Preprint

MFNets: Multi-fidelity data-driven networks for Bayesian learning and prediction

A. Gorodetsky[1], J.D. Jakeman[2], G. Geraci[2], and M.S. Eldred[2]

[1]University of Michigan, 3053 FXB, 1320 Beal Avenue, Ann Arbor, MI. 48109, USA
[2]Optimization and Uncertainty Quantification, Sandia National Laboratories, Albuquerque, NM, 87123

Abstract

This paper presents a multifidelity uncertainty quantification framework called MFNets. We seek to address three existing challenges that arise when experimental and simulation data from different sources are used to enhance statistical estimation and prediction with quantified uncertainty. Specifically, we demonstrate that MFNets can (1) fuse heterogeneous data sources arising from simulations with different parameterizations, e.g simulation models with different uncertain parameters or data sets collected under different environmental conditions; (2) encode known relationships among data sources to reduce data requirements; and (3) improve the robustness of existing multi-fidelity approaches to corrupted data. MFNets construct a network of latent variables (LVs) to facilitate the fusion of data from an ensemble of sources of varying credibility and cost. These LVs are posited as explanatory variables that provide the source of correlation in the observed data. Furthermore, MFNets provide a way to encode prior physical knowledge to enable efficient estimation of statistics and/or construction of surrogates via conditional independence relations on the LVs. We highlight the utility of our framework with a number of theoretical results which assess the quality of the posterior mean as a frequentist estimator and compare it to standard sampling approaches that use single fidelity, multilevel, and control variate Monte Carlo estimators. We also use the proposed framework to derive the Monte Carlo-based control variate estimator entirely from the use of Bayes rule and linear-Gaussian models – to our knowledge the first such derivation. Finally, we use two numerical examples to demonstrate robustness and the ability to deal with heterogeneous information sources.

Article Info

Correspondence:
A. Gorodetsky
goroda@umich.edu

Keywords:
machine learning, Bayesian network, multi-fidelity modelling, regression, uncertainty quantification, control-variate Monte Carlo, multi-level Monte Carlo, co-kriging

1 Introduction

Credible prediction, certification, and design of complex systems requires assessment of uncertainties in the models and data used. Unfortunately, accurately quantifying uncertainty using high-fidelity models and/or observational data is typically intractable due to large computational and experimental costs. Addressing this challenge requires moving beyond single source data analysis by leveraging multiple data sources of varying

cost and credibility. In the simulation context, multi-fidelity (MF) modeling approaches utilize an ensemble of models that enrich a small number of high-fidelity simulations with larger numbers of simulations from models of varying prediction accuracy and reduced cost. These approaches enable either (1) more accurate estimation of predictive statistics for the same cost or (2) equal accuracy estimation for a lower computational cost compared with high-fidelity analysis. Overall, MF modeling can reduce the amount of high-fidelity data required for UQ analyses by orders of magnitude.

Research for combining low- and high-fidelity information sources can be traced to at least the 1980s with the development of a large number of frequentist estimators — in particular control variate techniques [29, 31, 19]. These techniques facilitate the creation of an unbiased estimator that reduces the variance of a Monte Carlo (MC) estimate of an integral by introducing easier integrals with known values. These approaches have more recently been refined to address numerical simulation models via multilevel and multi-index Monte-Carlo techniques [6, 12], which has brought a resurgence in the idea of leveraging multiple model fidelities to enhance high fidelity modeling and uncertainty quantification [25, 3].

The vast majority of recent literature has typically focused on quantifying uncertainty only in the output of a numerical simulation. Here we consider multiple information sources more broadly, and the terms simulation outputs and data are used interchangeably. Consequently the term high-fidelity can refer to either highly-resolved simulation models with an incorporation of the most accurate physics models available, or to an actual physical process for which analysis, optimization, or control is desired. Similarly the term low fidelity is used to describe both models with simplified physics e.g., potential flow or Reynolds Averaged Navier Stokes (RANS) vs Direct Numerical Simulation (DNS) and low-resolution scattered data and/or under different, but related, experimental conditions.

The motivation for multi-fidelity approaches relies on the assumption that there exists a similar underlying cause for data obtained from different sources. As a result, the effectiveness of any multi-fidelity scheme depends primarily on the ability to discover and exploit the relationships between the causes for each data generating process. The precise definitions of these relationships can vary greatly. Primarily, the underlying causes are shared physics and the differences are attributed to different combinations of active physics, the nature of model form inadequacies, and/or the choices of numerical simulation schemes. There is very little existing literature that addresses how different relationships can be leveraged and/or learned during the inference process. This paper seeks to make a step in this direction.

Instead, a vast majority of existing MF approaches exploit structure that presumes an underlying hierarchy of simulations. For example, [26, 15, 10, 18] build approximations that leverage a hierarchy of models of increasing fidelity, with varying physics and/or numerical discretizations. Multi-level [6, 2, 21] and multi-index [12, 14] methods also leverage a set of models of increasing fidelity, with the additional assumption that a model sequence forms a convergent hierarchy due to the convergence of a numerical scheme. Some attention has been given to building multi-fidelity approximations using models that do not admit a strict ordering of fidelity [9, 24, 17]; however literature in this area is limited.

In the context of UQ, there are multiple target applications for a MF analysis, which encompass either sampling-based or surrogate-based approaches. Sampling algorithms [2, 6, 12, 24, 9] extract structure among data sources primarily through the computation of correlation. They then form statistical estimators that exploit this correlation for the purpose of *estimator variance reduction*. In practice, the observed variance reduction can result in orders of magnitude reduction in the computational cost of quantifying uncertainty, but like traditional MC sampling, the error decreases slowly as the number of samples is increased, which can still render this approach infeasible in some contexts. In contrast, surrogate methods [11, 14, 32] build an approximation of the input-output response. Provided the response surface is smooth enough, the error in surrogate methods can decrease rapidly, in some cases exponentially, as the amount of data increases. These approaches often take the form of deterministic surrogates or Gaussian process surrogates (co-Kriging [17]). Unfortunately it is not always clear a priori whether to use a sampling or surrogate based method on any given application.

Existing approaches have 3 major shortcomings: (1) they do not explicitly address noisy or corrupted data; (2) they do not provide a general mechanism to incorporate or exploit prior information about the relationships between data; and (3) they (generally) do not provide uncertainty estimates. The framework presented here, based upon multi-fidelity data-driven networks for Bayesian learning and prediction (MFNets), provides a consistent approach for both sampling and surrogate based MF modeling. We show that some existing sampling and surrogate based approaches are specific realizations of this general frame-

work. For example, we derive from first principles (Bayes rule and linear-Gaussian assumptions) the standard control variate scheme, as well as a number of other sampling-based estimators. We also show that recursive co-kriging [10] also has an interepretation in our context.

Our approach is inspired by the fact that, in the context of physical systems, engineers and scientists typically have some prior knowledge about the relationships between the models. For example, they may know that a high fidelity model may share some physical parameters with a lower fidelity model, or that the magnitude of the discrepancy between QoI computed with successive finite element models decreases as the mesh is refined. In this paper we develop a methodology by which this information can be encoded to create MF UQ algorithms.

To achieve this goal, our approach posits latent variables (LVs) to explain observed relationships in data sources. These latent variables could correspond to the parameters of a simulation model or the coefficients of a fully data-driven model used to explain experimental data. Once LVs and their presumed prior relationships are specified, the task of MF UQ becomes one of inferring the LVs and then using these data-informed estimates to make predictions. Adopting this approach allows data from every information source to inform the latent variables of every other information source. However, naively inferring latent variables can become very expensive, especially if the latent variables are high dimensional and there are a large number of information sources. To overcome this challenge, we will exploit conditional independence properties between each latent variable to reduce data requirements and use Bayesian networks to ground our approach in a widely used modeling and analysis framework.

Existing methods focus on exploiting the relationships between model outputs. In contrast inferring "explanatory" latent variables provides a mechanism to exploit not just model outputs (effects) but also common causes. Consequently, by leveraging disparate latent variable parameterizations for each model, MFNets enable MF analysis even in the case of heterogeneous data – simulation and experiment – that may involve different input/output spaces. For example, latent variables can be used to represent parameters of a simulation model or the coefficients of a regression model used to explain experimental data. Regardless of the interpretation of the LVs, inferring them only requires estimating the relationships between LVs and is agnostic to the mapping from LVs to the data.

To summarize, the major contributions presented in this paper are:

1. A description of a Bayesian modeling formulation for multi-fidelity UQ that incorporates prior knowledge and exploits connections between models;

2. Theoretical development and frequentist analysis of certain random sampling-based estimators that can be derived from the Bayesian model (Theorems 1 and 2);

3. A derivation of the control variate method from a Bayesian viewpoint (Theorem 3);

4. Numerical demonstration of overcoming robustness issues in MLMC and dealing with heterogeneous inputs.

The remainder of the paper is organized as follows. Section 2 introduces multi-fidelity modeling more formally and provides the relevant background for our proposed approach. Section 3 discusses our Bayesian modeling framework and algorithms. Section 4 establishes the connections between the proposed approach and existing sampling approaches. Numerical results are provided in Section 5 to support the theoretical findings in the previous sections and develop intuition regarding the effectiveness of MF using Bayesian networks. Conclusions are provided in Section 6.

2 Problem formulation and background

In this section we formalize the problems of multi-fidelity inference and prediction and provide background relevant to our proposed solution. Let $\mathbb{Z}_+$ denote the positive integers and $\mathbb{R}$ denote the set of real numbers. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability triple of sample space, sigma algebra, and probability measure. For continuous random variables, we will often abuse notation and have $\mathbb{P}$ denote both the probability density function and probability distribution. It will be clear from context as to the meaning. We will use capital letters, e.g., X, to denote $d \in \mathbb{Z}_+$ dimensional random variables $X : \Omega \to \mathbb{R}^d$ that are $\mathcal{F}$ measurable. In this

work the term *model* and *information source* are fully synonymous, as both apply to the concept of feeding an input into a black box (being a model or a sensor) and receiving an output.

2.1 Problem statement

Consider $M \in \mathbb{Z}_+$ models $(h_i : \mathcal{X}_i \to \mathcal{Y}_i)_1^M$ with possibly different input and output spaces across models. For simplicity we have assumed that the output of each model is scalar valued $\mathcal{Y}_i \subseteq \mathbb{R}$. The *inputs* to model i are random variables $X_i \in \mathcal{X}_i$ for $d_i \in \mathbb{Z}_+$ and $i = 1, \ldots, M$, and the output random variables are $Y_i \in \mathbb{R}$:

$$Y_i = h_i(X_i), \quad i = 1, \ldots, M. \tag{1}$$

The UQ problem seeks statistics of the output random variables Y_i. One wide ranging class of statistics of interest — that includes expectations, variances, failure probabilities, etc. — is formulated as expectations of functions $g : \mathbb{R} \to \mathbb{R}$ of the output from each model $\mathbb{E}_{X_1}\left[g\left(h\left(X_1\right)\right)\right]$. Concretely, examples include the average value of Y_1, for which $g(y) = y$ is the identity; the variance of Y_1, for which $g(y) = y^2 - \mathbb{E}_{X_1}[y]^2$; or the probability of an event $a < h(x) < b$, for which $g(y) = \mathbf{1}_{a<y<b}$ is the indicator function.

The following exposition focuses on situations for which a database of model evaluations $\mathcal{D}_i = \left(x_i^{(j)},\ y_i^{(j)}\right)_{j=1}^{m_i}$ for $m_i \in \mathbb{N}$ is provided, and that there is no possibility of acquiring new data. We arbitrarily designate the first model $i = 1$ as "high-fidelity" — denoting the model whose output uncertainty we would like to quantify. The last $M - 1 \in \mathbb{Z}_+$ are called "low fidelity" models. Again, they need not be worse at predicting, rather they are information sources we can leverage for the purpose of refining the estimate of the expectation of $g(Y_1)$.

Concretely, this paper proposes a methodology to evaluate the conditional expectation

$$q(\mathcal{D}_1, \ldots \mathcal{D}_M) = \mathbb{E}\left[g(Y_1)|\mathcal{D}_1, \ldots, \mathcal{D}_M\right] \tag{P}$$

with limited computational expense. As a specific example of (P), we focus on the expected value of Y_1

$$q_E(\mathcal{D}_1, \ldots \mathcal{D}_M) = \mathbb{E}\left[Y_1|\mathcal{D}_1, \ldots, \mathcal{D}_M\right]. \tag{P_E}$$

2.2 Latent variable modeling approach

We are interested in estimating the statistics of QoI using data from each model $\mathcal{D}_i$. To aid this task we will attempt to exploit structure relating the data of each model to one another. To this end, we essentially presume that for each model there exists a set of hidden variables that (1) captures model structure in the input-output map and (2) are somehow related among the models.

These hidden, or latent, variables $\theta_i \in \mathbb{R}^{p_i}$, are used to parameterize functions that map the input-output structures generically as

$$Y_i = \widehat{h}_i\left(X_i, \theta_i\right), \tag{2}$$

where $\widehat{h}_i$ can be a representation such as a linear basis expansion or a neural network. The goal then is to use all of the data to learn the parameters θ_i of each surrogate. In the single-model case, this is a standard supervised learning task: minimize an empirical risk measure to obtain an optimal θ_i (frequentist viewpoint) or use Bayesian inference to obtain a posterior distribution for θ_i. The connection between these approaches is highlighted in Section 4.2.2.

