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

Armin Eftekhari, Michael B. Wakin, Ping Li, Paul G. Constantine*

May 25, 2018

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 $x \in \mathbb{D}$ and $*$ stands for transpose. For instance, the span of the leading r eigenvectors of Σ_μ forms an *active subspace* of $f(\cdot)$, which contains the directions along which $f(\cdot)$ changes the most and is of particular interest in *ridge approximation*. In this work, we propose a simple algorithm for estimating Σ_μ from random point evaluations of $f(\cdot)$ *without* imposing any structural assumptions on Σ_μ. Theoretical guarantees for this algorithm are established with the aid of the same technical tools that have proved valuable in the context of covariance matrix estimation from partial measurements.

1 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 [1, 2, 3, 4, 5, 6].

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

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

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] The $[i, j]$th entry of this matrix, namely $\Sigma_\mu[i, j]$, 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 it is easy to check that the directional derivative of $f(\cdot)$ along v, itself a scalar function on $\mathbb{D}$, has the average energy of $v^* \Sigma_\mu v$ with respect to the measure μ. The directions with the most energy, that is the directions along which $f(\cdot)$ changes the most on average, are particularly important in *ridge approximation*, where we are interested in approximating (the possibly complicated function) $f(\cdot)$ with a (simpler) ridge function. More specifically, the leading r eigenvectors of Σ_μ span an r-dimensional *active subspace* of $f(\cdot)$ with respect to the measure μ [7], which contains the directions along which $f(\cdot)$ changes the most. 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 $h : \mathbb{R}^r \to \mathbb{R}$. In this sense, we might think of ridge approximation and active subspaces as the extensions of, respectively,

*AE is with the Alan Turing Institute in London. MBW is with the Electrical Engineering department at the Colorado School of Mines. PL is with the Statistics and Computer Science departments at Rutgers University. PGC is with Computer Science department at the University of Colorado-Boulder.

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

dimensionality reduction and principal components to high-dimensional functions. Beyond approximation theory, the significance of second-moment matrices (and related concepts) across a number of other disciplines is discussed in Section 4.

With this introduction, the main objective of this paper is the following, which will be made precise later in Section 2.

> **Objective:** *Design query points* $\{x_i\}_{i=1}^N$ *and learn from* $\{x_i, f(x_i)\}_{i=1}^N$ *the second-moment matrix of* $f(\cdot)$ *with respect to the measure* μ.

We must emphasize that we impose *no structural assumptions* on the second-moment matrix (such as being low rank or sparse), a point that we shall revisit later in Section 4. Our approach to this problem, alongside the results, is summarized next with minimal details for better accessibility. A rigorous account of the problem and our approach is then presented in Sections 2 and 3.

1.1 Approach

We assume in this paper that points in the domain $\mathbb{D}$ are observed randomly according to the probability measure μ. In particular, consider N random points drawn independently from μ and stored as the columns of a matrix $X \in \mathbb{R}^{n\times N}$. It is then easy to verify [8] that

$$\dot{\Sigma}_X := \frac{1}{N}\sum_{x\in X} \nabla f\,(x)\cdot \nabla f\,(x)^* \tag{2}$$

is an unbiased estimator of Σ_μ in (1).[2] In fact, a standard large deviation analysis reveals that $\|\dot{\Sigma}_X - \Sigma_\mu\| \propto \frac{1}{\sqrt{N}}$, with overwhelming probability and for any matrix norm $\|\cdot\|$.

Since we furthermore assume that only the point values of $f(\cdot)$ are at our disposal (rather than its gradients), it is not possible to directly calculate $\dot{\Sigma}_X$ as in (2). Thus, one might resort to using finite difference approximations of the partial derivatives, as we sketch here and formalize in Section 2. Our procedure for estimating the second-moment matrix of $f(\cdot)$ will in fact rely not only on $\{x_i, f(x_i)\}_{i=1}^N$ but also on a supplementary set of points (also drawn randomly) nearby those in X. In particular, for a sufficiently small $\epsilon > 0$ and arbitrary x, let $\mathbb{B}_{x,\epsilon}$ denote the Euclidean ball of radius ϵ about x, and set

$$\mathbb{B}_{X,\epsilon} = \bigcup_{x\in X} \mathbb{B}_{x,\epsilon}.$$

Let also $\mu_{X,\epsilon}$ be the conditional probability measure on $\mathbb{B}_{X,\epsilon}$ induced by μ. Consider $N_{X,\epsilon}$ random points drawn independently from $\mu_{X,\epsilon}$ and stored as the columns of $Y_{X,\epsilon} \in \mathbb{R}^{n\times N_{X,\epsilon}}$. Then 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 $Y_{X,\epsilon}$. This setup is illustrated in Figure 1.

For every $x \in X$, consider $\dot{\nabla}_{Y_{x,\epsilon}} f(x) \in \mathbb{R}^n$ as an estimate of the true gradient $\nabla f(x)$, where

$$\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y\in Y_{x,\epsilon}} \frac{f(y)-f(x)}{\|y-x\|_2}\cdot\frac{y-x}{\|y-x\|_2}; \tag{3}$$

the scaling with n will be shortly justified. Then we could naturally consider $\dot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n\times n}$ as an estimate of $\dot{\Sigma}_X$ in (2), and in turn an estimate of Σ_μ in (1), where

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

Algorithm 1, in fact, introduces a better estimate of $\dot{\Sigma}_X$, denoted throughout by $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$, which has a smaller bias than $\dot{\Sigma}_{X,Y_{X,\epsilon}}$. Indeed Theorem 1 in Section 3, roughly speaking, establishes that

$$\left\|\mathbb{E}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \Sigma_\mu\right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + \epsilon n^{3/2},$$

[2] As indicated above, we slightly abuse the standard notation by treating matrices and sets interchangeably. For example, the expression $x \in X$ can also be interpreted as x being a column of matrix X.

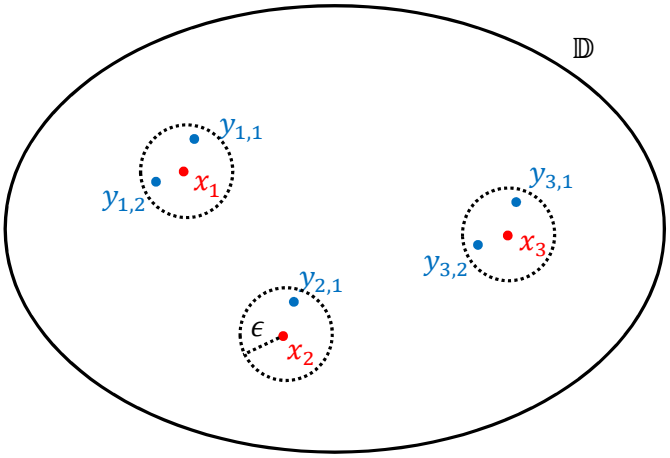

Figure 1: Visualization of the problem setup. The probability measure μ is supported on the domain $\mathbb{D} \subseteq \mathbb{R}^n$ of a smooth function $f(\cdot)$. Here, $N = 3$ and $X = \{x_i\}_{i=1}^N$ are drawn independently from μ. For sufficiently small ϵ, we let $\mathbb{B}_{x_i,\epsilon}$ denote the ϵ-neighborhood of each x_i and set $\mathbb{B}_{X,\epsilon} = \cup_{i=1}^N \mathbb{B}_{x_i,\epsilon}$. On $\mathbb{B}_{X,\epsilon}$, μ induces the conditional measure $\mu_{X,\epsilon}$, from which $N_{X,\epsilon}$ points are independently drawn and collected in $Y_{X,\epsilon} = \{y_{ij}\}_{i,j}$. Here, x_1 has $N_{x_1,\epsilon} = 2$ neighbors in $Y_{X,\epsilon}$ and we set $Y_{x_1,\epsilon} = \{y_{1,j}\}_{j=1}^{N_{x_1,\epsilon}}$. Similarly, $Y_{x_2,\epsilon}$ and $Y_{x_3,\epsilon}$ are formed. Note that $Y_{X,\epsilon} = \cup_{i=1}^N Y_{x_i,\epsilon}$. Our objective is to estimate the second-moment matrix of $f(\cdot)$ (with respect to the probability measure μ) given $\{x_i, f(x_i)\}$ and $\{y_{ij}, f(y_{ij})\}$.

where the expectation is over $X, Y_{X,\epsilon}$ and $\|\cdot\|_F$ stands for the Frobenius norm. Throughout, we will use $\lesssim$ and similarly $\gtrsim, \approx$ to suppress universal constants and simplify the presentation. Above, the quantity $\mathrm{B}_{\mu,\epsilon}$ depends in a certain way on the regularity of the measure μ and function $f(\cdot)$, with the dependence on $f(\cdot)$ suppressed as usual. Moreover, loosely speaking, it holds true that

$$\left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu \right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + \epsilon n^2 + \frac{n}{\sqrt{N_{X,\epsilon}}}, \tag{5}$$

with high probability, as described in Theorem 2 in Section 3.

Before turning to the details, consider also the following numerical example. Suppose that $n = 100, \epsilon = 0.01$. Let $\mathbb{D}$ be the unit sphere in $\mathbb{R}^n$ and take μ to be the uniform probability measure on $\mathbb{D}$. Also take $f(x) = (x[1]^2 + x[2]^2)/2$. For N ranging from 1 to 200 and with $N_{X,\epsilon} \approx 2N_{X,\min,\epsilon} \cdot N = 2\log(N) \cdot N$, we compute $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ following the procedure given in Algorithm 1. The estimation error, as measured by the left-hand side of (5), is plotted in Figure 2, after averaging over 100 trials. Note that, on average, each sample $x \in X$ has only $2\log N \leq 11 \ll n = 100$ neighbors in $Y_{X,\epsilon}$. This is far less than what would be required with a deterministic sampling scheme that used $n = 100$ neighbors for each $x \in X$ to estimate the local gradient $\nabla f(x)$ via a conventional finite difference approximation, and unlike such a deterministic scheme, it can be implemented when data is only available randomly according to the distribution μ. Observe also that the error decays roughly like $1/\sqrt{N_{X,\epsilon}}$, as suggested by (5) and detailed later in Theorem 2.

1.2 Contribution and Organization

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; see (1) and Algorithm 1. As argued earlier and also in Section 4, Σ_μ is a key quantity in ridge approximation and a number of related problems.

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 2. Imposing a specific structure on Σ_μ can lead to more efficient algorithms as we discuss in Section 4.

Algorithm 1 for estimating the second-moment matrix of the function $f(\cdot)$ with respect to the measure μ

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

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \left(1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \cdot \left(\sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \frac{\left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2}{\left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right), \quad (6)$$

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

$$\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2} \in \mathbb{R}^n. \quad (7)$$

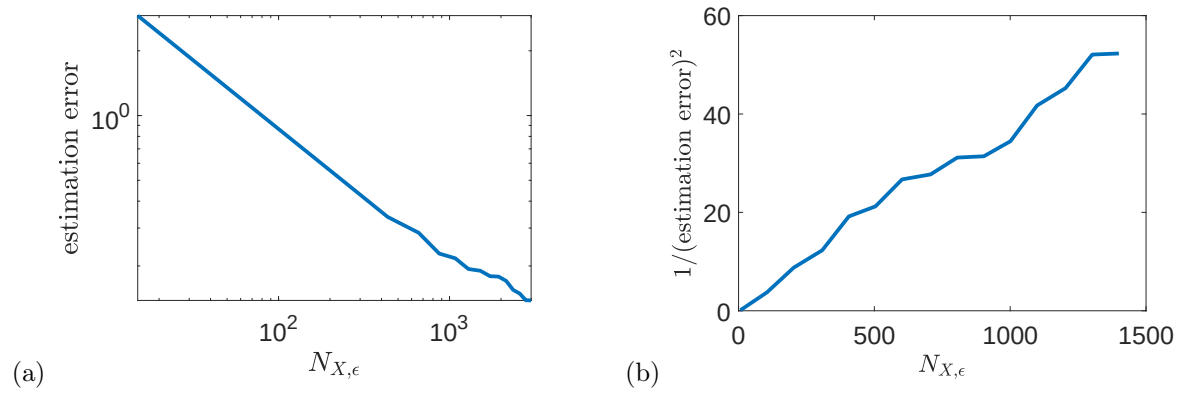

Figure 2: A numerical example for the proposed algorithm in Algorithm 1; see the last paragraph of Section 1.1 for details. (a) The average estimation error, as measured by the left-hand side of (5), plotted versus the number of secondary samples used $N_{X,\epsilon}$. (b) The same estimation error, squared and inverted, to emphasize the convergence rate given in (5).

At a very high level, there is indeed a parallel between estimating the second-moment matrix of a function from random point samples and estimating the covariance matrix of a random vector; Algorithm 1 in a sense produces an analogue of the *sample covariance matrix*, adjusted to handle missing data [9]. In this context, more efficient algorithms are available for estimating, for example, the covariance matrix with a sparse inverse [10]. In this sense, we feel that 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 2. Our approach to this problem, stated more formally, along with the theoretical guarantees, are described in Section 3. In Section 4, we sift through a large body of literature and summarize the relevant prior work. Proofs and technical details are deferred to Section 5 and the appendices.

2 Problem Statement and Approach

In this section, we formalize the problem outlined in Section 1. Consider an open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with subspace Borel σ-algebra and probability measure μ. We assume throughout that $f : \mathbb{D} \to \mathbb{R}$ is twice differentiable on $\mathbb{D}$, and that

$$\mathrm{L}_f := \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2 < \infty, \tag{8}$$

$$\mathrm{H}_f := \sup_{x \in \mathbb{D}} \left\|\nabla^2 f(x)\right\|_2 < \infty, \tag{9}$$

where $\nabla f(x) \in \mathbb{R}^n$ and $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ are the gradient and Hessian of $f(\cdot)$ at $x \in \mathbb{D}$, respectively, 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

$$\Sigma_\mu := \mathbb{E}_x \left[\nabla f(x) \cdot \nabla f(x)^*\right] = \int_{\mathbb{D}} \nabla f(x) \cdot \nabla f(x)^* \, \mu(dx), \tag{10}$$

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, as noted in Section 1.1, it is easy to verify that

$$\dot{\Sigma}_X := \frac{1}{N} \sum_{x \in X} \nabla f(x) \cdot \nabla f(x)^*, \tag{11}$$

is an unbiased estimator for Σ_μ in (10). To interpret (11), recall also 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. Its proof is included in Appendix B for completeness; see [8] for related results concerning the accuracy of $\dot{\Sigma}_X$ as an estimate of Σ_μ.

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 (10) and (11). Moreover, except for a probability of at most n^{-1}, it holds that*

$$\left\|\dot{\Sigma}_X - \Sigma_\mu\right\|_F \lesssim \frac{\mathrm{L}_f^2 \log n}{\sqrt{N}}. \tag{12}$$

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 the point cloud 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

$$\epsilon_X := \sup\left\{\epsilon' \,:\, X \subset \mathbb{D}_{\epsilon'} \text{ and } \|x - x'\|_2 \ge 2\epsilon',\ \forall x, x' \in X,\ x \ne x'\right\}. \tag{13}$$

Let

$$\mathbb{B}_{X,\epsilon} := \bigcup_{x\in X} \mathbb{B}_{x,\epsilon} \subseteq \mathbb{D} \tag{14}$$

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

$$\mu_{X,\epsilon} = \begin{cases} \mu/\mu\left(\mathbb{B}_{X,\epsilon}\right), & \text{inside } \mathbb{B}_{X,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{X,\epsilon}. \end{cases} \tag{15}$$

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 evaluations of $f(\cdot)$ at these points is proposed by $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ in Algorithm 1.

3 Theoretical Guarantees

Recalling (10) and (11), how well does $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ in Algorithm 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 3.2.

