MULTIFIDELITY DIMENSION REDUCTION VIA ACTIVE SUBSPACES[*]

REMI R. LAM[†], OLIVIER ZAHM[‡], YOUSSEF M. MARZOUK[†], AND
KAREN E. WILLCOX[§]

Abstract. We propose a multifidelity dimension reduction method to identify a low-dimensional structure present in many engineering models. The structure of interest arises when functions vary primarily on a low-dimensional subspace of the high-dimensional input space, while varying little along the complementary directions. Our approach builds on the gradient-based methodology of active subspaces, and exploits models of different fidelities to reduce the cost of performing dimension reduction through the computation of the active subspace matrix. We provide a nonasymptotic analysis of the number of gradient evaluations sufficient to achieve a prescribed error in the active subspace matrix, both in expectation and with high probability. We show that the sample complexity depends on a notion of intrinsic dimension of the problem, which can be much smaller than the dimension of the input space. We illustrate the benefits of such a multifidelity dimension reduction approach using numerical experiments with input spaces of up to two thousand dimensions.

Key words. dimension reduction, multifidelity, gradient-based, active subspace, intrinsic dimension, effective rank, matrix Bernstein inequality, control variate

AMS subject classifications. 15A18, 15A60, 41A30, 41A63, 65D15, 65N30

DOI. 10.1137/18M1214123

1. Introduction. Engineering models are typically parameterized by a large number of input variables, and can also be expensive to evaluate. Yet these models are often embedded in problems of global optimization or uncertainty quantification, whose computational cost and complexity increase dramatically with the number of model inputs. One strategy to circumvent this *curse of dimensionality* is to exploit, when present, some notion of low-dimensional structure and to perform *dimension reduction*. Doing so can significantly reduce the complexity of the problem at hand. In this paper, we consider the problem of identifying the low-dimensional structure that arises when an output of a model varies primarily on a low-dimensional subspace of the input space, while varying little along the complementary directions. This structure is commonly found in engineering problems and can be identified using the *active subspace* method [7, 39], among other methods. The active subspace method relies on the computation of a second moment matrix, a step that can be costly as it often involves many evaluations of the gradient of the model. In this work, we consider the common engineering setting where cheap low-fidelity approximations of an expensive high-fidelity model, and its gradients, are available. We propose a *multifidelity* gradient-based algorithm to reduce the cost of performing dimension

[†]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 (rlam@mit.edu, ymarz@mit.edu).

[‡]Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France (olivier.zahm@inria.fr).

[§]Oden Institute for Computational Engineering and Sciences, UT Austin, Austin, TX 78712 (kwillcox@oden.utexas.edu).

reduction via active subspaces. In particular, we present a multifidelity estimator of the second moment matrix used by the active subspace method and show, theoretically and empirically, that fewer evaluations of the expensive gradient are sufficient to perform dimension reduction.

Several approaches have been devised to identify a low-dimensional structure in the input space of a function. These methods include global sensitivity analysis [41], sliced inverse regression [27], basis adaptation [46], and low-rank matrix recovery [48]. Recent work has also explored combining dimension reduction in both the input and the state space of the associated model [2, 12, 16, 28, 40, 45]. Such methods typically require a large number of (potentially expensive) function evaluations. When derivative information is available (e.g., via adjoint methods or automatic differentiation), gradient-based methods have also been proposed to detect the low-dimensional structure of a smooth function with higher sample efficiency [42]. One way to leverage derivative information is to examine the spectral properties of the second moment matrix of the gradient of the function. The dominant eigenspace of that matrix contains the directions along which the function, loosely speaking, varies the most. This dominant eigenspace is called the *active subspace* [7, 9, 39]. More precisely, in [50], the second moment matrix is used to construct an upper bound for the function approximation error induced by dimension reduction. The active subspace's dimension is then chosen in order to satisfy a user-defined tolerance, allowing a rigorous control of the approximation error. Gradient-based methods have been successfully used to detect and exploit low-dimensional structure in engineering models [8, 18, 19, 29] as well as in Bayesian inverse problems [10, 11, 51]. The efficiency of these gradient-based methods depends upon the computation of the second moment matrix of the gradient. This can be an expensive step as it involves computing an integral, over the high-dimensional input space, of the gradient of an expensive function. Reducing the cost of the dimension reduction step is particularly important as it allows more computational resources to be allocated to the original task of interest (e.g., optimization or uncertainty quantification).

To reduce this computational cost, one strategy consists of replacing the expensive gradient with a cheap-to-evaluate approximation or surrogate. Surrogates with lower evaluation cost are widely available in engineering problems: they include models defined by numerically solving equations on coarser meshes, using simplified governing equations, imposing looser convergence criteria, or employing reduced-order models. In order to control the error induced by the use of a surrogate, *multifidelity methods* aim at combining cheap approximations with expensive but accurate information in an optimal way (see [35] for a survey). The goal of such approaches is to shift most of the work to the cheaper model, while querying the expensive model often enough to guarantee convergence to the desired quantity (in this case, the second moment matrix of the gradient). For instance, multigrid methods use a hierarchy of cheaper and coarser discretizations to solve systems of partial differential equations more efficiently [4, 5, 15]. In multilevel Monte Carlo, expected quantities and rare event probabilities are computed by distributing the computational work among several levels of approximation with known error rate and cost [3, 14, 22, 44, 49]. When no such information about error rates is available, or when there is no hierarchy among models, multifidelity techniques have been employed to accelerate Monte Carlo estimates [36] by solving an optimal resource allocation problem among a collection of models with varying fidelity. Multifidelity techniques have also been devised to accelerate optimization [1, 13, 20, 24, 30, 37, 43], global sensitivity analysis [38], or importance sampling and rare event estimation [25, 26, 33, 34]. While most multi-

fidelity techniques have focused on estimating the expectations of scalar quantities, high-dimensional objects such as the second moment matrix in the active subspace method—effectively, the expectation of a matrix-valued function—have received less attention. Because high-dimensional objects are typically more challenging to approximate, developing and analyzing multifidelity algorithms for their estimation could lead to significant computational savings.

In this paper, we use multifidelity techniques to reduce the computational cost of performing dimension reduction. We build on the gradient-based active subspace method, proposing a multifidelity estimator for the second moment matrix that uses the low-fidelity model as a control variate for the outputs of the high-fidelity model—thus providing variance reduction and reducing computational costs. We establish nonasymptotic error bounds for this estimator, both in expectation and in high probability. We show that the sample complexity depends on the intrinsic dimension of the second moment matrix, a quantity that can be much smaller than the dimension of the input space when the function of interest varies mostly along a few directions. Finally, we demonstrate the performance of our proposed multifidelity dimension reduction technique on several analytical and engineering examples.

The paper is organized as follows. In section 2, we give a brief review of the active subspace methodology. Then, we formalize the proposed active subspace multifidelity algorithm in section 3. Error bounds for the single-fidelity and multifidelity active subspace algorithms are provided in section 4. We illustrate the benefits of our approach with numerical examples in section 5 before summarizing our findings in section 6.

2. Active subspace. We consider a scalar-valued function $f : \mathcal{X} \to \mathbb{R}$, where the input space $\mathcal{X}$ is a subset of $\mathbb{R}^d$. We refer to the dimension $d \in \mathbb{N}$ as the *ambient dimension*. The active subspace method [7, 9] aims to compute a low-dimensional subspace of $\mathcal{X}$ in which most of the variations of f are concentrated. The active subspace method assumes that f is differentiable and that each component of ∇f is square integrable on the space $\mathcal{X}$, weighted by a user-defined probability density $\rho : \mathcal{X} \to \mathbb{R}^+$. This guarantees the well posedness of the second moment matrix

$$H = \mathbb{E}\left[\nabla f(X)\nabla f(X)^T\right],$$

where $X \sim \rho$ is a random variable taking values in $\mathcal{X}$ and $\mathbb{E}[\cdot]$ denotes the expectation. We refer to H as the active subspace matrix (AS matrix). The eigendecomposition of H yields information about the directions along which f varies. Specifically, for any unit norm vector $u \in \mathbb{R}^d$, the quantity $u^T H u = \mathbb{E}[(\nabla f(X)^T u)^2]$ corresponds to the L^2 norm of the gradient ∇f projected on $\mathrm{span}\{u\}$. Thus, the largest eigenvector of H, which is a maximizer of $u^T H u$ over unit norm vectors $u \in \mathbb{R}^d$, is aligned with the direction in which f has the largest (in squared magnitude) average derivative.

Another important property is that, under some mild assumptions on the probability density ρ, the AS matrix allows us to control the mean square error between $f(X)$ and a ridge approximation of the form of $h(U_r^T X)$, where $U_r \in \mathbb{R}^{d \times r}$ is a matrix with $r \leq d$ orthonormal columns. In particular, if h is defined to be the conditional expectation $h(U_r^T X) = \mathbb{E}[f(X)|U_r^T X]$, $\mathcal{X} = \mathbb{R}^d$, and ρ is the density of the standard normal distribution on $\mathcal{X}$, then Proposition 2.5 in [50] (with $P_r = U_r U_r^T$) guarantees that

$$(2.1) \qquad \mathbb{E}[(f(X) - h(U_r^T X))^2] \leq \mathrm{trace}(H) - \mathrm{trace}(U_r^T H U_r)$$

holds for any U_r such that $U_r^T U_r = I_r$. This result relies on Poincaré-type inequalities

and can be extended to more general densities ρ (see Corollary 2 in [51]). In order to obtain a good approximation of f in the L^2 sense, we can choose U_r as a matrix which minimizes the right-hand side of (2.1). This is equivalent to the problem

$$(2.2) \qquad \max_{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}} \text{trace}(U_r^T H U_r).$$

Any matrix U_r whose columns span the r-dimensional dominant eigenspace of H is a solution. The corresponding subspace is called the *active subspace*.

In practice, there is no closed-form expression for the AS matrix and H must be approximated numerically. The following Monte Carlo estimator requires evaluating ∇f at m_1 realizations of the input parameters, drawn independently from ρ. We refer to this estimator as a single-fidelity estimator (SF estimator).

DEFINITION 2.1 (SF estimator). *Let $m_1 \geq 1$ be the number of gradient evaluations. We define the SF estimator of H to be*

$$\widehat{H}_{SF} = \frac{1}{m_1} \sum_{i=1}^{m_1} \nabla f(X_i) \nabla f(X_i)^T,$$

where $X_1, \ldots, X_{m_1}$ are independent copies of $X \sim \rho$.

Computing an estimate of H with a satisfactory error can require a large number m_1 of gradient evaluations. In the following section, we propose a new multifidelity algorithm that leverages a cheap-to-evaluate approximation of ∇f to reduce the cost of estimating H.

3. Multifidelity dimension reduction. In this section, we describe a multifidelity approach for estimating the AS matrix H (section 3.1). We also characterize the impact of using such an approximation of H on the quality of the dimension reduction (section 3.2).

3.1. Multifidelity active subspace estimator. Suppose we are given a function $g : \mathcal{X} \to \mathbb{R}$ that is a cheap-to-evaluate approximation of f. We assume that g is differentiable and that each component of ∇g is square integrable. From now on, we refer to f as the *high-fidelity* function and to g as the *low-fidelity* function. Based on the identity

$$H = \mathbb{E}[\nabla f(X) \nabla f(X)^T - \nabla g(X) \nabla g(X)^T] + \mathbb{E}[\nabla g(X) \nabla g(X)^T],$$

we introduce the following unbiased multifidelity estimator (MF estimator).

DEFINITION 3.1 (MF estimator). *Let $m_1 \geq 1$ and $m_2 \geq 1$ be the numbers of gradient evaluations of f and g. We define the MF estimator of H to be*

$$\widehat{H}_{MF} = \frac{1}{m_1} \sum_{i=1}^{m_1} (\nabla f(X_i) \nabla f(X_i)^T - \nabla g(X_i) \nabla g(X_i)^T) + \frac{1}{m_2} \sum_{i=m_1+1}^{m_1+m_2} \nabla g(X_i) \nabla g(X_i)^T,$$

where $X_1, \ldots, X_{m_1+m_2}$ are independent copies of $X \sim \rho$.

Remark 3.2 (indefiniteness of $\widehat{H}_{MF}$). While the quantity of interest H is symmetric positive semidefinite, the MF estimator $\widehat{H}_{MF}$ is symmetric but not necessarily positive semidefinite. It is natural to ask whether a positive semidefinite estimator is necessary to yield good dimension reduction. In the following, we show that the

quality of the dimension reduction is controlled by the error between H and $\widehat{H}_{MF}$ (Corollary 3.4) which can be reduced arbitrarily close to zero with high probability (Proposition 4.1). In particular, those results do not require positive semidefiniteness from the estimator $\widehat{H}_{MF}$.

A realization of $\widehat{H}_{MF}$ can be obtained using Algorithm 3.1. First, $m_1 + m_2$ input parameter realizations are drawn independently from ρ. Then, the high-fidelity gradients are evaluated at the first m_1 input parameter values while the low-fidelity gradients are evaluated at all $m_1 + m_2$ input parameter values.

Algorithm 3.1 Multifidelity active subspace.

