Learning the Second-Moment Matrix of a Smooth Function From Point Samples

Armin Eftekhari, *Alan Turing Institute*
Michael B. Wakin, *Colorado School of Mines*
Ping Li, *Rutgers Univ.*
Paul G. Constantine, *Univ. Colorado–Boulder*
Rachel A. Ward, *Univ. Texas–Austin*

***Abstract*—Consider an open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with a probability measure μ. An important characteristic of a smooth function $f : \mathbb{D} \to \mathbb{R}$ is its *second-moment matrix* $\Sigma_\mu := \int \nabla f(x)(\nabla f(x))^* \mu(dx) \in \mathbb{R}^{n\times n}$, where $\nabla f(x) \in \mathbb{R}^n$ is the gradient of $f(\cdot)$ at $x \in \mathbb{D}$. For instance, the span of the leading r eigenvectors of Σ_μ forms an *active subspace* of $f(\cdot)$, thereby extending the concept of *principal component analysis* to the problem of *ridge approximation*. In this work, we propose and analyze a simple algorithm for estimating Σ_μ from point values of $f(\cdot)$ *without* imposing any structural assumptions on Σ_μ.**

***Index Terms*—Second moment matrix, active subspaces**

I. Introduction

Central to approximation theory, machine learning, and computational sciences in general is the task of learning a function given its finitely many point samples. More concretely, consider an open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with probability measure μ. The objective is to *learn* (approximate) a smooth function $f : \mathbb{D} \to \mathbb{R}$ from the query points $\{x_i\}_{i=1}^N \subset \mathbb{D}$, and evaluation of $f(\cdot)$ at these points [5], [12], [19], [20], [22].

An important quantity in this context is the *second-moment matrix* of $f(\cdot)$ with respect to the measure μ, defined as the $n \times n$ matrix

$$\Sigma_\mu := \mathbb{E}_x \left[\nabla f(x) \cdot (\nabla f(x))^*\right] = \int_{\mathbb{D}} \nabla f(x) \cdot (\nabla f(x))^* \ \mu(dx), \tag{1}$$

where $\nabla f(x) \in \mathbb{R}^n$ is the gradient of $f(\cdot)$ at $x \in \mathbb{D}$ and the superscript $*$ denotes vector and matrix transpose.[1] In this matrix, $\Sigma_\mu[i,j]$, the $[i,j]$th entry of Σ_μ, measures the expected product between the ith and jth partial derivatives of $f(\cdot)$. Note that Σ_μ captures key information about how $f(\cdot)$ changes along different directions. Indeed, for an arbitrary vector $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$, the *directional derivative* of $f(\cdot)$ at $x \in \mathbb{D}$ and along v is $v^*\nabla f(x)$, and the average energy of the directional derivative of $f(\cdot)$ along v and with respect to μ is $v^*\Sigma_\mu v$. Furthermore, in ridge approximation, the leading r eigenvectors of Σ_μ span an r-dimensional *active subspace* of $f(\cdot)$ with respect to the measure μ [7]. If $U_{\mu,r} \in \mathbb{R}^{n\times r}$ denotes an orthonormal basis for this active subspace, then it might be possible to reliably approximate $f(x)$ with $h(U_{\mu,r}^* x)$ for all $x \in \mathbb{D}$ and for some smooth function $h : \mathbb{R}^r \to \mathbb{R}$; one can estimate the L_2 error in such an approximation from the trailing r eigenvalues of Σ_μ [8]. Beyond approximation theory, the significance of second-moment matrices (and related concepts) across a number of other disciplines is discussed in Section IV.

The main contribution of this paper is the design and analysis of a simple algorithm to estimate the second-moment matrix Σ_μ of a smooth function $f(\cdot)$ from its point samples. This algorithm can be used, in particular, in situations where the gradients $\nabla f(x)$ appearing in (1) are not explicitly available. The key distinction of this work is the lack of any structural assumptions (such as small rank or sparsity) on Σ_μ; mild assumptions on f are specified at the beginning of Section II. Imposing a specific structure on Σ_μ can lead to more efficient algorithms, as we discuss in Section IV.