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

$$Y_{x,\epsilon} := Y_{X,\epsilon} \cap \mathbb{B}_{x,\epsilon}, \qquad \#Y_{x,\epsilon} = N_{x,\epsilon}. \tag{16}$$

Because $\epsilon \le \epsilon_X$ is small, see (13), these neighborhoods do not intersect, that is

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

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:

$$y|x, N_{x,\epsilon} \sim \mu_{x,\epsilon} := \begin{cases} \mu/\mu\left(\mathbb{B}_{x,\epsilon}\right), & \text{inside } \mathbb{B}_{x,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{x,\epsilon}. \end{cases} \tag{17}$$

3.1 Regularity of μ

In order to introduce the regularity condition imposed on μ here, consider first the special case where the domain $\mathbb{D} \subset \mathbb{R}^n$ is bounded and μ is the uniform probability measure on $\mathbb{D}$. Then, for $\epsilon > 0$ and arbitrary ϵ-interior point $x \in \mathbb{D}_\epsilon$, the conditional measure $\mu_{x,\epsilon}$ too is the uniform measure on $\mathbb{B}_{x,\epsilon}$, see (17). Draw y from $\mu_{x,\epsilon}$, that is, $y|x \sim \mu_{x,\epsilon}$ in our notation. Then it is easy to verify that $y - x$ is an *isotropic* random vector, namely

$$\mathbb{E}_{y|x}\left[(y-x)(y-x)^*\right] = C \cdot I_n,$$

for some factor C.[3] Above, $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix and $\mathbb{E}_{y|x}[\cdot] = \mathbb{E}_y[\cdot|x]$ stands for conditional expectation, given x. A similar property plays an important role in this paper, as captured by Assumption 1 below.

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 (17). Then, for every $\gamma_1 \ge 0$ and arbitrary (but fixed) $v \in \mathbb{R}^n$, it holds that

$$\Pr_{y|x}\left[\|P_{x,y}\cdot v\|_2^2 > \gamma_1 \cdot \frac{\|v\|_2^2}{n}\right] \lesssim e^{-\mathrm{K}_\mu \gamma_1}, \tag{18}$$

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 Appendix C. Moreover, Assumption 1 is not too restrictive. One would expect that a probability measure μ, if dominated by the uniform measure on $\mathbb{D}$ and with a smooth Radon-Nikodym derivative, satisfies Assumption 1 when restricted to sufficiently small neighborhoods.

Assumption 1 also controls the growth of the moments of $\|P_{x,y}v\|_2$ [11, Lemma 5.5]. Finally, given the point cloud X, we also conveniently set

$$\epsilon_{\mu,X} := \min\left[\epsilon_\mu, \epsilon_X\right]. \qquad \text{(see Assumption 1 and (13))} \tag{19}$$

We are now in position to present the main results.

3.2 Performance of Algorithm 1

With the setup detailed in Section 2, we now quantify the performance of Algorithm 1. In Theorems 1 and 2 below, for a fixed point cloud X, we focus on how well the output of Algorithm 1, namely $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$, 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. We now turn to the details.

Theorem 1 below states that $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ can be a nearly unbiased estimator of $\dot{\Sigma}_X$ given X, see (11). The proof is given in Section 5.1. 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.

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 (8,9). Assume that the columns of (fixed) $X \in \mathbb{R}^{n\times N}$ belong to $\mathbb{D}$, namely $X \subset \mathbb{D}$ in our notation. Fix also $\epsilon \in (0, \epsilon_{\mu,X}]$, see (19). For an integer N and integers $N_{X,\epsilon} \ge N$ and $N_{X,\min,\epsilon} \le N_{X,\epsilon}$, assume also that*

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

[3] A simple calculation shows that $C = 1/n$. See Appendix D.

where

$$\rho_{\mu,X,\epsilon} := N \cdot \min_{x \in X} \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}. \tag{21}$$

Then the output of Algorithm 1, namely the estimator $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ defined in (6), satisfies

$$\left\| \mathbb{E}_{Y_{X,\epsilon}|X} \left[\ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \dot{\Sigma}_X \right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + n^2 \, \mathrm{L}_f^2 \, N^{-10} + \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^{3/2} \max(\mathrm{K}_\mu^{-1/2}, 1) \log^{\frac{1}{2}} N_{X,\epsilon}, \tag{22}$$

where $\mathrm{B}_{\mu,\epsilon}$ is given explicitly in (43).

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 (20).

- The quantity $\rho_{\mu,X,\epsilon}$, defined in (21), 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 (20) ensure that every $x \in X$ has sufficiently many neighbors in $Y_{X,\epsilon}$, that is $N_{x,\epsilon}$ 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 (22) holds with a total of $N_{X,\epsilon} = O(N \log^2 N)$ samples.
- The requirement that $N_{X,\epsilon} \geq n^{\frac{1}{20}}$ in (20) is very mild and will be automatically satisfied in cases of interest, as we discuss below.

Let us next interpret the bound on the bias in (22).

- The first term on the right-hand side of (22), namely $\mathrm{B}_{\mu,\epsilon}$, is given explicitly in (43); 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, as explained in the proof of Theorem 1, 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, (22) can in fact be sharpened by replacing the definition of $\mathrm{B}_{\mu,\epsilon}$ in (43) 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 (22) is negligible, as we will generally have N growing at least with n^2, as explained below.
- The third and fourth terms on the right-hand side of (22) can be made arbitrarily small by choosing the neighborhood radius 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 (22) is small relative to the size of Σ_μ, it may be appropriate to normalize (22). A reasonable choice would be to divide both sides of (22) by L_f^2, where L_f bounds $\|\nabla f(x)\|_2$ on $\mathbb{D}$; see (8). 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 (22) 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 (22) remains zero, as also described above. More generally, however, $\mathrm{B}_{\mu,\epsilon} \,/\, \mathrm{L}_f^2$ will contain a term that scales like $\sqrt{n}/N_{X,\min,\epsilon}$, and to control this term it is necessary to choose $N_{X,\min,\epsilon} \gtrsim n^{1/2} \log^2 N$ so that (20) is also satisfied. Notably, though, this method can be implemented when fewer than n neighbors are available for each $x \in X$, whereas estimating the local gradients via a conventional finite difference approximation would require n neighbors per point using deterministic queries. For Algorithm 1, we revisit the impact of n on the choices of N and $N_{X,\epsilon}$ after presenting Theorem 2 below.

Remark 2. (Sampling strategy) In Algorithm 1, $N_{X,\epsilon}$ points are independently drawn from the conditional probability measure on the ϵ-neighborhood of the point cloud X and then stored as the columns of $Y_{X,\epsilon}$, namely

$$Y_{X,\epsilon} \overset{\text{i.i.d.}}{\sim} \mu_{X,\epsilon}. \qquad \text{(see (15))} \tag{23}$$

This sampling strategy appears to best fit our fixed budget of $N_{X,\epsilon}$ samples, as it "prioritizes" the areas of $\mathbb{D}$ with larger "mass." For example, suppose that $\mu(dx) \gg \mu(dx')$ and $\nabla f(x) \approx \nabla f(x')$ for a pair $x, x' \in \mathbb{D}$. Then, $\nabla f(x)\nabla f(x)^*\mu(dx) \gg \nabla f(x')\nabla f(x')^*\mu(dx')$, suggesting that a larger weight should be placed on x rather than x' when estimating Σ_μ (see (10)). In the same scenario, assume naturally that $\mu(\mathbb{B}_{x,\epsilon}) \gg \mu(\mathbb{B}_{x',\epsilon})$, so that it is more likely to sample from the neighborhood of x than x'. Then, given a fixed budget of $N_{X,\epsilon}$ samples, it is highly likely that $N_{x,\epsilon} \gg N_{x',\epsilon}$. That is, x likely has far more ϵ-neighbors in $Y_{X,\epsilon}$ compared to x'. Loosely speaking then, the contribution of x to $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ is calculated more accurately than that of x'. In other words, the sampling strategy used in Algorithm 1 indeed assigns more weight to areas of $\mathbb{D}$ with larger mass.

In some applications, however, sampling points according to the distribution $\mu_{X,\epsilon}$ may be a challenge. A rejection sampling strategy—where points are drawn i.i.d. from μ on $\mathbb{D}$ and those falling outside $\cup_{x\in X}\mathbb{B}_{x,\epsilon}$ are discarded—is one possibility but is not feasible in high dimensions. As an alternative, one can consider a two-phase approach where first a ball $\mathbb{B}_{x,\epsilon}$ with $x \in X$ is selected with probability proportional to $\mu(\mathbb{B}_{x,\epsilon})$, and second a point is selected from the *uniform* measure within this ball. Such *locally uniform* sampling is an approximation to sampling from the distribution $\mu_{X,\epsilon}$. We expect that similar performance bounds hold for this locally uniform sampling strategy—especially when ϵ is small—but we do not quantify this here.

Remark 3. (Proof strategy) At a high level, the analysis handles the possible non-uniformity of the measure μ and higher order terms in $f(\cdot)$ by introducing quantities that are simpler to work with but are similar to $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$. Moreover, if $N_{X,\epsilon}$ is sufficiently large, then each $x \in X$ has many neighbors in $Y_{X,\epsilon}$ and this observation aids the analysis. The rest of the calculations, in effect, remove the estimation bias of $\dot{\Sigma}_{X,Y_{X,\epsilon}}$ in (4) to arrive at $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$.

Our second result, proved in Section 5.2, 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 (20), and under the mild assumptions that* $\log(n) \geq 1$, $N \geq \log(n)$, *and* $N_{X,\epsilon} \geq n$, *it holds that*

$$\begin{aligned}\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X\right\|_F \lesssim{}& \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,n^2 + \mathrm{B}_{\mu,\epsilon} \\ &+ \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}]\,\mathrm{L}_f^2 + n^2 L_f^2 N_{X,\epsilon}^{-3}\end{aligned} \tag{24}$$

except with a probability of $O\left(n^{-3} + N^{-3}\right)$. *Here,* $O(\cdot)$ *is the standard Big-O notation, the probability is with respect to the selection of* $Y_{X,\epsilon}$ *conditioned on the fixed set* X, *and* $\mathrm{B}_{\mu,\epsilon}$ *is given explicitly in (43).*

Remark 4. (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 (20) were discussed in Remark 1. The requirement that $N_{X,\epsilon} \geq n^{\frac{1}{20}}$ in (20) has been strengthened to $N_{X,\epsilon} \geq n$ in the statement of Theorem 2. However, this will again be automatically satisfied in cases of interest, as we discuss below.

- Let us now dissect the estimation error, namely the right-hand side of (24). As discussed in Remark 1, $\mathrm{B}_{\mu,\epsilon}$ in effect captures the non-uniformity of measure μ. In particular, the right-hand side of (24) 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 on the right hand side of (24) 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 (24) can be controlled by making $N_{X,\epsilon}$ suitably large. We discuss this point further below.

- The final term on the right hand side of (24) is negligible compared to the fourth, and we omit this in our discussion below.

Remark 5. (Estimating Σ_μ) Combining Theorem 2 with Proposition 1 and omitting the negligible terms yields

$$\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu\right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}]\, \mathrm{L}_f^2 + \frac{\mathrm{L}_f^2 \log n}{\sqrt{N}}, \tag{25}$$

with high probability when both X and $Y_{X,\epsilon}$ are selected randomly, therefore quantifying how well the full algorithm in Algorithm 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:

$$\frac{\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu\right\|_F}{L_f^2} \lesssim \frac{\mathrm{B}_{\mu,\epsilon}}{L_f^2} + \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}] + \frac{\log n}{\sqrt{N}}. \tag{26}$$

We discuss the terms appearing on the right hand side of (26):

- 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 $N_{X,\min,\epsilon} \gtrsim \sqrt{n} \log^2 N$.
- The second and third terms in (26) dictate the convergence rate of the error as the number of samples increases. In particular, setting $N_{X,\epsilon}$ proportional to $N \log^2(N)$ gives $N \approx N_{X,\epsilon} / \log^2(N)$ 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 logarithmic 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, the second term in (26) will dominate the third. Controlling the second term in (26) will require ensuring that $N_{X,\epsilon}$ (and thus the overall number of samples) scales like n^2, neglecting logarithmic factors.

Remark 6. (Proof strategy) The estimation error here is decomposed into "diagonal" and "off-diagonal" terms. The diagonal term, we find, can be written as a sum of independent random matrices and controlled by applying a standard Bernstein inequality. The off-diagonal term, however, is a second-order chaos (a certain sum of products of random variables) and requires additional care.

Remark 7. (Possible improvements) In combination with Weyl's inequality [12], (25) might be used to control the distance between the spectrum of $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ and that of Σ_μ. Likewise, given an integer $r \leq n$, standard perturbation results [13] might be deployed to measure the principal angle between the span of the leading r eigenvectors of $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ and an r-dimensional active subspace of $f(\cdot)$. To obtain the sharpest bounds, both these improvements would require controlling the spectral norm of $\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu$ rather than its Frobenius norm (which is bounded in Theorem 2 above). Controlling the spectral norm of the error appears to be considerably more difficult to achieve. As an aside, let us point out that the spectrum of Σ_μ in relation to $f(\cdot)$ has been studied in [3, 14].

4 Related Work

As argued in Section 1, 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 [15, 16, 2, 17, 18, 19, 20, 21, 22, 23]. 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 [3], 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 [14, 24] 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. In this sense, the present work fills a gap in the literature of ridge approximation; see also Section 1.2. For completeness, we note that it is natural to ask whether the results in [14] 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 that work). As detailed in Section 5 of [14], however, the sampling complexity in this general case will scale with n^5 (ignoring log factors). In contrast, our bound (25) requires only that the total number of function samples $N + N_{X,\epsilon}$ scale with n^2.

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 [8, 7, 25]. 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 $r \times n$ 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* [26], the aim is to construct $f(\cdot)$ using functions of the form $\sum_i h_i(a_i^* x)$ [27, 28, 29]. Further connections with neural networks are studied in [30],[31, Chapter 11]. See also [32, 33] for connections with Gaussian process regression and uncertainty quantification. *Sufficient dimension reduction* and related topics [34, 35, 36, 37, 38, 39, 40, 41] 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.

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 [42, 43, 44, 45, 46]. In this context, it appears that [47, 48, 49, 50] 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 [47], 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 [47] is not applicable to our setup.[4] 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. It is also worth including a few other works [9, 51, 52] which also involve covariance estimation from partially observed random vectors.

Yet another related field is matrix completion and recovery [53, 54, 55] and subspace estimation from data with erasures [56], where typically a low-rank structure is imposed. Lastly, in numerical linear algebra, random projections are increasingly used to facilitate matrix operations [57, 58, 59]. As a result, a very similar mathematical toolbox is used in that line of research.

5 Theory

This section contains the proofs of the two main results of this paper.

[4] The use of finite differences will effectively replace $\Phi_t x_t$ in $\widehat{\Sigma}_1$ in [47, Section 3] with a sum of rank-1 projections of x_t.

5.1 Proof of Theorem 1

Let us begin by outlining the proof strategy.

- First, we introduce a new quantity: $\dddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n\times n}$. Conditioned on a certain "good" event $\mathcal{E}_1$, $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ is easier to work with than $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$.
- Then, for fixed $X \subset \mathbb{D}_\epsilon$, we define another "good" event $\mathcal{E}_2$ where each $x \in X$ has sufficiently many neighbors in $Y_{X,\epsilon}$. Lemma 2 below shows that $\mathcal{E}_2$ is very likely to happen if $N_{X,\epsilon} = \#Y_{X,\epsilon}$ is large enough. Conditioned on the event $\mathcal{E}_2$, Lemma 3 below shows that $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ is a nearly unbiased estimator of $\dot{\Sigma}_X$:
$$\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] \approx \dot{\Sigma}_X. \tag{27}$$
- Lastly, we remove the conditioning on $\mathcal{E}_1 \cap \mathcal{E}_2$ to complete the proof of Theorem 1.

We now turn to the details and introduce $\dddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n\times n}$:

$$\dddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N}\left(1+\frac{1-\frac{2}{n}}{1+\frac{2}{n}}\cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \cdot\left(\sum_{N_{x,\epsilon}\geq N_{X,\min,\epsilon}} \ddot{\nabla}_{Y_{x,\epsilon}} f(x)\cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \frac{\sum_{N_{x,\epsilon}\geq N_{X,\min,\epsilon}} \left\|\ddot{\nabla}_{Y_{x,\epsilon}} f(x)\right\|_2^2}{\left(1+\frac{2}{n}\right)N_{X,\min,\epsilon}+n+1-\frac{2}{n}}\cdot I_n\right), \tag{28}$$

Here,

$$\ddot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}}\sum_{y\in Y_{x,\epsilon}} P_{x,y}\cdot\nabla f(x) \in \mathbb{R}^n, \tag{29}$$

and $P_{x,y} \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto the direction of $y-x$. In order to relate $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ to $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$, we invoke the following result, proved in Appendix E.

