Accelerated Channel Operating Margin for Automated Context and Applications to Design Optimization

by

Zackary Gromko

S.B. Electrical Engineering and Computer Science and Physics
Massachusetts Institute of Technology, 2021

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by .....
Luca Daniel
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by .....
Dr. Shirin Farrahi
Senior Principle Engineer, Cadence Design Systems
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Accelerated Channel Operating Margin for Automated Context and Applications to Design Optimization

by

Zackary Gromko

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Rapidly becoming a cornerstone of signal integrity metrics, Channel Operating Margin (COM) provides a highly desirable single figure of merit encapsulating performance guarantees and providing a variety of byproducts yielding insight on channel behavior. This poignant metric has been anticipated as an effective tool for automated design problems from root cause analysis to design optimization. However, in practice it has largely been mired in the manual design regime, requiring a great depth of tooling and expert input to assess efficiently. The most substantial such bottleneck derives from the dual design problem of finding optimal equalizer settings for general transmitters and receivers given a channel description, which has historically been approached with an expert-guided grid search. We tackle this issue by introducing a practical method in Bayesian optimization accelerated by a lightweight transfer learning framework. We additionally present our methods in developing an unsupervised flow including derivation of channel descriptions using automated PowerSI tooling, fast time-domain assessment, and flexible frequency-space channel simulation. Finally, we discuss briefly the state of applications to design optimization. In sum, we develop a method for automated COM analysis from design to metric permitting the rapid analysis of entire PCBs.

Thesis Supervisor: Luca Daniel
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Dr. Shirin Farrahi
Title: Senior Principle Engineer, Cadence Design Systems

Acknowledgments

My deepest thanks to my partner, Tiffany Chinwuba, without whom I would not have been remotely able to find the successes I have. Also to the excellent Move37 team at Cadence, Professor Luca Daniel, and especially to Dr. Shirin Farrahi. You have all provided invaluable feedback, support, and encouragement while allowing my work to move in the direction that interests me most. Thank you for your time and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With rapidly increasing data transmission rates and the physical complexities they entail, signal channel assessment and design has become an increasingly arduous task. Channel operating margin (COM) was brought about to evaluate and specify minimal device capabilities and assist in design budget allocation. In particular, it supersedes signal integrity metrics based in multiple figures of merit, such as integrated inter-symbol interference (ISI) and crosstalk power sums. This brings a variety of compelling advantages, enabling rapid and more wholistic analysis as well as decision making criteria robust to design topology. These qualities are promising from the perspective of automated design and analysis, however the automated regime entails a more stringent sense of rapidity and a need for rolling the analysis practitioner's supporting work into the fold of the analysis flow.

With the goal of enabling automated design and analysis of arbitrary topology, we therefore present a comprehensive method of producing COM metrics in substantially less time given only a layout. To this end, we present a flow for derivation, assessment, and repair of channel S parameter models, our methods in single bit response construction, and novel work in Bayesian optimization on the intermediate problem of equalizer performance optimization.

An overview of the methods to be presented is shown in figure 1-1. While the overall structure is reflective of prior work, we note several substantive alterations. We include an automated modeling flow to form the S parameter and transfer matrix

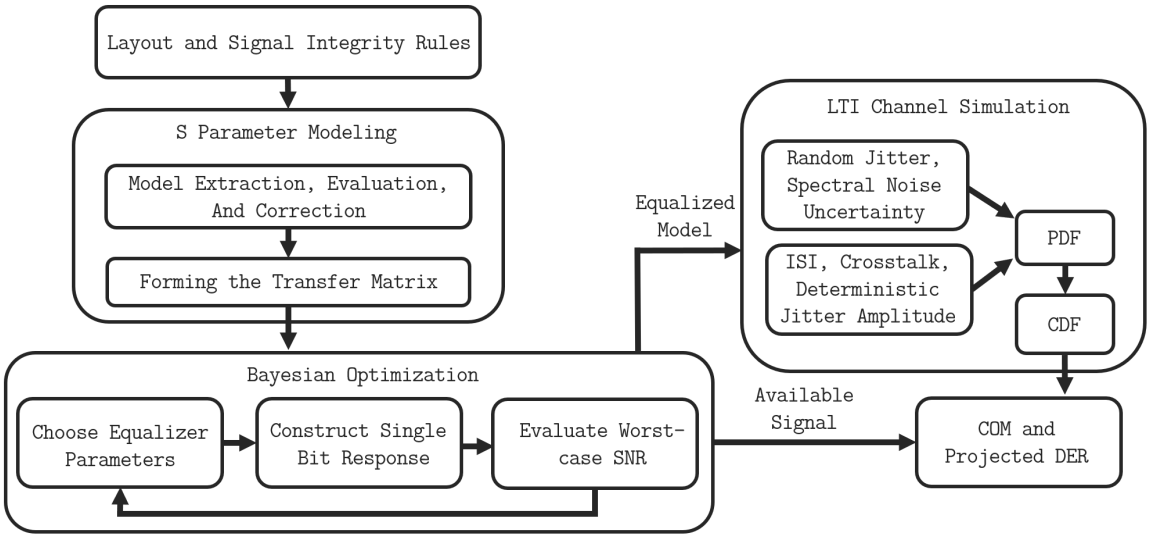


Figure 1-1: A general overview of the methodology that will be presented in performing automated COM analysis directly from layout.

models, in which we must address the issues of S parameter correction and channel identification. In the former we will detail methods in assessing and repairing issues of causality and passivity, partially induced by rapid model extraction. In the latter, we decide which differential pairs to model together and which crosstalk terms to sparsify out, in light of the difficulty inherent in modeling an entire PCB's signal network.

Next, we address the inefficiency of the grid search that has previously been employed for equalizer optimization. We instead apply Bayesian optimization. We note that this step is performed once for each channel on a board, many of which are likely to constitute inherently similar optimization problems. Therefore we further develop a multi-task methodology in order to transfer surrogate models between channels and mitigate the cold-start problem.

Finally, we briefly examine our implementation of the LTI channel simulation used in deriving COM as well as projected data error rate (DER), which we fall back on when target error rates are not available. However, the methods developed here are not generally unique. We focus instead on assessing the assumptions made on the channel and supply a flexible approach to admit efficient frequency-space modeling

for arbitrary specification of random noise forms.

Overall, we find the methods presented to be a substantially faster and reasonably reliable alternative to existing analysis flows, without need for supervision. While this constitutes a step toward applications to automated design, we anticipate that further work is likely needed to fully develop the role of COM analysis in rapid, automated design rollout.

Chapter 2

Prior Work

2.1 COM and Applications

We define the overall structure of our methodology in a similar vein to the works of those first involved in the development of COM analysis. In particular, we refer to "Channel Operating Margin (COM): Evolution of Channel Specifications for 25Gbps and Beyond" [22]. The strategy presented here consists of computing a transfer matrix model, optimizing on the time-domain response over parameterized equalizing filters, and finally using the resulting model in computing COM via LTI channel simulation. The authors demonstrate a simple, yet highly explanatory fit between COM and the traditional validation metric of eye height over a variety of design topologies and the methods presented here have become well accepted.

We therefore adopt the strategy, but aim to address several shortcomings related to our application goals. In particular, these methods and derivatives of them rely on an analysis practitioner to collect, validate, correct, and interpret the models specifying the channels under test. Further, the equalizer optimization routine as formulated requires a designed grid search, which represents a severe performance bottleneck and entails some oversight, particularly when extended to arbitrary equalization schema which may introduce dimensions of differing length scale, sensitivity, and non-uniform regions of interest.

2.2 Bayesian Optimization

In addressing the problem of a cold-start in Bayesian optimization, two methods are well regarded in terms of practical efficacy [30]. The first and simplest ranks similarity to previously completed tasks by examining context features and warm starting using previously meaningful sets of input [27]. This is applied to the problem of ML model hyperparameter tuning and the authors demonstrate success in training a random forest to determine task similarity. In so doing, they examine primarily meta-features of the dataset on which they intend to train the problem model. They then identify hyperparameter settings that produce found extrema and retry the same settings on the new task.

This method is comparatively simple and notably does not inflate the difficulty of acquisition for a given number of points sampled in the new task's domain. However, we consider function-driven rather than data-driven problems and our access to analogous meta-features is mostly limited to those we can determine analytically, which is to say fairly little. Further, the information transfer is somewhat limited without any direct transfer of our belief model. We are also hard-pressed to train any complex model in a reasonable time frame. Nevertheless, the strategy of heuristic selection on previous tasks to target information transfer is attractive for its limited impact on optimization overhead.

In the second method, known as the intrinsic model of coregionalization (ICM), the underlying Gaussian process regression (GPR) is altered to operate on the more general joint parameter-task space by defining a cross-task covariational structure [4]. In the case of ICM, we assume a kernel structure separable along parameter and task axes, implying a constant relational structure termed isotropy. This represents a deeper form of information transfer in that we aim to form a knowledge-discounted initial surrogate model directly from the sampling over all previous tasks. The ICM model also leaves the optimization routine with $\mathrm{O}(n^3)$ complexity, with n the number of samples taken. However, we notice that unlike in the single-task regime, n can grow rapidly in practical terms with the completion of many tasks.

In our problem regime of interest, we are somewhat sensitive to optimization overhead and may have dozens of tasks to consider. Further, each task is of similar evaluation time, preventing us from exploiting other methods in simultaneous optimization using this framework [29]. For these reasons, a full ICM transfer model is impractical for this problem, though we are motivated to explore the structure combined with a heuristic method of pairing down related tasks, thereby reducing model complexity and overhead.

Chapter 3

Channel Modeling

3.1 Channel Specification

In conjunction with optimization, the primary point of reliance on the analysis practitioner in current COM flows lies in derivation of the channel model. In general, this can be a highly complex process requiring a substantial depth of tooling, particularly in work designed to produce verification quality results. By relaxing this constraint and introducing assumptions on typical signal content and reflectivity, we aim to leverage available tooling to produce rapid approximate models that practically abstract away manual modeling assessment and correction.

3.1.1 S Parameter Derivation

The first hurdle we address in automated channel specification lies in the extraction of S parameter models from layout geometry. To this end, we employ the 2021 release of Sigrity's PowerSI. Its own performance claims [14] in the case of non-parallel differential pairs with layer transitions, as well as the efficiency and accuracy noted in independent assessment [21][5], are particularly motivating, though rigorous physical measurement validation of this particular workflow falls out of scope of this work. The analysis presented here targets rapid pre-verification assessment in the context of design tooling, and we therefore take interest largely in the speed of extraction and

ability to automate.

Before simulation, we begin by compiling desirable parameters from layout. First, differential pairs of interest are identified both by frequency and categorization. The traces in layout are then explored via SPICE modeling engine in order to derive models for the passive components that may bifurcate transmission lines and prevent complete simulation. Net-wise ports are then placed on each end of these differential pairs and models are specified to a distributed PowerSI service via TCL script control.

We note several important aspects of the simulation that is run. First, we enable both passivity and dielectric model causality enforcement, however we will check both properties ourselves as in section 3.1.2. This is partially because the discretized model we request introduces issues of causality in its own right. Specifically, it is desirable from an efficiency standpoint to both decide bandwidth a priori rather than on decay of the reflective terms and to simulate a loose adaptive frequency sampling which we may later interpolate into a uniform mesh. In fact, knowledge of the signaling we will apply to our channels of interest admits substantial alleviation of the difficulties inherent to working with discretized models. We may practically bound our frequency band of interest by presumption or specification of transmitter transition capability. We therefore allow PowerSI to AFS up to four times the signaling frequency under the presumption that the pessimistic rise-times we examine will not generally be less than 10% of the symbol width.

3.1.2 S Parameter Post-Processing

Given a touchstone file specifying an S parameter model, several steps must be taken to apply it toward a robust channel description. The quality of the S parameters serving as input into COM analysis generally has substantial effect on the quality of results [2][22]. In particular, we may be concerned with both passivity and causality. While passivity seems a clear reflection of physicality, non-causal components reflect both model physicality as well as the underlying methodology in dealing with the bandlimiting and sampling of the frequency-space mesh [23]. We aim to enforce constraints on both properties and ideally repair violations where possible, a topic of

substantial depth which can be seen characterized in works such as [31].

Before performing any substantive analysis, we first resample the S spectra onto an affine mesh. While more sophisticated treatment such as via time-domain vector fitting as described in [9] and specification via complete functional model are both potential avenues to alleviating resampling concerns, we opt for a swifter and more approximate method under the motivation of pre-verification rapid assessment. This raises the issue of time-domain aliasing and we must choose the mesh spacing so as to ensure any reflected waveforms are fully damped at the end of the associated time-domain sampling period. This is generally nontrivial, as the period necessary is a function of electrical length, reflectivity, and channel gain. The former is estimated via the phase delay found in unwrapping the phase response of the transmitting parameter. In practice, we typically choose some maximal number of meaningful reflections, particularly given that we take the terminal impedences to be reasonably well matched as examined more closely in equation 3.1. In principle however, tradeoff could made in the complexity of analysis to compute the channel transfer matrix and $\Gamma(s)$. For symbol width T, this multiple of electrical length gives a window of KT, or a sampling interval of $\frac{1}{KT}$.

