Multi-Fidelity Covariance Estimation in the Log-Euclidean Geometry

Aimee Maurais [1] **Terrence Alsup** [2] **Benjamin Peherstorfer** [2] **Youssef Marzouk** [1]

Abstract

We introduce a multi-fidelity estimator of covariance matrices that employs the log-Euclidean geometry of the symmetric positive-definite manifold. The estimator fuses samples from a hierarchy of data sources of differing fidelities and costs for variance reduction while guaranteeing definiteness, in contrast with previous approaches. The new estimator makes covariance estimation tractable in applications where simulation or data collection is expensive; to that end, we develop an optimal sample allocation scheme that minimizes the mean-squared error of the estimator given a fixed budget. Guaranteed definiteness is crucial to metric learning, data assimilation, and other downstream tasks. Evaluations of our approach using data from physical applications (heat conduction, fluid dynamics) demonstrate more accurate metric learning and speedups of more than one order of magnitude compared to benchmarks.

1. Introduction

Covariance estimation is a fundamental task for understanding the relationships between multiple features of data points. It arises in a wide range of machine learning applications such as metric learning (Zadeh et al., 2016; Huang et al., 2015), graphical models (Loh & Bühlmann, 2014; Ravikumar et al., 2011), and classification (Hastie et al., 2001), as well as in spatial statistics (Cressie, 2015) and in science and engineering applications such as filtering and data assimilation (Schillings & Stuart, 2017; Bergemann & Reich, 2010; Law et al., 2015). The canonical Monte Carlo approach to estimating the covariance matrix $\mathbf{\Sigma} = \mathbb{C}\mathrm{ov}[\mathbf{y}]$ of a random variable $\mathbf{y}$ with d components is to take n samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and compute the sample covariance matrix

$$\hat{\mathbf{\Sigma}}_n = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})^\top , \qquad (1)$$

where[1] $\bar{\mathbf{y}} = \mathbb{E}[\mathbf{y}]$. The mean-squared error (MSE) of the estimator $\hat{\mathbf{\Sigma}}_n$ in the Euclidean metric, i.e., $\mathbb{E}[d_{\mathrm{F}}(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma})^2]$ where $d_{\mathrm{F}}(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, decays as $\mathcal{O}(1/n)$ if the samples are independent and identically distributed (i.i.d.). Furthermore, if $\mathbf{\Sigma}$ lies in the space $\mathbb{S}^d_{++}$ of symmetric (strictly) positive definite (SPD) $d \times d$ matrices, then the sample covariance matrix $\hat{\mathbf{\Sigma}}_n$ is in $\mathbb{S}^d_{++}$ almost surely for $n > d$.

We are interested in situations where obtaining a realization of $\mathbf{y}$ incurs a high cost $c_0 > 0$ and thus straightforward application of the Monte Carlo estimator (1) becomes computationally intractable. High sampling cost is a pervasive issue in science and engineering, where sampling typically corresponds to measuring the response of processes and systems that must be either numerically simulated or physically observed. We note that there exist a wide range of techniques for covariance estimation in high dimensions (Bickel & Levina, 2008; Ledoit & Wolf, 2012; Loh & Wainwright, 2012; Fan et al., 2011; Tsiligkaridis & Hero, 2013; Loh & Bühlmann, 2014; Donoho et al., 2018). Our focus here instead is on situations where *drawing each sample incurs high cost*; see also the discussion in the Conclusions.

Multi-fidelity estimation The opportunity that we exploit is that in many applications there is a *multi-fidelity hierarchy* of data sources from which correlated auxiliary samples can be obtained; see Peherstorfer et al. (2018) for a survey. Multi-fidelity approaches have found increasing utility in machine learning—e.g., for optimization (Poloczek et al., 2017; Wu et al., 2020; Li et al., 2020), classification and sequential learning (Marques et al., 2018; Gundersen et al., 2021; Palizhati et al., 2022), and sampling (Cai & Adams, 2022; Prescott & Baker, 2020; Alsup et al., 2022).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple. In the multi-fidelity setting, in addition to the original "high-fidelity" random variable $\mathbf{y} \equiv \mathbf{y}^{(0)} : \Omega \to \mathbb{R}^d$ generating the sample set $\{\mathbf{y}^{(0)}(\omega_i)\}_{i=1}^{n_0}$, one has access to a number of low-fidelity surrogate random variables $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(\ell)} : \Omega \to \mathbb{R}^d$ and obtains corresponding sets of samples $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_\ell}$,

[1]Massachusetts Institute of Technology, Cambridge, MA USA [2]Courant Institute of Mathematical Sciences, New York University, New York, NY USA. Correspondence to: Aimee Maurais <maurais@mit.edu>, Terrence Alsup <alsup@cims.nyu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]If the mean of $\mathbf{y}$ is unknown, one can replace $\bar{\mathbf{y}}$ with the sample mean $\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$ and normalize (1) with $1/(n-1)$.

$\ell = 1, \ldots, L$. For a particular $\omega \in \Omega$, the realizations $\mathbf{y}^{(0)}(\omega)$, $\mathbf{y}^{(1)}(\omega), \ldots, \mathbf{y}^{(L)}(\omega)$ are correlated, but the realizations within each set $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_\ell}$ are i.i.d., for $\ell = 0, \ldots, L$. We assume that the surrogates $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}$ are sorted in order of decreasing fidelity to $\mathbf{y}^{(0)}$ as measured by correlation coefficients $1 \geq |\rho_1| \geq |\rho_2| \geq \cdots \geq |\rho_L|$. Here ρ_ℓ is a notion of multivariate correlation between $\mathbf{y}^{(0)}(\omega)$ and $\mathbf{y}^{(\ell)}(\omega)$, $\ell = 1, \ldots, L$, that will be formally defined in Section 3.3. Corresponding to decreasing fidelity, the surrogate random variables have decreasing sampling costs, $c_0 \geq c_1 \geq \cdots \geq c_L$.

Loss of definiteness in Euclidean multi-fidelity covariance estimation Direct application of multi-level and multi-fidelity Monte Carlo estimation (Giles, 2008; Cliffe et al., 2011; Teckentrup et al., 2013; Ng & Willcox, 2014) based on control variates to (co)variance estimation has been proposed in, e.g., Bierig & Chernov (2015); Qian et al. (2018); Mycek & De Lozzo (2019). These multi-fidelity estimators, however, rely on differences of single-fidelity Monte Carlo estimators, which can lead to a loss of definiteness of the estimated covariance matrix, as we will detail in Section 2. Ignoring this loss-of-definiteness can result in errors in downstream tasks (e.g., instabilities in Kalman filtering, distances becoming negative in a learned metric, inadmissible graph structures). Accordingly, post-processing strategies (Hoel et al., 2016; Chernov et al., 2021) have been proposed that eliminate negative eigenvalues, but elimination of eigenvalues introduces potential for large errors, requires hand-tuning of thresholds, and often, as the authors note, incurs high computational cost; see also Qi & Sun (2006); Popov et al. (2021).

Our contribution: Guaranteed positive multi-fidelity log-Euclidean covariance estimation We introduce a multi-fidelity covariance estimator based on control variates in the log-Euclidean geometry for $\mathbb{S}^d_{++}$ (Arsigny et al., 2006). The log-Euclidean geometry equips $\mathbb{S}^d_{++}$ with vector-space structure by placing it in one-to-one correspondence with $\mathbb{S}^d$, the vector space of $d \times d$ symmetric matrices, and defining notions of logarithmic addition and scalar multiplication for elements of $\mathbb{S}^d_{++}$. In particular, we make use of this vector space structure to safely "subtract" sample covariance matrices as part of a control variate construction in the tangent space to $\mathbb{S}^d_{++}$, which we identify with $\mathbb{S}^d$. The log-Euclidean vector space can be equipped with the log-Euclidean metric for SPD matrices, which itself induces Riemannian structure on $\mathbb{S}^d_{++}$. An additional advantage of the log-Euclidean geometry used here, as compared to many other geometries for $\mathbb{S}^d_{++}$ (see Section 3.1), is that computations remain efficient as the most complex operations required are matrix exponentials and logarithms.

Summary of main contributions The main contributions and key features of this work are summarized as follows:

(a) We introduce a multi-fidelity estimator that preserves definiteness of covariance matrices in the finite-sample regime and that has first-order minimal mean squared error (MSE) in the log-Euclidean metric.

(b) We provide analysis that leads to a first-order optimal sample allocation between the high-fidelity and surrogate models for gaining orders of magnitude speedups compared to using high-fidelity samples alone.

(c) We demonstrate the algorithmic benefits of the log-Euclidean geometry by showing that our estimator can be implemented with few lines of code (Appendix F), relying on standard numerical linear algebra routines. This low barrier to implementation allows the estimator to be readily plugged into existing code with minimal effort and even as a post-processing step.

2. Multi-Fidelity Covariance Estimation in Euclidean Geometry and Loss of Definiteness

Recall from the introduction that we generate a set of high-fidelity samples $\{\mathbf{y}^{(0)}(\omega_i)\}_{i=1}^{n_0}$ and sets of surrogate samples $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_\ell}$ over different fidelity levels $\ell = 1, \ldots, L$. The classical approach to multi-fidelity estimation, i.e., in the Euclidean geometry, is to use the surrogate samples to define control variates that reduce the variance of the standard Monte Carlo estimator (1) (Bierig & Chernov, 2015; Qian et al., 2018; Mycek & De Lozzo, 2019). This approach yields the Euclidean multi-fidelity (EMF) covariance estimator

$$\hat{\mathbf{\Sigma}}^{\mathrm{EMF}}_{\boldsymbol{n}} = \hat{\mathbf{\Sigma}}^{(0)}_{n_0} + \sum\nolimits_{\ell=1}^{L} \alpha_\ell \left(\hat{\mathbf{\Sigma}}^{(\ell)}_{n_\ell} - \hat{\mathbf{\Sigma}}^{(\ell)}_{n_{\ell-1}} \right), \quad (2)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^\top \in \mathbb{R}^L$ are the control variate weights and

$$\begin{aligned} \hat{\mathbf{\Sigma}}^{(\ell)}_{n_\ell} &= \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left(\mathbf{y}^{(\ell)}(\omega_i) - \bar{\mathbf{y}}^{(\ell)}_{n_\ell} \right) \left(\mathbf{y}^{(\ell)}(\omega_i) - \bar{\mathbf{y}}^{(\ell)}_{n_\ell} \right)^\top, \\ \hat{\mathbf{\Sigma}}^{(\ell)}_{n_{\ell-1}} &= \frac{1}{n_{\ell-1}} \sum_{i=1}^{n_{\ell-1}} \left(\mathbf{y}^{(\ell)}(\omega_i) - \bar{\mathbf{y}}^{(\ell)}_{n_\ell} \right) \left(\mathbf{y}^{(\ell)}(\omega_i) - \bar{\mathbf{y}}^{(\ell)}_{n_\ell} \right)^\top \end{aligned} \quad (3)$$

are sample covariance matrices computed from $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_\ell}$ and $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_{\ell-1}}$, $\ell = 1, \ldots, L$. We require for well-definedness that $n_\ell \geq n_{\ell-1}$, $\ell = 1, \ldots, L$. For levels $\ell, t \in \{0, \ldots, L\}$ and integers $m_i, m_j \in \mathbb{Z}^+$, notice that the sample covariance matrices

$$\hat{\mathbf{\Sigma}}^{(\ell)}_{m_i} = \frac{1}{m_i} \sum_{i'=1}^{m_i} \left(\mathbf{y}^{(\ell)}(\omega_{i'}) - \bar{\mathbf{y}}^{(\ell)}_{m_i} \right) \left(\mathbf{y}^{(\ell)}(\omega_{i'}) - \bar{\mathbf{y}}^{(\ell)}_{m_i} \right)^\top,$$

$$\hat{\mathbf{\Sigma}}^{(t)}_{m_j} = \frac{1}{m_j} \sum_{j'=1}^{m_j} \left(\mathbf{y}^{(t)}(\omega_{j'}) - \bar{\mathbf{y}}^{(t)}_{m_j} \right) \left(\mathbf{y}^{(t)}(\omega_{j'}) - \bar{\mathbf{y}}^{(t)}_{m_j} \right)^\top$$

are correlated because they are constructed from evaluations of $\mathbf{y}^{(\ell)}$ and $\mathbf{y}^{(t)}$ on the common set of events $\{\omega_i\}_{i=1}^{\min\{m_i,m_j\}}$. This correlation is the key ingredient to achieving variance reduction in (2) over equivalent-cost single-fidelity estimators; see Section 3.3.

Loss of definiteness The use of regular subtraction and scalar multiplication in Equation (2) treats the symmetric positive definite matrices $\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}$ and $\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}$, $\ell = 0, \ldots, L$, as elements of the Euclidean vector space $\mathbb{S}^d$. However, SPD matrices do not constitute a vector space under the Euclidean geometry and thus the Euclidean multi-fidelity estimator (2) may become indefinite due to the presence of subtraction. That is, positive definiteness of (2) in the finite-sample regime cannot be guaranteed even if its constituent single-fidelity sample covariance matrices are definite.

3. Log-Euclidean Multi-Fidelity (LEMF) Covariance Estimator

3.1. Log-Euclidean Geometry and its Benefits

We begin by orienting ourselves within the log-Euclidean geometry for $\mathbb{S}_{++}^d$ as defined in Arsigny et al. (2006). For $\mathbf{A}, \mathbf{B} \in \mathbb{S}_{++}^d$ and $\lambda \in \mathbb{R}$ we have notions of *logarithmic* addition $\oplus$ and scalar multiplication $\odot$,

$$\begin{aligned} \mathbf{A} \oplus \mathbf{B} &= \operatorname{Exp}(\operatorname{Log} \mathbf{A} + \operatorname{Log} \mathbf{B}) \,, \\ \lambda \odot \mathbf{A} &= \operatorname{Exp}(\lambda \cdot \operatorname{Log} \mathbf{A}) = \mathbf{A}^\lambda, \end{aligned} \tag{4}$$

corresponding to regular addition and scalar multiplication of $\log \mathbf{A}$ and $\log \mathbf{B}$ in $\mathbb{S}^d$. Here $\operatorname{Exp} : \mathbb{S}^d \to \mathbb{S}_{++}^d$ is the matrix-exponential and $\operatorname{Log} : \mathbb{S}_{++}^d \to \mathbb{S}^d$ is its inverse (Arsigny et al., 2007, Def. 2.1). The mappings Log and Exp place $\mathbb{S}^d$ and $\mathbb{S}_{++}^d$ in one-to-one correspondence, enabling $\mathbb{S}_{++}^d$ to make use of the vector space structure of $\mathbb{S}^d$. The definitions of $\oplus$ and $\odot$ (4) satisfy the axioms of a vector space, and we can equip this vector space $\mathbb{S}_{++}^d(\oplus, \odot)$ with the log-Euclidean metric,

$$d_{\mathrm{LE}}(\mathbf{A}, \mathbf{B}) = \|\operatorname{Log} \mathbf{A} - \operatorname{Log} \mathbf{B}\|_{\mathrm{F}} \,. \tag{5}$$

Computations related to the log-Euclidean geometry can be performed economically because efficient algorithms exist for computing matrix exponentials and logarithms. Efficient computation is a major advantage as compared to the geometries induced by other metrics such as the affine invariant metric (Bhatia, 2007; Huang et al., 2015).