Lemma 1. *Fix X and $\epsilon \in (0,\epsilon_{\mu,X}]$. It holds that*

$$\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F \leq \frac{1}{2}\epsilon^2\,\mathrm{H}_f^2\,n^2 + 2\epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,n^2. \tag{30}$$

Moreover, consider the event

$$\mathcal{E}_1 := \left\{\max_{x\in X}\max_{y\in Y_{x,\epsilon}} \|P_{x,y}\cdot\nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}\right\}, \tag{31}$$

for $Q_{X,\epsilon} > 0$ to be set later. Then, conditioned on the event $\mathcal{E}_1$, it holds that

$$\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F \leq \frac{1}{2}\epsilon^2\,\mathrm{H}_f^2\,n^2 + 2\epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,Q_{X,\epsilon}^{1/2}n^{3/2}. \tag{32}$$

Thanks to Assumption 1, the event $\mathcal{E}_1$ is very likely to happen for the right choice of $Q_{X,\epsilon}$. Indeed, if we set $Q_{X,\epsilon} = \gamma_2 \log N_{X,\epsilon}$ for $\gamma_2 \geq 1$, then

$$\Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_1^C\right] \lesssim N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2}, \tag{33}$$

which follows from (18) and an application of the union bound (similar to the slightly more general result in Lemma 8).

Roughly speaking, in light of Lemma 1, $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \approx \dddot{\Sigma}_{X,Y_{X,\epsilon}}$. It therefore suffices to study the bias of $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ in the sequel. As suggested earlier, if $\#Y_{X,\epsilon} = N_{X,\epsilon}$ is sufficiently large, then every $x \in X$ will likely have many neighbors in $Y_{X,\epsilon}$, namely $\#Y_{x,\epsilon} = N_{x,\epsilon} \gg 1$ for every $x \in X$. This claim is formalized below and proved in Appendix F.

Lemma 2. *Fix X and $\epsilon \in (0, \epsilon_X]$. With $\gamma_3 \geq 1$, assume that*

$$N_{X,\epsilon} \gtrsim \frac{\gamma_3^2 \log^2 N \cdot \mu\left(\mathbb{B}_{X,\epsilon}\right)}{\min_{x\in X} \mu\left(\mathbb{B}_{x,\epsilon}\right)}. \tag{34}$$

Then, except with a probability of at most $N^{1-\gamma_3}$, it holds that

$$\frac{1}{2} \cdot \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon} \leq N_{x,\epsilon} \leq \frac{3}{2} \cdot \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}, \qquad \forall x \in X. \tag{35}$$

To use Lemma 2 here, we proceed as follows. For $\gamma_3 \geq 1$, suppose that

$$N_{X,\min,\epsilon} \gtrsim \gamma_3^2 \log^2 N, \tag{36}$$

and consider the event

$$\mathcal{E}_2 := \bigcap_{x\in X} \left\{ N_{x,\epsilon} \geq \frac{1}{2} \cdot \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon} \geq N_{X,\min,\epsilon} \right\}, \tag{37}$$

where, in particular, each $x \in X$ has at least $N_{X,\min,\epsilon}$ neighbors in $Y_{X,\epsilon}$. In light of Lemma 2, $\mathcal{E}_2$ is very likely to happen. To be specific,

$$\Pr_{Y_{X,\epsilon}|X} \left[\mathcal{E}_2^C\right] \leq N^{1-\gamma_3}, \tag{38}$$

provided that

$$N_{X,\epsilon} \gtrsim \frac{N_{X,\min,\epsilon} \cdot \mu\left(\mathbb{B}_{X,\epsilon}\right)}{\min_{x\in X} \mu\left(\mathbb{B}_{x,\epsilon}\right)} = \frac{N_{X,\min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, \qquad \text{(see (36))} \tag{39}$$

where we conveniently defined

$$\rho_{\mu,X,\epsilon} = N \cdot \min_{x\in X} \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})}. \tag{40}$$

Conditioned on the event $\mathcal{E}_2$, $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ in (28) takes the following simplified form:

$$\dddot{\Sigma}_{X,Y_{X,\epsilon}} = \frac{1}{N} \left(1 + \frac{1-\frac{2}{n}}{1+\frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \left(\sum_{x\in X} \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \ddot{\nabla}_{Y_{X,\epsilon}} f(x)^* - \frac{\sum_{x\in X} \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2}{\left(1+\frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right). \tag{41}$$

Using the above simplified form, we will prove the following result in Appendix G. Roughly speaking it states that, conditioned on the event $\mathcal{E}_2$, $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ is a nearly-unbiased estimator of $\dot{\Sigma}_X$.

Lemma 3. *Fix X and $\epsilon \in (0, \epsilon_{\mu,X}]$. Then, it holds that*

$$\left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X \right\|_F \leq \mathrm{B}_{\mu,\epsilon}, \tag{42}$$

where

$$\mathrm{B}_{\mu,\epsilon} := \frac{2\,\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 4\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon} + 1\right) \mathrm{L}_f^2 + \frac{2\,\mathrm{L}_f^2\,(1+\sqrt{n})}{N_{X,\min,\epsilon}}, \tag{43}$$

$$\mathrm{B}'_{\mu,\epsilon} := n \cdot \sup_{x\in\mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x}\left[P_{x,y}\right] - \frac{I_n}{n} \right\|_2, \qquad (y|x \sim \mu_{x,\epsilon})$$

$$\mathrm{B}''_{\mu,\epsilon} := n^2 \cdot \sup_{x\in\mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x}\left[P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\right] - \left(\frac{2\nabla f(x)\nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right\|_F, \qquad (y|x \sim \mu_{x,\epsilon}).$$

Moreover, suppose that μ is the uniform probability measure on $\mathbb{D}$, and that $N_{x,\epsilon} = N_{x',\epsilon}$ for every pair $x, x' \in X$. Then, conditioned on $\mathcal{E}_2$, one can replace $\mathrm{B}_{\mu,\epsilon}$ with 0, and thus $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ is an unbiased estimator of $\dot{\Sigma}_X$.

Next, we remove the conditioning on the event $\mathcal{E}_2$, with the aid of the following bounds:

$$\begin{aligned}
\left\|\dot{\Sigma}_X\right\|_F &\leq \mathrm{L}_f^2, \qquad \text{(see (11) and (8))} \\
\left\|\ddot{\nabla}_{Y_{x,\epsilon}} f(x)\right\|_2 &\leq n\,\mathrm{L}_f, \qquad \forall x \in X, \qquad \text{(see (29) and (8))} \\
\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F &\lesssim n^2\,\mathrm{L}_f^2\,. \qquad \text{(see (28) and (8))}
\end{aligned} \tag{44}$$

Then, we write that

$$\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F \\
&= \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] \cdot \Pr_{Y_{X,\epsilon}|X}[\mathcal{E}_2] + \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2^C,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] \cdot \Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_2^C\right] - \dot{\Sigma}_X\right\|_F \\
&\leq \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F \cdot \Pr_{Y_{X,\epsilon}|X}[\mathcal{E}_2] + \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2^C,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F \cdot \Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_2^C\right] \\
&\leq \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F \cdot \Pr_{Y_{X,\epsilon}|X}[\mathcal{E}_2] + \left(\sup\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F + \sup\left\|\dot{\Sigma}_X\right\|_F\right) \cdot \Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_2^C\right] \\
&\lesssim \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F + n^2\,\mathrm{L}_f^2 \cdot N^{1-\gamma_3} \qquad \text{(see (44) and (38))} \\
&\leq \mathrm{B}_{\mu,\epsilon} + n^2\,\mathrm{L}_f^2 \cdot N^{1-\gamma_3}, \qquad \text{(see Lemma 3 and (11))}
\end{aligned} \tag{45}$$

which, to reiterate, holds with $N_{X,\min,\epsilon} \gtrsim \gamma_3^2 \log^2 N$ and under (39). Lastly, we reintroduce $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ by invoking Lemma 1 as follows:

$$\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]\right\|_F \\
&\leq \left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]\right\|_F \cdot \Pr_{Y_{X,\epsilon}|X}[\mathcal{E}_1] \\
&\qquad + \left(\sup\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F + \sup\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F\right) \cdot \Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_1^C\right] \qquad \text{(similar to (45))} \\
&\lesssim \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,Q_{X,\epsilon}^{1/2} n^{3/2} + \left(\sup\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F + \sup\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F\right) \cdot \Pr_{Y_{X,\epsilon}|X}\left[\mathcal{E}_1^C\right] \qquad \text{(see (32))} \\
&\lesssim \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,(\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}}\,n^{3/2} \\
&\qquad + \left(\sup\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F + \sup\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F\right) \cdot N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2} \qquad \text{(with the choice of } Q_{X,\epsilon} \text{ in (33)) } \\
&\leq \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,(\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}}\,n^{3/2} \\
&\qquad + \left(\sup\left\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F + 2\sup\left\|\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right\|_F\right) \cdot N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2} \qquad \text{(triangle inequality)} \\
&\lesssim \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,(\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}}\,n^{3/2} \\
&\qquad + \left(\epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,n^2 + \mathrm{L}_f^2\,n^2\right) \cdot N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2}. \qquad \text{(see (30) and (44))}
\end{aligned} \tag{46}$$

Combining the above bound with (45) yields that

$$\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F \\
&\leq \left\|\mathbb{E}_{Y_{X,\epsilon}|X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \dot{\Sigma}_X\right\|_F + \left\|\mathbb{E}_{Y_{X,\epsilon}|X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]\right\|_F \qquad \text{(triangle inequality)} \\
&\lesssim \mathrm{B}_{\mu,\epsilon} + n^2\,\mathrm{L}_f^2 \cdot N^{1-\gamma_3} + \epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,(\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}}\,n^{3/2} \\
&\qquad + \left(\epsilon^2\,\mathrm{H}_f^2\,n^2 + \epsilon\,\mathrm{L}_f\,\mathrm{H}_f\,n^2 + \mathrm{L}_f^2\,n^2\right) \cdot N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2} \qquad \text{(see (45) and (46))} \\
&= \mathrm{B}_{\mu,\epsilon} + n^2\,\mathrm{L}_f^2\left(N^{1-\gamma_3} + N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2}\right) + \epsilon^2\,\mathrm{H}_f^2\,n^2\left(1 + N_{X,\epsilon}^{1-\mathrm{K}_\mu\,\gamma_2}\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^2 \left(\left(\frac{\gamma_2 \log N_{X,\epsilon}}{n} \right)^{\frac{1}{2}} + N_{X,\epsilon}^{1-\mathrm{K}_\mu \, \gamma_2} \right) \\
&\lesssim \mathrm{B}_{\mu,\epsilon} + n^2 \, \mathrm{L}_f^2 \left(N^{-10} + N_{X,\epsilon}^{-10} \right) + \epsilon^2 \, \mathrm{H}_f^2 \, n^2 \left(1 + N_{X,\epsilon}^{-10} \right) \\
&\quad + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^2 \left(\left(\frac{\max(\mathrm{K}_\mu^{-1}, 1) \log N_{X,\epsilon}}{n} \right)^{\frac{1}{2}} + N_{X,\epsilon}^{-10} \right) \qquad \text{(setting } \gamma_2 = 11 \max(\mathrm{K}_\mu^{-1}, 1) \text{ and } \gamma_3 = 11) \\
&\lesssim \mathrm{B}_{\mu,\epsilon} + n^2 \, \mathrm{L}_f^2 \, N^{-10} + \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^2 \left(\left(\frac{\max(\mathrm{K}_\mu^{-1}, 1) \log N_{X,\epsilon}}{n} \right)^{\frac{1}{2}} + N_{X,\epsilon}^{-10} \right) \qquad (N_{X,\epsilon} \geq N \geq 1) \\
&\lesssim \mathrm{B}_{\mu,\epsilon} + n^2 \, \mathrm{L}_f^2 \, N^{-10} + \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^{3/2} \max(\mathrm{K}_\mu^{-1/2}, 1) \log^{\frac{1}{2}} N_{X,\epsilon}. \qquad (N_{X,\epsilon} \geq n^{\frac{1}{20}}) \qquad (47)
\end{aligned}
$$

This completes the proof of Theorem 1.

5.2 Proof of Theorem 2

At a high level, the proof strategy here matches that of Theorem 1. First, we replace $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ with the simpler quantity $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ defined in (28). More specifically, in light of Lemma 1, it suffices to study $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ in the sequel.

Next, for $N_{X,\min,\epsilon} > 0$ to be set later, recall the "good" event $\mathcal{E}_2$ in (37) whereby every $x \in X$ has at least $N_{X,\min,\epsilon}$ neighbors in $Y_{X,\epsilon}$. Conditioned on the event $\mathcal{E}_2$, $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ takes the simpler form of (41), using which we prove the following result in Appendix H.

Lemma 4. *Fix X and $\epsilon \in (0, \epsilon_{\mu,X}]$. If $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{X,\epsilon}) \geq \log(n)$, then conditioned on $\mathcal{E}_2$ and X, it holds that*

$$
\left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + \gamma_7 \gamma_2^2 \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}] \, \mathrm{L}_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}, \tag{48}
$$

for $\gamma_7 \geq 1$ and $\gamma_2 \geq 3$, except with a probability $\lesssim e^{-\gamma_7} + n^{2-\log \gamma_7} + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}$.

We next remove the conditioning on the event $\mathcal{E}_2$ by letting R denote the right hand side of (48) and by writing that

$$
\begin{aligned}
\Pr_{Y_{X,\epsilon}|X} \left[\left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \gtrsim R \right] &\leq \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[\left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \gtrsim R \right] + \Pr_{Y_{X,\epsilon}|X} \left[\mathcal{E}_2^C \right] \text{ (see (53))} \\
&\lesssim e^{-\gamma_7} + n^{2-\log \gamma_7} + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} + N^{1-\gamma_3}, \qquad \text{(see Lemma 4 and (38))}
\end{aligned} \tag{49}
$$

under (36). Lastly, we reintroduce $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ by invoking Lemma 1: it holds that

$$
\begin{aligned}
\left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F &\leq \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \qquad \text{(triangle inequality)} \\
&\lesssim \frac{1}{2} \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + 2\epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, n^2 + R \qquad \text{(see Lemma 1)}
\end{aligned} \tag{50}
$$

with a failure probability of the order of

$$
e^{-\gamma_7} + n^{2-\log \gamma_7} + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} + N^{1-\gamma_3} \quad \text{(see (49))}, \tag{51}
$$

assuming (36) holds and that $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{X,\epsilon}) \geq \log(n)$. Setting $\gamma_2 = 4$, $\gamma_3 = 4$, and $\gamma_7 = 149 \log(N)$ and noting that $N_{X,\epsilon} \geq N$ completes the proof of Theorem 2.

6 Acknowledgements

AE would like to thank Hemant Tyagi for many interesting conversations regarding ridge approximation. AE was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1 and also by the Turing Seed Funding grant SF019. MBW was partially supported by NSF grant CCF-1409258 and NSF CAREER grant CCF-1149225. PGC was partially supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under award de-sc0011077 and the Defense Advanced Research Projects Agency's Enabling Quantification of Uncertainty in Physical Systems.

References

[1] J. F. Traub and H. Wozniakowski. A general theory of optimal algorithms. Technical report, Academic Press New York, 1980.

[2] 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.

[3] 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.

[4] J. Haupt, R. M Castro, and R. Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 57(9):6222–6235, 2011.

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

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

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

[8] P. Constantine and D. Gleich. Computing active subspaces with Monte Carlo. *arXiv preprint arXiv:1408.0545*, 2014.

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

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

[11] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[12] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*. Princeton reference. Princeton University Press, 2009.

[13] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[14] 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.

[15] A. Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, 2015.

[16] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33(1):125–143, 2011.

[17] C. J. Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.

[18] S. Gaïffas and G. Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007.

[19] A. B. Juditsky, O. V. Lepski, and A. B. Tsybakov. Nonparametric estimation of composite functions. *The Annals of Statistics*, pages 1360–1404, 2009.

[20] G. K. Golubev. Asymptotic minimax estimation of regression in the additive model. *Problemy peredachi informatsii*, 28(2):3–15, 1992.

[21] E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems: Standard information for functionals*, volume 12. European Mathematical Society, 2010.

[22] E. J. Candes. *Ridgelets: Theory and applications*. PhD thesis, Stanford University, 1998.

[23] S. Keiper. Analysis of generalized ridge functions in high dimensions. In *International Conference on Sampling Theory and Applications (SampTA)*, pages 259–263. IEEE, 2015.