Next, we notice that by formulating our sampled spectra, we have implicitly introduced a hard bandlimit. This is liable to produce non-causal ringing and disrupt our ability to both set appropriate phase automatically as well as measure inter-symbol interference. To this end, we introduce a windowing function. However, we must be wary of any distortion we introduce into our frequency band of interest. To this end, we observe that the sampled bandwidth reaches $4f_s$ while a rise time of 10% the symbol width gives an effective bandlimit of $3.5f_s$ for any signal of interest. Therefore we may use the final $0.5f_s$ of the band for the transition of a high attenuation FIR filter as specified here.

$$|H_{fir}(j\omega)| = \begin{cases} > -0.01\text{dB} & \omega \leq \frac{3.5f_s}{2\pi} \\ < -60\text{dB} & \omega \geq \frac{4f_s}{2\pi} \end{cases}$$

Without any constraints of physical realizability and a reasonably wide transition band, we find a 51-tap Chebyshev filter given by the Parks-McClellan algorithm to satisfy these requirements for arbitrary horizontal scaling.

Next, we implement direct causality and passivity checking. Our formulation of band limitation admits direct analysis of Kramers-Kronig relations, but generally the products of these metrics do not yield a clear method of repair. Instead, we opt to examine the integrated quadratic constraints that directly give passivity, as examined commonly in works such as [12]. In particular, we must find for unitary U the hermitian form to be positive semidefinite, and therefore the singular values of S to be limited to 1.

$$\begin{aligned} &(U - S^*S) \succeq 0 \\ \implies &\sigma_i(S) < 1 \qquad \forall i \end{aligned}$$

Performing the SVD at each sample point then gives a straightforward, if inefficient, method for determining passivity. This further admits repair based on convex optimization [6], or more efficient methods such as those shown in [11]. Adaptive sampling schema and characterization using efficient Hamiltonian methods such as presented in [28] may be a similarly fruitful route given the methods discussed. However, in practice we find no structural issue with models produced via PowerSI when using passivity and dielectric causality enforcement along with the resampling and windowing described. For this reason, further development on this methodology lies outside the scope of this work, but serves as a natural extension for more general model derivation methods.

3.1.3 Equalization

The final substantive component of a channel we must specify in order to derive reasonable post-layout behavioral approximations lies in the equalization scheme. In general, we expect parameterized transfer functions specifying post-transmission and pre-reception equalization as well as a number of post-sampling DFE taps to

consider. The parameters describing a specific realization of these equalizers will then be derived so as to produce the best SNR over the channel as described in section 3.2.2. Initial work in the area assumed up to a 3-tap FFE at the transmitting end and a combination of a Bessel-Thomson filter and CTLE at the receiving end, plus an arbitrary number of DFE taps [22].

$$H_{TX}(s) = c_{-2}e^{2Ts} + c_{-1}e^{Ts} + c_0 + c_1e^{-Ts}$$
$$H_{CTLE}(s) = \frac{g_{DC}p_1p_2}{z}\frac{s+z}{(s+p_1)(s+p_2)}$$

As DFE tap gains can be determined analytically from the transient response, this particular scheme is reasonably general but falls on the side of few degrees of freedom with effectively four free parameters. In service of the tractability of optimization, CTLE pole placement is treated as a fixed function of signaling rate. When examining performance on more complex schemes, we introduce free parameters by way of p_1 and p_2 of the CTLE response, a constrained flat group delay Bessel-Thomson filter at the receiver, and two fractionally spaced emphasis terms in the FFE.

$$H_{BT}(s) = \frac{1}{1 + \frac{b_0}{f_b}s + \frac{b_1}{f_b}s^2 + \frac{b_2}{f_b}s^3 + s^4}$$
$$H_{RX}(s) = H_{BT}(s)H_{CTLE}(s)$$

We construct this as somewhat of a pathological case of the difficulty inherent to generalized equalizer optimization. This scheme now admits eight free parameters and presents a domain where a uniform affine mesh betrays issues of length scale, particularly with regard to pole placement.

While we intrinsically assume that the TX and RX devices are linear and time invariant in this work by way of representation by transfer function, we otherwise make no restriction on specific form. That is, we apply these particular schema are used for the purposed of evaluation and as reasonable defaults, but in general our

analysis will be agnostic to the particular form of these functions and indeed we take this as input to our analysis when available.

3.2 Single Bit Response

Construction of the single bit response (SBR) serves as a crucial intermediate result in examining higher order channel behavior. In particular, it will guide decisions on equalizer parameter setting and provide a basis for a generative channel simulation process. Our interest therefore moves beyond examination of the channel's transfer function to include descriptions of noise and interference as well as post-sampling receiver behavior.

3.2.1 Time Domain Response

To begin construction of a channel's time-domain response, we first wish to compute the THRU, NEXT, and FEXT transfer functions. To this end, we proceed in keeping with the original proposal material [22]. That is, we assume a simple rational form of the package impedance and compute the reflection coefficient resulting from equal line termination impedance, Z_T. We perform lookup on package impedance coefficients based on the channel type under scrutiny using typical values given in [2][7] and the referenced [22] in unknown cases. We take Z_T from channel specification and in channels specified by some differential impedance, Z_d, we assume symmetry to a common ground to take $Z_T = Z_d/2$.

$$\begin{aligned} Z_{\text{pkg}}(s) &= \frac{a_2 s^2 + a_1 s + a_0}{b_3 s^3 + b_2 s^2 + b_1 s + 1} \\ \Gamma(s) &= \frac{Z_T - Z_{\text{pkg}}(s)}{Z_T + Z_{\text{pkg}}(s)} \end{aligned} \tag{3.1}$$

We then compute a port-pair-wise transfer matrix. That is, for every differential pair of ports j_1 and j_2, we examine the submatrix of S parameters with every other port-pair k_1 and k_2 to compute transfer functions as shown in equation 3.3. As the S matrix as derived in section 3.1.1 is TX/RX agnostic, this encompasses all far- and

near-end crosstalk terms.

$$\begin{bmatrix} j_1 \\ j_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} \tag{3.2}$$

$$H_{kj}(s) = \frac{S_{21}}{2} \frac{(1-\Gamma(s))^2}{(1-S_{11}\Gamma(s))(1-S_{22}\Gamma(s)) - S_{12}S_{21}\Gamma(s)^2} \tag{3.3}$$

Finally, we must construct some test pulse from which we induce the transient response of interest. Starting with a square pulse disturbance from the 0 equilibrium, we primarily need to adjust the high-frequency content of the pulse in order to produce a waveform that is physically consistent with transmitter capability. In practice, we find that specific transmitter transition characteristics are seldom specified in-layout and in the absence of specification, we opt to presume a one-pole filter with a rise-time no longer than 40% of the symbol width, giving the following pulse with symbol width T.

$$p(s) = \frac{AT\text{sinc}(2\pi Ts)}{1+0.18Ts} \tag{3.4}$$

We now attempt to compensate for a shortcoming of our automated analysis in that our desire to encompass all relevant crosstalk terms across a PCB produces an S matrix of inflated size, incorporating terms of negligible effect and producing transfer functions of vanishing gain. While direct computation of crosstalk transients is viable in that it still saves the complexity associated with repeating transient derivation many times over the course of equalizer optimization, we still prefer to avoid time-domain analysis. For this reason, we instead seek to bound the sampled time-domain cross talk power series by frequency-domain analysis.

Two issues prevent application of Parseval's relation directly. First is lack of knowledge over very high-frequency content in the discretized S parameter model. However, this can be resolved via assumption on the aggressing transmitter transition capability and the observation that passivity and linearity of the channel implies we must see a cutoff of higher frequency content in the received signal. However, when analyzing crosstalk, which is generally sensitive to high-frequency content, the

pessimistic assumption becomes one of short rise time. We therefore suppose it unlikely for a well-budgeted transmitter to attain a rise time less than 5% of the symbol width.

Second, we might be concerned about pathological cases in which a sampling of the time response may occur at regions of large amplitude and arbitrarily small breadth. To remedy this, we apply the same supposition on bandwidth to Bernsteins inequality. Namely, for signal $f(t)$ with bandwidth B:

$$\sup|f(t)| = \sup|\frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega)e^{j\omega t}d\omega| \leq \frac{1}{2\pi}\int_{-B}^{B}|F(j\omega)|$$
$$|\frac{df(t)}{dt}| \leq B\int_{-B}^{B}|F(j\omega)|$$

Given this maximum rate of rise and falloff and taking the bandwidth implied by rise time assumption of $\frac{1}{50T}$, this gives the following relation bounding a cross talk power series Δ_{XT}^2 with a pessimistically unequalized receiver.

$$\int_{-\infty}^{\infty}\Delta_{XT}^2(t)\sum\frac{1}{B\pi}\delta(t-nT) \leq \int_{-\infty}^{\infty}\frac{AT\text{sinc}(2\pi Ts)}{1+0.02Ts}H_{XT}(s)ds$$
$$\sum\Delta_{XT}^2[nT] \leq \frac{\pi}{50T}\int_{\frac{-1}{50T}}^{\frac{1}{50T}}\frac{AT\text{sinc}(2\pi Ts)}{1+0.02Ts}H_{XT}(s)ds$$

This gives a fairly lax bound, but a fast method of sparsifying when H_{XT} is anticipated to be quite small in practical terms for non-interesting aggressors. We note that a sharper bound can be obtained if one is willing to make assumptions on signal form to admit analysis by Chebychev inequality, but we are concerned primarily with certainty that we do not eliminate meaningful contributions.

3.2.2 Intermediate Metrics and the Equalization Problem

The primary result we wish to achieve by examination of the SBR is a preliminary SNR-like metric we may use to evaluate a given equalization strategy. Specifically, we wish to calculate the signal available for some reasonable sampling phase, as well as the uncertainty in magnitude attributed to a variety of sources summarized in

A	Signal Amplitude
σ_{TX}	Transmitted Noise
σ_{ISI}	Inter-symbol Interference
σ_{XT}	Near- and Far-end Crosstalk
σ_S	Integrated Spectral Noise
σ_J	Random and Deterministic Jitter
$\mu[n]$	Derivative of the SBR
Δ_x	The Power Series of Source X

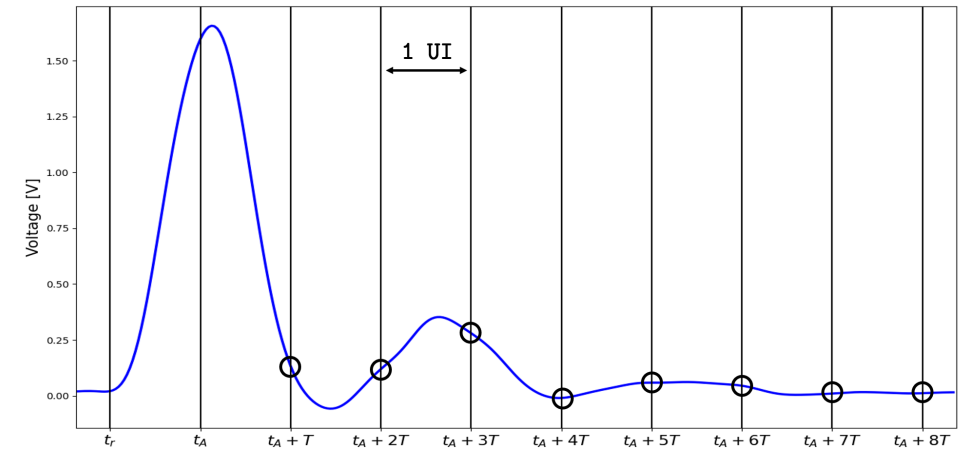
Table 3.1: The terms common to interference analysis presented here and in chapter 5. We note that ISI, crosstalk, and the component of jitter proportional to A_{dd} are referred to as the deterministic sources and modeled accordingly while transmitted noise, spectral noise, and the component of jitter described by σ_{RJ} are taken to be Gaussian in nature.

equation 3.5 and table 3.1

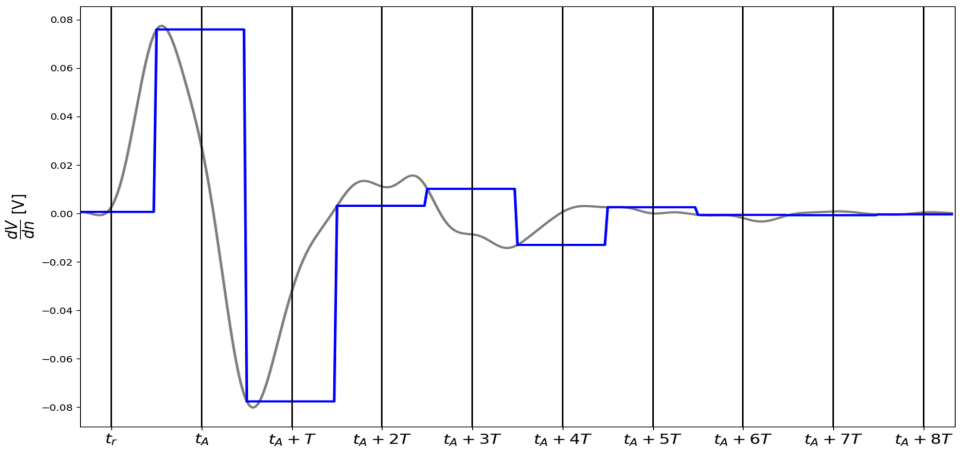
$$M = 10 \log \frac{A^2}{\sigma_{TX}^2 + \sigma_{ISI}^2 + \sigma_{XT}^2 + \sigma_S^2 + \sigma_J^2} \tag{3.5}$$

Beginning with the THRU channel SBR, we first define our sampling phase to begin at the initial signal rise, which is found agnostic to the maximum value attained to enforce a somewhat pessimistic causal, one-shot phase-setting policy. The amplitude is then derived from the sample occurring one UI later as shown at point t_A in figure 3-1a. We then assume, as in the proposal and following adaptations [22][7], a given transmitter SNR which will induce σ_{TX} as a proportion of transmitted signal. Typical values in the original proposal papers and following adaptations take a range of 25-32dB, with higher values assigned to PAM4 channels.

