ (Eq. 5.66). Figures 5.2 and 5.9 are in agreement with this result.
Table 5.2 shows the computed variance-reduction factor,
\[
\text{var} \left( \hat{C}^{(i)}(t) \right) / \sigma^2 = \text{var} \left( \hat{\phi}_{2n_j}(t) \right) / \sigma^2 = \sum_{i=1}^{R} \left[ u_i \hat{\beta} \right]^2, 
\]
of Eq. 5.44 for the reconstructed group-two flux in node \( n_j = 893 \), \( \hat{\phi}_{2893}(t) \).

### Table 5.2. Variance Reduction Factor for the Reconstructed Group-Two Flux in Node # 893 (Four-Mode Synthesis).

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>( \text{var} \left( \hat{\phi}_{2893}(t) \right) / \sigma^2 )</th>
<th>( \frac{1}{80} \sum_{j=1}^{80} \text{var} \left( \hat{\phi}_{2n_j}(t) \right) / \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 10 s]</td>
<td>0.0284</td>
<td>0.05</td>
</tr>
<tr>
<td>[10 s, 20 s]</td>
<td>0.0146</td>
<td>0.05</td>
</tr>
<tr>
<td>[20 s, 30 s]</td>
<td>0.0128</td>
<td>0.05</td>
</tr>
<tr>
<td>[30 s, 40 s]</td>
<td>0.0113</td>
<td>0.05</td>
</tr>
<tr>
<td>[40 s, 50 s]</td>
<td>0.0358</td>
<td>0.05</td>
</tr>
<tr>
<td>[50 s, 60 s]</td>
<td>0.0512</td>
<td>0.05</td>
</tr>
<tr>
<td>[60 s, 70 s]</td>
<td>0.0789</td>
<td>0.05</td>
</tr>
<tr>
<td>[70 s, 80 s]</td>
<td>0.0106</td>
<td>0.05</td>
</tr>
<tr>
<td>[80 s, 90 s]</td>
<td>0.0137</td>
<td>0.05</td>
</tr>
<tr>
<td>[90 s, 100 s]</td>
<td>0.0118</td>
<td>0.05</td>
</tr>
<tr>
<td>[100 s, 110 s]</td>
<td>0.0552</td>
<td>0.05</td>
</tr>
<tr>
<td>[110 s, 180 s]</td>
<td>0.0385</td>
<td>0.05</td>
</tr>
</tbody>
</table>
The computed average reduction is the predicted value \(0.05 = \frac{80}{4}\). It can be seen that the variance reduction in \(\hat{\phi}_{2893}(t)\) is generally better than that average value.

Figures 5.9 through 5.13 confirm several expected results:

(i) Zero-mean noise does not cause any bias in the \(\hat{\phi}_{2n}(t)\)'s;
(ii) The dispersion increases with time until \(t \approx 120\) s. Beyond that point, the \(C^{(i)}\)'s, and therefore the \(\sigma^{(i)}\)'s, are almost constant;
(iii) Nodes such as Node \# 649 and Node \# 244 in which the high-order mode \(\psi^{(4)}_{2n}\) is nonnegligible exhibit a stronger dispersion than predicted by the R-over-J rule.

Figure 4.28 shows that, for node \(n = 649\), \(\psi^{(4)}_{2n}\) is significant only between \(t = 40\) s and \(t = 70\) s. The largest deviations in Figure 5.11 occur within that time-interval too.

For node \(n = 244\), the contribution of \(\psi^{(4)}_{2n}\) is felt only between \(t = 0\) s and \(t = 20\) s (Figure 4.30). In Figure 5.13, the dispersion of the points is also more pronounced in that interval;

(iv) Nodes such as Node \# 495 in which \(\psi^{(4)}_{2n}\) does not contribute appreciably to the synthesis remain much less sensitive to noise (Figure 5.12).

In an attempt to reduce the sensitivity of nodes such as Node \# 649 and Node \# 244 to noise without sacrificing too much accuracy, the smallest singular value, \(s_4\), is set to zero. The effect of this zeroing is to reduce the four-mode synthesis (Eq. 3.71) to only three modes, the highest-order mode, \(\psi^{(4)}\), being eliminated. The new \(\underline{A}\) matrices are still 80-by-4, but have rank \(R = 3\) instead of 4. It was found in Chapter 4 that such a three-mode synthesis was capable of reconstructing nodal group-fluxes accurately. The condition numbers, \(\kappa(\underline{A})\), are now such that

\[
4.730 \times 10^{-3} \leq \frac{1}{\kappa(\underline{A})} \leq 1.439 \times 10^{-2}.
\]
Figure 5.14 shows that neglecting $\Psi^{(4)}$ has essentially no impact on the relative residual of the least-squares fit.

Figure 5.14. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).

Eq. 5.67 gives an estimate of the expected value of the square relative residual:

$$\mathbb{E} \left\{ \frac{\|E_{Ls}(t)\|}{\|C(t)\|} \right\} \approx \frac{80 - \frac{3}{80}(0.05)^2}{2.406 \times 10^{-3}} = (0.0491)^2.$$

With the variance neglected, one obtains for the expected relative residual:
\[ E \left\{ \frac{\| E_{LS}(t) \|}{\| C(t) \|} \right\} \approx 0.0491, \tag{5.72} \]

a value which agrees well with the experimental result of 0.0482 (variance = 3.17 \times 10^{-5}) in Figure 5.14, and which is only marginally greater than in the four-mode synthesis (0.0487).

The Euclidean norm of the least-squares solution-vector, \( \| T'_{LS}(t) \| \), is displayed in Figure 5.15.

![Euclidean Norm of Least-Squares Solution-Vector Versus Time](image)

**Figure 5.15.** Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
This figure should be compared with the result of the four-mode synthesis, Figure 5.4. Observe that the large variations in $\|T_L^*(t)\|$ caused previously by the fourth modal coefficient (Figure 5.8) have been considerably reduced.

Figures 5.16 through 5.20 compare the reconstructed group-two flux in node # 893, 651, 649, 495, and 244 with the corresponding reference values. These figures should be compared with Figures 5.9 through 5.13 to appreciate the effect of zeroing the smallest mode.

Figure 5.16. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.17. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.18. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.19. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.20. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).

Careful comparisons reveal that the dispersion of the points has been slightly reduced. This is indeed what the R-over-J rule predicts: R/J is now 3/80 instead of 4/80:

\[ \text{var} \{ \hat{\phi}_{2n}(t) \} = \frac{3}{80} \sigma^2 = 0.0375 \sigma^2 = \frac{1}{(5.2)^2} \sigma^2. \]  

(5.73)

The variance reduction factor for \( \hat{\phi}_{2893}(t) \) is reproduced in Table 5.3 for each of the twelve time-intervals.
Table 5.3. Variance Reduction Factor for the Reconstructed Group-Two Flux in Node # 893 (Three-Mode Synthesis).

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>( \text{var}\left(\hat{\phi}_{2893}(t)\right)/\sigma^2 )</th>
<th>( \frac{1}{80} \sum_{j=1}^{80} \text{var}\left(\hat{\phi}_{2893j}(t)\right)/\sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 10 s]</td>
<td>0.00980</td>
<td>0.0375</td>
</tr>
<tr>
<td>[10 s, 20 s]</td>
<td>0.0111</td>
<td>0.0375</td>
</tr>
<tr>
<td>[20 s, 30 s]</td>
<td>0.0124</td>
<td>0.0375</td>
</tr>
<tr>
<td>[30 s, 40 s]</td>
<td>0.00982</td>
<td>0.0375</td>
</tr>
<tr>
<td>[40 s, 50 s]</td>
<td>0.00674</td>
<td>0.0375</td>
</tr>
<tr>
<td>[50 s, 60 s]</td>
<td>0.0213</td>
<td>0.0375</td>
</tr>
<tr>
<td>[60 s, 70 s]</td>
<td>0.0708</td>
<td>0.0375</td>
</tr>
<tr>
<td>[70 s, 80 s]</td>
<td>0.0749</td>
<td>0.0375</td>
</tr>
<tr>
<td>[80 s, 90 s]</td>
<td>0.0914</td>
<td>0.0375</td>
</tr>
<tr>
<td>[90 s, 100 s]</td>
<td>0.0732</td>
<td>0.0375</td>
</tr>
<tr>
<td>[100 s, 110 s]</td>
<td>0.0494</td>
<td>0.0375</td>
</tr>
<tr>
<td>[110 s, 180 s]</td>
<td>0.0328</td>
<td>0.0375</td>
</tr>
</tbody>
</table>

Table 5.3 should be compared with Table 5.2 for the four-mode synthesis. The variance reduction factor of \( \hat{\phi}_{2893}(t) \) has decreased in some time-intervals, but has increased in others.
The variance of \( \hat{\phi}_{2n}(t) \) has been reduced in perturbed nodes \( n = 649 \) (Figures 5.11 and 5.18), and \( n = 244 \) (Figures 5.13 and 5.20). However, the sensitivity to noise remains appreciable in Node # 649 between \( t = 50 \) s and \( t = 60 \) s. This happens because the third mode, \( \psi^{(3)} \), is relatively large in that node and in that time interval, while being a fairly high-order mode since its magnitude is \( s_3 = \| \Sigma^T \psi^{(3)} \| = 4.73 \times 10^{-3} s_1 \ll s_1 \) (See Table 4.4). By eliminating this third mode, sensitivity to noise would be reduced. However, the price to pay would be large, systematic errors, as seen in Chapter 4 and as demonstrated by Figure 5.21.
In Section 5.4, an alternative means of reducing sensitivity to noise will be presented in the form of simple filtering algorithms.

Plots of amplitude function, $\hat{n}_{\text{eff}}(t)$, total fission-power, $\hat{P}(t)$, and reactivity, $\hat{\rho}(t)$, as reconstructed by the three-mode synthesis are shown in Figures 5.22, 5.23, and 5.24 respectively. Reference curves are also included.
Figure 5.22. Reference and Reconstructed Amplitude Function Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.23. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Noisy Detectors, One Singular Value Zeroed).
There is relatively little dispersion in amplitude function, and total power. Random (as well as systematic) errors are reduced by sign cancellations which occur in the calculation of these integral quantities from $\hat{\phi}(t)$ (Eqs. 2.17 and 2.18). In contrast, reactivity remains fairly noisy.

Notice that the slight bias in reconstructed total power (caused by the assumption of constant fission cross sections) is hardly visible in Figure 5.23: it is masked by random errors.
5.3.3. Tests with Four In-Core Detectors.

The importance of the number of detectors when noise is present can be best appreciated by repeating the above simulations with only $J = 4$ in-core detectors instead of eighty. The positions of these four detectors is the same as in the noise-free tests of Section 4.7.1.

Figure numbers for the forthcoming graphs are listed in Table 5.4.

Table 5.4. Figure Numbering for the Results of the Tests With 4 In-Core Detectors.

<table>
<thead>
<tr>
<th></th>
<th>Four-Mode Synthesis</th>
<th>Three-Mode Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Residual</td>
<td>Figure 5.25</td>
<td>-</td>
</tr>
<tr>
<td>Euclidean Norm of Solution Vector</td>
<td>Figure 5.26</td>
<td>-</td>
</tr>
<tr>
<td>Modal Coefficients</td>
<td>Figures 5.27, 5.28, 5.29, 5.30</td>
<td>Figures 5.27, 5.28, 5.29</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 893</td>
<td>Figure 5.31</td>
<td>Figure 5.34</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 651</td>
<td>Figure 5.32</td>
<td>Figure 5.35</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 649</td>
<td>Figure 5.33</td>
<td>Figure 5.36</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 495</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 244</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The matrix systems to be solved at every time-step are now square, four-by-four systems. There is no extra degree of freedom. If the systems are not singular, and if no singular value is zeroed, the solution vector, $T_{LS}(t)$, is exact, equal to $T(t) = A^{-1} [C(t) + \delta C(t)]$. The four components of this vector, $T^{(k)}(t)$, “follow” the four noisy signals, $C^{(j)}(t) + \delta C^{(j)}(t)$.

The relative residual is almost zero to within machine accuracy, as evidenced by Figure 5.25.

![Graph](image)

**Figure 5.25.** Relative Residual Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed).
The combination of noise and numerical roundoff causes a slight, upward bias.

The Euclidean norm of the solution vector is plotted in Figure 5.26.

Figure 5.26. Euclidean Norm of Solution Vector Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed).

As with 80 detectors (Figure 5.4), the dispersion of the points increases with time. Note that the variations are larger than before, consistently with the fact that the ill-conditioning of the matrices is more pronounced (condition numbers larger by a factor of roughly 5).
Individual modal coefficients, $T^{(i)}(t), i = 1,2,3,4$, are compared with the noise-free values in Figures 5.27 through 5.30.

![Graph](image)

Figure 5.27. First Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors).
Figure 5.28. Second Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors).
Figure 5.29. Third Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors).
Figure 5.30. Fourth Modal Coefficient Versus Time (Operational Transient, Four In-Core Noisy Detectors).

The patterns are similar to those of Figures 5.5 through 5.8, but the dispersion ($\sigma^2/s_i^2$ in Eq. 5.27) is more pronounced (larger condition numbers).

The anomalously large values for $T_{n}^{(3)}(t)$ in the interval [50 s, 60 s] are explained by the fact that $s_1/s_3$ is equal to 2.30 $10^3$ in that interval, a value greater than in any other interval by more than one order of magnitude.

Figures 5.31 through 5.33 show the reconstructed group-two flux in Nodes # 893, 651, and 649. These three figures should be compared with their analogs, Figures 5.9, 5.10, and 5.11, obtained with 80 detectors.
Figure 5.31. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.32. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.33. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Noisy Detectors, No Singular Value Zeroed).

For Node # 893, the reconstructed flux is identical to the output of the detector located in that node (Figure 5.2).

Figure 5.33 is a dramatic illustration of the extreme sensitivity of the present synthesis to noise: Negative values appear in $\hat{\phi}_{2n}(t)$ in the interval [50 s, 60 s].

When the smallest singular value, $s_4$, is zeroed, the sensitivity to noise is slightly reduced, as Figures 5.34, 5.35, and 5.36 illustrate for the same three nodes.
Figure 5.34. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, Four In-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.35. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, Four In-Core Noisy Detectors, One Singular Value Zeroed).
The R-over-J rule predicts a variance

$$\text{var} \left( \hat{\phi}_{2n}(t) \right) = \frac{3}{4} \sigma^2 = 0.75 \sigma^2 = \frac{1}{(1.15)^2} \sigma^2,$$

(5.74)

i.e. a slight reduction which is indeed observed. The rule does not apply, however, to nodes such as Node # 649 between $t = 50$ s and $t = 60$ s. The third mode, $\psi^{(3)}$, and its associated, small singular value, $s_3$, still cause excessive sensitivity to noise in that node and in that time interval. Note that the situation is much worse than with 80 detectors (Figure 5.18).
By zeroing $s_3$, negative fluxes disappear (Figure 5.37). However, the price to pay is large systematic errors (Table 4.7c).

Figure 5.37. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, Four In-Core Noisy Detectors, Two Singular Values Zeroed).
5.3.4. Tests with Forty-Eight Out-of-Core Detectors.

The same series of synthesis calculations was repeated with the $J = 48$ out-of-core detectors of Section 4.7.2 (Figure 4.52), with the noise of Eq. 5.53 added to all signals.

Figure numbers for important graphs are displayed in Table 5.5.

Table 5.5. Figure Numbering for the Results of the Tests With 48 Out-of-Core Detectors.

<table>
<thead>
<tr>
<th></th>
<th>Four-Mode Synthesis</th>
<th>Three-Mode Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Residual</td>
<td>Figure 5.39</td>
<td>Figure 5.48</td>
</tr>
<tr>
<td>Euclidean Norm of Solution Vector</td>
<td>Figure 5.40</td>
<td>-</td>
</tr>
<tr>
<td>Modal Coefficients</td>
<td>Figures 5.41, 5.42, 5.43, 5.44</td>
<td>Figures 5.41, 5.42, 5.43</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 893</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 651</td>
<td>Figure 5.45</td>
<td>Figure 5.49</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 649</td>
<td>Figure 5.46</td>
<td>Figure 5.50</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 495</td>
<td>Figure 5.47</td>
<td>Figure 5.51</td>
</tr>
<tr>
<td>Reconstructed Group-Two Flux in Node # 244</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The noisy output of the detector located in Node # 495 appears in Figure 5.38.

![Graph](image)

**Figure 5.38.** Noise-Free and Noisy Outputs of the Detector Located in Node # 495 Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors).

Processing of these 48 noisy signals through the four-mode synthesis program yields a fitted solution characterized by the relative residual of Figure 5.39.
Figure 5.39. Relative Residual Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors, No Singular Value Zeroed).

This figure provides another test for Eq. 5.67:

\[
\mathbb{E} \left\{ \left[ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right]^2 \right\} = \frac{48 - 4}{48} (0.05)^2 = \frac{2.292 \times 10^{-3}}{} = (0.0479)^2,
\]

giving approximately for the expected relative residual:

\[
\mathbb{E} \left\{ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right\} = 0.0479,
\]

(5.75)
a value slightly smaller than with 80 detectors (Eq. 5.69), but in good agreement with Figure 5.39 (mean value = 0.0445, variance = 8.47 \times 10^{-5}). A possible explanation for this slightly lower value may be that the least-squares fit is "easier" in reflector regions.

Figure 5.40 shows the Euclidean norm of the least-squares solution vector, \( \| T_{LS}(t) \| = \| T'_{LS}(t) \| \), plotted as a function of time.

![Graph showing Euclidean norm versus time](image)

**Figure 5.40.** Euclidean Norm of Least-Squares Solution-Vector Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors, No Singular Value Zeroed).
As in the previous cases and as predicted by Eq. 5.30, the dispersion increases with time at a rate governed by $\sigma^2 \sum_{i=1}^{R} \frac{1}{s_i^2} \approx \sigma^2 / s_R^2$.

Individual modal coefficients, $T^{(0)}_{LS}(t)$, are compared with the noise-free values in Figures 5.41 through 5.44.

---

**Figure 5.41.** First Modal Coefficient Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors).
Figure 5.42. Second Modal Coefficient Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors).
Figure 5.43. Third Modal Coefficient Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors).
Figure 5.44. Fourth Modal Coefficient Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors).

The increased dispersion with respect to the 80-detector case (Figures 5.5 through 5.8) reflects the fact that the condition numbers are slightly larger here (by a factor of approximately 2).

Figures 5.45 through 5.47 show the reconstructed group-two fluxes, $\hat{\phi}_{2n}(t)$, in Nodes # 651, 649, and 495, together with the corresponding reference values.
Figure 5.45. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.46. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors, No Singular Value Zeroed).
Figure 5.47. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-Of-Core Noisy Detectors, No Singular Value Zeroed).

According to the R-over-J rule, the variance of the reconstructed group-two fluxes should be approximately

\[
\text{var}\left\{\hat{\phi}_{2n}(t)\right\} = \frac{4}{48} \sigma^2 \approx 0.0833 \sigma^2 \approx \frac{1}{(3.5)^2} \sigma^2. \quad (5.76)
\]

Comparison of Figures 5.38 and 5.47 for the measured and synthesized group-two flux in Node # 495 indicates that this is approximately the case.

Again, the rule breaks down for Node # 649 between \( t = 50 \) s and
$t = 60$ s. The fourth mode, $\psi^{(4)}_{2n}$, in that node and in that time-interval is sensitive to the large fluctuations in the modal coefficient $T^{(4)}_{LS}(t)$ (Figure 5.44).

When $\psi^{(4)}_{2n}$ is eliminated by zeroing of $s_4$, little deterioration results in the relative residual of the least-squares fit, as can be seen from Figure 5.48.

![Graph showing relative residual versus time](image)

Figure 5.48. Relative Residual Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed).
The new expected value for \[ \left( \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right)^2 \] is approximately (Eq. 5.67):

\[
E \left\{ \left( \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right)^2 \right\} \approx \frac{48 - 3}{48} (0.05)^2 \approx 2.344 \times 10^{-3} \approx (0.0484)^2.
\]

Therefore, roughly,

\[
E \left\{ \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right\} \approx 0.0484,
\]  \hspace{1cm} (5.77)

in agreement with Figure 5.48 for which the average value is 0.0457 and the variance 8.50 \times 10^{-5}.