[24] I. Bogunovic, V. Cevher, J. Haupt, and J. Scarlett. Active learning of self-concordant like multi-index functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2189–2193. IEEE, 2015.

[25] P. G. Constantine, A. Eftekhari, and R. Ward. A near-stationary subspace for ridge approximation. *arXiv preprint arXiv:1606.01929*, 2016.

[26] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990.

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

[28] P. J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

[29] D. L. Donoho and I. M. Johnstone. Projection-based approximation and a duality with kernel methods. *The Annals of Statistics*, pages 58–106, 1989.

[30] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[31] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[32] F. Vivarelli and C. K. I. Williams. Discovering hidden features with Gaussian processes regression. *Advances in Neural Information Processing Systems*, pages 613–619, 1999.

[33] R. Tripathy, I. Bilionis, and M. Gonzalez. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191–223, 2016.

[34] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[35] X. Yin and B. Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, pages 3392–3416, 2011.

[36] R. D. Cook. Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the section on Physical and Engineering Sciences*, pages 18–25. American Statistical Association Alexandria, VA, 1994.

[37] Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[38] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, 5:73–99, 2004.

[39] A. M. Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

[40] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.

[41] A. T. Glaws, P. G. Constantine, and R. D. Cook. Inverse regression for ridge recovery. *arXiv preprint arXiv:1702.02227*, 2017.

[42] T. T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[43] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[44] G. Dasarathy, P. Shah, B. Narayan Bhaskar, and R. D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

[45] M. Kolar and E. P. Xing. Consistent covariance selection from data with missing values. In *Proceedings of the International Conference on Machine Learning (ICML-12)*, pages 551–558, 2012.

[46] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

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

[48] A. Krishnamurthy, M. Azizyan, and A. Singh. Subspace learning from extremely compressed measurements. *arXiv preprint arXiv:1404.0751*, 2014.

[49] F. P Anaraki and S. Hughes. Memory and computation efficient PCA via very sparse random projections. In *Proceedings of the International Conference on Machine Learning (ICML-14)*, pages 1341–1349, 2014.

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

[51] P. L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[52] A. Gonen, D. Rosenbaum, Y. Eldar, and S. Shalev-Shwartz. The sample complexity of subspace learning with partial information. *arXiv preprint arXiv:1402.4844*, 2014.

[53] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[54] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[55] A. Eftekhari, M. B. Wakin, and R. A. Ward. MC^2: A two-phase algorithm for leveraged matrix completion. *arXiv preprint arXiv:1609.01795*, 2016.

[56] A. Eftekhari, L. Balzano, and M. B. Wakin. What to expect when you are expecting on the Grassmannian. *arXiv preprint arXiv:1611.07216*, 2016.

[57] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

[58] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[59] E. Liberty, F. Woolfe, P. G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

[60] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[61] F. W. J. Olver. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[62] R. Adamczak. Logarithmic Sobolev inequalities and concentration of measure for convex functions and polynomial chaoses. *Bull. Pol. Acad. Sci. Math.*, 53(2):221–238, 2005.

A Toolbox

In this section, we list a few results that are repeatedly used in the rest of appendices. Recall the following inequalities for a random variable z and event $\mathcal{A}$ (with complement $\mathcal{A}^C$):[5]

$$\mathbb{E}_z^p[z] \leq \mathbb{E}_{z|\mathcal{A}}^p[z] + \sup|z| \cdot \left(\Pr_z\left[\mathcal{A}^C\right]\right)^{\frac{1}{p}}, \qquad (\text{if } \sup|z| < \infty),$$

$$\Pr_z[z > z_0] \leq \Pr_{z|\mathcal{A}}[z > z_0] + \Pr_z\left[\mathcal{A}^C\right], \qquad \forall z_0. \tag{53}$$

We also recall the Bernstein inequality [60].

Proposition 2. (Bernstein inequality) *Let $\{A_i\}_i$ be a finite sequence of zero-mean independent random matrices, and set*

$$b := \max_i \|A_i\|_F, \tag{54}$$

$$\sigma^2 := \sum_i \mathbb{E}\|A_i\|_F^2. \tag{55}$$

Then, for $\gamma \geq 1$ and except with a probability of at most $e^{-\gamma}$, it holds that

$$\left\|\sum_i A_i\right\|_F \lesssim \gamma \cdot \max[b, \sigma]. \tag{56}$$

B Proof of Proposition 1

Recalling the definition of $\dot{\Sigma}_X$ from (11), we write that

$$\begin{aligned}\mathbb{E}_X\left[\dot{\Sigma}_X\right] &= \frac{1}{N}\sum_{x\in X}\mathbb{E}_X\left[\nabla f(x)\nabla f(x)^*\right] \qquad (\text{see (11)})\\ &= \mathbb{E}_x\left[\nabla f(x)\nabla f(x)^*\right] \qquad (\#X = N)\\ &= \Sigma_\mu, \qquad (\text{see (10)})\end{aligned} \tag{57}$$

[5] To see why the first inequality holds, note that

$$\begin{aligned}\mathbb{E}_z^p[z] &= \mathbb{E}_z^p\left[z\cdot 1_{\mathcal{A}}(z) + z\cdot 1_{\mathcal{A}^C}(z)\right]\\ &\leq \mathbb{E}_z^p\left[z\cdot 1_{\mathcal{A}}(z)\right] + \sup|z|\cdot\mathbb{E}^p\left[z\cdot 1_{\mathcal{A}^C}(z)\right], \qquad (\text{triangle inequality})\end{aligned}$$

where $1_{\mathcal{A}}(\cdot)$ is the indicator function for the event $\mathcal{A}$. It is easily verified that

$$\mathbb{E}_z^p\left[z\cdot 1_{\mathcal{A}}(z)\right] \leq \mathbb{E}_{z|\mathcal{A}}^p[z], \qquad \mathbb{E}_z^p\left[z\cdot 1_{\mathcal{A}^C}(z)\right] \leq \sup|z|\cdot\Pr_z[\mathcal{A}^C]^{\frac{1}{p}}, \tag{52}$$

from which (53) follows immediately.

which proves the first claim. To control the deviation about the mean, we will invoke the standard Bernstein inequality, recorded in Proposition 2 for the reader's convenience. Note that

$$\begin{aligned}\dot{\Sigma}_X - \Sigma_\mu &= \dot{\Sigma}_X - \mathbb{E}_X\left[\dot{\Sigma}_X\right] \\ &= \frac{1}{N}\sum_{x\in X} \nabla f(x)\nabla f(x)^* - \mathbb{E}_x\left[\nabla f(x)\nabla f(x)^*\right] \\ &=: \sum_{x\in X} A_x, \end{aligned} \tag{58}$$

where $\{A_x\}_x \subset \mathbb{R}^{n\times n}$ are independent and zero-mean random matrices. To apply the Bernstein inequality (Proposition 2), we compute the parameters

$$\begin{aligned} b &= \max_{x\in X} \|A_x\|_F \\ &= \frac{1}{N}\max_{x\in X} \|\nabla f(x)\nabla f(x)^* - \mathbb{E}_x\left[\nabla f(x)\nabla f(x)^*\right]\|_F \qquad \text{(see (58))} \\ &\le \frac{1}{N}\max_{x\in X} \|\nabla f(x)\nabla f(x)^*\|_F + \frac{1}{N}\mathbb{E}_x\|\nabla f(x)\nabla f(x)^*\|_F \qquad \text{(triangle and Jensen's inequalities)} \\ &\le \frac{2}{N}\sup_{x\in\mathbb{D}} \|\nabla f(x)\nabla f(x)^*\|_F \\ &= \frac{2}{N}\sup_{x\in\mathbb{D}} \|\nabla f(x)\|_2^2 \\ &= \frac{2\,\mathrm{L}_f^2}{N} \qquad \text{(see (8))} \end{aligned}$$

and

$$\begin{aligned} \sigma^2 &= \sum_{x\in X} \mathbb{E}_x\|A_x\|_F^2 \\ &= \frac{1}{N}\mathbb{E}_x\|\nabla f(x)\nabla f(x)^* - \mathbb{E}_x\left[\nabla f(x)\nabla f(x)^*\right]\|_F^2 \qquad \text{(see (58) and } \#X = N) \\ &\le \frac{1}{N}\mathbb{E}_x\|\nabla f(x)\nabla f(x)^*\|_F^2 \qquad \left(\mathbb{E}\|Z - \mathbb{E}[Z]\|_F^2 \le \mathbb{E}\|Z\|_F^2 \text{ for a random matrix } Z\right) \\ &= \frac{1}{N}\mathbb{E}_x\|\nabla f(x)\|_2^4 \\ &\le \frac{\mathrm{L}_f^4}{N}, \qquad \text{(see (8))} \end{aligned}$$

and thus

$$\max[b,\sigma] \le \frac{2\,\mathrm{L}_f^2}{\sqrt{N}}. \tag{59}$$

Therefore, for $\gamma_4 \ge 1$ and except with a probability of at most $e^{-\gamma_4}$, Proposition 2 dictates that

$$\begin{aligned} \left\|\dot{\Sigma}_X - \Sigma_\mu\right\|_F &= \left\|\sum_{x\in X} A_x\right\|_F \qquad \text{(see (58))} \\ &\lesssim \gamma_4 \cdot \max[b,\sigma] \\ &\lesssim \gamma_4 \cdot \frac{\mathrm{L}_f^2}{\sqrt{N}}, \end{aligned}$$

which completes the proof of Proposition 1 when we take $\gamma_4 = \log n$.

C Uniform Measure Satisfies Assumption 1

We verify in this appendix that the uniform probability measure on $\mathbb{D}$ satisfies Assumption 1. Fix arbitrary $\epsilon > 0$ and x in the ϵ-interior of $\mathbb{D} \subseteq \mathbb{R}^n$, namely $x \in \mathbb{D}_\epsilon$, assuming that $\mathbb{D}_\epsilon \neq \emptyset$. The conditional measure in the neighborhood $\mathbb{B}_{x,\epsilon}$ too is uniform, so that $y|x \sim \text{uniform}(\mathbb{B}_{x,\epsilon})$. Then, for fixed $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$, observe that

$$\|P_{x,y}v\|_2^2 \sim \text{beta}\left(\frac{1}{2}, \frac{n-1}{2}\right). \tag{60}$$

To study the tail bound of the random variable $\|P_{x,y}v\|_2^2$, we proceed as follows. We note that (18) trivially holds for any $\gamma_1 > n$ since $\|P_{x,y}v\|_2^2 \leq \|v\|_2^2$ because $P_{x,y}$ is an orthogonal projection. Thus, it suffices to consider fixed $\gamma_1 \in (0, n]$. Recalling the moments of the beta distribution, write that

$$\begin{aligned}
\Pr{}_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] &= \Pr{}_{y|x}\left[\|P_{x,y}v\|_2^{2\lambda} > \left(\frac{\gamma_1}{n}\right)^\lambda\right] \qquad (\lambda > 0) \\
&\leq \left(\frac{\gamma_1}{n}\right)^{-\lambda} \mathbb{E}\left[\|P_{x,y}v\|_2^{2\lambda}\right] \qquad \text{(Markov's inequality)} \\
&= \left(\frac{\gamma_1}{n}\right)^{-\lambda} \frac{B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-1}{2}\right)},
\end{aligned} \tag{61}$$

where

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{62}$$

is the beta function. Above, $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the usual gamma function. In order to choose λ above, we rewrite (61) as

$$\begin{aligned}
\Pr{}_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] &\leq e^{-\lambda\log\left(\frac{\gamma_1}{n}\right) + \log\left(B\left(\lambda+\frac{1}{2}, \frac{n-1}{2}\right)\right) - \log\left(B\left(\frac{1}{2}, \frac{n-1}{2}\right)\right)} \\
&=: e^{l(\lambda)}.
\end{aligned} \tag{63}$$

In order to minimize $l(\cdot)$, we compute its derivative:

$$\begin{aligned}
l'(\lambda) &= -\log\left(\frac{\gamma_1}{n}\right) + \frac{d}{d\lambda}\log\left(B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)\right) - \frac{d}{d\lambda}\log\left(B\left(\frac{1}{2}, \frac{n-1}{2}\right)\right) \\
&= -\log\left(\frac{\gamma_1}{n}\right) + \frac{d}{d\lambda}\log\left(\Gamma\left(\lambda + \frac{1}{2}\right)\right) - \frac{d}{d\lambda}\log\left(\Gamma\left(\lambda + \frac{n}{2}\right)\right) \qquad \text{(see (62))} \\
&= -\log\left(\frac{\gamma_1}{n}\right) + \frac{\Gamma'\left(\lambda + \frac{1}{2}\right)}{\Gamma\left(\lambda + \frac{1}{2}\right)} - \frac{\Gamma'\left(\lambda + \frac{n}{2}\right)}{\Gamma\left(\lambda + \frac{n}{2}\right)} \\
&= -\log\left(\frac{\gamma_1}{n}\right) + \psi\left(\lambda + \frac{1}{2}\right) - \psi\left(\lambda + \frac{n}{2}\right),
\end{aligned} \tag{64}$$

where $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$ is the "digamma" function. It is well-known that $\psi(a) \approx \log(a)$ for large a (see, for example, [61]). To guide our choice of λ, note that if n is sufficiently large and we take λ such that $1 \ll \lambda \ll n$, we have that

$$\begin{aligned}
l'(\lambda) &= -\log\left(\frac{\gamma_1}{n}\right) + \psi\left(\lambda + \frac{1}{2}\right) - \psi\left(\lambda + \frac{n}{2}\right) \qquad \text{(see (64))} \\
&\approx -\log\left(\frac{\gamma_1}{n}\right) + \log\lambda - \log\left(\frac{n}{2}\right) \\
&= -\log\left(\frac{2\lambda}{\gamma_1}\right),
\end{aligned} \tag{65}$$

thereby suggesting the choice of $\lambda = \gamma_1/2$. With this choice, we find that

$$\Pr{}_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] \leq \left(\frac{\gamma_1}{n}\right)^{-\frac{\gamma_1}{2}} \frac{B\left(\frac{\gamma_1+1}{2}, \frac{n-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-1}{2}\right)} \qquad \text{(see (61))}$$

$$
\begin{aligned}
&= \left(\frac{\gamma_1}{n}\right)^{-\frac{\gamma_1}{2}} \frac{\Gamma\left(\frac{\gamma_1+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n+\gamma_1}{2}\right)} \qquad \text{(see (62))} \\
&\lesssim \left(\frac{\gamma_1}{n}\right)^{-\frac{\gamma_1}{2}} \frac{\left(\frac{\gamma_1+1}{2}\right)^{\frac{\gamma_1}{2}} e^{-\frac{\gamma_1+1}{2}} \left(\frac{n}{2}\right)^{\frac{n-1}{2}} e^{-\frac{n}{2}}}{\left(\frac{n+\gamma_1}{2}\right)^{\frac{n+\gamma_1-1}{2}} e^{-\frac{n+\gamma_1}{2}}} \qquad \left(1 < \frac{a^{\frac{1}{2}-a}e^a}{\sqrt{2\pi}}\Gamma(a) < e^{\frac{1}{12a}},\ \forall a>0\right) \\
&\lesssim \left(\frac{n}{n+\gamma_1}\right)^{\frac{n+\gamma_1-1}{2}} \\
&\le \left(\frac{n}{n+\gamma_1}\right)^{\frac{n-1}{2}} \qquad (\gamma_1 > 0) \\
&= \left(1+\frac{\gamma_1}{n}\right)^{-\frac{n-1}{2}} \\
&\le e^{-\frac{\gamma_1}{n}\cdot\frac{n-1}{2}} \qquad (1+a \le a^a) \\
&\le e^{-\frac{\gamma_1}{2}+\frac{1}{2}}. \qquad (\gamma_1 \le n)
\end{aligned} \tag{66}
$$

Therefore, Assumption 1 holds for the uniform probability measure with $\epsilon_\mu = \infty$ and $\mathrm{K}_\mu = 1/2$.

D Estimating $\nabla f(x)$

For fixed $x \in \mathbb{D}$, by drawing samples from the neighborhood of x and then applying the method of finite differences, we may estimate $\nabla f(x)$. This is described below for the sake of completeness.