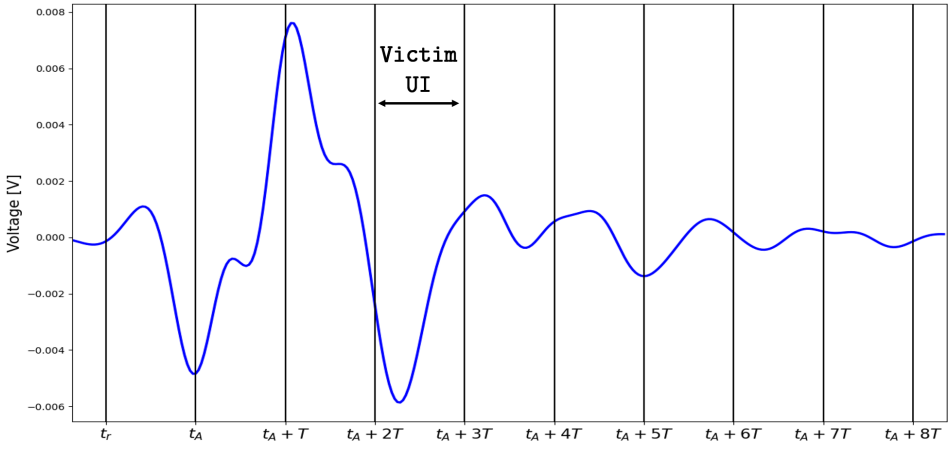
We continue to evaluate ISI within the THRU channel. Given time-independence of our model and an assumption of locally constant phase between symbols, we take uncertainty due to the n^{th} prior symbol to be given by the SBR amplitude n UI after the initial signal, as one might consider a time shifted version of the same SBR. This includes both time-domain smearing and reflected symbols as observed 1 and 3 UI



(a) Transient Response



(b) Transient Derivative, Jitter Susceptibility



(c) Crosstalk Transient Response

Figure 3-1: Example transient response, jitter susceptibility, and crosstalk response. The initial rise is given by t_r and the primary sample at t_A. The Δ_{ISI} series is circled in (a). A pessimistic half-sample zero-order hold on the response derivative is shown in (b), giving $\mu[n]$ which we apply to transform time-axis uncertainty into voltage noise. The crosstalk response of one impinging channel is shown in (c) with a maximally pessimistic phase alignment, yielding one Δ_{XT} series.

after t_A respectively in the transient shown in figure 3-1a, which was constructed with electrical length of 1.5λ. However, given the presence of one or more DFE taps parameterized by gains of $\{b_1 \dots b_k\}$, effective residual ISI is computed as follows.

$$\Delta_{ISI}(n) = \begin{cases} \text{SBR}(t_A + nT) - b_n A & n \leq k \\ \text{SBR}(t_A + nT) & \text{else} \end{cases} \tag{3.6}$$

$$\sigma^2_{ISI} = \sum \Delta^2_{ISI}$$

We proceed identically in our formulation of crosstalk, however we must also construct the interference responses. For a test pulse as constructed in section 3.2.1 consistent with the impinging channel $TP_{XT}(s)$ and transfer H_{kj} for each aggressing channel k, we formulate the time-domain response as follows.

$$\text{SBR}_{XT} = \mathbb{F}^{-1}[TP_{XT}(s)H_{TX}(s)H_{kj}(s)H_{RX}(s)]$$

However, we notice two issues with this formulation. First, while we are given parameters for the victim TX and RX equalization, H_{TX} in the context of the impinging channel is not defined. Further, for an impinging channel of different signaling rate, a pessimistic choice of phase alignment is not obvious. To resolve the former, we make the choice of setting H_{TX} to the identity, which is pessimistic in the primary response, though potentially optimistic in lagging symbols. We compute the latter directly by operating on the entire positive-time SBR_{XT} array, which may generally be performed in a single pass.

One example response is shown in figure 3-1c, though generally we should expect to encounter many such impinging channels indexed here with k. We also note that pessimistic assumptions of transmitter rise-time given in equation 3.4 are relaxed in the inducing pulses here, as we are now concerned with high frequency content.

$$\Delta^2_{XT,k} = \max_i \sum_n \text{SBR}^2_{XT}(i + nT)$$

We then derive σ_{XT}^2 by performing this somewhat expensive inverse Fourier transform and phase alignment extraction for each aggressor channel. Herein lies the need for the culling of crosstalk terms discussed in section 3.2.1.

$$\sigma_{XT}^2 = \sum_k \Delta_{XT,k}^2$$

Finally, we compute spectral noise and jitter terms. The former is relatively straightforward. Given spectral noise density η, we are concerned primarily with RMS amplification at the receiver along the imaginary axis. We therefore compute σ_S as follows:

$$\sigma_S^2 = \eta \int |H_{RX}(j\omega)|^2 d\omega$$

In the case of jitter however, we must transform time axis misalignment into amplitude uncertainty. The Jacobian of this transformation is trivially given by the derivative of the THRU transient response. This exposes a choice in method of finite difference approximation on the discretized response, which we resolve with the simple second-order, centered approximation. This is taken pessimistically at the half-sampling interval, which generally corresponds to the zero-crossing and measurements of horizontal eye diagram closure. Example derivative and zero-order hold is shown in figure 3-1b. We then reach the following expression given dual-Dirac amplitude and RMS random jitter A_{dd} and σ_{RJ}, though we note that these terms will be realized separately in later analysis.

$$\mu[n] = \frac{SBR(t_A + nT/2 + 1) - SBR(t_A + nT/2 - 1)}{2}$$
$$\sigma_J^2 = \sum_n (A_{dd}^2 + \sigma_{RJ}^2)\mu^2[n]$$

In sum, this analysis yields every component in the SNR-like FoM M described in 3.5 as a function of the victim channel equalization parameters specifying H_{TX}, H_{RX}, and the DFE taps utilized in equation 3.6. As per the specification summarized

in figure 1-1, we now must find such parameters that maximize M. This is noted in proposal materials as the most computationally expensive step in computing COM [22].

However, these materials use a grid search constructed by the analysis practitioner to find such maximizing parameters. While we may constrain several parameters as a function of the others, particularly DFE tap gains, we are still subject to tens of thousands of evaluations on M in the best case of a fairly sparse grid and simple schema. Worse, increasing signaling rate leads to physical complexity that must be compensated for via more sophisticated equalization methods. This is to say, we should generally expect dimensionality of this problem to increase over time. In practice, this already leads to many minutes of optimization per channel, even for fewer than 20 differential pairs and associated crosstalk terms under test. In addition to the culling of crosstalk terms presented in section 3.2.1, we aim to further improve this optimization step and wane off of practitioner guidance with directed search methods, namely Bayesian optimization, as presented in chapter 4.

Chapter 4

Multi-Task Bayesian Optimization

Here we examine more general methods in Bayesian optimization, though with application to equalizer parameter optimization and similar problem regimes in mind.

4.1 Background and Single Task Optimization

As the more sophisticated methods we go on to examine largely function as extensions of a single-task optimization routine, we begin by formulating such a method to serve as a basis for further work. We therefore begin with the goal of any Bayesian optimization routine, to build an analytically tractable surrogate model using our statistical knowledge of some true cost function based on previous observations of its behavior. While a variety of surrogate structures have been employed to this end, the two most common comprise tree-structured Parzen estimation and the Gaussian process regression, or Kriging.

Both processes are examined in [3]. However, our choice to give primacy to the latter is due more to conceptual constraint. In the form of some Bayesian update on our beliefs of function output $f(x) = y$ with input parameters x given by

$$p(x|y)p(y)/p(x) = p(y|x)$$

Parzen estimation endeavors to approximate the prior and likelihood on generative

parameters, $p(y)$ and $p(x|y)$ while GPR estimates $p(y|x)$ directly. As we have seen in prior work on multi-task formulation and will construct in 4.2, the latter lends great flexibility in cross-task information transfer while difficulty of establishing relational properties regarding $p(x|y)$ constrains similar methods built on Parzen estimators.

4.1.1 Gaussian Process Regression

We begin construction of our surrogate model's structure by noting that we wish to form a distribution over functions to serve as a prior belief on $f(x) = y$. However, to give stability in the structure of our model and to introduce minimal assumption on the behaviour of the cost function, we wish to choose a point-wise mixture of distributions that will act as its own conjugate prior, and ideally a maximal entropy distribution with support on $\mathbb{R}$. The natural product of this line of thinking is then to choose mean and covariance functions $\mu(x)$ and $k(x, x')$ that at every point in our input parameter space uniquely specify a Gaussian distribution.

$$\begin{aligned} \mu(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))] \\ f(x) &\sim \mathcal{N}(\mu(x), k(x, x')) \end{aligned}$$

Specifying our surrogate structure then reduces to choosing a form for $k(x, x')$ as well as specifying our Bayesian updates on these functions. A rigorous specification of $k(x, x')$ amounts to a choice of covariance function. We again find a crossroad of well-studied choices, of which we examine two: the squared exponential and Matérn classes of function. The former is given below, parameterized by a vector of length scales in each dimension Θ_d and an intrinsic variance scale θ_{var}. When applicable, this provides a desirable kernel as it carries a strong assumption of smoothness. In particular, any realization of a Gaussian process constructed via this kernel must be infinitely differentiable [24]. We may certainly be concerned that this is not appropriate for

our purposes, however.

$$K_{SE}(x, x') = \theta_{var}\exp\left(\frac{-||x - x'||^2_{\Theta_d}}{2}\right) \tag{4.1}$$

The Matérn kernel reproduced below makes a far less heavy-handed presumption of smoothness. Instead it produces models that are only given to be twice differentiable, which is frequently found to produce better results in practical applications [26]. We therefore assess our optimization routine using both kernels in section 4.1.3. Generally, we do indeed find the Matérn kernel more appropriate for our use.

$$K_M(x, x') = \theta_{var}(1 + \sqrt{5||x - x'||^2_{\Theta_d}} + \frac{5}{3}||x - x'||^2_{\Theta_d})\exp\left(-\sqrt{5||x - x'||^2_{\Theta_d}}\right) \tag{4.2}$$

With regards to updating $\mu(x)$ and $k(x, x')$ in a Bayesian way, we modify our notation thus far in two ways. We define the matrix containing all previously sampled parameter vectors X and the vector of all previously observed function evaluations Y. We then construct our entire covariance matrix over previously sampled points and any given unobserved test point x as follows.

$$K = \begin{bmatrix} K(x, x) & K(x, X) \\ K(X, x) & K(X, X) \end{bmatrix}$$

Then, similarly to texts such as [24], we form the joint and conditional distributions on $f(x)$ and Y given a prior mean of 0 and covariance matrix K.

$$\begin{aligned} f(x), Y &\sim \mathcal{N}(0, K) \\ f(x)|x, X, Y &\sim \mathcal{N}(K(x, X)K(X, X)^{-1}Y, \\ &\qquad K(x, x) - K(x, X)K(X, x)^{-1}K(X, x)) \end{aligned}$$

This gives the natural summary statistic functions below, which function as our posterior belief on $f(x)$ given observations $Y = f(X)$ and specify our surrogate model

at any given optimization step.

$$\mu(x) = K(x, X)K(X, X)^{-1}Y \tag{4.3}$$

$$k(x) = K(x, x) - K(x, X)K(X, x)^{-1}K(X, x) \tag{4.4}$$

Finally, we handle the hyperparameter setting of the optimization routine. As we model a noiseless process, for optimization over d parameters we find $d + 1$ hyperparameters by way of θ_{var} and Θ_d as in equations 4.1 and 4.2. With concern for optimization overhead, we opt for point-wise estimates for each. For example, we find a simple form shown in equation 4.5 for the maximum likelihood estimate of θ_{var} as a measurement of observed variance and n the number of samples.

$$\theta_{var} \sim^{MLE} \frac{Y^T K(X, X)^{-1} Y}{n} \tag{4.5}$$

This concludes our treatment of the underlying GP surrogate model. However we note that in practice, it is inefficient to compute the inverse of $K(X, X)$ as prescribed in equations 4.3, 4.4, and 4.5. Instead we leverage both symmetry and positive definiteness conferred by the nature of covariance matrices to perform efficient Cholesky decomposition and solve the system directly. This admits six times fewer operations, which is substantial as these systems generally present the aspect of our surrogate update with the greatest computational complexity with $\mathrm{O}(n^3)$.