Review of other geometries for $\mathbb{S}_{++}^d$ The set of $d \times d$ symmetric positive definite matrices $\mathbb{S}_{++}^d$ forms a Riemannian manifold embedded in the vector space of $d \times d$ symmetric matrices $\mathbb{S}^d$. One can equip $\mathbb{S}_{++}^d$ with a number of geometries in addition to the Euclidean (which we have shown to be problematic for multi-fidelity estimation) and log-Euclidean geometries discussed here. There is, e.g., the affine-invariant geometry (Bhatia, 2007), often billed as the "canonical choice," which gives rise to the affine-invariant metric,

$$d_{\mathrm{Aff}}(\mathbf{A},\ \mathbf{B}) = \left\|\operatorname{Log}\left(\mathbf{A}^{-1}\mathbf{B}\right)\right\|_{\mathrm{F}} \,. \tag{6}$$

Other choices include the Bures-Wasserstein geometry, arising from the L^2-Wasserstein distance between multivariate, nondegenerate, mean-zero Gaussians (Malagò et al., 2018); and the Log-Cholesky geometry, obtained by parametrizing SPD matrices in terms of their Cholesky factors and defining a Riemannian metric on the space of lower-triangular matrices with positive diagonal entries (Lin, 2019). Choice of geometry for $\mathbb{S}_{++}^d$ is application-dependent and often depends on factors such as cost to compute geodesic distance, availability of the Riemannian exponential and logarithmic maps in closed form, and whether the so-called "swelling effect" (Arsigny et al., 2006) is an issue. We choose the log-Euclidean geometry for its computational advantages, described at the end of the previous paragraph.

3.2. Log-Euclidean Multi-Fidelity Estimation

We construct the log-Euclidean multi-fidelity (LEMF) covariance estimator via linear control variates in the log-Euclidean geometry for $\mathbb{S}_{++}^d$,

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} &= \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L} \alpha_\ell \odot \left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right) = \\ &\operatorname{Exp}\left(\operatorname{Log} \hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L} \alpha_\ell \left(\operatorname{Log} \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log} \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right), \end{aligned} \tag{7}$$

where we have used the convention $\mathbf{A} \ominus \mathbf{B} = \mathbf{A} \oplus -1 \odot \mathbf{B}$. The estimator (7) has the same algebraic structure as (2), but is guaranteeably positive definite at finite $\boldsymbol{n} = (n_0, n_1, \ldots, n_L)$ as the following proposition shows:

Proposition 3.1. *The log-Euclidean multi-fidelity covariance estimator* (7) *exists and is positive definite almost surely whenever* $n_0, n_1, \ldots, n_L > d$.

In addition to interpreting (7) as a linear control variate construction in the log-Euclidean geometry for $\mathbb{S}_{++}^d$, we can equivalently view the LEMF estimator as a Euclidean linear control variate estimate of $\operatorname{Log} \mathbf{\Sigma}$ on $\mathbb{S}^d$,

$$\operatorname{Log} \hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} = \operatorname{Log} \hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L} \alpha_\ell \left(\operatorname{Log} \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log} \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right), \tag{8}$$

which we then map to $\mathbb{S}_{++}^d$ via the matrix exponential. Variance reduction in (8) is achieved on $\mathbb{S}^d$ via the standard Euclidean control variate framework and then propagated

to $\mathbb{S}^d_{++}$ via $\mathrm{Exp}(\cdot)$. Furthermore, (7) has the form of a log-Euclidean Fréchet average of $\hat{\mathbf{\Sigma}}^{(0)}_{n_0}$ and L small perturbations induced by the surrogate sample covariance matrices; see Appendix A for details.

The LEMF estimator requires computing $2L+1$ matrix logarithms and one matrix exponential. Note that obtaining the sample sets $\{\mathbf{y}^{(\ell)}(\omega_i)\}_{i=1}^{n_\ell}$, $\ell = 0, \ldots, L$, is typically much more computationally expensive than computing matrix logarithms and exponentials.

3.3. The MSE of the LEMF Estimator

For the following analysis, we assume without loss of generality that $\mathbf{y}^{(\ell)}, \ell = 0, \ldots, L$, have zero mean, for ease of exposition. As introduced in Section 1, the surrogates $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}$ are ordered by decreasing fidelity,

$$1 = \rho_0 > |\rho_1| > \cdots > |\rho_L| \geq \rho_{L+1} = 0,$$

where ρ_ℓ is a multivariate generalization of the Pearson correlation coefficient,

$$\rho_\ell = \frac{\mathrm{Tr}\left(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top,\ (\mathbf{y}^{(\ell)})(\mathbf{y}^{(\ell)})^\top]\right)}{\sigma_0 \sigma_\ell}, \quad (9)$$

with generalized variances defined as

$$\sigma_\ell^2 = \mathrm{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top]), \quad \ell = 0, \ldots, L. \quad (10)$$

The covariance of $\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top, \ell = 0, \ldots, L$ is a symmetric positive semidefinite linear operator from $\mathbb{S}^d$ to $\mathbb{S}^d$,

$$\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top] = \mathbb{E}\left[\left(\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top - \mathbf{\Sigma}^{(\ell)}\right) \otimes \left(\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top - \mathbf{\Sigma}^{(\ell)}\right)\right], \quad (11)$$

where we define $\mathbf{\Sigma}^{(\ell)} = \mathbb{E}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top]$. Similarly, the cross-covariance between $\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top$ and $\mathbf{y}^{(m)}(\mathbf{y}^{(m)})^\top$, $\ell, m \in \{0, \ldots, L\}$ is

$$\mathbb{C}\mathrm{ov}\left[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top,\ \mathbf{y}^{(m)}(\mathbf{y}^{(m)})^\top\right] = \mathbb{E}\left[\left(\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top - \mathbf{\Sigma}^{(\ell)}\right) \otimes \left(\mathbf{y}^{(m)}(\mathbf{y}^{(m)})^\top - \mathbf{\Sigma}^{(m)}\right)\right]. \quad (12)$$

In writing (11) and (12) we do not invoke any particular representation for linear operators on $\mathbb{S}^d$. Rather, we only assume that $\otimes$ is an outer product for symmetric matrices compatible with the Frobenius inner product, that is,

$$\mathrm{Tr}(\mathbf{A} \otimes \mathbf{B}) = \langle \mathbf{A}, \mathbf{B} \rangle_F, \quad \mathbf{A}, \mathbf{B} \in \mathbb{S}^d.$$

The following proposition derives the MSE of the log-Euclidean estimator in the log-Euclidean metric (5).

Proposition 3.2. *Assume that $\|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}} < h/4$ with $h < 1$ and $\|\mathbf{\Sigma} - \mathbf{\Sigma}^{(\ell)}\|_{\mathrm{F}} < h/4$ for $\ell = 1, \ldots, L$. Then, for coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ and sufficiently large numbers of samples $\boldsymbol{n} \in \mathbb{N}^{L+1}$, the MSE in the log-Euclidean metric of the LEMF estimator defined in* (7) *is*

$$\mathbb{E}\left[d_{\mathrm{LE}}\left(\hat{\mathbf{\Sigma}}^{\mathrm{LEMF}}_{\boldsymbol{n}},\ \mathbf{\Sigma}\right)^2\right] = \mathbb{E}\left[\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}^{\mathrm{LEMF}}_{\boldsymbol{n}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right] = \frac{\sigma_0^2}{n_0} + \sum_{\ell=1}^{L}\left(\frac{1}{n_{\ell-1}} - \frac{1}{n_\ell}\right)(\alpha_\ell^2\sigma_\ell^2 - 2\alpha_\ell\rho_\ell\sigma_\ell\sigma_0) + \mathcal{O}(h^2). \quad (13)$$

Note that the technical condition $\|\mathbf{\Sigma} - \mathbf{I}\|_F < h/4$ can always be established by re-scaling $\tilde{\mathbf{y}}^{(\ell)} = \mathbf{C}^{-1/2}\mathbf{y}^{(\ell)}$ with a matrix $\mathbf{C} \in \mathbb{S}^d_{++}$ and using $\mathbf{C}$ as the base of the matrix logarithm and exponential in the estimator (7); we consider $\mathbf{I}$ as the base here for ease of exposition.

Below we relate the MSE of the Euclidean estimator $\hat{\mathbf{\Sigma}}^{\mathrm{EMF}}_{\boldsymbol{n}}$ (2) in the Euclidean metric with the MSE of the log-Euclidean estimator (7) $\hat{\mathbf{\Sigma}}^{\mathrm{LEMF}}_{\boldsymbol{n}}$ in the log-Euclidean metric:

Corollary 3.3. *The MSE in the Euclidean metric of the Euclidean multi-fidelity estimator* (2) *with sample size vector $\boldsymbol{n}$ and coefficients $\boldsymbol{\alpha}$ is*

$$\mathbb{E}\left[d_{\mathrm{F}}\left(\hat{\mathbf{\Sigma}}^{\mathrm{EMF}}_{\boldsymbol{n}}, \mathbf{\Sigma}\right)^2\right] = \mathbb{E}\left[\left\|\hat{\mathbf{\Sigma}}^{\mathrm{EMF}}_{\boldsymbol{n}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right] = \frac{\sigma_0^2}{n_0} + \sum_{\ell=1}^{L}\left(\frac{1}{n_{\ell-1}} - \frac{1}{n_\ell}\right)(\alpha_\ell^2\sigma_\ell^2 - 2\alpha_\ell\rho_\ell\sigma_\ell\sigma_0). \quad (14)$$

Corollary 3.3 shows that the MSE of the LEMF estimator in the log-Euclidean metric is equal to the MSE of the EMF estimator in the Euclidean metric up to first order. Note that it does not comment on a cross-comparison (e.g., MSE of the LEMF in the Euclidean metric and vice-versa); we will explore such comparisons numerically in Section 4.

3.4. Optimal Sample Allocation for the LEMF

Proposition 3.2 and Corollary 3.3 show the dependence of the MSE in the log-Euclidean and Euclidean metrics of the LEMF and EMF estimators on the sample allocations $\boldsymbol{n}$ and coefficients $\boldsymbol{\alpha}$. We now derive $\boldsymbol{n}^*$ and $\boldsymbol{\alpha}^*$ that minimize the respective MSEs up to first order and so achieve an optimal sample allocation across the hierarchy of data sources.

We start with $\boldsymbol{\alpha}^*$: because the first-order approximation of the MSE in (13) and the MSE in (14) are both quadratic in the coefficients $\alpha_1, \ldots, \alpha_L$, we can directly minimize for $\alpha_1, \ldots, \alpha_L$ and obtain

$$\boldsymbol{\alpha}^\star = [\alpha_1^\star, \ldots, \alpha_L^\star], \quad \alpha_\ell^\star = \rho_\ell \frac{\sigma_0}{\sigma_\ell}, \quad (15)$$

which are independent of the sample sizes.

In preparation for finding $\boldsymbol{n}^*$, we define the costs of the multi-fidelity estimators as

$$c(\boldsymbol{n}) = \sum\nolimits_{\ell=0}^{L} n_\ell c_\ell \,. \tag{16}$$

Following Peherstorfer et al. (2016), Proposition 3.4 derives the optimal sample allocation $\boldsymbol{n}^*$ to obtain an estimator with first-order minimal MSE and costs $c(\boldsymbol{n}^*) \leq B$ for a fixed computational budget $B > 0$.

Note that in writing (16), we assume that the costs of generating *samples* of the random variables $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(L)}$ are much higher than the cost of constructing a multi-fidelity covariance estimate, and in particular eclipse the cost of computing the matrix exponential and the $2L+1$ matrix logarithms required by the LEMF estimator (7). This situation is often the case, e.g., in science and engineering applications when generating samples may correspond to evaluating computationally intensive physics-based models. One could explicitly account for the cost of constructing the LEMF estimator by inserting a modified budget of $\hat{B} = B - c_{\log/\exp}$ into Proposition 3.4, where $c_{\log/\exp}$ is the cost of the matrix logarithms and exponential, and subsequently obtain a result analogous to Corollary 3.5 with a more complicated condition for when the LEMF estimator outperforms an equivalent-cost high-fidelity estimator.

Proposition 3.4. *Let $B > 0$ denote the computational budget and let Proposition 3.2 apply. The first-order optimal sample allocation $\boldsymbol{n}^\star$ for the LEMF estimator* (2) *that solves*

$$\begin{aligned} \min_{\boldsymbol{n}\in\mathbb{R}^{L+1}} \quad & \mathbb{E}\left[d_{\mathrm{LE}}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}}, \boldsymbol{\Sigma}\right)^2 \right] \\ \textit{such that} \quad & c(\boldsymbol{n}) \leq B \end{aligned} \tag{17}$$

is

$$n_\ell^\star = B\sqrt{\frac{c_0(\rho_\ell^2 - \rho_{\ell+1}^2)}{c_\ell(1-\rho_1^2)}} \Bigg/ \sum_{i=0}^{L} c_i \sqrt{\frac{c_0(\rho_i^2 - \rho_{i+1}^2)}{c_i(1-\rho_1^2)}} \,. \tag{18}$$

Because the MSE of the Euclidean estimator (in $d_{\mathrm{F}}(\cdot,\cdot)$) is equal to the first-order approximation of the MSE of the log-Euclidean estimator (in $d_{\mathrm{LE}}(\cdot,\cdot)$), we directly obtain that (18) is the optimal sample allocation for the Euclidean estimator as well. In practice, we round either up or down and so use either $\lceil \boldsymbol{n}^\star \rceil$ or $\lfloor \boldsymbol{n}^\star \rfloor$.