At a high level, there is a parallel between estimating the second-moment matrix of a function and estimating the covariance matrix of a random vector; our algorithm might be considered as an analogue of the standard *sample covariance matrix*, adjusted to handle missing data [17]. In this context, more efficient algorithms are available for estimating, for example, the covariance matrix with a sparse inverse [13]. In this sense, this work fills an important gap in the literature of ridge approximation and perhaps dimensionality reduction by addressing the problem in more generality.

The rest of this paper is organized as follows. The problem of learning the second-moment matrix of a function is formalized in Section II. Our approach to this problem, along with the theoretical guarantees, is described in Section III. In Section IV, we sift through a large body of literature and summarize the relevant prior art. Proofs, numerics, and technical details are available in an online preprint [11].

II. Problem Statement and Approach

Consider an open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with subspace Borel σ-algebra and probability measure μ. We assume that $f : \mathbb{D} \to \mathbb{R}$ is twice differentiable on $\mathbb{D}$, and that

$$\mathrm{L}_f := \sup_{x\in\mathbb{D}} \|\nabla f(x)\|_2 < \infty, \tag{2}$$

$$\mathrm{H}_f := \sup_{x\in\mathbb{D}} \left\|\nabla^2 f(x)\right\|_2 < \infty, \tag{3}$$

AE was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. MBW was partially supported by NSF grant CCF-1409258 and NSF CAREER grant CCF-1149225. PGC was partially supported by the DOE Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under award DE-SC0011077 and DARPA's Enabling Quantification of Uncertainty in Physical Systems. RW was partially funded by NSF CAREER Grant CCF-1255631.

[1] As suggested above, we will often suppress the dependence on $f(\cdot)$ in our notation for the sake of brevity.

where $\nabla f(x) \in \mathbb{R}^n$ and $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ are the gradient and Hessian, respectively, of $f(\cdot)$ at $x \in \mathbb{D}$, and we use the notation $\|\cdot\|_2$ to denote both the ℓ_2-norm of vectors and the spectral norm of matrices. Moreover, for $\epsilon > 0$, let $\mathbb{D}_\epsilon \subset \mathbb{D}$ denote the ϵ-interior of $\mathbb{D}$, namely $\mathbb{D}_\epsilon = \{x \in \mathbb{D} : \mathbb{B}_{x,\epsilon} \subseteq \mathbb{D}\}$. Throughout, $\mathbb{B}_{x,\epsilon} \subset \mathbb{R}^n$ denotes the (open) Euclidean ball of radius ϵ centered at x.

Consider $\Sigma_\mu \in \mathbb{R}^{n\times n}$ defined as in (1), where $\mathbb{E}_x$ computes the expectation with respect to $x \sim \mu$. Our objective in this work is to estimate Σ_μ. To that end, consider N random points drawn independently from μ and stored as the columns of $X \in \mathbb{R}^{n\times N}$. Then, it is easy to verify that

$$\dot{\Sigma}_X := \frac{1}{N}\sum_{x\in X} \nabla f(x)\cdot \nabla f(x)^*, \tag{4}$$

is an unbiased estimator for Σ_μ in (1). To interpret (4), note that we treat matrices and sets interchangeably throughout, slightly abusing the standard notation. In particular, $x \in X$ can also be interpreted as x being a column of $X \in \mathbb{R}^{n\times N}$. The following result quantifies how well $\dot{\Sigma}_X$ approximates Σ_μ. See [6] for related results.

Proposition 1. *Let $X \in \mathbb{R}^{n\times N}$ contain N independent samples drawn from the probability measure μ. Then, $\dot{\Sigma}_X$ is an unbiased estimator for $\Sigma_\mu \in \mathbb{R}^{n\times n}$ (see (1) and (4)). Moreover, except for a probability of at most n^{-1}, it holds that*

$$\left\|\dot{\Sigma}_X - \Sigma_\mu\right\|_F \lesssim \frac{\mathrm{L}_f^2 \log n}{\sqrt{N}}. \tag{5}$$

Since only point values of $f(\cdot)$ are at our disposal, we cannot compute $\dot{\Sigma}_X$ directly. Instead, we will systematically generate random points near X and then estimate $\dot{\Sigma}_X$ by aggregating local information, as detailed next.