4.1.2 Acquisition

Given a surrogate model, our goal is to choose a new point to sample before repeating the process, generally known as acquisition. Intuitively, we would like to maximize the resulting real cost function evaluation. However, as we wish to locate a global maximum, we must develop a surrogate model to reasonable confidence throughout our parameter space as a safeguard against overexploitation of local maxima. A range of well-studied functions on the GPR surrogate exist that are regarded for balancing exploitation and exploration. In particular, we examine expectation maximization

(EI) [15] and entropy search [13], though we will proceed primarily with application of the former.

Entropy search formulates the optimization objective as the maximization of information gain on the distribution of the global optimum. Taking this distribution as P_{opt}, we hypothesize the addition of some estimated point (x, y) into our observation set to form an augmented belief on the location of the optimum, P^{y}_{opt}. The entropy search objective is then given as in 4.6.

This objective has become popular, particularly within multi-task formations such as [29], as it lends itself well to generalization over superspaces of the parameter space and generally outperforms heuristic methods within tasks. For our purposes however, the major tradeoff made by entropy search as noted in [13] lies in computation time. As we will see in section 4.4, our problem regime of interest leaves us sensitive to acquisition complexity.

$$\alpha(x) = \int \int [H(P_{\text{opt}}) - H(P^{y}_{\text{opt}})] p(y|x, X) dy dx \tag{4.6}$$

We will therefore generally prefer the EI method in this work as a faster-to-evaluate heuristic. Given the current best-performing point observed f^*, we define $\Delta(x) = \mu(x) - f^*$. Then, following derivation given in [16] or [15], we have the following formulation, where ϕ and Φ are the standard normal PDF and CDF respectively and $\sigma = \sqrt{k(x)}$ is the GPR standard deviation.

$$\text{EI}(x) = \max(0, \Delta(x)) + \sigma(x)\phi(\frac{\Delta(x)}{\sigma(x)}) - |\Delta(x)|\Phi(\frac{\Delta(x)}{\sigma(x)}) \tag{4.7}$$

As first order gradient information is readily available for this formulation, we optimize expected improvement over our bounded parameter space by computing the Jacobian and applying L-BFGS-B [34] using the physical parameter bounds given in section 3.1.3. Again, we cite our problem regime's sensitivity to acquisition complexity as compelling reason to avoid longer running multi-start optimization routines that are generally preferable when applied to problems of greater evaluation complexity.

4.1.3 Single-task Results

We begin by addressing our choice over the squared exponential and Matérn kernels. Convergence comparison is shown in figure 4-1 in which we compare performance on many different channel descriptions to optimal values derived via fine grid search with resolution 12 in all dimensions. Convergence of a random grid search is included for comparison, however we note that metrics on number of function evaluations do not necessarily constitute a practical criteria for assessment as it normalizes away the overhead associated with directed search.

We observe only a mild tradeoff in reliability for performance of the Matérn kernel. In fact, we find that they typically converge to statistically indistinguishable optima, indicating that the geometry of the equalization problem is reasonably smooth. We opt to continue with application of the Matérn kernel given its slight mean performance edge. We speculate that this may become all the more compelling a choice when the equalization scheme under study employs more complex post-sampling behavior that leads to sharper problem geometry.

Next we observe relative performance between several optimization methods in figure 4-2 for both the simple and complex equalization schema specified in section 3.1.3 on a smaller set of 27 channels, which were manually tuned for the optimal performance in computing loss. This was necessitated by the rapid increasing in runtime for finer grid searches in the high dimensional problems. In the case of the simpler scheme, we apply a grid search of resolution 6 in each dimension, modeled after the grid search specified in most default configurations of the available COM scripting [22], though with decreased resolution in the CTLE gain parameter. The Bayesian optimization routine runs to a fixed 200 Bayesian samples and the random grid search stops at 400 samples.

Motivated by the implied smoothness of the problem given the relative performance of the squared exponential kernel, we also include Jacobian estimation via L-BFGS-B. Unexpectedly, this outperformed all other optimization methods with some consistency. The primary drawback lies in runtime, as several function evalu-

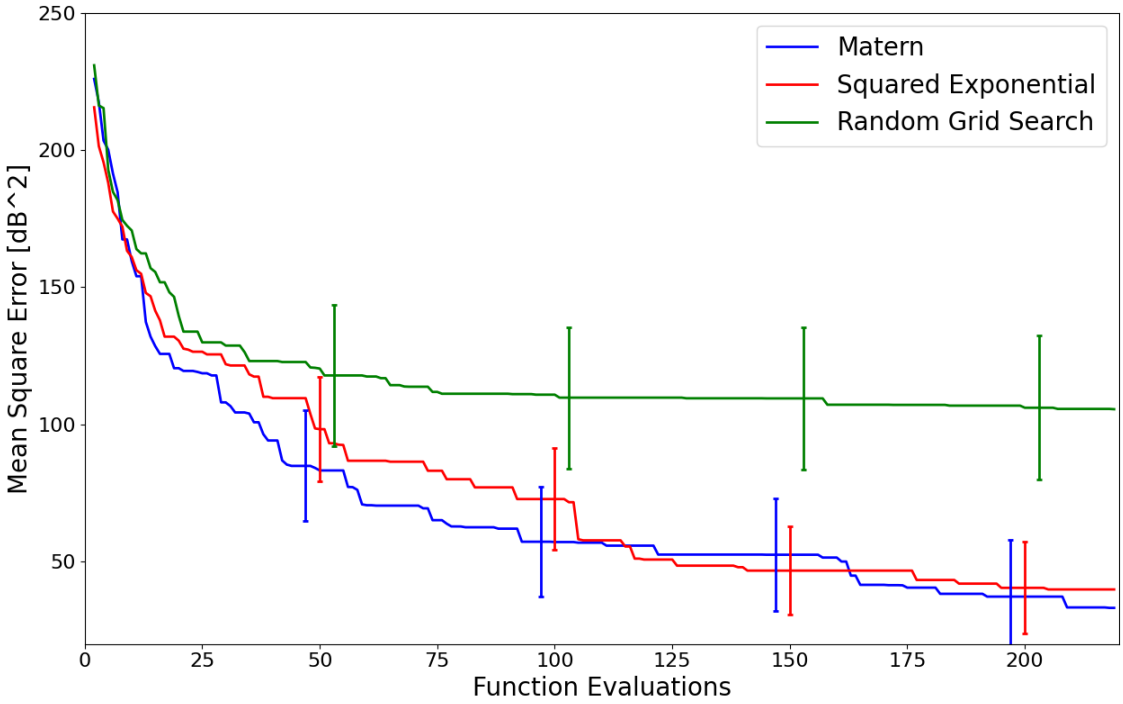


Figure 4-1: Comparison of convergence to optimal values derived by very fine mesh exhaustive grid search. Mean square losses were examined over a set of 40 different channel description, including the Raspberry Pi and Coral board designs discussed in section 6.1 and the simpler equalizer scheme detailed in 3.1.3. Each function evaluation corresponds to the transient analysis of one set of equalizing parameters. Data from the first two function evaluations are omitted. Error bars correspond to the 60% confidence interval.

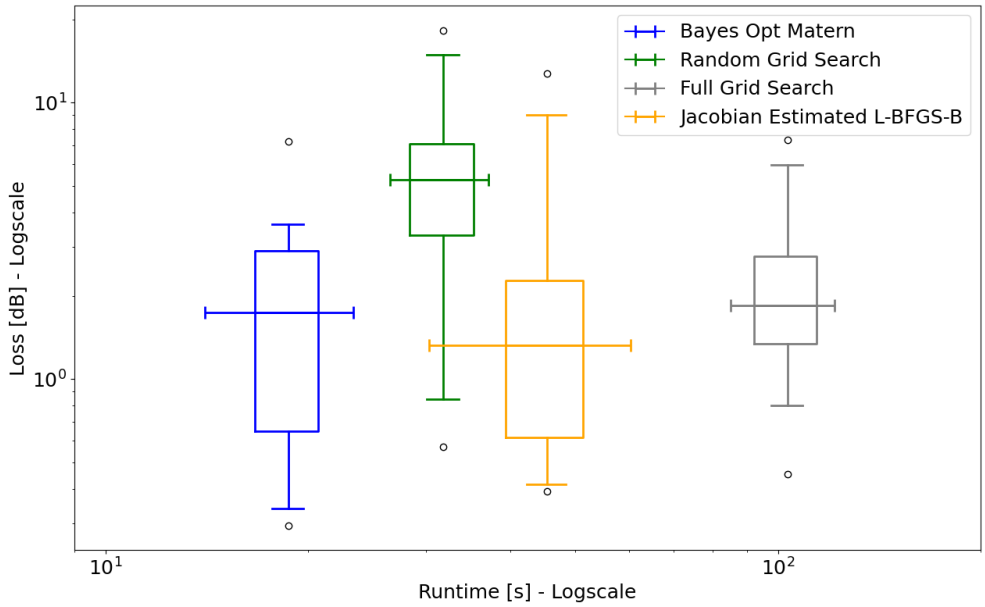
ations are needed for estimation of the Jacobian and several random starting points were also necessary to achieve consistency. Even with the three starting points used, grid search was still generally more consistent and we maintain the supposition that some schema may lead to sharper geometry that confounds gradient methods more readily, which is consistent with a rapid increase in the range of L-BFGS-B on the more complex problem.

The set of problems are optimized in figure 4-2b. Here, the exhaustive grid search has a resolution of only 4 per dimension, as constrained by total runtime. The random grid search runs to 600 samples, but operates on a mesh of resolution 6. Similarly, the Bayesian routine now takes 300 samples, but due to cubic overhead, approaches the random grid search runtime much more closely.

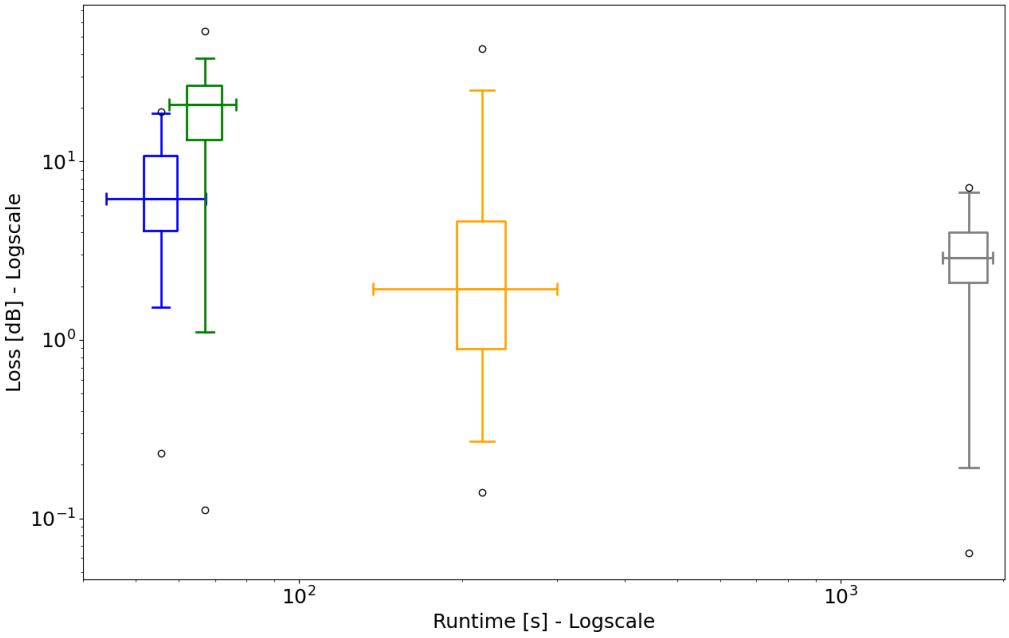
Generally we see Bayesian optimization perform nearly as well as methods of substantially longer runtime. We note that the divide in runtime grows with problem dimensionality, reaching 4 to 20 times faster than L-BFGS-B and grid search respectively, but so too does the gap in performance. For our purposes and likely schema falling between the more and less complex cases presented here, the modest performance loss with Bayesian optimization constitutes a desirable tradeoff, with loss typically 4 to 6 dB worse. Random search performs similarly with some frequency, we observe nearly a 50% improvement in median loss for similar runtime. While not the focus of our methods, the results of L-BFGS-B also demonstrate potential improvement of multi-start gradient approximation methods over grid search for contexts closer to verification.

4.2 Multi-Task Formulations

Observing results presented in section 4.1.3 and the potential multitude of victim channels we wish to analyze over any given analysis run, we aim toward addressing the cold-start problem of our Bayesian optimization routine. In particular, we notice that near the beginning of any such routine, we work with a mostly flat and poorly informed surrogate model. However, in later tasks we expect some relation to pre-



(a) Four Free-Parameter Equalization Problem



(b) Seven Free-Parameter Equalization Problem

Figure 4-2: Four optimization methods applied to either equalizer scheme discussed in section 3.1.3. Whiskers show the 99^{th} percentile with fliers giving the extrema. The medians are extended into horizontal error bounds giving the 95 percentile runtime range. Full grid search uses 6 samples per dimension in (a) and 4 in (b). The Bayesian routine uses 200 and 300 samples respectively while random grid search uses 400 and 600 on grids of resolution 6.

viously completed tasks which could represent an opportunity to begin with some knowledge-discounted surrogate model. We first present the general framework we will employ for this model transfer and then discuss realizing this transfer where prior work fails to serve the problem at-hand efficiently.