Function: `multifidelity_active_subspace`(m_1, m_2)
Input: m_1 and m_2
Draw $m_1 + m_2$ independent copies $\{X_i\}_{i=1}^{m_1+m_2}$ of $X \sim \rho$
for $i = 1$ **to** m_1 **do**
    Compute $\nabla f(X_i)$ and $\nabla g(X_i)$
end for
$\widehat{H}_{MF} \leftarrow \frac{1}{m_1} \sum_{i=1}^{m_1} (\nabla f(X_i)\nabla f(X_i)^T - \nabla g(X_i)\nabla g(X_i)^T)$
for $i = 1$ **to** m_2 **do**
    Compute $\nabla g(X_{m_1+i})$
end for
$\widehat{H}_{MF} \leftarrow \widehat{H}_{MF} + \frac{1}{m_2} \sum_{i=m_1+1}^{m_1+m_2} \nabla g(X_i)\nabla g(X_i)^T$
Output: $\widehat{H}_{MF}$

The definition of the proposed MF estimator of the AS matrix uses the low-fidelity gradient to construct a control variate $\nabla g(X)\nabla g(X)^T$ for $\nabla f(X)\nabla f(X)^T$. The MF estimator is written as the sum of two terms. The first one involves m_1 evaluations of the low-fidelity and high-fidelity gradients. This is an expensive quantity to compute, so the number of samples m_1 is typically set to a low value. Note that if ∇g is a good approximation of ∇f, then the control variate $\nabla g(X)\nabla g(X)^T$ is highly correlated with $\nabla f(X)\nabla f(X)^T$ and the first term of the estimator has low variance (in a sense yet to be made precise for matrices). The low variance of $\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T$ allows for a good estimator of $\mathbb{E}[\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T]$ despite the small number of samples m_1. The second term involves m_2 evaluations of the cheap low-fidelity gradient. Thus, m_2 can usually be set to a large value, allowing for a good estimation of $\mathbb{E}[\nabla g(X)\nabla g(X)^T]$ despite the possibly large variance of $\nabla g(X)\nabla g(X)^T$. Combining the two terms, the MF estimator $\widehat{H}_{MF}$ provides a good approximation of H with few evaluations of the expensive high-fidelity gradient ∇f. In section 4, we make this statement precise by providing an analysis of the error between H and $\widehat{H}_{MF}$ as a function of the number of samples m_1 and m_2.

3.2. Relationship to function approximation. The performance of our MF estimator (or that of any estimator for H) should be analyzed with respect to the end goal of the problem which, in this paper, is to perform dimension reduction. Computing a good approximation of H is an *intermediate step* in the dimension reduction process. To further motivate the use of an MF estimator to reduce the difference between H and $\widehat{H}_{MF}$ at low cost, we show how this matrix error impacts the quality of the dimension reduction. As shown in section 2, one way of performing dimension reduction is to minimize a bound on the function approximation error (2.1). This corresponds to maximizing $U_r \mapsto \text{trace}(U_r^T H U_r)$. Replacing the unknown H by

$\widehat{H}_{MF}$, we can compute the matrix $\widehat{U}_r$ defined by

$$(3.1) \qquad \widehat{U}_r \in \underset{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}}{\text{argmax}} \ \text{trace}(U_r^T \widehat{H}_{MF} U_r),$$

and ask how does $\text{trace}(\widehat{U}_r^T H \widehat{U}_r)$ compare to the maximal value of $\text{trace}(U_r^T H U_r)$ over all $U_r \in \mathbb{R}^{d \times r}$ such that $U_r^T U_r = I_r$. By definition, we have the inequality in the following direction:

$$\max_{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}} \text{trace}(U_r^T H U_r) \geq \text{trace}(\widehat{U}_r^T H \widehat{U}_r).$$

The next proposition shows that the difference between the two terms of the previous inequality can be controlled by means of the error $\|H - \widehat{H}_{MF}\|$, where $\|\cdot\|$ denotes the matrix operator norm. Note that the proof is not restricted to the MF estimator: the same result holds for any symmetric estimator of H.

PROPOSITION 3.3. *Let $\widehat{H}$ be a symmetric estimator of H and*

$$(3.2) \qquad \widetilde{U}_r \in \underset{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}}{\text{argmax}} \ \text{trace}(U_r^T \widehat{H} U_r),$$

then

$$(3.3) \qquad \text{trace}(\widetilde{U}_r^T H \widetilde{U}_r) \geq \max_{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}} \text{trace}(U_r^T H U_r) - 2r\|H - \widehat{H}\|.$$

Proof. Consider the eigenvalue decomposition of $H - \widehat{H} = V \Sigma V^T$, where $\Sigma = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$ is a diagonal matrix containing the eigenvalues of $H - \widehat{H}$ and $V \in \mathbb{R}^{d \times d}$ is a unitary matrix. For any matrix $U_r \in \mathbb{R}^{d \times r}$ such that $U_r^T U_r = I_r$, we have

$$\begin{aligned} |\text{trace}(U_r^T H U_r) - \text{trace}(U_r^T \widehat{H} U_r)| &= |\text{trace}(U_r^T (H - \widehat{H}) U_r)| \\ &= |\text{trace}(\Sigma V^T U_r U_r^T V)| \\ &\leq \max\{|\lambda_1|, \ldots, |\lambda_d|\} \ |\text{trace}(V^T U_r U_r^T V)| \\ &= \|H - \widehat{H}\| \ |\text{trace}(U_r U_r^T)| \\ (3.4) \qquad &= r\|H - \widehat{H}\|. \end{aligned}$$

Letting $U_r = \widetilde{U}_r$ in the above relation yields

$$\begin{aligned} \text{trace}(\widetilde{U}_r^T H \widetilde{U}_r) &\overset{(3.4)}{\geq} \text{trace}(\widetilde{U}_r^T \widehat{H} \widetilde{U}_r) - r\|H - \widehat{H}\| \\ &\overset{(3.2)}{\geq} \text{trace}(U_r^T \widehat{H} U_r) - r\|H - \widehat{H}\| \\ &\overset{(3.4)}{\geq} \text{trace}(U_r^T H U_r) - 2r\|H - \widehat{H}\|. \end{aligned}$$

Maximizing over $U_r \in \mathbb{R}^{d \times r}$ with $U_r^T U_r = I_r$, yields (3.3) and concludes the proof. $\square$

COROLLARY 3.4. *Let $\widehat{H}_{MF}$ be an MF estimator of H and $\widehat{U}_r$ be defined by (3.1). Then*

$$(3.5) \qquad \text{trace}(\widehat{U}_r^T H \widehat{U}_r) \geq \max_{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}} \text{trace}(U_r^T H U_r) - 2r\|H - \widehat{H}_{MF}\|.$$

Proof. This follows from applying Proposition 3.3 to $\widehat{H} = \widehat{H}_{MF}$ and $\widetilde{U}_r = \widehat{U}_r$. $\square$

We now establish the connection between the result of Corollary 3.4 and the quality of the dimension reduction. Assume that inequality (2.1) holds true for any $U_r^T U_r = I_d$ (this is in particular the case if $X \sim \mathcal{N}(0, I_d)$). Replacing U_r by $\widehat{U}_r$ in (2.1) and using Corollary 3.4, we can write

$$(3.6)$$
$$\mathbb{E}[(f(X) - h(\widehat{U}_r^T X))^2] \leq \text{trace}(H) - \max_{\substack{U_r \in \mathbb{R}^{d \times r} \\ \text{s.t. } U_r^T U_r = I_r}} \text{trace}(U_r^T H U_r) + 2r\|H - \widehat{H}_{MF}\|$$

$$(3.7) \qquad\qquad = (\lambda_{r+1} + \cdots + \lambda_d) + 2r\|H - \widehat{H}_{MF}\|,$$

where $h(\widehat{U}_r^T X) = \mathbb{E}[f(X)|\widehat{U}_r^T X]$. Here $\lambda_i \geq 0$ denotes the ith largest eigenvalue of H. This relation shows that a strong decay in the spectrum of H is favorable to efficient dimension reduction. Also, increasing the number r of active variables has competitive effects on the two terms in the right-hand side: the first term $(\lambda_{r+1} + \cdots + \lambda_d)$ is reduced whereas the second term $2r\|H - \widehat{H}_{MF}\|$ increases linearly in r. Given the importance of $\|H - \widehat{H}_{MF}\|$ in controlling the quality of the dimension reduction, we show in the next section how this error can be controlled at cheap cost using the proposed MF estimator.

Remark 3.5 (angle between subspaces). Another way of controlling the quality of the approximate active subspace (the span of the columns of $\widehat{U}_r$) is via the *principal angle* between the exact and the approximate active subspaces [6, 17]. This angle, denoted by $\angle(\widehat{\mathcal{S}}, \mathcal{S})$, is defined by

$$\sin(\angle(\widehat{\mathcal{S}}, \mathcal{S})) = \|\widehat{U}_r \widehat{U}_r^T - \tilde{U}_r \tilde{U}_r^T\|,$$

where $\widehat{\mathcal{S}} = \text{range}(\widehat{U}_r)$ is the approximate subspace and $\mathcal{S} = \text{range}(\tilde{U}_r)$ is the exact active subspace, $\tilde{U}_r$ being a solution to (2.2). This requires $\widehat{\mathcal{S}}$ and $\mathcal{S}$ to be uniquely defined, which might not be the case if there is a plateau in the spectra of H and $\widehat{H}_{MF}$. For instance, if the rth eigenvalue of H equals the $(r+1)$th, problem (2.2) admits infinitely many solutions and $\mathcal{S}$ is not uniquely defined. Note that in practice, r is chosen such that the spectral gap is large, by inspection of the spectrum. To our knowledge, all analyses focusing on controlling the principal angle $\angle(\widehat{\mathcal{S}}, \mathcal{S})$ rely on the spectral gap assumption $\lambda_r > \lambda_{r+1}$, where λ_r is the rth eigenvalue of H.

In contrast, the goal-oriented approach consisting of minimizing the upper bound of the functional error does not require the spectral gap assumption. This results from (2.2) and Corollary 3.4. In particular, the uniqueness of the active subspace is not required, as any solution to (2.2) yields an equally good active subspace for the purpose of function approximation. Therefore, in this paper, we do not further consider the principal angle $\angle(\widehat{\mathcal{S}}, \mathcal{S})$. Instead we focus on comparing $\text{trace}(\widehat{U}_r^T H \widehat{U}_r)$ to $\text{trace}(\tilde{U}_r^T H \tilde{U}_r)$. As illustrated by Corollary 3.4, this is sufficient to control the error $\|H - \widehat{H}_{MF}\|$ between the AS matrix and its estimator.

4. A nonasymptotic analysis of the estimator. In this section, we use results from nonasymptotic random matrix theory to express the number of gradient evaluations sufficient to control the error in an estimate of H, up to a user-defined tolerance. We present our main results in this section and defer the proofs to Appendices A.1 and A.2.

In general, Monte Carlo estimation of a high-dimensional object such as a $d \times d$ matrix can require a large number of samples. If the matrix does not have some

special structure, one can expect the sample complexity to scale with the large ambient dimension d. This is costly if each sample is expensive. However, the AS matrix H enjoys some structure when the problem has low effective dimension. In particular, when most of the variation of f is concentrated in a low-dimensional subspace, we expect the number of samples required to obtain a good approximation of H to depend on the dimension of this subspace, rather than on the ambient dimension d of the input space. One case of interest occurs when f is a ridge function that *only* depends on a small number of linear combinations of input variables. This leads to a rank-deficient matrix H. In such a case, we expect the number of samples to depend on the rank of H. Another important case occurs when a (possibly full-rank) matrix H has a quickly decaying spectrum. In such a case, we expect that the number of samples should depend on a characteristic quantity of the spectrum (e.g., the sum of the eigenvalues). To make a precise statement, we use the notion of *intrinsic dimension* [47, Definition 7.1.1], also called the *effective rank* [21, Definition 1]. The intrinsic dimension of H is defined by

$$\delta_H = \frac{\text{trace}(H)}{\|H\|}.$$

The intrinsic dimension is a measure of the spectral decay of H. It is bounded by the rank of H, i.e., $1 \leq \delta_H \leq \text{rank}(H) \leq d$.

Our main result, Proposition 4.1 below, establishes how many evaluations of the gradient are sufficient to guarantee that the error $\|H - \widehat{H}_{MF}\|$ is below some user-defined tolerance. In particular, the number of gradient evaluations from the low-fidelity and high-fidelity models is shown to be function of the intrinsic dimension of H and of two coefficients θ and β, characterizing the quality of the low-fidelity model and the maximum relative magnitude of the high-fidelity gradient, respectively.