Given the point cloud $X \subset \mathbb{D}$, fix $\epsilon > 0$, small enough so that X is a 2ϵ-separated point cloud that belongs to the ϵ-interior of $\mathbb{D}$. Formally, fix $\epsilon \le \epsilon_X$, where

$$\epsilon_X := \sup\{\epsilon' : X \subset \mathbb{D}_{\epsilon'} \text{ and } \|x - x'\|_2 \ge 2\epsilon', \\ \forall x, x' \in X,\ x \neq x'\}. \tag{6}$$

Let

$$\mathbb{B}_{X,\epsilon} := \bigcup_{x\in X} \mathbb{B}_{x,\epsilon} \subseteq \mathbb{D} \tag{7}$$

denote the ϵ-neighborhood of the point cloud X. Consider the conditional probability measure on $\mathbb{B}_{X,\epsilon}$ described as

$$\mu_{X,\epsilon} = \begin{cases} \mu/\mu(\mathbb{B}_{X,\epsilon}), & \text{inside } \mathbb{B}_{X,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{X,\epsilon}. \end{cases} \tag{8}$$

For an integer $N_{X,\epsilon}$, draw $N_{X,\epsilon}$ independent random points from $\mu_{X,\epsilon}$ and store them as the columns of $Y_{X,\epsilon} \in \mathbb{R}^{n\times N_{X,\epsilon}}$. Finally, an estimate of $\dot{\Sigma}_X$ (and in turn of Σ_μ) as a function of $X, Y_{X,\epsilon} \subset \mathbb{D}$ and $f(\cdot)$ evaluated at these points is proposed by $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ in Figure 1.

III. Theoretical Guarantees

Recalling (1) and (4), how well does $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ in Figure 1 approximate $\dot{\Sigma}_X$ and in turn Σ_μ? Parsing the answer requires introducing additional notation and imposing a certain regularity assumption on μ. All these we set out to do now, before stating the results in Section III-B.

For each $x \in X$, let the columns of $Y_{x,\epsilon} \in \mathbb{R}^{n\times N_{x,\epsilon}}$ contain the ϵ-neighbors of x in $Y_{X,\epsilon}$. In our notation, this can be written as

$$Y_{x,\epsilon} := Y_{X,\epsilon} \cap \mathbb{B}_{x,\epsilon}, \qquad \#Y_{x,\epsilon} = N_{x,\epsilon}. \tag{11}$$

Because $\epsilon \le \epsilon_X$ is small (see (6)), these neighborhoods do not intersect, that is

$$Y_{x,\epsilon} \cap Y_{x',\epsilon} = \emptyset, \quad \forall x, x' \in X, \quad x \neq x';$$

therefore, $Y_{X,\epsilon}$ is simply partitioned into $\#X = N$ subsets $\{Y_{x,\epsilon}\}_{x\in X}$. Observe also that, conditioned on $x \in X$ and $N_{x,\epsilon}$, each neighbor $y \in Y_{x,\epsilon}$ follows the conditional probability measure described as follows:

$$y|x, N_{x,\epsilon} \sim \mu_{x,\epsilon} := \begin{cases} \mu/\mu(\mathbb{B}_{x,\epsilon}), & \text{inside } \mathbb{B}_{x,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{x,\epsilon}. \end{cases} \tag{12}$$

A. Regularity of μ

Here we introduce the regularity condition imposed on μ.

Assumption 1. (Local near-isotropy of μ) *Throughout this paper, we assume that there exist $\epsilon_\mu, \mathrm{K}_\mu > 0$ such that for all $\epsilon \le \epsilon_\mu$, the following requirement holds for any arbitrary ϵ-interior point $x \in \mathbb{D}_\epsilon$.*

Given x, draw y from the conditional measure on the ϵ-neighborhood of x, namely $y|x \sim \mu_{x,\epsilon}$ with $\mu_{x,\epsilon}$ defined in (12). Then, for every $\gamma_1 \ge 0$ and arbitrary (but fixed) $v \in \mathbb{R}^n$, it holds that

$$\Pr_{y|x}\left[\|P_{x,y}\cdot v\|_2^2 > \gamma_1 \cdot \frac{\|v\|_2^2}{n}\right] \lesssim e^{-\mathrm{K}_\mu \gamma_1}, \tag{13}$$

where

$$P_{x,y} := \frac{(y-x)(y-x)^*}{\|y-x\|_2^2} \in \mathbb{R}^{n\times n}$$

is the orthogonal projection onto the direction of $y-x$. Above, $\Pr_{y|x}[\cdot] = \Pr_y[\cdot|x]$ stands for conditional probability.