Proposition 3. *Fix $x \in \mathbb{D}$ and take $\epsilon > 0$ small enough so that x belongs to ϵ-interior of $\mathbb{D}$, namely $x \in \mathbb{D}_\epsilon$. Draw y from the conditional measure on the neighborhood $\mathbb{B}_{x,\epsilon}$, namely $y|x \sim \mu_{x,\epsilon}$ (see (17)). For an integer $N_{x,\epsilon}$, let $Y_{x,\epsilon} \subset \mathbb{B}_{x,\epsilon}$ contain $N_{x,\epsilon}$ independent copies of y. Then, it holds that*

$$
\left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon},x}\left[\dot{\nabla}_{Y_{x,\epsilon}} f(x)\right] - \nabla f(x) \right\|_2 \le \mathrm{B}'_{\mu,\epsilon}\,\mathrm{L}_f + \frac{\epsilon\,\mathrm{H}_f\,n}{2}. \tag{67}
$$

where

$$
\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y)-f(x)}{\|y-x\|_2} \cdot \frac{y-x}{\|y-x\|_2} \in \mathbb{R}^n, \tag{68}
$$

$$
\mathrm{B}'_{\mu,\epsilon} := n \cdot \sup_{x \in \mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x}\left[P_{x,y}\right] - \frac{I_n}{n} \right\|. \qquad (y|x \sim \mu_{x,\epsilon}) \tag{69}
$$

In particular, if μ is the uniform probability measure on $\mathbb{D}$, then $\mathrm{B}'_{\mu,\epsilon} = 0$.

Proof. First, we replace $\dot{\nabla}_{Y_{x,\epsilon}} f(x)$ with the simpler quantity $\ddot{\nabla}_{Y_{x,\epsilon}} f(x)$, defined as

$$
\ddot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n, \tag{70}
$$

where $P_{x,y} \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto the direction of $y-x$. By definition, the two quantities are related as follows:

$$
\begin{aligned}
&\left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \\
&= \frac{n}{N_{x,\epsilon}} \left\| \sum_{y \in Y_{x,\epsilon}} \frac{y-x}{\|y-x\|_2^2} \left(f(y) - f(x) - (y-x)^* \nabla f(x)\right) \right\|_2 \qquad \left(P_{x,y} = \frac{(y-x)(y-x)^*}{\|y-x\|_2^2}\right) \\
&\le \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \left\| \frac{y-x}{\|y-x\|_2^2} \left(f(y) - f(x) - (y-x)^* \nabla f(x)\right) \right\|_2 \qquad \text{(triangle inequality)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{|f(y) - f(x) - (y-x)^* \nabla f(x)|}{\|y-x\|_2} \\
&\leq n \cdot \sup_{y \in \mathbb{B}_{x,\epsilon}} \frac{|f(y) - f(x) - (y-x)^* \nabla f(x)|}{\|y-x\|_2} \qquad (\# Y_{x,\epsilon} = N_{x,\epsilon}) \\
&\leq n \cdot \sup_{y \in \mathbb{B}_{x,\epsilon}} \frac{\mathrm{H}_f \|y-x\|_2}{2} \qquad \text{(Taylor's expansion and (9))} \\
&\leq n \cdot \frac{\mathrm{H}_f \cdot \epsilon}{2}. \qquad (y \in \mathbb{B}_{x,\epsilon}) \qquad (71)
\end{aligned}
$$

Loosely speaking then, $\dot{\nabla}_{Y_{x,\epsilon}} f(x) \approx \ddot{\nabla}_{Y_{x,\epsilon} f(x)}$ and it therefore suffices to study the estimation bias of $\ddot{\nabla}_{Y_{x,\epsilon}} f(x)$. To that end, we simply note that

$$
\begin{aligned}
\left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} \left[\ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right] - \nabla f(x) \right\|_2 &= \left\| n \cdot \mathbb{E}_{y|x} \left[P_{x,y} \nabla f(x) \right] - \nabla f(x) \right\|_2 \qquad (y|x \sim \mu_{x,\epsilon}) \\
&= \left\| n \cdot \mathbb{E}_{y|x} \left[P_{x,y} \right] \cdot \nabla f(x) - \nabla f(x) \right\|_2 \\
&\leq n \cdot \sup_{x \in \mathbb{D}} \left\| \mathbb{E}_{y|x} \left[P_{x,y} \right] - \frac{I_n}{n} \right\| \cdot \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2 \\
&=: \mathrm{B}'_{\mu,\epsilon} \cdot \mathrm{L}_f, \qquad \text{(see (8))} \qquad (72)
\end{aligned}
$$

which, in turn, implies that

$$
\begin{aligned}
&\left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} \left[\dot{\nabla}_{Y_{x,\epsilon}} f(x) \right] - \nabla f(x) \right\|_2 \\
&\leq \left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} \left[\dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right] \right\|_2 + \left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} \left[\ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right] - \nabla f(x) \right\|_2 \qquad \text{(triangle inequality)} \\
&\leq \frac{n \, \mathrm{H}_f \, \epsilon}{2} + \mathrm{B}'_{\mu,\epsilon} \, \mathrm{L}_f \, . \qquad \text{(see (71) and (72))} \qquad (73)
\end{aligned}
$$

In particular, when μ is the uniform probability measure on $\mathbb{D}$, $P_{x,y}$ is an isotropic random matrix (for fixed $x \in \mathbb{D}$). Therefore, $\mathbb{E}_{y|x}[P_{x,y}] = C \cdot I_n$ for some scalar C. To find C, we note that

$$
\mathrm{trace} \left[\mathbb{E}_{y|x} \left[P_{x,y} \right] \right] = \mathbb{E}_{y|x} \left[\mathrm{trace} \left[P_{x,y} \right] \right] = 1 = C \cdot \mathrm{trace}[I_n] = C \cdot n \Longrightarrow C = \frac{1}{n},
$$

where we used the fact that $P_{x,y}$ is a rank-1 orthogonal projection. Consequently, when μ is the uniform measure, $\mathrm{B}'_{\mu,\epsilon} = 0$. This completes the proof of Proposition 3. □

E Proof of Lemma 1

We only verify the second claim, as the other proof is similar. Conditioned on the event $\mathcal{E}_1$, note that

$$
\begin{aligned}
\left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 &\leq \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \|P_{x,y} \cdot \nabla f(x)\|_2 \qquad \text{(see (29))} \\
&\leq n \cdot \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \|P_{x,y} \cdot \nabla f(x)\|_2 \qquad (\# Y_{x,\epsilon} = N_{x,\epsilon}) \\
&\leq n \cdot \sqrt{\frac{Q_{X,\epsilon} \, \mathrm{L}_f^2}{n}}. \qquad \text{(see (31))} \qquad (74)
\end{aligned}
$$

Using the inequality $\|aa^* - bb^*\|_2 \leq \|a-b\|(\|a\|_2 + \|b\|_2)$ for any $a, b \in \mathbb{R}^n$ in the third line below, it follows that

$$
\frac{1}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F
$$

$$
\begin{aligned}
&\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \qquad \text{(triangle inequality)} \\
&\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left(\left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 + \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \right) \\
&\leq \max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left(\max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 + \max_{x \in X} \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \right) \qquad (\#X = N) \\
&\leq \max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left(\max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 + 2 \max_{x \in X} \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \right) \quad \text{(triangle ineq.)} \\
&\leq \frac{\epsilon \, \mathrm{H}_f \, n}{2} \left(\frac{\epsilon \, \mathrm{H}_f \, n}{2} + 2\sqrt{Q_{X,\epsilon} \, \mathrm{L}_f^2 \, n} \right) \qquad \text{(see (71) and (74))} \\
&\leq \frac{1}{4} \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, Q_{X,\epsilon}^{1/2} n^{3/2}, \qquad (75)
\end{aligned}
$$

which, in turn, immediately implies that

$$
\begin{aligned}
&\frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right| \\
&= \frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \operatorname{trace} \left[\dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right] \right| \\
&\leq \frac{\sqrt{n}}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \\
&\leq \frac{1}{4} \epsilon^2 \, \mathrm{H}_f^2 \, n^{5/2} + \epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, Q_{X,\epsilon}^{1/2} n^2, \qquad \text{(see (75))} \qquad (76)
\end{aligned}
$$

where the third line above uses the fact that $|\operatorname{trace}(A)| \leq \sqrt{n} \|A\|_F$ for any $A \in \mathbb{R}^{n \times n}$. Recall the definitions of $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ and $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ in (6) and (28), respectively. Then, by combining (75) and (76), it follows that

$$
\begin{aligned}
&\left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \\
&\leq \frac{1}{N} \left\| \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \\
&\quad + \frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right| \cdot \frac{\|I_n\|_F}{n} \qquad \text{(see (6) and (28))} \\
&\leq \frac{1}{2} \epsilon^2 \, \mathrm{H}_f^2 \, n^2 + 2\epsilon \, \mathrm{L}_f \, \mathrm{H}_f \, Q_{X,\epsilon}^{1/2} n^{3/2}. \qquad \text{(see (75) and (76))} \qquad (77)
\end{aligned}
$$

This completes the proof of Lemma 1.

F Proof of Lemma 2

Our objective is to establish that, given X and neighborhood radius ϵ, each $x \in X$ has many neighbors in $Y_{X,\epsilon}$ provided that $N_{X,\epsilon} = \# Y_{X,\epsilon}$ is sufficiently large. To that end, we proceed as follows. Recall that $\mu_{X,\epsilon}$ is the conditional distribution on the ϵ-neighborhood of the point cloud X (see (15)). With $y \sim \mu_{X,\epsilon}$ and for fixed $x \in X$, observe that y belongs to the ϵ-neighborhood of x (namely, $y \in \mathbb{B}_{x,\epsilon}$) with the following probability:

$$
\Pr{}_{y|x} \left[y \in \mathbb{B}_{x,\epsilon} \right] = \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}. \qquad (78)
$$

Equivalently, the indicator function $1_{y\in\mathbb{B}_{x,\epsilon}}$ follows a Bernoulli distribution:

$$1_{y\in Y_{x,\epsilon}}|x \sim \text{Bernoulli}\left(\frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\right). \tag{79}$$

Then,

$$\mathbb{E}_{Y_{X,\epsilon}|X}\left[N_{x,\epsilon}\right] = \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} \cdot \#Y_{X,\epsilon} = \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} \cdot N_{X,\epsilon}, \tag{80}$$

and, to investigate the concentration of $N_{x,\epsilon}$ about its expectation, we write that

$$\begin{aligned}
N_{x,\epsilon} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} \cdot N_{X,\epsilon} &= N_{x,\epsilon} - \mathbb{E}_{Y_{X,\epsilon}|X}\left[N_{x,\epsilon}\right] \qquad \text{(see (80))} \\
&= \sum_{y\in Y_{X,\epsilon}} \left(1_{y\in\mathbb{B}_{x,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|X}\left[1_{y\in\mathbb{B}_{x,\epsilon}}\right]\right) \\
&= \sum_{y\in Y_{X,\epsilon}} \left(1_{y\in\mathbb{B}_{x,\epsilon}} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\right) \\
&=: \sum_{y\in Y_{X,\epsilon}} a_y,
\end{aligned} \tag{81}$$

where $\{a_y\}_y$ are independent zero-mean random variables (for fixed $x \in X$). In order to apply the Bernstein's inequality (Proposition 2) to the last line of (81), we write that

$$\begin{aligned}
b &= \max_y |a_y| \\
&= \max_y \left|1_{y\in\mathbb{B}_{x,\epsilon}} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\right| \qquad \text{(see (81))} \\
&\leq 1,
\end{aligned} \tag{82}$$

$$\begin{aligned}
\sigma^2 &= \sum_{y\in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|x}\left[a_y^2\right] \\
&= \sum_{y\in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|x}\left[\left(1_{y\in\mathbb{B}_{x,\epsilon}} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\right)^2\right] \qquad \text{(see (81))} \\
&= \sum_{y\in Y_{X,\epsilon}} \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\left(1 - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)}\right) \qquad \text{(see (79))} \\
&\leq \sum_{y\in Y_{X,\epsilon}} \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} = \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} \cdot N_{X,\epsilon}, \qquad (\#Y_{x,\epsilon} = N_{x,\epsilon})
\end{aligned} \tag{83}$$

$$\max\left[b,\sigma\right] = \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}. \qquad \left(\text{if } N_{X,\epsilon} \geq \frac{\mu\left(\mathbb{B}_{X,\epsilon}\right)}{\mu\left(\mathbb{B}_{x,\epsilon}\right)}\right) \tag{84}$$

From Proposition 2, then, it follows that

$$\begin{aligned}
\left|N_{x,\epsilon} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}\right| &\lesssim \gamma_5 \cdot \max\left[b,\sigma\right] \\
&= \gamma_5 \cdot \sqrt{\frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}}, \qquad \text{(see (84))}
\end{aligned} \tag{85}$$

for $\gamma_5 \geq 1$ and except with a probability of at most $e^{-\gamma_5}$. Recall that $\#X = N$. Then, an application of the union bound with the choice of $\gamma_5 = \gamma_3 \log N$ (with $\gamma_3 \geq 1$) yields that

$$\max_{x\in X}\left|N_{x,\epsilon} - \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}\right| \lesssim \gamma_3 \log N \cdot \sqrt{\frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}}, \tag{86}$$

except with a probability of at most $Ne^{-\gamma_3 \log N} = N^{1-\gamma_3}$. For the bound above to hold, we assume that $N_{X,\epsilon}$ is sufficiently large (so that the requirement in (84) hold for every $x \in X$). In fact, if

$$N_{X,\epsilon} \gtrsim \frac{\gamma_3^2 \log^2 N \cdot \mu\left(\mathbb{B}_{X,\epsilon}\right)}{\min_{x\in X} \mu\left(\mathbb{B}_{x,\epsilon}\right)}, \tag{87}$$

then (86) readily yields that

$$\frac{1}{2} \cdot \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon} \le N_{x,\epsilon} \le \frac{3}{2} \cdot \frac{\mu\left(\mathbb{B}_{x,\epsilon}\right)}{\mu\left(\mathbb{B}_{X,\epsilon}\right)} N_{X,\epsilon}, \qquad \forall x \in X, \tag{88}$$

except with a probability of at most $N^{1-\gamma_3}$. This completes the proof of Lemma 2.

G Proof of Lemma 3

Throughout, X and $\epsilon \in (0, \epsilon_{\mu,X}]$ are fixed, and we further assume that the event $\mathcal{E}_2$ holds (see (37)). For now, suppose in addition that the neighborhood structure $\overline{N}_{X,\epsilon} := \{N_{x,\epsilon}\}_{x\in X}$ is fixed too. Recalling the definition of $\ddot{\nabla}_{Y_{x,\epsilon}} f(\cdot)$ from (29), we first set

$$\mathbb{R}^{n\times n} \ni \dddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \sum_{x\in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^*, \tag{89}$$

for short, and then separate the "diagonal" and "off-diagonal" components of the expectation of $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ as follows:

$$\begin{aligned}
&\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right] \\
&= \frac{1}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\sum_{x\in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^*\right] \qquad \text{(see (89))} \\
&= \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'}\right] \qquad \text{(see (29))} \\
&= \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y\in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\right] \\
&\quad + \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} 1_{y\neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'}\right] \\
&= \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y\in Y_{x,\epsilon}} \mathbb{E}_{y|x}\left[P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\right] \\
&\quad + \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} \mathbb{E}_{y,y'|x}\left[1_{y\neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'}\right] \qquad (y, y' \sim \mu_{x,\epsilon}) \\
&= \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}} \cdot \mathbb{E}_{y|x}\left[P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\right] \\
&\qquad + \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} 1_{y\neq y'} \cdot \mathbb{E}_{y|x}\left[P_{x,y} \nabla f(x)\right] \cdot \mathbb{E}_{y'|x}\left[\nabla f(x)^* P_{x,y'}\right]. \qquad (90)
\end{aligned}$$

The last line above uses the fact that distinct elements of $Y_{x,\epsilon}$ are statistically independent. We next replace both the diagonal and off-diagonal components (namely, the first and second sums in the last line above) with simpler expressions. We approximate the diagonal term with another sum as follows:

$$\left\| \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}} \mathbb{E}_{y|x}\left[P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\right] - \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}} \left(\frac{2\nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right\|_F$$

$$
\begin{aligned}
&\leq \frac{n^2}{\min_{x\in X} N_{x,\epsilon}} \cdot \sup_{x\in\mathbb{D}} \left\| \mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\right] - \left(\frac{2\nabla f(x)\nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)}\cdot I_n\right)\right\|_F \qquad (\#X = N)\\
&=: \frac{\mathrm{B}''_{\mu,\epsilon}}{\min_{x\in X} N_{x,\epsilon}}\\
&\leq \frac{\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}}. \qquad \text{(see (37))}
\end{aligned} \tag{91}
$$