Given that we are seeking a surrogate model for representing each model output and that each surrogate is entirely defined by a set of parameters θ_i and a fixed approximation form $\widehat{h}_i$, the only sources that could correlate or relate the outputs across the models are the parameters. The relationship between these variables can be captured in a probabilistic form via a joint distribution

$$\mathbb{P}_\theta\left(\theta_1, \ldots, \theta_M\right). \tag{3}$$

In the Bayesian framework, this joint distribution represents the prior over the parameters of each model, and we seek a posterior distribution

$$\mathbb{P}_\theta\left(\theta_1, \ldots, \theta_M|\mathcal{D}_1, \ldots, \mathcal{D}_M\right) \propto \mathbb{P}\left(\mathcal{D}_1, \ldots, \mathcal{D}_M|\theta_1, \ldots, \theta_M\right)\mathbb{P}_\theta\left(\theta_1, \ldots, \theta_M\right). \tag{4}$$

This setting introduces a new source of uncertainty into the surrogate model beyond the inputs. In other words, we now can consider estimated statistics as a function of the parameters

$$\tilde{q}_E(\theta_1) = \mathbb{E}_X \left[\widehat{h}_1 \left(X_1, \theta_1 \right) \right]. \tag{5}$$

Here, our estimate of the target statistic is itself a random variable (because the parameters θ_i are uncertain). There are several options for extracting a point-estimate of the statistic. Options include marginalizing over the latent variables (MAR) and choosing the most likely variable θ_1 (MAX) while marginalizing over the rest:

$$q_E(\mathcal{D}_1, \ldots, \mathcal{D}_M) = \mathbb{E}_{\theta_1} \left[\tilde{q}_E(\theta_1) \right] = \int \mathbb{E}_{X_1} \left[\widehat{h}_1(X_1, \theta_1) \right] \mathbb{P}_\theta \left(\theta_1, \ldots, \theta_M | \mathcal{D}_1, \ldots, \mathcal{D}_M \right) d\theta_1 \ldots d\theta_M, \tag{MAR}$$

$$q_E(\mathcal{D}_1, \ldots, \mathcal{D}_M) = \mathbb{E}_{\theta_1} \left[\tilde{q}_E(\theta_1^*) \right] = \mathbb{E}_{X_1} \left[\widehat{h}_1 \left(X_1, \theta_1^* \right) \right] \tag{MAX}$$

where $\theta_1^* = \arg\max_{\theta_1} \int \mathbb{P}_\theta \left(\theta_1, \ldots, \theta_M | \mathcal{D}_1, \ldots, \mathcal{D}_M \right) d\theta_2 \ldots d\theta_M$ is the parameter which maximizes the marginal posterior over θ_1.

The challenge of inferring all the parameters stems from scalability with the number of models M and the number of parameters representing each one p_i. In other words, inference over $\mathcal{O}(Mp)$ for $p_i < p$ may be computationally challenging. From an information viewpoint, there are also many unknowns and therefore a large amount of data is potentially required to sufficiently reduce the uncertainty represented by the prior. To counteract both these computational and data-based difficulties, we restrict the allowable joint distributions in a manner that is synergistic with the aims of multi-fidelity UQ. Specifically, we impose conditional independence relationships between models.

2.3 Conditional independence assumptions for multi-fidelity UQ

In this paper we use conditional independence to construct compact representations of the joint probability distribution functions and facilitate data-efficient inference. To aid in this task we employ Bayesian networks to provide an abstraction built on graph theory for describing conditional independence relations among large groups of random variables. A deeper discussion of the following concepts can be found in [16].

Definition 1 provides a precise notion of conditional independence.

Definition 1 (Conditional independence of random variables). *Let A, B and C denote three random variables. A is conditionally independent of B given C in distribution $\mathbb{P}$, denoted by $A \perp\!\!\!\perp B|C$, if*

$$\mathbb{P}(A = a, B = b|C = c) = \mathbb{P}(A = a|C = c)\mathbb{P}(B = b|C = c). \tag{6}$$

A Bayesian network (BN) structure is a directed acyclic graphs (DAG) whose nodes represent random variables and whose edges represent conditional probability distributions (CPDs). A graph can be represented by a tuple of vertices (or nodes) and edges $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. A node $\theta_j \in \mathbf{V}$ is a parent of a random variable $\theta_i \in \mathbf{V}$ if there is a directed edge $[\theta_j \to \theta_i] \in \mathbf{E}$. The set of parents of random variable θ_i is denoted by $\theta_{\text{pa}(i)}$. The set of all CPDs is given by $\mathbf{p} = \{\mathbb{P}(\theta_i|\theta_{\text{pa}(i)}) : \theta_i \in \mathbf{V}\}$ and can be used to define a factorization of the graph.

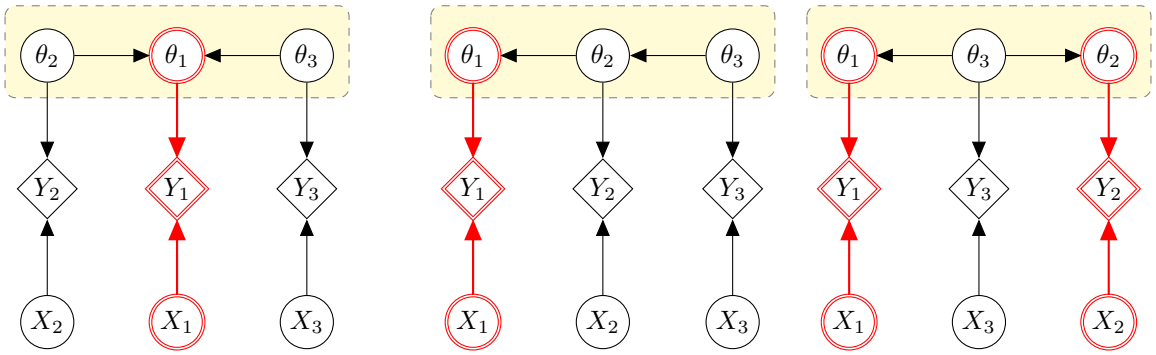
Definition 2 (Factorizing). *Let $\mathcal{G}$ be a BN structure over the set of random variables $\{\theta_1, \ldots, \theta_M\}$. The distribution $\mathbb{P}$ factorizes according to $\mathcal{G}$ if it can be expressed as a product over the set of conditional probability distributions*

$$\mathbb{P}(\theta_1, \ldots \theta_M) = \prod_{i=1}^{M} \mathbb{P}(\theta_i|\theta_{\text{pa}(i)}). \tag{7}$$

These factored forms can be leveraged to compute required statistics (MAR) and (MAX). For example using this factorization the MAR objective can be written

$$q_E = \int \mathbb{E}_{X_1} \left[\widehat{h}_1(X_1, \theta_1 \right] \int \theta_1 \mathbb{P}(\theta_1|\theta_{\text{pa}(1)}) \ldots \mathbb{P}(\theta_M|\theta_{\text{pa}(M)}) d\theta_1 \ldots d\theta_M, \tag{8}$$

where the data is omitted but understood to be implicit in the distribution of the parameters. A similar modification can be done for MAX.



(a) Peer low-fidelities ↔ V-structure (b) Model hierarchy ↔ Serial (c) Peer high-fidelities ↔ diverging

Figure 1: Multifidelity modeling interpretations of basic elements of Bayesian networks (surrounded in yellow). The random variables corresponding to high fidelity models and data are outlined in red. The X_i are input random variables representing the sources of uncertainty in model h_i, The quantities of interest for each model are Y_i, and the hidden variables are denoted by θ_i.

Bayesian networks encode sets of conditional independence assumptions, called local Markov assumptions, and is composed of three fundamental building blocks: serial, diverging, and v-structure connections. These three building blocks are are shown in Figure 1. We now describe their interpretation in the multi-fidelity UQ setting.

2.3.1 Peer low-fidelity models

Figure 1a shows the situation where two models are peers with respect to a single high fidelity model. This type of network structure is one where θ_1 is interpreted as a *common effect* of the two low fidelity models. This particular model introduces a relationship called a *v-structure* $\theta_2 \rightarrow \theta_1 \leftarrow \theta_3$. This structure encodes the notion that the two lower-fidelity models are marginally independent. However, if θ_1 is known then they become conditionally dependent on one another. As an example, consider the case where a high fidelity simulation model incorporates two sets of active physics, and the two low-fidelity peer models each contain one of these components. If the high-fidelity model is given, then the low-fidelity models are no longer independent of one another. In other words, information about the parameters used in one set of physics will inform the other set of physics because they are coupled together in a known way through the high-fidelity model.

2.3.2 Ordered hierarchy of fidelities

Another example, and one that is used most commonly in the literature is one of a chain, or hierarchy, of models. Figure 1b shows the graphical representation of such a network. This network represents a causal trail, where an *ordered set* of fidelities is assumed and the highest and lowest fidelity models are independent given the middle fidelity models. As an example, suppose that the models are related through grid coarsening. In this case, given the middle level of grid refinement, the high and low fidelity models are clearly independent. The high fidelity model is obtained through a prolongation of the middle one, and the low fidelity model is obtained through restriction. However, if only the solution at the high-fidelity model is given then the two lower fidelity models are clearly dependent. The lowest fidelity model is obtained through a restriction of the middle fidelity model.

When all variables are Gaussian, the latent variable description we have described is a parametric form of a multi-output Gaussian process, or co-Kriging model. The parametric and non-parametric forms of Gaussian process have a precise relationship defined through the eigenfunctions of the kernel and Mercers theorem; we refer the reader to [28, 8] for a discussion of this relationship. Co-kriging distinguishes itself from regular kriging by building a model for the interaction between each of the sub-models θ_i. This approach has been used for modeling in multiple experiment environments [27], multi-fidelity optimization [4], as well as for multi-fidelity computer experiments [10]. Different kernel specifications, e.g., one based

on Markov random fields [26], have also been investigated. However, as far as the authors are aware, all of these methods operate in the case where the relationship between different models is "hierarchical" in two different but related senses. The first is the case where the actually computer experiments are ordered from most accurate to least accurate, this is the assumption in [26]. The second is an enforced hierarchy that occurs because the latent variable models are ordered in a chain. This ordering is used in [10] and utilized to construct a recursive estimation scheme for learning the hyper-parameters and performing experimental design.

2.3.3 Peer high-fidelity models

Figure 1c shows one possible case where two models are not marginally independent, but knowledge of a third model breaks their dependence; as a result, this structure is termed a *common cause*. An example application from multi-fidelity modeling is one where θ_3 represents a boundary, or interface, condition shared by models θ_1 and θ_2. Given this interface condition, there is no other shared information between the peer high fidelity models. As a result, any additional information used to inform one model does not provide any information about the other high fidelity model. These models effectively become decoupled and independent from one another.

Peer low-fidelity, hierarchies and peer high-fidelity graphs are multi-fidelity interpretations of the fundamental structures of Bayesian networks, namely, v-structure, serial and diverging respectively. These three structures can be combined to describe a rich set of model relationships. In the following section we demonstrate how arbitrary and potentially complex networks can be used to reduce the data requirements and computational cost of inferring latent variables.

3 Multi-Fidelity UQ through inference on Bayesian networks

In this section, we describe the parametric Bayesian formulation for assimilating disparate data sets into a single multi-fidelity model.

3.1 Linear-Gaussian models

Bayesian networks can be used for inference and prediction using a broad classes of prior distributions and models relating latent variables of the different data sources. In this section we describe a linear-Gaussian formulation. Given a set of M basis functions (features) $\phi = \{\phi_1, \ldots, \phi_M : \phi_i : \mathcal{X}_i \to \mathbb{R}^{p_i}\}$, we seek approximations of each model h_i of the form

$$Y_i = \widehat{h}_i(X_i, \theta_i) = \phi_i^T(X_i)\theta_i \equiv \sum_{k=1}^{p_i} \phi_{ik}(X_i)\theta_{ik} \tag{9}$$

This expansion can represent any true model h_i in L_2 as $p_i \to \infty$, if the basis for each model is complete in L_2. Furthermore, the linearity of this expansion, with respect to the latent variables, makes it easy to obtain statistics of the quantity of interest. For instance, Equation (5) becomes

$$\tilde{q}_E(\theta_1) = \mathbb{E}_{X_1}\left[\phi_1^T(X_1)\right]\theta_1, \tag{10}$$

which simplifies even further if ϕ_{ik} are orthogonal w.r.t the measure of X_1 and $\phi_{i1} = 1$ (as in the case of polynomial chaos expansions [5]) so that

$$\tilde{q}_E(\theta_1) = \theta_{1k}. \tag{11}$$

Given the model in (9) we use a Gaussian distribution to encapsulate uncertainty in the latent variables. That is

$$\theta = (\theta_1, \theta_2, \ldots, \theta_M) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_M \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1M} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{M1} & \cdots & \boldsymbol{\Sigma}_{MM} \end{bmatrix}, \tag{12}$$

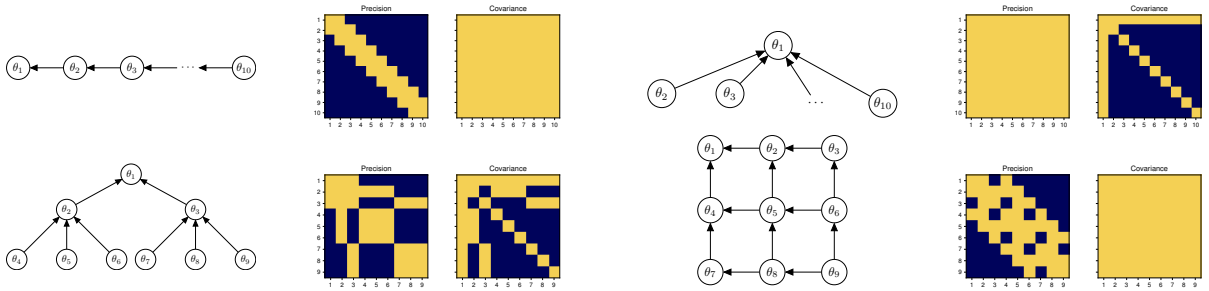


Figure 2: Sample networks and associated sparsity patterns of their precision $\mathbf{\Sigma}^{-1}$ and covariance matrices $\mathbf{\Sigma}$. Upper left is a hierarchical model, upper right is a peer model, lower left is a tree network, and lower-right is a multi-hierarchical network.

where the means of each parameter are $\boldsymbol{\mu}_i \in \mathbb{R}^{p_i}$ and covariances between parameters are $\mathbf{\Sigma}_{ij} \in \mathbb{R}^{p_i \times p_j}$. Under this assumption both the MAR and MAX estimators are identical and marginalizing over θ_1 in Equation (10) yields

$$q_E = \mathbb{E}_{X_1}\left[\phi_1^T(X_1)\right]\boldsymbol{\mu}_1, \tag{13}$$

where we have used the marginalization properties and linearization properties of Gaussians, and have again explicitly assumed that the effect of the data are implicitly captured in $\boldsymbol{\mu}_1$ — for instance, the Gaussian distribution is the posterior distribution.