Roughly speaking, under Assumption 1, μ is locally isotropic. Indeed, this assumption is met when μ is the uniform probability measure on $\mathbb{D}$, as shown in [11].

Given the point cloud X, we also set $\epsilon_{\mu,X} := \min[\epsilon_\mu, \epsilon_X]$.

B. Theoretical Guarantees

In Theorems 1 and 2, for a fixed point cloud X, we focus on how well $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ in Figure 1 approximates $\dot{\Sigma}_X$. Then, in the ensuing remarks, we remove the conditioning on X, using Proposition 1 to see how well $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ approximates Σ_μ.

Theorem 1 states that $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ can be a nearly unbiased estimator of $\dot{\Sigma}_X$ given X (see (4)). Throughout, $\mathbb{E}_{z_1|z_2}[\cdot] = \mathbb{E}_{z_1}[\cdot|z_2]$ stands for conditional expectation over z_1 and conditioned on z_2 for random variables z_1, z_2.

Estimating the second-moment matrix

Input:

- Open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with probability measure μ.
- An oracle that returns $f(x)$ for a query point $x \in \mathbb{D}$.
- Neighborhood radius $\epsilon > 0$, sample sizes N, $N_{X,\epsilon}$, and integer $N_{X,\min,\epsilon} \leq N_{X,\epsilon}$.

Output:

- $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$, as an estimate of Σ_μ.

Body:

- Draw N random points independently from μ and store them as the columns of $X \in \mathbb{R}^{n \times N}$.
- Draw $N_{X,\epsilon}$ random points independently from $\mu_{X,\epsilon}$ and store them as the columns of $Y_{X,\epsilon} \in \mathbb{R}^{n \times N_{X,\epsilon}}$. Here, $\mu_{X,\epsilon}$ is the conditional probability measure induced by μ on $\mathbb{B}_{X,\epsilon} = \cup_{x \in X} \mathbb{B}_{x,\epsilon}$. In turn, $\mathbb{B}_{x,\epsilon} \subset \mathbb{R}^n$ is the Euclidean ball of radius ϵ about x. Partition $Y_{X,\epsilon}$ according to X by setting $Y_{x,\epsilon} = Y_{X,\epsilon} \cap \mathbb{B}_{x,\epsilon}$, so that $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$ contains all ϵ-neighbors of x in $Y_{X,\epsilon}$.
- Compute and return

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N}\left(1 + \frac{1-\frac{2}{n}}{1+\frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \cdot \left(\sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \frac{\left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2}{\left(1+\frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right), \quad (9)$$

where I_n denotes the $n \times n$ identity matrix, and

$$\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y)-f(x)}{\|y-x\|_2} \cdot \frac{y-x}{\|y-x\|_2} \in \mathbb{R}^n. \quad (10)$$

Fig. 1. The proposed algorithm for estimating the second-moment matrix of the function $f(\cdot)$ with respect to the measure μ.

Theorem 1. (Bias) *Consider an open set $\mathbb{D} \subseteq \mathbb{R}^n$ equipped with probability measure μ satisfying Assumption 1, and consider a twice differentiable function $f : \mathbb{D} \to \mathbb{R}$ satisfying (2),(3). Assume that the (fixed) columns of $X \in \mathbb{R}^{n \times N}$ belong to $\mathbb{D}$. Fix also $\epsilon \in (0, \epsilon_{\mu,X}]$. For an integer N and integers $N_{X,\epsilon} \geq N$ and $N_{X,\min,\epsilon} \leq N_{X,\epsilon}$, assume also that*

$$N_{X,\epsilon} \geq \max\left(\frac{N_{X,\min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, n^{\frac{1}{20}}\right) \text{ and } N_{X,\min,\epsilon} \gtrsim \log^2 N, \quad (14)$$

where

$$\rho_{\mu,X,\epsilon} := N \cdot \min_{x \in X} \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}. \quad (15)$$