To replace the off-diagonal term in the last line of (90), first recall the inequality

$$
\|ab^* - cd^*\|_F \leq 2\max\left[\|a-c\|_2, \|b-d\|_2\right]\cdot\max\left[\|b\|_2, \|c\|_2\right], \qquad a,b,c,d\in\mathbb{R}^n, \tag{92}
$$

and then note that

$$
\begin{aligned}
&\left\| \frac{n^2}{N}\sum_{x\in X}\frac{1}{N_{x,\epsilon}^2}\sum_{y,y'\in Y_{x,\epsilon}} 1_{y\neq y'}\cdot\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right]\cdot\mathbb{E}_{y'|x}\left[\nabla f(x)^* P_{x,y'}\right] - \frac{1}{N}\sum_{x\in X}\frac{N_{x,\epsilon}-1}{N_{x,\epsilon}}\nabla f(x)\nabla f(x)^*\right\|_F\\
&= \left\| \frac{n^2}{N}\sum_{x\in X}\frac{1}{N_{x,\epsilon}^2}\sum_{y,y'\in Y_{x,\epsilon}} 1_{y\neq y'}\left(\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right]\cdot\mathbb{E}_{y'|x}\left[\nabla f(x)^* P_{x,y'}\right] - \frac{\nabla f(x)\nabla f(x)^*}{n^2}\right)\right\|_F \qquad (\#Y_{x,\epsilon} = N_{x,\epsilon})\\
&\leq n^2 \max_{x\in X}\max_{y,y'\in Y_{x,\epsilon}}\left\|\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right]\cdot\mathbb{E}_{y'|x}\left[\nabla f(x)^* P_{x,y'}\right] - \frac{\nabla f(x)\nabla f(x)^*}{n^2}\right\|_F \qquad (\#X = N, \#Y_{X,\epsilon} = N_{x,\epsilon})\\
&\leq 2n^2 \max_{x\in X}\left[\left\|\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right] - \frac{\nabla f(x)}{n}\right\|_2 \cdot \max\left[\left\|\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right]\right\|_2, \frac{\|\nabla f(x)\|_2}{n}\right]\right] \qquad \text{(see (92))}\\
&\leq 2n^2 \max_{x\in X}\left[\left\|\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right] - \frac{\nabla f(x)}{n}\right\|_2 \left(\left\|\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\right] - \frac{\nabla f(x)}{n}\right\|_2 + \frac{\|\nabla f(x)\|_2}{n}\right)\right] \qquad \text{(triangle ineq.)}\\
&\leq 2n^2\left(\frac{\mathrm{B}'_{\mu,\epsilon}}{n}\cdot\mathrm{L}_f\right)\left(\frac{\mathrm{B}'_{\mu,\epsilon}}{n}\cdot\mathrm{L}_f + \frac{\mathrm{L}_f}{n}\right) \qquad \text{(see (8) and (69))}\\
&= 2\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon}+1\right)\mathrm{L}_f^2\,.
\end{aligned} \tag{93}
$$

We may now replace the diagonal and off diagonal components in the last line of (90) with simpler expressions while incurring a typically small error. More specifically, in light of (91) and (93), (90) now implies that

$$
\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \frac{1}{N}\sum_{x\in X}\left(1+\frac{n-2}{N_{x,\epsilon}(n+2)}\right)\nabla f(x)\nabla f(x)^* - \frac{n}{N(n+2)}\sum_{x\in X}\frac{\|\nabla f(x)\|_2^2}{N_{x,\epsilon}}\cdot I_n\right\|_F\\
&= \left\|\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \frac{n^2}{N}\sum_{x\in X}\frac{1}{N_{x,\epsilon}}\left(\frac{2\nabla f(x)\nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)}\cdot I_n\right)\right.\\
&\qquad\qquad \left. - \frac{1}{N}\sum_{x\in X}\frac{N_{x,\epsilon}-1}{N_{x,\epsilon}}\nabla f(x)\nabla f(x)^*\right\|_F\\
&\leq \frac{\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon}+1\right)\mathrm{L}_f^2\,. \qquad \text{(see (91) and (93))}
\end{aligned} \tag{94}
$$

We can further simplify the first line of (94) by replacing $N_{x,\epsilon}$ with $N_{X,\min,\epsilon}$ as follows. By invoking (11) in the second line below, we note that

$$
\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \left(1+\frac{n-2}{N_{X,\min,\epsilon}(n+2)}\right)\dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)}\cdot\mathrm{trace}\left[\dot{\Sigma}_X\right]\cdot I_n\right\|_F\\
&= \left\|\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X}\left[\ddot{\Sigma}_{X,Y_{X,\epsilon}}\right] - \left(1+\frac{n-2}{N_{X,\min,\epsilon}(n+2)}\right)\frac{1}{N}\sum_{x\in X}\nabla f(x)\nabla f(x)^*\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \frac{1}{N} \sum_{x \in X} \|\nabla f(x)\|_2^2 \cdot I_n \Bigg\|_F \\
&\leq \left(\frac{\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon}+1\right)\mathrm{L}_f^2 \right) + \max_{x \in X} \left| \frac{1}{N_{x,\epsilon}} - \frac{1}{N_{X,\min,\epsilon}} \right| \cdot \max_{x \in X} \|\nabla f(x)\|_2^2 \cdot (1 + \|I_n\|_F) \qquad \text{(see (94))} \\
&\leq \left(\frac{\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon}+1\right)\mathrm{L}_f^2 \right) + \max_{x \in X} \left| \frac{1}{N_{x,\epsilon}} - \frac{1}{N_{X,\min,\epsilon}} \right| \cdot \mathrm{L}_f^2 \left(1 + \sqrt{n}\right) \qquad \text{(see (8))} \\
&\leq \left(\frac{\mathrm{B}''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2\,\mathrm{B}'_{\mu,\epsilon}\left(\mathrm{B}'_{\mu,\epsilon}+1\right)\mathrm{L}_f^2 \right) + \frac{\mathrm{L}_f^2\,(1+\sqrt{n})}{N_{X,\min,\epsilon}} \qquad \text{(see (37))} \\
&=: \frac{1}{2}\,\mathrm{B}_{\mu,\epsilon}\,. \qquad (95)
\end{aligned}
$$

Next, we replace $\mathrm{trace}[\dot{\Sigma}_X]$ in the first line of (95) with $\mathrm{trace}[\dddot{\Sigma}_{X,Y_{X,\epsilon}}]$. To that end, we first notice the following consequence of (95):

$$
\begin{aligned}
&\left| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\mathrm{trace}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right] - \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \mathrm{trace}\left[\dot{\Sigma}_X \right] - \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \cdot \mathrm{trace}\left[\dot{\Sigma}_X \right] \right| \\
&= \left| \mathrm{trace}\left[\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \mathrm{trace}\left[\dot{\Sigma}_X \right] \cdot I_n \right] \right| \\
&\leq \sqrt{n} \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \mathrm{trace}\left[\dot{\Sigma}_X \right] \cdot I_n \right\|_F \\
&\leq \frac{\sqrt{n}}{2}\,\mathrm{B}_{\mu,\epsilon}, \qquad \text{(see (95))} \qquad (96)
\end{aligned}
$$

where the second line uses the fact that $\mathrm{trace}\,[I_n] = n$. Also, the third line follows from the inequality $|\mathrm{trace}[A]| \leq \sqrt{n}\|A\|_F$ for an arbitrary matrix $A \in \mathbb{R}^{n \times n}$. After rearranging, (96) immediately implies that

$$
\begin{aligned}
&\left| \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\mathrm{trace}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right] - \mathrm{trace}\left[\dot{\Sigma}_X \right] \right| \\
&\quad \leq \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \frac{\sqrt{n}}{2}\,\mathrm{B}_{\mu,\epsilon}\,. \qquad (97)
\end{aligned}
$$

The above inequality enables us to remove $\mathrm{trace}[\dot{\Sigma}_X]$ from the first line of (95):

$$
\begin{aligned}
&\Bigg\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X \\
&\qquad - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[\mathrm{trace}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right] \cdot I_n \Bigg\|_F \\
&\leq \frac{1}{2}\,\mathrm{B}_{\mu,\epsilon} + \frac{n}{N_{X,\min,\epsilon}(n+2)} \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \frac{\sqrt{n}}{2}\,\mathrm{B}_{\mu,\epsilon} \cdot \|I_n\|_F \qquad \text{(see (95) and (97))} \\
&= \frac{1}{2} \left(1 + \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \right) \mathrm{B}_{\mu,\epsilon}\,. \qquad \left(\|I_n\|_F = \sqrt{n} \right) \qquad (98)
\end{aligned}
$$

Lastly, (98) can be rewritten as follows by introducing $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$:

$$
\begin{aligned}
&\left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X} \left[\ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \dot{\Sigma}_X \right\|_F \\
&\leq \frac{1}{2} \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \left(1 + \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \right) \mathrm{B}_{\mu,\epsilon}
\end{aligned}
$$

$$\leq \mathrm{B}_{\mu,\epsilon}, \qquad \text{(the factor in front of } \mathrm{B}_{\mu,\epsilon} \text{ does not exceed 1)} \tag{99}$$

where, above, we set

$$\begin{aligned}
&\dddot{\Sigma}_{X,Y_{X,\epsilon}} \\
&:= \left(1+\frac{n-2}{N_{X,\min,\epsilon}(n+2)}\right)^{-1}\left(\ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \frac{n}{N_{X,\min,\epsilon}(n+2)}\left(1+\frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)}\right)^{-1}\cdot \operatorname{trace}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}}\right]\cdot I_n\right) \\
&= \left(1+\frac{1-\frac{2}{n}}{1+\frac{2}{n}}\cdot N_{X,\min,\epsilon}^{-1}\right)^{-1}\left(\ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \left(\left(1+\frac{2}{n}\right)N_{X,\min,\epsilon}+n+1-\frac{2}{n}\right)^{-1}\operatorname{trace}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}}\right]\cdot I_n\right) \\
&= \left(1+\frac{1-\frac{2}{n}}{1+\frac{2}{n}}\cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \\
&\quad\cdot\left(\frac{1}{N}\sum_{x\in X}\ddot{\nabla}_{Y_{X,\epsilon}}f(x)\ddot{\nabla}_{Y_{X,\epsilon}}f(x)^* - \left(\left(1+\frac{2}{n}\right)N_{X,\min,\epsilon}+n+1-\frac{2}{n}\right)^{-1}\frac{1}{N}\sum_{x\in X}\left\|\ddot{\nabla}_{Y_{x,\epsilon}}f(x)\right\|_2^2\cdot I_n\right) \\
&= \frac{1}{N}\left(1+\frac{1-\frac{2}{n}}{1+\frac{2}{n}}\cdot N_{X,\min,\epsilon}^{-1}\right)^{-1}\cdot\left(\sum_{N_{x,\epsilon}\geq N_{X,\min,\epsilon}}\ddot{\nabla}_{Y_{X,\epsilon}}f(x)\ddot{\nabla}_{Y_{X,\epsilon}}f(x)^*\right. \\
&\qquad\left. - \left(\left(1+\frac{2}{n}\right)N_{X,\min,\epsilon}+n+1-\frac{2}{n}\right)^{-1}\sum_{N_{x,\epsilon}\geq N_{X,\min,\epsilon}}\left\|\ddot{\nabla}_{Y_{x,\epsilon}}f(x)\right\|_2^2\cdot I_n\right),
\end{aligned} \tag{100}$$

where the third identity uses (89) and the last line above follows from (37). Because $\mathrm{B}_{\mu,\epsilon}$ does not depend on $N_{X,\epsilon}$, it is easy to remove the conditioning on $N_{X,\epsilon}$ in (99):

$$\begin{aligned}
\left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]-\dot{\Sigma}_X\right\|_F &= \left\|\mathbb{E}\left[\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]-\dot{\Sigma}_X\right]\right\|_F \\
&\leq \mathbb{E}\left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X}\left[\dddot{\Sigma}_{X,Y_{X,\epsilon}}\right]-\dot{\Sigma}_X\right\|_F \qquad \text{(Jensen's inequality)} \\
&\leq \mathbb{E}\,\mathrm{B}_{\mu,\epsilon} \qquad \text{(see (99))} \\
&= \mathrm{B}_{\mu,\epsilon}\,. \qquad \text{(see (95))}
\end{aligned} \tag{101}$$

Consider also the following special case. Let μ be the uniform probability measure on $\mathbb{D}$ and fix x within the ϵ-interior of $\mathbb{D}$, namely $x\in\mathbb{D}_\epsilon$. Also draw y from $\mu_{x,\epsilon}$, namely $y|x\sim\mu_{x,\epsilon}$ (see (17)). Then, as stated in Proposition 3, $\mathrm{B}'_{\mu,\epsilon}=0$. Furthermore, it is known [47] that

$$P_{x,y}\cdot\nabla f(x) \overset{\text{dist.}}{=} \omega\cdot\nabla f(x)+\sqrt{\omega-\omega^2}\,\|\nabla f(x)\|_2\cdot A\alpha, \tag{102}$$

where ω follows the beta distribution, α is uniformly distributed on the unit sphere in $\mathbb{R}^{n-1}$, and the two variables are independent, i.e.,

$$\omega\sim\operatorname{beta}\left(\frac{1}{2},\frac{n-1}{2}\right), \qquad \alpha\sim\operatorname{uniform}\left(\mathbb{S}^{n-2}\right), \qquad \omega\perp\!\!\!\perp\alpha.$$

Finally, $A\in\mathbb{R}^{n\times(n-1)}$ in (102) is an orthonormal basis for the directions orthogonal to $\nabla f(x)\in\mathbb{R}^n$, namely

$$A^*\nabla f(x)=0, \qquad A^*A=I_{n-1}. \tag{103}$$

Using the expressions for the first and second moments of the beta distribution in the fourth line below, we write that

$$\begin{aligned}
&\mathbb{E}_{y|x}\left[P_{x,y}\nabla f(x)\nabla f(x)^*P_{x,y}\right] \\
&= \mathbb{E}\left[\left(\omega\nabla f(x)+\sqrt{\omega-\omega^2}\,\|\nabla f(x)\|_2\cdot A\alpha\right)\cdot\left(\omega\nabla f(x)+\sqrt{\omega-\omega^2}\,\|\nabla f(x)\|_2\cdot A\alpha\right)^*\right] \qquad \text{(see (102))} \\
&= \mathbb{E}\left[\omega^2\right]\cdot\nabla f(x)\nabla f(x)^*+\mathbb{E}\left[\omega-\omega^2\right]\|\nabla f(x)\|_2^2\cdot A\cdot\mathbb{E}\left[\alpha\alpha^*\right]\cdot A^* \qquad (\omega\perp\!\!\!\perp\alpha, \quad \mathbb{E}\alpha=0)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{n-1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot A \cdot \frac{I_{n-1}}{n-1} \cdot A^* \qquad \left(\mathbb{E}\left[\alpha \alpha^*\right] = \frac{I_{n-1}}{n-1} \right) \\
&= \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot A A^* \\
&= \frac{2}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot \left(\frac{\nabla f(x)}{\|\nabla f(x)\|_2} \cdot \frac{\nabla f(x)^*}{\|\nabla f(x)\|_2} + A A^* \right) \\
&= \frac{2}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot I_n, \qquad \text{(see (103))}
\end{aligned} \tag{104}
$$

and, consequently, $\mathrm{B}''_{\mu,\epsilon} = 0$. Furthermore, assume that $N_{x,\epsilon} = N_{x',\epsilon}$ for every pair $x, x' \in X$. Then, we observe that the upper bound in (95) can be improved to $\mathrm{B}_{\mu,\epsilon} = 0$, namely $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ is an unbiased estimator of $\dot{\Sigma}_X$, conditioned on the event $\mathcal{E}_2$. This completes the proof of Lemma 3.

H Proof of Lemma 4

Throughout, X is fixed and we assume that the event $\mathcal{E}_2$ holds (see (37)). We also consider $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$ (see (16)) to be any fixed neighborhood structure consistent with $\mathcal{E}_2$.