PROPOSITION 4.1. *Assume there exist positive constants $\beta < \infty$ and $\theta < \infty$ such that the relations*

$$(4.1) \qquad \qquad \|\nabla f(X)\|^2 \leq \beta^2 \, \mathbb{E}[\|\nabla f(X)\|^2],$$

$$(4.2) \qquad \|\nabla f(X) - \nabla g(X)\|^2 \leq \theta^2 \, \mathbb{E}[\|\nabla f(X)\|^2],$$

hold almost surely. Let $\widehat{H}_{MF}$ be the MF estimator introduced in Definition 3.1 and assume

$$(4.3) \qquad m_2 \geq m_1 \max \left\{ \frac{(\theta + \beta)^2 (1 + \theta)^2}{\theta^2 (2 + \theta)^2} \, ; \, \frac{(\theta + \beta)^2}{\theta(2\beta + \theta)} \right\}.$$

Then, for any $\varepsilon > 0$, the condition

$$(4.4) \qquad m_1 \geq \varepsilon^{-2} \delta_H \, \theta \, \log(2d) \max \left\{ 4\delta_H \theta (2 + \theta)^2 \, ; \, 2/3(2\beta + \theta) \right\}$$

is sufficient to ensure

$$(4.5) \qquad \mathbb{E}[\|H - \widehat{H}_{MF}\|] \leq (\varepsilon + \varepsilon^2)\|H\|.$$

Furthermore for any $0 < \varepsilon \leq 1$ and $0 < \eta < 1$, the condition

$$(4.6) \qquad m_1 \geq \varepsilon^{-2} \delta_H \, \theta \, \log(2d/\eta) \left(4\delta_H \theta (2 + \theta)^2 + \varepsilon 4/3(2\beta + \theta) \right)$$

is sufficient to ensure

$$(4.7) \qquad \mathbb{P}\big\{ \|H - \widehat{H}_{MF}\| \leq \varepsilon \|H\| \big\} \geq 1 - \eta.$$

Proof. See Appendix A.1 for the proof.   □

Similarly, we can derive the number of high-fidelity gradient evaluations m_1 sufficient to control the SF estimator error in expectation and with high probability. This is the purpose of the following proposition (see also [23]). Note that a high-probability bound similar to (4.11) and (4.12) is established by Corollary 2.2 from [17].

PROPOSITION 4.2. *Assume there exists $\beta < \infty$ such that*

$$\|\nabla f(X)\|^2 \leq \beta^2 \, \mathbb{E}[\|\nabla f(X)\|^2] \tag{4.8}$$

holds almost surely. Then for any $0 < \varepsilon \leq 1$ the condition

$$m_1 \geq C\varepsilon^{-2}\delta_H \log(1 + 2\delta_H)(1 + \beta^2) \tag{4.9}$$

is sufficient to ensure

$$\mathbb{E}[\|H - \widehat{H}_{SF}\|] \leq (\varepsilon + \varepsilon^2)\|H\|. \tag{4.10}$$

Here C is an absolute (numerical) constant. Furthermore for any $0 < \varepsilon \leq 1$ and any $0 < \eta \leq 1$, the condition

$$m_1 \geq 2\varepsilon^{-2}\delta_H \log(8\delta_H/\eta)(\beta^2 + \varepsilon(1 + \beta^2)/3) \tag{4.11}$$

is sufficient to ensure

$$\mathbb{P}\{\|H - \widehat{H}_{SF}\| \leq \varepsilon\|H\|\} \geq 1 - \eta. \tag{4.12}$$

Proof. See Appendix A.2 for the proof.   □

The previous propositions show that when θ^2 is small (i.e., when ∇g is a good approximation of ∇f), the number of samples m_1 of the high-fidelity function can be significantly reduced compared to the single-fidelity approach. The number of samples m_2 has to be adjusted according to (4.3). The guarantees provided for the MF estimator are different than those for the SF estimator. For the MF estimator, the bound on m_1 is a function of the intrinsic dimension but is also weakly (i.e., logarithmically) dependent on the ambient dimension d.

These results are especially interesting when β^2 has no dependency (or weak dependency) on the ambient dimension d. In such a case, the number of evaluations sufficient to obtain a satisfactory relative error depends only on the intrinsic dimension δ_H and the parameter β^2, and does not depend (or only weakly depends) on the ambient dimension d. Recall that β^2 quantifies the variation of the square norm of the gradient, $\|\nabla f\|^2$, relative to its mean $\mathbb{E}[\|\nabla f(X)\|^2]$. In section 5, we provide an example of gradient function ∇f for which β^2 is independent of d.

The proofs of Propositions 4.1 and 4.2 use a similar strategy. The key ingredient is the use of a concentration inequality to bound the error between the AS matrix and its estimator. These inequalities are applicable to matrices expressed as the sum of independent matrices (i.e., independent summands), a condition met by the MF and SF estimators of H. The error bounds depend on two characteristics of the matrix of interest: the variance of the estimator and an upper bound on the norm of the summands. Those two characteristic quantities are functions of the number of samples used to construct the estimator. Once established, those bounds are used to express sufficient conditions on the number of samples to guarantee a user-defined tolerance for the error, both in expectation and with high probability. The full proofs of Propositions 4.1 and 4.2 are given in Appendix A.1 and Appendix A.2.

We conclude this section by summarizing the connection between our main results and a quantity of interest in dimension reduction: the functional error. As shown in (2.1), for any matrix $U_r \in \mathbb{R}^{d \times r}$ with $r \leq d$ orthonormal columns, the functional error $\mathbb{E}[(f(X) - h(U_r^T X))^2]$ is upper bounded by $\text{trace}(H) - \text{trace}(U_r^T H U_r)$. Thus, the quality of the dimension reduction can be controlled by finding the maximizer U_r^* of $U_r \mapsto \text{trace}(U_r^T H U_r)$, as U_r^* yields the tightest upper bound on the functional error. However, U_r^* cannot be computed because H is unknown. Instead, we compute an approximator $\widehat{H}_{MF}$ of H and its associated $\widehat{U}_r$ (see (3.1)). Corollary 3.4 shows that using $\widehat{U}_r$ instead of U_r^* yields an upper bound close to the tightest one when $\|H - \widehat{H}_{MF}\|$ is small. As a result, the functional error $\mathbb{E}[(f(X) - h(\widehat{U}_r^T X))^2]$ incurred by the ridge approximation built with $\widehat{U}_r$ is at most $2r\|H - \widehat{H}_{MF}\|$ larger than the tightest bound defined by the unknown U_r^* (see (3.6)). Finally, Proposition 4.1 shows how $\|H - \widehat{H}_{MF}\|$ can be controlled by increasing the number of gradient evaluations. Therefore, those results establish a direct link between the number of gradient evaluations and the upper bound on the functional error of the ridge function built with $\widehat{U}_r$.

5. Numerical results. In this section, we conduct numerical experiments illustrating the performance of the proposed MF estimator. We first demonstrate the algorithm on synthetic examples for which we can compute errors; we conduct a parametric study and compare the SF and MF estimator performances. Second, we consider a high-dimensional engineering case: performing dimension reduction on a linear elasticity problem involving parameterized material properties of a wrench. Finally, we consider an expensive engineering problem: finding the active subspaces associated with the shape optimization of the ONERA M6 wing in a turbulent flow.

5.1. Analytical problem. In this section, we consider an example for which all characteristic quantities (δ_H, $\|H\|$, $\mathbb{E}[\|\nabla f(X)\|^2]$, β^2) are known with closed-form expressions.[1]

5.1.1. Problem description. We consider the input space $\mathcal{X} = [-1, 1]^d$ and define ρ to be the uniform distribution over $\mathcal{X}$. We introduce the function $f : \mathcal{X} \to \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathcal{X}$

$$f(\boldsymbol{x}) = \frac{\sqrt{3}}{2} \sum_{i=1}^d a_i\, x_i^2 \quad \text{and} \quad \nabla f(\boldsymbol{x}) = \sqrt{3} \begin{pmatrix} a_1 x_1 \\ \vdots \\ a_d x_d \end{pmatrix},$$

where $\boldsymbol{a} = (a_1, \ldots, a_d)^T \in \mathbb{R}^d$ is a user-defined vector with $|a_1| \geq |a_2| \geq \cdots \geq 0$. With $X \sim \rho$, we have

$$H = \mathbb{E}[\nabla f(X) \nabla f(X)^T] = \begin{pmatrix} a_1^2 & & 0 \\ & \ddots & \\ 0 & & a_d^2 \end{pmatrix}$$

so that $\|H\| = a_1^2$, $\text{trace}(H) = \|\boldsymbol{a}\|^2$, and

$$\delta_H = \frac{\text{trace}(H)}{\|H\|} = \frac{\|\boldsymbol{a}\|^2}{a_1^2}.$$

---

[1]Code for the analytical problems available at https://github.com/remilam/MultifidelityDimensionReduction.

We define β as the smallest parameter that satisfies $\|\nabla f(X)\|^2 \leq \beta^2 \, \mathbb{E}[\|\nabla f(X)\|^2]$. Given that $\mathbb{E}[\|\nabla f(X)\|^2] = \text{trace}(H) = \|\boldsymbol{a}\|^2$, we have

$$\beta^2 = \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{\|\nabla f(\boldsymbol{x})\|^2}{\mathbb{E}[\|\nabla f(X)\|^2]} = \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{3 \sum_{i=1}^{d} (a_i x_i)^2}{\sum_{i=1}^{d} a_i^2} = 3,$$

where the supremum is attained by maximizing each term in the numerator. This corresponds to $x_i^2 = 1$ for all $1 \leq i \leq d$. Notice that $\beta = \sqrt{3}$ is independent of the ambient dimension d and the user-defined vector $\boldsymbol{a}$.

We define the low-fidelity function $g : \mathcal{X} \to \mathbb{R}$ such that, for all $\boldsymbol{x} \in \mathcal{X}$, $g(\boldsymbol{x}) = f(\boldsymbol{x}) - bT\|\boldsymbol{a}\| \cos(x_d/T)$, where $b \geq 0$ and $T > 0$ are two user-defined parameters. In other words, g is a perturbation of f such that $g - f$ depends only on the last component. We have

$$\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b\|\boldsymbol{a}\| \sin(x_d/T) \end{pmatrix}$$

for all $\boldsymbol{x} \in \mathcal{X}$. We let θ be the smallest parameter such that $\|\nabla f(\boldsymbol{x}) - \nabla g(\boldsymbol{x})\|^2 \leq \theta^2 \mathbb{E}[\|\nabla f(X)\|^2]$ for all $\boldsymbol{x} \in \mathcal{X}$. We obtain

$$\theta^2 = \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{\|\nabla f(\boldsymbol{x}) - \nabla g(\boldsymbol{x})\|^2}{\mathbb{E}[\|\nabla f(X)\|^2]} = \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{(b\|\boldsymbol{a}\| \sin(x_d/T))^2}{\|\boldsymbol{a}\|^2} = \begin{cases} b^2 \sin(1/T)^2 & \text{if } T \geq \frac{2}{\pi}, \\ b^2 & \text{otherwise.} \end{cases}$$

For the numerical experiments, we set $b = \sqrt{0.05}$ and $T = 0.1$, leading to a parameter $\theta = \sqrt{0.05}$. The number of samples m_2 is set using the criteria of (4.3), leading to

$$m_2 = 63m_1 \ \geq \ m_1 \max \left\{ \frac{(\theta + \beta)^2 (1 + \theta)^2}{\theta^2 (2 + \theta)^2} \ ; \ \frac{(\theta + \beta)^2}{\theta(2\beta + \theta)} \right\}.$$

In the two following subsections, we consider the case where H is rank deficient, and the case where H is full rank but has a small intrinsic dimension. From now on, the ambient dimension is set to $d = 100$.

5.1.2. Rank-deficient matrices. In this section, we consider the parameter $\boldsymbol{a} \in \mathbb{R}^d$ defined by

$$\boldsymbol{a} = (\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{d-k})$$

for some $k \in \{1, 3, 10, 30, 100\}$. With this choice, the function f only depends on the k first variables of the input $\boldsymbol{x}$. This yields

$$H = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \delta_H = k \in \{1, 3, 10, 30, 100\},$$

where I_k is the identity matrix of size k. We study the dependence of the relative error $\|\hat{H} - H\|/\|H\|$ on the intrinsic dimension δ_H and the number of samples m_1.