Then, the estimator $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ defined in (9) satisfies

$$\left\| \mathbb{E}_{Y_{X,\epsilon}|X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X \right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + n^2 \mathrm{L}_f^2 N^{-10} + \epsilon^2 \mathrm{H}_f^2 n^2 + \epsilon \mathrm{L}_f \mathrm{H}_f n^{3/2} \max(\mathrm{K}_\mu^{-1/2}, 1) \log^{\frac{1}{2}} N_{X,\epsilon}, \quad (16)$$

where $\mathrm{B}_{\mu,\epsilon}$ is defined explicitly in [11].

A few remarks are in order.

Remark 1. (Discussion) Theorem 1 describes how well $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ approximates $\dot{\Sigma}_X$, in expectation. To form a better understanding of this result, let us first study the conditions listed in (14).

- The quantity $\rho_{\mu,X,\epsilon}$, which appears in (14) and is defined in (15), reflects the non-uniformity of μ over the set $\mathbb{D}$. In particular, if $\mathbb{D} \subset \mathbb{R}^n$ is bounded and μ is the uniform probability measure on $\mathbb{D}$, then $\rho_{\mu,X,\epsilon} = 1$. Non-uniform measures could yield $\rho_{\mu,X,\epsilon} < 1$.
- The requirements on $N_{X,\epsilon}$ and $N_{X,\min,\epsilon}$ in (14) ensure that every $x \in X$ has sufficiently many neighbors in $Y_{X,\epsilon}$, i.e., $N_{x,\epsilon}$ is large enough for all x. For example, if μ is the uniform probability measure on $\mathbb{D}$ and $\rho_{\mu,X,\epsilon} = 1$, we might take $N_{X,\min,\epsilon} \approx \log^2 N$ so that (16) holds with a total of $N_{X,\epsilon} = O(N \log^2 N)$ samples. Here, $O(\cdot)$ is the standard Big-O notation.

Let us next interpret the bound on the bias in (16).

- The first term on the right-hand side of (16), $\mathrm{B}_{\mu,\epsilon}$, is given explicitly and further discussed in [11]; it depends on both the probability measure μ and the function $f(\cdot)$, and it can also be viewed as a measure of the non-uniformity of μ. In fact, in the special case where μ is the uniform probability measure on a bounded and open set $\mathbb{D}$ and every $x \in X$ has the same number of neighbors $N_{x,\epsilon} = N_{X,\epsilon}/N$ within $Y_{X,\epsilon}$, then conditioned on this event, (16) can in fact be sharpened by replacing the definition of $\mathrm{B}_{\mu,\epsilon}$ simply with $\mathrm{B}_{\mu,\epsilon} = 0$. In general, the more isotropic μ is (in the sense described in Assumption 1), the smaller $\mathrm{B}_{\mu,\epsilon}$ will be.
- The second term on the right-hand side of (16) is negligible, as we will generally have $N \gtrsim n^2$.
- The third and fourth terms on the right-hand side of (16) can be made arbitrarily small by choosing ϵ appropriately small (as a function of L_f, H_f, n, K_μ, and $N_{X,\epsilon}$). In computational applications, however, choosing ϵ too small could raise concerns about numerical precision.