This three-mode synthesis gives the group-two fluxes of Figures 5.49, 5.50, and 5.51 for the same three nodes as in Figures 5.45, 5.46, 5.47.
Figure 5.49. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.50. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed).
Figure 5.51. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, One Singular Value Zeroed).

The average variance-reduction factor is now

\[
\frac{R}{J} = \frac{3}{48} = 0.0625 = \frac{1}{(4.0)^2},
\]

(5.78)

instead of \(1 / (3.5)^2\) for the four-mode synthesis. Nevertheless, as in the four-detector case, this average decrease cannot eliminate the large fluctuations in nodes such as Node # 649, between \(t = 50\) s and \(t = 60\) s.

Zeroing \(s_3\) permits achieving this objective (Figure 5.52), but at the expense of a net increase in systematic errors.
Figure 5.52. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 48 Out-of-Core Noisy Detectors, Two Singular Values Zeroed).

In summary, these simple simulations confirm the predictions made in Chapter 3, namely that measurement noise may be amplified by the fitted synthesis because of ill-conditioning in the matrix systems being solved. However, this amplification is limited to a few nodes and to certain time-intervals. For the vast majority of the nodes, and most of the time, the variance in reconstructed fluxes is actually reduced by the least-squares fit by an average factor of R / J, i.e. rank of the matrices over number of detectors.

The coefficients of the high-order perturbation-modes introduced in the synthesis are the most sensitive to noise, reflecting the fact that
extracting small, detailed information from noisy signals may be difficult, especially when these signals are in limited number or come from detectors located far away from the perturbations. If one is willing to accept the trade-off of small, extra, systematic errors in the results, the highest-order modes can be eliminated. This elimination reduces the sensitivity to noise. In some cases, however, this reduction may not be sufficient, prompting the need for some independent noise-filtering means.


The function of any digital filter receiving as input a sequence of noisy measurements, \( \{x_k\} \), is to find an estimate, \( y_n \), for the value of the actual, unknown, noise-free signal at some specified time \( t = t_n \) (Figure 5.53). The filter is efficient if the variance of \( y_n \) is smaller than that of \( x_n = x(t_n) \). The ratio of the two variances (variance reduction factor) is a measure of the efficiency of the filter.

![Diagram](image)

Figure 5.53. Filtered Output \( y_n \) Evaluated from a Sequence of Measurements \( x_k \).
To reduce random errors in the GN nodal group-fluxes, $\hat{\phi}_{gn}(t)$, reconstructed by the fitted-synthesis method, filtering algorithms can be applied either to the $\hat{\phi}_{gn}(t)$'s themselves, or to quantities from which they are computed, i.e. the R modal coefficients, $T_{LS}^{(i)}(t)$, or the J detector readouts, $C^{(j)}(t)$. From a computing-efficiency standpoint, the $T_{LS}^{(i)}(t)$'s are the quantities to be preferred because of their relatively small number, $R \ll J \ll GN$. However, these modal coefficients (as well as the $\hat{\phi}_{gn}(t)$'s) are extremely difficult to filter because they are discontinuous functions of time in the synthesis: They change abruptly at the beginning of every new time-interval. In these conditions, the only alternative is direct filtering of the detector signals, $C^{(j)}(t)$, $j = 1,2,...,J$.

Each detector should possess its own filter so that individual adjustments are possible if necessary. In the following, the focus is on one arbitrary filter and its associated detector.

The generic notation of Figure 5.53 is used in the presentation. The sequence $\{x_k\}$ denotes the inputs of the filter. (Variations in the subscript conventions may lead to simpler or more convenient formulas.) $t_0$ is the instant at which the most recent measurement, $x_0$, has been received. This sequence is obtained by periodic sampling of the analog signal received from the flux sensor. The sampling period is a constant denoted $\tau$.

$y_n$ is the filter output at $t = t_n$, the selected instant of filtering or "validity instant". Note that $t_n$ does not need to be equal to $t_0$. Retrodicting ($t_n < t_0$), updating or zero-step estimation ($t_n = t_0$), and predicting ($t_n > t_0$) are different forms of filtering. The terminology "smoothing" is generally reserved for off-line filtering, but it is sometimes used in lieu of retrodicting.

The objective is to design a filter capable of achieving a specified variance reduction factor (VRF) or "smoothing factor". The filter outputs should be available in real time (time scale $\sim 1$ s), or with only a small time-delay $t_0 - t_n$ ($\sim 1$ s).
5.4.1. Trade-Offs in Filtering.

It is important to realize that filtering a signal introduces not only complications, but also constraints and errors.

One constraint which is imposed by filtering is the requirement of a high-enough sampling rate to avoid aliasing of some high-frequency component of the signal into a low-frequency component. Another constraint is the appearance of a time delay when the filter is used for retrodicting.

Errors caused by filtering can be of two kinds: systematic errors, and transient errors. Systematic errors arise from assumptions made in designing the filter. These assumptions concern either the shape of the "true", unknown, noise-free signal (in the time domain), or the system transfer function (in the frequency domain). Transient errors, on the other hand, arise when the filter uses its own, past estimates, \( y_{n-1}, y_{n-2}, \ldots \), to compute \( y_n \) recursively.

Attempting to reduce random errors excessively may lead to either intolerable constraints (e.g. an excessive lag), or unacceptably large systematic or transient errors. Filtering always implies a compromise between noise reduction on the one hand, and restrictive constraints, systematic errors, and transient errors on the other hand. This compromise depends on the particular application. There is no "perfect" filter, nor is there an optimum filter for all situations.

The minimum-sampling-rate constraint does not pose serious difficulties here. Indeed, as mentioned in Section 2.5, sampling periods of the order of 0.1 s are achievable with flux sensors of the fission-chamber or prompt-responding-SPND type. This is less than the bandwidth (i.e. highest frequency, \( f_{\text{max}} \)) of the system during normal operation. Recall that the fastest decay-constant of all families of delayed-neutron precursors is \( \lambda_{\text{max}} \approx 3 \text{ s}^{-1} \).
Real-time filtering calculations should also be feasible without requiring large computing resources since there will be at most a few hundred detectors.

Limiting the time-delay on the one hand, and controlling systematic and transient errors on the other hand will be the main objectives in designing and “tuning” the filters of the fitted-synthesis method.

5.4.2. Filter Design.

a. FIR Versus IIR Filters.

Designing digital filters for the detectors of the flux-synthesis method is a challenging task because very little information is available. In particular, little is known about

(i) the “true” shape of the signal being processed, except perhaps for the fact that, during normal operation, it should be a fairly smooth curve in a time scale of a few seconds;

(ii) the physical phenomena responsible for the variations of this signal with time since, as explained in Chapters 1 and 2, there is no model of system dynamic behavior.

In fact, the only information available at any time $t_0$ is:

(i) The samples $x_k$ for $t_k \leq t_0$;

(ii) The bandwidth of the system, $f_{\text{max}}$, on the order of $10 \text{ s}^{-1}$;

(iii) Statistical data about the noise. As in the previous sections, measurement noise is assumed to be additive, stationary, and to have zero mean. An estimate, $\sigma^2$, of its variance is available.

This limited amount of information is not sufficient in itself to construct a filter. To proceed with the actual design, it is necessary to postulate a relationship between the filter output, $y_n$, and the input
sequence, \{x_k\}. (This is a source of systematic errors.) Common practice is to assume a simple, linear relationship of the type

\[ y_n = \sum_{k=0}^{L} h_k x_{n-k}, \quad (5.79) \]

or

\[ y_n = \sum_{k=1}^{L'} g_k y_{n-k} + \sum_{k=0}^{L} h_k x_{n-k} \quad (5.80) \]

where \( g_k \) and \( h_k \) are real, constant, weighting coefficients to be determined. \( L' \) and \( L \) are positive integers. \( L' \) is imposed, whereas \( L \) is an adjustable parameter. Note that Eq. 5.79 is actually a particular case of Eq. 5.80.

Postulating Eq. 5.80 amounts to assuming that the \( y_n \)'s obey a linear difference equation with constant coefficients. In general, the "exact", noise-free values will not verify such an equation. Therefore, the \( y_n \)'s will contain systematic errors. The parameters of the filters (e.g. \( L \)) should be adjusted such that these systematic errors are sufficiently small.

Filters designed on the basis of Eq. 5.79 (or analogs of it) are called finite impulse-response (FIR) or nonrecursive filters, while filters designed on the basis of Eq. 5.80 (with \( g_k \neq 0 \) for some \( k \)) are called infinite impulse-response (IIR) or recursive filters [P6, O4, H3]. The latter use their past estimates, \( y_{n-1}, y_{n-2}, ..., y_{n-L'} \), to compute \( y_n \), whereas the former do not. The main advantage of a recursive formulation is that the filtering algorithm can be implemented in a very compact form which makes the computations very efficient. However, such a formulation has two drawbacks. The first one is that it requires that initial values be specified for \( y_{n-1}, y_{n-2}, ..., y_{n-L'} \). If poorly chosen, these initial values may affect the filter outputs for a fairly long time. The other, more serious drawback of a recursive formulation is that it introduces spurious transient solutions corresponding to "homogeneous modes" (or "natural modes"), the solutions of the homogeneous equation:
\[ y_n - \sum_{k=1}^{L'} g_k y_{n-k} = 0. \] (5.81)

It is important that any recursive filter be designed in such a way that it is not only stable (no growing homogeneous mode), but also such that homogeneous modes die out sufficiently rapidly so as not to cause lasting transient-errors.

Figure 5.54 is an interpretation of the FIR filter of Eq. 5.79.

![FIR Filter Diagram](image)

Figure 5.54. FIR Filter of Eq. 5.79.

Note that, with such a formula, the validity instant is positioned at the end of the filtering interval, thus making the filter *causal*. Causality
needs not be an overriding concern with FIR filters if a time delay is acceptable.

There are many different ways to express $y_n$ in terms of the $x_k$'s. An alternative to Eq. 5.79 is (for even $L$)

$$y_n = \sum_{k=-\frac{L}{2}}^{+\frac{L}{2}} h_k x_{n-k},$$

which is illustrated in Figure 5.55.

![Diagram of FIR Filter](image)

**Figure 5.55.** FIR Filter of Eq. 5.82.

In this case, the validity instant is at the center of the filtering interval. The weights $h_k$ are different from those of Eq. 5.79.
In practical applications, \( n \) should be chosen in such a way that these weights are easily calculated. Transition formulas can then be used to find the \( h_k \)'s corresponding to a shifted validity-instant. By sufficient shifting, i.e. by introducing a sufficient time-delay, the filter can be made causal.

b. Direct Method Versus Indirect Method.

Discrete-time systems obeying linear difference equations with constant coefficients such as Eq. 5.80 belong to the class of linear, time-invariant systems, i.e. systems in which the output sequence \( y_n, y_{n-1}, y_{n-2}, \) at time \( t_n \) is related to the input values \( x_n, x_{n-1}, x_{n-2} \) by a linear operator which is invariant under translation in time: \( n \) can be replaced by \( n - n_0 \) where \( n_0 \) is an arbitrary origin [O4].

A fundamental property of linear, time-invariant systems is that their eigenfunctions are the complex exponential sequences, \( e^{i\omega_m n} \), where \( \omega_m \) is an arbitrary real number. For such systems, the discrete input signal:

\[
x_n = \sum_m \alpha_m e^{i\omega_m n}, \tag{5.83}
\]

produces as output:

\[
y_n = \sum_m \alpha_m H(\omega_m) e^{i\omega_m n}, \tag{5.84}
\]

which has the same frequency-composition as the input signal, with amplitude and phase determined by the system response-function or Fourier spectrum, \( H(\omega) \), for each frequency \( \omega_m \) [O4]. \( H(\omega_m) \) is the (complex) eigenvalue corresponding to the eigenfunction \( e^{i\omega_m n} \).

For FIR filters, \( H(\omega) \) is a polynomial function of \( z \equiv e^{i\omega} \). For IIR filters, \( H(\omega) \) is a rational function of \( z \). Indeed, for Eq. 5.79:

356
\[ H(\omega) = \sum_{k=0}^{L} h_k e^{-i\omega k}, \quad (5.85) \]

while for Eq. 5.80:

\[ H(\omega) = \frac{\sum_{k=0}^{L} h_k e^{-i\omega k}}{1 - \sum_{k=1}^{L'} g_k e^{-i\omega k}}, \quad (5.86) \]

In either case, \( H(\omega) \) is periodic with period \( 2\pi \). The definition is restricted to one period, usually the interval \([-\pi, \pi]\).

The great value of the function \( H(\omega) \) stems from the fact that, as the eigenvalue spectrum of the linear time-invariant system, it gives a complete characterization of that system.

There are essentially two methods for designing a filter, i.e. for determining the coefficients \( h_k \) of Eq. 5.79, or \( g_k \) and \( h_k \) of Eq. 5.80:

(i) The first method is the "direct method" of mathematicians and statisticians ("Regression Analysis"). It consists in making an assumption about the shape of the actual, continuous-time, noise-free signal as a function of time, t. No reference needs to be made to the system response-function \( H(\omega) \). For example, a low-degree polynomial profile may be postulated and fitted to the measurements over some interval of length \( L\tau \);

(ii) The second method is the "indirect method" of electrical engineers and signal-processing specialists. It relies on assumptions about the amplitude and phase of \( H(\omega) \) as a function of frequency, \( \omega \), i.e., in effect, about the characteristics of the actual physical system from which the signal arises.

The direct method is simple and straightforward. To see why the indirect method is also well-founded and of considerable appeal, consider its application to the design of the FIR filter given by Eq. 5.82.
Eq. 5.82 may in fact be viewed as an approximation for the infinite sum

\[ y_n = \sum_{k=-\infty}^{+\infty} h_k x_{n-k}. \]  

(5.87)

The response function of this linear time-invariant system is:

\[ H(\omega) = \sum_{k=-\infty}^{+\infty} h_k e^{-i\omega k}. \]  

(5.88)

\( H(\omega) \) in Eq. 5.88 should be recognized as the discrete Fourier transform of the impulse response of the system, \( h_n \), i.e. the response of the filter to a unit impulse at \( t = t_n \). Eq. 5.87 can also be written as a convolution sum:

\[ y_n = h_n \ast x_n, \]  

(5.89)

with the following relationship between discrete Fourier transforms:

\[ Y(\omega) = H(\omega) X(\omega). \]  

(5.90)

The weighting coefficients of Eq. 5.87 now have the meaning of Fourier coefficients which can be obtained by the inverse discrete Fourier transform integral:

\[ h_k = \frac{1}{2\pi} \int_{-\pi}^{+\pi} H(\omega) e^{i\omega k} d\omega. \]  

(5.91)

Eq. 5.91 makes it possible to determine the \( h_k \)'s from \( H(\omega) \). The reason why such a procedure for finding the \( h_k \)'s is appealing is that it is fairly simple to construct an efficient filter directly from the function \( H(\omega) \), without having to use an infinite number of Fourier coefficients \( h_k \).
To illustrate this construction procedure, assume that the filter of Eq. 5.82 has *symmetric* coefficients: \( h_k = h_{-k}, k = 1,2,...,L/2 \). The Fourier transform of this filter is simply

\[
H(\omega) = h_0 + 2 \sum_{k=1}^{+\infty} h_k \cos(k\omega),
\]

i.e. a *real, even* function of \( \omega \). (The phase is linear, i.e. is a simple time-shift.) In this particular case, there is a simple correspondence between the weights \( h_k \) of Eq. 5.82 and \( H(\omega) \).

\( H(\omega) \) of Eq. 5.92 constitutes an "ideal" low-pass filter if it has the shape of Figure 5.56 [O4].

![Diagram](image)

Figure 5.56. Response Function of Ideal Low-Pass Filter.
The frequency axis in Figure 5.56 is normalized in such a way that the actual frequency, \( f \), of the signal is \( \omega \) divided by \( 2\pi \tau \):

\[
\omega = 2 \pi f \tau . \quad (5.93)
\]

With such a convention:
(i) \( \omega = \pi \) corresponds to \( f = 1 / 2 \tau \), i.e. half the sampling rate \( 1 / \tau \);
(ii) values of \( \omega \) near 0 correspond to low frequencies in the signal, whereas values of \( \omega \) just above - \( \pi \) or just below + \( \pi \) are high frequencies;
(iii) \( \omega = 0 \) corresponds to the DC component of the signal.

The justification for the qualifier "ideal" is related to the **Nyquist sampling theorem**: A bandlimited signal (\( f_{\text{max}} < \infty \)) can be reconstructed from an infinite sequence of its samples provided the sampling rate is at least twice the frequency, \( f_{\text{max}} \), of the highest component in the signal.

Therefore, if an infinite sequence \( \{x_k\} \) has been obtained by sampling of a signal at a rate higher than \( 2 f_{\text{max}} \), then the values \( y_n \) reconstructed by Eq. 5.87 will be equal to the original values of the bandlimited signal. Sampling at \( 1 / \tau > 2 f_{\text{max}} \) assures that two samples of the highest frequency in the signal are obtained, which prevents any risk of aliasing.

A natural way of eliminating high-frequency noise from a signal which would be bandlimited in the absence of noise is to introduced a cutoff frequency, \( \omega_c \), equal to \( 2\pi f_{\text{max}} \tau \) (Figure 5.56). With this choice, there is no signal component outside the interval \([ - \omega_c, + \omega_c \]) except for high frequencies arising from noise. Any frequency component above \( f_c = \omega_c / 2\pi \tau \) is eliminated from \( y_n \) by the filter (Eq. 5.91). This filter is ideal in the sense that no valuable information is lost from filtering; only high-frequency noise is eliminated.

Some amounts of noise from the frequency range \([0, f_c]\) will inevitably remain in \( y_n \), however. In practice, if this noise is expected to be a nuisance, it may be a good idea to lower \( \omega_c \) below the ideal value
$2\pi f_{\text{max}} \tau$. This may permit a significant reduction of random errors in $y_n$ with minimal loss of important physical information.

There exists an infinite sequence of Fourier coefficients, $h_k$, such that, when used in Eq. 5.88, the function $H(\omega)$ of Figure 5.56 is obtained. In practice, this infinite series of Fourier coefficients must be truncated to obtain a causal, finite impulse-response (Eq. 5.82).

Truncating the infinite series is equivalent to multiplying it by a rectangular unit-window extending from $k = -L/2$ to $k = +L/2$. Unfortunately, such an abrupt windowing gives rise to small oscillations in $H(\omega)$ (Gibbs ripples) near the cutoff points $\omega = \pm \omega_c$. To avoid such a phenomenon, the response function is generally approximated, not by Eq. 5.82, but instead by the following formula:

$$
H(\omega) = \sum_{k = -L/2}^{+L/2} \sigma_k h_k e^{-i\omega k} = \sigma_0 h_0 + 2 \sum_{k = 1}^{+L/2} \sigma_k h_k \cos(\omega k) \quad (5.94)
$$

where the $\sigma_k$'s are introduced to "soften" the abrupt windowing. The original $h_k$'s have been simply multiplied by $\sigma_k$. A common choice for these smoothing factors $\sigma_k$ is \[H3\]:

$$
\sigma_k \equiv \frac{\sin \left( \frac{2\pi k}{L} \right)}{2\pi k/L}, \quad k = 0, \pm 1, \pm 2, \ldots, \pm \frac{L}{2}. \quad (5.95)
$$

The result of this windowing is a smooth graph of the form shown in Figure 5.57.
Figure 5.57. Response Function of Approximate Low-Pass Filter.

The cutoff is not as sharp as with the ideal low-pass filter, but there are no spurious ripples near $\omega = \pm \omega_c$.

In summary, the design of an FIR filter by the indirect method consists in the following successive steps:

(i) A shape such as that of Figure 5.56 or Figure 5.57 is postulated for $H(\omega)$. A symmetric profile facilitates the algebra and eliminates the need for specifying a phase profile;

(ii) The corresponding Fourier coefficients, $h_k$, are found from Eq. 5.91. These coefficients are multiplied by smoothing factors $\sigma_k$ if $H(\omega)$ contains discontinuities;

(iii) The result is substituted into Eq. 5.82.

In the following, simple FIR and IIR filters are constructed for the detectors of the fitted-synthesis method. The construction is based on the direct method and adapted from Reference M8. Both types of filters are tested and the results are compared. The FIR filters are subsequently modified in the spirit of the indirect method, according to a procedure described in Reference H3.
5.4.3. Fitted Synthesis with Filtered Noisy Signals.

a. Fixed-Memory FIR Filters.