Figure 1 shows the average relative error of the SF estimator (left panel) and the MF estimator (right panel) as a function of the number of samples m_1 for $\delta_H \in \{1, 10, 100\}$. The numerical results are in accordance with the theoretical bounds: the errors are dominated by the theoretical bounds and the slopes are similar. Comparing
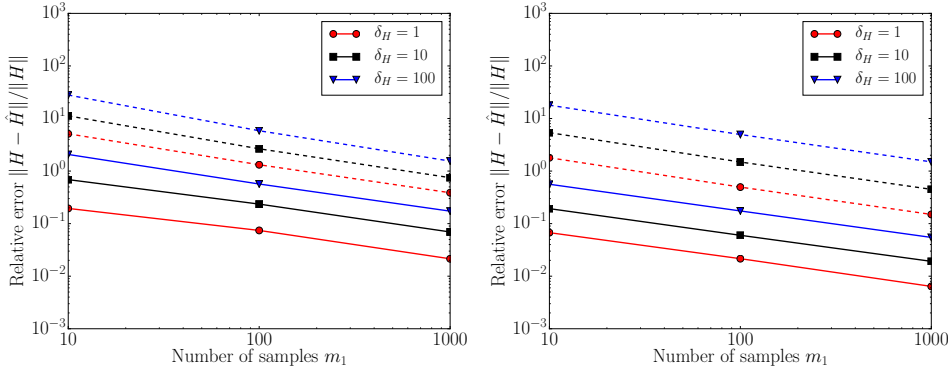
FIG. 1. *Rank-deficient example. Average relative error (solid line) as a function of the number of samples m_1 for the SF estimator (left panel) and the MF estimator (right panel). Average is computed over 100 trials. Dashed lines represent theoretical bounds.*
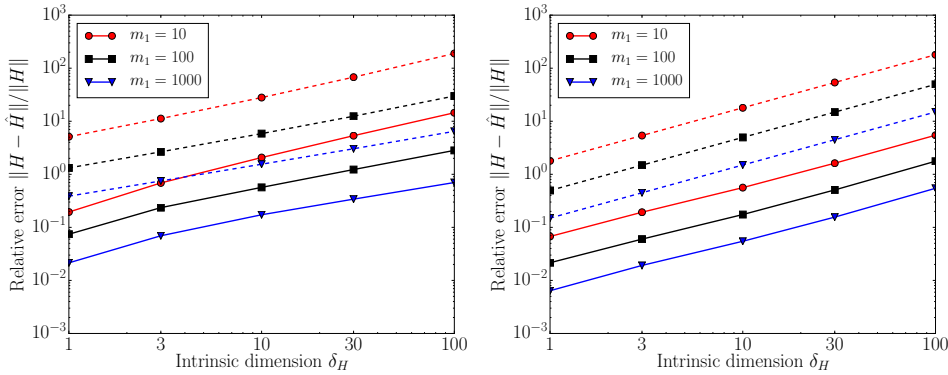

FIG. 2. *Rank-deficient example. Average relative error (solid line) as a function of the intrinsic dimension δ_H for the SF estimator (left panel) and the MF estimator (right panel). Average is computed over 100 trials. Dashed lines represent theoretical bounds.*

the left and right panels, we conclude that the MF estimator outperforms the SF estimator for a given number of high-fidelity evaluations. For example, in the case $\delta_H = 1$, with $m_1 = 10$ high-fidelity samples, the MF estimator achieves a relative error of 0.07 while the SF relative error is 0.23, i.e., a relative error 3.47 times lower. Similarly, Figure 2 shows the average relative error as a function of the intrinsic dimension of H for $m_1 \in \{10, 100, 1000\}$. We notice that the difference between the theoretical bound and the actual estimator error is larger for the MF estimator than for the SF estimator.

5.1.3. Full-rank matrices. In this section, we define the parameter $\boldsymbol{a} \in \mathbb{R}^d$ to be $a_i = \exp(-Ci)$ for all $1 \leq i \leq d$, where the value of $C \geq 0$ is set to match the intrinsic dimensions $\delta_H \in \{2, 5, 10, 50, 100\}$. The function f now depends on every component of the input $\boldsymbol{x}$ and the matrix H is full rank. Again, we study the dependence of the relative error $\|\widehat{H} - H\|/\|H\|$ on the intrinsic dimension δ_H and the number of samples m_1.

Figure 3 shows the average relative error of the SF estimator (left panel) and the MF estimator (right panel) as a function of the number of samples m_1 for $\delta_H \in \{2, 10, 100\}$. The numerical results agree with the theoretical bounds: the errors are
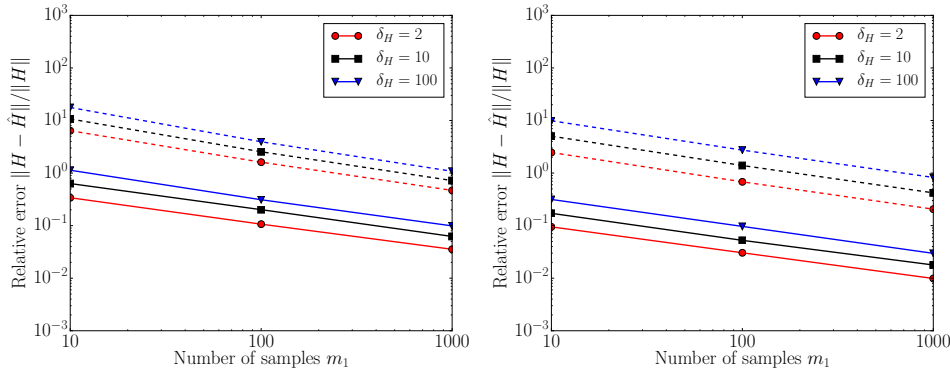
FIG. 3. *Full-rank example. Average relative error (solid line) as a function of the number of samples m_1 for the SF estimator (left panel) and the MF estimator (right panel). Average computed over 100 trials. Dashed lines represent theoretical bounds.*
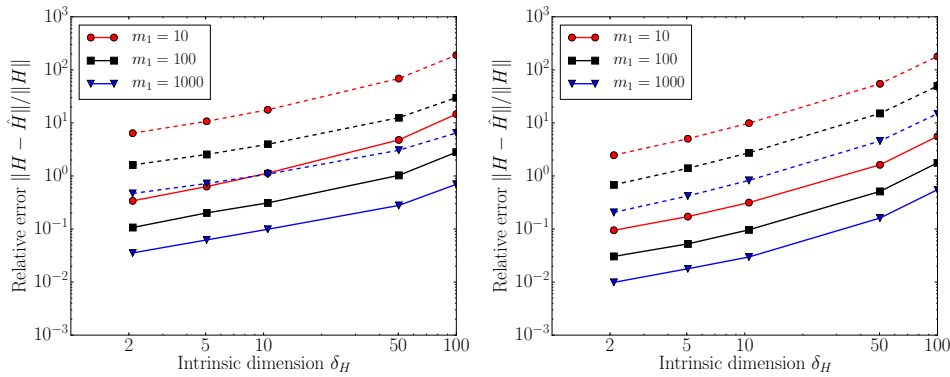

FIG. 4. *Full-rank example. Average relative error (solid line) as a function of the intrinsic dimension δ_H for the SF estimator (left panel) and the MF estimator (right panel). Average computed over 100 trials. Dashed lines represent theoretical bounds.*

dominated by the theoretical bounds and the slopes are similar. Comparing the left and right panels, we conclude that the MF estimator outperforms the SF estimator for a given number of high-fidelity evaluations. For example, in the case $\delta_H = 1$, with $m_1 = 10$ high-fidelity samples, the MF estimator achieves a relative error of 0.09 while the SF relative error is 0.37, leading to a relative error 3.86 times lower. Similarly, Figure 4 shows the average relative error as a function of the intrinsic dimension of H for $m_1 \in \{10, 100, 1000\}$. Again, we observe that the empirical results satisfy the theoretical bounds.

5.2. Linear elasticity analysis of a wrench. We now demonstrate the proposed MF estimator on a high-dimensional engineering example.

5.2.1. Problem description. We consider a two-dimensional linear elasticity problem which consists of finding a smooth displacement field $u : \Omega \to \mathbb{R}^2$ that satisfies

$$\mathrm{div}(K : \varepsilon(u)) = 0 \quad \text{on } \Omega \subset \mathbb{R}^2,$$

where $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T)$ is the strain field and K is the Hooke tensor. The boundary conditions are depicted in Figure 5(a). Using the plane stress assumption,
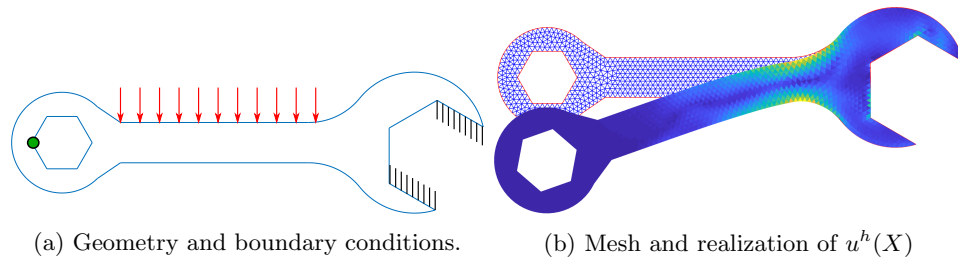
(a) Geometry and boundary conditions.   (b) Mesh and realization of $u^h(X)$

FIG. 5. *Left panel: Dirichlet condition $u = 0$ (black chopped lines), vertical unitary linear forcing (red arrows), and point of interest* PoI *(green dot). Right panel: mesh and finite element solution $u^h(X)$ associated with one realization of X, the random Young's modulus. The color represents the von Mises stress.*

the Hooke tensor K satisfies

$$K : \varepsilon(u) = \frac{E}{1+\nu}\varepsilon(u) + \frac{\nu E}{1-\nu^2}\operatorname{trace}(\varepsilon(u))I_2,$$

where $\nu = 0.3$ is the Poisson's ratio and $E \geq 0$ the Young's modulus. We model E as a random field such that $\log(E) \sim \mathcal{N}(0, C)$ is a zero-mean Gaussian field over Ω with covariance function $C : \Omega \times \Omega \to \mathbb{R}$ such that $C(s,t) = \exp(-\|s-t\|_2^2/l_0)$ with $l_0 = 1$. We consider the output defined by the vertical displacement of a point of interest (PoI) represented by the green point in Figure 5(a). We denote by $u_2(\text{PoI})$ this scalar output, where the subscript "2" refers to the vertical component of u.

We denote by u^h the finite element solution defined as the Galerkin projection of u on a finite element space comprising continuous piecewise linear functions over the mesh depicted in Figure 5(b). To compute u^h, we approximate the Young's modulus $\log(E)$ by a piecewise constant field $\log(E^h)$ that has the same statistics as $\log(E)$ at the center of the elements. Denoting by $d = 2197$ the number of elements in the mesh, we define $f : \mathbb{R}^d \to \mathbb{R}$ as the map from the log Young's modulus to the quantity of interest

$$f : X = \log(E^h) \mapsto u_2^h(\text{PoI}).$$

The gradients of f are computed with the adjoint method; see, for instance, [31]. The complexity of computing the gradient increases with the number of elements in the mesh. For this reason, we introduce a low-fidelity model g that relies on the coarser mesh depicted in Figure 6(d). This coarse mesh contains 423 elements, so that the function g and its gradient can be evaluated with less computational effort compared to f. We now explain in detail how g is defined. First, the log of the Young's modulus on the coarse mesh is defined as the piecewise constant field whose value at a given (coarse) element equals the spatial mean of $X = \log(E^h)$ restricted to that coarse element; see the illustration from Figures 6(a) to 6(b). Then, we compute the coarse finite element solution and we define $g(X)$ as the vertical displacement of the PoI. Figures 6(c) and 6(d) show that the fine and coarse solutions are similar. With our implementation,[2] evaluating the gradient of the low-fidelity model ∇g is approximately 7 times faster than computing ∇f.

Since the error bound (2.1) holds when X is a standard Gaussian random vector, we employ the following change of variables. Denoting by Σ the covariance matrix

---

[2]Code available at https://gitlab.inria.fr/ozahm/wrenchmark.git.

© 2020 Remi Lam, Olivier Zahm, Youssef Marzouk, Karen Willcox

(a) Realization of $X = \log(E^h)$ on the fine mesh

(b) Projection of X on the coarse mesh

(c) Finite element solution on the fine mesh
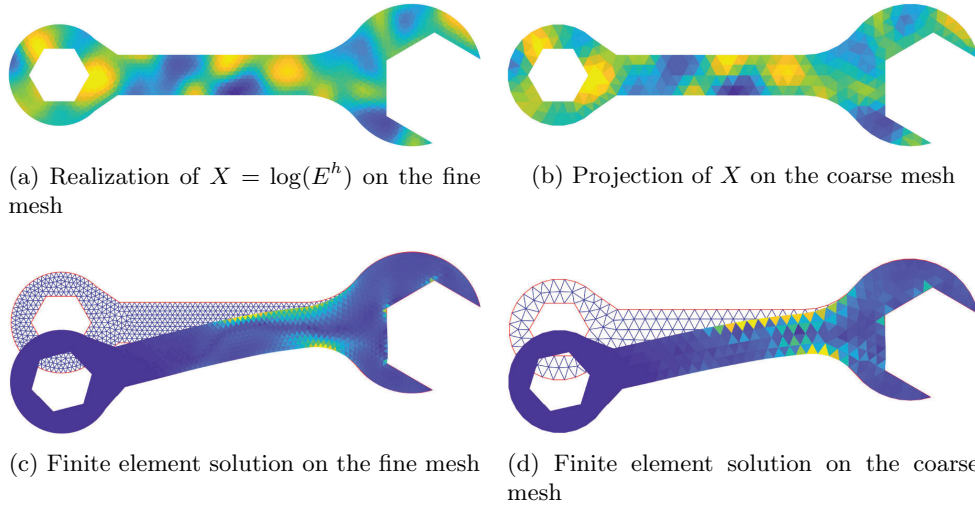
(d) Finite element solution on the coarse mesh

FIG. 6. *Construction of the low-fidelity approximation g of f using a coarse mesh. Figure 6(a) represents a realization of the Young's modulus on the fine mesh and Figure 6(b) its projection onto the coarser mesh. The corresponding finite element solution on the fine mesh (resp., coarse) and von Mises stress are represented in Figure 6(c) (resp., Figure 6(d)).*

of $X \sim \mathcal{N}(0, \Sigma)$, we write $X = \Sigma^{1/2} X_{\mathrm{std}}$, where $X_{\mathrm{std}} \sim \mathcal{N}(0, I_d)$ and where $\Sigma^{1/2}$ is a positive square root of Σ. After this change of variables, the high-fidelity and low-fidelity models are, respectively, $f_{\mathrm{std}} : X_{\mathrm{std}} \mapsto f(\Sigma^{1/2} X_{\mathrm{std}})$ and $g_{\mathrm{std}} : X_{\mathrm{std}} \mapsto g(\Sigma^{1/2} X_{\mathrm{std}})$ and the standardized AS matrix is $H_{\mathrm{std}} = \mathbb{E}[\nabla f_{\mathrm{std}}(X_{\mathrm{std}}) \nabla f_{\mathrm{std}}(X_{\mathrm{std}})^T]$. This change of variable can be interpreted as a preconditioning step of the AS matrix $H = \mathbb{E}[\nabla f(X) \nabla f(X)^T]$ since we can write $H_{\mathrm{std}} = \Sigma^{T/2} H \Sigma^{1/2}$. For the sake of simplicity, we omit the subscript "std" and we use the notations H, f, and g for the standardized version of the AS matrix, the high-fidelity model, and low-fidelity model.