- To get a sense of when the bias in (16) is small relative to the size of Σ_μ, it may be appropriate to normalize (16). A reasonable choice would be to divide both sides of (16) by L_f^2, where L_f (defined in (2)) bounds $\|\nabla f(x)\|_2$ on $\mathbb{D}$. In particular, such a normalization accounts for the possible scaling behavior of $\|\Sigma_\mu\|_F$ if one were to consider a sequence of problems with n increasing. For example, in the case where n increases but the new variables in the domain of $f(\cdot)$ do not affect its value, then L_f^2 and $\|\Sigma_\mu\|_F$ are both constant. On the other hand, in the case where n increases and $f(\cdot)$ depends uniformly on the new variables, then L_f^2 and $\|\Sigma_\mu\|_F$ both increase with n. In any case, one can show that $\|\Sigma_\mu\|_F \leq L_f^2$. With this choice of normalization, the second, third, and fourth terms on the right-hand side of (16) can still be made arbitrarily small as described above. In the special case where μ is uniform on $\mathbb{D}$ and every $x \in X$ has the same number of neighbors $N_{x,\epsilon} = N_{X,\epsilon}/N$, the first term on the right-hand side of (16) remains zero, as also described above. More generally, however, $\mathrm{B}_{\mu,\epsilon} \,/\, \mathrm{L}_f^2$ will contain a term that scales like $n^{1/2} N_{X,\min,\epsilon}^{-1}$, and to control this term it is necessary to choose $N_{X,\min,\epsilon} \gtrsim n^{1/2} \log^2 N$ so that (14) is also satisfied. We revisit the impact of n on the choices of N and $N_{X,\epsilon}$ after presenting Theorem 2 below.

Our second result is a finite-sample bound for $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$.

Theorem 2. (Finite-sample bound) *Under the same setup as in Theorem 1 including the conditions in* (14)*, it holds that*

$$\begin{aligned}\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X\right\|_F &\lesssim \frac{1}{2}\epsilon^2 \,\mathrm{H}_f^2\, n^2 + 2\epsilon\,\mathrm{L}_f\,\mathrm{H}_f\, n^2 + \mathrm{B}_{\mu,\epsilon} \\ &+ \log^3(N_{X,\epsilon}) \cdot \frac{n^{3/2}}{\sqrt{\min[\rho_{\mu,X,\epsilon} N_{X,\epsilon}, N]}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-1/2}]\,\mathrm{L}_f^2 \\ &+ 4n^2 L_f^2 N_{X,\epsilon}^{-10} \end{aligned} \tag{17}$$

except with a probability of $O\left(N^{-10}\right)$*. Here, the probability is with respect to the selection of* $Y_{X,\epsilon}$*, conditioned on the fixed set* X*.*

Remark 2. (Discussion) Theorem 2 states that $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ can reliably estimate $\dot{\Sigma}_X$ with high probability. We offer several remarks to help interpret this result.

- The conditions in (14) were discussed in Remark 1.
- Let us now dissect the estimation error, namely the right-hand side of (17). As discussed in Remark 1, $\mathrm{B}_{\mu,\epsilon}$ in effect captures the non-uniformity of μ. In particular, the right-hand side of (17) can be sharpened by setting $\mathrm{B}_{\mu,\epsilon} = 0$ in the setting described in that remark.
- Similar to Remark 1, the terms involving ϵ in (17) can be made negligible by choosing ϵ to be suitably small. We omit these terms in the discussion below.
- The fourth term on the right hand side of (17) can be controlled by making N and $N_{X,\epsilon}$ suitably large. We discuss this point further below.
- The final term on the right hand side of (17) is negligible and we omit this in our discussion below.

Remark 3. (Estimating Σ_μ) Combining Theorem 2 with Proposition 1 (and omitting negligible terms) yields

$$\begin{aligned}\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu\right\|_F &\lesssim \mathrm{B}_{\mu,\epsilon} + \frac{\mathrm{L}_f^2 \log n}{\sqrt{N}} \\ &+ \log^3(N_{X,\epsilon}) \cdot \frac{n^{3/2}}{\sqrt{\min[\rho_{\mu,X,\epsilon} N_{X,\epsilon}, N]}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-1/2}]\,\mathrm{L}_f^2, \end{aligned} \tag{18}$$

with high probability when both X and $Y_{X,\epsilon}$ are selected randomly, therefore quantifying how well the full algorithm in Figure 1 estimates the second-moment matrix of $f(\cdot)$.