If the correlation coefficients ρ_l and costs $c_\ell, \ell = 0, \ldots, L$ are known, one can determine the optimal $\boldsymbol{\alpha}^\star$ and $\boldsymbol{n}^\star$ for fixed budget B; or, alternatively, determine $\boldsymbol{\alpha}^\star$ and $\boldsymbol{n}^\star$ so that the MSE of the LEMF estimator is below a threshold $\epsilon > 0$. In situations when ρ_ℓ and c_ℓ are unknown, it is common to estimate them in pilot studies (Cliffe et al., 2011; Peherstorfer et al., 2016) and even to reuse the pilot samples in the actual estimator (Konrad et al., 2022).

3.5. Discussion of the LEMF Estimator

Interpretation Proposition 3.4 shows that when the correlation coefficients (9) are close to 1, then the LEMF estimator (7) allocates more of the budget to the cheaper surrogate samples. In the extreme case where $\rho_\ell = 0$, the entire budget is allocated to obtaining high-fidelity samples and one recovers the single-fidelity estimator (1) with $n = B/c_0$.

The following corollary shows that the corrrelation coefficients ρ_ℓ and costs c_ℓ dictate whether the proposed LEMF estimator leads to lower MSE than the equivalent-cost single (high)-fidelity estimator.

Corollary 3.5. *Let Proposition 3.4 apply. Then, the first-order term of the MSE of the LEMF estimator with $\boldsymbol{n}^*$ is*

$$\mathbb{E}\left[d_{\mathrm{LE}}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}^\star}^{\mathrm{LEMF}}, \boldsymbol{\Sigma}\right)^2 \right] \doteq \frac{\sigma_0^2}{B}\left(\sum_{\ell=0}^{L} \sqrt{c_\ell(\rho_\ell^2 - \rho_{\ell+1}^2)} \right)^2 , \tag{19}$$

which will be smaller than the first-order log-Euclidean MSE of the cost-equivalent single-fidelity estimator (1) *that uses $n_0 = B/c_0$ high-fidelity samples if and only if*

$$\sum\nolimits_{\ell=0}^{L} \sqrt{\frac{c_\ell}{c_0}(\rho_\ell^2 - \rho_{\ell+1}^2)} < 1 \,. \tag{20}$$

Equation (20) demonstrates that for fixed c_0, the benefit of using the LEMF estimator (7) over the equivalent-cost high-fidelity estimator (1) is a function of both the surrogate model correlations with the high-fidelity model, $\rho_1, \ldots, \rho_L$, and the surrogate model costs, $c_1, \ldots, c_L$. It is particularly instructive to consider the bifidelity case with only one surrogate model ($L = 1$). In this setting, condition (20) simplifies to

$$\sqrt{1-\rho_1^2} + \sqrt{\frac{c_1}{c_0}\rho_1^2} < 1 \iff 2\sqrt{\frac{1-\rho_1^2}{\rho_1^2}} < \frac{c_0 - c_1}{\sqrt{c_0 c_1}} . \tag{21}$$

If the normalized reduction in cost $\frac{c_0 - c_1}{\sqrt{c_0 c_1}}$ achieved by the low-fidelity model is large, then the minimum required generalized correlation ρ_1 such that condition (21) is satisfied decreases. Conversely, in the limit $\rho_1 \to 1$ we only require $c_1 < c_0$ in order to see a benefit from using the LEMF estimator. In Figure 1 we visually demonstrate this relationship by plotting the region in $(\rho, c_1/c_0)$ space for which (21) holds and the LEMF estimator has a lower first-order MSE than the equivalent-cost high-fidelity estimator.

Truncated multi-fidelity estimator In addition to the EMF estimator and equivalent-cost single-fidelity estimators, we

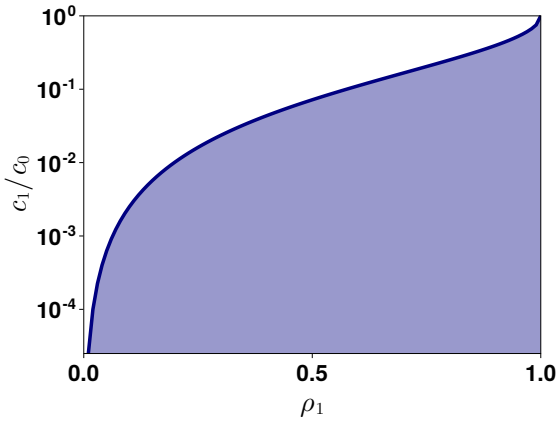


Figure 1. The shaded region corresponds to pairs $(\rho_1, c_1/c_0)$ of correlation coefficient and cost ratio for which (21) holds and thus the LEMF estimator has lower first-order error than the equivalent-cost single-fidelity estimator. Note that a lower correlation ρ_1 requires a smaller cost ratio c_1/c_0 for the LEMF estimator to achieve lower error than the single-fidelity estimator.

will compare the proposed LEMF estimator to the EMF estimator with positive definiteness enforced via eigenvalue truncation. Following Hoel et al. (2016), we define the truncated multi-fidelity covariance estimator as

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{+\delta} = \mathcal{T}_\delta\left(\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}\right), \tag{22}$$

where $\delta > 0$ is a small threshold and the truncation operator $\mathcal{T}_\delta$ acts on symmetric matrices via $\mathcal{T}_\delta(\mathbf{A}) = \mathbf{Q}\left(\mathbf{\Lambda} \vee \delta\right)\mathbf{Q}^\top$ with $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ denoting the eigen-decomposition of $\mathbf{A}$ and $\mathbf{\Lambda} \vee \delta$ denoting the diagonal matrix whose diagonal entries are $\max(\lambda_{ii}, \delta)$; see the Introduction for a discussion about drawbacks of the truncated estimator.

4. Numerical Examples

4.1. Motivating Example with Gaussians

Consider the problem of estimating the covariance of a Gaussian random variable $\mathbf{y} \sim N(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} \in \mathbb{S}^4_{++}$. For this motivating example, we obtain low-fidelity samples $\mathbf{y}^{(\ell)} = \mathbf{y} + \boldsymbol{\epsilon}^{(\ell)}$ by corrupting a high-fidelity sample $\mathbf{y}$ with independent noise at different noise levels $\boldsymbol{\epsilon}^{(\ell)} \sim N(\mathbf{0}, \mathbf{\Gamma}^{(\ell)})$ for $\ell = 1, 2, 3$. The matrices $\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(3)}$ are chosen such that noise levels lead to generalized correlations of $\rho_1 \approx 0.93$, $\rho_2 \approx 0.74$, and $\rho_3 \approx 0.58$; see Appendix G. We impose unit costs for drawing a high-fidelity sample $\mathbf{y}$ and costs $c_1 = 10^{-2}, c_2 = 10^{-3}$, and $c_3 = 10^{-4}$ for drawing the surrogate samples. Because the samples are Gaussian, their empirical second moments follow a Wishart distribution and the generalized variances (10) and correlations (9) can be computed exactly; see Appendix G. Table 1 shows the MSE of the multi-fidelity estimators compared with that of single-fidelity covariance estimators using high-fidelity or surrogate samples alone with the same budget $B = 15$. The LEMF estimator has $2\times$ lower MSE than the other estimators in the log-Euclidean and affine-invariant metric, while being competitive in the Euclidean metric.

4.2. Uncertainty Quantification of Steady-State Heat Flow

Physics model The steady-state heat equation over the domain $\mathcal{X} = (0, 1)$ with a variable heat conductivity is

$$-\operatorname{div}\left(\exp(\kappa(x;\boldsymbol{\theta}))\nabla u(x;\boldsymbol{\theta})\right) = f(x), \quad x \in \mathcal{X}, \tag{23}$$

where we impose Dirichlet boundary conditions $u(0, \boldsymbol{\theta}) = 0, u(1, \boldsymbol{\theta}) = 1$ and constant source $f(x) = 1$. We model the log heat conductivity κ as a degree-4 sine-polynomial $\kappa(x;\theta) = \sum_{k=1}^4 \theta_k \sin(2\pi k x)$, whose coefficients are given by the components of $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_4]^\top \in \mathbb{R}^4$.

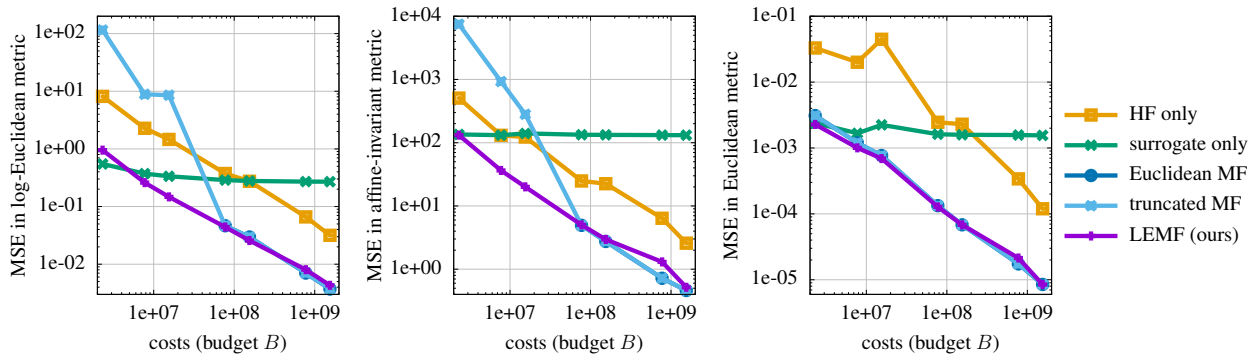
We obtain a high-fidelity observation model $G^{(0)} : \mathbb{R}^4 \to \mathbb{R}^{10}$ by approximating the solution $u(\cdot, \boldsymbol{\theta})$ via a second-order finite difference scheme with 65,536 grid points and observing the resulting temperature at $d = 10$ equally spaced points $x_i = i/11$, $i \in \{1, \ldots, 10\}$, in the interior of the domain $\mathcal{X}$ (see Figure 8 in Appendix H). The surrogate model $G^{(1)}$ is obtained by approximating the solution $u(\cdot, \boldsymbol{\theta})$ with only 1,024 grid points and performing the same measurement process. We identify the costs of the high-fidelity and surrogate models with the number of grid points involved, $c_0 = 2^{16}$ and $c_1 = 2^{10}$.

Covariance estimation We model $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I}_{4\times 4})$ and define high-fidelity samples as $\mathbf{y}^{(0)} = G^{(0)}(\boldsymbol{\theta})$ and surrogate samples as $\mathbf{y}^{(1)} = G^{(1)}(\boldsymbol{\theta})$. Our aim is to estimate the covariance of $\mathbf{y}^{(0)}$, which is a 10×10 matrix. We first estimate the generalized variances and correlations $\sigma_0, \sigma_1, \rho_1$ in a pilot study to compute the optimal multi-fidelity sample allocations $\boldsymbol{n}^\star = [n_\star^{(0)}, n_\star^{(1)}]$ and $\boldsymbol{\alpha}^\star$ for seven values of budget B in the interval $[10^6, 2 \times 10^9]$; see Appendix H. We then obtain $n_\star^{(1)} \geq n_\star^{(0)}$ realizations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n_\star^{(1)}}$ of the input random variable $\boldsymbol{\theta}$ and compute the $n_\star^{(0)}$ high-fidelity samples $\{\mathbf{y}^{(0)}(\boldsymbol{\theta}_i)\}_{i=1}^{n_\star^{(0)}}$ and $n_\star^{(1)}$ surrogate samples $\{\mathbf{y}^{(1)}(\boldsymbol{\theta}_i)\}_{i=1}^{n_\star^{(1)}}$. Note that the first $n_\star^{(0)}$ surrogate samples use the same realizations of $\boldsymbol{\theta}$ as the high-fidelity samples, creating statistical coupling, and thus correlation, between the high-fidelity and surrogate samples.

Results Figure 2 compares the MSE in the log-Euclidean, affine-invariant, and Euclidean metrics (6) of the single-fidelity and multi-fidelity estimators over 100 trials. Using the surrogate samples alone leads to estimates with large bias with respect to the true (high-fidelity) covariance. For example, the bias of the surrogate-only estimator can be seen in Figure 3c, which displays the convergence of each estimate of $\mathbf{\Sigma}_{2,5}$ to the truth; see also Appendix H.

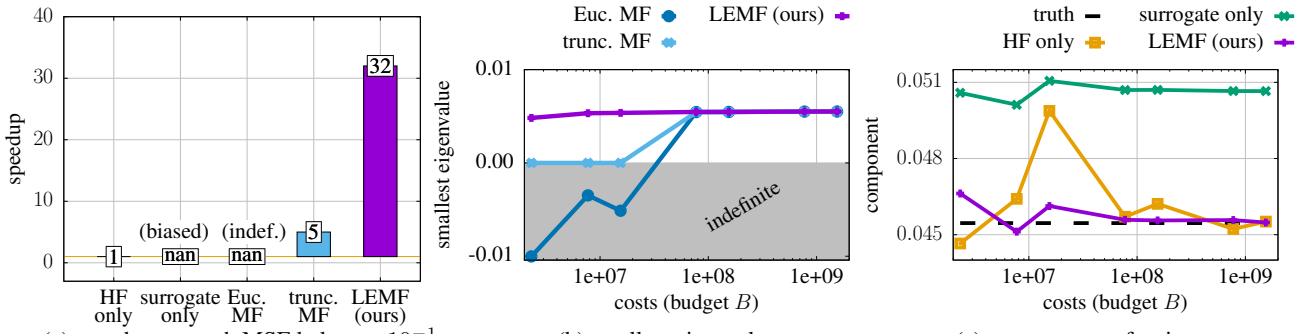
MSE	high-fidelity only	surrogate only	Euclidean MF	truncated MF ($\delta = 10^{-16}$)	LEMF (ours)
log-Euclidean	1.72	5.82	NaN	66.88	0.83
affine-invariant	4.99	5.82	NaN	125.11	2.69
Euclidean	3.00	4.00	0.51	0.51	0.71

Table 1. Motivating example: The LEMF estimator leverages cheap surrogate samples to reduce the MSE compared to cost-equivalent single-fidelity estimators and other multi-fidelity estimators that rely on post-processing via truncation.



(a) log-Euclidean metric (b) affine-invariant metric (c) Euclidean metric

Figure 2. Heat flow: In the log-Euclidean metric, our LEMF estimator achieves the MSE tolerance below $\approx 10^{-1}$ with $\approx 30\times$ speedup compared to the single-fidelity estimator that uses high-fidelity samples alone. Only using surrogate samples leads to a large bias that prevents reaching an MSE below 10^{-1}. The Euclidean multi-fidelity estimator is indefinite in more than 10% of the 100 trials used here and therefore does not provide a valid covariance matrix in this example. (Plot with min/max over 100 trials is shown in Appendix H.)