5.2.2. Numerical results. We compute 10^4 evaluations of the high-fidelity gradient ∇f and use them to form an SF estimator $\widehat{H}_{SF}^{\mathrm{ref}}$ that we consider as a reference AS matrix. We compare the performance of the MF and SF estimators computed on ten cases characterized by different computational budgets. For a given case $\gamma \in \{1, \ldots, 10\}$, the SF estimator is computed with $m_1 = 3\gamma$ gradient evaluations while the MF estimator uses $m_1 = 2\gamma$ and $m_2 = 5\gamma$. Given that evaluating ∇g is 7 times faster than evaluating ∇f, the computational costs of the SF and MF estimators are equivalent. In the following, we refer to γ as the cost coefficient, as we have set the computational budget to increase linearly with γ.

Figure 7 (left panel) shows the relative error with respect to the reference estimator $\widehat{H}_{SF}^{\mathrm{ref}}$ as a function of the cost coefficient γ, averaged over 100 independent experiments. The MF estimator outperforms the SF estimator for all the budgets tested. In particular, the MF estimator (respectively, SF estimator) reaches a relative error of 0.4 for $\gamma = 4$ (respectively, $\gamma = 7$). This represents a 42.8% reduction in computational resources for the MF estimator. Figure 7 (right panel) shows the eigenvalues of the AS matrix estimators, compared to those of the reference, for cost coefficient $\gamma = 10$. The MF estimator provides a better approximation of the spectral decay of the AS matrix than the SF estimator. We note that the spectral decay suggests that a few modes (≈ 5) are sufficient to describe the behavior of the function f.
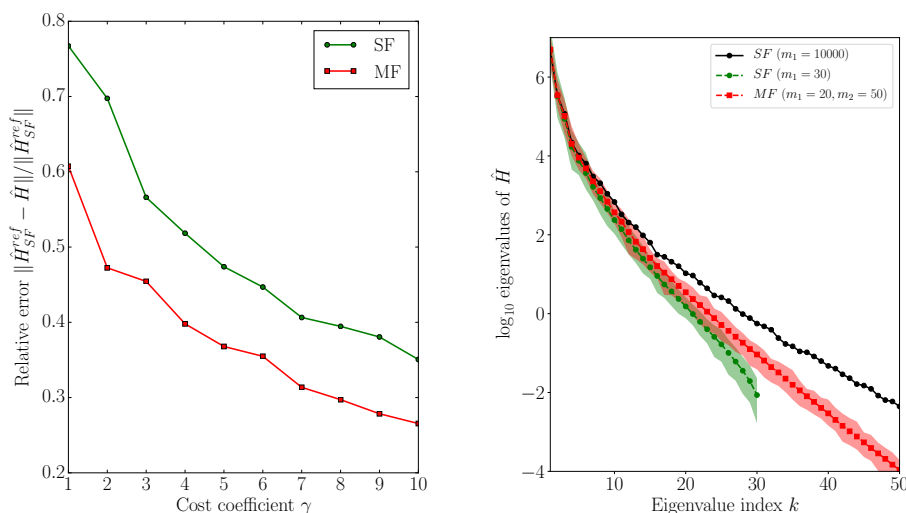
© 2020 Remi Lam, Olivier Zahm, Youssef Marzouk, Karen Willcox

FIG. 7. *Left panel: Relative error as a function of cost coefficient γ, averaged over 100 independent experiments. At an equivalent computational budget, the MF estimator outperforms the SF estimator. Right panel: Eigenvalues of the AS matrix estimators for the vertical displacement of the wrench PoI averaged over 100 independent experiments. Shadings represent the minimum and maximum values over the 100 independent experiments. The high-fidelity SF estimator used $m_1 = 30$ samples and the MF estimator is constructed with with $m_1 = 20$ and $m_2 = 50$ samples. This corresponds to a similar computational budget.*
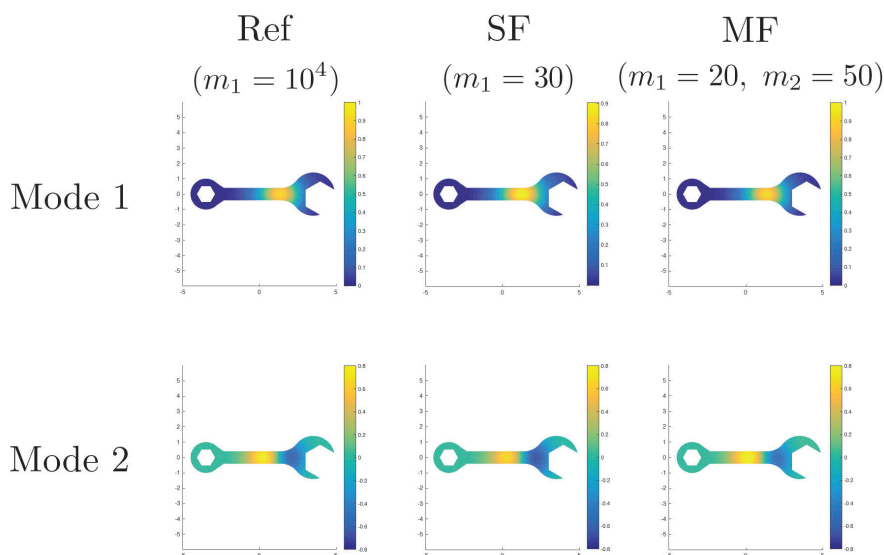

FIG. 8. *First (top row) and second leading modes (bottom row) of the AS matrix for the reference estimator (left column), SF estimator (middle column), and MF estimator (right column) for the cost coefficient $\gamma = 10$. Both the SF and the MF estimators correctly capture the two leading modes of the reference AS matrix.*

Finally, Figure 8 shows the two leading modes (eigenvectors) from the reference, the SF, and the MF estimators for cost coefficient $\gamma = 10$. We note that both the SF and the MF estimators recover the leading modes correctly.

5.3. ONERA M6 wing shape optimization. In this section, we illustrate the proposed MF estimator on an expensive-to-evaluate engineering example.

We consider the problem of finding the active subspace associated with the shape optimization of the ONERA M6 wing. The shape of the wing is parameterized by free form deformation (FFD) box control points. In our experiment we used $5 \times 8 \times 2 = 80$ FFD boxes. Imposing second-order continuity of surfaces with the FFD fixes $5 \times 3 \times 2 = 30$ variables, leaving $d = 50$ control points. These correspond to the input variable $\boldsymbol{x} \in \mathcal{X} = [-0.05, 0.05]^{50}$. The functions of interest are the drag and the lift coefficients. Those quantities are computed using expensive-to-evaluate computational fluid dynamics (CFD) tools. We use the SU2 package[3] [32] to solve the Reynolds-averaged Navier–Stokes (RANS) equations and compute the gradients using the continuous adjoint method. The flow conditions are such that the Mach number is $M_\infty = 0.8395$, the angle of attack of the wing is $\alpha = 3.03°$, and the Reynolds number is $Re = 11.72 \times 10^6$. We use the Spalart–Allmaras turbulence model.

The high-fidelity function uses the aforementioned CFD model with the following stopping criteria: the solution of the RANS equations and the associated adjoint are computed with a limit of 10^4 iterations or fewer if the Cauchy convergence criterion reaches 10^{-5} within this limit (i.e., maximum variation of the quantity of interest over 100 iterations is lower than 10^{-5}). For the low-fidelity function g, we use the same model as the high-fidelity function f but reduce the maximum number of iterations allowed for convergence from 10^4 to 10^3. An evaluation of the high-fidelity model (drag and lift coefficients and associated gradients) is thus approximately 10 times more expensive than the low-fidelity model.

We compute a total of 100 high-fidelity evaluations and 500 low-fidelity evaluations. We use the 100 evaluations of the high-fidelity model to compute a first "reference" SF estimator ($m_1 = 100$) that we denote $\widehat{H}_{SF}^{(100)}$. We split the collected data into 5 independent experimental batches and construct three estimators per batch. The first is an SF estimator constructed with $m_1 = 10$ high-fidelity evaluations. The second is an SF estimator built with $m_1 = 20$ high-fidelity evaluations. The last estimator is an MF estimator with $m_1 = 10$ and $m_2 = 90$ evaluations. For these three estimators, the m_1 high-fidelity evaluations are common. We note that the computational cost of the second SF estimator ($m_1 = 20$) and the MF estimator ($m_1 = 10$ and $m_2 = 90$) are similar.

Figure 9 shows the 20 first eigenvalues of the three estimators (averaged over 5 independent batches) and $\widehat{H}_{SF}^{(100)}$. Note that the SF estimators (except $\widehat{H}_{SF}^{(100)}$) are rank deficient (rank 10 and rank 20). The MF estimator is full rank (rank 50). The MF estimator is closer to $\widehat{H}_{SF}^{(100)}$ than the two SF estimators. In particular, the MF estimator outperforms the second SF estimator with a similar computational budget. The error bars show the maximum and the minimum values over the 5 independent experiments for each estimator and confirm the robustness of the proposed method.

Figure 10 shows the 50 eigenvalues of the SF estimator $\widehat{H}_{SF}^{(100)}$ ($m_1 = 100$) and a MF estimator using all the available evaluations ($m_1 = 100$, $m_2 = 400$). The leading eigenvalues are similar for both estimators. The difference between the two estimators increases for lower eigenvalues (higher indices). For both estimators, we compute approximations of the characteristic quantities δ_H, $\mathbb{E}[\|\nabla f(X)\|^2]$, $\|H\|$, β^2, and θ^2 and summarize them in Table 1. The SF quantities are computed based on the $m_1 = 100$ high-fidelity evaluations: $\|H\|$ and δ_H are computed using

---

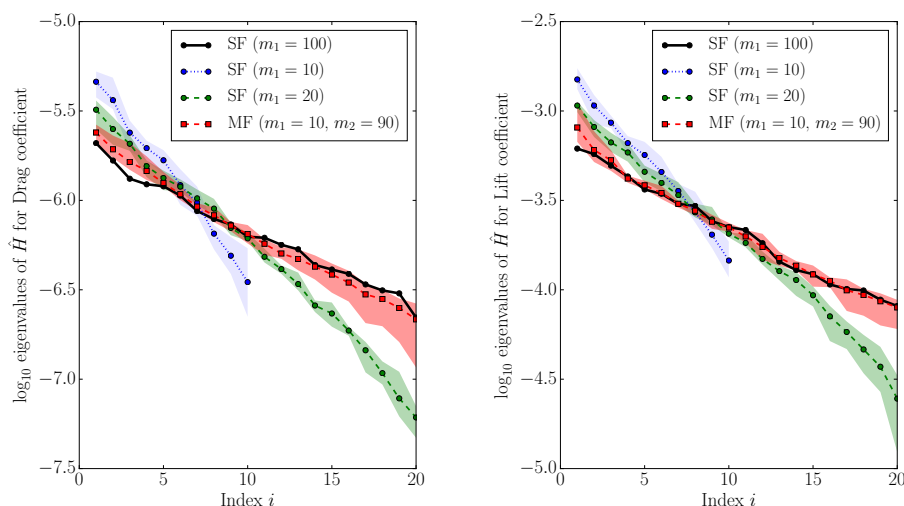[3]Available at https://github.com/su2code/SU2/tree/ec551e427f20373511432e6cd87402304cc46baa