To bound the estimation error, we write that

$$
\begin{aligned}
&\left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \\
&\leq \left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F + \left\| \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \dot{\Sigma}_X \right\|_F \qquad \text{(triangle inequality)} \\
&\leq \left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F + \mathrm{B}_{\mu,\epsilon} \,. \qquad \text{(see (99))}
\end{aligned} \tag{105}
$$

It therefore suffices to study the concentration of $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ about its expectation. In fact, as we show next, it is more convenient to first study the concentration of $\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$ instead, where

$$
\ddddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \sum_{x \in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^*, \tag{106}
$$

$$
\ddot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x) \ \in \mathbb{R}^n, \qquad \forall x \in X.
$$

Indeed, conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$, the expression for $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ in (28) simplifies to

$$
\dddot{\Sigma}_{X,Y_{X,\epsilon}} = \left(1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1} \right)^{-1} \left(\ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \frac{\operatorname{trace}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right]}{\left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right). \tag{107}
$$

Consequently, the deviation of $\dddot{\Sigma}_{X,Y_{X,\epsilon}}$ about its expectation can be bounded as:

$$
\begin{aligned}
&\left\| \dddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\dddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \\
&\leq \left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \\
&\quad + \left(\left(1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \cdot \left| \operatorname{trace}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\operatorname{trace}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right] \right| \cdot \|I_n\|_F \\
&\leq \left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \\
&\quad + \left(\left(1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \cdot \sqrt{n} \left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \cdot \sqrt{n} \qquad \left(\|I_n\|_F = \sqrt{n} \right) \\
&= \left(1 + \frac{n}{\left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \right) \left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon} | \mathcal{E}_2, \overline{N}_{X,\epsilon}, X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F
\end{aligned}
$$

$$\leq 2 \left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F . \qquad \text{(the factor above does not exceed 2)} \tag{108}$$

Above, the first inequality uses (107). We also used the linearity of trace and the inequality $|\operatorname{trace}[A]| \leq \sqrt{n}\|A\|_F$ for arbitrary $A \in \mathbb{R}^{n\times n}$. Thanks to (108), it suffices to study the concentration of $\ddddot{\Sigma}_{X,Y_{X,\epsilon}}$ about its expectation. The following result is proved in Appendix I.

Lemma 5. *Fix X and $\epsilon \in (0, \epsilon_{\mu,X}]$. If $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{X,\epsilon}) \geq \log(n)$, then conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$,*

$$\left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \lesssim \gamma_7\gamma_2^2 \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}]\, \mathrm{L}_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}, \tag{109}$$

for $\gamma_7 \geq 1$ and $\gamma_2 \geq 3$, except with a probability

$$\lesssim e^{-\gamma_7} + n^{2-\log\gamma_7} + N_{X,\epsilon}^{(1-\gamma_2\log(N_{X,\epsilon}))}. \tag{110}$$

Combining (105), (108), and Lemma 5 tells us that if $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{X,\epsilon}) \geq \log(n)$, then conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$,

$$\left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \lesssim \mathrm{B}_{\mu,\epsilon} + \gamma_7\gamma_2^2 \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}]\, \mathrm{L}_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))},$$

except with the probability appearing in (110). We observe that this expression and probability do not depend on $\overline{N}_{X,\epsilon}$, and so the same statement holds with the same probability when we condition only on $\mathcal{E}_2, X$. This completes the proof of Lemma 4.

I Proof of Lemma 5

Throughout, X is fixed and the event $\mathcal{E}_2$ holds. We will also use $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x\in X}$ to summarize the neighborhood structure of data (see (16)). As in Appendix G, we again decompose $\ddddot{\Sigma}_{X,Y_{X,\epsilon}}$ into "diagonal" and "off-diagonal" components:

$$\begin{aligned}
\ddddot{\Sigma}_{X,Y_{X,\epsilon}} &= \frac{1}{N} \sum_{x\in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \qquad \text{(see (106))} \\
&= \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \qquad \text{(see (29))} \\
&= \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y\in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} + \frac{n^2}{N} \sum_{x\in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y'\in Y_{x,\epsilon}} 1_{y\neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \\
&=: \ddddot{\Sigma}^d_{X,Y_{X,\epsilon}} + \ddddot{\Sigma}^o_{X,Y_{X,\epsilon}}.
\end{aligned} \tag{111}$$

This decomposition, in turn, allows us to break down the error into the contribution of the diagonal and off-diagonal components:

$$\begin{aligned}
&\left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F \\
&\leq \left\| \ddddot{\Sigma}^d_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\ddddot{\Sigma}^d_{X,Y_{X,\epsilon}} \right] \right\|_F + \left\| \ddddot{\Sigma}^o_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\ddddot{\Sigma}^o_{X,Y_{X,\epsilon}} \right] \right\|_F .
\end{aligned} \tag{112}$$

We bound the norms on the right-hand side above separately in Appendices J and K, respectively, and report the results below.

Lemma 6. *Fix X and $\epsilon \in (0, \epsilon_{\mu,X}]$. Consider the event*

$$\mathcal{E}_1 := \left\{ \max_{x\in X} \max_{y\in Y_{X,\epsilon}} \|P_{x,y}\nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \right\}, \tag{113}$$

for $Q_{X,\epsilon} > \mathrm{K}_\mu^{-1}$ to be set later. Then, conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$, it holds that

$$\left\| \dddot{\Sigma}^d_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}\right] \right\|_F \lesssim \gamma_6 \cdot \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2\, n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} + 2n^2 L_f^2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right], \tag{114}$$

for $\gamma_6 \geq 1$ and except with a probability of at most $e^{-\gamma_6} + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}[\mathcal{E}_1^C]$.

Lemma 7. *Fix X and $\epsilon \in (0, \epsilon_{\mu,X}]$. Let $\widetilde{Y}_{X,\epsilon}$ contain $Y_{X,\epsilon}$ and three independent copies of it. That is, $\widetilde{Y}_{X,\epsilon} = \cup_{x\in X}\widetilde{Y}_{x,\epsilon}$, where each $\widetilde{Y}_{x,\epsilon}$ contains $Y_{x,\epsilon}$ and three independent copies of it. Consider the event $\mathcal{E}_1$ defined in (113) for $Q_{X,\epsilon} > \mathrm{K}_\mu^{-1}$ to be set later. Consider also the event*

$$\mathcal{E}_3 := \left\{ \max_{x\in X} \max_{y\in\widetilde{Y}_{x,\epsilon}} \max_{i\in[1:n]} \|P_{x,y}e_i\|_2^2 \leq \frac{Q_{X,\epsilon}}{n} \right\} \bigcap \left\{ \max_{x\in X} \max_{y\in\widetilde{Y}_{x,\epsilon}} \|P_{x,y}\nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \right\}. \tag{115}$$

Here, $e_i \in \mathbb{R}^n$ is the ith canonical vector. Assume that

$$\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] \lesssim \left(\frac{\log n}{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}} \right)^{\frac{\log n}{2}}, \tag{116}$$

and $1 \leq \log n \leq N$. Then, conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$, it holds that

$$\left\| \ddddot{\Sigma}^o_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\ddddot{\Sigma}^o_{X,Y_{X,\epsilon}}\right] \right\|_F \lesssim \gamma_7 \cdot \sqrt{\log n} \cdot \frac{n \cdot \max[Q_{X,\epsilon}, Q_{X,\epsilon}^2] \cdot \mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 2n^2 L_f^2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right], \tag{117}$$

for $\gamma_7 \geq 1$ and except with a probability of at most $e^{-\gamma_7} + n^2 \cdot n^{-\log\gamma_7} + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]$.

Before we can apply Lemmas 6 and 7 to the right-hand side of (112), however, we must show that the events $\mathcal{E}_1$ and $\mathcal{E}_3$ are very likely to happen. Owing to Assumption 1, this is indeed the case for the right choice of $Q_{X,\epsilon}$ as shown in Appendix L and summarized below.

Lemma 8. *Fix $\epsilon \in (0, \epsilon_\mu]$ and X. Suppose that $Q_{X,\epsilon} = \gamma_2\,\mathrm{K}_\mu^{-1}\log^2(N_{X,\epsilon})$ for $\gamma_2 \geq 3$. Then, conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$, it holds that*

$$\Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right] \lesssim N_{X,\epsilon}^{(1-\gamma_2\log(N_{X,\epsilon}))}. \tag{118}$$

Moreover, if $1 \leq \log(n) \leq \log(N_{X,\epsilon})$ and if $N_{X,\epsilon}$ is large enough such that $\Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right] \leq \frac{1}{2}$, then conditioned on $\mathcal{E}_1, \mathcal{E}_2, \overline{N}_{X,\epsilon}, X$, the requirement in (116) is satisfied.

Revisiting (112), we put all the pieces together to conclude that if $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{X,\epsilon}) \geq \log(n)$, then conditioned on $\mathcal{E}_2, \overline{N}_{X,\epsilon}, X$,

$$\begin{aligned}
&\left\| \ddddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\ddddot{\Sigma}_{X,Y_{X,\epsilon}}\right] \right\|_F \\
&\leq \left\| \dddot{\Sigma}^d_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}\right] \right\|_F + \left\| \ddddot{\Sigma}^o_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\ddddot{\Sigma}^o_{X,Y_{X,\epsilon}}\right] \right\|_F \qquad \text{(see (112))} \\
&\lesssim \gamma_6 \cdot \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2\, n}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + \gamma_7 \cdot \sqrt{\log n} \cdot \frac{n \cdot \max[Q_{X,\epsilon}, Q_{X,\epsilon}^2] \cdot \mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 4n^2 L_f^2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right] \qquad \text{(see Lemmas 6 and 7)} \\
&\lesssim \gamma_7 \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} \cdot \max[Q_{X,\epsilon}, Q_{X,\epsilon}^2]\,\mathrm{L}_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2\log(N_{X,\epsilon}))} \qquad \text{(set } \gamma_6 = \gamma_7\text{; see Lemma 8)}
\end{aligned}$$

$$\lesssim \gamma_7\gamma_2^2 \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[\mathrm{K}_\mu^{-1}, \mathrm{K}_\mu^{-2}]\, \mathrm{L}_f^2$$
$$+ 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} \qquad \text{(choice of } Q_{X,\epsilon} \text{ in Lemma 8)} \tag{119}$$

except with a probability of at most

$$e^{-\gamma_6} + e^{-\gamma_7} + n^2 \cdot n^{-\log\gamma_7} + 2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X} \left[\mathcal{E}_1^C\right] \qquad \text{(see Lemmas 6 and 7)}$$
$$\lesssim e^{-\gamma_6} + e^{-\gamma_7} + n^{2-\log\gamma_7} + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} \qquad \text{(see Lemma 8)}$$
$$\lesssim e^{-\gamma_7} + n^{2-\log\gamma_7} + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}. \qquad \text{(choice of } \gamma_6 \text{ in (119))} \tag{120}$$

This completes the proof of Lemma 5.

J Proof of Lemma 6

Throughout, X and the neighborhood structure $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x\in X}$ (see (16)) are fixed. Moreover, we assume that the event $\mathcal{E}_2$ holds (see (37)). In addition, for $Q_{X,\epsilon} \geq \mathrm{K}_\mu^{-1}$ to be set later, we condition on the following event:

$$\mathcal{E}_1 := \left\{ \max_{x\in X} \max_{y\in Y_{x,\epsilon}} \|P_{x,y}\nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \right\}. \tag{121}$$

By the definition of $\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}$ in (111), we observe that

$$\left\| \dddot{\Sigma}^d_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}\right] \right\|_F$$
$$= \frac{n^2}{N} \left\| \sum_{x\in X} \sum_{y\in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \left(P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y} - \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x} \left[P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\right] \right) \right\|_F \qquad \text{(see (111))}$$
$$=: \frac{n^2}{N} \left\| \sum_{x\in X} \sum_{y\in Y_{x,\epsilon}} A_{x,y} \right\|_F, \tag{122}$$

where $\{A_{x,y}\}_{x,y} \subset \mathbb{R}^{n\times n}$ are zero-mean independent random matrices. To bound this sum, we appeal to Proposition 2 by computing the b and σ parameters below. For arbitrary $x \in X$ and $y \in Y_{x,\epsilon}$, note that

$$\|A_{x,y}\|_F$$
$$= \frac{1}{N_{x,\epsilon}^2} \left\| P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y} - \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x} \left[P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\right] \right\|_F \qquad \text{(see (122))}$$
$$\leq \frac{1}{N_{x,\epsilon}^2} \|P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\|_F + \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x} \|P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\|_F \qquad \text{(Jensen's inequality)}$$
$$= \frac{1}{N_{x,\epsilon}^2} \|P_{x,y}\nabla f(x)\|_2^2 + \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x} \|P_{x,y}\nabla f(x)\|_2^2$$
$$\leq \frac{1}{N_{x,\epsilon}^2} \|P_{x,y}\nabla f(x)\|_2^2 + \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_2,x} \|P_{x,y}\nabla f(x)\|_2^2 \qquad \text{(see (121))}$$
$$\lesssim \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \max_{x\in X} \max_{y\in Y_{x,\epsilon}} \|P_{x,y}\nabla f(x)\|_2^2 + \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \max_{x\in X} \frac{\|\nabla f(x)\|_2^2}{\mathrm{K}_\mu\, n} \qquad \text{(see [11, Lemma 5.5])}$$
$$\leq \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \max_{x\in X} \max_{y\in Y_{x,\epsilon}} \|P_{x,y}\nabla f(x)\|_2^2 + \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \frac{\mathrm{L}_f^2}{\mathrm{K}_\mu\, n} \qquad \text{(see (8))}$$
$$=: \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} + \frac{1}{\min_{x\in X} N_{x,\epsilon}^2} \cdot \frac{\mathrm{L}_f^2}{\mathrm{K}_\mu\, n} \qquad \text{(see (121))}$$

$$\lesssim \frac{1}{\min_{x\in X} N_{x,\epsilon}^2}\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \qquad \left(\text{when } Q_{X,\epsilon}\geq \mathrm{K}_\mu^{-1}\right)$$
$$=: b. \tag{123}$$

On the other hand, note that

$$\begin{aligned}
&\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}} \mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x}\left\|A_{x,y}\right\|_F^2\\
&=\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\cdot\mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x}\left\|P_{x,y}\nabla f(x)\nabla f(x)^*P_{x,y}-\mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x}\left[P_{x,y}\nabla f(x)\nabla f(x)^*P_{x,y}\right]\right\|_F^2\\
&\leq\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\cdot\mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x}\left\|P_{x,y}\nabla f(x)\nabla f(x)^*P_{x,y}\right\|_F^2\\
&=\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\cdot\mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_1,\mathcal{E}_2,x}\left\|P_{x,y}\nabla f(x)\right\|_2^4\\
&\leq\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\cdot\mathbb{E}_{y|\overline{N}_{x,\epsilon},\mathcal{E}_2,x}\left\|P_{x,y}\nabla f(x)\right\|_2^4 \qquad \text{(see (121))}\\
&\lesssim\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\cdot\frac{1}{\mathrm{K}_\mu^2\,n^2}\cdot\max_{x\in X}\left\|\nabla f(x)\right\|_2^4 \qquad \text{(see [11, Lemma 5.5])}\\
&\leq\frac{N}{\min_{x\in X}N_{x,\epsilon}}\cdot\frac{\mathrm{L}_f^4}{\mathrm{K}_\mu^2\,n^2}. \qquad (\#X=N,\ \#Y_{x,\epsilon}=N_{x,\epsilon},\text{ see (8)})\\
&=:\sigma^2.
\end{aligned}\tag{124}$$

where the second line uses (122). The third line above uses the fact that $\mathbb{E}\left\|Z-\mathbb{E}[Z]\right\|_F^2\leq\mathbb{E}\left\|Z\right\|_F^2$ for a random matrix Z. It follows that

$$\max[b,\sigma]\lesssim\sqrt{\frac{N}{\min_{x\in X}N_{x,\epsilon}}}\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}, \qquad \text{if } Q_{X,\epsilon}\geq\mathrm{K}_\mu^{-1}. \tag{125}$$