(a) speedup to reach MSE below $\approx 10^{-1}$ (b) smallest eigenvalue (c) convergence of estimators

Figure 3. Heat flow: Plot (a): Our LEMF estimator achieves a $32\times$ speedup. Plot (b): Our LEMF estimator remains positive definite even for small sample sizes, whereas the Euclidean estimator becomes indefinite. Plot (c): Entries of our LEMF estimates converge quickly to the entries of the true covariance, whereas surrogate-only estimators incur a large bias and high-fidelity-only estimators incur high costs.

The EMF estimator is unbiased and achieves competitively low MSE in the Euclidean metric but becomes indefinite in more than 10% of trials (see Appendix H). For this reason its MSE in the log-Euclidean and affine-invariant metrics is infinite/nonexistent (see Figure 3b), as the log-Euclidean and affine-invariant distances between an element of $\mathbb{S}^d_{++}$ and any non positive-definite matrix are ∞. When we "fix" the EMF estimator via eigenvalue truncation (using a threshold of $\delta = 10^{-16}$ in (22)), obtaining the positive-definite truncated EMF estimator, we introduce the possibility of large MSE in the log-Euclidean and affine-invariant metrics because we have modify the spectrum, which can be seen well for small budgets in Figure 2.

By contrast, the proposed log-Euclidean multi-fidelity estimator uses surrogate samples in conjunction with high-fidelity samples and inherently retains positive definiteness, thus achieving about one order of magnitude speedup across metrics as compared to the estimator that uses the high-fidelity samples alone; Figure 3a. The MSE of the LEMF estimator decays as $\mathcal{O}(1/B)$, in consort with the other Monte Carlo estimators which use high-fidelity samples.

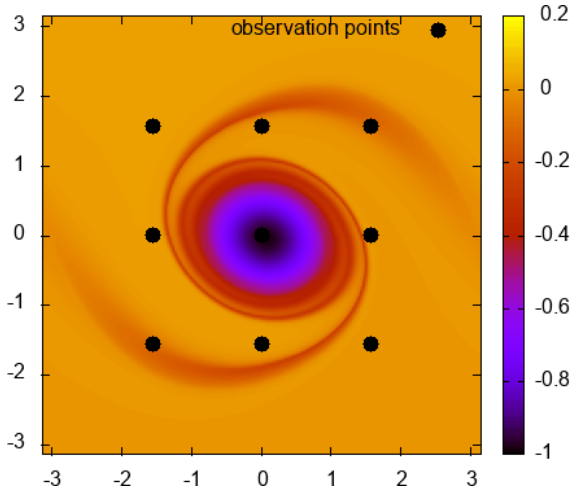


Figure 4. Metric learning: An example of the final-time buoyancy for class $i = 1$. Our measurements consist of buoyancy values at nine spatial locations in the domain, indicated by black dots in the plot above. We use the observations to estimate a metric which distinguishes between observations corresponding to $\boldsymbol{\theta}$ sampled from class $i = 0$ and from class $i = 1$.

4.3. Metric Learning: Surface Quasi-Geostrophy

We now apply our multifidelity covariance estimators in an intermediate step in the metric learning of Zadeh et al. (2016) to enable easier classification of fluid regimes from buoyancy measurements given by a surface quasi-geostrophic model (Held et al., 1985).

Physics model In the model presented in Held et al. (1985), the evolution of the buoyancy $b(\mathbf{x}, t)$ over a periodic spatial domain $\mathcal{X} = [-\pi, \pi] \times [-\pi, \pi]$ is governed by

$$\partial_t b(\mathbf{x}, t; \boldsymbol{\theta}) + J(\psi, b) = 0\,, \tag{24}$$

where ψ is the streamfunction and $\mathbf{x} = (x, y)$. The initial buoyancy is $b_0(\mathbf{x}; \boldsymbol{\theta}) = -\frac{1}{(2\pi/|\theta_5|)^2} \exp\left(-x^2 - \exp(2\theta_1)y^2\right)$, the contours of which form ellipses parametrized by $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_5]^\top \in \mathbb{R}^5$. θ_1 is the log-aspect ratio and controls the origination of vortices in the buoyancy over time, and θ_5 controls the amplitude of b_0. The parameters θ_2, θ_3, and θ_4 determine additional aspects of the dynamics (24); see Appendix I for details. We sample the parameters $\boldsymbol{\theta} \in \mathbb{R}^5$ from a two-component Gaussian mixture whose means differ only in the coordinate corresponding to the log-aspect ratio; see Appendix I. We use $i \in \{0, 1\}$ to denote the mixture component from which a given realization of $\boldsymbol{\theta}$ is sampled.

The high-fidelity parameter-to-observable map $G^{(0)}$ maps $\boldsymbol{\theta}$ onto $d = 9$ observations of the numerical solution of equation (24) in the spatial domain; see Figure 4 as well as Figure 12 in Appendix I. For $G^{(0)}$, numerical solution of (24) is computed up to time $T = 24$ using finite differences with 256 grid points along each coordinate dimension and with a time-step of $\Delta t = 0.005$ [2]. High-fidelity samples $\mathbf{y}^{(0)}$ are thus given by $\mathbf{y}^{(0)} = G^{(0)}(\boldsymbol{\theta})$ and have covariance $\boldsymbol{\Sigma} \in \mathbb{S}^9_{++}$.

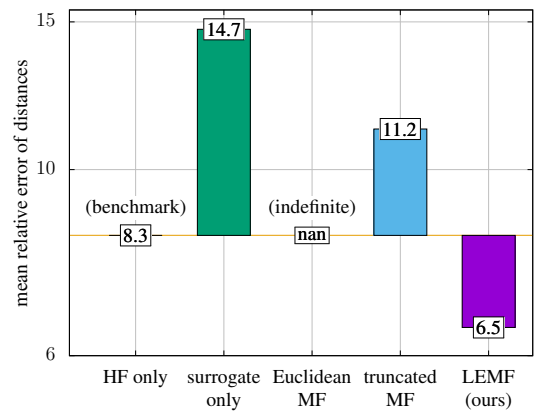


Figure 5. Metric learning: When using the learned metric to compute distances between observations of the surface quasi-geostrophic model, our LEMF estimator achieves a $> 20\%$ lower error in the distances than using high-fidelity samples alone. The Euclidean estimator provides an indefinite metric in this example.

The surrogate parameter-to-observable map $G^{(1)}$ and random variable $\mathbf{y}^{(1)}$ are defined in the same way except that in solving Equation (24) we only use 64 grid points along each coordinate dimension. Evaluating $G^{(1)}$ is 16 times faster than evaluating $G^{(0)}$, with $c_0 = 256^2$ and $c_1 = 64^2$.

Metric learning We now apply our covariance estimator (7) to learn a metric on the data manifold. We use a slight modification of the geometric mean metric learning approach introduced in Zadeh et al. (2016), which learns an SPD matrix $\mathbf{A}$ to define the metric

$$d_{\mathbf{A}}(\mathbf{y}_1, \mathbf{y}_2) = \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)^\top \mathbf{A} (\mathbf{y}_1 - \mathbf{y}_2)}\,.$$

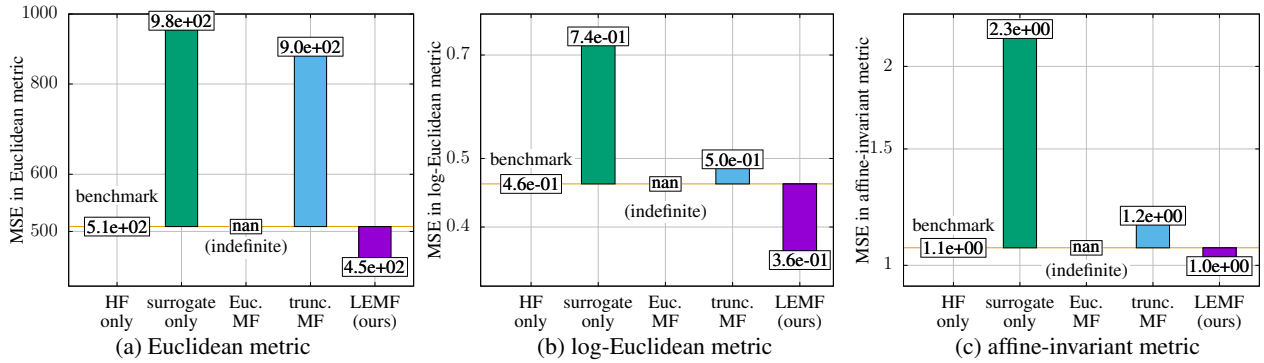
The matrix $\mathbf{A}$ is an interpolation in the affine-invariant geometry for $\mathbb{S}^d_{++}$ between the inverse of a similarity matrix $\mathbf{S} = \mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 0] + \mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 1]$ and a dissimilarity matrix $\mathbf{D} = \mathbf{S} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$, where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y} \mid i = 0] - \mathbb{E}[\mathbf{y} \mid i = 1]$. Following Zadeh et al. (2016), this interpolation is given by

$$\mathbf{A} = \mathbf{S}^{-1/2}(\mathbf{S}^{1/2}\mathbf{D}\mathbf{S}^{1/2})^t\mathbf{S}^{-1/2}\,, \quad t \in [0, 1]\,.$$

We apply our multifidelity techniques to estimating the two covariance matrices that comprise $\mathbf{S}$, $\mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 0]$ and $\mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 1]$. In constructing $\mathbf{A}$ we set $t = 0.1$, motivated by (Zadeh et al., 2016, Figure 3).

Results Prior to applying our multi-fidelity estimators in this setting, we take a total of $24,000$ pilot samples between

[2] We use the Python package `pyqg` to compute our solutions; see http://pyqg.readthedocs.org/



(a) Euclidean metric (b) log-Euclidean metric (c) affine-invariant metric

Figure 6. Metric learning: Our LEMF estimator achieves the lowest MSE compared to the cost-equivalent benchmark estimator that uses high-fidelity samples alone. The Euclidean multi-fidelity estimator is indefinite over 100 trials and therefore fails to learn a valid metric.

$(\mathbf{y}^{(0)}, \mathbf{y}^{(1)}) \mid i = 0$ and $(\mathbf{y}^{(0)}, \mathbf{y}^{(1)}) \mid i = 1$ in order to estimate the generalized variances σ_0^2 and σ_1^2 and correlation ρ_1 in both classes. We additionally use these pilot samples to obtain a reference estimate of $\mathbf{A}$, which we denote $\mathbf{A}_0$.

We estimate $\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)} \,|\, i = 0]$ and $\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)} \,|\, i = 1]$ as follows: we construct multifidelity estimators from a combination of 15 high-fidelity samples and an additional number $n_{\mathrm{lo},i}$ of low-fidelity samples computed according to the optimal ratio in Proposition 3.4. We correspondingly construct equivalent-cost single-fidelity estimators with budgets

$$B_i = 15c_{\mathrm{hi}} + n_{\mathrm{lo},i}c_{\mathrm{lo}}, \quad i \in \{0, 1\},$$

spent entirely on high- or low-fidelity samples. We estimate $\mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 0]$ and $\mathbb{C}\mathrm{ov}[\mathbf{y} \mid i = 1]$ using the EMF, LEMF, and truncated EMF ($\delta = 10^{-16}$) estimators and the high-fidelity- and low-fidelity-only estimators with the sample allocations described above. The two covariance estimates are combined to obtain estimates of the metric matrix $\mathbf{A}$.

We first study the performance of the learned metrics relative to the reference metric. Let $\{\mathbf{y}_j^{(0)}\}_{j=1}^{5000} \in \mathbb{R}^9$ be a test set of 5,000 points and consider the mean relative error of distances $\mathrm{MRE}(\mathbf{A})$ with respect to the reference metric $\mathbf{A}_0$; see Appendix I for a definition. Figure 5 shows the MRE for each estimator averaged over 50 independent trials. The LEMF estimator outperforms the single-fidelity estimator and the truncated MF estimator (22) by more than 20%. The Euclidean multi-fidelity estimator gives indefinite estimates of $\mathbf{A}$, which cannot constitute valid metrics.

We further compare the MSEs of the learned metric matrices relative to the reference $\mathbf{A}_0$. Figure 6 shows that the LEMF estimator (7) yields a closer approximation of the reference metric; in particular the LEMF estimator gives a more accurate estimate of $\mathbf{A}$ relative to the high-fidelity-only estimator while the other estimators give less accurate estimates. The surrogate-only estimator is limited by its large bias. The EMF estimator (2) is again indefinite in $>10\%$ of trials.

5. Conclusions

We showed that formulating multi-fidelity estimation in a non-Euclidean geometry can be beneficial for enforcing structure: The proposed LEMF estimator leverages the log-Euclidean geometry for SPD matrices to guarantee that the resulting covariance matrix estimates are positive definite. This property is in contrast to state-of-the-art multi-fidelity estimators that can lose definiteness, especially in the small-sample regime, which is typical in science and engineering applications in which drawing samples is expensive. Our LEMF estimator preserves definiteness and leverages surrogates to achieve speedups of more than one order of magnitude in our experiments. The experiments further show that preserving definiteness is key for enabling tasks such as metric learning and classification, for which indefinite Euclidean multi-fidelity estimators are invalid. Future work includes combining the multi-fidelity approach with, e.g., shrinkage and other regularization schemes for high-dimensional covariance estimation.

Acknowledgements

AM and YM acknowledge support from the Office of Naval Research, SIMDA (Sea Ice Modeling and Data Assimilation) MURI, award number N00014-20-1-2595 (Dr. Reza Malek-Madani and Dr. Scott Harper). AM additionally acknowledges support from the NSF Graduate Research Fellowship under Grant No. 1745302. TA and BP acknowledge support from the AFOSR under Award Number FA9550-21-1-0222 (Dr. Fariba Fahroo).

Code Availability

Python implementations of the LEMF, EMF, and truncated EMF estimators are available at https://github.com/amaurais/LEMF.

References

Alsup, T., Venturi, L., and Peherstorfer, B. Multilevel Stein variational gradient descent with applications to bayesian inverse problems. In Bruna, J., Hesthaven, J., and Zdeborova, L. (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 93–117. PMLR, 16–19 Aug 2022.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

Bergemann, K. and Reich, S. A localization technique for ensemble Kalman filters. *Quarterly Journal of the Royal Meteorological Society*, 136(648):701–707, 2010.

Bhatia, R. *Positive Definite Matrices*. Princeton University Press, 2007.

Bickel, P. J. and Levina, E. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199 – 227, 2008. doi: 10.1214/009053607000000758.

Bierig, C. and Chernov, A. Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems. *Numerische Mathematik*, 130 (4):579–613, Aug 2015.