The direct method of filter design was used to construct a simple, fixed-memory polynomial FIR filter [M8]. The procedure consists in approximating the unknown, noise-free signal by a polynomial in \( t \) of degree \( M \), over an interval \([t_L, t_0]\) of length \( L\tau \) \((L \geq M)\), as shown in Figure 5.58. \( t_0 \) corresponds to the most recent measurement, \( x_0 \).

![Diagram of Polynomial of Degree M](image)

Figure 5.58. Notational Conventions for Fixed-Memory Polynomial Filter.
The $r$-axis is introduced for convenience. It avoids many minus signs in the formulas. It replaces the $k$-axis of Figures 5.54 and 5.55.

The coefficients of the polynomial, $p$, are determined from the requirement that $p$ agrees with the $L+1$ measurements $x_{-L+r}$ in a least-squares sense. Mathematically, this means:

$$\text{Minimize } \sum_{r=0}^{L} \left[ x_{-L+r} - p(r) \right]^2. \quad (5.96)$$

An elegant method [M8] for solving this least-squares problem is to write $p(r)$ as a linear combination of the first $M+1$ discrete Legendre polynomials, $p_j(r)$, $j = 0, 1, 2, \ldots, M$. The orthogonality relationship between these polynomials:

$$\sum_{r=0}^{L} \frac{p_i(r) p_j(r)}{c_i c_j} = \delta_{ij}, \quad \begin{cases} i = 0, 1, \ldots, M, \\ j = 0, 1, \ldots, M, \end{cases} \quad (5.97)$$

with

$$c_j^2 = \frac{(2j+1) \, L! \, L!}{(L-j)! \, (L+j+1)!} = \sum_{r=0}^{L} \left[ p_j(r) \right]^2, \quad j = 0, 1, \ldots, M, \quad (5.98)$$

can be used to circumvent any matrix (pseudo-) inversion as well as possible numerical difficulties in solving the $(L+1)$-by-$(M+1)$ least-squares problem. Formulas for the $p_j(r)$'s can be found in Appendix A2. The result of this fitting procedure is a filtered estimate, $y_n$, of the form:

$$y_n = p(r)_{r=n} = \sum_{r=0}^{L} w_r(n) \, x_{-L+r} \quad (5.99)$$

where
\[ w_r(n) = \sum_{j=0}^{M} \frac{p_i(n)}{c_j} \frac{p_j(r)}{c_j}, \quad r = 0,1,\ldots,L. \] (5.100)

The weights \( w_r(n) \) depend on both \( L \) and \( M \) (implicitly), and on the validity instant, \( t_n \), yet to be chosen. Note that, even though the set of measurements \( x_{-L+r} \) changes at every time-step, the weights \( w_r(n) \) do not have to be recomputed (time-invariance property).

A very simple formula for the variance, \( \sigma_n^2 \), in \( y_n \) can be derived:

\[ \sigma_n^2 = \sigma^2 \sum_{r=0}^{L} \left[ w_r(n) \right]^2 = \sigma^2 \sum_{j=0}^{M} \left[ \frac{p_j(L+n)}{c_j} \right]^2. \] (5.101)

For \( n = -L/2 \), i.e. filtering at the center of the interval, Table 5.6a shows the \( w_r \)'s written in row-vector form:

\[ \begin{bmatrix} w_0 \left( -\frac{L}{2} \right), \ w_1 \left( -\frac{L}{2} \right), \ldots, \ w_{L-1} \left( -\frac{L}{2} \right), \ w_L \left( -\frac{L}{2} \right) \end{bmatrix} \]

for zero- and first-order polynomials. The variance reduction factors, \( \sigma_{-L/2}^2 / \sigma^2 \), are also shown.
Table 5.6a. Coefficients $w_r(-L/2)$ and Variance Reduction Factors (VRF) for Fixed-Memory Polynomial Filters of Order $M = 0$ and 1.

<table>
<thead>
<tr>
<th>L</th>
<th>$w_r\left(-\frac{L}{2}\right)$, $r = 0, 1, ..., L$</th>
<th>VRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$</td>
<td>$\frac{1}{(1.732)^2}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}$</td>
<td>$\frac{1}{(2.236)^2}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}$</td>
<td>$\frac{1}{(2.646)^2}$</td>
</tr>
<tr>
<td>8</td>
<td>$\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}$</td>
<td>$\frac{1}{(3.000)^2}$</td>
</tr>
<tr>
<td>10</td>
<td>$\frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}$</td>
<td>$\frac{1}{(3.317)^2}$</td>
</tr>
</tbody>
</table>

As could have been predicted for such trivial polynomials, the $w_r$'s are simply the weights of a standard arithmetic average over $L+1$ points.

Table 5.6b is the analog of Table 5.6a for polynomials of degree two or three.
Table 5.6b. Coefficients \( w_r(\pm L/2) \) and Variance Reduction Factors (VRF) for Fixed-Memory Polynomial Filters of Order \( M = 2 \) and 3.

<table>
<thead>
<tr>
<th>( L )</th>
<th>( w_r(\pm L/2) )</th>
<th>VRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0, 1, 0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{3}{35}, \frac{12}{35}, \frac{17}{35}, \frac{12}{35}, \frac{3}{35} )</td>
<td>( \frac{1}{(1.435)^2} )</td>
</tr>
<tr>
<td>6</td>
<td>( \frac{2}{21}, \frac{3}{21}, \frac{6}{21}, \frac{7}{21}, \frac{6}{21}, \frac{3}{21}, \frac{2}{21} )</td>
<td>( \frac{1}{(1.732)^2} )</td>
</tr>
<tr>
<td>8</td>
<td>( \frac{21}{231}, \frac{14}{231}, \frac{39}{231}, \frac{54}{231}, \frac{59}{231}, \frac{54}{231}, \frac{39}{231}, \frac{14}{231}, \frac{21}{231} )</td>
<td>( \frac{1}{(1.979)^2} )</td>
</tr>
<tr>
<td>10</td>
<td>( \frac{36}{429}, \frac{9}{429}, \frac{44}{429}, \frac{69}{429}, \frac{84}{429}, \frac{89}{429}, \frac{84}{429}, \frac{69}{429}, \frac{44}{429}, \frac{9}{429}, \frac{36}{429} )</td>
<td>( \frac{1}{(2.196)^2} )</td>
</tr>
</tbody>
</table>

Observe from these tables that, for a given interval-length, \( L \), the VRF increases as the order \( M \) of the polynomial increases.

The VRF, \( \sigma_{-L/2}^2 / \sigma^2 \), is plotted as a function of \( L \) in Figure 5.59a for \( M = 0, 1, 2, \) and 3. For comparisons, the VRFs, \( \sigma_0^2 / \sigma^2 \), obtained when filtering is performed at the end of the interval (\( n = 0 \)), i.e. without time-delay, are plotted in Figure 5.59b.
Figure 5.59a. Variance Reduction Factor for Fixed-Memory Polynomial Filters of Various Degrees $M$ Versus Length $L$ of the Filtering Interval (Estimation at the Center of the Interval).
Figure 5.59b. Variance Reduction Factors for Fixed-Memory Polynomial Filters of Various Degrees Versus Length L of Filtering Interval (Estimation at the End of the Interval).

Figures 5.59a and 5.59b confirm an intuitive result: For a given polynomial and a fixed interval, a smaller VRF is obtained by filtering at the center of the interval \([t_L, t_0]\), rather than at the end.

The general expression for \(\sigma_n^2\) (Appendix A2) indicates that \(\sigma_n^2\) is a polynomial in \(n\) of degree \(2M\) with all its zeroes inside the interval \([-L, 0]\). Therefore, \(\sigma_n^2\) increases for \(n\) outside that interval. The VRF is minimum at the center of the filtering interval. It can also be shown that filtering at the center of the interval is preferable from a systematic-error reduction standpoint \([M8]\). The penalty for such retrodicting is a time-delay equal to \(L\tau / 2\).
Table 5.7 summarizes qualitatively the consequences on \( y_n \) of an increase in the filter parameters, L and M, and in the sampling rate, \( 1 / \tau \). This table is a good illustration of the trade-offs involved in the design of a filter.

<table>
<thead>
<tr>
<th></th>
<th>Random Errors (VRF)</th>
<th>Systematic Errors</th>
<th>Lag ((L \tau / 2))</th>
<th>Computing Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
</tr>
<tr>
<td>( L )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
</tr>
<tr>
<td>( 1/\tau )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
</tr>
<tr>
<td>( 2f_{\text{max}} )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>

80 identical fixed-memory polynomial filters were coupled with the \( J = 80 \) detectors of the PWR model of Chapter 4 (Figures 4.2a,b). The tests of the fitted-synthesis method performed in Section 5.3.2 were repeated. The zero-mean Gaussian noise of Section 5.3.1 was added to all signals.

The methodology of the tests is still that of Figure 4.1 with both the "Noise" and "Filter" boxes in operation. Note that there are now two successive least-squares-fit processes in the method: The first one is a fit of a polynomial function of time to noisy measurements for each detector; the
second one is a fit of a linear combination of functions of space and energy to the \( J \) filtered detector-readouts.

The sampling period, \( \tau = 0.25 \) s, and filter parameters (\( L \) and \( M \)) are the same for all filters.

It is assumed that the design specifications for the filters are:

(i) \( VRF = 0.25 = 1 / (2.0)^2 \);

(ii) \( \text{Lag} \leq 1 \) s.

It follows from Item ii that \( L \tau / 2 \leq 1 \) s, and therefore \( L \leq 8 \).

On the basis of these requirements, of Figures 5.59a,b, and of Tables 5.6b and 5.7, the values \( L = 8 \), \( M = 2 \), and \( n = - L / 2 \) are chosen. The result is a set of nine-point, second-order, fixed-memory polynomial filters with centered estimation. Their variance reduction factor is:

\[
VRF = \frac{59}{231} \approx 0.2554 \approx \frac{1}{(1.979)^2}, \tag{5.102}
\]

and the lag in the output is

\[
\text{Lag} = \frac{L \tau}{2} = 4 \tau = 1.0 \text{ s}.
\]

Figure 5.60 shows the filtered readout of the detector located in Node \# 893. This Figure should be compared with Figure 5.2 for the corresponding unfiltered signal.
Figure 5.60. Noise-Free and Filtered Noisy Outputs of the Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval).

Figure 5.61 shows the relative residual, $\|E_{LS}(t)\| / \|C(t)\|$, obtained with the three-mode synthesis formula (smallest singular value zeroed) fitted to the 80 filtered signals. The corresponding plots obtained with noise-free signals on the one hand, and with noisy, unfiltered signals on the other hand are also shown.
Figure 5.61. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).

As expected, the mean value of \( \| \mathbf{E}_{LS}(t) \| / \| \mathbf{C}(t) \| \) is now smaller than when no filter is used.

In the formula for the expected value of the square relative residual (Eq. 5.67), the noise variance must now be multiplied by the VRF of the filters, 1 / (1.979)^2:

\[
\mathbb{E} \left\{ \left[ \frac{\| \mathbf{E}_{LS}(t) \|}{\| \mathbf{C}(t) \|} \right]^2 \right\} \approx \frac{80 - 3}{80} \left(0.05\right)^2 \frac{1}{(1.979)^2} \approx (0.0248)^2.
\]
Therefore, the expected value of the relative residual should be approximately:

\[
E \left\{ \frac{\|E_{ls}(t)\|}{\|C(t)\|} \right\} \approx 0.0248,
\] (5.103)

instead of 0.0491 without filters (Eq. 5.72). This theoretical estimate of 0.0248 is in agreement with the value obtained from Figure 5.61: 0.0276 (variance = \(9.61 \times 10^{-6}\)).

The effect of the filters on the four modal coefficients is shown in Figures 5.62 through 5.65. (The fourth modal coefficient does not enter the three-mode synthesis.) These graphs should be compared with their unfiltered analogs (Figures 5.5 through 5.8).
Figure 5.62. First Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval).
Figure 5.63. Second Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval).
Figure 5.64. Third Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval).
Figure 5.65. Fourth Modal Coefficient Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval).

Reference and reconstructed group-two fluxes in Nodes # 893, 651, and 649 appear in Figures 5.66, 5.67, and 5.68 respectively. Figures 5.16, 5.17, and 5.18 are the corresponding unfiltered results.
Figure 5.66. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).
Figure 5.67. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).
Figure 5.68. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).

The performance of the filters can be best appreciated by comparing Figure 5.66 with Figures 5.2 and 5.60. All three graphs correspond to Node # 893.

The variance of $\hat{\phi}_{2n}(t)$ in node n is again given by the R-over-J rule (Eq. 5.51) on the average, but with the noise variance, $\sigma^2$, reduced by the filter VRF, i.e.:
\[
\text{var} \{ \hat{\phi}_n(t) \} = \frac{3}{80} \frac{1}{(1.979)^2} \sigma^2 = 0.00958 \sigma^2 \approx \frac{1}{(10.2)^2} \sigma^2. \quad \text{(5.104)}
\]

The value without filters was \( 1/(5.2)^2 \) (Eq. 5.73).

In perturbed nodes such as Node # 649, residual random errors are still visible when perturbations occur (Figure 5.68). These residual errors attest to the difficulty of recovering detailed spatial information from noisy signals, even with the help of filters.

The impact of signal filtering on total power and reactivity can be judged from Figures 5.69 and 5.70. (Compare with the results of Figures 5.23 and 5.24 obtained without filters.)
Figure 5.69. Reference and Reconstructed Total Power Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).
The eighty, nine-point, fixed-memory filters do not increase the overall computing-cost of the synthesis appreciably in spite of the push-down tables needed to store the L+1 sets of most recent measurements. For the 720 time-steps, the total CPU time with filters is 89 s as opposed to 81 s without filters.

In fact, the only real drawback of the fixed-memory filters is the lag of $4 \tau = 1$ s that they introduce. In an attempt to eliminate this time-delay while preserving the computational efficiency, recursive filters giving 0-step estimates were designed.
b. Fading-Memory IIR Filters.

The direct method of filter design was applied to another type of filter: a *fading-memory polynomial IIR filter* [M8]. The construction relies again on the approximation of the noise-free signal by a smooth polynomial profile. The notation is that of Figure 5.71. For this filter, the $k$-axis shown is more convenient than the $r$-axis of Figure 5.58.

![Diagram](image)

**Figure 5.71.** Notational Conventions for Fading-Memory Polynomial Filter.

The fitting interval is no longer finite: it is now $]-\infty, 0]$. (In practice, the width of the interval starts from zero and increases each time a new measurement is received.) The new least-squares criterion is:
Minimize \[ \sum_{k=0}^{+\infty} [x_{-k} - q(k)]^2 \theta^k \] (5.104)

where \( q \) is the polynomial to be determined, and the \( \theta^k \)'s are weighting factors giving decreasing importance to old measurements in the fit. \( \theta \) is a real number restricted to the interval \([0, 1]\) to guarantee filter stability. The order of the polynomial \( q \) is \( M \). \( M \) and \( \theta \) are free parameters to be adjusted.

As with the fixed-memory polynomial filter, there exists an elegant method for solving the \( \infty \)-by-\( M+1 \) least-squares problem without resorting to any matrix inversion. It is to write \( q(k) \) as a linear combination of the first \( M+1 \) discrete Laguerre polynomials, \( q_j(k), k = 0,1,2,...,M \), and to make use of the orthogonality relationship between these polynomials:

\[
\sum_{k=0}^{\infty} \frac{q_i(k)}{d_i} \frac{q_j(k)}{d_j} \theta^k = \delta_{ij}, \quad \begin{cases} i = 0,1,...,M, \\ j = 0,1,...,M, \end{cases} (5.105)
\]

with

\[
d_j^2 = \sum_{k=0}^{\infty} [q_j(k)]^2 = \frac{\theta^j}{1 - \theta}, \quad j = 0,1,...,M. (5.106)
\]

Formulas for the \( q_j(k) \)'s are given in Appendix A2.

The output, \( y_n \), of this fading-memory filter is

\[
y_n = q(k)|_{k=n} = \sum_{k=0}^{+\infty} w_k(n) x_{-k} \quad (5.107)
\]

where
\[ w_k(n) = \sum_{j=0}^{M} \frac{q_j(n)}{d_j} \frac{q_j(k)}{d_j}, \quad k = 0,1,\ldots,\infty. \] (5.108)

The filter coefficients, \( w_k(n) \), depend on \( \theta \), \( M \), and on the validity instant, \( t_n \). They need be computed only once.

Comparison of the above formulas with those of the fixed-memory filter (Eqs. 5.96 through 5.100) shows that, in many respects, the role of \( \theta \) in the fading-memory filter is similar to that of \( L \) in the fixed-memory filter.

The validity instant is chosen as \( t_n = t_0 \). In these conditions, the filter operates as a 0-step estimator. Contrary to the fixed-memory filter, the output, \( y_0 \), of this fading-memory filter can be expressed in a compact, recursive form which makes the computations more efficient. There is no need to keep track of all the measurements \( x_{-k} \) between \( -k = -\infty \) and \( -k = 0 \). Indeed, \( y_0 \) can be written in terms of (i) the estimate, \( y_{-1} \), obtained at the previous sampling-time, \( t_{-1} \), and (ii) an error estimate, \( e_0 = x_0 - z_0 \), where \( z_0 \) is a 1-step prediction for \( y_0 \) obtained at \( t_{-1} \). (See Appendix A2).

The drawback of such a recursive formulation is that it gives rise to transient errors, as explained in Section 5.4.2.a. Figures 5.72a,b,c show the response of fading-memory filters of order \( M = 1, 2, \) and 3 respectively to a step decrease in input signal from 1.0 to 0.1 at \( t = 0 \).
Figure 5.72a. Time Response of a Fading-Memory Polynomial Filter of Order One to a Step Change in Input Signal for Different Weighting-Factors Theta.
Figure 5.72b. Time Response of a Fading-Memory Polynomial Filter of Order Two to a Step Change in Input Signal for Different Weighting-Factors Theta.
Figure 5.72c. Time Response of a Fading-Memory Polynomial Filter of Order Three to a Step Change in Input Signal for Different Weighting-Factors Theta.

These figures show that transient errors
(i) increase with the order $M$ of the polynomial for a given value of $\theta$;
(ii) increase with $\theta$ for a given polynomial.

A general formula for the variance of $y_0$, $\sigma_0^2$, can be found in Appendix A2. Figure 5.73 shows plots of the variance reduction factor, $\sigma_0^2 / \sigma^2$, as a function of the weighting factor $\theta$ for $M = 1, 2, 3, \text{and } 4$. 

390
Figure 5.73. Variance Reduction Factors for Fading-Memory Polynomial Filters of Various Degrees $M$ Versus Weighting Factor Theta (0-Step Estimation).

It can be seen from this figure that a small VRF is achieved only for fairly high values of $\theta$, i.e. in regions where, unfortunately, transient errors tend to be large. Also, for a given value of $\theta$, a low-order polynomial gives better performance.

Figure 5.73 should be compared with Figure 5.59b which shows analogous curves obtained for fixed-memory polynomial filters with $n = 0$ (filtering at the end of the interval). From these two figures and from the corresponding formulas given in Appendix A2, it is possible to define, for each fading-memory filter characterized by $M$ and $\theta$, an “equivalent”, fictitious, fixed-memory polynomial filter. The order of this equivalent
filter is M, while its characteristic interval has length, \( L_{eq} \), obtained by equating the variance reduction factors of the two filters. Approximate (asymptotic) formulas for \( L_{eq} \) are given in Table 5.8.

Table 5.8. Length \( L_{eq} \) of the Filtering Interval of a Fictitious Fixed-Memory Polynomial Filter Equivalent to a Fading-Memory Polynomial Filter of Identical Order and Weighting Factor \( \theta \) (\( \theta \geq 0.8 \)).

<table>
<thead>
<tr>
<th>( L_{eq} )</th>
<th>( M = 0 )</th>
<th>( M = 1 )</th>
<th>( M = 2 )</th>
<th>( M = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1 - \theta )</td>
<td>( 2.00 )</td>
<td>( 3.20 )</td>
<td>( 4.36 )</td>
<td>( 5.51 )</td>
</tr>
</tbody>
</table>

The equivalence just described is convenient to compare the two types of filters. For example, a second-order fading-memory filter characterized by \( \theta = 0.8 \) has an equivalent interval length, \( L_{eq} \), approximately equal to \( 4.36 \) / \( (1 - 0.8) \) = 22 sampling periods. (A more accurate calculation gives \( L_{eq} = 18 \).)