To combat the explosive growth in the size of the covariance matrix due to large M and/or large p_i we seek to restrict the class of covariance matrices. For the case of Gaussian distributions, the restriction from conditional independence relations imposes *sparsity* patterns on the covariance and/or precision matrices. The particular type of sparsity pattern is determined by the graph structure among the θ_i. Figure 2 shows several examples of graphs and the sparsity patterns they create. Furthermore, these patterns can be exploited during inference to improve computational efficiency.

To enable tractable inference we restrict the form of the CPDs themselves to be linear-Gaussian so that

$$\theta_i|\theta_{\text{pa}(i)} \sim \mathcal{N}\left(\boldsymbol{A}_{i|\text{pa}(i)}\theta_{\text{pa}(i)} + \boldsymbol{b}_{i|\text{pa}(i)}, \mathbf{\Gamma}_{i|\text{pa}(i)}\right), \tag{14}$$

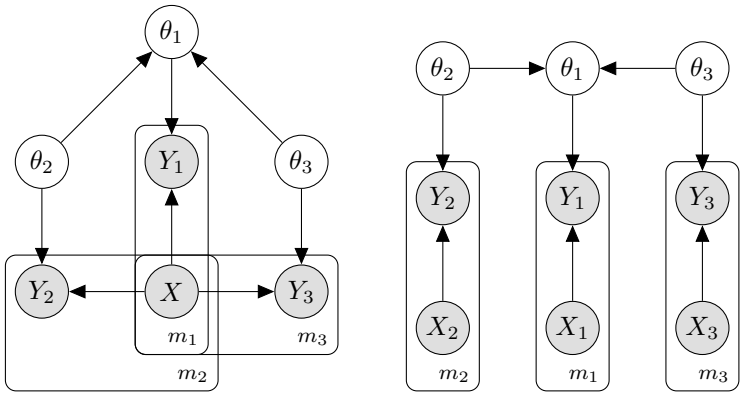
where $\boldsymbol{A}_{i|\text{pa}(i)} \in \mathbb{R}^{p_i \times \sum_{j\in\text{pa}(i)} p_j}$, $\boldsymbol{b}_{i|\text{pa}(i)} \in \mathbb{R}^{p_i}$, and $\mathbf{\Gamma}_{i|\text{pa}(i)} \in \mathbb{R}^{p_i \times p_i}$. In other words, the mean of the conditional distribution of a parameter θ_i is a linear combination of its parents values plus an offset. Clearly, if the parents are Gaussian the marginal distribution of the random variable is also Gaussian.

3.2 Data sources

Bayesian networks provide a mechanism to encode how the data $\mathcal{D}_i$ associated with each model in an ensemble are related. Figures 1 and 3b depict one type of relationship between inputs X_i and outputs Y_i; the input-output pairs of each model are independent (there are no shared input samples). However BN can be used to encode other relationships. In typical multi-fidelity UQ setups at least some of the inputs are shared across models and some are independent. For example, in control variate-based methods [9] a shared set of input samples must be used across information sources to estimate model correlations, and a separate set of input samples is used for each model to refine statistics of the individual models. This case is depicted in Figure 3a. Regardless of the relationship between data the computational procedure for inferring latent variables remains the same. In Section 3.3 we formulate the inverse problem and in Section 3.3.2 we discuss how to leverage conditional independence to construct an efficient algorithm for solving this inverse problem.

3.3 Linear-Gaussian inference

In this section we describe the inverse problem of inferring latent variables under the linear-Gaussian assumptions of Section 3.1.



(a) Shared inputs (b) Independent inputs

Figure 3: Sampling-based methods like multilevel Monte-Carlo or control variates require a certain number of shared input realizations in order to measure correlation between the output of each model (left). Introducing latent variable networks poses an underlying cause for observed correlation and therefore omits the requirement for shared data by introducing an alternative model for describing the relationship of model outputs.

3.3.1 Problem specification

The likelihood in Bayes rule (4) is obtained by specifying a statistical model relating the model outputs to the latent variables. Let $\boldsymbol{y}_i = \left[y_i^{(1)}, \ldots, y_i^{m_i}\right]$ and $\boldsymbol{x}_i = \left[x_i^{(1)}, \ldots, x_i^{(m_i)}\right]$ denote vectors containing the outputs and inputs for each model. Then the data for each model is then modeled with zero mean Gaussian noise

$$\boldsymbol{y}_i = \boldsymbol{\Phi}_i \theta_i + \xi, \quad \xi \sim \mathcal{N}(0, \sigma_i^2 I_{m_i}), \qquad \text{for } i = 1, \ldots, M, \tag{15}$$

where $\boldsymbol{\Phi}_i \in \mathbb{R}^{m_i \times p_i}$ are Vandermonde-like matrices

$$\boldsymbol{\Phi}_i = \begin{bmatrix} - & \phi_i^T(x_i^{(1)}) & - \\ & \vdots & \\ - & \phi_i^T(x_i^{(m_i)}) & - \end{bmatrix}$$

Under this statistical error model the likelihood for each model is Gaussian

$$\mathbb{P}(\boldsymbol{y}_i | \boldsymbol{x}_i, \theta_i) \propto \exp\left[-\frac{1}{2\sigma_i^2}\left(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\theta_i\right)^T\left(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\theta_i\right)\right], \tag{16}$$

where we have assumed that the data for each model are independent of each other for simplicity of presentation.

The marginal likelihood (likelihood integrated over the parameters) of the output given the inputs is

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \;\sim\; \mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$$

where

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{\Phi}_1\boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\Phi}_M\boldsymbol{\mu}_M \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{\Phi}_1\boldsymbol{\Sigma}_{11}\boldsymbol{\Phi}_1^T + \sigma_1^2\boldsymbol{I}_{m_1} & \cdots & \boldsymbol{\Phi}_1\boldsymbol{\Sigma}_{1M}\boldsymbol{\Phi}_M^T \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}_M\boldsymbol{\Sigma}_{M1}\boldsymbol{\Phi}_1^T & \cdots & \boldsymbol{\Phi}_M\boldsymbol{\Sigma}_{MM}\boldsymbol{\Phi}_M^T + \sigma_M^2\boldsymbol{I}_{m_M} \end{bmatrix}. \tag{17}$$

If we specify

$$\boldsymbol{\Phi} = \text{block-diag}\left[\boldsymbol{\Phi}_1 \;\cdots\; \boldsymbol{\Phi}_M\right] \qquad \text{and} \qquad \Sigma_n = \text{block-diag}\left[\sigma_1^2 I_{m_1} \;\cdots\; \sigma_M^2 I_{m_M}\right]$$

then $\boldsymbol{C} = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T + \Sigma_n$ and the joint density of the latent variables $\boldsymbol{\theta}$ and the data $\boldsymbol{y}$ is a Gaussian random variable with mean and covariance

$$\theta, \boldsymbol{y} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{m}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\\ \boldsymbol{\Phi}\boldsymbol{\Sigma} & \boldsymbol{C}\end{bmatrix}\right).$$

Using the Schur complement, we can obtain the conditional distribution for the parameters given observations is $\theta \mid \boldsymbol{y} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ where

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{C}^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Phi}^T\Sigma_n^{-1}\boldsymbol{\Phi}\right)^{-1} \tag{18}$$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}) = \bar{\boldsymbol{\Sigma}}\left(\boldsymbol{\Phi}^T\Sigma_n^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = \bar{\boldsymbol{\Sigma}}\boldsymbol{\Phi}^T\Sigma_n^{-1}\boldsymbol{y} + \bar{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{19}$$

where alternative but equivalent formulas (related by the Woodbury Identity) are given for both the posterior covariance and posterior mean.

The computational complexity of direct inference is dominated by two processes; inverting $\boldsymbol{C}$ requires $\mathcal{O}\left(\left(\sum_{i=1}^M m_i\right)^3\right)$ operations and computing $\boldsymbol{\Phi}\boldsymbol{\Sigma}$ requires $\mathcal{O}\left(p^2\sum_{i=1}^M m_i\right)$ where $p = \max_i p_i$. Furthermore the number of basis functions often grows exponentially with dimension of the input space $p = \mathcal{O}(N^d)$. Thus for problems with a large number of information sources or many input dimensions this approach becomes computationally prohibitive. However, the structured form $\boldsymbol{\Sigma}$ (and therefore $\boldsymbol{C}$) can be leveraged for efficient inference via variable elimination described next.

3.3.2 Bayesian network inference algorithm

In this paper we use variable elimination to reduce the computational cost and data reduction of inferring latent variables for large problem sizes. For a detailed explanation of variable elimination algorithms we refer the reader to [16]. In the following we only provide details relevant to our implementation context and discussion of computational complexity.

Algorithms 1 and 2 respectively describe algorithms for computing marginal distributions of a random variable and (unnormalized) conditional distributions of a subset of variables. Both of these algorithms work with distributions represented in a factor form rather than the CPD form of Equation (7). Bayesian networks can be trivially converted to this form constructing the factors $\psi_i(\theta_i, \theta_{\mathrm{pa}(i)}) = \mathbb{P}(\theta_i|\theta_{\mathrm{pa}(i)})$ so that we have $\mathbb{P}(\theta_1, \ldots, \theta_M) = \psi_1 \cdots \psi_M$. Each factor is a function of some subset of variables in the graph, and we denote the entire set of factors as $\Psi = \{\psi_1, \ldots \psi_M\}$. Algorithm 1 marginalizes over a subset of variables Z_i to obtain a factor corresponding to the marginal distribution of the remaining variable. For example, if $Z_i = \theta_{i+1}$ for $i = 1 \ldots, M-1$, then the output of this algorithm will be a marginal distribution over θ_1 — precisely what we require for evaluate (8).

The conditioning Algorithm (Algorithm 2) uses this marginalization algorithm following a preprocessing step that fixes the observed variables O to their given data points $\boldsymbol{y}$. Once these variables are fixed this algorithm returns factors representing the posterior distribution over query variables Q. This is achieved by eliminating all other variables by the sum-product variable elimination algorithm. As an example, if the network consisted of all the parameters θ_i and observations $\mathcal{D}_i$, then we can query a posterior over θ_1 by setting $Q = \theta_1$, $O = \{Y_1, \ldots, Y_M\}$ with $\boldsymbol{y}$ being the corresponding data.

Finally, we note that these variable elimination algorithms work with general distributions, but they can be performed efficiently in the linear-Gaussian case by relying on the *Canonical Form* for CPDs and associated Gaussian distributions. Recall that the canonical form of the density of a random variable X is given by $C(X; K, h, g) = \exp -\frac{1}{2}X^TKX + h^TX + g$. For Gaussian distributions of dimension d with mean μ and covariance Σ we have $K = \Sigma^{-1}$, $h = \Sigma^{-1}\mu$, and $g = -\frac{1}{2}\mu^T\Sigma^{-1}\mu - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\right)$. The operations of marginalization and computing factor products are required for these algorithms, and both of these operations have straightforward approaches in canonical form, and the computational complexity is linear in the number of cliques (nodes in the graph) and at most cubic in the size of the largest clique [16].

Algorithm 1 Sum-Product-VE variable elimination algorithm [16, Alg 9.1] (for computing marginals)

Inputs Set of factors Ψ; Set of variables to be eliminated $(Z_i)_{i=1}^d$; Ordering on Z $\prec$

Order Z_i such that $Z_i \prec Z_j$ iff $i < j$
for $i = 1$ to d **do**
  $\Psi \leftarrow$ sum-product-eliminate-var(Ψ, Z_i)
end for
$\psi^* \leftarrow \Pi_{\psi \in \Psi} \psi$
return ψ^*

Procedure sum-product-eliminate-var(Set of factors Ψ; Variable to be eliminated Z)

$\Psi' \leftarrow \{\psi \in \Psi : Z \in Scope[\psi]\}$
$\Psi'' \leftarrow \Psi \setminus \Psi'$
$\psi \leftarrow \Pi_{\psi \in \Psi'} \psi$
$\tau \leftarrow \sum_Z \psi$
return $\Psi \cup \{\tau\}$

Algorithm 2 Cond-Prob-VE [16, Alg 9.2] (for computing conditionals distributions of query variables Q)

Inputs Set of variables $\mathbf{V}$, Set of factors Ψ; Set of query variables Q; Observed variables O, their values $\boldsymbol{y}$

Replace each $\psi \in \Psi$ by $\psi(O = \boldsymbol{y})$
Select an elimination ordering $\prec$
$Z \leftarrow \mathbf{V} \setminus Q \setminus O$
$\psi^* \leftarrow$ Sum-Product-VE$(\Psi, Z, \prec)$
return ψ^*

4 Relation to frequentist sampling approaches

In this section, we develop Bayesian sampling-based estimators that can be compared to popular frequentist approaches. A major current theme in the multi-fidelity UQ literature is the idea of using low-fidelity models as a tool for *variance reduction* of Monte Carlo estimators of the statistics of the high fidelity model. Indeed this is the approach taken by the multilevel MC [6], multi-index MC [12], multi-fidelity Monte Carlo [23, 20], and more generally approximate control variate algorithms [9]. While reducing the variance of point estimators is a frequentist pursuit by its very nature, it turns out that estimators obtained from Bayes rule (specifically the posterior mean) can exhibit good frequentist properties.