4.2.1 Surrogate Relational Structure

As initially discussed in section 2.2, the intrinsic model of coregionalization (ICM) reconceptualizes the underlying GPR to include a concept of task-space, over which any given task may be thought of as a transformation of a more general GPR. Our GPR then operates on a superspace of the input parameter space that will be referred to here as the joint input-task space. We can then express a relationship between points in the input space of different tasks using an extended covariance kernel as a composition of the input-space kernel and a similar kernel encoding the covariational structure between tasks, $k_t(t, t')$.

$$k(x, t, x', t') = k^t(t, t') \otimes k^x(x, x') \tag{4.8}$$

However, we face several issues with such a kernel. For M tasks and N samples per task, the complexity of our model grows with M^3N^3, which is generally unsustainable for large numbers of tasks. Second, we observe that the assumption of separability is quite strong. Further, our problem domain of interest lies in a noise-free regime, over which this assumption of kernel separability implies a cancellation of inter-task transfer for some measurement of $k(t, t')$ computed from a (generally sparse) sampling in both tasks [4] [32]. This is a product of this assumption entailing an isotropic covariational structure between tasks [25].

This leaves two potential options. Either we may form a more complex model injecting anisotropic measurements of the relational structure or we might attempt to glean relational information prior to any sampling in our new task. Again we find the former to be likely inappropriate for the problem of interest due to computational overhead. The latter therefore comprises our focus, wherein we aim to apply heuristic

analysis of victim channels to glean a correlation structure.

4.3 Relational Assessment and Task Selection

Here we present the bulk of novel methodology. Our aim is to formulate a sequential process of optimization routines that exhibits reduced complexity associated with performing acquisition on the joint input-task space. In particular we aim to do this by examining each task sequentially in the context of previously completed tasks, choosing to transfer the generalized surrogate model from only the most closely related tasks.

4.3.1 Heuristics

Our primary method of interest in examining the relationship between tasks lies in a heuristic derived from observation regarding the problem of equalizer parameter setting. Namely, we observe that the problem itself can be viewed as a projection of the inverse channel transfer function onto the parameterized equalization transfer functions. We therefore suppose that a reasonable approximation to task similarity may be derived from similarity of the forward, unequalized transformation of the channel. Under this view, we use the normalized cross-correlation of the worst-case phase alignment of the SBRs, h and g, subject to crosstalk. We note that assuming phase alignment does not generally decorrelate on electrical length, as differences in reflection delay still push leading and lagging symbols out of alignment.

$$k(t,t') \propto \frac{1}{N\sigma_h\sigma_g} \max_n (h * g)[n] = \frac{1}{N\sigma_h\sigma_g} \max_n \sum_{k\in\mathcal{Z}} (h^*[k] - \mu_h)(g[k+n] - \mu_g) \quad (4.9)$$

However, this incurs the up-front cost of essentially performing $N+1$ function evaluations to relate a new task to N previous tasks, plus the $\mathrm{O}(n^2)$ cross-correlation operation reduced to $\mathrm{O}(n \log n)$ via conversion to frequency space. In section 4.3.2 we therefore examine methods to reduce the number of previous tasks we must relate to in order to find a strong candidate for our prior surrogate. We also observe that the

relationship this yields is fairly proximal. For this reason, we also briefly investigate a random forest model that might be pre-trained on data across topologies to infer actual task correlation. While this has not been implemented in our COM analysis flow, we will give preliminary results for such a model using as input this heuristic, as well as ISI, crosstalk, jitter, and signaling rate as features taken from the assessment of the unequalized channels.

4.3.2 Clustering

Here we explore a novel type of clustering on tasks to serve the rapid identification of likely similar problems. While determining $k(t, t')$ through the use of clustering input data for machine learning hyperparameter optimization problems has been explored [1], we instead attempt to use our previously derived $k(t, t')$ to sparsify relations to a new task. We interpret $k(t, t')$ as a sparse, weighted adjacency matrix representing an undirected similarity graph.

Our principle concern with the application of graphical clustering methods lies in our definition of connectivity. In an intuitive sense, this means we must define the levels of correlation for which we should consider tasks inherently similar. To this end, we find a thresholding of correlation to be a natural inclination. However, we also note that strong anti-correlation should be similarly useful in an information-theoretic sense. Therefore an initial binary connectivity measure might take the form of a symmetric threshold for some $0 \leq b \leq 1$, using which we form the connectivity matrix C.

$$C(i,j) = \begin{cases} 1 & |k(i,j)| \geq b \\ 0 & \text{else} \end{cases} \tag{4.10}$$

There is a deeper issue to address with this structure, however. In particular, it encourages an implicit assumption on the transitivity of similarity. Unfortunately, only strong measures of correlation can be shown to exhibit a sense of transitivity in general [17]. To give one potential bound demonstrating this relationship, we observe the transitivity inequalities derived via the angular separation of the underlying vec-

torized signals in [18]. In terms of SBRs f, g, and h where f may represent a new task, g a central task in a cluster, and h a previously completed task connected to g, we have the following.

$$\rho_{f,g}\rho_{g,h} - \sqrt{1-\rho_{f,g}^2}\sqrt{1-\rho_{g,h}^2} \leq \rho_{f,h} \leq \rho_{f,g}\rho_{g,h} + \sqrt{1-\rho_{f,g}^2}\sqrt{1-\rho_{g,h}^2} \tag{4.11}$$

This implies that if we wish the threshold to be great enough such that transitive connectivity is given, we are required to set our threshold to the undesirable value of 1. This all but guarantees an edgeless graph, however. Examining the converse, in which we hope to discount connectivity to h given connection of g and h and some poor relation between f and g initially gives similarly problematic results. However, we can partially resolve these issues analytically by introducing a slackness factor $\gamma > 1$ for which we expect some actual observation of connectivity to satisfy $\rho > \gamma b$ and accept that we may find a weaker transitive conclusion that still satisfies the bound b. This prevents us from applying this analysis to arbitrary length walks of the similarity graph, but does give rigorous and convenient means of examining transitivity over short walks.

In effect, examining these transitivity relations gives a means for sparsifying relational statements over a shallow clustering graph on tasks built via existing $k(t, t')$. The setting of b boils down to a tradeoff between sparsity and overhead and is therefore highly problem dependent. On The equalizer problem of interest, we will generally apply $b = 0.9$, given the distribution of problem correlation that will be examined in figure 4-3.

We then propose the greedy clustering method as follows. First, we compute $C(i, j)$ as shown. Then, we order the vertices of the graph by their degree. We then check each vertex by descending degree attempting to form a new connection. For each vertex examined, we decide to discount connection to future vertices using the given transitivity bounds. When we do find a vertex our current task connects to with correlation at least b, we greedily take this task as the best to transfer.

By preferentially selecting tasks with the greatest degree, the goal is first to favor

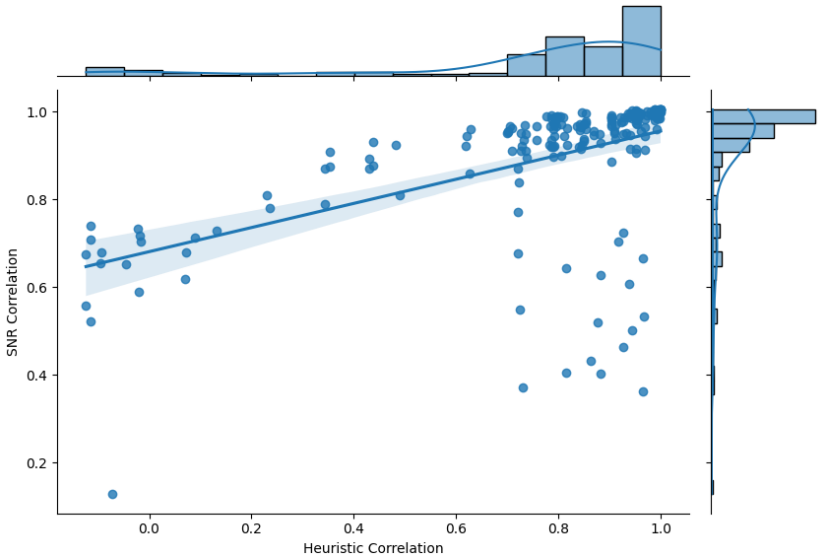
tasks which relate well to many other tasks in their cohort. Second, we aim to favor larger connected clusters that will then allow us to discount more tasks per evaluation in the future.

4.4 Results on Equalizer Optimization

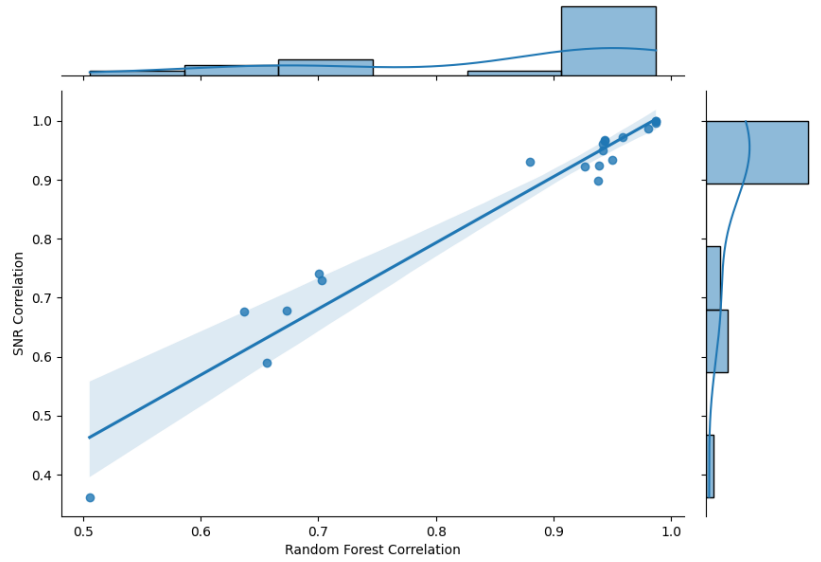
We begin with examination of our ability to assess the similarity of optimization problems. We measure the actual relation between two problems by the directly computed correlation between their outputs M given a set of common inputs uniformly sampled on the equalizer parameter space. In our assessment here, 200 samples were taken independently for each distinct pair of problems in a 20 task cohort. The tasks are derived from the Raspberry Pi design discussed in section 6.1 as it presents a large number of interest channels in a single design. Notably, this dataset produces longer task evaluation times than other boards tested, as the associated transfer matrix is not particularly sparse. The heuristic given in equation 4.9 was then computed on the unequalized channels to form figure 4-3a. The relationship is clearly not one-to-one and this metric appears to miss a significant trend. However, considering the distribution in probability mass, the majority of task pairs lie along the shown regression line and we find it to be a somewhat reasonable metric for sorting tasks by similarity.

The described random forest model is able to do substantially better. Its performance on a randomly selected 10% holdout set is shown in figure 4-3b where we see an r^2 of 0.95. This may provide a much more reliable method for task correlation estimation, however a larger data set that requires it to generalize across design topologies is necessary before it can be fully assessed and implemented. Assessment in this work relies only on the given heuristic.

The runtime distribution for several methods on the same set of tasks is shown in figure 4-4. In (a), we see ICM slow substantially due to increasing acquisition complexity as expected. With every task, 120 additional samples are included in the GP, which requires a cubic investment in computing the Cholesky decomposition proxy for $K((x,t),(x',t'))^{-1}$. After only a few tasks, this accumulating difficulty slows



(a) Heuristic



(b) Random forest model

Figure 4-3: Examination of heuristic and random forest regression for task correlation. Notably for this problem, most probability mass lies near the upper right, generally giving an abundance of related tasks, which are in turn frequently found even by the poor regression given in (a). Regression lines are shown with r^2 values of 0.32 and 0.95 respectively.

Method	# Samples	Mean Loss	Sample Std Dev in Loss	Total Runtime
Full ICM	120	-0.32 dB	9.42 dB	2196 s
Clustering	200	+3.202 dB	6.70 dB	822 s
Single Task	200	-	-	795 s
Best Task	120	-0.09 dB	5.09 dB	514 s
Clustering	120	-0.06 dB	6.72 dB	473 s
Single Task	120	-1.01 dB	6.22 dB	452 s

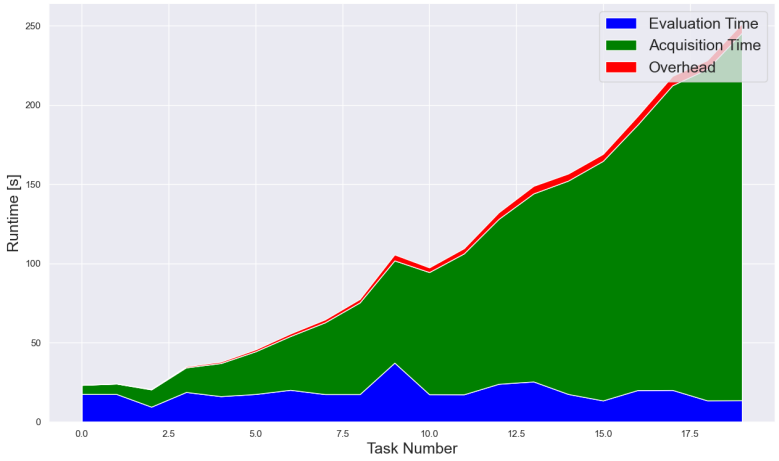
Table 4.1: Performance summaries of each of the optimization methods as well as a single-task run using 120 Bayesian samples. Each is compared against a single task run taking 200 Bayesian samples per task. Loss is computed per task.

the method beyond single-task optimization. Herein lies the need for a thinning of the transferred information.