This equivalence can also be applied to the problem of initializing the recursive fading-memory filter. Details about a “near-optimal” initialization procedure can be found in Reference M8 and will not be described here.

Table 5.9 is the analog of Table 5.7. It shows the impact of variations in \( M \), \( \theta \), and \( 1 / \tau \) on the performance of a fading-memory filter.
Table 5.9. Influence of the Order of the Polynomial, M, Weighting Factor, θ, and Sampling Rate, 1/τ, on the Performance of a Fading-Memory Polynomial Filter.

<table>
<thead>
<tr>
<th></th>
<th>Random Errors (VRF)</th>
<th>Systematic Errors</th>
<th>Transient Errors</th>
<th>Computing Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>▶️   ⤓</td>
<td>▶️</td>
<td>▶️</td>
<td></td>
</tr>
<tr>
<td>θ</td>
<td>▶️   ⤓</td>
<td>▶️</td>
<td>▶️</td>
<td>-</td>
</tr>
<tr>
<td>1/τ</td>
<td>▶️</td>
<td>▶️</td>
<td>▶️</td>
<td>▶️</td>
</tr>
<tr>
<td>2f_{max}</td>
<td>▶️</td>
<td>▶️</td>
<td>▶️</td>
<td>▶️</td>
</tr>
</tbody>
</table>

As with the fixed-memory filter, compromise values must be found for M and θ. In practice, θ is generally less than 0.9 while M seldom exceeds 2.

The eighty, second-order, fixed-memory filters previously paired with the detectors of the PWR model are replaced by eighty, second-order, fading-memory filters.

In order to achieve the specified VRF of 0.25 as before, a value of θ = 0.8703 would be necessary (Figure 5.73). However, it is feared that such a high value could lead to excessive transient errors. Therefore, the smaller value of θ = 0.8000 is chosen instead.

All filters are “tuned” with the same two parameters: M = 2 and θ = 0.8. The corresponding variance reduction factor is:
VRF = 0.3822 ≈ \frac{1}{(1.618)^2} \quad (5.109)

This value is larger than the targeted one, but, contrary to the fixed-memory filter, it is obtained without any time delay.

The output of the detector-filter pair associated with Node # 893 is shown in Figure 5.75. This figure should be compared with Figures 5.2 and 5.60.

Figure 5.75. Noise-Free and Filtered Noisy Outputs of the Detector Located in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8).
The same three-mode synthesis calculation was repeated with the 80 noisy signals preprocessed through the fading-memory polynomial filters. The resulting relative residual, $\frac{\|E_{LS}(t)\|}{\|C(t)\|}$, appears in Figure 5.76.

![Figure 5.76](image)

Figure 5.76. Relative Residual Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed).

The expected value for the square relative residual can be estimated:

$$
\mathbb{E} \left( \frac{\|E_{LS}(t)\|}{\|C(t)\|} \right)^2 \approx (0.0491)^2 \frac{1}{(1.618)^2} \approx (0.0303)^2.
$$
Therefore, roughly,

\[ \mathbb{E} \left\{ \frac{\| E_{LS}(t) \|}{\| C(t) \|} \right\} = 0.0303. \] (5.110)

This is in excellent agreement with the experimental value of 0.0298 (variance = 1.10 \times 10^{-5}) obtained from Figure 5.76.

Note the anomalously-large values which occur during the filter initialization phase.

Figures 5.77 through 5.79 contain plots of the reconstructed group-two flux in Nodes # 893, 651, 649.
Figure 5.77. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed)
Figure 5.78. Reference and Reconstructed Group-Two Flux in Node # 651 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed)
Figure 5.79. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Fading-Memory Filters of Order Two, Weighting Factor = 0.8, One Singular Value Zeroed)

When the above graphs are compared with those obtained with fixed-memory filters (Figures 5.66 through 5.68), only minor differences are visible, except during the “learning-phase” of the filters. If the VRFs were not known, it would be difficult to assess the relative performance of the two types of filters on the basis of the plots only.

Total CPU time with the fading-memory filters was 85 s. This cost would have been substantially larger had fixed-memory filters with $M = 2$, $L = 18$, and $n = 0$ been used to achieve the same VRF of 0.38 without time-delay.
The next section explains how these results, albeit already very good, can be further improved.

b. Modified Fixed-Memory FIR Filters.

No use has been made thus far of the concept of response function in the design of the filters. In general, one should always verify the adequacy of a particular filter designed by the direct method by computing its response function, \( H(\omega) \). This is done below for the fixed-memory polynomial FIR filter of Section 5.4.3.a with estimation at the center of the filtering interval \([-L, 0]\).

The output, \( y_{L/2} \), of such a filter is given by Eq. 5.99:

\[
y_{L/2} = \sum_{r=0}^{L} w_r \left( -\frac{L}{2} \right) x_{L+r}
\]  
(5.111)

It is possible to reconcile this formula with the result of Eq. 5.82,

\[
y_{L/2} = \sum_{k=-\frac{L}{2}}^{\frac{L}{2}} h_k x_{\frac{L}{2} - k}
\]  
(5.112)

by introducing the subscript correspondence: \( k = L/2 - r \), i.e.:

\[
h_k = w_{k-r} \left( -\frac{L}{2} \right), \quad k = 0,1,2,...,L.
\]  
(5.113)

The weights \( h_k \) defined by Eq. 5.113 possess the symmetry property:

\[ h_k = h_{-k} \]  

because the \( w_r \left( -\frac{L}{2} \right) \)'s verify that
\[ w_r \left( -\frac{L_c}{2} \right) = w_{L - r} \left( -\frac{L_c}{2} \right), \quad r = 0, 1, 2, ..., L, \quad (5.114) \]

an identity which is itself a consequence of a symmetry property of the discrete Legendre polynomials:

\[ p_j(r) = (-1)^j p_j(L - r), \quad j = 0, 1, 2, ..., M. \quad (5.115) \]

As a result, the response function of the filter is real and can be obtained directly from Eq. 5.94 (without the \( \sigma_k \)'s):

\[
H(\omega) = h_0 + 2 \sum_{k=1}^{L_c/2} h_k \cos(k\omega)
\]

\[
= w_{L_c/2} + 2 \sum_{r=L_c/2+1}^{L} w_r \left( -\frac{L_c}{2} \right) \cos \left[ \left( r - \frac{L_c}{2} \right) \omega \right]. \quad (5.116)
\]

Substitution of the \( w_r \left( -\frac{L_c}{2} \right) \)'s of Table 5.6b for \( M = 2 \) and \( 3 \) gives the expressions for \( H(\omega) \) listed in Table 5.10a.
Table 5.10a. Transfer Functions of Second- and Third-Order Fixed-Memory Polynomial Filters.

<table>
<thead>
<tr>
<th>L</th>
<th>Transfer Function H(ω)</th>
<th>VRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cos(ω)</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>(\frac{17 + 24 \cos(ω) - 6 \cos(2ω)}{35})</td>
<td>(\frac{1}{(1.435)^2})</td>
</tr>
<tr>
<td>6</td>
<td>(\frac{7 + 12 \cos(ω) + 6 \cos(2ω) - 4 \cos(3ω)}{21})</td>
<td>(\frac{1}{(1.732)^2})</td>
</tr>
<tr>
<td>8</td>
<td>(\frac{59 + 108 \cos(ω) + 78 \cos(2ω) + 28 \cos(3ω) - 42 \cos(4ω)}{231})</td>
<td>(\frac{1}{(1.979)^2})</td>
</tr>
<tr>
<td>10</td>
<td>(\frac{89 + 168 \cos(ω) + 138 \cos(2ω) + 88 \cos(3ω) + 18 \cos(4ω) - 72 \cos(5ω)}{429})</td>
<td>(\frac{1}{(2.196)^2})</td>
</tr>
</tbody>
</table>

Transfer functions corresponding to L = 4, 6, 8, and 10 have been plotted in Figure 5.80a.
Figure 5.80a. Response Function of Second-Order Fixed-Memory Polynomial Filter Versus Normalized Frequency for Different Values of the Filtering-Interval Length, L.

It can be seen that these Fourier spectra differ appreciably from the shape of the typical low-pass filter (Figure 5.57). Most notable is the presence of oscillations or "side lobes", and the nonzero values for $\omega = \pi$.

These observations suggest that the $w_r\left(-\frac{L}{2}\right)$'s used in Eq. 5.111 are far from "optimal" weights, and that simply changing L (or M) will not improve $H(\omega)$.
The oscillations are caused by the highest-order term in the formula for $H(\omega)$, i.e. by the term

$$w_0 \left( \frac{L}{2} \right) \cos \left[ \left( \frac{L}{2} \right) \omega \right] = w_L \left( \frac{L}{2} \right) \cos \left[ \left( \frac{L}{2} \right) \omega \right].$$

A simple change which can be expected to reduce the "tails" of $H(\omega)$ consists in modifying the coefficient $w_0 \left( \frac{L}{2} \right) = w_L \left( \frac{L}{2} \right)$ of this high-order term in such a way that $H(\pi) = 0$. When this is done, and after the algebraic sum of the coefficients has been renormalized to unity (so that $H(0) = 1$), the modified transfer-functions of Table 5.10b and Figure 5.80b are obtained.

Table 5.10b. Transfer Functions of Second- and Third-Order, Modified, Fixed-Memory Polynomial Filters.

<table>
<thead>
<tr>
<th>$L$</th>
<th>Transfer Function $H(\omega)$</th>
<th>VRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$\cos(\omega)$</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{17 + 24 \cos(\omega) + 7 \cos(2\omega)}{48}$</td>
<td>$\frac{1}{(1.957)^2}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{7 + 12 \cos(\omega) + 6 \cos(2\omega) + \cos(3\omega)}{26}$</td>
<td>$\frac{1}{(2.201)^2}$</td>
</tr>
<tr>
<td>8</td>
<td>$\frac{59 + 108 \cos(\omega) + 78 \cos(2\omega) + 28 \cos(3\omega) - \cos(4\omega)}{272}$</td>
<td>$\frac{1}{(2.409)^2}$</td>
</tr>
<tr>
<td>10</td>
<td>$\frac{89 + 168\cos(\omega) + 138\cos(2\omega) + 88\cos(3\omega) + 18\cos(4\omega) - 11\cos(5\omega)}{490}$</td>
<td>$\frac{1}{(2.595)^2}$</td>
</tr>
</tbody>
</table>
Figure 5.80b. Response Function of Second-Order, Modified, Fixed-Memory, Polynomial Filters Versus Normalized Frequency for Different Values of the Filtering-Interval Length, L.

Observe that the profiles are now in better agreement with the typical low-pass shape, except for the absence of a sharp cutoff-frequency. As expected, the tails have been reduced, at the expense of a slightly larger "main lobe" (distance from $\omega = 0$ to the first zero of $H(\omega)$).

Comparisons between Tables 5.10a and 5.10b reveal that the VRFs have been reduced; therefore, the modified filters should be more efficient. In particular, for $L = 8$, the VRF is now $1 / (2.409)^2 = 0.1723$ instead of $1 / (1.979)^2 = 0.2554$ prior to the changes.
In the least-squares fit to the measurements, the effect of changing the coefficients \( w_0 \left( \frac{-L}{2} \right) \) and \( w_L \left( \frac{-L}{2} \right) \) is to reduce the relative importance of the end points, \( x_{-L} \) and \( x_0 \), so that they do not influence the fit as much as the inner points (for \( L \geq 6 \)). The original, standard least-squares fit is in fact replaced by a *weighted* least-squares fit.

These modified fixed-memory filters were applied to the 80 detectors signals used as inputs to the synthesis code, and the same calculations were carried out one more time. The parameters \( M = 2 \) and \( L = 8 \) were chosen for all filters.

Figures 5.81 and 5.82 show the synthesized group-two flux in Nodes # 893 and 649 respectively. The beneficial effect of the modification in the filters can be appreciated by comparing these figures with Figures 5.66 and 5.68.
Figure 5.81. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Operational Transient, 80 In-Core Detectors With Modified Fixed-Memory Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).
The variance of $\hat{\phi}_{2n}(t)$ is now

$$\text{var} \left( \hat{\phi}_{2n}(t) \right) = \frac{3}{80} \frac{1}{(2.409)^2} \sigma^2 = 0.00646 \sigma^2 \approx \frac{1}{(12.4)^2} \sigma^2.$$  (5.117)

The quality of these results attests to the power of the indirect method of filter design.
As a final illustration, a plot of the reconstructed reactivity is shown in Figure 5.83. Figure 5.70 is the analog of this graph before filter modification.

Figure 5.83. Reference and Reconstructed Reactivity Versus Time (Operational Transient, 80 In-Core Noisy Detectors With Modified Fixed-Memory Polynomial Filters of Order Two, 9-Point Interval, One Singular Value Zeroed).

In summary, simple FIR and IIR filtering algorithms were designed in the time domain by least-squares fitting of polynomial profiles to the measurements. These filters were "tuned" to achieve a specified variance reduction factor. Caution was exercised in adjusting the filter parameters
so as not to generate an unreasonably large time-delay, large systematic errors, or large transient errors.

Numerical tests confirmed that these filters, although not very sophisticated, were nonetheless capable of improving the signals supplied to the fitted-synthesis method. The result was a reduction of random errors in all reconstructed quantities by the desired variance reduction factor.

A slight modification of the FIR filters motivated by transfer-function considerations in frequency space lead to even better agreement with reference, noise-free results.

5.5. Summary.

In this chapter, the effect of measurement noise on the discontinuous fitted-synthesis method was investigated. A statistical analysis was performed to confirm quantitatively a qualitative prediction made in the previous chapters: The presence of high-order modes in the synthesis causes extreme sensitivity of the results to noise in the detector readouts. The analysis also revealed that, provided the J neutron-detectors are reasonably distributed throughout the core, a reduction of the noise variance by an average factor of $R/J$ can be expected in the nodal group-fluxes reconstructed by an R-mode synthesis, except possibly in the regions where, and at the time when, perturbations occur.

Numerical simulations were performed with the same reactor model and instrumentation as in Chapter 4. The same operational transient was studied with artificial, zero-mean noise added to all detector outputs. The conditions of the tests were made deliberately difficult by imposing a nonstationary noise of fairly high intensity. Despite these difficulties, the statistical predictions were confirmed. The best results were obtained with the set of 80 in-core detectors. The use of only four in-core sensors, or of a side instrumentation, lead to larger random errors.

One natural way of decreasing the sensitivity of the results to noise was to eliminate high-order terms from the synthesis. This simple
procedure was limited, however, since discarding too many modes lead to unacceptable systematic errors.

Another way of reducing the impact of measurement noise on the results of the fitted synthesis was to filter the detectors signals prior to their use by the synthesis program. Simple, low-pass, digital filtering algorithms were designed for that purpose.

Satisfactory performance was achieved through a combination of these two methods of noise reduction: Filtering of all detector signals, and zeroing of the highest-order mode(s) in the synthesis. The reduction in random errors agreed with the predictions.

Since both methods introduce systematic errors in the reconstructed nodal group-fluxes, it is important that they be used with restraint. Some residual amount of noise must be accepted. Signal filtering, although not as simple as truncating the synthesis formula, has the distinctive advantage of offering more flexibility. In some cases, it may even allow the synthesis method to operate with more modes. In addition, if used with moderation, it may not introduce any significant systematic errors.

In Chapter 6, the discontinuous fitted-synthesis method is applied to the analysis of two different transients: an inlet coolant-temperature transient, and a very severe, startup transient.

Chapter 7 contains conclusions and recommendations for future research.
Chapter 6

SECOND AND THIRD APPLICATIONS OF THE METHOD

6.1. Introduction.

In Chapter 2, a time-synthesis method for reconstructing three-dimensional, instantaneous, nodal flux-distributions in light-water reactors was described. The essential idea of this method is a least-squares fit of a linear combination of K, precalculated, 3D, static expansion-functions to J local flux measurements. The J-by-K least-squares fit problem is solved by the technique of singular value decomposition described in detail in Chapter 3. This orthogonalization technique rearranges the fitted combination of basis functions into an equivalent, more convenient combination of R (≤ K ≤ J) orthogonal modes of decreasing magnitude.

In Chapter 4, this fitted-synthesis method was applied to a severe operational transient in a PWR model. Numerical tests showed that the method was capable of reconstructing nodal group-fluxes with negligibly small errors in most nodes. The results were further improved by restricting the number of basis functions used at any instant in the synthesis to a small subset of carefully selected flux-shapes. The effect of this modification was to make the synthesis discontinuous in time. In this discontinuous mode of operation, and in the absence of measurement noise, the number and positions of the neutron detectors were found to have a relatively small impact on the quality of the results.

The effect of measurement noise on the method was studied in Chapter 5. The analysis showed that the least-squares fit of the R-mode synthesis formula to the J detector outputs reduces the noise variance by an average factor of R / J. However, the presence of high-order modes in the synthesis was found to cause random-error amplification in perturbed
nodes. This amplifying effect was not as pronounced when many in-core detectors \( J \gg R \) distributed throughout the core were used. In that case, the method was still able to yield satisfactory results when the smallest modes were zeroed, and when the detector noisy signals were replaced by filtered estimates.

In this chapter, the fitted-synthesis method is applied to two different reactor transients: an accident-type, coolant inlet-temperature transient, and a reactor-startup transient. The PWR model and instrumentation are the same as those described in Section 4.3. The 80 in-core detectors of Figures 4.2a and 4.2b are used in the simulations. The testing methodology is the same as in Chapters 4 and 5 (See Section 4.2). Reference solutions, expansion functions, and detector signals are all provided by the QUANDRY code.

The objectives of these complementary tests are (i) to verify that the conclusions drawn in Chapter 4 and 5 are not restricted to the operational transient studied in those chapters; (ii) to provide additional illustrations of the shape-selection and singular-value-zeroing procedures which are crucial to the success of the fitted-synthesis method; and (iii) to identify possible physical conditions in which the synthesis approximation may break down.

6.2. Application to a Coolant Inlet-Temperature Transient.

6.2.1. Description of the Transient.

The reactor is initially in steady-state, critical conditions at nominal power, \( P_{\text{nominal}} = 3338 \text{ MWe} \). This critical configuration is achieved by dividing all fission cross sections by \( k_{\text{eff}} = 1.047818 \). All rod-banks are fully withdrawn except rod-bank D which is partially inserted (as in Figure 4.3 for \( t \geq 120 \text{ s} \)). The total coolant flow-rate, \( \dot{m} \), is constant, equal to \( 3.87 \times 10^3 \text{ kg s}^{-1} \).
A transient is initiated thermally by a two-second exponential
decrease in the coolant inlet-temperature, \( \overline{T}_{\text{inlet}}^{(c)}(t) \), from 555 K to 535 K,
followed by an exponential increase to 555 K. The exact form of the
perturbation is

\[
\overline{T}_{\text{inlet}}^{(c)}(t) = \overline{T}_{\text{inlet}}^{(c)}(0) \left[ \exp \left( -\frac{t}{\tau_1} \right) + 1 - \exp \left( -\frac{t}{\tau_2} \right) \right]
\]  

(6.1)

where \( \overline{T}_{\text{inlet}}^{(c)}(0) = 555 \text{K}, \tau_1 = 2.000 \text{ s}, \) and \( \tau_2 = 2.206 \text{ s}. \) The transient is
followed until \( t = 10 \text{ s}. \)

Figure 6.1 is a plot of \( \overline{T}_{\text{inlet}}^{(c)}(t) \) as a function of time. Also shown in
the same figure is the QUANDRY-computed core-averaged fuel-
temperature, \( \overline{T}_{\text{core}}^{(f)}(t) \).
The decrease in coolant temperature improves neutron moderation, which in turn leads to a higher fission-rate, and therefore an increase in fuel temperature. (Values for the cross section temperature-feedback coefficients can be found in Appendix A1). Note the delay of approximately two seconds between the two temperature peaks.

The numerical parameters used to generate the QUANDRY reference solution appear in Table 6.1.
Table 6.1. Numerical Parameters Used in the QUANDRY Reference Calculation (Coolant Inlet-Temperature Transient).

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Number of Time-Steps</th>
<th>Time-Step Size (s)</th>
<th>$\varepsilon_\phi$ in Eq. 4.2</th>
<th>Number of Time-Steps per Matrix Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 10 s]</td>
<td>400</td>
<td>0.025</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
</tbody>
</table>

The minimum number of inner iterations per time-step was set to two.