We define sampling approaches to be the cases for which $\phi_i(x) = 1$, i.e., the parameters are taken to be the observations themselves, and the overall approach seeks to learn and refine the joint distribution amongst the outputs of each model. Our analysis consists of three steps: (1) we briefly review the analytical solution to a single fidelity inverse problem in this sampling case; (2) we extend this analysis to an arbitrary number of low-fidelity models; and finally (3) we show the connections between our Bayesian multi-fidelity posterior estimates with frequentist control-variate Monte Carlo estimators.

4.1 Single fidelity Bayesian sampling

In this section, we review a Bayesian single-fidelity sampling-based estimate of the expectation of a high-fidelity model output Y_1. In the single-fidelity sampling context we have

$$Y_1 = \theta_1 + \xi, \tag{20}$$

with prior $\theta_1 \sim \mathcal{N}(\mu, c_1^2)$and likelihood $Y_1|\theta_1 \sim \mathcal{N}(\theta, \sigma_1^2)$. Given this model the following proposition, which can be found in any standard text e.g., [30], provides an expression for the posterior mean of the model $\mathbb{E}[\theta_1|\boldsymbol{y}] = \bar{\mu}_1$.

Proposition 1 (Bayesian Sampler). *Let μ_1 and c_1^2 denote the prior mean and variance for parameter θ_1, $\boldsymbol{y}_1 \in \mathbb{R}^{m_1}$ denote a vector of m_1 i.i.d data points, and let the likelihood be Gaussian according to*

$y_1^{(i)}|\theta_1 \sim \mathcal{N}(\theta_1, \sigma_1^2)$. *Then the posterior* $\theta_1|\boldsymbol{y}_1 \sim \mathcal{N}(\bar{\mu}, \bar{c}_1^2)$ *with*

$$\bar{\mu}_1 = \nu\mu_1 + (1-\nu)\widehat{y}_1^{MC},$$
$$\bar{c}_1^2 = \nu c_1^2,$$

where $\nu \equiv \frac{\eta}{\eta+1} \in [0,1]$ *and* $\eta \equiv \sigma_1^2/(m_1 c_1^2)$ *denotes the non-dimensional variable describing the balance between the data and the prior uncertainty, and*

$$\widehat{y}_1^{MC} = \frac{1}{m_1}\sum_{i=1}^{m_1} y_1^{(i)}, \tag{21}$$

is a Monte Carlo estimate of the mean.

The proof is given in Appendix 8.2. As expected, the posterior mean is a convex combination of the prior mean and a Monte Carlo estimate of the data. In other words, it carefully balances the prior beliefs and the observed data. As data becomes more informative, either because $m_1 \to \infty$ or because there is less noise $\sigma_1 \to 0$, the weight converges $\nu \to 0$.

It is instructive to compare the behavior of the Bayesian posterior to the behavior of the Monte Carlo estimator. First, the fact that the posterior mean converges to the MC estimator indicates that the Bayesian sampling setup is *consistent.* We also see that the posterior covariance decays at the same rate $\mathcal{O}(1/m_1)$ as the MC estimator variance

$$\frac{\bar{c}_1^2}{c_1^2} = \frac{\eta}{\eta+1} = \frac{\sigma_1^2/c_1^2}{\sigma_1^2/c_1^2 + m_1}. \tag{22}$$

If one treats the posterior mean as an estimator, one can analyze the frequentist properties of this estimator. For finite sample sizes the posterior mean is biased, in contrast to the unbiased Monte Carlo estimator, because the prior mean affects the answer at finite sample sizes. Consequently, to compare this estimator to MC we have to look at the full mean squared error. The bias and variance of this estimator is given by

$$\mathbb{E}\left[\bar{\mu}_1\right] - \mathbb{E}[Y_1] = \nu\mu_1 + (1-\nu)\mathbb{E}\left[\widehat{y}_1^{MC}\right] - \mathbb{E}[Y] = \nu\mu_1 + (1-\nu)\mathbb{E}[Y_1] - \mathbb{E}[Y] = \nu(\mu_1 - \mathbb{E}[Y]),$$
$$\mathbb{V}\mathrm{ar}\left[\bar{\mu}_1\right] = (1-\nu)^2\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right] m$$

respectively, where the expectation and variance are understood to be over hypothetical realizations of the data. Interestingly the variance of this estimator is smaller than the variance of the MC estimator. This fact suggests that, if the bias is small enough, the posterior mean can be a more effective estimator than the MC estimator. To determine if/when this is the case we must compare the mean squared errors (MSE) of these two estimators. Since MC is unbiased, its MSE is equal to its variance. The MSE of the posterior mean is

$$\mathrm{MSE}(\bar{\mu}_1) = \mathrm{Bias}^2 + \mathrm{Var} = \nu^2\left(\mathbb{E}\left[Y_1\right] - \mu_1\right)^2 + (1-\nu)^2\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right] \tag{23}$$

Intuitively, if the prior mean is a good estimate of the expectation, we should see a smaller MSE. Clearly, this is the case since the Bias term becomes negligible so that the posterior MSE is smaller than the MSE of the Monte Carlo estimator. Although this situation is unrealistic, it reflects the fact that if the prior mean is not too far from the truth, this estimator may perform better than Monte Carlo. Alternatively, the estimator variance will suffer if the prior mean is too far from the truth. The threshold of this value is given by the following theorem.

Theorem 1 (Posterior mean as a frequentist estimator). *Let* μ_1 *and* c_1^2 *denote the prior mean and variance for parameter* θ_1, $\boldsymbol{y}_1 \in \mathbb{R}^{m_1}$ *denote a vector of* m_1 *i.i.d data points, and* σ_1 *denote the likelihood standard deviation so that the likelihood be given as* $y_1^{(i)}|\theta_1 \sim \mathcal{N}(\theta_1, \sigma_1^2)$. *If* μ_1 *and* c_1^2 *are chosen so that*

$$\left(\mathbb{E}\left[Y_1\right] - \mu_1\right)^2 \leq \left(\frac{1}{m_1} + 2\frac{c_1^2}{\sigma_1^2}\right)\mathbb{V}\mathrm{ar}\left[Y_1\right] \tag{24}$$

then the mean squared error of the posterior mean is smaller than that of the Monte Carlo estimator.

The proof is given in Appendix 8.3. We can interpret this result in the following way. If the prior covariance is much smaller than the noise $c_1^2 \ll \sigma_1^2$ then the right-hand side is essentially equal to the MC estimator variance. Therefore, to beat MC, the prior mean has to be *within* the estimator variance. On the other hand, if the prior covariance is large $c_1^2 \gg \sigma_1^2$, then the prior mean does not have to be within the MC estimator variance. Indeed in this case, the small noise can rapidly move the prior mean to the true value.

The goal for multi-fidelity UQ in this setting is to have the marginal posterior variance $\bar{c}_1^2$ be smaller when using more models, even if the prior marginal variance c_1^2 is identical. We show this fact in the next section.

4.2 Multi-fidelity Bayesian sampling

We now extended the single fidelity results in the previous section to the multi-fidelity setting by deriving the form of the marginal posterior distribution over the high-fidelity model. The multi-fidelity sampling case is defined by $\phi_i = 1$ for $i = 1, \ldots, M$. Under this assumption Equations (18) and (19) become

$$\bar{\boldsymbol{\Sigma}} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Gamma}\right)^{-1}, \tag{25}$$

and

$$\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\begin{bmatrix} \widehat{y}_1^{MC} \\ \vdots \\ \widehat{y}_M^{MC} \end{bmatrix} + \bar{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_M \end{bmatrix} \tag{26}$$

respectively, where $\boldsymbol{\Gamma} = \text{diag}\begin{bmatrix} m_1/\sigma_1^2 & \cdots & m_M/\sigma_M^2 \end{bmatrix}$. Clearly, the posterior mean in the multi-fidelity setting is again a convex combination of the Monte Carlo estimator and the prior mean. Broadly speaking as each $m_i \to \infty$, then the matrix $\boldsymbol{\Gamma}$ begins to dominate the posterior so that the Monte Carlo estimator becomes more heavily weighted. In the subsequent discussion, it will be useful to partition the noise matrix as

$$\boldsymbol{\Gamma} = \begin{bmatrix} m_1/\sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{>1} \end{bmatrix}, \tag{27}$$

where $\boldsymbol{\Gamma}_{>1} = \text{diag}\begin{bmatrix} m_2/\sigma_2^2 & \cdots & m_M/\sigma_M^2 \end{bmatrix}$.

We are actually interested in the posterior of only the high-fidelity model to determine whether the addition of low-fidelity models reduce the uncertainty of the high-fidelity model. Specifically, we care about comparing the marginal posterior of the high-fidelity parameters in the multi-fidelity case to the posterior of the single fidelity case, given identical marginal priors.

Theorem 2 (Multi-Fidelity Bayesian Sampler)**.** *Let $\mu_1, \ldots, \mu_M$ denote the prior mean and*

$$\boldsymbol{\Sigma} = \begin{bmatrix} c_1^2 & \boldsymbol{c}^T \\ \boldsymbol{c} & \boldsymbol{\Sigma}_{>1} \end{bmatrix}, \tag{28}$$

denote the prior covariance. Let $\boldsymbol{y}_i \in \mathbb{R}^{m_i}$ denote a vector of m_i i.i.d data points for models $i = 1 \ldots, M$. Then the marginal posterior $\theta_1|\boldsymbol{y}_1 \sim \mathcal{N}(\bar{\mu}, \bar{c}_1^2)$ with

$$\bar{\mu}_1 = \lambda\mu_1 + (1-\lambda)\widehat{y}_1^{MC} + \nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1}\right) \tag{29}$$

$$\bar{c}_1^2 = \nu^{MF}c_1^2 \tag{30}$$

where $\nu^{MF} \equiv \frac{\eta}{1+\eta(1+\gamma)} \in [0,1]$, $\lambda = (1+\gamma)\nu^{MF} \in [0,1]$, and $\eta \equiv \sigma_1^2/(m_1 c_1^2)$ is the ratio of the noise to prior uncertainty for the high-fidelity model as in Proposition 1. Furthermore, we have

$$\boldsymbol{B} = \left(\boldsymbol{\Gamma}_{>1}^{-1} + \boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)^{-1}, \quad \text{and} \quad \gamma = \frac{1}{c_1^2}\boldsymbol{c}^T\boldsymbol{B}\boldsymbol{c}. \tag{31}$$

Interestingly the expression for the mean takes approximately a form that is similar to that of a control variate (CV) [22]. The first two terms are convex combination of the prior and MC means, and the third term is a correction factor that feeds back a discrepancy between the low fidelity models and their prior means. We cover the connections to control variates explicitly in the next section. Furthermore, we see that the posterior variance is actually smaller than for the case of the single-fidelity model due to the $1+\gamma$ term in the denominator. Here γ is a variable that accounts for the covariance between all the low-fidelity models and between the low-fidelity models and high-fidelity models.

4.2.1 Posterior mean as a frequentist estimator

In this section we treat the multi-fidelity posterior mean as a frequentist estimator and develop expressions for its bias and variance. Let Y_1 be a random variable from which we treat the data $y_1^{(i)}$ as realizations, and let $Y_{>1}$ have a corresponding relationship to $y_{>1}^{(i)}$. Since $Y_{>1}$ is multivariate, we simplify the analysis by assuming $m_k = m_{>1}$ for $k = 2, \ldots, M$. Under these settings the bias is

$$\begin{aligned}
\mathbb{E}\left[\bar{\mu}_1\right] - \mathbb{E}\left[Y_1\right] &= \lambda\mu_1 + (1-\lambda)\mathbb{E}\left[\widehat{y}_1^{MC}\right] + \nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\mathbb{E}_{Y>1}\left[\left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1}\right)\right] - \mathbb{E}\left[Y_1\right] && (32)\\
&= \lambda\left(\mu_1 - \mathbb{E}\left[Y_1\right]\right) + \nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\left(\mathbb{E}\left[\widehat{y}_{>1}^{MC}\right] - \boldsymbol{\mu}_{>1}\right), && (33)\\
&= \lambda\left(\mu_1 - \mathbb{E}\left[Y_1\right]\right) + \nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\left(\mathbb{E}\left[Y_{>1}\right] - \boldsymbol{\mu}_{>1}\right), && (34)
\end{aligned}$$

and the variance is

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\bar{\mu}_1\right] &= (1-\lambda)^2\frac{\mathbb{V}\mathrm{ar}\left[Y_1\right]}{m_1} + \frac{{\nu^{MF}}^2}{m_{>1}}\boldsymbol{c}^T\boldsymbol{B}\mathbb{C}\mathrm{ov}\left[Y_{>1}, Y_{>1}\right]\boldsymbol{B}\boldsymbol{c} + 2\left(1-\lambda\right)\nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\mathbb{C}\mathrm{ov}\left[\widehat{y}_{>1}^{MC}, \widehat{y}_1^{MC}\right]\\
&= (1-\lambda)^2\frac{\mathbb{V}\mathrm{ar}\left[Y_1\right]}{m_1} + \frac{{\nu^{MF}}^2}{m_1}\boldsymbol{c}^T\boldsymbol{B}\mathbb{C}\mathrm{ov}\left[Y_{>1}, Y_{>1}\right]\boldsymbol{B}\boldsymbol{c} + 2\left(1-\lambda\right)\nu^{MF}\boldsymbol{c}^T\boldsymbol{B}\frac{1}{m_1}\mathbb{C}\mathrm{ov}\left[Y_1, Y_{>1}\right] \qquad (35)
\end{aligned}$$

where the expectations, variances, and covariances are understood to be with respect to the random variables Y_1 and $Y_{>1}$. The last equality used the following fact.

$$\mathbb{C}\mathrm{ov}\left[\frac{1}{m_1}\sum_{i=1}^{m_1} y_1^{(i)}, \frac{1}{m_1}\sum_{i=1}^{m_1} y_{>1}^{(i)}\right] = \frac{1}{m_1^2}\mathbb{C}\mathrm{ov}\left[\sum_{i=1}^{m_1} y_1^{(i)}, \sum_{i=1}^{m_1} y_{>1}^{(i)}\right] = \frac{1}{m_1}\mathbb{C}\mathrm{ov}\left[Y_1, Y_{>1}\right].$$

The bias can be made zero in at least two ways: (1) set the prior means of the low and high fidelity models to be equal to the truth, or more generally (2) set $\nu^{MF} = 0$ and/or $\lambda = 0$. Setting the prior mean of the high-fidelity model to the true expectation is an unreasonable requirement for practical applications. However, setting the prior low-fidelity means (close) to the true low-fidelity model expectations is possible when using low-fidelity models with known statistics, which is assumed when using control variate approaches.