Cai, D. and Adams, R. P. Multi-fidelity Monte Carlo: a pseudo-marginal approach. *arXiv*, 2210.01534, 2022.

Capet, X., Klein, P., Hua, B. L., Lapeyre, G., and McWilliams, J. C. Surface kinetic energy transfer in surface quasi-geostrophic flows. *Journal of Fluid Mechanics*, 604:165 – 174, 2008.

Chernov, A., Hoel, H., Law, K. J. H., Nobile, F., and Tempone, R. Multilevel ensemble Kalman filtering for spatio-temporal processes. *Numerische Mathematik*, 147(1): 71–125, Jan 2021.

Cliffe, K. A., Giles, M. B., Scheichl, R., and Teckentrup, A. L. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3–15, 2011.

Cressie, N. *Statistics for spatial data*. John Wiley & Sons, 2015.

Donoho, D., Gavish, M., and Johnstone, I. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742 – 1778, 2018. doi: 10.1214/17-AOS1601.

Fan, J., Liao, Y., and Mincheva, M. High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320 – 3356, 2011.

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

Gundersen, G. W., Cai, D., Zhou, C., Engelhardt, B. E., and Adams, R. P. Active multi-fidelity Bayesian online changepoint detection. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1916–1926. PMLR, 27–30 Jul 2021.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer, 2001.

Held, I., Pierrehumbert, R., Garner, S., and Swanson, K. Surface quasi-geostrophic dynamics. *Journal of Fluid Mechanics*, 282:1–20, 1985.

Hoel, H., Law, K. J., and Tempone, R. Multilevel ensemble Kalman filtering. *SIAM Journal on Numerical Analysis*, 54(3):1813–1839, 2016.

Huang, Z., Wang, R., Shan, S., Li, X., and Chen, X. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 720–729, Lille, France, 07–09 Jul 2015. PMLR.

Konrad, J., Farcaş, I.-G., Peherstorfer, B., Di Siena, A., Jenko, F., Neckel, T., and Bungartz, H.-J. Data-driven low-fidelity models for multi-fidelity Monte Carlo sampling in plasma micro-turbulence analysis. *Journal of Computational Physics*, 451:110898, 2022.

Law, K., Stuart, A., and Zygalakis, K. *Data Assimilation*. Springer, 2015.

Ledoit, O. and Wolf, M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012. doi: 10.1214/12-AOS989.

Li, S., Xing, W., Kirby, R., and Zhe, S. Multi-fidelity Bayesian optimization via deep neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8521–8531. Curran Associates, Inc., 2020.

Lin, Z. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(88): 3065–3105, 2014.

Loh, P.-l. and Wainwright, M. J. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Malagò, L., Montrucchio, L., and Pistone, G. Wasserstein Riemannian geometry of gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

Marques, A., Lam, R., and Willcox, K. Contour location via entropy reduction leveraging multiple information sources. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mycek, P. and De Lozzo, M. Multilevel Monte Carlo covariance estimation for the computation of Sobol' indices. *SIAM/ASA Journal on Uncertainty Quantification*, 7(4): 1323–1348, 2019.

Ng, L. and Willcox, K. Multifidelity approaches for optimization under uncertainty. *International Journal for Numerical Methods in Engineering*, 100(10):746–772, 2014.

Palizhati, A., Torrisi, S. B., Aykol, M., Suram, S. K., Hummelshøj, J. S., and Montoya, J. H. Agents for sequential learning using multiple-fidelity data. *Scientific Reports*, 12(1):4694, Mar 2022.

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

Peherstorfer, B., Willcox, K., and Gunzburger, M. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

Poloczek, M., Wang, J., and Frazier, P. Multi-information source optimization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Popov, A. A., Mou, C., Sandu, A., and Iliescu, T. A multifidelity ensemble Kalman filter with reduced order control variates. *SIAM Journal on Scientific Computing*, 43(2): A1134–A1162, 2021.

Prescott, T. P. and Baker, R. E. Multifidelity approximate Bayesian computation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):114–138, 2020.

Qi, H. and Sun, D. A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications*, 28 (2):360–385, 2006.

Qian, E., Peherstorfer, B., O'Malley, D., Vesselinov, V. V., and Willcox, K. Multifidelity Monte Carlo estimation of variance and sensitivity indices. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):683–706, 2018.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(none):935 – 980, 2011.

Schillings, C. and Stuart, A. M. Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, 2017.

Teckentrup, A. L., Scheichl, R., Giles, M. B., and Ullmann, E. Further analysis of multilevel Monte Carlo methods for elliptic pdes with random coefficients. *Numerische Mathematik*, 125(3):569–600, Nov 2013.

Tsiligkaridis, T. and Hero, A. O. Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360, 2013. doi: 10.1109/TSP.2013.2279355.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy

1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wu, J., Toscano-Palmerin, S., Frazier, P. I., and Wilson, A. G. Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 788–798. PMLR, 22–25 Jul 2020.

Zadeh, P. H., Hosseini, R., and Sra, S. Geometric mean metric learning. In *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, pp. 2464–2471, June 2016.

A. Connection to Fréchet Averaging

In addition to being a control-variate estimator in the log-Euclidean geometry, Equation (7) can also be viewed as a Fréchet average. Recall that the LEMF estimator takes the form

$$\operatorname{Log}\hat{\mathbf{\Sigma}}_{\mathbf{n}}^{\mathrm{LEMF}} = \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right).$$

Defining the SPD "difference" matrices

$$\mathbf{D}_\ell = \operatorname{Exp}(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}) = \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}, \quad \ell = 1,\dots,L$$

we equivalently have

$$\operatorname{Log}\hat{\mathbf{\Sigma}}_{\mathbf{n}}^{\mathrm{LEMF}} = \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell \operatorname{Log}\mathbf{D}_\ell. \tag{25}$$

Assuming that $\alpha_1,\dots,\alpha_L > 0$ [3], we see in (25) by definition of the Fréchet mean in the log-Euclidean geometry (Arsigny et al., 2006) that $\hat{\mathbf{\Sigma}}_{\mathbf{n}}^{\mathrm{LEMF}}$ satisfies

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{\mathbf{n}}^{\mathrm{LEMF}} &\equiv \mathbb{E}_{\mathrm{LE},\boldsymbol{\alpha}}(\hat{\mathbf{\Sigma}}_{n_0}^{(0)}, \mathbf{D}_1,\dots,\mathbf{D}_L)\\ &= \operatorname*{arg\,min}_{\mathbf{\Sigma}\in\mathbb{S}_{++}^d}\left(d_{\mathrm{LE}}^2(\mathbf{\Sigma},\hat{\mathbf{\Sigma}}_{n_0}^{(0)}) + \sum_{\ell=1}^{L}\alpha_\ell\, d_{\mathrm{LE}}^2(\mathbf{\Sigma},\mathbf{D}_\ell)\right)\end{aligned}$$

where $\boldsymbol{\alpha} = \begin{bmatrix}1 & \alpha_1 & \cdots & \alpha_L\end{bmatrix}^\top \in \mathbb{R}^{L+1}$. Thus $\hat{\mathbf{\Sigma}}_{\mathbf{n}}^{\mathrm{LEMF}}$ can be interpreted as a *weighted Fréchet average* between high-fidelity $\hat{\mathbf{\Sigma}}_{n_0}^{(0)}$ and the low-fidelity perturbations $\mathbf{D}_\ell$, $\ell = 1,\dots,L$, with weights assigned according to how much $\hat{\mathbf{\Sigma}}_{n_0}^{(0)}$ is correlated in a generalized sense with each $\mathbf{D}_\ell$.

B. Proof of Proposition 3.1

We have defined the log-Euclidean multi-fidelity estimator by

$$\hat{\mathbf{\Sigma}}_{n}^{\mathrm{LEMF}} = \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L}\alpha_\ell\odot\left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right) = \operatorname{Exp}\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right). \tag{26}$$

[3] If $\alpha_\ell < 0$ for some ℓ, we can define $\beta_\ell = -\alpha_\ell$ and redefine $\mathbf{D}_\ell = \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}$ by reversing the order of subtraction in order to ensure that the Fréchet weights are all positive.

Because we have assumed $n_0, n_1, \dots, n_L > d$, the sample covariance matrices $\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}$, and $\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}$, $\ell = 0,\dots,L$ are positive definite almost surely. That is, they are members of $\mathbb{S}_{++}^d$. Because $\mathbb{S}_{++}^d(\oplus,\odot)$ is a vector space, it follows immediately that the top line of (26)

$$\hat{\mathbf{\Sigma}}_{n}^{\mathrm{LEMF}} = \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L}\alpha_\ell\odot\left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right) \in \mathbb{S}_{++}^d$$

is positive definite. It remains to show that the bottom line of (26) is equivalent to the top line. This we accomplish by straightforward algebra,

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{n}^{\mathrm{LEMF}} &= \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L}\alpha_\ell\odot\left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \ominus \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\\ &= \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L}\alpha_\ell\odot\left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} \oplus \operatorname{Exp}(-\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)})\right)\\ &= \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \bigoplus_{\ell=1}^{L}\operatorname{Exp}\left(\alpha_\ell\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right)\\ &= \hat{\mathbf{\Sigma}}_{n_0}^{(0)} \oplus \operatorname{Exp}\left(\sum_{\ell=1}^{L}\alpha_\ell\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right)\\ &= \operatorname{Exp}\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell\left(\operatorname{Log}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \operatorname{Log}\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right).\end{aligned} \tag{27}$$

C. Proof of Corollary 3.3

We first note that, by properties of inner and outer products on $\mathbb{S}^d$, for two random matrices $\mathbf{A},\mathbf{B}\in\mathbb{S}^d$ with $\mathbb{E}[\mathbf{A}] = \mathbb{E}[\mathbf{B}] = \mathbf{0}$ we have

$$\begin{aligned}\mathbb{E}\left[\langle\mathbf{A},\ \mathbf{B}\rangle_{\mathrm{F}}\right] &= \mathbb{E}[\operatorname{Tr}(\mathbf{A}\otimes\mathbf{B})]\\ &= \operatorname{Tr}(\mathbb{E}[\mathbf{A}\otimes\mathbf{B}]) = \operatorname{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{A},\mathbf{B}]),\end{aligned} \tag{28}$$

where we have used linearity of trace. In the special case where $\mathbf{A} = \mathbf{B}$ we have

$$\mathbb{E}[\|\mathbf{A}\|_{\mathrm{F}}^2] = \mathbb{E}[\langle\mathbf{A},\ \mathbf{A}\rangle_{\mathrm{F}}] = \operatorname{Tr}\left(\mathbb{C}\mathrm{ov}[\mathbf{A}]\right). \tag{29}$$

We now proceed with the remainder of the proposition.

Recall that the EMF estimator is defined

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} = \hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum\nolimits_{\ell=1}^{L}\alpha_\ell\left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right),$$

where $\boldsymbol{\alpha} = [\alpha_1,\dots,\alpha_L]^\top \in \mathbb{R}^L$ are the control variate weights and

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} &= \frac{1}{n_\ell}\sum\nolimits_{i=1}^{n_\ell}\mathbf{y}_i^{(\ell)}\left(\mathbf{y}_i^{(\ell)}\right)^\top,\\ \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} &= \frac{1}{n_{\ell-1}}\sum\nolimits_{i=1}^{n_{\ell-1}}\mathbf{y}_i^{(\ell)}\left(\mathbf{y}_i^{(\ell)}\right)^\top\end{aligned} \tag{30}$$

are sample covariance matrices computed from $\{\mathbf{y}_i^{(\ell)}\}_{i=1}^{n_\ell}$, $\ell = 0, \ldots, L$ and we have assumed that $\mathbb{E}[\mathbf{y}^{(\ell)}]$ is known and without loss of generality equal to $\mathbf{0}$, $\ell = 0, \ldots, L$. Because the means of $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(L)}$ are known, in defining the SCMs in Equation (30) we have normalized by n_ℓ and $n_{\ell-1}$ to obtain unbiased sample covariance estimates. Under this assumption the EMF estimator (2) with SCMs defined as in (30) is unbiased,

$$\mathbb{E}\left[\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}\right] = \mathbb{E}\left[\hat{\mathbf{\Sigma}}_{n_0}^{(0)} + \sum\nolimits_{\ell=1}^{L} \alpha_\ell \left(\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right)\right] = \mathbf{\Sigma}.$$

Because of the unbiasedness, the Frobenius norm MSE of $\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}$ is thus equal to the trace of the covariance of $\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}$,

$$\mathbb{E}[||\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}||_F^2] = \mathbb{E}[||\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbb{E}[\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}]||_F^2] = \mathrm{Tr}(\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}]).$$

The covariance of $\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}$ can be decomposed

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}] &= \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_0}^{(0)}] \\
&\quad + \sum_{\ell=1}^{L} \alpha_\ell^2 (\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}] + \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}]) \\
&\quad + \sum_{\ell=1}^{L} \alpha_\ell \Big(\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_0}^{(0)}] + \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_0}^{(0)}, \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}] \\
&\quad - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_0}^{(0)}] - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_0}^{(0)}, \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}]\Big) \\
&+ \sum_{\ell=1}^{L} \sum_{m=\ell+1}^{L} \alpha_\ell \alpha_m \Big(\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_m}^{(m)}] - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_m}^{(m)}] \\
&\quad - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_{m-1}}^{(m)}] + \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_{m-1}}^{(m)}]\Big) \\
&+ \sum_{\ell=1}^{L} \sum_{m=\ell+1}^{L} \alpha_\ell \alpha_m \Big(\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_m}^{(m)}, \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}] - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_m}^{(m)}, \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}] \\
&\quad - \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{m-1}}^{(m)}, \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}] + \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{m-1}}^{(m)}, \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}]\Big) \\
&\quad - \sum_{\ell=1}^{L} \alpha_\ell^2 \Big(\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)})] + \mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)}, \hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)}]\Big)
\end{aligned} \tag{31}$$

We will need a result analogous to Lemma 3.2 in Peherstorfer et al. (2016) in order to simplify (31).