With these parameters, the total CPU time for the calculation was 3 hrs 50 min. (1 hr 25 min with the average convergence criterion of Eq. 4.1.)

Reference plots of total fission power, $P(t)$, and reactivity, $\rho(t)$, appear in Figure 6.2. The weight functions used in the calculation of reactivity are adjoint fluxes and net leakages corresponding to the initial, critical condition of the reactor.
Figure 6.2. QUANDRY-Computed Reactivity and Total Fission Power Versus Time (Coolant Inlet-Temperature Transient).

Note that the reactivity peak is higher than in the case of the operational transient of Chapter 4 (Figure 4.4a). Therefore, from a reactivity standpoint, this transient is more severe than the operational transient, even though localized perturbations in flux shape such as those caused by control-rod motions are not expected here.

Figure 6.3 shows the evolution with time of the coolant and fuel temperature in Node # 893, located in Plane # 12 (See Figure 4.11).
Figure 6.3. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (Coolant Inlet-Temperature Transient).

The maximum change in coolant temperature experienced by Node # 893 is 13 K as opposed to 20 K for nodes in the bottom plane.

6.2.2. Synthesis Calculations.

The reactor nuclear instrumentation consists of the 80 in-core flux-sensors of Figures 4.2a and 4.2b. As before, the characteristic weight-
functions, $\Sigma^{(j)}$, are such that the detector output, $C^{(j)}(t)$, is equal to the group-two flux in the node containing detector $j$ (Eq. 4.3).

A sampling-period of $\tau = 0.025$ s was chosen. The reason for this fairly small value is to obtain a sufficiently dense distribution of points in the tests with noise (presented in Section c below).

a. Expansion Functions.

Nine expansion-functions were generated by criticality calculations corresponding to constant inlet temperatures and constant flow rate. Calculations were performed for three different power-levels: 100 % $P_{\text{nominal}}$, 120 % $P_{\text{nominal}}$, 140 % $P_{\text{nominal}}$, and three different values of coolant inlet-temperature: 535 K, 545 K, and 555 K. The static module of QUANDRY was used for these calculations. The nine eigenvalues returned by the code are listed in Table 6.2.

Table 6.2. Eigenvalues $k_{\text{eff}}$ for Three Different Power-Levels and Three Different Values of Coolant Inlet-Temperature.

<table>
<thead>
<tr>
<th></th>
<th>$P = 100$ %</th>
<th>$P = 120$ %</th>
<th>$P = 140$ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{T}^{(c)}_{\text{inlet}} = 535$ K</td>
<td>1.05086</td>
<td>1.04860</td>
<td>1.04635</td>
</tr>
<tr>
<td>$\bar{T}^{(c)}_{\text{inlet}} = 545$ K</td>
<td>1.04934</td>
<td>1.04708</td>
<td>1.04484</td>
</tr>
<tr>
<td>$\bar{T}^{(c)}_{\text{inlet}} = 555$ K</td>
<td>1.04782</td>
<td>1.04557</td>
<td>1.04334</td>
</tr>
</tbody>
</table>

The main challenge presented by this transient is its fast dynamics. Identifying a reduced subset of expansion functions in real time appears quite difficult unless accurate measurements of coolant inlet-temperature
can be obtained in a fraction of a second. For that reason, contrary to the operational transient, no attempt is made to identify instantaneous subsets of "appropriate" shapes. All nine expansion functions, \( \Psi^{(k)} \), are included in the synthesis for the duration of the transient (continuous synthesis). As explained in Section 4.6.1, this amounts to "lumping" without discrimination all the information provided by the nine basis-functions into a single matrix \( \Delta \). Only one rectangular matrix system, \( \Delta T(t) = C(t) \), needs to be solved, but this indiscriminate amalgam causes the singular-value clustering problem and, in general, does not permit a synthesis as accurate as with a sequence of smaller matrices formed from only a few expansion-functions. This unfavorable situation is accepted here.

Table 6.3 shows the normalized singular-value spectrum of the single 80-by-9 matrix \( \Delta \) used throughout the transient.

Table 6.3. Normalized Singular Value Spectrum of the 80-by-9 Matrix \( \Delta \).

<table>
<thead>
<tr>
<th></th>
<th>( s_i / s_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000 ( 10^0 )</td>
</tr>
<tr>
<td>2</td>
<td>1.614 ( 10^{-2} )</td>
</tr>
<tr>
<td>3</td>
<td>4.768 ( 10^{-4} )</td>
</tr>
<tr>
<td>4</td>
<td>2.449 ( 10^{-4} )</td>
</tr>
<tr>
<td>5</td>
<td>1.020 ( 10^{-4} )</td>
</tr>
<tr>
<td>6</td>
<td>7.015 ( 10^{-6} )</td>
</tr>
<tr>
<td>7</td>
<td>4.881 ( 10^{-7} )</td>
</tr>
<tr>
<td>8</td>
<td>3.824 ( 10^{-7} )</td>
</tr>
<tr>
<td>9</td>
<td>1.345 ( 10^{-7} )</td>
</tr>
</tbody>
</table>
As expected, the condition number of $\mathbf{A}$, $\kappa(\mathbf{A}) = s_1 / s_9 = 1 / 1.345 \times 10^{-7}$, is much larger than unity: $\mathbf{A}$ is very ill-conditioned. For consistency with the limited, relative, numerical accuracy of approximately $5 \times 10^{-5}$ in the $\psi^{(k)}$'s, and to avoid numerical difficulties (see Section 3.3.3.d), a zeroing threshold equal to $\varepsilon = 5 \times 10^{-5}$ is imposed. Accordingly, the four smallest singular values of $\mathbf{A}$ are zeroed. The resulting, reduced-rank matrix has rank $R = 5$ and condition number $s_1 / s_5 = 1 / 1.020 \times 10^{-4}$. The synthesis formula (Eq. 5.69) contains only five modes.


Least-squares fitting of the five-mode synthesis formula to the $J = 80$ noise-free detector readouts gives the relative residual, $\|E_{LS}(t)\| / \|C(t)\| = \|C(t) - \mathbf{A}T_{LS}(t)\| / \|C(t)\|$, shown in Figure 6.4. The Euclidean norm of the minimum-norm, least-squares solution vector of $\mathbf{A}T(t) = C(t)$, $\|T_{LS}(t)\|$, is also shown in that figure.
Figure 6.4. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).

Observe that the relative residual remains very small with respect to unity ($\leq 0.0016$), in spite of a peak during the first second of the transient.

No component of $T_{LS}(t)$ has abnormally large magnitude since $\|T_{LS}(t)\|$ remains less than 6 at all times. This confirms that no numerical difficulty is to be feared.

Evidence of the success of the synthesis is given by Figures 6.5a and 6.5b which show the node-averaged and maximum relative errors, $e_g^{(ave)}(t)$
and $e^{(\text{max})}(t)$ (Eqs. 4.4 and 4.5), in reconstructed group-one and group-two fluxes respectively.

Figure 6.5a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).
Figure 6.5b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core detectors, Four Singular Values Zeroed).

Node-averaged errors are always less than 0.25 %, while the maximum local error over all nodes and all time-steps is only 1.27 % in group one (Node # 1429, \( t = 0.6 \) s), and 1.25 % in group two (Node # 1429, \( t = 0.6 \) s).

Reference and reconstructed group-two flux in Nodes # 893, 649, 495, and 244 are plotted in Figures 6.6 through 6.9. Recall that Node # 893 is an instrumented node, and Node # 495 a reflector node.
Figure 6.6. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).
Figure 6.7. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).
Figure 6.8. Reference and Reconstructed Group-Two Flux in Node # 495 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).
Figure 6.9. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Four Singular Values Zeroed).

Error peaks appear in the interval [0 s, 2 s]. However, their magnitude remains very small. The agreement between reference and synthesized fluxes is excellent throughout the transient.

Total CPU time for this 400-step, five-mode synthesis calculation is 71 s. This is more than the 10 seconds of the transient, but considerably less than the 3 hrs 50 min required by QUANDRY. Real-time synthesis calculations could be achieved with a sampling-period of 0.2 s or larger.
c. Test With Noisy Signals.

As explained in Chapter 5, noise in the detector readouts may be a concern if very small modes (i.e. singular values) are present in the synthesis. Random errors may be amplified in some regions. One way of limiting this effect is to eliminate small modes from the synthesis by zeroing the corresponding singular values.

Zeroing small modes must be done with caution, however, especially when the modes correspond to local perturbations in flux shape. Eliminating such high-order perturbation-modes may lead to substantial systematic errors locally. In the present transient, however, one does not expect any local distortion in flux-shape. Instead, a global axial “tilt” is expected. Therefore, the synthesis formula should not contain very high-order terms. (This can be verified by comparing the $\psi_{n}$ s node by node and group by group.) In such conditions, zeroing small singular values should not cause large errors in a few nodes, but rather small errors in many nodes. As a consequence, random errors affecting the coefficients $T^{(i)}_{L3}(t)$ of the smallest modes in the synthesis formula (Eq. 3.69) should have a small, “distributed” rather than a large, “concentrated” impact on the reconstructed flux-shapes. The results presented below confirm these expectations.

In an attempt to limit the sensitivity of the synthesis to noise, two additional singular values, $s_{4}$ and $s_{5}$, are zeroed. Table 6.3 shows that the resulting rank-three matrix has condition number $s_{1} / s_{3} = 1 / 4.768 \times 10^{-4}$.

It may be argued that $s_{3}$ could be zeroed too since $s_{2} / s_{3} \approx 34$ and $s_{3} / s_{4} = 2$. The conditioning of the resulting rank-two matrix would be considerably better. However, this small singular value and its associated information (the mode $\psi^{(3)}$) are retained for two reasons. The first reason is to demonstrate that, as predicted, small modes such as $\psi^{(3)}$ do not cause excessive, local sensitivity to noise. The other reason is to show that the
application of one of the filters developed in Chapter 5 can reduce random errors to the point that zeroing $s_3$ becomes unnecessary.

Figures 6.10, 6.11, and 6.12 confirm that, in the absence of noise, neglecting the small modes $\psi^{(4)}$ and $\psi^{(5)}$ does not lead to a significant increase in systematic errors.

![Graph showing Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time](image)

Figure 6.10. Euclidean Norm of Least-Squares Solution-Vector and Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed).
Figure 6.11a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed).
Figure 6.11b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Detectors, Six Singular Values Zeroed).

The relative residual remains much smaller than unity (maximum value = 0.0048), while maximum errors in local fluxes are only 2.1 % in group one (Node # 1307, t = 0.55 s), and 2.0 % in group two (Node # 1307, t = 0.55 s).

The same artificial, zero-mean, Gaussian noise as in Section 5.3.1 was added to all detector signals. The variance is again \((0.05)^2\) times the local flux measurement, \(C^{(j)}(t)\) (nonstationary noise). Figure 6.12 shows the output signal of the detector located in Node # 893.
Figure 6.12. Noise-Free and Noisy Outputs of the Detector Located in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors).

Observe that the magnitude of the fluctuations is relatively high with respect to the variations of the noise-free signal itself. In fact, the noise has essentially the same variance as in the last part \((t \geq 120 \text{ s})\) of the operational transient (Figure 5.2).

When the 80 noisy (unfiltered) signals are supplied to the fitted-synthesis program, the relative residual of Figure 6.13 is obtained.
Figure 6.13. Relative Residual Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed).

Eq. 5.67 may be used to obtain an estimate of the expected value of the square of this residual. Since R, J, and the noise characteristics are the same as in Section 5.3, the result is identical to that of Eq. 5.72:

\[
E \left\{ \frac{||E_{LS}(t)||}{||C(t)||} \right\} \approx 0.0491.
\]  

(5.72)

The value computed from the 401 points of Figure 6.13 is 0.0483 (variance = 3.40 \times 10^{-5}), in excellent agreement with the above theoretical estimate.
The reconstructed group-two flux in Node # 893 appears in Figure 6.15, together with the QUANDRY reference solution.

![Graph showing flux versus time with data points and line representing QUANDRY and Synthesis]

Figure 6.14. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed).

Comparison of this figure with Figure 6.12 permits verifying the R-over-J rule which predicts an average variance-reduction factor of $(1 / 5.164)^2$.

A detailed analysis reveals that no anomalously large errors occur in any reconstructed flux value, $\hat{\phi}_{gn}(t)$, during this transient. In fact, Figure
6.14 is typical of the group-two flux in any node. This is in contrast with the operational transient of Chapter 4 for which relatively large errors were detected in isolated nodes (clusters of three and four points). Those nodes were nodes in which perturbations (rod motions) occurred. These observations confirm that, in this coolant inlet-temperature transient, no very high-order mode is generated, which greatly facilitates the task of the synthesis method.

Plots of total fission power, \( \hat{P}(t) \), and reactivity, \( \hat{\rho}(t) \), reconstructed from the \( \hat{\phi}_{\text{gn}}(t) \)'s according to Eqs. 2.17 and 2.25, are compared with reference curves in Figures 6.15 and 6.16.
Figure 6.15. Reference and Reconstructed Total Power Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed).
Residual random errors in these plots may be judged too high. If so, it is possible to reduce them further by zeroing another singular value ($s_3$). Alternatively, and to avoid introducing additional systematic errors, the detector readouts can be filtered, as explained in Section 5.4.

If the simple, second-order, nine-point, fixed-memory polynomial filters of Section 5.4.3.a are applied to the 80 detector signals, the variance of any measurement is reduced by the factor $\sigma_{L/2}^2 / \sigma^2 = 1 / (1.979)^2$ (See Table 5.6b), as the comparison of Figures 6.17 and 6.16 illustrates for $\hat{\phi}_{2893}(t)$. 

Figure 6.16. Reference and Reconstructed Reactivity Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors, Six Singular Values Zeroed).
Figure 6.17. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed).

The price to pay for the improvement is a time-delay of $4 \tau = 0.1 \text{ s}$ in the outputs of the synthesis program.

The benefit of signal filtering on power and reactivity can be appreciated from the comparison of Figures 6.18 and 6.19 below with Figures 6.15 and 6.16.
Figure 6.18. Reference and Reconstructed Total Power Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed).
Figure 6.19. Reference and Reconstructed Reactivity Versus Time (Coolant Inlet-Temperature Transient, 80 In-Core Noisy Detectors With Fixed-Memory Filters of Order Two, Six Singular Values Zeroed).

6.2.3. **Conclusion.**

The results of this test suggest that thermally-induced transients should not present any major difficulty to the fitted-synthesis method. The reason for this is that such transients do not generate very high-order perturbation modes. In such conditions, the fitted-synthesis method gives good results, even when it operates in a less efficient, time-continuous mode, and even if a significant amount of noise is present in the measurements.
6.3. Motivations for the Startup Transient.

In view of the quality of the results obtained thus far in all the numerical simulations, it is natural to wonder whether there exist conditions in which the fitted-synthesis method may break down.

It has already been pointed out in Chapter 2 that errors in the flux-to-measurement operator, Σ^T, i.e. in the detector weight functions, Σ^{(j)}, would cause difficulties (systematic errors). The unnoticed presence of even a single, “wild” measurement, C^{(j)}(t), would certainly be detrimental to the least-squares-fit process. An accurate evaluation of the Σ^{(j)}'s (a difficult task), and a good signal-validation algorithm are essential prerequisites to a successful application of the synthesis method.

Another situation where the method may experience difficulties is when perturbations of multiple origins occur simultaneously, forcing the shape-selection algorithm described in Section 4.6.3 to include many expansion-functions in the synthesis. It was shown in Chapter 4 that best results are obtained when only a small subset of “presorted” basis-functions is used at any time. Retaining many functions leads to larger errors while giving the deceiving impression of a better synthesis because of a smaller least-squares-fit residual.

A third situation in which the synthesis method may not perform satisfactorily is when the instantaneous, physical state of the reactor is “far away” from any equilibrium state. This happens when strong dynamic effects occur. There are at least two regimes of operation in which such effects may arise. The first one is the prompt-critical regime, i.e. during very rapid transients (time scale on the order of 10^{-5} s). In this high-frequency regime, the right-hand side of Eq. 1.2 contributes significantly to the overall neutron-balance equation. However, operation in this regime is prohibited in commercial plants.

The second regime of strong dynamic effects is the low-power, subcritical-to-critical range when both neutrons from external sources and
multiplication neutrons contribute appreciably to the instantaneous neutron-balance equation. It can be seen from Eq. 2.25 for instance that no equilibrium state exists when reactivity is positive and when the external-source term is significant.

The availability of an external-source option in the QUANDRY code makes it possible to study this interesting second regime. With this option, it is possible, for example, starting from low-power, subcritical, steady-state conditions, to simulate a rapid approach to criticality. The transient states of the reactor should be difficult to approach with a linear combination of neutron flux shapes obtained from steady-state configurations. The startup transient of the next section is an attempt to create such transient conditions which should constitute a serious challenge to the synthesis method. Admittedly, some of the assumptions made in this startup simulation are not realistic. However, the objective of the test is not to simulate a realistic reactor-startup, but rather to create physical conditions under which the fitted-synthesis method may break down.

6.4. Application to a Reactor-Startup Transient.

6.4.1. Description of the Transient.

The same PWR model as in the previous tests is used in this simulation. The reactor is initially in a "zero-power" equilibrium state. Rod-banks C and D are in their "fully-inserted" positions as shown in Figures 4.2b. The coolant inlet-temperature and total mass flow-rate are $T_{\text{inlet}}^{(c)} = 555 \text{ K}$ and $\dot{m} = 3.87 \times 10^3 \text{ kg s}^{-1}$ respectively. They remain constant throughout the transient.

A source-free criticality-calculation for this initial state yields an eigenvalue of $k_{\text{eff}} = 0.953177$. 

443
Small, primary startup-sources are present in Plane # 5, in Nodes # 330 and 370 (Figure 4.11). Both sources are identical. They are made of californium, a material which emits neutrons spontaneously.

Nodes # 330 and 370 are located in the ‘X-Z’ and ‘Y-Z’ planes of Figures 4.2a,b, and are symmetrically positioned across the ‘X = Y’ diagonal. This choice preserves octant symmetry, which makes the generation of the reference solution by QUANDRY less time-consuming. However, it also implies that the entire core contains four primary startup sources. Actual Westinghouse PWRs contain only two such sources located approximately at the elevation of Plane # 5, but not exactly in Nodes # 330 and 370 [W8]. This difference is accepted here.

The node-averaged intensity of each startup source is taken as $Q_1 = 3.7275 \times 10^7$ neutrons cm$^{-3}$ s$^{-1}$, and $Q_2 = 0$ (no thermal neutrons). With such a spectrum, QUANDRY gives a corewide fission power of 333.8 W, i.e. $10^{-7} P_{\text{nominal}}$.

The startup procedure consists of five distinct phases:

(i) A ‘boron-dilution’ phase, from $t = 0$ s to $t = 40$ s;
(ii) A first ‘rod-withdrawal’ phase, from $t = 40$ s to $t = 60$ s;
(iii) A first ‘no-action’ phase, between $t = 60$ s and $t = 200$ s;
(iv) A second ‘rod-withdrawal’ phase, from $t = 200$ s to $t = 300$ s;
(v) A second ‘no-action’ phase between $t = 300$ s and $t = 360$ s.

Figure 6.20 is a schematic illustration of these five successive phases.
The boron-dilution phase is simulated by a linear decrease in the group-two neutron-absorption cross-section by $\Delta \Sigma_{a2} = -0.020$ cm$^{-1}$ in all reflector nodes, and by $\Delta \Sigma_{a2} = -0.010$ cm$^{-1}$ in all fuel nodes. For convenience, the boric-acid concentration in the core is denoted $c_B = 100\%$ at $t = 0\ s$, and $c_B = 0\%$ at $t = 40\ s$.

Following this (unrealistically rapid) boron-dilution phase, rod-banks C and D are withdrawn by 40 cm at the constant speed of $v_{rod} = 2\ cm\ s^{-1}$. All rod motions stop temporarily at $t = 60\ s$, and resume at $t = 200\ s$, at the same speed of 2 cm s$^{-1}$. At $t = 300\ s$, rod-banks C and D reach their final positions (corresponding to $t = 120\ s$ in Figure 4.3). The values for the various cross-section perturbations have been chosen in such a way that, in this final rod-configuration, the reactor is exactly critical. The (net) reactivity is zero when $P = 100\% \ P_{nominal}$.