As suggested in Remark 1, we can normalize this bound by dividing both sides by L_f^2:

$$\begin{aligned}\frac{\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu\right\|_F}{L_f^2} &\lesssim \frac{\mathrm{B}_{\mu,\epsilon}}{L_f^2} + \frac{\log n}{\sqrt{N}} \\ &+ \log^3(N_{X,\epsilon}) \cdot \frac{n^{3/2}}{\sqrt{\min[\rho_{\mu,X,\epsilon} N_{X,\epsilon}, N]}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-1/2}]. \end{aligned} \tag{19}$$

We discuss the terms appearing on the right hand side of (19):

- As described in Remark 1, in some settings $\mathrm{B}_{\mu,\epsilon} \,/\, \mathrm{L}_f^2$ will be zero, while in other settings controlling $\mathrm{B}_{\mu,\epsilon} \,/\, \mathrm{L}_f^2$ will require choosing $N_{X,\min,\epsilon} \gtrsim n^{1/2} \log^2 N$.
- The third term in (19) dominates the second and therefore dictates the convergence rate of the error as the number of samples increases. In particular, setting $N_{X,\epsilon}$ proportional to $N \log^2(nN)$ gives $N \approx N_{X,\epsilon} / \log^2(nN)$ and an overall convergence rate (perhaps to a nonzero bias $\mathrm{B}_{\mu,\epsilon} \,/\, \mathrm{L}_f^2$) of $\log^4(N_{X,\epsilon})/\sqrt{N_{X,\epsilon}}$ as the number $N_{X,\epsilon}$ of secondary samples (which dominates the total) increases. Up to log terms, this is the same as the convergence rate appearing in Proposition 1 where perfect knowledge of gradients was available.
- As a function of the ambient dimension n, controlling the third term in (19) will require ensuring that N scales like n^3. In cases where $N_{X,\epsilon}$ is set proportional to $N \log^2(nN)$ the overall number of samples therefore also must scale like n^3 (neglecting log factors). In cases where $N_{X,\min,\epsilon} \gtrsim n^{1/2} \log^2 N$, however, the overall number of samples must scale like $n^{3.5}$.

IV. Related Work

As argued in Section I, the second-moment matrix (or its leading eigenvectors) is of particular relevance in the context of ridge approximation. A ridge function $f(\cdot)$ is one for which $f(x) = h(A^* x)$ for all $x \in \mathbb{D}$, where A is an $n \times r$ matrix with $r < n$ and $h : \mathbb{R}^r \to \mathbb{R}$. Such a function varies only along the r-dimensional subspace spanned by the columns of A and is constant along directions in the $(n-r)$-dimensional orthogonal complement of this subspace. A large body of work exists in the literature of approximation theory on learning ridge functions from point samples. Most of these works focus on finding an approximation to the underlying function h and/or the dimensionality-reducing matrix A (or its column span).

When $f(\cdot)$ is a ridge function, the r-dimensional column span of A coincides with the span of the eigenvectors of Σ_μ, which will have rank r. This illuminates the connection between ridge approximation and second-moment matrices.

In [12], the authors develop an algorithm to learn the column span of A when its basis vectors are (nearly) sparse. The sparsity assumption was later removed in [4], [21] and replaced with an assumption that this column span is low-dimensional (r is small). For learning such a low-dimensional subspace, these models allow for algorithms with better sample complexities compared to Theorem 2 which, in contrast, provides a guarantee on learning the entire second-moment matrix Σ_μ and holds without any assumption (such as low rank) on Σ_μ. For completeness, we note that it is natural to ask whether the results in [21] could simply be applied in the "general case" where the subspace dimension r approaches the ambient dimension n (thus relaxing the critical structural assumption in [21]). As detailed in Section 5 of [21], however, the sampling complexity in this general case will scale with n^5 (ignoring log factors). In contrast, (18) requires only that the total number of function samples $N + N_{X,\epsilon}$ scale with n^3.

A ridge-like function is one for which $f(x) \approx h(A^*x)$. The framework of *active subspaces* provides a mechanism for detecting ridge-like structure in functions and reducing the dimensionality of such functions [6], [7], [9]. For example, in scientific computing $f(x)$ may represent the scalar-valued output of some complicated simulation that depends on a high-dimensional input parameter x. By finding a suitable $n \times r$ matrix A, one can reduce the complexity of parameter studies by varying inputs only in the r-dimensional column space of A^*. The term *active subspace* refers to the construction of A via the r leading eigenvectors of Σ_μ.