Fig. 9. *Eigenvalues of the AS matrix estimators for the drag coefficient (left panel) and for the lift coefficient (right panel) averaged over 5 independent experiments. The SF estimator with $m_1 = 20$ and the MF estimator (dashed lines) used the same computational budget to evaluate the gradients. At an equal budget, the MF estimator provides a better estimate of the spectrum of $\widehat{H}_{SF}^{(100)}$ than the SF estimator. Shadings represent maximum and minimum values over the 5 independent experiments.*
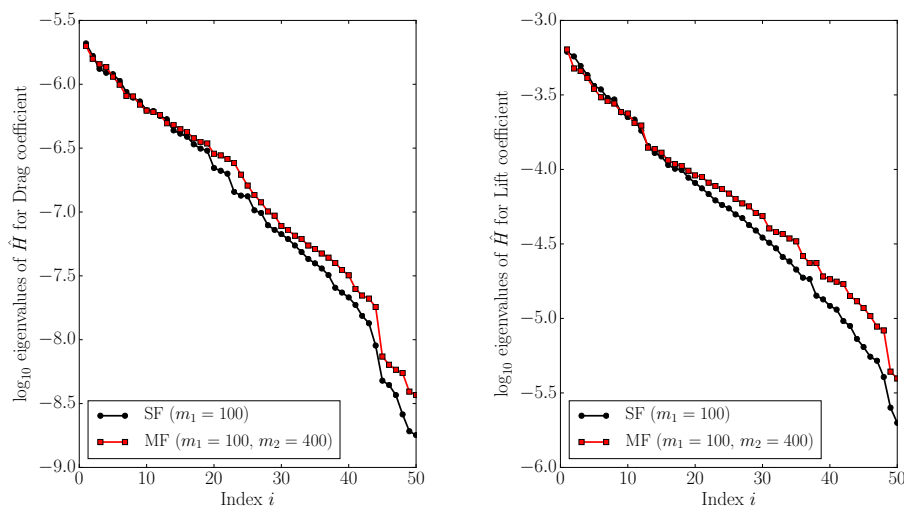

Fig. 10. *Eigenvalues of the AS matrix estimators for the drag coefficient (left panel) and for the lift coefficient (right panel). The high-fidelity SF estimator used $m_1 = 100$ samples and the MF estimator is constructed with $m_1 = 100$ and $m_2 = 400$ samples.*

$\widehat{H}_{SF}^{(100)}$ in lieu of H, $\mathbb{E}[\|\nabla f(X)\|^2]$ is approximated by $\frac{1}{m_1}\sum_{i=1}^{m_1}\|\nabla f(X_i)\|^2$, while β^2 is approximated by $\max_i \|\nabla f(X_i)\|^2/\mathbb{E}[\|\nabla f(X)\|^2]$ for $i \in \{1,\ldots,m_1\}$. The MF quantities are computed based on the $m_1 = 100$ high-fidelity evaluations and $m_1 + m_2 = 500$ low-fidelity evaluations: $\|H\|$ and δ_H are computed using the MF estimator in lieu of H, $\mathbb{E}[\|\nabla f(X)\|^2]$ is approximated by the (scalar) MF estimator

TABLE 1

Approximation of the characteristic quantities for the SF and MF estimators for the drag and lift coefficients.

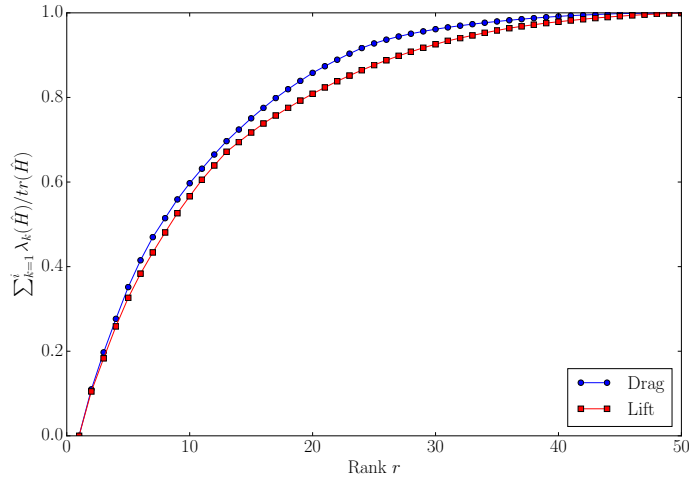|	Drag coefficient C_d		Lift coefficient C_l	
	SF	MF	SF	MF
$\mathbb{E}[\|\nabla f(X)\|^2]$	1.74×10^{-5}	1.81×10^{-5}	5.96×10^{-3}	6.08×10^{-3}
$\|H\|$	2.09×10^{-6}	1.99×10^{-6}	6.16×10^{-4}	6.38×10^{-4}
δ_H	8.34	9.10	9.67	9.53
β^2	1.76	-	1.54	-
θ^2	-	1.30×10^{-5}	-	1.03×10^{-4}

FIG. 11. *Percentage of the spectrum $\sum_{k=1}^{i} \lambda_r(\widehat{H})/tr(\widehat{H})$ captured by the best rank-r approximation of $\widehat{H}$, where $\lambda_k(\widehat{H})$ is the kth eigenvalue of $\widehat{H}$.*

$\frac{1}{m_1}\sum_{i=1}^{m_1}(\|\nabla f(X_i)\|^2 - \|\nabla g(X_i)\|^2) + \frac{1}{m_2}\sum_{i=m_1+1}^{m_1+m_2}\|\nabla g(X_i)\|^2$, while θ^2 is approximated by $\max_i \|\nabla f(X_i) - \nabla g(X_i)\|^2/\mathbb{E}[\|\nabla f(X)\|^2]$ for $i \in \{1, \ldots, m_1\}$. From Table 1, it can be seen that the intrinsic dimension of the AS matrix is approximately 9 for both the drag and lift coefficients. Table 1 also shows that the parameter β^2 is not large (≤ 2), which indicates, along with the low intrinsic dimension, that computing a good estimator of H requires relatively few evaluations. We also note that the low-fidelity model is a good approximation of the high-fidelity model, with $\theta^2 \leq 1.5 \times 10^{-5}$ for the drag and $\theta^2 \leq 1.1 \times 10^{-4}$ for the lift. Figure 11 shows the normalized sum of the eigenvalues of the best rank-r approximation of $\widehat{H}$. Selecting the 20 first eigenvectors of $\widehat{H}$ allows us to capture more than 80% of the spectral content and decreases by 30 the dimensionality of the shape optimization problem (a 60% decrease).

6. Conclusions. We proposed a multifidelity approach to identify low-dimensional subspaces capturing most of the variation of a function of interest. Our approach builds on the gradient-based active subspace methodology, which seeks to compute the matrix H containing the second moments of the gradient function. The proposed approach reduces the computational cost of Monte Carlo methods used to estimate this matrix, by using a low-fidelity, cheap-to-evaluate gradient approximation as a control variate. The performance improvements of the resulting multifidelity

(MF) estimator $\widehat{H}_{MF}$ are demonstrated on two engineering examples governed by partial differential equations: a high-dimensional linear elasticity problem defined over more than two thousand input variables and an expensive shape optimization problem defined over 50 input variables.

Analysis of the performance of the multifidelity technique yields error bounds for the matrix error $\|H - \widehat{H}_{MF}\|$ both in expectation and with high probability. These error bounds depend on the intrinsic dimension of the problem, which is related to the spectral decay of the active subspace matrix. When a function varies mostly along a few directions, the intrinsic dimension is low. In such a case, approximating H may require only a small number of gradient evaluations. This relationship was confirmed empirically by a parametric study conducted on two analytical problems: lower intrinsic dimension led to a lower matrix error for the same number of high-fidelity evaluations.

The performance improvements of the multifidelity approach are threefold. First, we showed that the MF estimator reduces the cost of performing dimension reduction. This was illustrated on the linear elasticity problem, where the multifidelity approach needed about 43% less computational effort than its single-fidelity counterpart to achieve the same relative error. Second, we showed that the multifidelity approach was able to recover eigenvectors qualitatively similar to the exact ones. Third, the MF estimator led to better estimates of the spectral decay of H. On the linear elasticity problem, which is characterized by a low intrinsic dimension (≤ 5), the multifidelity approach led to better estimates, especially for smaller eigenvalues. On the shape optimization problem, which is characterized by a higher intrinsic dimension (≈ 9), the MF estimator led to better estimates for all eigenvalues. This behavior is in contrast to the single-fidelity method, which overestimated the leading eigenvalues and underestimated the lower eigenvalues, thereby underestimating the intrinsic dimension of H and overestimating the spectral decay. Recovering the spectral decay of H is particularly important since one popular way of choosing the dimension r of the active subspace is based on the spectral gap between consecutive eigenvalues. If the spectral decay is overestimated, r might be chosen too small to correctly capture the behavior of the high-fidelity function. By providing a better estimate of the spectral decay, the MF estimator reduces this risk.

The dimension of the active subspace can also be chosen to control an error bound on the associated function approximation error, i.e., when using the active subspace to construct a ridge approximation of the original function. We showed that the approximation of this bound, which depends on the unknown H, improves as the matrix error decreases. Because the MF estimator reduces the error in estimating H for a given computational effort, the proposed multifidelity approach leads to a better selection of the active subspace dimension r.

Appendix A. Proofs of main results. In this section, we present the details of the proofs of Proposition 4.1 (Appendix A.1) and Proposition 4.2 (Appendix A.2).

A.1. Proof of Proposition 4.1. The proof of Proposition 4.1 relies on the concentration inequality known as the matrix Bernstein theorem. It corresponds to Theorem 6.1.1 from [47] restricted to the case of real symmetric matrices.

THEOREM A.1 (Matrix Bernstein: real symmetric case [47]). *Let $S_1, \ldots, S_m$ be m independent zero-mean random symmetric matrices in $\mathbb{R}^{d \times d}$. Assume there exists $L < \infty$ such that*

$$\|S_i\| \leq L$$

almost surely for all $1 \leq i \leq m$, and let v be such that

$$\|\mathbb{E}[(S_1 + \cdots + S_m)^2]\| \leq v.$$

Then,

$$\mathbb{E}[\|S_1 + \cdots + S_m\|] \leq \sqrt{2v \log(2d)} + \frac{1}{3} L \log(2d).$$

Furthermore, for any $t \geq 0$ we have

$$\mathbb{P}\{\|S_1 + \cdots + S_m\| \geq t\} \leq 2d \exp\left(\frac{-t^2/2}{v + Lt/3}\right).$$

In order to apply the matrix Bernstein theorem, we need to express the difference between the AS matrix H and the MF estimator $\widehat{H}_{MF}$ as the sum of independent matrices. We write $H - \widehat{H}_{MF} = S_1 + \cdots + S_m$, where $m = m_1 + m_2$ and

$$S_i = \begin{cases} \frac{1}{m_1}\left(H - G - (\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T)\right) & \text{if } 1 \leq i \leq m_1, \\ \frac{1}{m_2}\left(G - \nabla g(X)\nabla g(X)^T\right) & \text{otherwise,} \end{cases}$$

where $G = \mathbb{E}[\nabla g(X)\nabla g(X)^T]$. The following property provides bounds for $\|S_i\|$ and $\mathbb{E}[\|S_1 + \cdots + S_m\|]$. Those bounds are expressed as functions of m_1, m_2, β, θ, and $\mathbb{E}[\|\nabla f(X)\|^2]$.

PROPERTY A.2. *Assumptions* (4.1) *and* (4.2) *yield*

$$(A.1) \quad \|\mathbb{E}[(S_1 + \cdots + S_m)^2]\| \leq \left(\frac{\theta^2(2+\theta)^2}{m_1} + \frac{(\theta+\beta)^2(1+\theta)^2}{m_2}\right)\mathbb{E}[\|\nabla f(X)\|^2]^2,$$

and

$$(A.2) \quad \|S_i\| \leq \max\left\{\frac{2\theta(2\beta+\theta)}{m_1}; \frac{2(\theta+\beta)^2}{m_2}\right\}\mathbb{E}[\|\nabla f(X)\|^2]$$

almost surely for all $1 \leq i \leq m$.

Proof. We first derive the bound (A.1) for the variance of the estimator before proving the bound (A.2) for the norm of the summands.