Note that the last 160 seconds of this transient are very similar to the last 160 seconds of the operational transient of Chapter 4.

The transient module of the QUANDRY code was used to generate a reference solution to this problem. The numerical parameters used in this calculation are presented in Table 6.4.
Table 6.4. Numerical Parameters Used in the QUANDRY Reference Calculation (Startup Transient).

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Number of Time-Steps</th>
<th>Time-Step Size (s)</th>
<th>$\varepsilon_\Phi$ in Eq. 4.2</th>
<th>Number of Time-Steps per Matrix Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 40 s]</td>
<td>160</td>
<td>0.25</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[40 s, 60 s]</td>
<td>160</td>
<td>0.125</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[60 s, 200 s]</td>
<td>560</td>
<td>0.25</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[200 s, 300 s]</td>
<td>800</td>
<td>0.125</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>[300 s, 360 s]</td>
<td>240</td>
<td>0.25</td>
<td>$10^{-3}$</td>
<td>1</td>
</tr>
</tbody>
</table>

A minimum of two inner-iterations per time-step was imposed.

With these values, the total CPU time was 25 hrs 50 min (but only 7 hrs 25 min when the average convergence criterion of Eq. 4.1 was used.)

Figures 6.21a,b show the evolution with time of total fission power, $P(t)$, and reactivity, $\rho(t)$. Figure 6.21a contains only the first 180 seconds of the transient. The last 180 seconds appear in Figure 6.21b. The same subdivision of the time axis will be used in subsequent illustrations. Notice the logarithmic scale for power.
Figure 6.21a. QUANDRY-Computed Reactivity and Total Power Versus Time (First Half of Startup Transient).
Figure 6.21b. QUANDRY-Computed Reactivity and Total Power Versus Time (Second Half of Startup Transient),

The weight functions used to evaluate reactivity and other point-kinetics parameters are adjoint fluxes and net leakages corresponding to the final, critical state of the reactor. With such a choice:

\[ \rho_0 = -0.04564 = 7021.8 \text{ millibeta}; \]
\[ n_{eff_0} = 1.1920 \times 10^1 \text{ neutrons cm}^{-3}; \]
\[ q_{eff_0} = 2.4414 \times 10^4 \text{ neutrons cm}^{-3} \text{ s}^{-1}; \]
\[ \Lambda_0 = 2.2286 \times 10^{-5} \text{ s}. \]

If, instead, adjoint quantities corresponding to the initial, artificially-critical, source-free state (eigenvalue \( k_{eff} = 0.953177 \)) were chosen, \( \rho_0 \) would be equal to \(-0.04912 = 1 - 1/k_{eff} \). (Provided the eigenvalue-
dependent coupling matrices in QUANDRY are computed with an
eigenvalue set to one.)

Figure 6.21a shows that $\rho(t)$ increases very rapidly during the
boron-dilution phase. (Again, such a rapid dilution of boric acid is
unrealistic in practice.) $\rho(t)$ becomes positive at $t = 40 \text{ s}$, and is equal to
394 millibeta at $t = 60 \text{ s}$, when the end of the first rod-motion phase is
reached. At that time, total fission power is only $P = 0.00145 \% P_{\text{nominal}}$, and there are no feedback effects. For $t \geq 60 \text{ s}$, $P(t)$ increases at a nearly
constant logarithmic rate. In the absence of any inherent feedback
mechanism, the power excursion would continue in an unbounded fashion.
Approximately 150 s into the transient ($P \approx 5 \% P_{\text{nominal}}$), thermal-
hydraulic feedback effects become significant, as the inflection in the
power profile indicates. Reactivity starts to decrease. At this point,
external-source contributions are completely negligible.

If control-rod motion did not resume at $t = 200 \text{ s}$, $P(t)$ would
eventually level off at some value $P_1 (< P_{\text{nominal}})$ such that negative
reactivity-feedback effects compensate for the positive external-reactivity
insertion. There are in fact many stable equilibrium points in this high-
power range [G17, A1, H10]. In normal conditions, the reactor operates at
its design equilibrium-point which is such that $\rho = 0$, $P = 100 \% P_{\text{nominal}}$, and the power distribution is sufficiently “flat” (See Chapter 1).

The most interesting portion of this transient is the interval
[0 s, 80 s] in which $\rho(t)$ is positive, but external-source effects are still
significant. The neutron-flux distribution (in space and energy) evolves
from a distribution peaked in the region of the startup sources to a more
homogeneous distribution imposed by multiplication neutrons. Between
these two extreme distributions, a variety of intermediate, “dynamic” flux-
shapes are encountered. Some of these instantaneous shapes cannot be well
approximated by any static shape. Such conditions can be expected to be
very challenging for the synthesis method.

Figure 6.23 is a plot of the fuel and coolant temperatures in Node #
893 as a function of time.
Figure 6.22a. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (First Half of Startup Transient).
Figure 6.22b. QUANDRY-Computed Fuel and Coolant Temperatures in Node # 893 Versus Time (Second Half of Startup Transient).

It can be seen from Figure 6.22a that there is no significant temperature-change before $t = 120$ s. On the other hand, large variations occur for $t \geq 200$ s. As for the operational transient, these large variations make the validity of the linear feedback-model highly questionable. However, possible modeling errors are not a concern here since, as explained in Chapter 4, they do not affect the consistency of the tests.
6.4.2. Synthesis Calculations.

In principle, application of the fitted-synthesis method to the analysis of the above startup transient requires the availability of wide-range neutron-flux sensors, or, alternatively, of several sets of sensors having overlapping ranges of operation (see Section 2.5) to cover the seven decades of power variations. For simplicity, it is assumed that the 80 detectors of Figures 4.2a and 4.2b are wide-range sensors capable of covering these seven decades. (Some sophisticated neutron-detectors are capable of such performance [B10].) In addition, possible difficulties which may arise from the very low, initial flux-levels are ignored. The detector outputs, $C^{(i)}(t)$, are assumed to be noise-free and, as before, exactly equal to the group-two fluxes in the nodes containing the sensors.

a. Expansion Functions.

28 expansion-functions are generated for the analysis of this startup transient by the synthesis method:

(i) Three flux-shapes obtained from steady-state calculations with the startup sources present for three different values of boron concentration: $c_B = 100 \%$, $c_B = 50 \%$, and $c_B = 0 \%$;

(ii) Three flux-shapes obtained from steady-state criticality calculations at $P = 0$ and $c_B = 0 \%$ for the three control-rod patterns at $t = 0 \text{s (or 40 s), 50 s, and 60 s}$;

(iii) Twenty-two flux-shapes obtained from steady-state criticality calculations at $c_B = 0 \%$. These shapes are computed for two different power-levels: $P = 15 \% P_{\text{nominal}}$ and $P = 100 \% P_{\text{nominal}}$, and for the eleven control-rod patterns at $t = 200 \text{s, 210 s, 220 s, ..., 290 s, and 300 s}$.

For application of the discontinuous synthesis method, the transient is divided into 14 time-intervals, as shown in Table 6.5.
Table 6.5. Time-Intervals and Corresponding Expansion-Functions for the Discontinuous Synthesis (Startup Transient).

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Total Number of Expansion Functions</th>
<th>Boron Concentration $c_B$</th>
<th>Control-Rod Pattern</th>
<th>External Source Present? If No, Specified Power-Level in Criticality Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 40 s]</td>
<td>3</td>
<td>100%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td>[40 s, 50 s]</td>
<td>5</td>
<td>100%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 0 s</td>
<td>No, P = 0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 50 s</td>
<td>No, P = 0%</td>
</tr>
<tr>
<td>[50 s, 60 s]</td>
<td>5</td>
<td>100%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 0 s</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 50 s</td>
<td>No, P = 0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 60 s</td>
<td>No, P = 0%</td>
</tr>
<tr>
<td>[60 s, 200 s]</td>
<td>2</td>
<td>0%</td>
<td>t = 60 s</td>
<td>No, P = 0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 60 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td>[200 s, 210 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 200 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 200 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 210 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 210 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[210 s, 220 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 210 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 210 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 220 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 220 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[220 s, 230 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 220 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 220 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 230 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 230 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[230 s, 240 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 230 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 230 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 240 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 240 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[240 s, 250 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 240 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 240 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 250 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 250 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[250 s, 260 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 250 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 250 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 260 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 260 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>[260 s, 270 s]</td>
<td>4</td>
<td>0%</td>
<td>t = 260 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 260 s</td>
<td>No, P = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 270 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0%</td>
<td>t = 270 s</td>
<td>No, P = 100%</td>
</tr>
<tr>
<td>Time-interval</td>
<td>$%$</td>
<td>$t$-value</td>
<td>Note</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>----------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>[270 s, 280 s]</td>
<td>4</td>
<td>0 %</td>
<td>$t = 270$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 270$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 280$ s</td>
<td>No, $P = 100$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 280$ s</td>
<td>No, $P = 100$ %</td>
</tr>
<tr>
<td>[280 s, 290 s]</td>
<td>4</td>
<td>0 %</td>
<td>$t = 280$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 280$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 290$ s</td>
<td>No, $P = 100$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 290$ s</td>
<td>No, $P = 100$ %</td>
</tr>
<tr>
<td>[290 s, 360 s]</td>
<td>4</td>
<td>0 %</td>
<td>$t = 290$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 290$ s</td>
<td>No, $P = 15$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 300$ s</td>
<td>No, $P = 100$ %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>$t = 300$ s</td>
<td>No, $P = 100$ %</td>
</tr>
</tbody>
</table>

Each time-interval is characterized by a subset of at most five expansion-functions.

Since measuring the boron concentration takes on the order of minutes in actual PWRs [B3, M9, J2], it is not reasonable to subdivide the boron-dilution phase into several time-intervals and to use flux-shapes corresponding to selected values of $c_B$ in each of these time-intervals.

In contrast, information about the positions of the control rods is available in real time. It is therefore legitimate to break the two rod-motion phases into intervals in which only the nearest flux-shapes are used, as was done for the operational transient in Section 4.6.

Note that the three flux-shapes obtained from steady-state calculations with external sources are included not only in the first time-interval, but also in the second and third ones. The reason for this is that one can expect external-source effects to persist in those time-intervals. Retaining only flux shapes obtained from criticality calculations would probably not be sufficient to permit a good synthesis. (This was confirmed by numerical tests.) An automatic shape-selection algorithm may have difficulty to recognize such "persisting effects".
b. Results.

Application of the synthesis program produces the normalized singular value spectra of Table 6.6 for the 14 successive 80-by-K matrices $\Delta$.

**Table 6.6. Singular Value Spectra of the Fourteen 80-by-K Matrices $\Delta$ of the Discontinuous Synthesis.**

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>K</th>
<th>$s_i / s_1, i = 1,2,\ldots,R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 s, 40 s]</td>
<td>3</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.090 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$9.568 \times 10^{-5}$</td>
</tr>
<tr>
<td>[40 s, 50 s]</td>
<td>5</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.090 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$9.678 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$3.261 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2.214 \times 10^{-7}$</td>
</tr>
<tr>
<td>[50 s, 60 s]</td>
<td>5</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.090 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.045 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$5.327 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$9.458 \times 10^{-6}$</td>
</tr>
<tr>
<td>[60 s, 200 s]</td>
<td>2</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.396 \times 10^{-8}$</td>
</tr>
<tr>
<td>[200 s, 210 s]</td>
<td>4</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.816 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.166 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2.401 \times 10^{-4}$</td>
</tr>
<tr>
<td>[210 s, 220 s]</td>
<td>4</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2.734 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.027 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2.617 \times 10^{-4}$</td>
</tr>
<tr>
<td>[220 s, 230 s]</td>
<td>4</td>
<td>$1.000 \times 10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.988 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$8.091 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.537 \times 10^{-4}$</td>
</tr>
<tr>
<td>Interval</td>
<td>Value</td>
<td>Value</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>[230 s, 240 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.601 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.321 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.919 $10^{-4}$</td>
</tr>
<tr>
<td>[240 s, 250 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.818 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.369 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.901 $10^{-4}$</td>
</tr>
<tr>
<td>[250 s, 260 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.437 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.500 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.266 $10^{-4}$</td>
</tr>
<tr>
<td>[260 s, 270 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.023 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.385 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.860 $10^{-4}$</td>
</tr>
<tr>
<td>[270 s, 280 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.571 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.368 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.077 $10^{-4}$</td>
</tr>
<tr>
<td>[280 s, 290 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.186 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.294 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.644 $10^{-4}$</td>
</tr>
<tr>
<td>[290 s, 360 s]</td>
<td>4</td>
<td>1.000 $10^0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.024 $10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.664 $10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.081 $10^{-4}$</td>
</tr>
</tbody>
</table>

A linear-dependence tolerance of $\varepsilon = 3 \times 10^{-5}$ is imposed for consistency with the limited numerical accuracy in the expansion functions. The smallest singular values in the second, third, and fourth time-intervals are therefore zeroed.

Figures 6.23a and 6.23b show the relative residual returned by the synthesis program. Also shown is the Euclidean norm of the minimum-norm, least-squares solution vector, $\|T_{LS}(t)\|$. 

456
Figure 6.23a. Euclidean Norm of Least-Squares Solution Vector and Relative Residual Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.23b. Euclidean Norm of Least-Squares Solution Vector and Relative Residual Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).

Observe the peaks in $\|T_{LS}(t)\|$ between $t = 40$ s and $t = 60$ s. This implies that some of the components of $T_{LS}(t)$ have fairly large magnitude, a characteristic of the solution of ill-conditioned systems of equations. Table 6.6 confirms that $A$ is indeed quite ill-conditioned in the second and third time-intervals, despite the zeroing of $s_5$. However, this ill-conditioning has no detrimental consequence on the synthesis in the absence of noise in the $C^{(i)}$'s.
The relative residual does not exceed 0.021, a remarkably low value. The most "difficult" interval for the least-squares fit appears to be between \( t = 30 \text{ s} \) and \( t = 180 \text{ s} \).

Node-averaged and maximum relative errors in reconstructed group-fluxes are presented in Figures 6.24a,b and 6.25a,b.

Figure 6.24a. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.24b. Node-Averaged and Maximum Relative Errors in Group-One Flux Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.25a. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.25b. Node-Averaged and Maximum Relative Errors in Group-Two Flux Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).

The node-averaged relative error, $\varepsilon_{g}^{(ave)}(t)$, is always less than 5 % in both groups, and is less than 0.6 % in the second half of the transient.

The magnitude of the maximum error, $\varepsilon_{g}^{(max)}(t)$, is 5 % or less in both groups for $t \geq 70$ s. As expected, larger errors occur between $t = 0$ s and $t = 70$ s, the "difficult" portion of the transient. Maximum values over all nodes and all time-steps are -47.5 % in group one (Node # 330, $t = 44$ s), and 39.3 % in group two (Node # 1060, $t = 50$ s).
Figures 6.26a,b and 6.27a,b show plots of reference and reconstructed group-fluxes in Node # 330 versus time. Figures 6.26a,b are for group one, while Figures 6.27a,b are for group two. A logarithmic scale is used for the fluxes, whereas a linear scale is used for the relative error.

Figure 6.26a. Reference and Reconstructed Group-One Flux in Node # 330 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.26b. Reference and Reconstructed Group-One Flux in Node # 330 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.27a. Reference and Reconstructed Group-Two Flux in Node # 330 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.27b. Reference and Reconstructed Group-Two Flux in Node # 330 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).

The presence of a startup source in Node # 330 explains the relatively large errors in that node and its immediate neighbors between $t = 30$ s and $t = 70$ s. However, regions near the external sources are not the only locations where fairly large errors occur. Figures 6.28a,b show that substantial discrepancies also affect nodes located in the upper half of the core, such as Node # 1060 located in Plane # 14.
Figure 6.28a. Reference and Reconstructed Group-Two Flux in Node # 1060 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
An sharp peak in the relative-error curve occurs in that node for $t$ in the interval [50 s, 60 s]. One should notice, however, that the absolute error remains small. In addition, discrepancies do not exceed a few percent in all the other intervals.

In fact, the discontinuous fitted-synthesis method performs surprisingly well in many nodes, as Figures 6.29a,b, 6.30a,b, and 6.31a,b illustrate for Nodes # 893, 649, and 244.
Figure 6.29a. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.29b. Reference and Reconstructed Group-Two Flux in Node # 893 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.30a. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.30b. Reference and Reconstructed Group-Two Flux in Node # 649 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.31a. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.31b. Reference and Reconstructed Group-Two Flux in Node # 244 Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).

As with Nodes # 330 and 1060, the largest errors occur between \( t = 30 \) s and \( t = 70 \) s.

As in the operational transient (Figure 4.33), the relative error in \( \hat{\phi}_{2649}(t) \) takes on larger values when the tip of rod-bank D moves across the node, i.e. between \( t = 230 \) s and \( t = 240 \) s.

Total CPU time for the 720 time-steps of this synthesis calculation is only 74 s.
Reconstructed amplitude-function, $\tilde{n}_{\text{eff}}(t)$, total power, $\tilde{P}(t)$, and reactivity, $\tilde{\rho}(t)$, are compared with the corresponding reference solutions in Figures 6.32a,b, 6.33a,b, and 6.34a,b respectively.

Figure 6.32a. Reference and Reconstructed Amplitude Function Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.32b. Reference and Reconstructed Amplitude Function Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.33a. Reference and Reconstructed Total Power Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.33b. Reference and Reconstructed Total Power Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).
Figure 6.34a. Reference and Reconstructed Reactivity Versus Time (First Half of Startup Transient, 80 In-Core Detectors).
Figure 6.34b. Reference and Reconstructed Reactivity Versus Time (Second Half of Startup Transient, 80 In-Core Detectors).

As in the operational transient, the slight variations in fission cross-sections are neglected in the evaluation of $\hat{P}(t)$. This is the reason for the growing discrepancy between reference and reconstructed power in Figure 6.33b.

In the computation of $\hat{\rho}(t)$ (Eq. 2.25), the value of the effective external source, $q_{\text{eff}}$, is assumed to be known. If, instead, the synthesis calculation is performed with $q_{\text{eff}}$ set to zero, the resulting reactivity profile is that of Figure 6.35.
Figure 6.35. Reference and Reconstructed Reactivity Versus Time With Effective External Source Neglected (Startup Transient, [40 s, 120 s] Interval, 80 In-Core Detectors).

\( \hat{\rho}(t) \) differs substantially from the reference curve for \( t < 60 \) s. However, little difference can be seen for \( t \geq 80 \) s, a sign that, beyond that point (which corresponds to \( P = 0.01 \) \% \( P_{\text{nominal}} \) in Figure 6.21a), external-source effects have become negligibly small.

6.4.3. Conclusion.
The results of this simulated startup-transient confirm that the discontinuous fitted-synthesis method leads to relatively large errors when applied to reactor dynamic conditions for which "nearby" static flux-shapes cannot be generated. However, such conditions do not exist in the power range during normal operation, and may not even be encountered in the low-power range during an actual reactor startup. It should also be pointed out that application of the method in the source range may not be possible because of the limited nuclear instrumentation available in that range.

6.5. Summary.

In this chapter, the fitted-synthesis method was applied to two severe PWR transients: a coolant inlet-temperature transient, and a startup transient.

Excellent results were obtained for the first of these two transients, even though the synthesis method was restricted to a nonoptimal, continuous mode of operation. The presence of noise in the detector readouts did not cause serious difficulties because of the absence of very high-order modes in the synthesis.

The startup-transient simulation was intended to challenge the synthesis method by exhibiting strong dynamic effects in neutron flux-shape in the low-power range. The simulation achieved its objective in the sense that fairly large errors in reconstructed group-fluxes were observed locally. However, a number of unrealistic assumptions had to be made to create the extreme physical conditions in which these errors occurred. In an actual reactor-startup, such conditions would not be realized. Application of the synthesis method to reactor startups may nevertheless be precluded by the lack of a sufficient instrumentation in the low-power range.

The next chapter contains a summary of this investigation, conclusions, and recommendations for future research.
Chapter 7

SUMMARY AND CONCLUSIONS


In Chapter 1, the main challenges and objectives of light-water reactor core analyses were presented. The importance of the on-line (real-time) determination of the instantaneous neutron-flux distribution in space (3-D) and energy (two groups) to the optimal performance of the plant was stressed. Reasons were given for the difficulty of obtaining this information in real time with sufficient accuracy.

Various techniques available for solving the time-dependent, group, neutron diffusion equations were briefly described. Special consideration was given to the most efficient of these methods: synthesis methods, and nodal methods.