4.2.2 Optimality of the "correct" covariance structure

Frequentist estimators $g(\boldsymbol{y})$ are functions mapping data $\boldsymbol{y} = [\boldsymbol{y}_1\ \boldsymbol{y}_>1]$ to an estimate $\hat{\theta}$, i.e., $\hat{\theta} = g(\boldsymbol{y})$. A general procedure for comparing and choosing estimators is through a decision theoretic approach based on risk and loss functions [1]. These approaches requires choosing a loss function $L(\hat{\theta}, \theta)$ that penalizes deviation of an estimate $\hat{\theta}$ from a baseline θ. As an example, this loss could be the L_2 loss $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$. Comparing estimators also requires specification of a risk functional $R(g(\boldsymbol{y}), \theta)$ that measures how much loss is obtained by choosing an estimator g for every parameter θ. As an example, the frequentist risk of choosing an estimator g is given by taking the expectation of the loss over all possible data.

$$R(g, \theta) = \mathbb{E}_{\boldsymbol{y}|\theta}\left[L\left(g(\boldsymbol{y}), \theta\right)\right]. \qquad (36)$$

Since the true θ is in general unknown, there are several options for using the risk R within an optimization problem over g. Two of the most popular ways are *minimax risk* which seeks to minimize the worst case risk over the space Θ

$$\min_g\left[\max_{\theta\in\Theta}\left[R(g, \theta)\right]\right], \qquad (37)$$

and the *Bayes risk* which specifies a distribution (weight) over θ and minimizes

$$\min_g \mathbb{E}_\theta\left[R(g, \theta)\right] = \min_g \mathbb{E}_\theta\left[\mathbb{E}_{\boldsymbol{Y}|\theta}\left[L(g(\boldsymbol{y}), \theta)\right]\right] \qquad (38)$$

where the equality arises the definition of risk (36). Interestingly, this risk can be viewed from a Bayesian perspective as minimizing the expectation of the risk over the prior distribution. It is well known that when using the L_2 loss, the estimator g that minimizes the Bayes risk is the *posterior mean* of a Bayesian inference

problem that uses $p(\theta)$ as the prior [1, 30]. In other words $g(\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}[\theta|\boldsymbol{y}]$. These common settings lead to a straightforward result for our problem

If our prior belief is that θ is distributed with a specified distribution that encodes some conditional independence structure, e.g., those in Figure 2, then the estimator which minimizes the Bayes risk (38) *is the posterior of a Bayesian inference problem which uses this same prior.* In other words, if one used a different set of conditional independence properties that was not consistent with the specified distribution, then one would obtain a suboptimal estimator for the parameters of all the models.

4.3 Relation to control variates

In this section, our aim is to compare the variance of the optimal control variate estimator (41) to the variance of the posterior mean (35) under identical assumptions. Here the setup is a bit different from that already introduced. The goal of the control variate is to estimate the expectation of some random variable Y_1 given knowledge about: the expectation of a set of low-fidelity models $Y_{>1}$, the covariance between these models and Y_1, and the covariance amongst the low-fidelity models themselves. Explicitly, we make the following assumptions

1. $\mathbb{E}[Y_{>1}]$ is known
2. $\mathbb{C}\text{ov}[Y_{>1}, Y_{>1}]$ is known
3. $\mathbb{C}\text{ov}[Y_1, Y_{>1}]$ is known
4. Data for all models is generated via the same input realizations.

Definition 3 (Monte Carlo based (Optimal) Control Variate [9]). *Let $Y_1 \in \mathbb{R}$ and $Y_{>1} \in \mathbb{R}^{M-1}$ be two random variables and let $\mathbb{E}[Y_{>1}] = \boldsymbol{\mu}_{>1}$. The control variate estimator for the mean of Y_1 is*

$$\widehat{y}_1^{CV} = \widehat{y}_1^{MC} + \boldsymbol{\alpha}^T\left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1}\right), \tag{39}$$

where $\boldsymbol{\alpha}^T$ are the control variate weight, and $m_1 = m_{>1}$ samples are used from each random variable for the MC estimates. Moreover

$$\boldsymbol{\alpha}^* = -\mathbb{C}\text{ov}[Y_{>1}, Y_{>1}]^{-1}\,\mathbb{C}\text{ov}[Y_{>1}, Y_1] \tag{40}$$

is the optimal weight that minimizes the variance of $\widehat{y}_1^{CV}$. Leading to the variance

$$\mathbb{V}\text{ar}\left[\widehat{y}_1^{CV}\right] = \frac{\mathbb{V}\text{ar}[Y_1]}{m_1}\left(1 - \mathbb{C}\text{ov}[Y_1, Y_{>1}]\,\frac{\mathbb{C}\text{ov}[Y_{>1}, Y_{>1}]^{-1}}{\mathbb{V}\text{ar}[Y]}\mathbb{C}\text{ov}[Y_{>1}, Y_1]\right) \tag{41}$$

Our aim is to show that if this knowledge is incorporated into the Bayesian setting, we achieve the same results. This setting differs in three ways from the one we recently considered. First, the likelihood does not have any noise terms, i.e. the noise is already accounted for in the known covariance. Second, our inference task is to infer the mean of the high-fidelity model, i.e the low fidelity parameters are already known. Third, the same amount of high-fidelity and low-fidelity data is available. The result is summarized by the following theorem.

Theorem 3 (Bayesian formulation of the optimal control variate). *Assume that bullets (1-4) above hold. Let $\theta_1 \in \mathbb{R}$ be a random variable denoting uncertainty corresponding to the marginal mean of the random variable Y_1 and let $Y_{>1} \in \mathbb{R}^{M-1}$ denote a random variable denoting low-fidelity models. Furthermore let the likelihood be*

$$[Y_1, Y_{>1}]\,|\theta_1 \sim \mathcal{N}\left(\begin{bmatrix}\theta_1 \\ \boldsymbol{\mu}_{>1}\end{bmatrix}, \boldsymbol{\Sigma}\right), \tag{42}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} c_1^2 & \boldsymbol{c}^T \\ \boldsymbol{c} & \boldsymbol{\Sigma}_{>1}\end{bmatrix} \equiv \begin{bmatrix} \mathbb{V}\text{ar}[Y_1] & \mathbb{C}\text{ov}[Y_1, Y_{>1}] \\ \mathbb{C}\text{ov}[Y_{>1}, Y_1] & \mathbb{C}\text{ov}[Y_{>1}, Y_{>1}]\end{bmatrix}, \tag{43}$$

and let m_1 i.i.d samples of each Y_1 and $Y_{>1}$ be denoted by $(y_1^{(i)}, y_{>1}^{(i)})_{i=1}^{m_1}$. Now letting the prior be

$$\theta_1 \sim \mathcal{N}(\mu_1, c_\mu^2), \tag{44}$$

where c_μ^2 is the prior covariance of the mean, then the posterior mean is

$$\bar{\mu}_1 = (1-\lambda)\mu_1 + \lambda \widehat{y}_1^{MC} - \frac{\eta_1 m_1}{c_\mu^{-2} + m_1 \eta_1} \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right), \tag{45}$$

where

$$\lambda = \frac{\eta_1 m_1}{c_\mu^{-2} + m_1 \eta_1}, \quad \textit{and} \quad \eta_1 = c_1^2 - \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}, \tag{46}$$

so that in the limit $c_\mu^2 \to \infty$, we exactly recover the optimal control variate of Definition 3.

The proof is given in Appendix 8.5. The recovery of the optimal control variate occurs in the natural limit of maximum uninformitiveness $c_\mu^2 \to \infty$. Notice that in this limit the specification of the covariance c_1^2 does not impact the final result.

Our aim here has been to show that we do not lose performance in cases where we have the same information as an equivalent frequentist procedure. Indeed, this Bayesian framework is able to actually recover the control variate as the result of the inference problem. To our knowledge, this is the first derivation of the control variate from a Bayesian perspective. Moreover, we have a natural way to incorporate prior information without re-deriving new and ad-hoc estimators.

5 Numerical Examples

In this section we demonstrate the performance and algorithmic work-flow of the Bayesian approach on several examples. In particular, we show how the consideration of several graphs can resolve robustness issues in Multilevel Monte Carlo that arise due to a recursive assumption which breaks down when one of the levels is corrupted by noise. As in the previous section, in the context of purely sampling-based algorithms we assume $p_i = 1$ and $\phi_i = 1$ in the linear model of Equation (9). We first begin by describing constraints on the hyperparameters that ensure a fair comparison amongst different graphs, we then provide an example application to corrupted data for a stochastic differential equation model. Finally, we provide an example application on models with differing inputs.

5.1 Setup

Our sampling examples will primarily consider of four models with parameters $(\theta_1, \theta_2, \theta_3, \theta_4)$ relating to random variables Y_1, Y_2, Y_3, and Y_4, according to Equation (9). We will compare two graph structures

1. Peer structure $G_p = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = (\theta_1, \theta_2, \theta_3, \theta_4)$ and $\mathcal{E} = \{\theta_4 \to \theta_1,\ \theta_3 \to \theta_1,\ \theta_2 \to \theta_1\}$ with hyperparameters $(p_1, p_2, p_3, m_p, \sigma_p^2)$ and CPD defined by

$$\theta_1 = p_1\theta_2 + p_2\theta_3 + p_3\theta_4 + \xi, \quad \xi \sim \mathcal{N}(m_p, \sigma_p^2)$$

2. Recursive structure $G_r = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = (\theta_1, \theta_2, \theta_3, \theta_4)$ and $\mathcal{E} = \{\theta_4 \to \theta_3, \theta_3 \to \theta_2,\ \theta_2 \to \theta_1\}$ with hyperparameters $(r_1, r_2, r_3, m_{r1}, m_{r2}, m_{r3}, \sigma_{r1}^2, \sigma_{r2}^2)$ and CPD defined by

$$\begin{aligned} \theta_1 &= r_1\theta_2 + \xi_1 \quad \xi_1 \sim \mathcal{N}(m_{r1}, \sigma_{r1}^2) \\ \theta_2 &= r_2\theta_3 + \xi_2 \quad \xi_2 \sim \mathcal{N}(m_{r2}, \sigma_{r2}^2) \\ \theta_3 &= r_3\theta_4 + \xi_3 \quad \xi_3 \sim \mathcal{N}(m_{r3}, \sigma_{r3}^2) \end{aligned}$$

From data $\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3$, and $\boldsymbol{y}_4$, we will compare the posteriors of the peer graph

$$\hat{\theta}_{1p} \equiv \mathbb{E}\left[\theta_1 | \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4, G_p\right] \tag{47}$$

with the posterior of the recursive graph

$$\hat{\theta}_{1r} \equiv \mathbb{E}\left[\theta_1 | \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4, G_r\right]. \tag{48}$$

To create a fair approximation we must ensure that, regardless of the choice of hyperparameters, the resulting prior covariance for each choice of graph is in some sense equal. This type of constraint is typically not considered in the literature; where the hyperparameters are tuned *independently* of acknowledging the prior information. Such tuning procedures which produce more informative priors for certain values of the tuning parameters during optimization, can potentially result to misleading conclusions about the effectiveness of an approach. Here we seek to address this concern.

While different graph structure prohibits the construction of identical prior covariances, some common properties should still be enforced. Foremost we *ensure that the marginal distributions of each parameter are identically distributed regardless of the graph.* Let these distributions be denoted Gaussian with varying means but identical variances

$$\theta_1 \sim \mathcal{N}(m_1, \sigma_1) \quad \theta_2 \sim \mathcal{N}(m_2, \sigma_2) \quad \theta_3 \sim \mathcal{N}(\sigma_3, 1) \quad \theta_4 \sim \mathcal{N}(\sigma_4, 1). \tag{49}$$

For convenience we set $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$. These fixed marginals lead to the following constraints. The allowable parameters for the peer structure lie in the set $(p_1, p_2, p_3, \sigma_p^2) \in \mathcal{H}_p$ where

$$\mathcal{H}_p = \left\{ (p_1, p_2, p_3, \sigma_p^2) \, ; \quad p_1^2 + p_2^2 + p_3^2 + \sigma_p^2 = 1, \quad m_1 = p_1 m_2 + p_2 m_3 + p_3 m_4 + m_p \right\}. \tag{50}$$

This set ensures that the marginal variance of θ_1 is 1 and the mean is m_1. Similarly, the parameters for the recursive structure lie in the set $(r_1, r_2, r_3, \sigma_{r1}^2, \sigma_{r2}^2, \sigma_{r3}^2) \in \mathcal{H}_p$ defined by

$$\begin{aligned} \mathcal{H}_r = \Bigg\{ (r_1, r_2, r_3, \sigma_{r1}^2, \sigma_{r2}^2, \sigma_{r3}^2) \, ; \quad & r_1^2 + \sigma_{r1}^2 = 1 \quad m_2 = r_1 m_1 + m_{r1}, \\ & r_2^2 + \sigma_{r2}^2 = 1, \quad m_2 = r_2 m_3 + m_{r2}, r_3^2 + \sigma_{r3}^2 = 1, \quad m_3 = r_3 m_4 + m_{r3}. \Bigg\} \end{aligned} \tag{51}$$

There are many approaches to tune hyperparameters in these linear Gaussian models [28]. Here we pick a cross validation scheme that minimizes the MSE (Bias of the posterior mean plus the posterior variance). This approach can directly leverage Algorithm 1 to compute the marginal of the data, and then evaluate this marginal at the given data.