Lemma C.1. *Consider* $\hat{\mathbf{\Sigma}}_{m_i}^{(\ell)} = \frac{1}{m_i}\sum_{i'=1}^{m_i} \mathbf{y}_{i'}^{(\ell)}(\mathbf{y}_{i'}^{(\ell)})^\top$ *and* $\hat{\mathbf{\Sigma}}_{m_j}^{(t)} = \frac{1}{m_j}\sum_{j'=1}^{m_j} \mathbf{y}_{j'}^{(t)}(\mathbf{y}_{j'}^{(t)})^\top$, *where* $\ell, t \in \{0, \ldots, L\}$ *and* $m_i, m_j \in \mathbb{Z}^+$. *It holds that*

$$\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{m_i}^{(\ell)},\ \hat{\mathbf{\Sigma}}_{m_j}^{(t)}] = \frac{1}{\max\{m_i, m_j\}} \mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top,\ \mathbf{y}^{(t)}(\mathbf{y}^{(t)})^\top]$$

Proof. Proceeding algebraically, we see

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{m_i}^{(\ell)},\ \hat{\mathbf{\Sigma}}_{m_j}^{(t)}] = \\
&\mathbb{C}\mathrm{ov}\left[\frac{1}{m_i}\sum_{i'=1}^{m_i} \mathbf{y}_{i'}^{(\ell)}(\mathbf{y}_{i'}^{(\ell)})^\top,\ \frac{1}{m_j}\sum_{j'=1}^{m_j} \mathbf{y}_{j'}^{(t)}(\mathbf{y}_{j'}^{(t)})^\top\right] \\
&= \frac{1}{m_i m_j}\mathbb{C}\mathrm{ov}\left[\sum_{i'=1}^{m_i} \mathbf{y}_{i'}^{(\ell)}(\mathbf{y}_{i'}^{(\ell)})^\top,\ \sum_{j'=1}^{m_j} \mathbf{y}_{j'}^{(t)}(\mathbf{y}_{j'}^{(t)})^\top\right].
\end{aligned}$$

$\mathbf{y}_{i'}^{(\ell)}(\mathbf{y}_{i'}^{(\ell)})^\top$ and $\mathbf{y}_{j'}^{(t)}(\mathbf{y}_{j'}^{(t)})^\top$ are independent except in the case that $i' = j'$. Thus the terms that remain are

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{m_i}^{(\ell)},\ \hat{\mathbf{\Sigma}}_{m_j}^{(t)}] = \\
&\frac{1}{m_i m_j}\sum_{i'=1}^{\min\{m_i,m_j\}} \mathbb{C}\mathrm{ov}[\mathbf{y}_{i'}^{(\ell)}(\mathbf{y}_{i'}^{(\ell)})^\top,\ \mathbf{y}_{i'}^{(t)}(\mathbf{y}_{i'}^{(t)})^\top] \\
&= \frac{\min\{m_i, m_j\}}{m_i m_j}\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top,\ \mathbf{y}^{(t)}(\mathbf{y}^{(t)})^\top] \\
&= \frac{1}{\max\{m_i, m_j\}}\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top,\ \mathbf{y}^{(t)}(\mathbf{y}^{(t)})^\top].
\end{aligned}$$

□

We return to Equation (31) with this result and simplify, seeing that the doubly-indexed sums are zero,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}] &= \tfrac{1}{n_0}\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top] \\
&\quad + \sum_{\ell=1}^{L} \alpha_\ell^2 (\tfrac{1}{n_\ell} + \tfrac{1}{n_{\ell-1}})\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top] \\
&+ \sum_{\ell=1}^{L} \alpha_\ell (\tfrac{1}{n_\ell} - \tfrac{1}{n_{\ell-1}}) \Big(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top,\ \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top] + \\
&\quad + \mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top,\ \mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top]\Big) \\
&\quad - \sum_{\ell=1}^{L} \alpha_\ell^2 \tfrac{2}{n_\ell}\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top].
\end{aligned}$$

We rearrange and write the covariance of $\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}$ compactly

as

$$\begin{aligned}\mathbb{C}\mathrm{ov}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}] &= \tfrac{1}{n_0}\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top] \\ &+ \sum_{\ell=1}^{L}(\tfrac{1}{n_{\ell-1}} - \tfrac{1}{n_\ell})\Big(\alpha_\ell^2\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top] \\ &- \alpha_\ell\Big(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top, \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top] \\ &\quad + \mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top, \mathbf{y}^{(\ell)}(\mathbf{y}^{(\ell)})^\top]\Big)\Big).\end{aligned}$$

The MSE of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}$ is equal to the trace of its covariance,

$$\begin{aligned}\mathbb{E}[||\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \boldsymbol{\Sigma}||] &= \mathrm{Tr}(\mathbb{C}\mathrm{ov}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}}]) \\ &= \frac{\sigma_0^2}{n_0} + \sum_{\ell=1}^{L}(\frac{1}{n_{\ell-1}} - \frac{1}{n_\ell})(\alpha_\ell^2\sigma_\ell^2 - 2\alpha_\ell\rho_\ell\sigma_0\sigma_\ell)\end{aligned}$$

which is the result we sought to show.

D. Proof of Proposition 3.2

Proof. Canceling the matrix logarithm with the matrix exponential in the LEMF estimator (7) yields

$$\begin{aligned}&\left\|\mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\boldsymbol{\Sigma}\right\|_{\mathrm{F}} = \\ &\left\|\mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell\left(\mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right) - \mathrm{Log}\,\boldsymbol{\Sigma}\right\|_{\mathrm{F}}.\end{aligned}$$

Since $\|\boldsymbol{\Sigma} - \mathbf{I}\|_{\mathrm{F}} < h/4 < 1$ we may define the matrix logarithm through its Taylor series

$$\mathrm{Log}(\boldsymbol{\Sigma}) = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}(\boldsymbol{\Sigma} - \mathbf{I})^k, \tag{32}$$

which converges absolutely. Similarly, by the triangle inequality

$$\left\|\boldsymbol{\Sigma}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}} \le \|\boldsymbol{\Sigma} - \mathbf{I}\|_{\mathrm{F}} + \left\|\boldsymbol{\Sigma}^{(\ell)} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} \le \frac{h}{2} < 1. \tag{33}$$

For the sample covariance matrices $\boldsymbol{\Sigma}_{n_\ell}^{(\ell)}$ and $\boldsymbol{\Sigma}_{n_{\ell-1}}^{(\ell)}$, the law of large numbers implies almost surely that for any $\delta_\ell > 0$ sufficiently large sample size n_ℓ guarantees

$$\max\left\{\left\|\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \boldsymbol{\Sigma}^{(\ell)}\right\|_{\mathrm{F}}, \left\|\hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \boldsymbol{\Sigma}^{(\ell)}\right\|_{\mathrm{F}}\right\} < \delta_\ell, \tag{34}$$

for $\ell = 0, \ldots, L$. Setting $\delta_\ell = h/4$ and applying the triangle inequality again with the result of (33) gives

$$\begin{aligned}&\max\left\{\left\|\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}, \left\|\hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}\right\} < \\ &\qquad \delta_\ell + \left\|\boldsymbol{\Sigma}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}} < \delta_\ell + \frac{h}{2} < \frac{3h}{4} < 1.\end{aligned} \tag{35}$$

Expanding the Taylor series for each matrix logarithm up to first-order (i.e. $\mathrm{Log}(\boldsymbol{\Sigma}) \approx \boldsymbol{\Sigma} - \mathbf{I}$) and canceling out the resulting identity matrices gives

$$\begin{aligned}&\left\|\mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\boldsymbol{\Sigma}\right\|_{\mathrm{F}} = \\ &\left\|\hat{\boldsymbol{\Sigma}}_{n_0}^{(0)} + \sum_{\ell=1}^{L}\alpha_\ell\left(\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)}\right) - \boldsymbol{\Sigma} + \mathbf{M}\right\|_{\mathrm{F}},\end{aligned} \tag{36}$$

where the matrix $\mathbf{M}$ is the remainder of the first-order Taylor expansions

$$\begin{aligned}\mathbf{M} = &\sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}(\hat{\boldsymbol{\Sigma}}_{n_0}^{(0)} - \mathbf{I})^k \\ &+ \sum_{\ell=1}^{L}\alpha_\ell\left(\sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}(\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I})^k\right) \\ &- \sum_{\ell=1}^{L}\alpha_\ell\left(\sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}(\hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I})^k\right) \\ &- \sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}(\boldsymbol{\Sigma} - \mathbf{I})^k.\end{aligned} \tag{37}$$

Applying both the triangle and reverse-triangle inequalities gives

$$\left|\left\|\mathrm{Log}\,\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\boldsymbol{\Sigma}\right\|_{\mathrm{F}} - \left\|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}\right| \le \|\mathbf{M}\|_{\mathrm{F}}. \tag{38}$$

To bound the right-hand-side of (38), we again use the triangle inequality and that $\left\|\mathbf{A}^k\right\|_{\mathrm{F}} \le \|\mathbf{A}\|_{\mathrm{F}}^k$ to obtain

$$\begin{aligned}\|\mathbf{M}\|_F \le &\sum_{k=2}^{\infty}\frac{1}{k}\left\|\hat{\boldsymbol{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}^k \\ &+ \sum_{\ell=1}^{L}\alpha_\ell\left(\sum_{k=2}^{\infty}\frac{1}{k}\left\|\hat{\boldsymbol{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^k\right) \\ &+ \sum_{\ell=1}^{L}\alpha_\ell\left(\sum_{k=2}^{\infty}\frac{1}{k}\left\|\hat{\boldsymbol{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^k\right) \\ &+ \sum_{k=2}^{\infty}\frac{1}{k}\|\boldsymbol{\Sigma} - \mathbf{I}\|_{\mathrm{F}}^k.\end{aligned} \tag{39}$$

By (35) and the assumption $\|\boldsymbol{\Sigma} - \mathbf{I}\|_{\mathrm{F}} < h/4$, each series in (39) converges absolutely. Bounding $1/k$ with 1 and

factoring out a quadratic power for each term in (39) gives

$$\begin{aligned}\|\mathbf{M}\|_F \leq & \left\|\hat{\mathbf{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \sum_{k=0}^{\infty} \left\|\hat{\mathbf{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}^k \\ & + \sum_{\ell=1}^{L} \alpha_\ell \left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \left(\sum_{k=0}^{\infty} \left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^k\right) \\ & + \sum_{\ell=1}^{L} \alpha_\ell \left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \left(\sum_{k=0}^{\infty} \left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^k\right) \\ & + \|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}^2 \sum_{k=0}^{\infty} \|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}^k \,. \end{aligned} \tag{40}$$

Evaluating each of these geometric series gives the bound

$$\begin{aligned}\|\mathbf{M}\|_{\mathrm{F}} \leq & \frac{\left\|\hat{\mathbf{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}^2}{1 - \left\|\hat{\mathbf{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}} \\ & + \sum_{\ell=1}^{L} \alpha_\ell \frac{\left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2}{1 - \left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}} \\ & + \sum_{\ell=1}^{L} \alpha_\ell \frac{\left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2}{1 - \left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}} \\ & + \frac{\|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}^2}{1 - \|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}} \,. \end{aligned} \tag{41}$$

By (35) we have that

$$\max\left\{\frac{1}{1 - \left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}}, \frac{1}{1 - \left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}}\right\} < \frac{1}{1 - 3h/4} < 4\,, \tag{42}$$

and similarly

$$\frac{1}{1 - \|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}} < \frac{1}{1 - h/4} < \frac{4}{3}\,. \tag{43}$$

Substituting the bounds (42) and (43) into (41) gives

$$\begin{aligned}\|\mathbf{M}\|_{\mathrm{F}} \leq & \, 4 \left\|\hat{\mathbf{\Sigma}}_{n_0}^{(0)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \\ & + 4 \sum_{\ell=1}^{L} \alpha_\ell \left\|\hat{\mathbf{\Sigma}}_{n_\ell}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \\ & + 4 \sum_{\ell=1}^{L} \alpha_\ell \left\|\hat{\mathbf{\Sigma}}_{n_{\ell-1}}^{(\ell)} - \mathbf{I}\right\|_{\mathrm{F}}^2 \\ & + \frac{4}{3} \|\mathbf{\Sigma} - \mathbf{I}\|_{\mathrm{F}}^2 \,. \end{aligned} \tag{44}$$

Applying the bound (35) with (44) gives

$$\begin{aligned}\|\mathbf{M}\|_{\mathrm{F}} &\leq 4\left(\frac{3h}{4}\right)^2 + 8\sum_{\ell=1}^{L} \alpha_\ell \left(\frac{3h}{4}\right)^2 + \frac{4}{3}\left(\frac{h}{4}\right)^2 \\ &\leq \left(\frac{7}{3} + \frac{9}{2}\sum_{\ell=1}^{L} \alpha_\ell\right) h^2 = Kh^2\,. \end{aligned} \tag{45}$$

Factoring using the difference of squares formula ($a^2 - b^2 = (a-b)(a+b)$) gives

$$\begin{aligned}&\left|\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 - \left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right| = \\ &\left|\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}} + \left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}\right| \\ &\times \left|\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}} - \left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}\right|\,, \end{aligned} \tag{46}$$

Applying the law of large numbers again, we know that almost surely for sufficiently large sample sizes

$$\max\left\{\left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}, \left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}\right\} < 1\,. \tag{47}$$

Combining the bound (45) with (38) and substituting into (46) gives

$$\left|\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 - \left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right| \leq 2Kh^2\,. \tag{48}$$

Taking the expectation and applying Jensen's inequality gives the result.

$$\begin{aligned}&\left|\mathbb{E}\left[\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right] - \mathbb{E}\left[\left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right]\right| \\ &\leq \mathbb{E}\left[\left|\left\|\mathrm{Log}\,\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{LEMF}} - \mathrm{Log}\,\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 - \left\|\hat{\mathbf{\Sigma}}_{\boldsymbol{n}}^{\mathrm{EMF}} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right|\right] \\ &\leq 2Kh^2\,. \end{aligned} \tag{49}$$

□

E. Proof of Proposition 3.4 and Corollary 3.5

Since the Euclidean MSE (Corollary 3.3) has the exact same form as the MSE of the scalar multi-fidelity Monte Carlo estimator in Peherstorfer et al. (2016), the optimal allocations and weights there apply directly to the Euclidean multi-fidelity covariance estimator (2) and in first-order to the LEMF estimator (7), with the exception that our definitions of σ_ℓ^2 and ρ_ℓ are generalized for matrices. Equation 20 directly follows from (19).

F. Algorithmic Description and Code Sketch of LEMF Estimator

The code shown in Figure 7 sketches an implementation of the LEMF estimator. It is important to note that it only

```
from numpy import cov
from scipy.linalg import logm, expm

# compute sample-size vector n[] and coefficients a[] as in Proposition 3.4
# collect samples n_l x d of level l = 0, ..., L in array samps[l]

S = logm(cov(samps[0].T))
for i in range(len(a)):
  S = S + a[i]*(logm(cov(samps[i+1].T))-logm(cov(samps[i+1][:n[i]].T)))
S = expm(S)
```

Figure 7. This code sketches an implementation of the LEMF estimator. It builds on standard numerical linear algebra routines that are readily available in NumPy/SciPy (Harris et al., 2020; Virtanen et al., 2020).

requires numerical linear algebra functions that are readily available in, e.g., NumPy/SciPy (Harris et al., 2020; Virtanen et al., 2020). A detailed implementation to reproduce the shown numerical results is provided in the supplemental material.