In high-dimensional statistics and machine learning, similar structures arise in the task of regression, where given a collection of data pairs (x_i, z_i), the objective is to construct a function $z = f(x)$ that is a model for the relationship between x and z. One line of work in this area is *projection pursuit* where, spurred by the interest in *generalized additive models* [15], the aim is to construct $f(\cdot)$ using functions of the form $\sum_i h_i(a_i^* x)$ [10], [14], [16]. *Sufficient dimension reduction* and related topics are still other lines of related work in statistics. In this context, a collection of data pairs (x_i, z_i) are observed having been drawn independently from some unknown joint density. The assumption is that z is conditionally independent of x, given A^*x for some $n \times r$ matrix A. The objective is then to estimate the column span of A, known as the *effective subspace for regression* in this literature; see, e.g., [1] for a review.

Finding the second-moment matrix of a function is also closely related to covariance estimation (see (1)), which is widely studied in modern statistics often under various structural assumptions on the covariance matrix, e.g., sparsity of its inverse. In this context, it appears that [2], [3], [18] are the most relevant to the present work, in part because of their lack of any structural assumptions. For the sake of brevity, we focus on [3], which offers an unbiased estimator for the covariance matrix of a random vector x given few measurements of multiple realizations of x in the form of $\{\Phi_i x_i\}_i$ for low-dimensional (and uniformly random) orthogonal projection matrices $\{\Phi_i\}_i$. It is important to point out that, by design, the estimator in [3] is not applicable to our setup.[2] Our framework might be interpreted as sum of rank-1 projections. To further complicate matters, the probability measure μ on $\mathbb{D}$ is not necessarily uniform; we cannot hope to explicitly determine the distribution of the crucial components of the estimator. Instead, we rely on the standard tools in empirical processes to control the bounds.

REFERENCES

[1] K. P. Adragni and R. D. Cook. Sufficient dimension reduction and prediction in regression. *Philos. Trans. Royal Soc. A*, 367(1906), 2009.

[2] F. P Anaraki and S. Hughes. Memory and computation efficient PCA via very sparse random projections. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 1341–1349, 2014.

[3] M. Azizyan, A. Krishnamurthy, and A. Singh. Extreme compressive sampling for covariance estimation. *arXiv preprint arXiv:1506.00898*, 2015.

[4] I. Bogunovic, V. Cevher, J. Haupt, and J. Scarlett. Active learning of self-concordant like multi-index functions. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, pages 2189–2193, 2015.

[5] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2012.

[6] P. Constantine and D. Gleich. Computing active subspaces with Monte Carlo. *arXiv preprint arXiv:1408.0545v2*, 2015.

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

[8] P. G. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4), 2014.

[9] P. G. Constantine, A. Eftekhari, J. Hokanson, and R. A. Ward. A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engineering*, 326, 2017.

[10] D. L. Donoho and I. M. Johnstone. Projection-based approximation and a duality with kernel methods. *Ann. Stat.*, pages 58–106, 1989.

[11] A. Eftekhari, M. B. Wakin, P. Li, and P. G. Constantine. Learning the second-moment matrix of a smooth function from point samples. *arXiv preprint arXiv:1612.06339*, 2016.

[12] M. Fornasier, K. Schnass, and J. Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[14] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

[15] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Taylor & Francis, 1990.

[16] P. J. Huber. Projection pursuit. *Ann. Stat.*, pages 435–475, 1985.

[17] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[18] F. Pourkamali-Anaraki. Estimation of the sample covariance matrix from compressive measurements. *arXiv preprint arXiv:1512.08887*, 2015.

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

[20] J. F. Traub and A. G. Werschulz. *Complexity and Information*. Cambridge University Press, 1998.

[21] H. Tyagi and V. Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, 2014.

[22] H. Wendland. *Scattered data approximation*, volume 17. Cambridge University Press, 2004.

[2] The use of finite differences will effectively replace $\Phi_t x_t$ in $\widehat{\Sigma}_1$ in [3, Section 3] with a sum of rank-1 projections of x_t.