Recent developments in the field of on-line core-power monitoring and control were reviewed. It was noted that most of these developments take the form of very complicated computer-programs. Many of these programs are constructed around nodal kinetics codes which, although very fast-running, are still unable to operate in real time (time scale ~ 1 s) on desktop machines.

The need for simpler, more efficient yet sufficiently accurate methods for evaluating the neutron-flux distribution was identified. The objective of the present investigation was stated as the development of one such method.

In Chapter 2, the general characteristics of a semi-experimental, nodal synthesis method for the real-time reconstruction of the three-dimensional neutron-flux distribution were given. A distinctive
characteristic of this method is that it avoids the solution of the time-dependent, neutron diffusion equations. The essential idea of the method is to approximate instantaneous, nodal, group fluxes by a linear combination of \( K \), three-dimensional, static, nodal flux-shapes. These static flux-shapes or expansion functions are generated off-line, either from steady-state, criticality calculations (fundamental \( \lambda \)-modes), or from steady-state, nonhomogeneous calculations with external sources. The time-dependent coefficients, \( T^{(k)}(t) \), of the linear combination are determined from the requirement that the reconstructed flux-distribution agree in a least-squares sense with local flux-measurements, \( C^{(i)}(t) \), provided by \( J \) neutron detectors. The least-squares fit problem was written in vector form as
\[
\mathbf{A} \mathbf{T}(t) = \mathbf{C}(t)
\]
in which \( \mathbf{A} \) is a \( J \)-by-\( K \) matrix.

Possible difficulties with this fitted-synthesis method were identified. Of particular concern were

(i) Amplification of numerical-roundoff errors in the solution of the least-squares-fit problem because this amplification may lead to meaningless results;

(ii) Systematic errors caused by the synthesis approximation because there is no general prescription other than physical insight for selecting the number of basis functions and the reactor conditions for which they are generated;

(iii) Random errors caused by measurement noise because of the unavoidable character of such errors.

The "inverse kinetics" procedure for estimating reactivity from the synthesized flux-distribution was described.

Lastly, a review of neutron-flux detectors currently in use in light-water reactors was given.

In Chapter 3, the general properties of rectangular systems of linear algebraic equations, \( \mathbf{A} \mathbf{T} = \mathbf{C} \), were reviewed. Central to the presentation was the interpretation of the \( J \)-by-\( K \) matrix \( \mathbf{A} \) as a linear map from \( \mathbb{R}^K \) onto \( \mathbb{R}^J \).

Three least-squares-theory methods for solving the rectangular system \( \mathbf{A} \mathbf{T} = \mathbf{C} \) were compared: the Normal Equations, QR Factorization, and Singular Value Decomposition. In spite of its higher
cost, the SVD method was chosen because of the wealth of information that it provides about the $A$ matrix, and because of its unmatched ability to detect ill-conditioning in $A$, an unfortunate but very likely characteristic of the matrix of the fitted-synthesis equations.

The SVD process was described geometrically and algebraically in considerable detail. At the heart of SVD is a complete orthogonalization procedure in which orthonormal bases are constructed for (i) the row space of $A$ and its orthogonal complement in $R^K$ on the one hand, and (ii) the column space of $A$ and its orthogonal complement in $R^J$ on the other hand. The result is a "diagonalized" map, i.e. a one-to-one correspondence between basis vectors of the row space of $A$, and basis vectors of its column space. Both subspaces have dimension $R = \text{rank}(A)$. The image of each basis-vector of the row space of $A$ is in the direction of one of the basis vectors of the column space of $A$, and its length is equal to one of the singular values of $A$. The minimum-norm, least-squares solution-vector of $A \mathbf{T} = C$, $T_{LS}$, is easily found from these orthogonal bases and from the simple vector-to-vector correspondence. In addition, a measure of the degree of ill-conditioning in $A$ is available from the ratio of the largest to the smallest of its singular values. Possible roundoff-error amplification by ill-conditioning can be prevented by zeroing exceedingly small singular values.

Least-squares theory and SVD were applied specifically to the system of equations of the fitted-synthesis method. Overdetermination, i.e. more detectors than expansion functions, was found to be highly desirable for reliable performance. Under the assumption of unbiased, noise-free detector-readings, the relative residual of the least-squares-fit problem was shown to be a measure of the systematic errors caused by the synthesis approximation. It was also demonstrated that solving $A \mathbf{T}(t) = C(t)$ by the SVD technique amounts to replacing the original, linear combination of $K$ expansion functions by an equivalent, linear combination of $R$ ($\leq K$) new basis functions or "modes". These modes are "felt" by the instrumentation as orthogonal vectors of $R^J$. The magnitude of these vectors are the singular values of $A$. Zeroing small singular values was therefore interpreted as eliminating high-order modes from the synthesis.
In Chapter 4, numerical simulations were performed to evaluate the accuracy and computing efficiency of the fitted-synthesis method. Ideal experimental conditions were assumed.

The general testing methodology was first presented. In the tests, the expansion-functions are generated by the static module of the nodal diffusion code QUANDRY. The transient module of the code is used to compute a reference solution to the transient under study. Detector signals are formed by combining instantaneous local-flux values from the reference solution with known, characteristic, detector weight-functions.

The synthesis method was implemented in the form of an experimental FORTRAN computer-program. The program receives as inputs expansion functions, flux measurements, and detector characteristics; it returns as outputs reconstructed group-fluxes and reactivity.

This program was applied to the analysis of a severe operational transient involving control-rod motions and thermal-hydraulic feedback effects. A PWR model was used in this simulation. The reactor was assumed to be equipped with $J = 80$, fixed, in-core detectors asymmetrically distributed in four horizontal planes.

A first synthesis calculation using $K = 26$ expansion functions, but only $R = 15$ modes (11 singular values were zeroed), yielded good results. The CPU time required by the synthesis calculation was on the order of minutes as opposed to hours for the QUANDRY reference calculation. Errors in reconstructed group-two fluxes were a few percent or less, except in nodes in which perturbations (rod motions) occurred. In these perturbed nodes, errors as large as 25% in magnitude were observed.

The modal form of the synthesis approximation was used to explain the anomalously large local errors in reconstructed group-fluxes. The cause of these errors was identified as the indiscriminate “lumping” of the modal information contained in the 26 static flux-shapes into a single matrix $\mathbf{A}$. Because of this lumping, there was no obvious cutoff-value in the singular-value spectrum of $\mathbf{A}$, which made it unclear whether some of these singular values had to be zeroed or not.

Improved performance resulted when the synthesis program was modified to allow the use of a sequence of small subsets of expansion
functions instead of a single large set. Time intervals were defined, and different subsets of flux-shapes were attributed to each of them. Each subset was made of four flux-shapes corresponding to core conditions closely bracketing the instantaneous state of the reactor. A new, better conditioned, 80-by-4 matrix $A$ was formed at the beginning of each time-interval, and its SVD was computed. By presenting the information in this selective manner to the synthesis program, the singular-value clustering problem disappeared, total computing-cost was reduced, and the accuracy of the results was improved. Maximum errors in group fluxes were reduced to approximately 5%. These maximum values occurred in partially-rodded nodes. Errors in reactivity were less than one millibeta.

This idea of a time-discontinuous, fitted-synthesis method was then generalized. An on-line, automatic shape-selection algorithm was proposed as an alternative to the time-interval method used in the tests. This algorithm makes use of the reactor nuclear and nonnuclear instrumentation, and does not require a priori knowledge of the future evolution of the reactor.

Lastly, the influence of the number and positions of the detectors on the time-discontinuous, fitted-synthesis method was studied. Surprisingly good results were obtained when the 80 in-core detectors were replaced by either four 4 in-core detectors, or 48 out-of-core detectors.

In Chapter 5, the influence of noise in the detector readouts was investigated. The same testing methodology, reactor model, and operational transient as in Chapter 4 were used.

The presence of random errors in the measurements was shown to be a possible source of difficulties for syntheses containing high-order modes. A statistical analysis was performed to quantify the potential effect of zero-mean, stationary noise on various outputs of the synthesis program: relative residual, modal coefficients, and reconstructed fluxes. The orthonormal bases constructed by SVD were found to be most convenient to express covariance matrices. It was established that (i) the variance of any modal coefficient is inversely proportional to the square of the corresponding singular value, and that (ii) the fitted-synthesis method has an inherent noise-reduction effect: The noise variance is reduced by an average factor
of $R / J$, a small number when $J \gg R$. This variance reduction factor applies to most of the nodes if the detectors are reasonably distributed throughout the core. It was also predicted that this reduction could be offset for a limited time in nodes affected by local perturbations. The reason for this net amplifying effect in a few nodes is the extreme sensitivity to noise of the coefficients of the high-order modes in the synthesis formula.

Numerical simulations confirmed these theoretical predictions despite the fact that the artificial noise added to the detectors signals was not stationary. Accrued sensitivity to noise was observed when only 4 in-core detectors, or 48 out-of-core detectors were used. Evidence was given that zeroing the smallest singular values lead to a reduction in random errors at the expense of a small increase in systematic errors. In the simulation, zeroing more than one singular value, although beneficial from a noise-reduction standpoint, caused unacceptably large systematic errors in some nodes.

Simple, low-pass, filtering algorithms were developed as a more flexible alternative to the procedure of reducing random errors by truncating the synthesis formula. Nonrecursive, fixed-memory polynomial filters on the one hand, and recursive, fading-memory polynomial filters on the other hand, were constructed and “tuned” so as to achieve a reasonably small variance-reduction factor. Applications of these filters to the 80 noisy detector-signals lead to the desired variance reduction. Further improvement was even possible by modifying the transfer function of the fixed-memory filters.

The combination of on-line filtering of the detector readouts and zeroing of high-order modes resulted in an acceptably accurate synthesis.

In Chapter 6, the time-discontinuous, fitted-synthesis method was applied to two other transients: a coolant inlet-temperature transient, and a reactor-startup transient. The same PWR model and 80 in-core flux-sensors as in Chapter 4 and 5 were used in these simulations.

A single set of nine basis-functions was supplied to the synthesis program for the analysis of the coolant inlet-temperature transient. In spite of the rapid dynamics of this transient, a five-mode synthesis was still able to reconstruct instantaneous, nodal group -fluxes with errors of less
than 1.3% in magnitude throughout the transient. Performance was also very good with noisy signals. This was explained by the absence of very high order perturbation-modes.

The excellent results obtained in all the numerical simulations motivated the application of the synthesis program to a very challenging startup transient. Several (unrealistic) assumptions were made in this simulated startup to create difficult conditions for the fitted-synthesis method. Expansion functions obtained from steady-state, nonhomogeneous calculations with startup sources were included in the first time-intervals of the transient. Strong dynamic effects in the source range succeeded in causing significant relative errors in some nodes. However, these fairly large errors were not grounds for rejecting the method since the physical conditions created numerically are not encountered in actual reactor startups.

7.2. Conclusion.

A semi-experimental method for "unfolding" the detailed, instantaneous neutron-flux distribution in light-water reactors from precomputed flux-shapes and local flux measurements has been developed. This method avoids the solution of the time-dependent neutron diffusion equations. Instantaneous nodal group-fluxes are approximated by a linear combination of precomputed, static, three-dimensional, nodal flux-shapes. The time-dependent coefficients of the combination are found from the requirement that the approximation agree as closely as possible with the readings of fixed, fast-responding neutron-detectors. Potential numerical difficulties arising from the ill-conditioning of the system of least-squares equations are brought under complete control by the use of a singular-value-decomposition technique. This technique also makes it possible (i) to obtain a solution that is consistent with the limited accuracy of the data, and (ii) to interpret the fitted-synthesis formula as a form of modal expansion containing orthogonalized modes of decreasing importance. The smallest of these modes are discarded when their physical meaning cannot be
ascertained. Additional modes may be eliminated if it is desirable to reduce the sensitivity of the solution to measurement noise.

Numerical tests performed for severe transients in a PWR model demonstrated that the method performs best when it is allowed to be discontinuous in time, i.e. when different subsets of expansion functions are used rather than a single large set. These subsets should be formed from only a few basis-functions closely-bracketing the instantaneous reactor conditions.

In the absence of random and systematic errors in the detector outputs, and with the use of appropriate basis-functions in the synthesis, the residual of the least-squares fit gives a reliable, integral measure of the systematic errors caused by the synthesis approximation. Local error estimates can be obtained from the highest-order mode present in the synthesis formula or from the first neglected mode.

If (i) the detectors are sufficiently numerous and reasonably distributed throughout the core, (ii) their characteristics are perfectly known, (iii) the flux measurements are unbiased and noise-free, and (iv) only small subsets of carefully-selected expansion-functions are used at any time-step in the synthesis, then the method is capable of reconstructing instantaneous nodal group-fluxes in real-time with errors of less than 1 %, except possibly in a few locally-perturbed nodes where errors on the order of 5 % may occur. Reactivity can be inferred with an accuracy better than one millibeta.

Actual implementation of the method would require the availability of a static nodal diffusion code, and of fixed in-core neutron-detectors having known response-functions. An assembly code would also be necessary to determine two-group, homogenized, detector cross sections. (The assembly code used for generating the equivalent, two-group, nodal diffusion-theory parameters supplied to the nodal code could be used.) The accuracy of the detector weight-functions and the reliability of the signals returned by the instrumentation conditions the quality of the results.

The main advantages of the method are that it is simple, easy to implement (assuming that the required instrumentation is already in place), and very efficient from a computational standpoint. In addition, its
memory requirements are very modest. Another important advantage is that it does not require detailed, instantaneous information about the thermal-hydraulic conditions of the core. Thermal-hydraulic data is needed only for generating the static expansion-functions.

The main disadvantages of the method are that it has limited predictive capabilities because it is not model-based, and that it lacks robustness because of its strong reliance on the reactor instrumentation. Also, the bookkeeping effort associated with the generation of the expansion functions may be substantial if a wide range of reactor operating conditions is to be covered.

7.3. Recommendations for Future Research.

The fact that the fitted-synthesis method was able to reproduce reference results very accurately for the severe transients studied in this investigation augurs even better performance in less extreme conditions more typical of actual reactor transients. Additional numerical simulations should be performed for a variety of realistic operating conditions to verify this conjecture. Different reactor models should be used in these tests. Confidence in the method could be gained by the accumulation of systematically good results.

Numerical simulations, however, no matter how successful, cannot be a substitute for experimental evidence. An experimental evaluation of the method in actual reactor conditions would be extremely valuable. Such an evaluation could be most easily performed in a plant already equipped with fixed in-core flux-sensors. Simple experiments would require little or no investment in hardware, and minimum software developments. Yet, substantial benefits would result from such experimental tests. In particular, possible unforeseen difficulties with the method could be identified.

In an actual implementation of the method, a library of expansion functions would have to be constructed (and updated regularly to reflect
long-term changes in core conditions), and some means of identifying and retrieving “best” flux-shapes from that library on-line and in real time would have to be available, as explained in Section 4.6.3. The systematic generation and storage of a large number of expansion functions is a simple but cumbersome approach. Unless this bookkeeping task can be automated, more economical alternatives should be sought.

As mentioned repeatedly throughout this investigation, serious practical difficulties with the method may be caused by the presence of errors in the detector characteristic weight-functions. These errors arise from the evaluation of the detector response-functions or from the subsequent homogenization calculations. The first of these two sources of errors is of most concern. The evaluation of the response-function of a neutron detector is a very difficult task in general, in part because the physical characteristics of the detector (material properties and geometry) are never perfectly known (they change with time and with the operating conditions), and also because of the complexity of the physical interactions taking place in the sensor. As a result, the computed response-function may not be representative of the actual sensor. When this happens, the signal returned by the detector differs from the signal which would be obtained from the “exact” response-function. In some cases, the disagreement may be such that the detector outputs are considered “wild” and the sensor “unreliable”.

Unnoticed, biased measurements (outliers) would have a very detrimental effect on the least-squares fit of the synthesis. An efficient signal-validation algorithm is therefore necessary for a successful synthesis. However, it may not always be sufficient. A Chebyshev fit rather than a least-squares fit (i.e. the use of the infinity-norm rather than the two-norm) would provide considerably more robustness to outliers. Alternatively, a weighting procedure in which “doubtful” measurements are given low weights could be incorporated in the least-squares fit.

It may be argued that the most serious drawback of the method is its theoretical weakness because of the absence of a dynamic model and of a rigorous, theoretical error-bound on the results. As a partial remedy to this difficulty, a more conventional procedure for determining the mixing
coefficients could be adopted. For example, the synthesis formula could be substituted into the time-dependent, nodal, neutron balance equation, and the result required to be true in some weighted integral sense [H8, L5]. A set of nonlinear, ordinary differential equations ("integral model") would result. A foreseeable difficulty with this procedure is the evaluation of some of the coefficients (time-dependent integrals) in these equations because of the presence of cross sections having complicated functional dependences. Additional equations and variables would probably have to be introduced, thereby complicating the method. However, this weighted-residual procedure would have the advantage of producing "pure" theoretical estimates, which would make the application of standard control-theory methods possible.
REFERENCES


APPENDIX A1

DESCRIPTION OF THE PWR MODEL

a. Geometry: See Figures 4.2a,b.

Node Size = 21.591 cm \times 21.591 cm \times 20.000 cm

\[
\phi_y = \alpha_y I_y
\]

\[
\phi_x = \alpha_x I_x
\]

Figure A1.1. X-Y Rodded Cross-Section of the PWR Model (1/4 Core) Showing Boundary Conditions and Composition Numbers.
In unrodded planes, Composition # 1 replaces Composition # 16.

b. Material Properties:

Table A1.1. Two-Group Nodal Cross Sections and Diffusion Coefficients for the PWR Model.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Group g</th>
<th>(D_g) (cm)</th>
<th>(\Sigma_{ag}) (cm(^{-1}))</th>
<th>(\nu\Sigma_{fg}) (cm(^{-1}))</th>
<th>(\Sigma_{g'g}) (cm(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>g = 1</td>
<td>1.3648</td>
<td>0.008887</td>
<td>0.005550</td>
<td>0.017245</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4826</td>
<td>0.130772</td>
<td>0.185823</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>g = 1</td>
<td>1.3603</td>
<td>0.009661</td>
<td>0.006267</td>
<td>0.015942</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4776</td>
<td>0.169403</td>
<td>0.229195</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>g = 1</td>
<td>1.3396</td>
<td>0.009957</td>
<td>0.006267</td>
<td>0.015398</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4798</td>
<td>0.181915</td>
<td>0.230258</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>g = 1</td>
<td>1.3592</td>
<td>0.010104</td>
<td>0.006269</td>
<td>0.015728</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4810</td>
<td>0.188426</td>
<td>0.230923</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>g = 1</td>
<td>1.3594</td>
<td>0.009509</td>
<td>0.006890</td>
<td>0.016386</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4673</td>
<td>0.169073</td>
<td>0.264760</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>g = 1</td>
<td>1.35898</td>
<td>0.0096925</td>
<td>0.0068905</td>
<td>0.0160495</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.46853</td>
<td>0.1762888</td>
<td>0.2653975</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>g = 1</td>
<td>1.33890</td>
<td>0.009730</td>
<td>0.006890</td>
<td>0.015981</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.46875</td>
<td>0.177654</td>
<td>0.265512</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>g = 1</td>
<td>1.3576</td>
<td>0.010252</td>
<td>0.006892</td>
<td>0.015022</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4728</td>
<td>0.200287</td>
<td>0.267778</td>
<td>0.0</td>
</tr>
<tr>
<td>9</td>
<td>g = 1</td>
<td>1.3572</td>
<td>0.010399</td>
<td>0.006894</td>
<td>0.014752</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.4740</td>
<td>0.206951</td>
<td>0.268552</td>
<td>0.0</td>
</tr>
<tr>
<td>10, 14</td>
<td>g = 1</td>
<td>1.4957</td>
<td>0.002683</td>
<td>0.0</td>
<td>0.022923</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.3637</td>
<td>0.051595</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>11</td>
<td>g = 1</td>
<td>1.3933</td>
<td>0.003541</td>
<td>0.0</td>
<td>0.017943</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.3659</td>
<td>0.068149</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>12, 15</td>
<td>g = 1</td>
<td>1.6701</td>
<td>0.001220</td>
<td>0.0</td>
<td>0.031408</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.3621</td>
<td>0.039330</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>13</td>
<td>g = 1</td>
<td>1.7446</td>
<td>0.005960</td>
<td>0.0</td>
<td>0.035032</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.3614</td>
<td>0.034208</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>16</td>
<td>g = 1</td>
<td>1.321964</td>
<td>0.013482</td>
<td>0.005567</td>
<td>0.015178</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.486198</td>
<td>0.211003</td>
<td>0.194976</td>
<td>0.0</td>
</tr>
</tbody>
</table>
c. Discontinuity Factors (All Compositions):

Table A1.2. Discontinuity Factors for all Nodes, Groups, and Directions.