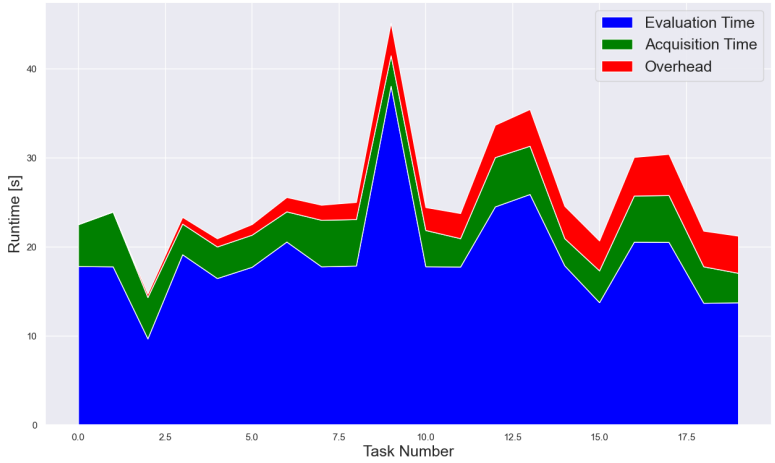
In (b), we see the method that makes all comparisons to previous tasks and selects the best. The selection of only the best task limits acquisition difficulty to slightly beyond typical. However, as we must still make every comparison to select the best, we still incur a linearly increasing overhead as we perform two function evaluations in computing the heuristic. The clustering method presented here incurs no such increase in overhead. This will not generally be true when the threshold b is set larger than we typically expect the correlation between tasks. As we observe the probability mass in figure 4-3a concentrated in the top right, above our selected b of 0.9, we typically find a sufficiently related task among the first several we check. An exception to this occurs in task number 9, in which we don't find any prior tasks that are sufficiently related. In this case, we check every vertex before selecting the best one and incur overhead proportional to the best task method. While the set of tasks to be optimized over the analysis of a board appears too small for this to make a substantial difference, these results are encouraging enough to suggest that maintaining a task cohort over the entire design process could be beneficial, over which this method may give more meaningful results.

Table 4.1, shows a summary of the methods' performance on the same set of tasks. The best task and clustering methods using 120 Bayesian samples performed nearly as well as the the single task run using 200, indicating a viable method for reducing runtime. Between the two, we see that their performance is quite similar,

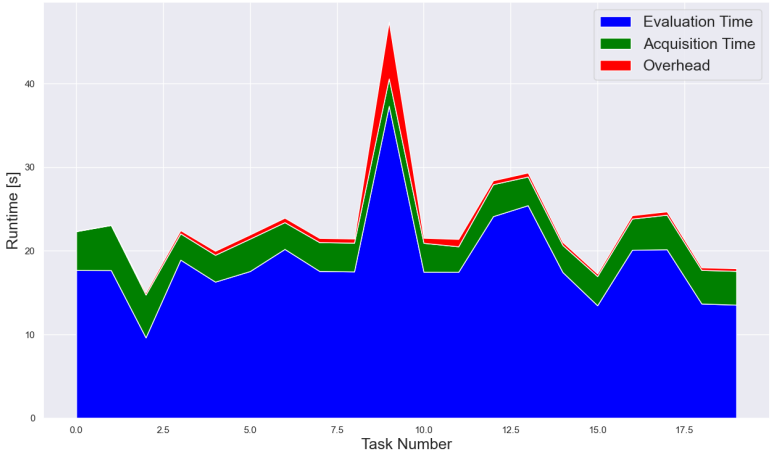
and we appear to be primarily trading off runtime for consistency, which tracks with the observation that taking a vertex greedily may mean missing a closer connection. We also see the clustering method using 200 samples outperform the single-task formulation for only a moderate increase in runtime, which may be a desirable tradeoff for accuracy-focused assessment. The full ICM formulation performs well with access to comparatively thorough exploratory information, but remains nonviable. Interestingly, it also appears to be the least consistent, which might indicate that the ICM kernel separability assumption fails in regions of good performance on this problem.



(a) All Tasks Transferred



(b) All Tasks Related, Best Transferred



(c) Clustering Method

Figure 4-4: Runtimes on successive tasks in a cohort for each of (a) the full ICM model, (b) transfer of the best task after comparing all, and (c) the clustering method presented here. All take 200 evaluations on the first two tasks, then drop to 120. Function evaluations inherently vary in difficulty due to the number of crosstalk terms included, but are presented in the same order to each method.

Chapter 5

Channel Simulation

Given channel specification and optimized parameters, we now reach our ultimate goal of deriving operating margin. In this section, we aim to apply the results of our analysis thus far toward the derivation of concrete performance metrics. In particular, we now explore the topic of channel simulation as well as the assumptions under which we may extract COM efficiently.

5.1 Modeling Assumptions

First, we approach the issue of developing a generative channel model capturing the statistical nature of the various noise sources. In general, two distinct types of analysis stem from assumptions made on the form of the channel. Primarily, we may either consider a channel linear and time invariant (LTI) or belonging to a more general class of system. Nonlinear devices in drivers and receivers and single-ended signaling are both reasonably common situations that cause non-LTI behavior [8].

While COM is a generally valid approach for specifying both LTI and non-LTI channels, our analysis thus far has included two assumptions that focus on LTI behavior. First, the easily modified form of the transfer function presented in section 3.2.1 and second, the more ingrained assumption of linearity in the superposition of noise terms and time invariance in the computation of ISI and multi-symbol crosstalk in section 3.2.2. In particular, we note that the latter assumption is critical to rapid

computation of the SBR and therefore to equalizer optimization. The novel methods presented in section 4.2 are all the more relevant to more complex non-linear treatments of the pessimistic SBR such as in [8], however these generally move our analysis out of the rapid prototyping metrics we ultimately hope to achieve. Therefore the main focus of our interest will be on LTI regimes.

5.2 Eye Characteristics

The focus of channel simulation is on the production and analysis of an eye diagram such as presented in figure 5-1. Eye closure represents a commonly accepted validation metric [20], though direct computation of an eye diagram is fairly expensive. We commonly wish to validate to data error rate (DER) thresholds on the order of 10^{-12}, requiring exhaustive transient simulation. However, as we target pre-validation assessment and generally take the channel of interest to be LTI, we make the simplifying step of examining only a cross section at the height of the eye.

As shown, this fails to directly consider horizontal closure due to jitter and associated zero-crossing variance as well as diagonal closure owing to slew rate. Partially for this reason, we recall that our computation of noise due to jitter has been severely pessemistic in the sense that we compute vertical variation using the transient slope near the zero-crossing rather than near the height. For most cases of reasonable horizontal closure, we presume that a good choice of sampling phase setting policy admits analysis through this vertical slice. Computing the PDF model of this slice is then the focus for the remainder of our work in channel simulation.

5.3 LTI Channel Simulation

Given the assumption that our channels under scrutiny are linear and time invariant, we may bypass much of the difficulty inherent to general channel simulation. In particular, we may reuse many of the techniques presented in computation of the M in section 3.2.2 and directly form a PDF on the magnitude of noise via the convolution

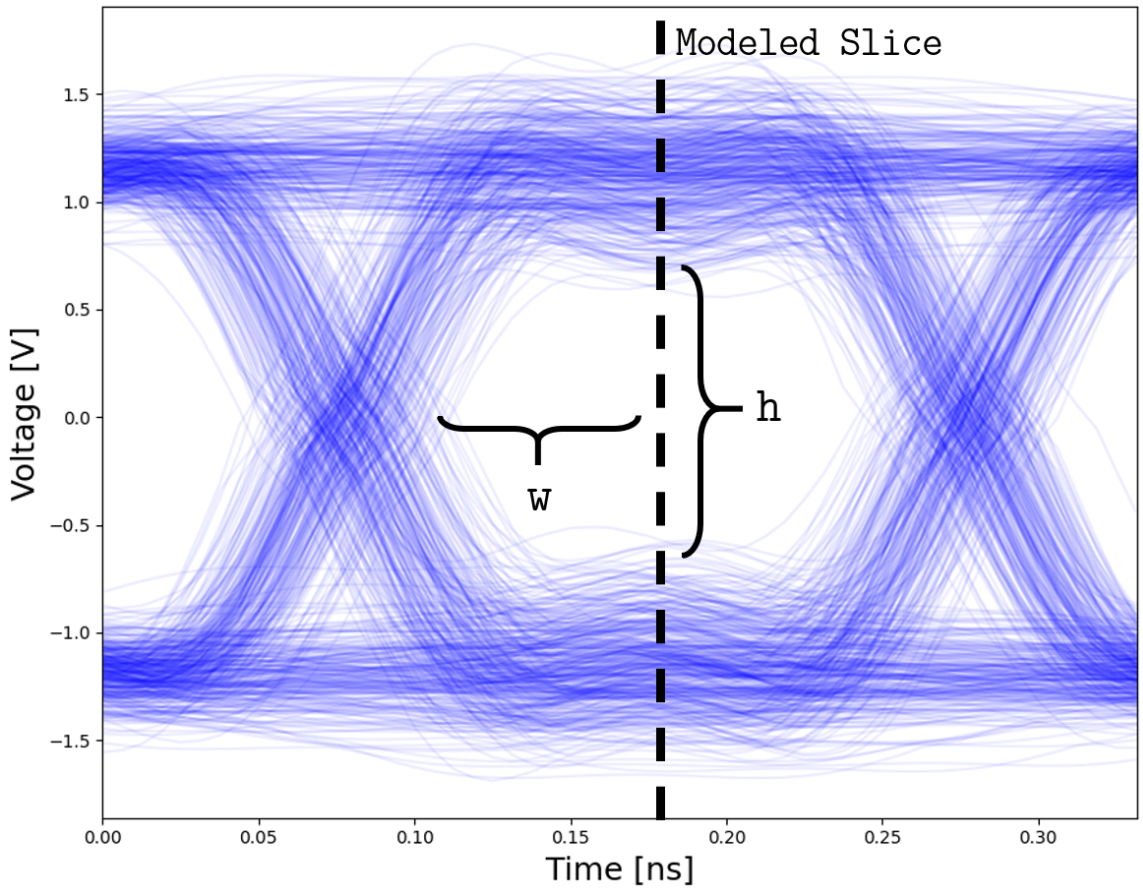


Figure 5-1: A sample eye diagram produced from an optimally equalized USB channel on the Coral board discussed in section 6.1. The slice to be modeled as a PDF is dotted with the target metric of eye height labeled as h. Also labeled is the eye width resulting from variation in zero crossing.

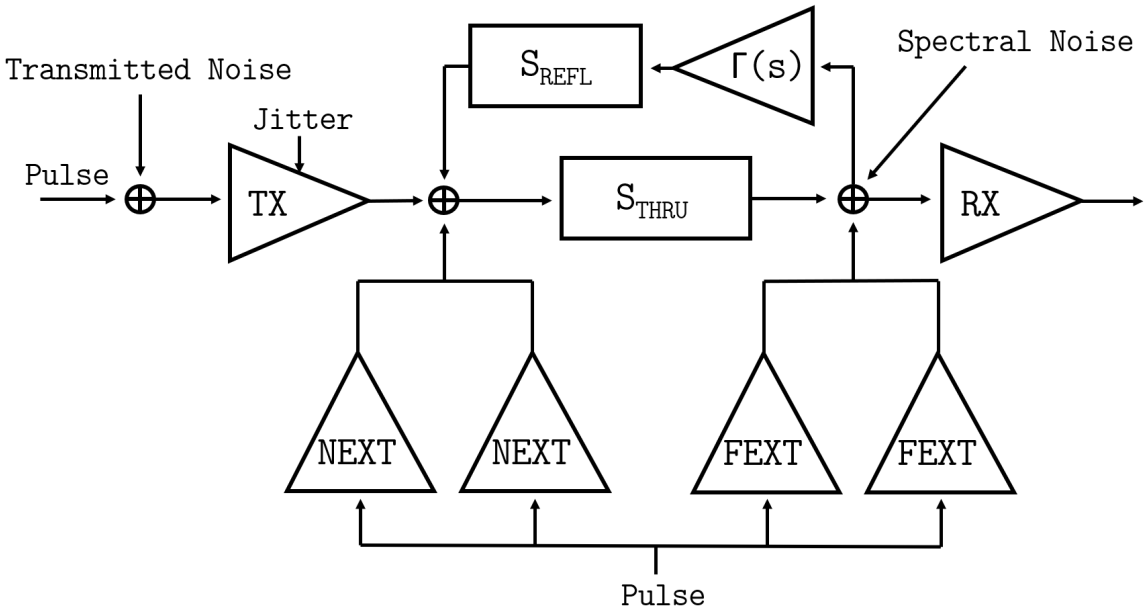


Figure 5-2: Review of LTI channel formulation. Multi-symbol terms including ISI, crosstalk, and jitter are brought about both via time-domain smearing and reflective feedback, wherein we use linearity and time invariance to describe the resulting noise as a superposition of time shifted waveforms. Note that reflective behavior as well as the disparity in NEXT and FEXT terms are rolled into the transfer matrix and S parameter model definitions respectively, flattening the system we construct in practice.

of all relevent terms. We then extract COM metrics directly from the associated CDF.