5.2 Multilevel model with corrupt data

Multilevel Monte Carlo [6] enhances the standard Monte Carlo approach by rearranging the target expectation of Y_1 via a recursive sum

$$\mathbb{E}\left[Y_1\right] = \mathbb{E}\left[Y_M\right] + \sum_{i=M-1}^{2} \mathbb{E}\left[Y_{i-1} - Y_i\right]. \tag{52}$$

This rearrangement can yield a very effective algorithm when the variance of the discrepancy between model outputs at successive discretization levels $\mathbb{V}\text{ar}[Y_{i-1} - Y_i]$ decreases. In this situation, the number of samples allocated to each model decreases with increasing fidelity. This resource allocation can result in a large reduction in the amount of data required when compared to single fidelity MC [7].

MLMC assumes that the model data exhibit a hierarchical graph structure and when this structure is not present the effectiveness of MLMC deteriorates. For example consider three finite element models corresponding to increasing mesh resolution. The performance of MLMC will deteriorate medium fidelity mesh is not perfectly adapted to the geometry, or if there are in general poorly tuned tolerances in the solver. Here, we will mimic these effects by corrupting the medium fidelity data.

Consider the model of geometric Brownian motion used in [6] as a surrogate for option pricing

$$dX_t = rX_t dt + \sigma X_t dW_t, \quad X_0 = 1 \quad t \in [0, 1] \tag{53}$$

where W_t is a Wiener process. This stochastic differential equation (SDE) via the Euler-Maruyama method, where successive iterates are obtained according to

$$X_{t_{k+1}} = (1 + \Delta t r) X_{t_k} + \sigma X_{t_k} \xi, \quad \xi \sim \mathcal{N}(0, \Delta t), \tag{54}$$

and we consider a final European option payoff

$$P = \exp(-r)\max(0, X_1 - 1), \tag{55}$$

where $r = 0.05$ and $\sigma = 0.2$. We consider four resolution levels $\Delta t = (1, \frac{1}{4}, \frac{1}{16}, \frac{1}{64})$. The high fidelity model output is denoted by Y_1 corresponding to $\Delta t = 1/64$, the medium fidelity output is denoted by Y_2 and uses $\Delta t = 1/16$, the second medium fidelity fidelity output is denoted by Y_3 and uses $\Delta t = \frac{1}{4}$, and the coarsest model is Y_4 using $\Delta t = 1$. For reference, this setup can be used as the *inner loop* of an adaptive MLMC scheme that then also refines the time instance [7].

We will model corruption via the following process. With frequency f we will add Gaussian noise with an offset of five and a standard deviation one to the output of a simulation of the medium fidelity model. We will then consider what happens for all $f \in (0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2})$. Figure 4 shows the variance of the discrepancy for each level of MLMC. We see that when there is no corruption ($f = 0$), then we get the standard decay of the variance, indicating a highly effective MLMC scheme. When $f = 10^{-2}$, and only one out of every 100 samples is corrupted, the variance decay is no longer present and MLMC loses its efficiency. The noise level of each model is assumed to have a standard deviation of one.

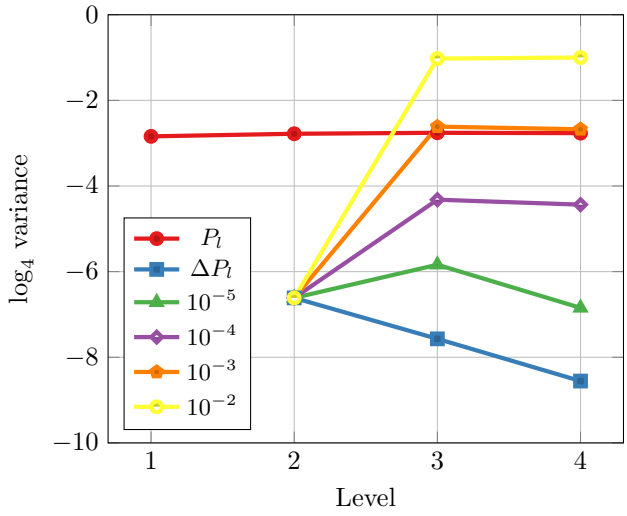


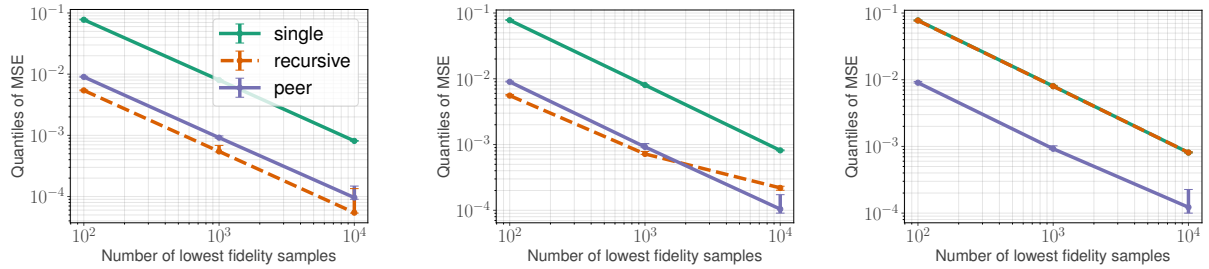
Figure 4: Variance of the discrepancy between levels for the European option payoff with the Euler Maruyama Scheme. The red line indicates the variance of each individual level, and the blue line indicates the variance of the discrepancies when there is no corrupted data. Since it decays with level (level 4 is the finest grid), MLMC is highly effective here. However, when the frequency of corruption increases (green, purple, orange and yellow lines), the discrepancy between levels no longer decays and MLMC performance suffers.

Figure 5 shows the MSE of three competing models as corruption increases: a single high-fidelity estimator, the peer graph, and (3) the recursive graph. Here, the MSE refers to the Bias squared of the posterior mean plus the posterior variance. Here we vary the number of low fidelity samples, and in each case the number of samples in the next finest grid is halved.

Clearly, the recursive approaches begin to break down, while the peer graph consistently performs well. This performance enhancement occurs because the peer graph can still use the lowest fidelity model to inform the high fidelity model. Whereas the recursive graph has to work just with the high fidelity data. To summarize this example, we show that considering a variety of network structures allows us to obtain better performance than maintaining a single structure. Furthermore, we are able to detect when one structure begins to significantly outperform another.

5.3 Fusing heterogeneous data

In this section we demonstrate that latent variable inference enables MF analysis using heterogeneous data, that is different input/output spaces. This is achieved by leveraging disparate latent variable parameteriza-



(a) No corrupted data (b) 1% corruption (c) 10% corruption

Figure 5: Performance of a single high-fidelity, recursive, and peer graph models. For the lowest corruption level we see that the recursive model, as expected, has the lowest MSE. As the corruption increases, the recursive model can no longer leverage the intermediate fidelities and defaults to the performance of a single high-fidelity model. The peer model performance does not degrade.

tions for each model. Consider the following three models

$$f_3(x_1, x_2) = \cos(3\pi x_1 + 0.1x_2) \tag{56}$$
$$f_2(x_1) = \exp\left(-2(x - 0.5)^2\right) \tag{57}$$
$$f_1(x_1) = f_2(x_1) + \cos(3\pi x_1) \tag{58}$$

The "low-fidelity" model f_3 has two inputs x_1 and x_2 whereas the other models are functions of only x_1. In the following we will build an approximation of the high fidelity model f_1 using both low-fidelity models. Existing methods are unable to build such an approximation as they assume all models have the same inputs.

To build a MF approximation we must first define the Vandermonde-like matrices $\mathbf{\Phi}_i$ for each model. We use total-degree Legendre polynomial approximations and set the degree of $\mathbf{\Phi}_1$ to 3 and the degree of the other polynomials to 5. The first first polynomial consists of 10 terms and the latter polynomials 6. Next we randomly sample the inputs x_i for each model from independent uniform distributions on the interval $[0, 1]$ and add mean zero independent Gaussian noise with standard deviation $\sigma_i = 0.01$. We use 10 samples from f_2 and f_3 and 2 samples of the high-fidelity model f_1. The approximation of the high-fidelity data is dependent on the marginal prior covariance of each model ,which we set to 1 for all models, and the matrices $\boldsymbol{A}_{i|\mathrm{pa}(i)}$ and vectors $\boldsymbol{b}_{i|\mathrm{pa}(i)}$ in (14). Here we set $\boldsymbol{A}_{i|\mathrm{pa}(i)} = \mathrm{diag}(0.5, -0.5)$ and $\boldsymbol{b}_{i|\mathrm{pa}(i)} = \mathbf{0}$ for both CPDs connecting models f_3 and f_2 to f_1.

Figure 6 compares a single fidelity Bayesian approximation against multi-fidelity approximations using two and three models. The two model MF approximation uses models with homogeneous parameterizations where as the three model approximation uses models with different numbers of parameters. It is clear that the three model approximation is more accurate than both the single fidelity and two fidelity approximations. These results are dependent on the chosen values of $\boldsymbol{A}_{i|\mathrm{pa}(i)}$ and the number of samples of each model. Our points here is to illustrate that use of heterogeneous data is possible. A extensive exploration of surrogate based MFNets is beyond the scope of this paper. In future work we will investigate how to tune the entries of $\boldsymbol{A}_{i|\mathrm{pa}(i)}$ and how to optimally allocate samples to each model minimize the error in the final surrogate of the high-fidelity model.

6 Conclusions

In this paper we developed a Bayesian framework for multi-fidelity data-driven modeling, MFNets for short. The approach presented uses data from an ensemble of sources of varying credibility and cost to compute statistics and/or construct data-driven surrogates. State-of-the-art MF strategies have three fundamental limitations: (1) data sources must have the same parameterization to enable the fusing of simulation and experimental data; and (2) approaches fail when a priori assumptions about the relationships among data are invalid; and (3) they (generally) do not provide uncertainty estimates.

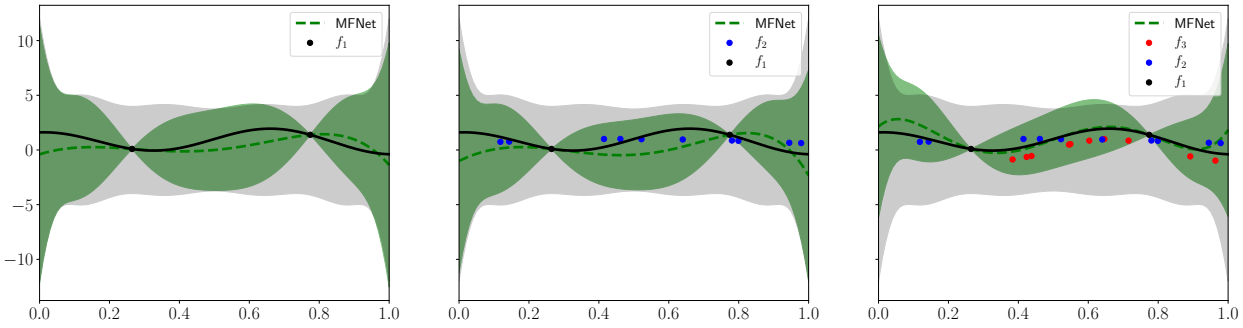


Figure 6: Single fidelity Bayesian approximation vs multi-fidelity approximation using heterogeneous data. (Left) Single fidelity using high-fidelity data only. (Middle) Two model MFNet with homogeneous data. (Right) Three model peer-based MFNet with heterogeneous data. Dots represent training data of x_1, the dashed line the MF approximation and the solid line the true high-fidelity function. The training values of x_2 used to evaluate f_1 are not plotted. The bounds of the gray region represent the prior marginal mean of all models plus and minus two standard deviations of the marginal prior and the bounds on the green region the posterior marginal mean of the high-fidelity model plus and minus two standard deviations of the high-fidelity marginal posterior.

To address the limitations of current MF approaches, we developed a generalized MF framework based upon latent variables and conditional independence. Firstly, the use of latent variables enables the fusion of heterogeneous data sources with potentially different parameterizations, e.g simulation models with different uncertain parameters, or data sets collected under different environmental conditions. Secondly, the use of conditional independence as a modeling tool allows the practitioner to both encode prior knowledge, as well as obtain efficient inference procedures. Finally, the use of Bayesian inference provides estimates of uncertainties in estimated statistics and surrogates.

The ability to leverage knowledge about the underlying structure of the data generation process represents a fundamental advancement relative to existing MF approaches. The state-of-the-art Bayesian approaches, multi-fidelity co-kriging and recursive co-kriging, respectively restrict attention to the extreme cases when all models are correlated and when these models admit a natural hierarchy. Here, we provided upon a general set of tools that Bayesian inference provides for handling a wider variety of problems and problem structures. Moreover, our framework is flexible enough that it provides an alternative to pure Monte Carlo, Multilevel Monte Carlo, and control variate approaches without the usual smoothness assumptions that are required by kriging/Gaussian process-based approaches.