G. Supplemental to "Motivating Example with Gaussian"

Consider high-fidelity samples $\mathbf{y}^{(0)} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \in \mathbb{S}^d_{++}$ and low-fidelity samples which correspond to high-fidelity samples corrupted by additional independent noise,

$$\mathbf{y}^{(\ell)} = \mathbf{y} + \boldsymbol{\epsilon}^{(\ell)}\,, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Gamma}^{(\ell)})\,. \tag{50}$$

Because the samples are Gaussian we may compute explicitly the generalized variances (10) and generalized correlations (9) needed for the optimal sample allocation in Theorem 3.4.

Generalized variance For the generalized variance we have that

$$\begin{aligned}
&\mathrm{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top]) \\
&= \mathbb{E}\left[\left\|\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top\right\|_{\mathrm{F}}^2\right] - 2\mathbb{E}\left[\left\langle \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top,\, \mathbf{\Sigma}\right\rangle_{\mathrm{F}}\right] + \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 \\
&= \mathbb{E}\left[\mathrm{Tr}(\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top)\right] - \|\mathbf{\Sigma}\|_{\mathrm{F}}^2\,.
\end{aligned}$$

Focusing on the first term we see

$$\mathrm{Tr}\left(\left\|\mathbf{y}^{(0)}\right\|^2 \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top\right) = \left\|\mathbf{y}^{(0)}\right\|^2 \sum_{i=1}^d \left(y_i^{(0)}\right)^2 = \left\|\mathbf{y}^{(0)}\right\|^4,$$

so that

$$\mathrm{Tr}\left(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top]\right) = \mathbb{E}\left[\left\|\mathbf{y}^{(0)}\right\|^4\right] - \|\mathbf{\Sigma}\|_{\mathrm{F}}^2\,,$$

and is analogous to the formula $\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ for scalar random variables. For a multivariate Gaussian random variable with zero mean we have that

$$\mathbb{E}\left[\left\|\mathbf{y}^{(0)}\right\|^4\right] = \mathrm{Tr}(\mathbf{\Sigma})^2 + 2\mathrm{Tr}(\mathbf{\Sigma}^2)\,. \tag{51}$$

Therefore,

$$\mathrm{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top]) = \mathrm{Tr}(\mathbf{\Sigma}^2) + \mathrm{Tr}(\mathbf{\Sigma})^2\,. \tag{52}$$

Similarly, we may apply (52) for the low-fidelity samples $\mathbf{y}^{(\ell)} \sim N(\mathbf{0}, \mathbf{\Sigma} + \mathbf{\Gamma}^{(\ell)})$.

Generalized correlation In order to compute the generalized correlation we compute the cross-covariance,

$$\begin{aligned}
&\mathrm{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top,\, (\mathbf{y}^{(\ell)})(\mathbf{y}^{(\ell)})^\top]) \\
&= \mathbb{E}\left[\left\langle \mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top - \mathbf{\Sigma},\, (\mathbf{y}^{(\ell)})(\mathbf{y}^{(\ell)})^\top - \mathbf{\Sigma}^{(\ell)}\right\rangle_{\mathrm{F}}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top(\mathbf{y}^{(\ell)})(\mathbf{y}^{(\ell)})^\top\right)\right] - \mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^{(\ell)}) \\
&= \mathbb{E}\left[\left((\mathbf{y}^{(0)})^\top(\mathbf{y}^{(\ell)})\right)^2\right] - \mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^{(\ell)})\,,
\end{aligned}$$

where we have used the notation $\mathbf{\Sigma}^{(\ell)} = \mathbf{\Sigma} + \mathbf{\Gamma}^{(\ell)}$. Writing $\mathbf{y}^{(\ell)} = \mathbf{y}^{(0)} + \boldsymbol{\epsilon}^{(\ell)}$ and expanding the squared inner product gives

$$\begin{aligned}
\left((\mathbf{y}^{(0)})^\top(\mathbf{y}^{(\ell)})\right)^2 &= \left(\left\|\mathbf{y}^{(0)}\right\|^2 + (\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}\right)^2 \\
&= \left\|\mathbf{y}^{(0)}\right\|^4 + 2\left\|\mathbf{y}^{(0)}\right\|^2 (\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)} + \\
&\qquad \left((\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}\right)^2\,.
\end{aligned}$$

For the first term we may use (51) while for the second term we have by independence that

$$\mathbb{E}\left[\left\|\mathbf{y}^{(0)}\right\|^2 (\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}\right] = \mathbb{E}\left[\left\|\mathbf{y}^{(0)}\right\|^2 \mathbf{y}^{(0)}\right]^\top \mathbb{E}\left[\boldsymbol{\epsilon}^{(\ell)}\right] = 0\,.$$

For the last term we rely on properties of the trace operator.

$$\begin{aligned}\mathbb{E}\left[\left((\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}\right)^2\right] &= \mathbb{E}\left[\mathrm{Tr}\left((\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}(\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}\right)\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left((\mathbf{y}^{(0)})^\top\boldsymbol{\epsilon}^{(\ell)}(\boldsymbol{\epsilon}^{(\ell)})^\top\mathbf{y}^{(0)}\right)\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top(\boldsymbol{\epsilon}^{(\ell)})(\boldsymbol{\epsilon}^{(\ell)})^\top\right)\right] \\ &= \mathrm{Tr}\left(\mathbb{E}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top]\mathbb{E}[(\boldsymbol{\epsilon}^{(\ell)})(\boldsymbol{\epsilon}^{(\ell)})^\top]\right) \\ &= \mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{(\ell)})\,.\end{aligned}$$

Combining everything gives the cross-covariance

$$\begin{aligned}&\mathrm{Tr}(\mathbb{C}\mathrm{ov}[\mathbf{y}^{(0)}(\mathbf{y}^{(0)})^\top,\ (\mathbf{y}^{(\ell)})(\mathbf{y}^{(\ell)})^\top]) \\ &\qquad = \mathrm{Tr}(\boldsymbol{\Sigma})^2 + \mathrm{Tr}(\boldsymbol{\Sigma}^2) = \sigma_0^2\,, \quad (53)\end{aligned}$$

which is analogous to the scalar setting where $\mathbb{C}\mathrm{ov}[y,\epsilon]=0$ so that $\mathbb{C}\mathrm{ov}[y, y+\epsilon] = \mathrm{Var}[y]$. Using the notation σ_0, σ_ℓ, and ρ_ℓ defined in Section 3.3 we have that the generalized correlation is given by

$$\rho_\ell = \frac{\sigma_0}{\sigma_\ell}\,. \quad (54)$$

Notice that $\sigma_\ell > \sigma$ and so $\rho_\ell < 1$.

Example set up For the example presented in Section 4.1, we select $\boldsymbol{\Sigma}$ randomly from the Wishart(4) ensemble, that is, $\boldsymbol{\Sigma} = \mathbf{A}^\top\mathbf{A}$ where $\mathbf{A} \in \mathbb{R}^{4\times 4}$ has i.i.d. standard normal entries. The particular realization of $\boldsymbol{\Sigma}$ we obtained was

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.52 & -0.17 & 0.67 & -0.98 \\ -0.17 & 0.64 & -0.29 & 0.35 \\ 0.67 & -0.29 & 0.49 & -0.52 \\ -0.98 & 0.35 & -0.52 & 1.31 \end{bmatrix}.$$

The noise covariance matrices defining the low-fidelity samples at each level $\ell \in \{1,2,3\}$ are given by

$$\boldsymbol{\Gamma}^{(1)} = 0.1\mathbf{I}_{4\times 4}, \quad \boldsymbol{\Gamma}^{(2)} = 0.5\mathbf{I}_{4\times 4}, \quad \boldsymbol{\Gamma}^{(3)} = \mathbf{I}_{4\times 4}\,,$$

and result in generalized correlations (9)

$$\rho_1 \approx 0.93, \quad \rho_2 \approx 0.74, \quad \rho_3 \approx 0.58\,.$$

We impose sampling costs $c_0 = 1$ and $c_\ell = 10^{-\ell-1}$ for $\ell \in \{1,2,3\}$. Setting the computational budget to $B = 15$ and taking sample sizes $\lfloor \boldsymbol{n}^* \rfloor$ from Theorem 3.4 yields

$$n = 12, \quad n_1 = 199, \quad n_2 = 505, \quad n_3 = 2073\,.$$

Using the derived sample allocation we draw the corresponding number of samples from each level to obtain the three multi-fidelity estimators: EMF (2), LEMF (7), and the truncated estimator (22). For the truncated multi-fidelity estimator, we set a threshold of $\delta = 10^{-16}$ on the eigenvalues. Additionally, we also draw $n_{\text{high}} = B/c = 15$ high-fidelity samples and $n_{\text{low}} = B/c_3 = 1.5\times 10^5$ low-fidelity samples $\mathbf{y}^{(3)}$ to compute single-fidelity estimates with the same computational budget. The results reported in Table 1 are estimated MSE averaged over 100 independent trials. Across all 100 trials, about 5% of the EMF estimates were indefinite.

H. Supplemental to "Uncertainty Quantification of Steady-State Heat Flow"

For demonstration purposes, we show in Figure 8 the numerical solution of the steady-state heat-flow problem (23) for three different, randomly sampled parameters $\boldsymbol{\theta}$. Observations $\mathbf{y}^{(\ell)}$, $\ell \in \{0,1\}$, consist of temperature measurements at 10 equidistant locations over the spatial domain, where ℓ corresponds to the number of grid points used to numerically solve (23). Our goal in this example is to estimate the covariance of $\mathbf{y}^{(0)}$, the observations corresponding to the highest number of solver grid points.

We perform a pilot study to estimate the generalized variances and correlations $\sigma_0, \sigma_1, \rho_1$ needed to compute the optimal multi-fidelity sample allocations $\boldsymbol{n}^\star = [n_\star^{(0)}, n_\star^{(1)}]$ and coefficients $\boldsymbol{\alpha}^\star$. We use 10^5 samples of $\boldsymbol{\theta}$ for the pilot study to avoid mixing errors of the optimal sample allocation with the MSEs of the estimators.

Figure 9 is analogus to Figure 2 and shows the MSEs of our LEMF estimator compared to the MSEs of single-fidelity and other multi-fidelity estimators, except that now the minimum and maximum MSEs over all 100 trials are shown with error bars.

Figure 10a-b shows speedups of our LEMF estimator with respect to the Euclidean and affine-invariant metric. As the budget is increased, the Euclidean estimator becomes indefinite less frequently, which can be seen in Figure 10c.

Figure 11 visualizes the convergence behavior of various estimators. Our LEMF estimator converges more quickly to the true value of $\boldsymbol{\Sigma}$ element-wise than the single-fidelity estimators; this is particularly true for the diagonal elements of $\boldsymbol{\Sigma}$. The y-scale is comparable to that of Figure 3c and the x-scale shows the costs and is the same as the x-axis of Figure 3c.

I. Supplemental to "Metric Learning: Surface Quasi-Geostrophy"

We follow the surface quasi-geostrophic (SQG) model for the evolution of surface buoyancy $b : \mathcal{X}\times[0,\infty) \to \mathbb{R}$ in the periodic domain $\mathcal{X} = [-\pi,\pi]^2 \times (-\infty, 0]$ presented

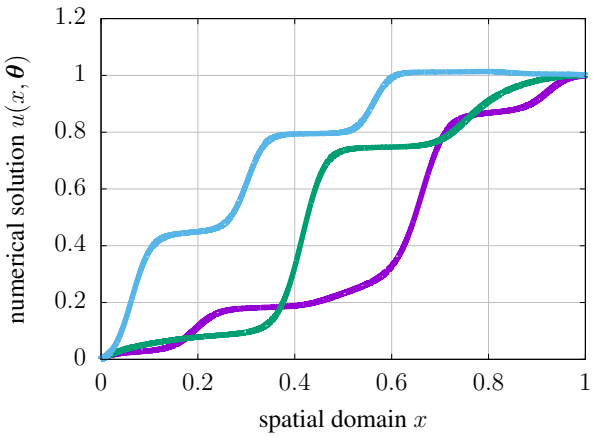


Figure 8. Supplemental to heat flow example: Three sample solutions (temperature) of the steady-state heat equation (23) corresponding to different random parameters $\boldsymbol{\theta} \in \mathbb{R}^4$.

in Held et al. (1985) and Capet et al. (2008).

$$\begin{aligned} \frac{\partial}{\partial t} b(\mathbf{x}, t; \boldsymbol{\theta}) + J(\psi, b) &= 0\,, \quad z = 0\,, \\ b &= \frac{\partial}{\partial z}\psi \\ \Delta\psi &= 0\,, \quad z < 0 \\ \psi &\to 0\,, \quad z \to -\infty\,. \end{aligned} \tag{55}$$

Here $\mathbf{x} = [x, y, z]^\top \in \mathcal{X}$ and the Jacobian is

$$J(\psi, b) = \left(\frac{\partial \psi}{\partial x}\right)\left(\frac{\partial b}{\partial y}\right) - \left(\frac{\partial b}{\partial x}\right)\left(\frac{\partial \psi}{\partial y}\right)\,. \tag{56}$$

The SQG equation (55) is solved by applying a Fourier transform and integrating in time with finite differences. The parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_5]^\top$ defines the initial buoyancy $b_0 : \mathbb{R}^2 \to \mathbb{R}$ at the surface $z = 0$ as well as the flow. The initial buoyancy is given by the Gaussian

$$b_0(x, y; \boldsymbol{\theta}) = -\frac{1}{(2\pi/|\theta_5|)^2} \exp\left(-x^2 - \exp(2\theta_1) y^2\right)\,, \tag{57}$$

where θ_1 is the log aspect ratio that determines the shape of the ellipse, see Figure 12, and θ_5 controls the amplitude. Of the remaining parameters θ_2 is the gradient Coriolis parameter, θ_3 is the log buoyancy frequency, and θ_4 is background zonal flow along the x-axis.

The high-fidelity random variable consists of nine equally-spaced observations of the solution $b(\mathbf{x}, T; \boldsymbol{\theta})$ with $T = 24$, where $b(\cdot, \cdot\,; \boldsymbol{\theta})$ is computed numerically using a time step of size $\Delta t = 0.005$ and 256 grid points along each coordinate axis. The surrogate random variable consists of observations of the solution at the same locations and time except with $b(\cdot, \cdot\,; \boldsymbol{\theta})$ computed numerically using only 64 grid points along each coordinate axis. The costs of each sampling each random variable are taken to be proportional to the total number of spatial grid points used in the solvers, $c_0 = 256^2 = 65,536$ and $c_1 = 64^2 = 4,096$. The observation points are shown in plot (c) and (d) in Figure 12.