Using the independence of the summands S_i and the fact that $\mathbb{E}[S_i] = 0$, we have

$$\mathbb{E}[(S_1 + \cdots + S_m)^2] = \mathbb{E}[S_1^2] + \cdots + \mathbb{E}[S_{m_1}^2] + \mathbb{E}[S_{m_1+1}^2] + \cdots + \mathbb{E}[S_m^2]$$

$$(A.3) \qquad\qquad = \frac{1}{m_1}\mathbb{E}[(A - \mathbb{E}[A])^2] + \frac{1}{m_2}\mathbb{E}[(B - \mathbb{E}[B])^2],$$

where

$$A = \nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T,$$
$$B = \nabla g(X)\nabla g(X)^T.$$

Notice that $0 \preccurlyeq \mathbb{E}[(A - \mathbb{E}[A])^2] = \mathbb{E}[A^2] - \mathbb{E}[A]^2 \preccurlyeq \mathbb{E}[A^2]$ so that $\|\mathbb{E}[(A - \mathbb{E}[A])^2]\| \leq \|\mathbb{E}[A^2]\|$, where $\preccurlyeq$ denotes the Loewner partial order. Similarly, one has

$$\|\mathbb{E}[(B - \mathbb{E}[B])^2]\| \leq \|\mathbb{E}[B^2]\|.$$

Taking the norm of (A.3) and using a triangle inequality yields

$$\|\mathbb{E}[(S_1 + \cdots + S_m)^2]\| \leq \frac{1}{m_1}\|\mathbb{E}[A^2]\| + \frac{1}{m_2}\|\mathbb{E}[B^2]\|.$$

To obtain (A.1), it remains to show (i) that $\|\mathbb{E}[A^2]\| \leq \theta^2(2+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2]^2$ and (ii) that $\|\mathbb{E}[B^2]\| \leq (\beta+\theta)^2(1+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2]^2$. Let $u \in \mathbb{R}^d$ such that $\|u\| \leq 1$. We have

$$
\begin{aligned}
u^T\mathbb{E}[A^2]u &= \mathbb{E}[u^T(\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T)^2 u] \\
&= \mathbb{E}[\|(\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T)u\|^2] \\
&= \mathbb{E}[\|\nabla f(X)(\nabla f(X) - \nabla g(X))^T u - (\nabla g(X) - \nabla f(X))\nabla g(X)^T u\|^2].
\end{aligned}
$$

Using triangle inequalities, we can write

$$
\begin{aligned}
u^T\mathbb{E}[A^2]u &\leq \mathbb{E}[(\|\nabla f(X)(\nabla f(X) - \nabla g(X))^T u\| + \|(\nabla g(X) - \nabla f(X))\nabla g(X)^T u\|)^2] \\
&\leq \mathbb{E}[(\|\nabla f(X)\|\,\|\nabla f(X) - \nabla g(X)\| + \|\nabla g(X) - \nabla f(X)\|\,\|\nabla g(X)\|)^2] \\
&\leq \mathbb{E}[\|\nabla f(X) - \nabla g(X)\|^2(\|\nabla f(X)\| + \|\nabla g(X)\|)^2] \\
&\leq \mathbb{E}[\|\nabla f(X) - \nabla g(X)\|^2(2\|\nabla f(X)\| + \|\nabla f(X) - \nabla g(X)\|)^2] \\
(\text{A.4}) \quad &\leq \theta^2\mathbb{E}[\|\nabla f(X)\|^2]\,\mathbb{E}[(2\|\nabla f(X)\| + \theta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]})^2],
\end{aligned}
$$

where for the last inequality we used Assumption (4.2). Expanding the last term yields

$$
\begin{aligned}
\mathbb{E}[(2\|\nabla f(X)\| &+ \theta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]})^2] \\
&= 4\mathbb{E}[\|\nabla f(X)\|^2] + 4\theta\mathbb{E}[\|\nabla f(X)\|]\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]} + \theta^2\mathbb{E}[\|\nabla f(X)\|^2] \\
&\leq 4\mathbb{E}[\|\nabla f(X)\|^2] + 4\theta\mathbb{E}[\|\nabla f(X)\|^2] + \theta^2\mathbb{E}[\|\nabla f(X)\|^2] \\
(\text{A.5}) \quad &= (2+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2].
\end{aligned}
$$

Here, we used the relation $\mathbb{E}[\|\nabla f(X)\|]^2 \leq \mathbb{E}[\|\nabla f(X)\|^2]$, which holds true by Jensen's inequality. Combining (A.4) and (A.5) and taking the supremum over $\|u\| \leq 1$ yields

$$\|\mathbb{E}[A^2]\| \leq \theta^2(2+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2]^2,$$

which gives (i). To show (ii), we first notice that Assumptions (4.1) and (4.2) yield

$$(\text{A.6}) \quad \|\nabla g(X)\|^2 \leq (\|\nabla f(X)\| + \|\nabla g(X) - \nabla f(X)\|)^2 \leq (\beta+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2]$$

almost surely. Then, for any $u \in \mathbb{R}^d$ such that $\|u\| \leq 1$, we have

$$
\begin{aligned}
u^T\mathbb{E}[B^2]u &= \mathbb{E}[u^T(\nabla g(X)\nabla g(X)^T)^2 u] \\
&= \mathbb{E}[\|\nabla g(X)\|^2(\nabla g(X)^T u)^2] \\
(\text{A.7}) \quad &\leq (\beta+\theta)^2\mathbb{E}[\|\nabla f(X)\|^2]\mathbb{E}[(\nabla g(X)^T u)^2].
\end{aligned}
$$

Using similar arguments, the last term in the above relation satisfies

$$
\begin{aligned}
\mathbb{E}[(\nabla g(X)^T u)^2] &= \mathbb{E}[(\nabla g(X)^T u)^2 - (\nabla f(X)^T u)^2] + \mathbb{E}[(\nabla f(X)^T u)^2] \\
&= \mathbb{E}[((\nabla g(X) + \nabla f(X))^T u)(\nabla g(X) - \nabla f(X)^T u)] + \mathbb{E}[(\nabla f(X)^T u)^2] \\
&\leq \mathbb{E}[\|\nabla g(X) + \nabla f(X))\| \ \|\nabla g(X) - \nabla f(X)\|] + \mathbb{E}[\|\nabla f(X)\|^2] \\
&\leq \mathbb{E}[(2\|\nabla f(X)\| + \|\nabla g(X) - \nabla f(X)\|)\|\nabla g(X) - \nabla f(X)\|] + \mathbb{E}[\|\nabla f(X)\|^2] \\
&\leq 2\mathbb{E}[\|\nabla f(X)\|]\theta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]} + \theta^2 \mathbb{E}[\|\nabla f(X)\|^2] + \mathbb{E}[\|\nabla f(X)\|^2] \\
(\text{A.8}) \quad &\leq (2\theta + \theta^2 + 1)^2 \mathbb{E}[\|\nabla f(X)\|^2] = (1+\theta)^2 \mathbb{E}[\|\nabla f(X)\|^2].
\end{aligned}
$$

Combining (A.7) with (A.8) and taking the supremum over $\|u\| \leq 1$ yields

$$
\|\mathbb{E}[B^2]\| \leq (\beta + \theta)^2 (1+\theta)^2 \mathbb{E}[\|\nabla f(X)\|^2]^2,
$$

which establishes (ii). This proves (A.1).

Now we prove the second part of the property: the bound on the summand (A.2). Recall that S_i is defined as $\frac{1}{m_1}(\mathbb{E}[A] - A)$ if $1 \leq i \leq m_1$ and as $\frac{1}{m_2}(\mathbb{E}[B] - B)$ if $m_1 \leq i \leq m$. To obtain (A.2), it is then sufficient to show (iii) that

$$
\|\mathbb{E}[A] - A\| \leq 2\theta(2\beta + \theta)\mathbb{E}[\|\nabla f(X)\|^2]
$$

almost surely and (iv) that $\|\mathbb{E}[B] - B\| \leq 2(\beta + \theta)^2 \mathbb{E}[\|\nabla f(X)\|^2]$ almost surely. To show (iii), we first notice that

$$
\begin{aligned}
\|A\| &= \|\nabla f(X)\nabla f(X)^T - \nabla g(X)\nabla g(X)^T\| \\
&= \|\nabla f(X)(\nabla f(X) - \nabla g(X))^T - (\nabla g(X) - \nabla f(X))\nabla g(X)^T\| \\
&\leq \|\nabla f(X)\|\|\nabla f(X) - \nabla g(X))\| + \|\nabla g(X) - \nabla f(X)\|\|\nabla g(X)\| \\
&\leq \|\nabla f(X) - \nabla g(X))\|(2\|\nabla f(X)\| + \|\nabla f(X) - \nabla g(X)\|) \\
&\leq \theta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]}(2\beta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]} + \theta\sqrt{\mathbb{E}[\|\nabla f(X)\|^2]}) \\
&\leq \theta(2\beta + \theta)\mathbb{E}[\|\nabla f(X)\|^2].
\end{aligned}
$$

Then, we can write

$$
\|\mathbb{E}[A] - A\| \leq \|\mathbb{E}[A]\| + \|A\| \leq 2\theta(2\beta + \theta)\mathbb{E}[\|\nabla f(X)\|^2],
$$

which gives (iii). Finally we have

$$
\|\mathbb{E}[B] - B\| \leq \mathbb{E}[\|\nabla g(X)\|^2] + \|\nabla g(X)\|^2 \overset{(\text{A.6})}{\leq} 2(\beta + \theta)^2 \mathbb{E}[\|\nabla f(X)\|^2],
$$

which gives (iv) and therefore (A.2). This concludes the proof of Property A.2. $\square$

To prove our main result, Proposition 4.1, it remains to express the bounds of the estimator variance and the summands as a function of m_1 and to apply the matrix Bernstein theorem. By Assumption (4.3), we have $m_2 \geq m_1 \frac{(\theta+\beta)^2}{\theta(2\beta+\theta)}$ so that, using (A.2) of Property A.2, we have that

$$
\|S_i\| \leq \frac{2\theta(2\beta + \theta)}{m_1}\mathbb{E}[\|\nabla f(X)\|^2] =: L,
$$

holds almost surely for all $1 \leq i \leq m$. By Assumption (4.3), we also have $m_2 \geq m_1 \frac{(\theta+\beta)^2(1+\theta)^2}{\theta^2(2+\theta)^2}$ so that (A.1) yields

$$\|\mathbb{E}[(S_1 + \cdots + S_m)^2]\| \leq \frac{2\theta^2(2+\theta)^2}{m_1}\mathbb{E}[\|\nabla f(X)\|^2]^2 =: v.$$

Applying the matrix Bernstein theorem (Theorem A.1) gives

$$\begin{aligned}
\mathbb{E}[\|H - \widehat{H}_{MF}\|] &= \mathbb{E}[\|S_1 + \cdots + S_m\|] \\
&\leq \sqrt{2v\log(2d)} + \frac{1}{3}L\log(2d) \\
&= \left(\frac{2\theta(2+\theta)\sqrt{\log(2d)}}{\sqrt{m_1}} + \frac{2\theta(2\beta+\theta)\log(2d)}{3m_1}\right)\mathbb{E}[\|\nabla f(X)\|^2] \\
&\overset{(4.4)}{\leq} (\varepsilon + \varepsilon^2)\frac{\mathbb{E}[\|\nabla f(X)\|^2]}{\delta_H} = (\varepsilon + \varepsilon^2)\|H\|,
\end{aligned}$$

where, for the last equality, we used the definition of $\delta_H = \text{trace}(H)/\|H\|$ and the fact that $\text{trace}(H) = \text{trace}\,\mathbb{E}(\nabla f(X)\nabla f(X)^T) = \mathbb{E}[\|\nabla f(X)\|^2]$. This proves (4.5).

To show (4.7), we apply the high-probability bound of Theorem A.1 with $t = \varepsilon\|H\|$. We obtain

$$\begin{aligned}
\mathbb{P}\{\|H - \widehat{H}_{MF}\| \geq \varepsilon\|H\|\} &\leq 2d\exp\left(\frac{-\varepsilon^2\|H\|^2/2}{v + L\varepsilon\|H\|/3}\right) \\
&= 2d\exp\left(\frac{-\varepsilon^2 m_1}{4\theta^2(2+\theta)^2\delta_H^2 + 4/3\theta(2\beta+\theta)\delta_H\varepsilon}\right) \overset{(4.6)}{\leq} \eta,
\end{aligned}$$

which is (4.7). This concludes the proof of Proposition 4.1.

A.2. Proof of Proposition 4.2. The proof of Proposition 4.2 relies on the concentration inequality known as the intrinsic dimension matrix Bernstein theorem. It corresponds to Property 7.3.1 and Corollary 7.3.2 from [47], restricted to the case of real symmetric matrices.

THEOREM A.3 (Matrix Bernstein: intrinsic dimension, real symmetric case). *Let* $S_1, \ldots, S_m$ *be* m *zero-mean random symmetric matrices in* $\mathbb{R}^{d \times d}$. *Assume the existence of* $L < \infty$ *such that*

$$\|S_i\| \leq L$$

almost surely for all $1 \leq i \leq m$. *Let* $V \in \mathbb{R}^{d \times d}$ *be a symmetric matrix such that*

$$\mathbb{E}[(S_1 + \cdots + S_m)^2] \preccurlyeq V,$$

and let $\delta_V = \text{trace}(V)/\|V\|$ *and* $v = \|V\|$. *Then, we have*

$$\mathbb{E}[\|S_1 + \cdots + S_m\|] \leq C_0\left(\sqrt{v\ln(1 + 2\delta_V)} + L\ln(1 + 2\delta_V)\right),$$

where C_0 *is an absolute (numerical) constant. Furthermore, for any* $t \geq \sqrt{v} + L/3$, *we have*

$$\mathbb{P}\{\|S_1 + \cdots + S_m\| \geq t\} \leq 8\delta_V\exp\left(\frac{-t^2/2}{v + Lt/3}\right).$$

Proof. Using the definition of V and the symmetry of the matrices S_i, we have

$$(A.9) \qquad V \succcurlyeq \mathbb{E}[(S_1 + \cdots + S_m)^2] = \mathbb{E}[(S_1 + \cdots + S_m)(S_1 + \cdots + S_m)^T],$$

$$(A.10) \qquad V \succcurlyeq \mathbb{E}[(S_1 + \cdots + S_m)^2] = \mathbb{E}[(S_1 + \cdots + S_m)^T (S_1 + \cdots + S_m)].$$

Defining the matrix $M = \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}$, we have $\delta_M = 2\delta_V$. A direct application of Property 7.3.1 and Corollary 7.3.2 from [47] yields the theorem. □