Theoretically, our contributions include the derivation of analytical expressions for the marginal posterior of a high-fidelity model for arbitrary covariance structures. Furthermore, we provide the first (as far as we are aware) derivation of the control variate algorithm simply through Bayes rule on a linear-Gaussian model and show that *under the same assumptions* used in a frequentist procedure, we can recover the frequentist estimator from the Bayesian posterior. Unlike control variate MC, however, our approach is much more flexible in that the prior allows for naturally encoding a wide variety of information and through the use of Bayes rule. MFNets removes the need to derive new estimators for every new situation, which has contributed to the large number of MF algorithms in the existing literature.

To build upon our theoretical results we empirically demonstrated how learning the graphical connections for a given problem can improve robustness. In particular, we show how the proposed framework can address the degradation of accuracy in Multilevel Monte Carlo estimators that arises when data from models in the hierarchy are corrupted. We also demonstrated how to construct surrogate models that are parametric equivalents to the co-kriging approaches, but unlike co-kriging approaches can be applied to data sources with heterogeneous parameterizations.

Our proposed framework provides a general way to create a generative model of a multi-fidelity system. As a result, there are many avenues for future work, both in expanding the Bayesian machinery and building more complex models. In particular, we can consider experimental design for developing sample allocation strategies; graph learning algorithms for updating the structure in a systematic way in response to data; and

nonlinear models (such as artificial neural networks) for representing more complex model relationships and data generating processes.

7 Acknowledgments

The authors were supported by the Laboratory Directed Research Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[2] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. *Computing and Visualization in Science*, 14(1):3, 2011.

[3] M. G. Fernández-Godino, C. Park, N.-H. Kim, and R. T. Haftka. Review of multi-fidelity models. *arXiv preprint arXiv:1609.07196*, 2016.

[4] A. I. Forrester, A. Sóbester, and A. J. Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 463(2088):3251–3269, 2007.

[5] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag New York, Inc., New York, NY, USA, 1991.

[6] M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[7] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

[8] A. Gorodetsky and Y. Marzouk. Mercer kernels and integrated variance experimental design: Connections between gaussian process regression and polynomial approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):796–828, 2016.

[9] A. A. Gorodetsky, G. Geraci, M. Eldred, and J. D. Jakeman. A generalized framework for approximate control variates. *arXiv preprint arXiv:1811.04988*, 2018.

[10] L. L. Gratiet and J. Garnier. Recursive co-kriging model for design of experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5):365–386, 2014.

[11] A. Haji-Ali, F. Nobile, L. Tamellini, and R. Tempone. Multi-index stochastic collocation for random pdes. *Computer Methods in Applied Mechanics and Engineering*, 306:95 – 122, 2016.

[12] A.-L. Haji-Ali, F. Nobile, and R. Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numerische Mathematik*, 132(4):767–806, Apr 2016.

[13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[14] J. Jakeman, M. Eldred, G. Geraci, and A. Gorodetsky. Adaptive multi-index collocation and sensitivity analysis. *International Journal for Numerical Methods in Engineering*, 2018. Accepted, arXiv:1909.13845.

[15] M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[16] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[17] D. E. Myers. Matrix formulation of co-kriging. *Journal of the International Association for Mathematical Geology*, 14(3):249–257, 1982.

[18] A. Narayan, C. Gittelson, and D. Xiu. A stochastic collocation algorithm with multifidelity models. *SIAM Journal on Scientific Computing*, 36(2):A495–A521, 2014.

[19] B. L. Nelson. On control variate estimators. *Computers and Operations Research*, 14(3):219 – 225, 1987.

[20] L. W. Ng and K. E. Willcox. Multifidelity approaches for optimization under uncertainty. *International Journal for numerical methods in Engineering*, 100(10):746–772, 2014.

[21] L. W.-T. Ng and M. Eldred. Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA*, page 1852, 2012.

[22] A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[23] R. Pasupathy, B. W. Schmeiser, M. R. Taaffe, and J. Wang. Control-variate estimation using estimated control means. *IIE Transactions*, 44(5):381–385, 2012.

[24] B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016.

[25] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Preprint*, pages 1–57, 2016.

[26] P. Perdikaris, D. Venturi, J. O. Royset, and G. E. Karniadakis. Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

[27] P. Z. Qian and C. J. Wu. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204, 2008.

[28] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[29] R. Y. Rubinstein and R. Marcus. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677, 1985.

[30] L. Tenorio. *An introduction to data analysis and uncertainty quantification for inverse problems*, volume 3. SIAM, 2017.

[31] S. Venkatraman and J. R. Wilson. The efficiency of control variates in multiresponse simulation. *Operations Research Letters*, 5(1):37 – 42, 1986.

[32] Y. Yang and P. Perdikaris. Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems. *Computational Mechanics*, 64(2):417–434, Aug 2019.

8 Appendix

8.1 Inverse of 2×2 matrices

The following two properties hold for the matrix inverse of a partitioned matrix [13] into blocks of appropriate sizes

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix} \tag{59}$$

$$= \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix} \tag{60}$$

8.2 Proof of Proposition 1

Proof The proof is uses Equations (19) and (18). For the mean we have

$$\begin{aligned} \bar{\mu}_1 &= \mu_1 + \frac{1}{\sigma_1^2/c_1^2 + m_1}\left(\sum_{i=1}^{m_1} y_1^{(i)} - m_1\mu_1\right) \\ &= \mu_1 + \frac{m_1}{\sigma_1^2/c_1^2 + m_1}\left(\widehat{y}_1^{MC} - \mu_1\right) = \frac{\sigma_1^2/(m_1c_1^2)}{\sigma_1^2/(m_1c_1^2)+1}\mu_1 + \frac{1}{\sigma_1^2/(m_1c_1^2)+1}\widehat{y}_1^{MC} \\ &= \frac{\eta}{\eta+1}\mu_1 + \frac{1}{\eta+1}\widehat{y}_1^{MC} = \nu\mu_1 + (1-\nu)\,\widehat{y}_1^{MC}. \end{aligned}$$

For the covariance we have

$$\bar{c}_1^2 = \left(1 - \frac{m_1}{\sigma_1^2/c_1^2 + m_1}\right)c_1^2 = \left(1 - \frac{1}{\sigma_1^2/(m_1c_1^2)+1}\right)c_1^2 = \frac{\eta}{\eta+1}c_1^2 = \nu c_1^2. \tag{61}$$

□

8.3 Proof of Theorem 1

Proof Since the MSE of the MC estimator is its variance, we seek to determine when $\frac{\mathrm{MSE}(\bar{\mu}_1)}{\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right]} < 1$. Using Equation (23) we obtain

$$\frac{\mathrm{MSE}(\bar{\mu}_1)}{\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right]} = a\nu^2 + (1-\nu)^2,$$

where $a = \frac{(\mathbb{E}[Y]-\mu_1)^2}{\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right]}$. Thus we need $a\nu^2 + (1-\nu)^2 \leq 1$ leading to

$$a \leq \frac{1-(1-\nu)^2}{\nu^2} = \frac{2\nu-\nu^2}{\nu^2} = \frac{2-\nu}{\nu} = \frac{2}{\nu} - 1 = 2\frac{\eta+1}{\eta} - 1 = \frac{\eta+2}{\eta} = 1 + 2m_1\frac{c_1^2}{\sigma_1^2}$$

and therefore

$$\left(\mathbb{E}\left[Y_1\right] - \mu_1\right)^2 \leq \left(1 + 2m_1\frac{c_1^2}{\sigma_1^2}\right)\mathbb{V}\mathrm{ar}\left[\widehat{y}_1^{MC}\right] = \left(1 + 2m_1\frac{c_1^2}{\sigma_1^2}\right)\frac{1}{m_1}\mathbb{V}\mathrm{ar}\left[Y_1\right] = \left(\frac{1}{m_1} + 2\frac{c_1^2}{\sigma_1^2}\right)\mathbb{V}\mathrm{ar}\left[Y_1\right].$$

□

8.4 Proof of Theorem 2

Proof The proof occurs in three phases. First, we obtain an analytic expression for the inverse prior covariance, then we compute certain elements of the posterior covariance, and finally we compute the posterior mean.

We begin by partitioning and inverting the prior covariance matrix. Let

$$\eta_1 = \left(c_1^2 - \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)^{-1}, \tag{62}$$

then the partitioned-matrix inverse property Equation (60) implies

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} c_1^2 & \boldsymbol{c}^T \\ \boldsymbol{c} & \boldsymbol{\Sigma}_{>1} \end{bmatrix}^{-1} = \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \\ Sym & \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)^{-1} \end{bmatrix}. \tag{63}$$

We can now use Equation (18) with this inverse prior and $\boldsymbol{\Gamma}$ to obtain a block partitioned posterior covariance

$$\bar{\boldsymbol{\Sigma}} = \begin{bmatrix} \bar{c}_1^2 & \bar{\boldsymbol{c}}^T \\ \bar{\boldsymbol{c}} & \bar{\boldsymbol{\Sigma}}_{>1} \end{bmatrix} = \begin{bmatrix} \eta_1 + m_1/\sigma_1^2 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \\ Sym & \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)^{-1} + \boldsymbol{\Gamma}_{>1} \end{bmatrix}^{-1} \tag{64}$$

where $\bar{c}_1^2$ is the marginal posterior covariance corresponding to the high-fidelity model θ_1.

Covariance Calculation: Next we consider computing the upper left and upper right terms of the posterior covariance. A closed form expression for $\bar{c}_1^2$ by again using Equation (60)

$$\bar{c}_1^2 = \left[\eta_1 + m_1/\sigma_1^2 - \eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \left(\left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)^{-1} + \boldsymbol{\Gamma}_{>1}\right)^{-1} \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right]^{-1} \tag{65}$$

We now work to simplify the third term in this equation. Let

$$\boldsymbol{A} = \left(\left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)^{-1} + \boldsymbol{\Gamma}_{>1}\right)^{-1} \quad \text{and} \tag{66}$$

The Woodbury identity provides the following relationship between these two matrices.

$$\boldsymbol{A} = \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} - \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right) \boldsymbol{B} \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2}\right)\right). \tag{67}$$

Using $\boldsymbol{A}$ in Equation (65) yields

$$\bar{c}_1^2 = \left[\eta_1 + m_1/\sigma_1^2 - \eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A} \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right]^{-1}. \tag{68}$$

If we replace A using Equation (67) and distribute the multiplications we obtain

$$\begin{aligned}
\eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A} \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} &= \eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} - \frac{\eta_1^2}{c_1^2} \left(\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)^2 - \eta_1^2 \left(\boldsymbol{c}^T - \frac{\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}}{c_1^2} \boldsymbol{c}^T\right) \boldsymbol{B} \left(\boldsymbol{c} - \boldsymbol{c} \frac{\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}}{c_1^2}\right) \\
&= \eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} - \frac{\eta_1^2}{c_1^2} \left(\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)^2 - \eta_1^2 \left(1 - \frac{\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}}{c_1^2}\right)^2 \boldsymbol{c}^T \boldsymbol{B} \boldsymbol{c} \\
&= \eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} - \frac{\eta_1^2}{c_1^2} \left(\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)^2 - \eta_1^2 \left(1 - \frac{\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}}{c_1^2}\right)^2 \gamma c_1^2.
\end{aligned}$$

The following simplifications can now be made

$$\begin{aligned}
\eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} &= \eta_1^2 \left(c_1^2 - \eta_1^{-1}\right) = \eta_1^2 c_1^2 - \eta_1 \\
\frac{\eta_1^2}{c_1^2} \left(\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)^2 &= \frac{\eta_1^2}{c_1^2} \left(c_1^2 - \eta_1^{-1}\right)^2 = \frac{\eta_1^2}{c_1^2} \left(c_1^4 - 2\frac{c_1^2}{\eta_1} + \frac{1}{\eta_1^2}\right) = \eta_1^2 c_1^2 - 2\eta_1 + \frac{1}{c_1^2} \\
\eta_1^2 \gamma c_1^2 \left(1 - \frac{\left(\boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}\right)}{c_1^2}\right)^2 &= \eta_1^2 \gamma c_1^2 \left(1 - \frac{c_1^2 - \eta_1^{-1}}{c_1^2}\right)^2 = \eta_1^2 \gamma c_1^2 \left(1 - 1 + \frac{1}{\eta_1 c_1^2}\right)^2 = \frac{\gamma}{c_1^2}.
\end{aligned}$$

Combining these terms yields

$$\eta_1^2 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A} \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} = \eta_1 - \frac{1}{c_1^2} - \frac{\gamma}{c_1^2} = \eta_1 - \frac{1+\gamma}{c_1^2}. \tag{69}$$

Using this expression within Equation (68) we obtain the marginal posterior covariance of the high-fidelity model

$$\bar{c}_1^2 = \left[m_1/\sigma_1^2 + \frac{1+\gamma}{c_1^2} \right]^{-1} = c_1^2 \left[c_1^2 m_1/\sigma_1^2 + 1 + \gamma \right]^{-1} = c_1^2 \left[\eta^{-1} + 1 + \gamma \right]^{-1} = c_1^2 \frac{\eta}{1+\eta(1+\gamma)}. \tag{70}$$

For the subsequent calculations of the posterior mean, it will also be useful to have an expression for $\bar{\boldsymbol{c}}^T$. Again using the inverse of a partitioned matrix (60) we obtain

$$\bar{\boldsymbol{c}}^T = \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \left(\left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} + \boldsymbol{\Gamma}_{>1} \right)^{-1} = \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A}. \tag{71}$$