The input parameters $\boldsymbol{\theta}$ are drawn from a Gaussian mixture model where $i \in \{0, 1\}$ is a latent variable determining the mixture component sampled from. The full distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta} \mid i) = N(\boldsymbol{\mu}_i, \mathbf{C})\,, \quad i \sim \text{Bernoulli}(1/2)\,,$$

where

$$\boldsymbol{\mu}_0 = [1, 0, 0, 0, 4]^\top\,, \quad \boldsymbol{\mu}_1 = [0.1, 0, 0, 0, 4]^\top\,,$$

and

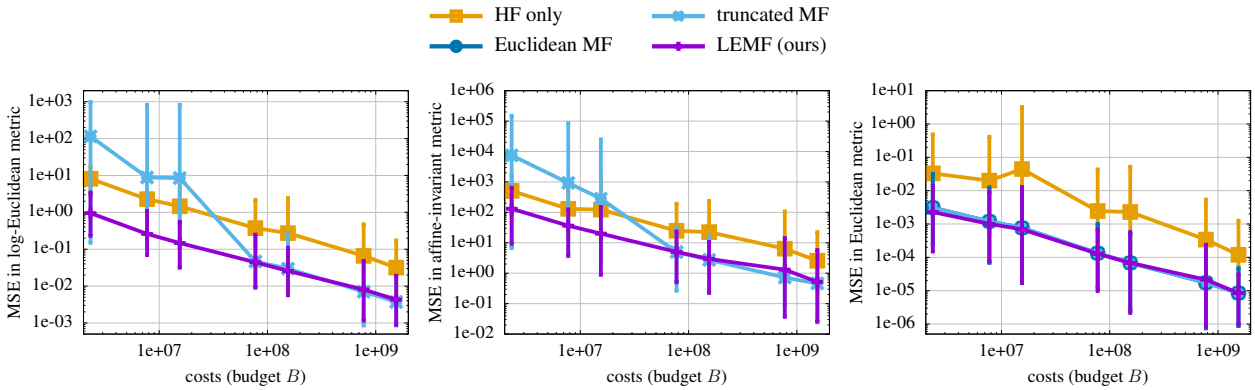
$$\mathbf{C} = \begin{bmatrix} 0.3^2 & & & & \\ & 0.003^2 & & & \\ & & 0 & & \\ & & & 0.08^2 & \\ & & & & 0.3^2 \end{bmatrix}.$$

Note that we fix the third parameter $\theta_3 = 0$, but the covariance matrix of the output observations $\mathbf{y}^{(0)}$ and $\mathbf{y}^{(1)}$ will still be positive definite.

The first class $i = 0$ corresponds to an initial buoyancy with a larger log-aspect ratio, which introduces an instability, while the second class $i = 1$ corresponds to a smaller log-aspect ratio. Figure 12 shows the buoyancy at initial and final time for stochastic inputs corresponding to both classes. We sample the class variable i from a Bernoulli distribution with parameter $1/2$, i.e., $i = 0$ and $i = 1$ are drawn with equal probability.

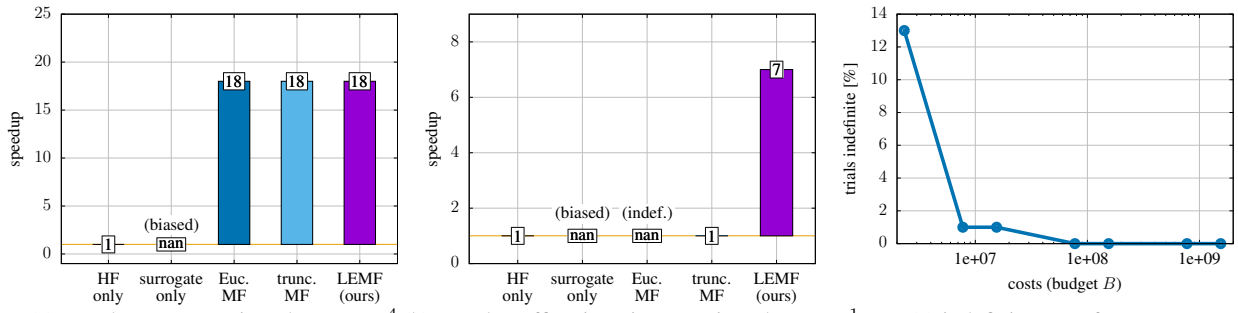
We define the mean relative error of distances $\text{MRE}(\mathbf{A})$ for the metric $\mathbf{A}$ with respect to the reference metric $\mathbf{A}_0$ as

$$\text{MRE}(\mathbf{A}) = \frac{1}{5000} \sum_{i=1}^{5000} \frac{|d_{\mathbf{A}}(\mathbf{y}_i, \mathbf{0}) - d_{\mathbf{A}_0}(\mathbf{y}_i, \mathbf{0})|}{d_{\mathbf{A}_0}(\mathbf{y}, \mathbf{0})}\,.$$



(a) log-Euclidean metric (b) affine-invariant metric (c) Euclidean metric

Figure 9. Supplemental for heat-flow problem: Plots analogous to Figure 2, except that the minimum and maximum MSE over all 100 trials are shown as error bars. Note that the surrogate-only estimator is not shown to ease exposition. The Euclidean multi-fidelity estimator is shown only in plot (c) because it becomes indefinite and therefore the MSE cannot be computed in the log-Euclidean and the affine-invariant metric.



(a) speedup Euc. metric, tol. 2×10^{-4} (b) speedup affine-invariant metric, tol. 3×10^{1} (c) indefiniteness of Euc. MF

Figure 10. Supplemental for heat flow problem: Plots (a) and (b) show the speedups of our LEMF estimator compared to the single-fidelity estimator that uses the high-fidelity samples alone. Note that the Euclidean estimator is indefinite for more than 10% of the trials and thus the speedup with respect to the affine-invariant metric cannot be computed. Plot (c) shows the percentage of the trails of the Euclidean multi-fidelity estimator that lead to an indefinite covariance matrix estimate.

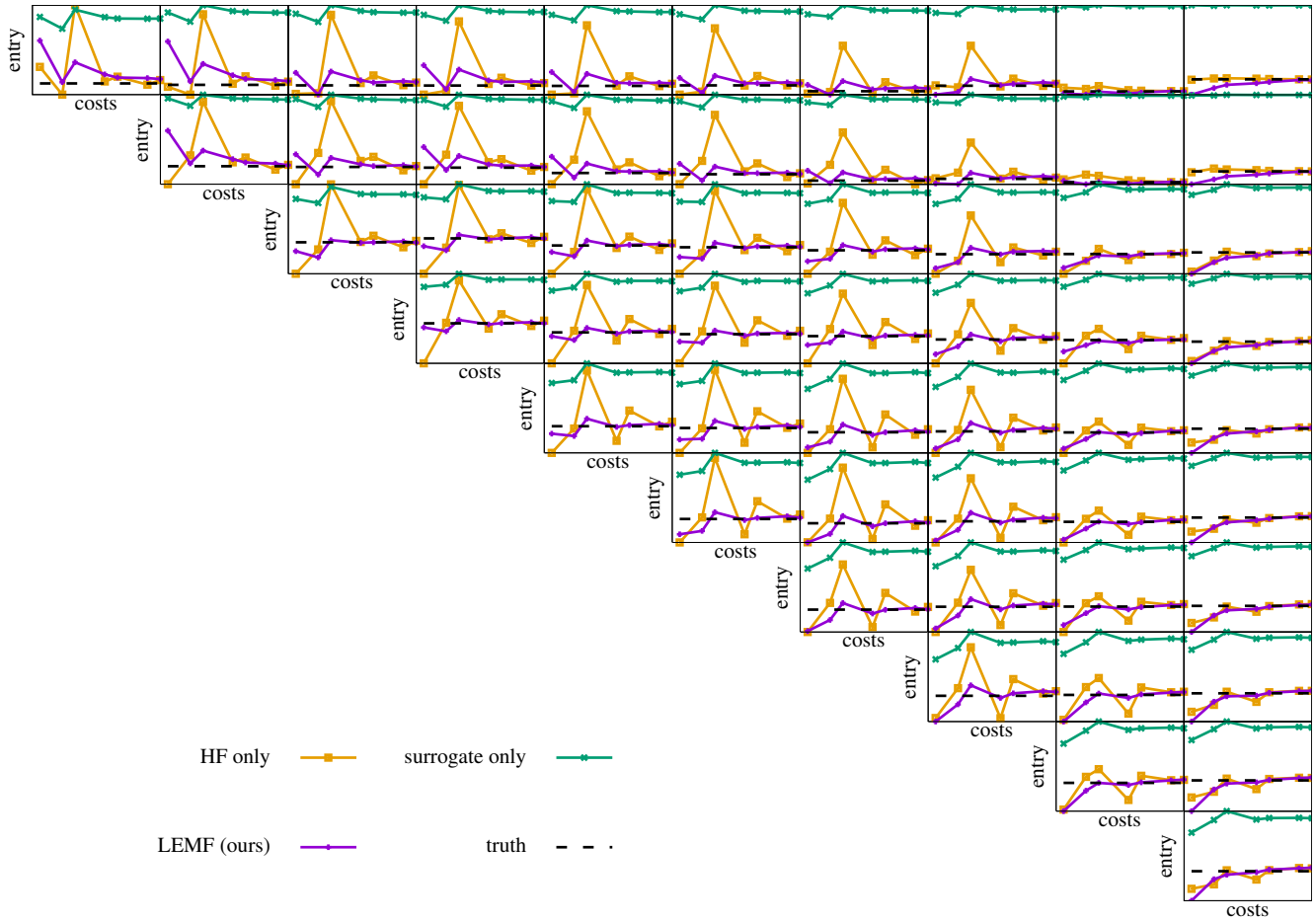
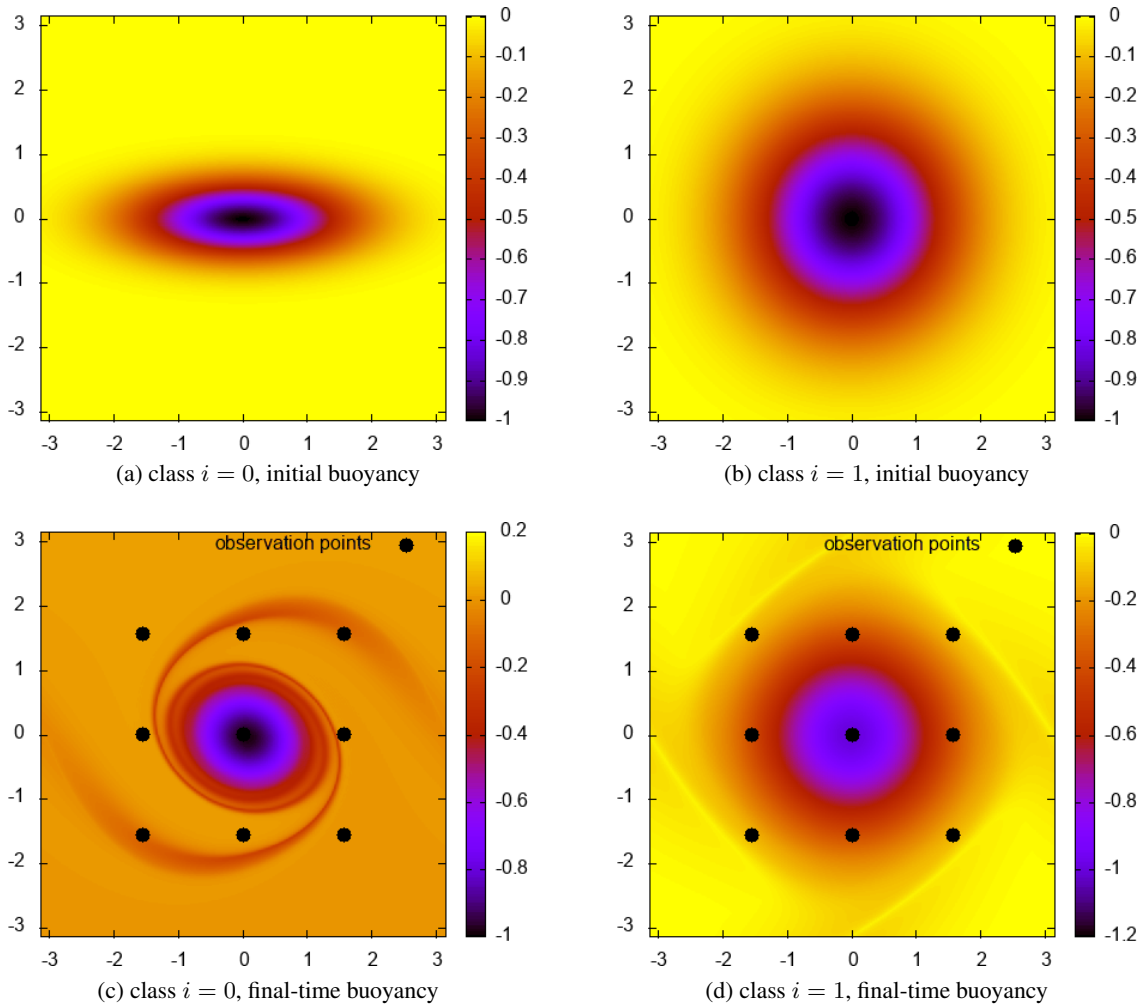


Figure 11. Supplemental for heat-flow problem: The plot shows the convergence behavior of the estimators. Our LEMF estimator is closer to the true value (dashed black) than the single-fidelity estimator, especially for low costs and for diagonal entries of the covariance matrix.



(a) class $i = 0$, initial buoyancy

(b) class $i = 1$, initial buoyancy

(c) class $i = 0$, final-time buoyancy

(d) class $i = 1$, final-time buoyancy

Figure 12. Supplemental for metric learning: Plots show examples of the buoyancy at initial (top) and final time (bottom) for $\boldsymbol{\theta}$ sampled from class $i = 0$ (left) and $i = 1$ (right). Observations consist of solution values at nine spatial locations in the domain, as depicted in plots (c) and (d). We use the observations to estimate a metric which will distinguish between observations corresponding to $\boldsymbol{\theta}$ sampled from class $i = 0$ and $\boldsymbol{\theta}$ sampled from class $i = 1$.