5.3.1 Computing The Noise PDF

In figure 5-2, we observe that the signal present prior to sampling is a linear superposition of several "deterministic" noise sources of various delay including ISI and crosstalk, as well as several random sources such as random jitter and spectral noise. While the former is commonly referred to as deterministic, we do note that the resulting noise term is effectively a random variable parameterized by the inducing signal. For instance, the sign of any particular ISI term in an NRZ channel is determined by the originating symbol. Therefore, given that we have an unbiased and stationary encoding, we might consider this term to be a Bernoulli process with outcomes of $\pm\Delta_{ISI}(n)$ with $\Delta_{ISI}(n)$ computed from the SBR via the process presented in section

3.2.2. In general, we find ISI, crosstalk, and deterministic jitter as follows.

$$I_n = \frac{\delta(i - \Delta_{ISI}(n))}{2} + \frac{\delta(i + \Delta_{ISI}(n))}{2} \qquad J_n = \frac{\delta(i - A_{dd}\mu_n)}{2} + \frac{\delta(i + A_{dd}\mu_n)}{2}$$
$$X_{k,n} = \frac{\delta(i - \Delta_{XT,k}(n))}{2} + \frac{\delta(i + \Delta_{XT,k}(n))}{2}$$

Proceeding as in [22], we then formulate the overall deterministic noise in the channel as a summation over each such random variable. Further, adding in sources of random noise, we augment this sum with three Gaussian forms of mean zero parameterized by variance given as the RMS of the source. We take μ_n as the derivative series of the SBR as computed in 3.2.2.

$$RJ_n = \frac{1}{\sqrt{2\pi\sigma_{RJ}^2\mu_n^2}}\exp(-\frac{i^2}{2\sigma_{RJ}^2\mu_n^2}) \qquad S = \frac{1}{\sqrt{2\pi\sigma_S^2}}\exp(-\frac{i^2}{2\sigma_S^2})$$
$$T = \frac{1}{\sqrt{2\pi\sigma_{TX}^2}}\exp(-\frac{i^2}{2\sigma_{TX}^2})$$
$$I = \sum_{n=0}^{N} I_n + \sum_{n=0}^{N} J_n + \sum_{k=0}^{K}\sum_{n=0}^{N} I_n + \sum_{n=0}^{N} RJ_n + S + T \tag{5.1}$$

As summation of random variables is isomorphic to convolution of their PDFs, and further a convolution of Gaussians is Gaussian, it is trivial to compute a closed form representation of the aggregate PDF as in [7] reproduced below.

$$\sigma_G^2 = \sigma_{TX}^2 + \sigma_S^2 + \sum_{n=0}^{N} \sigma_{RJ}^2\mu_n^2$$
$$p(I = i) = \frac{1}{N^3K\sqrt{2\pi\sigma_G^2}} \sum_{q=0}^{N^3K-1} \exp(-\frac{i - i_q}{2\sigma_G^2})$$

However, in practice, we wish to provide flexibility over the form of each noise term. For instance, we may wish to model the smearing of each deterministic source or drop-in a more specific profile of spectral or transmitted noise according to available specification. Therefore, we opt instead to compute numerically the resulting PDF.

In doing so, we notice that for N symbols appearing in the crosstalk, ISI, and jitter terms and K aggressing channels, equation 5.1 implies $N(K+3)+2$ convolutions, each

of which are $O(n^2)$ in the resolution of our probability space when performed directly. Instead, we operate in frequency space wherein each corresponding multiplication is performed in $O(n)$ and a single inverse Fourier transform is performed at the end in $O(n \log n)$. We then find our PDF given as below, in the case of the same deterministic and random forms given prior.

$$\begin{aligned}\mathcal{F}[p_{\text{det}}(I=i)] =& \prod_{n=0}^{N}(e^{2j\pi\Delta_{ISI}(n)\omega} + e^{-2j\pi\Delta_{ISI}(n)\omega})(e^{2j\pi A_{dd}\mu_n\omega} + e^{-2j\pi A_{dd}\mu_n\omega}) \\ & \prod_{n=0}^{N}\prod_{k=0}^{K}(e^{2j\pi\Delta_{XT,k}(n)\omega} + e^{-2j\pi\Delta_{XT,k}(n)\omega}) \\ p(I=i) =& \mathcal{F}^{-1}[\mathcal{F}[p_{\text{det}}(I=i)]\exp(-2\omega^2\pi^2(\sigma_{TX}^2 + \sigma_S^2 + \sum_{n=0}^{N}\sigma_{RJ}^2\mu_n^2))]\end{aligned}$$

5.3.2 Extracting COM

Finally, with an explicit PDF representing the vertical slice of an approximate eye diagram at its peak, we may extract the amplitude of noise corresponding to a desired data error rate. Specifically, we find A_n such that,

$$\text{DER} = \sum_{i=-A_n}^{A_n} p(I=i)$$

For most reasonable PDFs governing noise terms and small DER, we may leverage both symmetry and the size of the tail to compute this more efficiently as follows.

$$\text{DER} = 1 - 2\sum_{i=A_n}^{\infty} p(I=i)$$

Finally, taking A_s as produced in 3.2.2 for the equalizer-optimized channel, we may compute COM following the nominal formula as below. However we note that in some designs and channel types, we may not be given a specific target for data error rate. In this instance, we may instead compute the projected data error rate such that we expect a reasonable COM value. The summation limit used below corresponds to a

COM value of around 4.1, which is generally considered acceptable [22][19].

$$\mathrm{COM} = 20\log_{10}(\frac{A_s}{A_n})$$
$$\mathrm{DER}_{\mathrm{proj}} = \sum_{i=-\infty}^{0.625A_s} p(I=i)$$

In practice, we report these metrics as well as contributing metrics that might aid in channel analysis. In particular, these include RMS measurements for each contributing noise term as well as signal gain and optimal equalizer parameters. Also available for reporting are channel reflectivity metrics as described in section 3.2.1 and confidence bounds on the optimality of equalization computed from entropy in the underlying GPR as described in section 4.1.2.

Chapter 6

Results and Discussion

6.1 Coral Board and RPI Case Studies

Here we walk through our analysis pipeline examining two different PCB designs as well as a comparison of our results to those generated using publicly available COM scripts. While no distributions considered normative for our application exist, we conduct our analysis using derivative of Annex 93a [10] notably also distributed alongside the Serial Link Designer in MATLAB R2022a.

6.1.1 Designs Under Test

The first design we follow is a simple subsection of the publicly available Coral Board. We examine the composition of two channels, USB1 and USB2. Each of these are two-lane, uni-directional channels operating at a signaling rate of 10Gbd along with a single USB2.0 compliant differential pair which will be included, but not a target for analysis. Their layouts are depicted in figure 6-1 via PowerSI.

Next, we examine a larger subsection of a Raspberry Pi design composed of several channels of differing signal rates, the top layer of which is shown in figure 6-2. This subsection consists of two USB 3.1 channels, an HDMI 1.3a channel, a CSI-2 channel, and a much lower frequency DSI channel. We note that a tighter parallel grouping and parallel layer transitions indicates crosstalk will likely be a much more substantial

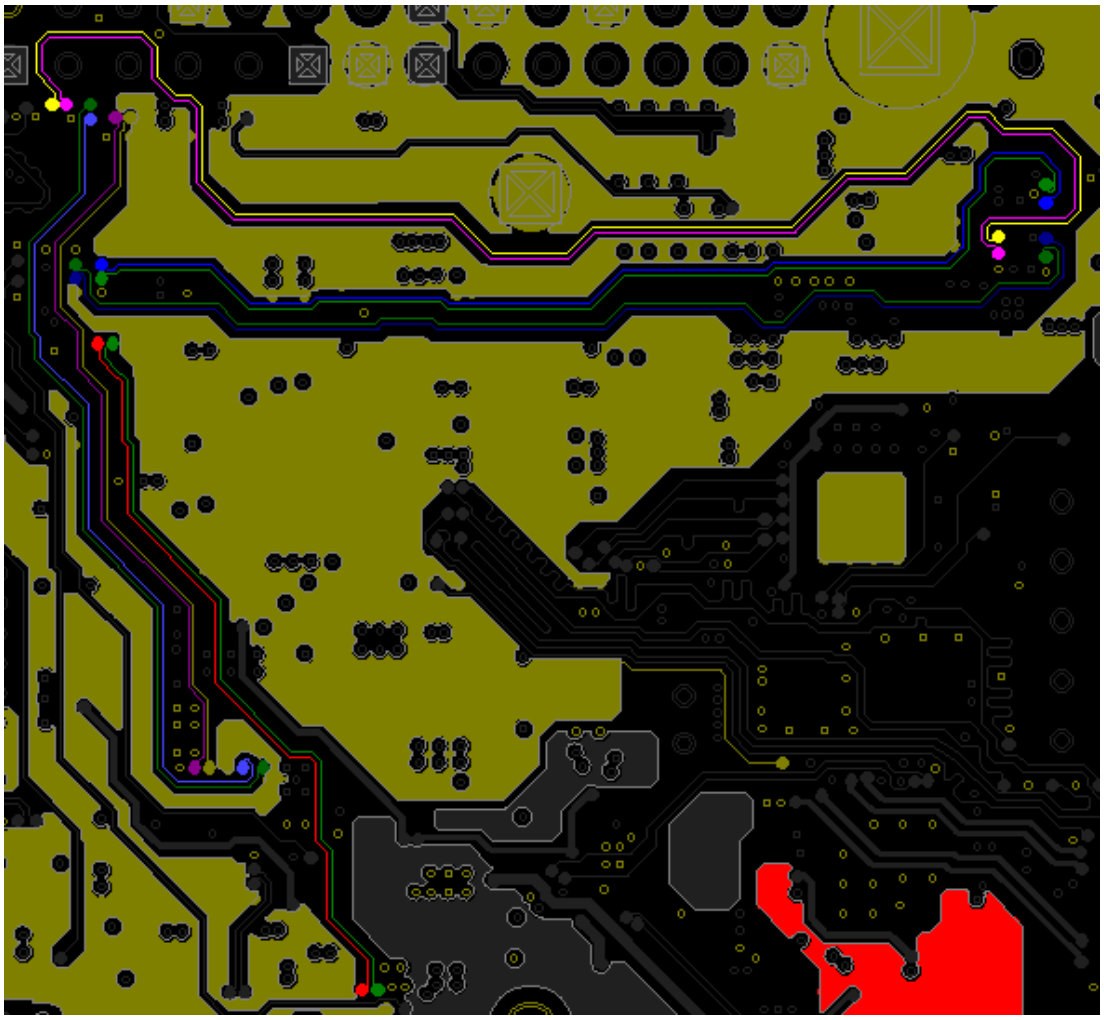


Figure 6-1: Two USB 3.1 channels under test on the Coral Board. USB1 is shown vertically with the USB 2.0 compliant diff pair on the right. USB2 is shown horizontally with its USB 2.0 diff pair on top. The shaded region in dark yellow is ground metal.