In order to apply the intrinsic dimension matrix Bernstein theorem, we express the difference between the AS matrix H and the SF estimator $\widehat{H}_{SF}$ as the sum of independent matrices. We write $H - \widehat{H}_{SF} = S_1 + \cdots + S_{m_1}$, where

$$S_i = \frac{1}{m_1}\left(H - \nabla f(X_i)\nabla f(X_i)^T \right)$$

for all $1 \le i \le m_1$. Since $H = \mathbb{E}[\nabla f(X)\nabla f(X)]$, we have

$$\|S_i\| \le \frac{\|H\| + \|\nabla f(X_i)\nabla f(X_i)^T\|}{m_1} \le \frac{\mathbb{E}[\|\nabla f(X)\|^2] + \|\nabla f(X_i)\|^2}{m_1}$$
$$\overset{(4.8)}{\le} \frac{\mathbb{E}[\|\nabla f(X)\|^2] + \beta^2 \mathbb{E}[\|\nabla f(X)\|^2]}{m_1} = \frac{1+\beta^2}{m_1}\mathbb{E}[\|\nabla f(X)\|^2] =: L$$

almost surely for all $1 \le i \le m_1$. By independence of the summands S_i and given $\mathbb{E}[S_i] = 0$, we have

$$\mathbb{E}[(S_1 + \cdots + S_{m_1})^2] = \mathbb{E}[S_1^2] + \cdots + \mathbb{E}[S_{m_1}^2]$$
$$= \frac{1}{m_1}\mathbb{E}[(\nabla f(X)\nabla f(X)^T - \mathbb{E}[\nabla f(X)\nabla f(X)^T])^2]$$
$$\preccurlyeq \frac{1}{m_1}\mathbb{E}[(\nabla f(X)\nabla f(X)^T)^2] = \frac{1}{m_1}\mathbb{E}[\|\nabla f(X)\|^2 \nabla f(X)\nabla f(X)^T]$$
$$\overset{(4.8)}{\preccurlyeq} \frac{\beta^2 \mathbb{E}[\|\nabla f(X)\|^2]}{m_1}\mathbb{E}[\nabla f(X)\nabla f(X)^T] = \frac{\beta^2 \mathbb{E}[\|\nabla f(X)\|^2]}{m_1}H =: V.$$

With the above definition of V, we have

$$v := \|V\| = \frac{\beta^2 \mathbb{E}[\|\nabla f(X)\|^2]\|H\|}{m_1},$$
$$\delta := \frac{\text{trace}(V)}{\|V\|} = \frac{\text{trace}(H)}{\|H\|} = \delta_H.$$

Applying the expectation bound of Theorem A.3 gives

$$\mathbb{E}[\|H - \widehat{H}_{SF}\|] = \mathbb{E}[\|S_1 + \cdots + S_m\|] \le C_0\left(\sqrt{v\ln(1+2\delta)} + L\ln(1+2\delta) \right)$$

$$\le C_0\left(\sqrt{\frac{\beta^2 \mathbb{E}[\|\nabla f(X)\|^2]\|H\|}{m_1}\ln(1+2\delta_H)} + \frac{1+\beta^2}{m_1}\mathbb{E}[\|\nabla f(X)\|^2]\ln(1+2\delta_H) \right)$$

$$\overset{(4.9)}{\le} C_0\left(\sqrt{\frac{\varepsilon^{-2}\|H\|^2}{C}} + \frac{\varepsilon^{-2}\|H\|}{C} \right) \le (\varepsilon + \varepsilon^2)\|H\|,$$

where the last inequality is obtained by defining $C = \max\{C_0, C_0^2\}$. This yields (4.10).

Finally, letting $t = \varepsilon\|H\|$, the high-probability bound of Theorem A.3 ensures that

$$\mathbb{P}\{\|H - \widehat{H}_{SF}\| \geq \varepsilon\|H\|\} \leq 8\delta_H \exp\left(\frac{-\varepsilon^2\|H\|^2/2}{v + L\varepsilon\|H\|/3}\right)$$

$$= 8\delta_H \exp\left(\frac{-\varepsilon^2\|H\|^2/2}{\frac{\beta^2\mathbb{E}[\|\nabla f(X)\|^2]\|H\|}{m_1} + \frac{1+\beta^2}{m_1}\mathbb{E}[\|\nabla f(X)\|^2]\varepsilon\|H\|/3}\right)$$

$$= 8\delta_H \exp\left(\frac{-m_1\varepsilon^2/2}{\beta^2\delta_H + (1+\beta^2)\delta_H\varepsilon/3}\right) \overset{(4.11)}{\leq} \eta,$$

which is (4.12). This concludes the proof of Proposition 4.2.

REFERENCES

[1] N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green, and P. A. Newman, *Approximation and model management in aerodynamic optimization with variable-fidelity models*, J. Aircraft, 38 (2001), pp. 1093–1101.

[2] F. Ballarin, A. D'Amario, S. Perotto, and G. Rozza, *A POD-selective inverse distance weighting method for fast parametrized shape morphing*, Internat. J. Numer. Methods Engrg., 117 (2019), pp. 860–884.

[3] A. Beskos, A. Jasra, K. Law, Y. Marzouk, and Y. Zhou, *Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 762–786.

[4] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[5] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.

[6] P. Constantine and D. Gleich, *Computing Active Subspaces with Monte Carlo*, preprint, https://arxiv.org/abs/1408.0545, 2015.

[7] P. G. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*, SIAM. Philadelphia, 2015.

[8] P. G. Constantine and A. Doostan, *Time-dependent global sensitivity analysis with active subspaces for a lithium ion battery model*, Stat. Anal. Data Min., 10 (2017), pp. 243–262.

[9] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524.

[10] P. G. Constantine, C. Kent, and T. Bui-Thanh, *Accelerating Markov chain Monte Carlo with active subspaces*, SIAM J. Sci. Comput., 38 (2016), pp. A2779–A2805.

[11] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 30 (2014), 114015.

[12] T. Cui, Y. Marzouk, and K. Willcox, *Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction*, J. Comput. Phys., 315 (2016), pp. 363–387.

[13] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Multi-fidelity optimization via surrogate modelling*, R. Soc. Proc. A Math. Phys. Eng., 463 (2007), pp. 3251–3269.

[14] M. B. Giles, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617.

[15] W. Hackbusch, *Multi-grid Methods and Applications*, Springer, Berlin, 2013.

[16] C. Himpe and M. Ohlberger, *Data-driven combined state and parameter reduction for inverse problems*, Adv. Comput. Math., 41 (2015), pp. 1343–1364.

[17] J. T. Holodnak, I. C. F. Ipsen, and R. C. Smith, *A probabilistic subspace bound with application to active subspaces*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1208–1220.

[18] J. L. Jefferson, R. M. Maxwell, and P. G. Constantine, *Exploring the sensitivity of photosynthesis and stomatal resistance parameters in a land surface model*, J. Hydrometeorol., 18 (2017), pp. 897–915.

[19] W. Ji, J. Wang, O. Zahm, Y. Marzouk, B. Yang, Z. Ren, and C. K. Law, *Shared low-dimensional subspaces for propagating kinetic uncertainty to multiple outputs*, Combust. Flame, 190 (2018), pp. 146–157.

[20] K. Kandasamy, G. Dasarathy, J. Schneider, and B. Póczos, *Multi-fidelity Bayesian optimisation with continuous approximations*, in International Conference on Machine Learning, International Machine Learning Society, Stroudsburg, PA, 2017, pp. 1799–1808.

[21] V. KOLTCHINSKII AND K. LOUNICI, *Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance*, Ann. Inst. Henri Poincaré, Probab. Stat., 52, 2016, pp. 1976–2013.

[22] F. KUO, R. SCHEICHL, C. SCHWAB, I. SLOAN, AND E. ULLMANN, *Multilevel quasi-Monte Carlo methods for lognormal diffusion problems*, Math. Comp., 86 (2017), pp. 2827–2860.

[23] R. LAM, *Scaling Bayesian Optimization for Engineering Design: Lookahead Approaches and Multifidelity Dimension Reduction*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2018.

[24] R. LAM, D. ALLAIRE, AND K. E. WILLCOX, *Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources*, in 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, AIAA, Reston, VA, 2015, 0143.

[25] J. LI, J. LI, AND D. XIU, *An efficient surrogate-based method for computing rare failure probability*, J. Comput. Phys., 230 (2011), pp. 8683–8697.

[26] J. LI AND D. XIU, *Evaluation of failure probability via surrogate models*, J. Comput. Phys., 229 (2010), pp. 8966–8980.

[27] K. C. LI, *Sliced inverse regression for dimension reduction*, J. Amer. Statist. Assoc., 86 (1991), pp. 316–327.

[28] C. LIEBERMAN, K. WILLCOX, AND O. GHATTAS, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.

[29] T. LUKACZYK, F. PALACIOS, J. J. ALONSO, AND P. G. CONSTANTINE, *Active subspaces for shape optimization*, in Proceedings of the 10th AIAA Multidisciplinary Design Optimization Conference, AIAA, Reston, VA, 2014, pp. 1–18.

[30] A. MARCH AND K. E. WILLCOX, *Provably convergent multifidelity optimization algorithm not requiring high-fidelity derivatives*, AIAA J., 50 (2012), pp. 1079–1089.

[31] A. A. OBERAI, N. H. GOKHALE, AND G. R. FEIJÓO, *Solution of inverse problems in elasticity imaging using the adjoint method*, Inverse Problems, 19 (2003), pp. 297–313.

[32] F. PALACIOS, J. J. ALONSO, K. DURAISAMY, M. COLONNO, J. HICKEN, A. ARANAKE, A. CAMPOS, S. COPELAND, T. D. ECONOMON, A. LONKAR, ET AL., *Stanford university unstructured (SU2): An open-source integrated computational environment for multi-physics simulation and design*, in 51st AIAA Aerospace Sciences Meeting and Exhibit, AIAA, Reston, VA, 2013.

[33] B. PEHERSTORFER, T. CUI, Y. MARZOUK, AND K. E. WILLCOX, *Multifidelity importance sampling*, Comput. Methods Appl. Mech. Engrg., 300 (2016), pp. 490–509.

[34] B. PEHERSTORFER, B. KRAMER, AND K. E. WILLCOX, *Combining multiple surrogate models to accelerate failure probability estimation with expensive high-fidelity models*, J. Comput. Phys., 341 (2017), pp. 61–75.

[35] B. PEHERSTORFER, K. WILLCOX, AND M. GUNZBURGER, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Rev., 60 (2017), pp. 550–591.

[36] B. PEHERSTORFER, K. WILLCOX, AND M. GUNZBURGER, *Optimal model management for multifidelity Monte Carlo estimation*, SIAM J. Sci. Comput., 38 (2016), pp. A3163–A3194.

[37] M. POLOCZEK, J. WANG, AND P. FRAZIER, *Multi-information source optimization*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2017, pp. 4289–4299.

[38] E. QIAN, B. PEHERSTORFER, D. O'MALLEY, V. V. VESSELINOV, AND K. WILLCOX, *Multifidelity Monte Carlo estimation of variance and sensitivity indices*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 683–706.

[39] T. M. RUSSI, *Uncertainty Quantification with Experimental Data and Complex System Models*, PhD thesis, University of California, Berkeley, CA, 2010.

[40] F. SALMOIRAGHI, F. BALLARIN, L. HELTAI, AND G. ROZZA, *Isogeometric analysis-based reduced order modelling for incompressible linear viscous flows in parametrized shapes*, Adv. Model. Simul. Eng. Sci., 3 (2016), 21.

[41] A. SALTELLI, M. RATTO, T. ANDRES, F. CAMPOLONGO, J. CARIBONI, D. GATELLI, M. SAISANA, AND S. TARANTOLA, *Global Sensitivity Analysis: The Primer*, Wiley, Chichester, England, 2008.

[42] A. M. SAMAROV, *Exploring regression structure using nonparametric functional estimation*, J. Amer. Statist. Assoc., 88 (1993), pp. 836–847.

[43] K. SWERSKY, J. SNOEK, AND R. P. ADAMS, *Multi-task Bayesian optimization*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2013, pp. 2004–2012.

[44] A. L. TECKENTRUP, P. JANTSCH, C. G. WEBSTER, AND M. GUNZBURGER, *A multilevel stochastic collocation method for partial differential equations with random input data*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1046–1074.

[45] M. TEZZELE, F. BALLARIN, AND G. ROZZA, *Combined parameter and model reduction of cardiovascular problems by means of active subspaces and POD-Galerkin methods*, in Mathematical and Numerical Modeling of the Cardiovascular System and Applications, Springer, Cham, Switzerland, 2018, pp. 185–207.

[46] R. TIPIREDDY AND R. GHANEM, *Basis adaptation in homogeneous chaos spaces*, J. Comput. Phys., 259 (2014), pp. 304–317.

[47] J. A. TROPP, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learn., 8 (2015), pp. 1–230.

[48] H. TYAGI AND V. CEVHER, *Learning non-parametric basis independent models from point queries via low-rank methods*, Appl. Comput. Harmon. Anal., 37 (2014), pp. 389–412.

[49] E. ULLMANN AND I. PAPAIOANNOU, *Multilevel estimation of rare events*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 922–953.

[50] O. ZAHM, P. CONSTANTINE, C. PRIEUR, AND Y. MARZOUK, *Gradient-based dimension reduction of multivariate vector-valued functions*, SIAM J. Sci. Comput., 42 (2020), pp. A534–A558.

[51] O. ZAHM, T. CUI, K. LAW, A. SPANTINI, AND Y. MARZOUK, *Certified Dimension Reduction in Nonlinear Bayesian Inverse Problems*, preprint, https://arxiv.org/abs/1807.03712, 2018.