Also useful will be the following expression

$$\begin{aligned}
\bar{\boldsymbol{c}}^T \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} &= \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A} \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} = \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \left(\boldsymbol{I} - \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right) \boldsymbol{B} \right) && (72) \\
&= \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} - \bar{c}_1^2 \eta_1 \left(\boldsymbol{c}^T \boldsymbol{B} - \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \boldsymbol{B} \right) && (73) \\
&= \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} - \frac{\bar{c}_1^2}{c_1^2} \eta_1 \left(c_1^2 - \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} \right) \boldsymbol{c}^T \boldsymbol{B} && (74) \\
&= \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} - \frac{\bar{c}_1^2}{c_1^2} \boldsymbol{c}^T \boldsymbol{B}, && (75)
\end{aligned}$$

where the second equality follows from Equation (67) and the fourth inequality follows from Equation (62).
Mean Calculation: Substituting the appropriate quantities into Equation (26) leads to

$$\begin{aligned}
\begin{bmatrix} \bar{\mu}_1 \\ \bar{\boldsymbol{\mu}}_{>1} \end{bmatrix} &= \begin{bmatrix} \bar{c}_1^2 & \bar{\boldsymbol{c}}^T \\ \bar{\boldsymbol{c}} & \bar{\boldsymbol{\Sigma}}_{>1} \end{bmatrix} \begin{bmatrix} m_1/\sigma_1^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{>1} \end{bmatrix} \begin{bmatrix} \widehat{y}_1^{MC} \\ \widehat{\boldsymbol{y}}_{>1}^{MC} \end{bmatrix} + \begin{bmatrix} \bar{c}_1^2 & \bar{\boldsymbol{c}}^T \\ \bar{\boldsymbol{c}} & \bar{\boldsymbol{\Sigma}}_{>1} \end{bmatrix} \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \\ Sym & \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix} \\
&= \begin{bmatrix} \bar{c}_1^2 m_1/\sigma_1^2 \widehat{y}_1^{MC} + \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1} \widehat{\boldsymbol{y}}_{>1}^{MC} \\ \text{notneeded} \end{bmatrix} + \begin{bmatrix} \bar{c}_1^2 & \bar{\boldsymbol{c}}^T \\ \bar{\boldsymbol{c}} & \bar{\boldsymbol{\Sigma}}_{>1} \end{bmatrix} \begin{bmatrix} \eta_1 \mu_1 - \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{\mu}_{>1} \\ -\eta_1 \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} \mu_1 + \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \boldsymbol{\mu}_{>1} \end{bmatrix}
\end{aligned}$$

Our focus is on the marginal posterior mean of the high fidelity model. We can extract this equation and partition it into two terms corresponding to the prior and MC of the HF and LF models separately

$$\bar{\mu}_1 = \left(\bar{c}_1^2 m_1/\sigma_1^2 \right) \widehat{y}_1^{MC} + \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1} \widehat{\boldsymbol{y}}_{>1}^{MC} + \bar{c}_1^2 \eta_1 \left(\mu_1 - \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{\mu}_{>1} \right) + \bar{\boldsymbol{c}}^T \left(-\eta_1 \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} \mu_1 + \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \boldsymbol{\mu}_{>1} \right) \tag{76}$$

$$= A_1 \left(\mu_1, \widehat{y}_1^{MC} \right) + A_2 \left(\boldsymbol{\mu}_{>1}, \widehat{\boldsymbol{y}}_{>1}^{MC} \right) \tag{77}$$

where A_1 and $A_{>1}$ are the functions

$$A_1(\mu_1, \widehat{y}_1^{MC}) = \eta_1 \left(\bar{c}_1^2 - \bar{\boldsymbol{c}}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} \right) \mu_1 + \left(\bar{c}_1^2 m_1/\sigma_1^2 \right) \widehat{y}_1^{MC} \tag{78}$$

$$A_{>1}(\boldsymbol{\mu}_{>1}, \widehat{\boldsymbol{y}}_{>1}^{MC}) = \left(-\bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} + \bar{\boldsymbol{c}}^T \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \right) \boldsymbol{\mu}_{>1} + \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1} \widehat{\boldsymbol{y}}_{>1}^{MC} \tag{79}$$

We first consider A_1. Using Equation (71), $\bar{\boldsymbol{c}}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c}$ becomes

$$\bar{\boldsymbol{c}}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} = \bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{A} \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{c} = \bar{c}_1^2 \left[1 - \frac{1+\gamma}{c_1^2 \eta_1} \right], \tag{80}$$

where the second equality uses Equation (69). Substituting this result into A_1 yields

$$A_1 = \eta_1 \left(\bar{c}_1^2 - \bar{c}_1^2 \left[1 - \frac{1+\gamma}{c_1^2 \eta_1} \right] \right) \mu_1 + \bar{c}_1^2 m_1 / \sigma_1^2 \widehat{y}_1^{MC} = \bar{c}_1^2 \left(\frac{1+\gamma}{c_1^2} \right) \mu_1 + \bar{c}_1^2 m_1 / \sigma_1^2 \widehat{y}_1^{MC} \tag{81}$$

Using the definition $\eta = \sigma_1^2/(m_1 c_1^2)$ and the expression for $\bar{c}_1^2$ given in Equation (70) we can write A_1 as a convex combination of the prior mean and MC estimate

$$A_1 \left(\mu_1, \widehat{y}_1^{MC} \right) = \frac{\eta(1+\gamma)}{1+\eta(1+\gamma)} \mu_1 + \frac{1}{1+\eta(1+\gamma)} \widehat{y}_1^{MC} = \lambda \mu_1 + (1-\lambda) \widehat{y}_1^{MC}. \tag{82}$$

Next we considers $A_{>1}$ by first using Equation (71) to obtain

$$\bar{c}_1^2 \eta \bar{\boldsymbol{c}}^T \boldsymbol{\Sigma}_{>1}^{-1} = \bar{\boldsymbol{c}}^T \left(\left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} + \boldsymbol{\Gamma}_{>1} \right), \tag{83}$$

which yields

$$A_{>1} = \left(-\bar{\boldsymbol{c}}^T \left(\left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} + \boldsymbol{\Gamma}_{>1} \right) + \bar{\boldsymbol{c}}^T \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \right) \boldsymbol{\mu}_{>1} + \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1} \widehat{y}_{>1}^{MC} = \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1} \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right). \tag{84}$$

We can generally rewrite Equation (79) as

$$A_{>1}(\boldsymbol{\mu}_{>1}, \widehat{y}_{>1}^{MC}) = \boldsymbol{C} \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right) \tag{85}$$

where

$$\boldsymbol{C} = \bar{\boldsymbol{c}}^T \boldsymbol{\Gamma}_{>1}, \tag{86}$$

and

$$\boldsymbol{C} = \left(\bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} - \bar{\boldsymbol{c}}^T \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \right) \tag{87}$$

Comparing this expression with Equation (79), we see that we can rewrite it equivalently as

$$A_{>1} = \left(\bar{c}_1^2 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} - \bar{\boldsymbol{c}}^T \left(\boldsymbol{\Sigma}_{>1} - \frac{\boldsymbol{c}\boldsymbol{c}^T}{c_1^2} \right)^{-1} \right) \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right) \tag{88}$$

Now using Equation (75) we obtain

$$A_{>1} = \frac{\bar{c}_1^2}{c_1^2} \boldsymbol{c}^T \boldsymbol{B} \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right) = \nu^{MF} \boldsymbol{c}^T \boldsymbol{B} \left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1} \right). \tag{89}$$

Here again we see the direct connection to control variates. If the prior mean $\boldsymbol{\mu}_{>1}$ is the true mean of $Y_{>1}$, then the last term drops out because the MC is unbiased, and we are left with only the first term. However, if the prior mean is chosen to depend on the data, e.g., the MC estimate of the data, then the first term also drops out and the posterior mean is an unbiased estimator.

These expressions for A_1 and $A_{>1}$ complete the proof. □

8.5 Proof of Theorem 3

Proof The strategy is to write this problem in the linear-Gaussian format and to apply Equation (19) for the mean. We will also rely on the form of the inverse of the covariance matrix $\boldsymbol{\Sigma}$ given by Equation (63) in Appendix 8.4.

We begin by reformulating the likelihood into an equivalent linear-Gaussian format

$$\begin{bmatrix} y_1 \\ y_{>1} \end{bmatrix} = \begin{bmatrix} 1 \\ \boldsymbol{0}_{M-1} \end{bmatrix} \theta + \xi, \quad \xi \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ \boldsymbol{\mu}_{>1} \end{bmatrix}, \boldsymbol{\Sigma} \right) \tag{90}$$

Equivalently, for all m_1 points we can write

$$\boldsymbol{y} = (\mathbf{1}_{m_1} \otimes \phi)\theta + \boldsymbol{\xi}, \quad \xi \sim \mathcal{N}\left(\mathbf{1}_{m_1} \otimes \boldsymbol{\mu}, \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}\right) \tag{91}$$

If we let $\boldsymbol{\phi} = \mathbf{1}_{m_1} \otimes \phi$. Then the marginal covariance of the data becomes

$$\boldsymbol{C} = c_\mu^2 \boldsymbol{\phi}\boldsymbol{\phi}^T + \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}. \tag{92}$$

Now we need the inverse of this marginal covariance and the Woodbury identity yields

$$\boldsymbol{C}^{-1} = \left(c_\mu^2 \boldsymbol{\phi}\boldsymbol{\phi}^T + \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}\right)^{-1} = \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1} - \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\phi} \left(c_\mu^{-2} + \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\phi}\right)^{-1} \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right).$$

Two components of this identity can be evaluated as

$$\begin{aligned}
\boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) &= \left(\mathbf{1}_{m_1}^T \otimes \phi^T\right) \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) = \left(\mathbf{1}_{m_1}^T \boldsymbol{I}_{m_1}\right) \otimes \left(\phi^T \boldsymbol{\Sigma}^{-1}\right) = \mathbf{1}_{m_1}^T \otimes \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \end{bmatrix} \\
\boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\phi} &= \left(\mathbf{1}_{m_1}^T \otimes \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \end{bmatrix}\right) \left(\mathbf{1}_{m_1} \otimes \phi\right) = m_1 \eta_1
\end{aligned}$$

where we used Equation (63) for the inverse of the prior covariance. The next required component is

$$\begin{aligned}
\boldsymbol{\phi}^T \boldsymbol{C}^{-1} &= \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) - \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\phi} \left(c_\mu^{-2} + \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\phi}\right)^{-1} \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \\
&= \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) - m_1 \eta_1 \left(c_\mu^{-2} + m_1 \eta_1\right)^{-1} \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \\
&= \left(1 - m_1 \eta_1 \left(c_\mu^{-2} + m_1 \eta_1\right)^{-1}\right) \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \\
&= \frac{c_\mu^{-2}}{c_\mu^{-2} + m_1 \eta_1} \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right)
\end{aligned}$$

so that Equation (19) for the posterior mean becomes

$$\bar{\mu}_1 = \mu_1 + c_\mu^2 \boldsymbol{\phi}^T \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{m}) = \mu_1 + \frac{1}{c_\mu^{-2} + m_1 \eta_1} \boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \left(\boldsymbol{y} - \left(\mathbf{1}_{m_1} \otimes \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix}\right)\right). \tag{93}$$

Then the prior term becomes

$$\boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \left(\mathbf{1}_{m_1} \otimes \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix}\right) = \left(\mathbf{1}_{m_1}^T \otimes \phi^T\right) \left(\mathbf{1}_{m_1} \otimes \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix}\right) = m_1 \begin{bmatrix} 1 & \mathbf{0}_{M-1}^T \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix} \tag{94}$$

$$= m_1 \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{>1} \end{bmatrix} = m_1 \eta_1 \mu_1 - m_1 \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \boldsymbol{\mu}_{>1}, \tag{95}$$

and the data term becomes

$$\boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{y} = \begin{bmatrix} 1 & \mathbf{0}_{M-1} & 1 & \mathbf{0}_{M-1} & \dots \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & 0 & \dots \\ 0 & \boldsymbol{\Sigma}^{-1} & \ddots \\ \vdots & 0 & \ddots \end{bmatrix} \begin{bmatrix} y_1^{(1)} \\ y_{>1}^{(1)} \\ \vdots \\ y_1^{(m_1)} \\ y_{>1}^{(m_1)} \end{bmatrix} \tag{96}$$

$$= \begin{bmatrix} \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} & \eta_1 & -\eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} & \dots \end{bmatrix} \begin{bmatrix} y_1^{(1)} \\ y_{>1}^{(1)} \\ \vdots \\ y_1^{(m_1)} \\ y_{>1}^{(m_1)} \end{bmatrix}. \tag{97}$$

The data term simplifies to

$$\boldsymbol{\phi}^T \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{y} = \eta_1 \sum_{i=1}^{m_1} y_1^{(i)} - \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} \sum_{i=1}^{m_1} y_{>1}^{(i)} = \eta_1 m_1 \widehat{y}_1^{MC} - \eta_1 \boldsymbol{c}^T \boldsymbol{\Sigma}_{>1}^{-1} m_1 \widehat{y}_{>1}^{MC}. \tag{98}$$

Combining these terms into the posterior yields our stated result

$$\begin{aligned}
\bar{\mu}_1 &= \mu_1 - \frac{m_1\eta_1}{c_\mu^{-2} + m_1\eta_1}\mu_1 + \frac{m_1\eta_1}{c_\mu^{-2} + m_1\eta_1}\boldsymbol{c}^T\boldsymbol{\Sigma}_{>1}^{-1}\boldsymbol{\mu}_{>1} + \frac{m_1\eta_1}{c_\mu^{-2} + m_1\eta_1}\left(\widehat{y}_1^{MC} - \boldsymbol{c}^T\boldsymbol{\Sigma}_{>1}^{-1}\widehat{y}_{>1}^{MC}\right) \\
&= (1-\lambda)\,\mu_1 + \lambda\widehat{y}_1^{MC} - \frac{\eta_1 m_1}{c_\mu^{-2} + m_1\eta_1}\boldsymbol{c}^T\boldsymbol{\Sigma}_{>1}^{-1}\left(\widehat{y}_{>1}^{MC} - \boldsymbol{\mu}_{>1}\right).
\end{aligned}$$

□