<table>
<thead>
<tr>
<th>Direction u</th>
<th>Group g</th>
<th>$f_{g1-}$</th>
<th>$f_{g2+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>g=1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>g=2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>y</td>
<td>g=1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>g=2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>z</td>
<td>g=1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>g=2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

d. Albedoes:

Table A1.3. Albedo Values for Radial and Axial Boundaries.

<table>
<thead>
<tr>
<th>Direction u</th>
<th>$\alpha_{11}$</th>
<th>$\alpha_{12}$</th>
<th>$\alpha_{21}$</th>
<th>$\alpha_{22}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>y</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>z</td>
<td>4.011</td>
<td>0.0</td>
<td>2.805</td>
<td>8.993</td>
</tr>
</tbody>
</table>

e. Thermal-Hydraulic Data:

Specific heat of the fuel = 2.46 $10^6$ erg g$^{-1}$ K$^{-1}$ \ (1 erg = 10$^{-7}$ J)
Specific heat of the coolant = 5.43 $10^7$ erg g$^{-1}$ K$^{-1}$
Density of the fuel = 10.3 g cm$^{-3}$
Total Mass Flow-Rate = 3.868 $10^6$ g s$^{-1}$
Film coefficient = 3.293 $10^7$ erg cm$^{-2}$ s$^{-1}$ K$^{-1}$
Conductivity / Conduction length of clad = 2.2 $10^6$ erg cm$^{-2}$ s$^{-1}$ K$^{-1}$
Surface area of clad / Volume of coolant = 3.097 cm$^{-1}$
Volume fraction of the coolant = 0.542
Inlet coolant temperature = 555 K
Fraction of fission energy released in coolant = 0.026
Coolant pressure = 1.551 \times 10^7 \text{ Pa}
Partial derivative of product density times enthalpy
with respect to coolant temperature = 1.6 \times 10^7 \text{ erg cm}^{-3} \text{ K}^{-1}

f. Cross-Section Feedback Coefficients:

Table A1.4a. Coolant-Density Feedback Coefficients.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Group g</th>
<th>$\frac{d}{d\rho_c} \frac{1}{D_g}$ (cm$^2$ g$^{-1}$)</th>
<th>$\frac{d}{d\rho_c} \Sigma_{cg}$ (cm$^2$ g$^{-1}$)</th>
<th>$\frac{d}{d\rho_c} \Sigma_{g'g}$ (cm$^2$ g$^{-1}$)</th>
<th>$\frac{d}{d\rho_c} v \Sigma_{fg}$ (cm$^2$ g$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 through 9 and 16</td>
<td>g = 1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>10 to 15</td>
<td>g = 1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Reference Coolant-Density = 0.79755 g cm$^{-3}$

Table A1.4b. Coolant-Temperature Feedback Coefficients.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Group g</th>
<th>$\frac{d}{dT_c} \frac{1}{D_g}$ (cm$^{-1}$ K$^{-1}$)</th>
<th>$\frac{d}{dT_c} \Sigma_{cg}$ (cm$^{-1}$ K$^{-1}$)</th>
<th>$\frac{d}{dT_c} \Sigma_{g'g}$ (cm$^{-1}$ K$^{-1}$)</th>
<th>$\frac{d}{dT_c} v \Sigma_{fg}$ (cm$^{-1}$ K$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 through 9 and 16</td>
<td>g = 1</td>
<td>- 8.0 \times 10^{-5}</td>
<td>3.0 \times 10^{-6}</td>
<td>- 1.5 \times 10^{-6}</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>- 1.3 \times 10^{-3}</td>
<td>- 8.2 \times 10^{-6}</td>
<td>0.0</td>
<td>- 2.075 \times 10^{-5}</td>
</tr>
<tr>
<td>10 to 15</td>
<td>g = 1</td>
<td>- 8.0 \times 10^{-5}</td>
<td>3.0 \times 10^{-6}</td>
<td>- 1.5 \times 10^{-6}</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>g = 2</td>
<td>- 1.3 \times 10^{-3}</td>
<td>- 8.2 \times 10^{-6}</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Reference Coolant-Temperature = 533 K

520
Table A1.4c. Fuel-Temperature Feedback Coefficients.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Group g</th>
<th>$\frac{d}{dT^f} \frac{1}{D_g}$</th>
<th>$\frac{d}{dT^f} \Sigma_{c_g}$</th>
<th>$\frac{d}{dT^f} \Sigma_{g_g}$</th>
<th>$\frac{d}{dT^f} v_{\Sigma_{f_g}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 through 9 and 16</td>
<td>$g = 1$</td>
<td>$-6.6 \times 10^{-6}$</td>
<td>$3.3 \times 10^{-7}$</td>
<td>$-8.5 \times 10^{-8}$</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$g = 2$</td>
<td>$-2.6 \times 10^{-6}$</td>
<td>$-3.8 \times 10^{-7}$</td>
<td>0.0</td>
<td>$-2.5 \times 10^{-6}$</td>
</tr>
<tr>
<td>10 to 15</td>
<td>$g = 1$</td>
<td>$-6.6 \times 10^{-6}$</td>
<td>$3.3 \times 10^{-7}$</td>
<td>$-8.5 \times 10^{-8}$</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$g = 2$</td>
<td>$-2.6 \times 10^{-6}$</td>
<td>$-3.8 \times 10^{-7}$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Reference Fuel-Temperature = 533 K

g. Energy Spectra:

\[ \chi_{p_1} = \chi_{d_1} = 1.0 \]
\[ \chi_{p_2} = \chi_{d_2} = 0.0 \]

h. Neutron Homogenized Speeds (All Compositions):

\[ v_1 = 1.25 \times 10^7 \text{ cm s}^{-1} \]
\[ v_2 = 2.50 \times 10^5 \text{ cm s}^{-1} \]

i. Delayed-Neutron Data:

Table A1.5. Delayed-Neutron Precursor Fractions and Decay Constants.

<table>
<thead>
<tr>
<th>Family d</th>
<th>$\beta_d$ (-)</th>
<th>$\lambda_d$ (s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000247</td>
<td>0.0127</td>
</tr>
<tr>
<td>2</td>
<td>0.0013845</td>
<td>0.0317</td>
</tr>
<tr>
<td>3</td>
<td>0.001222</td>
<td>0.115</td>
</tr>
<tr>
<td>4</td>
<td>0.0026455</td>
<td>0.311</td>
</tr>
<tr>
<td>5</td>
<td>0.000832</td>
<td>1.40</td>
</tr>
<tr>
<td>6</td>
<td>0.000169</td>
<td>3.87</td>
</tr>
</tbody>
</table>
APPENDIX A2

POLYNOMIAL FILTERS

This appendix contains formulas for the fixed-memory and fading-memory polynomial filters [M8] described in Section 4.3 of Chapter 5.

A2.1. Fixed-Memory Polynomial Filters.

a. Notational Conventions: See Figure 5.58.

b. Least-Squares Criterion:

\[
\text{Minimize } \sum_{r=0}^{L} (x_{-L+r} - p(r))^2
\]  

(5.96)

c. Filtered Estimate:

\[
y_n = p(r)_{r=n} = \sum_{r=0}^{L} w_r(n) x_{-L+r}
\]  

(5.99)

d. Discrete Legendre Polynomials:

\[
p_j(r) = \sum_{\nu=0}^{j} (-1)^\nu \binom{j}{\nu} \binom{j+\nu}{\nu} \frac{r!}{(r-\nu)!} \frac{(L-\nu)!}{L!} ,
\]

\[j = 0, 1, \ldots, M \quad (A2.1)\]
e. Properties:

\[
\sum_{r=0}^{L} \frac{p_i(r)}{c_i} \frac{p_j(r)}{c_j} = \delta_{ij}, \quad \begin{cases} i = 0,1,\ldots,M \\ j = 0,1,\ldots,M \end{cases}
\] (5.97)

\[p_j(0) = 1, \quad j = 0,1,\ldots,M \] (A2.2)

\[p_j(L) = (-1)^j, \quad j = 0,1,\ldots,M \] (A2.3)

\[p_j \left( \frac{L}{2} \right) = 0, \quad j = 1,3,5,\ldots \] (A2.4)

f. Normalization Constants:

\[c_j^2 = \frac{(2j+1) L! L!}{(L-j)! (L+j+1)!} = \sum_{r=0}^{L} \left[ p_j(r) \right]^2, \quad j = 0,1,\ldots,M \] (5.98)

Table A2.1. Normalization Constants of the First Discrete Legendre Polynomials.

<table>
<thead>
<tr>
<th>j</th>
<th>(c_j^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(\frac{1}{L+1})</td>
</tr>
<tr>
<td>1</td>
<td>(\frac{3L}{(L+2)(L+1)})</td>
</tr>
<tr>
<td>2</td>
<td>(\frac{5L(L-1)}{(L+3)(L+2)(L+1)})</td>
</tr>
<tr>
<td>3</td>
<td>(\frac{7L(L-1)(L-2)}{(L+4)(L+3)(L+2)(L+1)})</td>
</tr>
</tbody>
</table>

523
g. Filter Coefficients:

\[ w_r(n) = \sum_{j=0}^{M} \frac{p_j(n) \cdot p_j(r)}{c_j} , \quad r = 0,1,\ldots,L \]  

(5.100)

Table A2.2. Coefficients of Fixed-Memory Polynomial Filters of Order M = 1, 2, 3, and 4.

<table>
<thead>
<tr>
<th>M</th>
<th>( w_r(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \frac{1}{c_0^2} )</td>
</tr>
<tr>
<td>1</td>
<td>( \frac{1}{c_0^2} + \frac{1}{c_1^2} \left( 1 - \frac{n}{L} \right) \left( 1 - \frac{r}{L} \right) )</td>
</tr>
<tr>
<td>2</td>
<td>( \frac{1}{c_0^2} + \frac{1}{c_1^2} \left( 1 - \frac{n}{L} \right) \left( 1 - \frac{r}{L} \right) + \frac{1}{c_2^2} \left( 1 - \frac{6 \cdot n}{L} + 6 \frac{n-1}{L \cdot L-1} \right) \left( 1 - \frac{6 \cdot r}{L} + 6 \frac{r-1}{L \cdot L-1} \right) )</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{1}{c_0^2} + \frac{1}{c_1^2} \left( 1 - \frac{n}{L} \right) \left( 1 - \frac{r}{L} \right) + \frac{1}{c_2^2} \left( 1 - \frac{6 \cdot n}{L} + 6 \frac{n-1}{L \cdot L-1} \right) \left( 1 - \frac{6 \cdot r}{L} + 6 \frac{r-1}{L \cdot L-1} \right) + \frac{1}{c_3^2} \left( 1 - \frac{12 \cdot n}{L} + 30 \frac{n-1}{L \cdot L-1} - 20 \frac{n-2}{L \cdot L-1} \cdot \frac{n-2}{L-2} \right) \left( 1 - \frac{12 \cdot r}{L} + 30 \frac{r-1}{L \cdot L-1} - 20 \frac{r-1}{L \cdot L-1} \cdot \frac{r-2}{L-2} \right) )</td>
</tr>
</tbody>
</table>

d. Variance of \( y_n \):

\[ \sigma_n^2 = \sigma^2 \sum_{r=0}^{L} [w_r(n)]^2 = \sigma^2 \sum_{j=0}^{M} \left[ \frac{p_j(L+n)}{c_j} \right]^2 \]  

(5.101)
Table A2.3. Variance $\sigma_{L/2}^2$ for Fixed-Memory Filters of Order $M = 0, 1, 2, \text{ and } 3$ (Filtering at the Center of the Interval).

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\sigma_{L/2}^2$</th>
<th>$\sigma_{L/2}^2$ for $L &gt;&gt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{L + 1}$</td>
<td>$\frac{1}{L}$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{L + 1}$</td>
<td>$\frac{1}{L}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{3(3L^2 + 6L - 4)}{4(L + 3)(L + 1)(L - 1)}$</td>
<td>$\frac{9}{4L}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{3(3L^2 + 6L - 4)}{4(L + 3)(L + 1)(L - 1)}$</td>
<td>$\frac{9}{4L}$</td>
</tr>
</tbody>
</table>

Table A2.4. Variance $\sigma_0^2$ for Fixed-Memory Filters of Order $M = 0, 1, 2, \text{ and } 3$ (Filtering at the End of the Interval).

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\sigma_0^2$</th>
<th>$\sigma_0^2$ for $L &gt;&gt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{L + 1}$</td>
<td>$\frac{1}{L}$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{2(2L + 1)}{(L + 2)(L + 1)}$</td>
<td>$\frac{4}{L}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{3(3L^2 + 3L + 2)}{(L + 3)(L + 2)(L + 1)}$</td>
<td>$\frac{9}{L}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{8(2L^3 + 3L^2 + 7L + 3)}{(L + 4)(L + 3)(L + 2)(L + 1)}$</td>
<td>$\frac{16}{L}$</td>
</tr>
</tbody>
</table>

A2.2. Fading-Memory Polynomial Filters.

a. Notational Conventions: See Figure 5.71.

b. Least-Squares Criterion:
Minimize $\sum_{k=0}^{+\infty} [x_{-k} - q(k)]^2 \theta^k \quad (5.104)$

c. Filtered Estimate:

$$y_n = q(k)\Big|_{k=n} = \sum_{k=0}^{+\infty} w_k(n) x_{-k} \quad (5.107)$$

d. Discrete Laguerre Polynomials:

$$q_j(k) = \theta^j \sum_{v=0}^{j} (-1)^v \binom{j}{v} \binom{k}{v} \left(\frac{1-\theta}{\theta}\right)^v, \quad j = 0, 1, \ldots, M \quad (A2.5)$$

e. Property:

$$\sum_{k=0}^{\infty} \frac{q_i(k) q_j(k)}{d_i d_j} \theta^k = \delta_{ij}, \begin{cases} i = 0, 1, \ldots, M \\ j = 0, 1, \ldots, M \end{cases} \quad (5.105)$$

f. Normalization Constants:

$$d_j^2 = \sum_{k=0}^{\infty} [q_j(k)]^2 = \frac{\theta^j}{1 - \theta}, \quad j = 0, 1, \ldots, M \quad (5.106)$$

g. Filter Coefficients:

$$w_k(n) = \sum_{j=0}^{M} \frac{q_j(n) q_j(k)}{d_j d_j}, \quad k = 0, 1, \ldots, \infty \quad (5.108)$$
h. Recursive Formulation:

0-Step Estimate: \( \mathbf{Y} = [y_0, y_1, \ldots, y_M]^T \)

1-Step Estimate: \( \mathbf{Z} = [z_0, z_1, \ldots, z_M]^T \)

\( \mathbf{Y} = \Phi(-1) \mathbf{Z} \)

\( \Phi(n) = \text{Transition Matrix} = \begin{bmatrix} 1 & n & n^2 & n^3 \\ 0 & 1 & 2n & 3n^2 \\ 0 & 0 & 1 & 3n \\ 0 & 0 & 0 & 1 \end{bmatrix} \) for \( M = 3 \)

\[ [\Phi(n)]_{ij} = \begin{cases} 
\binom{j}{i} n^{j-i}, & 0 \leq i \leq j \leq M \\
0, & \text{otherwise}
\end{cases} \]

Error Estimate: \( \varepsilon = x^{\text{new}} - z_0^{\text{old}} \)

\( M = 0 \)
\( z_0^{\text{new}} = z_0^{\text{old}} + \rho_0 \varepsilon \)
\( y_0 = z_0^{\text{new}} \)

\( M = 1 \)
\( z_1^{\text{new}} = z_1^{\text{old}} + \rho_1 \varepsilon \)
\( z_0^{\text{new}} = z_0^{\text{old}} + z_1^{\text{new}} + \rho_0 \varepsilon \)
\( y_1 = z_1^{\text{new}} \)
\( y_0 = z_0^{\text{new}} - y_1 \)
\[ M = 2 \]
\[
z_2^{\text{new}} = z_2^{\text{old}} + \rho_2 \varepsilon
\]
\[
z_1^{\text{new}} = z_1^{\text{old}} + 2 z_2^{\text{new}} + \rho_1 \varepsilon
\]
\[
z_0^{\text{new}} = z_0^{\text{old}} + z_1^{\text{new}} - z_2^{\text{new}} + \rho_0 \varepsilon
\]
\[
y_2 = z_2^{\text{new}}
\]
\[
y_1 = z_1^{\text{new}} - 2 y_2
\]
\[
y_0 = z_0^{\text{new}} - y_1 - y_2
\]

\[ M = 3 \]
\[
z_3^{\text{new}} = z_3^{\text{old}} + \rho_3 \varepsilon
\]
\[
z_2^{\text{new}} = z_2^{\text{old}} + 3 z_2^{\text{new}} + \rho_2 \varepsilon
\]
\[
z_1^{\text{new}} = z_1^{\text{old}} + 2 z_2^{\text{new}} - 3 z_3^{\text{new}} + \rho_1 \varepsilon
\]
\[
z_0^{\text{new}} = z_0^{\text{old}} + z_1^{\text{new}} - z_2^{\text{new}} + z_3^{\text{new}} + \rho_0 \varepsilon
\]
\[
y_3 = z_3^{\text{new}}
\]
\[
y_2 = z_2^{\text{new}} - 3 y_3
\]
\[
y_1 = z_1^{\text{new}} - 2 y_2 - 3 y_3
\]
\[
y_0 = z_0^{\text{new}} - y_1 - y_2 - y_3
\]

Table A2.5. Weights of Error Estimate in Recursive Formulation of Fading-Memory Filters of Order \( M = 0, 1, 2, \) and 3.

<table>
<thead>
<tr>
<th></th>
<th>( \rho_0 )</th>
<th>( \rho_1 )</th>
<th>( \rho_2 )</th>
<th>( \rho_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M = 0 )</td>
<td>1 - ( \theta )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M = 1 )</td>
<td>1 - ( \theta^2 )</td>
<td>(1 - ( \theta ))^2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M = 2 )</td>
<td>1 - ( \theta^3 )</td>
<td>( \frac{3}{2} (1 + \theta)(1 - \theta)^2 )</td>
<td>( \frac{1}{2} (1 - \theta)^3 )</td>
<td></td>
</tr>
<tr>
<td>( M = 3 )</td>
<td>1 - ( \theta^4 )</td>
<td>( \frac{11 + 14\theta + 11\theta^2}{6} )(1 - ( \theta ))^2</td>
<td>(1 + ( \theta ))(1 - ( \theta ))^3</td>
<td>( \frac{1}{6} (1 - \theta)^4 )</td>
</tr>
</tbody>
</table>

i. Variance of \( y_0 \):

\[
\sigma_0^2 = \sigma^2 \sum_{\lambda=0}^{M} \sum_{\mu=0}^{M} \left( \begin{array}{c} \lambda + \mu \\ \lambda \end{array} \right) \frac{1 - \theta}{(1 + \theta)^{\lambda+\mu+1}} \theta^{\lambda+\mu} 
\]  
(A2.6)
Table A2.6. Variance $\sigma_0^2$ for Fading-Memory Filters of Order $M = 0, 1, 2,$ and $3$ (Filtering at the End of the Interval).

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\sigma_0^2$</th>
<th>$\sigma_0^2$ for $\theta \to 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1 - \theta}{1 + \theta}$</td>
<td>$0.5 \ (1 - \theta)$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{(1 - \theta) \ (5 + 4\theta + \theta^2)}{(1 + \theta)^3}$</td>
<td>$1.25 \ (1 - \theta)$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{(1 - \theta) \ (19 + 24\theta + 16\theta^2 + 6\theta^3 + \theta^4)}{(1 + \theta)^5}$</td>
<td>$2.062 \ (1 - \theta)$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{(1 - \theta) \ (69 + 104\theta + 97\theta^2 + 64\theta^3 + 29\theta^4 + 8\theta^5 + \theta^6)}{(1 + \theta)^7}$</td>
<td>$2.906 \ (1 - \theta)$</td>
</tr>
</tbody>
</table>