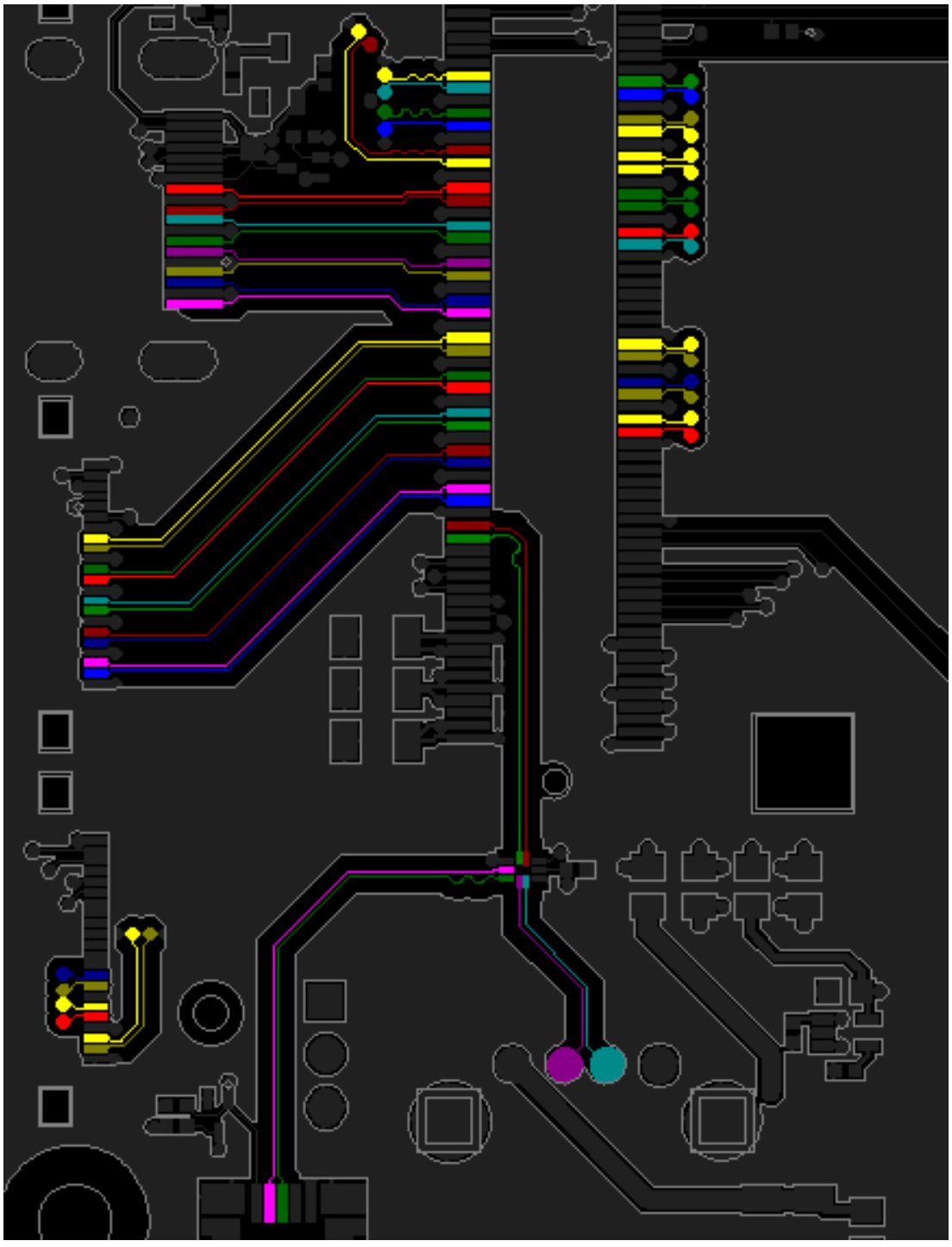


Figure 6-2: The top signal layer of the Raspberry Pi design under test. In contrast to the Coral board layout, we observe a much tighter grouping and greater total parallel length for most channels. We also observe tight groupings of parallel layer transitions to consider.

consideration and that further, there are many more crosstalk terms to evaluate. We therefore have a greater complexity of transient analysis and many more channels to optimize.

6.1.2 Comparison of Metrics

Beginning our analysis, we extract the S parameters as computed by the methodology specified here to use as input to the public COM script. We also apply the same equalization schema and take the same set of values for inputs described in sections 3.2.1 and 3.2.2 such as spectral noise density η, nominal terminal impedance, etc. The results of analysis by either method of the Coral board are shown in table 6.1 along with several intermediate results. Also shown are the partial results from applying a different workflow using Cadence's Topology Explorer (Top Xp) using a flow in the same vein as [33], though without direct IBIS model specification. This tooling affords similar measurement achieved using a different methodology based on directly on eye analysis.

We observe that there is fairly significant structural difference between the implementations. As this is a physically realized and functional board, we should generally expect COM of greater than 4 dB and substantially so considering the channel simplicity. The largest difference between the methods presents are relative optimism of the direct methods employed by Top Xp. This is not necessarily a result of inaccuracy, rather we adopt many of the same pessimistic assumptions in package modeling and per-source noise assessment as those of the reference script. Yet in comparison, we find our results still more pessimistic. The substantial obvious cause for this discrepancy is likely in our assumptions of rise-time discussed in sections 3.1.1 and 3.1.2, which are considerably more limiting, though convenient for automated modeling. We do find another interesting discrepancy, however. The ideal equalizer gain we find via our optimization routine appears to be pathologically greater than that of the reference script. Considering our agreement with Top Xp, this may be an artifact of choosing samples off-grid, wherein we may compensate with finer adjustments on emphasis.

Diff Pair	Metric	Our Methods	Reference Script	Top Xp
RX_N/P	COM	8.06 dB	9.62 dB	16.21 dB
	Signal After EQ	462 mV	247 mV	501 mV
	Interference at BER	183 mV	81 mV	(78 mV)
	Peak ISI	104 mV	68 mV	-
TX_N/P	COM	12.15 dB	12.21 dB	17.82 dB
	Signal After EQ	390 mV	206 mV	436 mV
	Interference at BER	96 mV	51 mV	(56 mV)
	Peak ISI	46 mV	38 mV	-

Table 6.1: Metrics derived on either of the primary USB1 channels on the Coral Board. Results presented here correspond to the direct, fine log-linear S parameter extraction. Parenthetical values are calculated rather than reported.

Overall, we find our methods to fall within reasonable agreement of the reference. While this generally presents the relative behavior we observe, we note that there is no ground truth to compare against barring measurement validation. Given the opportunity, this would be a desirable next step in performance assessment. Additionally, publicly available designs that are known to fail to demonstrate negative assessment will be desirable for further analysis.

Presented in table 6.2 are the results of our methodology as applied to the Raspberry Pi for further consideration. These results appear comparatively optimistic, though we generally expect these channels to pass as well. We do note an uncharacteristically low-scoring $DSI1\_2$ channel, notably this appears to be due to a fairly poor equalizer setting and in consequent runs the channel's score was only moderately diminished compared to its neighbors. This may be symbolic of some consistency trade-off inherent in failing to perform a full grid sweep. Whereas our crosstalk metrics agreed fairly well with the reference script, we do see unexpectedly low crosstalk throughout the Raspberry Pi design given the long parallel length, close spacing, and nearby layer transitioning. Barring further measurement validation however, it is difficult to conclude currently whether the unexpectedly insignificant crosstalk is accurate, or one source of optimism on this topology. We do note that generally very few crosstalk terms were sparsified out in this analysis with at least 10 aggressing channels included in the simulation of every differential pair.

Diff Pair	COM	Gain	Interference at BER	Peak ISI	Peak Crosstalk
HDMI_D1	12.05 dB	0.211	52.7 mV	39.8 mV	9.3 mV
HDMI_D2	12.06 dB	0.211	52.6 mV	40.2 mV	9.1 mV
CAM1_1	16.25 dB	0.577	88.9 mV	70.1 mV	7.0 mV
CAM1_2	10.35 dB	0.117	35.6 mV	28.8 mV	2.4 mV
CAM1_3	16.25 dB	0.577	88.9 mV	65.7 mV	10.1 mV
DSI1_0	9.12 dB	0.459	160.7 mV	103.0 mV	15.9 mV
DSI1_1	8.43 dB	0.291	110.3mV	63.2 mV	12.6 mV
DSI1_2	3.85 dB	0.141	90.5 mV	48.2 mV	20.7 mV
USB_B_1	14.96 dB	0.290	51.8 mV	50.6 mV	0.2 mV
USB_B_2	13.70 dB	0.546	112.7 mV	86.3 mV	8.9 mV

Table 6.2: The results of our methodology on the variety of signal channels present in the Raspberry Pi.

6.2 Future Work

Overall, we find the methods described in chapter 4 to provide substantial improvement in equalizer optimization performance. However, further work must be undertaken to assess the accuracy of our overall analysis flow. Generally we find the logical operation of our methods not to differ terribly from prior work and our results to generally agree with design intuition, barring some potentially concerning apparent optimism. In this vein, the methodology presented appears to be a reasonable basis for further work in rapid pre-verification design assessment.

There are many potential directions for future work beyond deeper accuracy assessment. The most immediate lies in the aforementioned application to automated design. While COM represents an attractive target for reinforcement learning applications, current implementations have largely required practitioner involvement in at least the topological level and basic equalizer optimization routines bog down performance. These factors are substantial obstacles for any type of value function assessment, but particularly so for techniques such as Monte Carlo tree search which rely on rapid design rollout and assessment. The runtime on complex designs presented here may remain a lessened obstacle to this line of work, however we consider a few potential improvements to this end.

The work presented here aims to tackle the cold-start problem in our Bayesian

optimization routine by transferring learning between channels over the course of an analysis routine. However, the lessened computational overhead of locating related channels may provide an opportunity for transfer across designs in a cohort, wherein we may expect previous iterations of the same channel to be very informative for current designs. Further, while little benefit was found from logically partitioned S parameter extractions in the development of these methods due to extraction overhead, we may expect severely reduced complexity in crosstalk assessment from logical partitioning of our entire analysis routine. However, this necessitates substantially deeper analysis on COM sensitivity to choose reasonable subdivisions.

In terms of the work presented on multi-tasking in Bayesian optimization, the primary concern we are left with is in applicability. The problem of equalizer parameter setting conditioned on SBR assessment precipitates the methods presented because it lies in a difficult runtime gap. The problem is poorly served by simpler Jacobian estimation methods, but remains sensitive to methods of substantial overhead. For this reason, the methods presented here serve the problem well, but may not generalize desirably to problems of different complexity regimes. However, it does appear likely that there may be a variety of similar optimization tasks over short-lived physical simulations that may benefit from similar methodology.

Bibliography

[1] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 2003.

[2] L. Ben-Artsi and R. Mellitz. Package and interconnect com impact analysis and what-ifs. 2013.

[3] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kegl. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 2011.

[4] B. Chai and Williams. Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, 2008.

[5] Po-Wei Chiu, Chun-Hung Lin, and Hong Shi. Highly efficient and accurate package electrical modeling automation. In *2016 IEEE 18th Electronics Packaging Technology Conference (EPTC)*, pages 289–291, 2016.

[6] C. P. Coelho, J. Phillips, and L. M. Silveira. A convex programming approach for generating guaranteed passive approximations to tabulated frequency data. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2004.

[7] Anritsu Company. Measuring channel operating margin, 2016.

[8] M. A. Dolatsara, J. A. Hejase, W. D. Becker, J. Kim, S. K. Lim, and M. Swaminathan. Worst-case eye analysis of high-speed channels based on bayesian optimization. *IEEE Transactions on Electromagnetic Compatability*, 63, 2021.

[9] S. Grivet-Talocia. Package macromodeling via time-domain vector fitting. *IEEE Microwave and Wireless Components Letters*, 13(11):472–474, 2003.

[10] 802.3 WG Ethernet Working Group and D. Law. Ieee standard for ethernet: Amendment 2, physical layer specifications and management parameters for 100 gb/s operation over backplanes and copper cables, 2014.

[11] B. Gustavsen. Fast passivity enforcement for s-parameter models by perturbation of residue matrix eigenvalues. *IEEE Transactions on Advanced Packaging*, 2010.

[12] B. Gustavsen. Fast passivity assessment for s-parameter rational models via a half-size test matrix. *IEEE Transaction on Microwave Theory and Techniques*, 2018.

[13] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 2012.

[14] COMPAL Electronics Inc. Channel based methods for signal integrity evaluation. 2013.

[15] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

[16] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black box functions. *Journal of Global Optimization*, 13, 1998.

[17] E. Langford, N. Schwertman, and M. Owens. Is the property of being positively correlated transitive. *The American Statistitian*, 2001.

[18] A. Mahmood and S. Khan. Exploiting transitivity of correlation for fast template matching. *IEEE Transactions on Image Processing*, 2010.

[19] A Manukovsky, Y. Shlepnev, Z. Khasidashvili, and E. Zalianski. Machine learning applications for com based simulation of 112gbs systems. *DesignCon*, 2020.

[20] M. Marin and Y. Shlepnev. Systematic approach to pcb interconnects analysis to measurement validation. In *2018 IEEE Symposium on Electromagnetic Compatibility, Signal Integrity and Power Integrity (EMC, SI PI)*, pages 228–233, 2018.

[21] R. Mayder, R. Anderson, and N. Kamdar. Touchstone v2.0 si/pi s-parameter models for simultaneous switching noise analysis of ddr4 memory interface applications. *DesignCon*, 2014.

[22] R. Mellitz, A. Ran, M. P. Li, and V. Ragavassamy. Channel operating margin (com): Evolution of channel specifications for 25gbps and beyond. *DesignCon*, 2013.

[23] P. J. Pupalaikis. The relationship between discrete-frequency s-parameters and continuous-frequency responses. *DesignCon*, 2012.

[24] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[25] B. J. Reich, J. Eidsvik, M. Guindani, J. A. Nail, and A. M. Schmidt. A class of covariate dependent spatiotemporal covariance function. *The Annals of Applied Statistics*, 2011.

[26] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[27] F. Springenberh and Hutter. Initializing bayesian hyperparameter optimization via meta-learning. *National Conference on Artificial Intelligence*, 2015.

[28] M. Stefano, S. Grivet-Talocia, T. Wendt, C. Yang, and C. Schuster. A multi-stage adaptive sampling scheme for passivity characterization of large-scale macromodels. *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, 2021.

[29] Swersky, Snoek, and Adams. Multi-task bayesian optimization. *Neural Information Processing Systems*, 2013.

[30] S. Swersky, W. Adams, and Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proc. IEEE 104*, 2016.

[31] P. Triverio, S. Grivet-Talocia, M. S. Nakhla, F. G. Canavero, and R. Achar. Stability, causality, and passivity in electrical interconnect models. *IEEE Transactions on Advanced Packaging*, 2007.

[32] C. K. I. Williams, K. M. A. Chai, and E. V. Bonilla. A note on noise-free gaussian process prediction with separable covariance functions and grid design. 2007.

[33] K. Willis, A. Varma, and K. Keshavan. Backchannel modeling and simulation using recent enhancements to the ibis standard, 2019.

[34] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23:550–560, 1997.