Thus, in light of Proposition 2, and conditioned on $\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X$, it follows that

$$\begin{aligned}
\left\|\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}-\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}\right]\right\|_F &= \frac{n^2}{N}\left\|\sum_{x\in X}\sum_{y\in Y_{x,\epsilon}}A_{x,y}\right\|_F \qquad \text{(see (122))}\\
&\lesssim\frac{n^2}{N}\cdot\gamma_6\cdot\max[b,\sigma]\\
&\lesssim\frac{n^2}{N}\cdot\gamma_6\sqrt{\frac{N}{\min_x N_{x,\epsilon}}}\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \qquad \text{(see (125))}\\
&=\gamma_6\cdot\frac{n}{\sqrt{N\cdot\min_x N_{x,\epsilon}}}\cdot Q_{X,\epsilon}\,\mathrm{L}_f^2\\
&\lesssim\gamma_6\cdot\frac{n}{\sqrt{N\cdot\min_x\frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})}\cdot N_{X,\epsilon}}}\cdot Q_{X,\epsilon}\,\mathrm{L}_f^2 \qquad \text{(see (37))}\\
&=\gamma_6\cdot\frac{n}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\cdot Q_{X,\epsilon}\,\mathrm{L}_f^2, \qquad \text{(see (21))}
\end{aligned}\tag{126}$$

for $\gamma_6\geq 1$ and except with a probability of at most $e^{-\gamma_6}$. Before we can remove the conditioning on the event $\mathcal{E}_1$, we use the law of total expectation to write

$$\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^d_{X,Y_{X,\epsilon}}\right]$$

$$= \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}[\mathcal{E}_1] + \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1^C,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right],$$

from which it follows that

$$\begin{aligned}
&\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] \\
&= \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\left(\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1^C,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right). \qquad (127)
\end{aligned}$$

Since for any X, $Y_{X,\epsilon}$, we have

$$\begin{aligned}
\left\|\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right\|_F &\le \frac{n^2}{N}\sum_{x\in X}\sum_{y\in Y_{Xx,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\left\|P_{x,y}\nabla f(x)\nabla f(x)^* P_{x,y}\right\|_F \qquad \text{(see (111))} \\
&\le \frac{n^2}{N}\sum_{x\in X}\sum_{y\in Y_{Xx,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\left\|P_{x,y}\nabla f(x)\right\|_2^2 \\
&\le \frac{n^2}{N}\sum_{x\in X}\sum_{y\in Y_{Xx,\epsilon}}\frac{1}{N_{x,\epsilon}^2}\left\|\nabla f(x)\right\|_2^2 \\
&\le \frac{n^2}{N}\sum_{x\in X}\sum_{y\in Y_{Xx,\epsilon}}\frac{1}{N_{x,\epsilon}^2}L_f^2 \qquad \text{(see (8))} \\
&= \frac{n^2}{N}\sum_{x\in X}\frac{1}{N_{x,\epsilon}}L_f^2 \\
&\le n^2 L_f^2, \qquad (128)
\end{aligned}$$

we conclude that

$$\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right\|_F \\
&\le \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1^C,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right\|_F \qquad \text{(see (127))} \\
&\le 2n^2 L_f^2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]. \qquad \text{(triangle inequality and (128))} \qquad (129)
\end{aligned}$$

Lastly, we remove the conditioning on the event $\mathcal{E}_1$ as follows:

$$\begin{aligned}
&\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right\|_F \gtrsim \gamma_6\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2\,n}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 2n^2L_f^2\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\right] \\
&\le \Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right\|_F \gtrsim \gamma_6\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2\,n}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 2n^2L_f^2\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\right] \\
&\quad + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right] \qquad \text{(see (53))} \\
&\le \Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{d}_{X,Y_{X,\epsilon}}\right]\right\|_F \gtrsim \gamma_6\cdot\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2\,n}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\right] \\
&\quad + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right] \qquad \text{(see (129))} \\
&\le e^{-\gamma_6} + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]. \qquad \text{(see (126))} \qquad (130)
\end{aligned}$$

The proof of Lemma 6 is now complete.

K Proof of Lemma 7

Throughout, X and the neighborhood structure $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x\in X}$ (see (16)) are fixed. Moreover, we assume that the event $\mathcal{E}_2$ holds (see (37)). In addition, for $Q_{X,\epsilon} \geq \mathrm{K}_\mu^{-1}$ to be set later, we condition on $\mathcal{E}_1$ as defined in (121).

Let us index X as $X = \{x_s\}_{s=1}^N$. For each $x_s \in X$, we index its neighbors $Y_{x_s,\epsilon}$ as $Y_{x_s,\epsilon} = \{y_{sk}\}_{k=1}^{N_{x_s,\epsilon}}$, where $N_{x_s,\epsilon} = \#Y_{x_s,\epsilon}$ is the number of neighbors of x_s (within radius of ϵ). Recalling the definition of $\ddot{\Sigma}^o_{X,Y_{X,\epsilon}}$ from (111), we aim to find an upper bound for

$$
\begin{aligned}
&\left\| \ddot{\Sigma}^o_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\ddot{\Sigma}^o_{X,Y_{X,\epsilon}}\right]\right\|_F \\
&= \frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k\neq l} \frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}}\nabla f(x_s)\nabla f(x_s)^* P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left[P_{x_s,y_s}\right]\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left[P_{x_s,y_s}\right]\right)\right\|_F \\
&\leq \frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k\neq l} \frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\right\|_F \\
&\quad + \frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k\neq l} \frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F \\
&\quad + \frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k\neq l} \frac{1}{N_{x_s,\epsilon}^2}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\right\|_F, \qquad (131)
\end{aligned}
$$

after which we will remove the conditioning on $\mathcal{E}_1$. Above, $y_s|\mathcal{E}_1, x_s$ is distributed according to the restriction of $\mu_{x_s,\epsilon}$ to the event $\mathcal{E}_1$. In the following subsections, we separately bound each of the three norms in the last line above.

K.1 First norm

In this section, we bound the first norm in the last line of (131) by writing it as a chaos random variable. Let us first write that

$$
\begin{aligned}
&\frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k\neq l} \frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\right\|_F \\
&=: \frac{n^2}{N}\left\| \sum_{s=1}^N \sum_{k,l=1}^{N_{x_s,\epsilon}} A_{skl}\right\|_F \\
&= \frac{n^2}{N}\sqrt{\sum_{i,j=1}^n \left| \sum_{s=1}^N \sum_{k,l=1}^{N_{x_s}} A_{skl}[i,j]\right|^2}. \qquad (132)
\end{aligned}
$$

Above, we also conveniently defined the matrices $\{A_{skl}\}_{s,k,l} \subset \mathbb{R}^{n\times n}$ as

$$
A_{skl} := \frac{1}{N_{x_s,\epsilon}^2}\begin{cases}\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right), & k\neq l, \\ 0, & k = l, \end{cases} \qquad (133)
$$

for every $s \in [1:N]$ and $k, l \in [1 : N_{x_s,\epsilon}]$. By their definition above, the random matrices $\{A_{skl}\}_{s,k,l}$ enjoy the following properties:

$$
A_{skk} = 0, \qquad \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[A_{skl}\right] = 0, \qquad s \in [1:N],\ k,l \in [1:N_{x_s,\epsilon}]. \qquad (134)
$$

With fixed $s \in [1:N]$ and $i,j \in [1:n]$, we may use $\{A_{skl}[i,j]\}_{k,l}$ to form a new matrix A_{sij} as

$$A_{sij} := [A_{skl}[i,j]]_{k,l} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}},$$

or, equivalently,

$$A_{sij}[k,l] := A_{skl}[i,j], \qquad k,l \in [1 : N_{x_s,\epsilon}]. \tag{135}$$

Let A_{ij} be the block-diagonal matrix formed from $\{A_{sij}\}_s \subset \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}}$, i.e.,

$$A_{ij} = \begin{bmatrix} A_{1ij} & & & \\ & A_{2ij} & & \\ & & \ddots & \\ & & & A_{Nij} \end{bmatrix} \in \mathbb{R}^{N_{X,\epsilon} \times N_{X,\epsilon}}. \tag{136}$$

where we used the fact that $N_{X,\epsilon} = \sum_{s=1}^{N} N_{x_s,\epsilon}$ to calculate the dimensions of A_{ij}. In particular, (134) implies that

$$\begin{aligned} A_{ij}[sk,sk] &= 0, \\ \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[A_{ij}[sk,tl]\right] &= 0, \qquad s,t \in [1:N],\ k \in [1:N_{x_s,\epsilon}],\ l \in [1:N_{x_t,\epsilon}], \end{aligned} \tag{137}$$

where, ignoring the standard convention, we indexed the entries of A_{ij} so that sk corresponds to the kth row of the sth block (and hence does not stand for the product of s and k). With this new notation, we revisit (132) to write that

$$\begin{aligned} &\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\right\|_F \\ &= \frac{n^2}{N}\sqrt{\sum_{i,j=1}^{n}\left|\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} A_{ij}[sk,tl]\right|^2} \qquad \text{(see (132))} \\ &=: \frac{n^2}{N}\sqrt{\sum_{i,j=1}^{n} a_{ij}^2}. \end{aligned} \tag{138}$$

For fixed $i,j \in [1:n]$, let us next focus on the random variable a_{ij}.

K.1.1 Tail Bound for a_{ij}

Recall that the pth moment of a random variable z is defined as $\mathbb{E}^p[z] := (\mathbb{E}[|z|^p])^{\frac{1}{p}}$. Fix $i,j \in [1:n]$. In order to bound a_{ij}, we

- First control its moments, namely

$$\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right] = \mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left|\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} A_{ij}[sk,tl]\right|, \qquad \forall p \geq 1. \tag{139}$$

- Second we use Markov's inequality to find a tail bound for a_{ij} (given its moments).

Each step is discussed in a separate subsection below.

K.1.2 Moments of a_{ij}

In order to control the moments of a_{ij}, we take the following steps:

- *symmetrization*,

- *decoupling*,
- *modulation* with *Rademacher sequences*, and finally
- bounding the moments of the resulting *decoupled chaos* random variable.

Each of these steps is detailed in a separate paragraph below.

Symmetrization To control the moments of a_{ij}, we first use a symmetrization argument as follows. With $s \in [1:N]$ and conditioned on x_s and $\mathcal{E}_1$, let $Y^{\mathrm{i}}_{x_s,\epsilon} \in \mathbb{R}^{n \times N_{x_s,\epsilon}}$ be an independent copy of $Y_{x_s,\epsilon}$. Then note that

$$
\begin{aligned}
&\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}[a_{ij}] \\
&= \mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} A_{ij}[sk,tl]\right] \\
&= \mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N^2_{x_s,\epsilon}} e_i^*\left(P_{x_s,y_{sk}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{y_s|\mathcal{E}_1,x_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{y_s|\mathcal{E}_1,x_s}]\right)e_j\right] \\
&= \mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N^2_{x_s,\epsilon}} e_i^*\left(P_{x_s,y_{sk}} - \mathbb{E}_{y^{\mathrm{i}}_{sk}|\mathcal{E}_1,x_s}[P_{x_s,y^{\mathrm{i}}_{sk}}]\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - \mathbb{E}_{y^{\mathrm{i}}_{s,l}|\mathcal{E}_1,x_s}[P_{x_s,y^{\mathrm{i}}_{s,l}}]\right)e_j\right] \\
&= \mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathbb{E}_{Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N^2_{x_s,\epsilon}} e_i^*\left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}}\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - P_{x_s,y^{\mathrm{i}}_{s,l}}\right)e_j\right] \\
&\qquad \text{(independence)} \\
&\leq \mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N^2_{x_s,\epsilon}} e_i^*\left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}}\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - P_{x_s,y^{\mathrm{i}}_{s,l}}\right)e_j\right] \\
&\qquad \text{(Jensen's inequality)} \\
&= \mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} B_{ij}[sk,tl]\right], \qquad (140)
\end{aligned}
$$

where we defined the block-diagonal matrix $B_{ij} \in \mathbb{R}^{N_{X,\epsilon}\times N_{X,\epsilon}}$ such that

$$
\begin{aligned}
&B_{ij}[sk,tl] \\
&= \begin{cases} N^{-2}_{x_s,\epsilon}\cdot e_i^*\left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}}\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}} - P_{x_s,y^{\mathrm{i}}_{sl}}\right)e_j, & s=t \text{ and } k\neq l, \\ 0, & s\neq t \text{ or } k=l, \end{cases} \qquad (141)
\end{aligned}
$$

for every $s,t \in [1:N]$, $k \in [1:N_{x_s,\epsilon}]$, $l \in [1:N_{x_t,\epsilon}]$. Above, $e_i \in \mathbb{R}^n$ is the ith coordinate vector. Note that, by construction, each $B_{ij}[sk,tl]$ is a *symmetric random variable* (in the sense that its distribution is symmetric about the origin). Moreover, similar to (137), it holds that

$$
\begin{aligned}
B_{ij}[sk,sk] &= 0, \\
\mathbb{E}_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[B_{ij}[sk,tl]\right] &= 0, \qquad s,t\in[1:N],\ k\in[1:N_{x_s,\epsilon}],\ l\in[1:N_{x_t,\epsilon}]. \qquad (142)
\end{aligned}
$$

Our next step is to decouple the sum in the last line of (140).

Decoupling Let $\Xi = \{\xi_{sk}\}_{s,k}$ (with $s \in [1:N]$ and $k \in [1:N_{x_s,\epsilon}]$) be a sequence of independent standard Bernoulli random variables: each ξ_{sk} independently takes one and zero with equal probabilities. We will shortly use the following simple observation:

$$\mathbb{E}_{\Xi}\left[\xi_{sk}\left(1-\xi_{tl}\right)\right] = \frac{1}{4}, \qquad sk \neq tl. \tag{143}$$

We now revisit (140) and write that

$$\begin{aligned}
&\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right] \\
&\leq \mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} B_{ij}[sk,tl]\right] \qquad \text{(see (140))} \\
&= 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} \mathbb{E}_{\Xi}\left[\xi_{sk}\left(1-\xi_{tl}\right)\right]\cdot B_{ij}[sk,tl]\right] \qquad (B_{ij}[sk,sk]=0, \text{ and (143)}) \\
&\leq 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},\Xi|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} \xi_{sk}\left(1-\xi_{tl}\right)\cdot B_{ij}[sk,tl]\right]. \qquad \text{(Jensen's inequality)}
\end{aligned} \tag{144}$$

In particular, there must exist $\Xi_0 = \{\xi_{0sk}\}_{s,k}$ that exceeds the expectation in the last line above, so that

$$\begin{aligned}
&\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right] \\
&\leq 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}} \xi_{0sk}\left(1-\xi_{0tl}\right)\cdot B_{ij}[sk,tl]\right] \\
&= 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{\xi_{0sk}=1,\,\xi_{0tl}=0} B_{ij}[sk,tl]\right] \\
&= 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{sk\in S_0,\,tl\notin S_0} B_{ij}[sk,tl]\right]. \qquad (S_0 := \{sk \,:\, \xi_{0sk}=1\} \subseteq [1:N_{X,\epsilon}])
\end{aligned} \tag{145}$$

Let $\{Y^{\mathrm{ii}}_{X,\epsilon}, Y^{\mathrm{iii}}_{X,\epsilon}\} \subset \mathbb{R}^{n\times \#N_{X,\epsilon}}$ be an independent copy of $\{Y_{X,\epsilon}, Y^{\mathrm{i}}_{X,\epsilon}\}$. For the sake of brevity, we will use the following short hand:

$$\begin{aligned}
\widetilde{Y}_{X,\epsilon} &:= Y_{X,\epsilon}\cup Y^{\mathrm{i}}_{X,\epsilon}\cup Y^{\mathrm{ii}}_{X,\epsilon}\cup Y^{\mathrm{iii}}_{X,\epsilon}, \\
\widetilde{Y}_{x_s,\epsilon} &:= Y_{x_s,\epsilon}\cup Y^{\mathrm{i}}_{x_s,\epsilon}\cup Y^{\mathrm{ii}}_{x_s,\epsilon}\cup Y^{\mathrm{iii}}_{x_s,\epsilon}, \qquad \forall x_s \in X.
\end{aligned} \tag{146}$$

Equipped with the construction above, we revisit (145) and write that

$$\begin{aligned}
&\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right] \\
&\leq 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{sk\in S_0,\,tl\notin S_0} B_{ij}[sk,tl]\right] \qquad \text{(see (145))} \\
&= 4\cdot\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{sk\in S_0,\,sl\notin S_0} N^{-2}_{x_s,\epsilon}\cdot e_i^*\left(P_{x_s,y_{sk}}-P_{x_s,y^{\mathrm{i}}_{sk}}\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}}-P_{x_s,y^{\mathrm{i}}_{sl}}\right)e_j\right] \\
&\qquad \text{(see (141))} \\
&= 4\cdot\mathbb{E}^p_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\Bigg[\sum_{sk\in S_0,\,sl\notin S_0} \underbrace{N^{-2}_{x_s,\epsilon}\cdot e_i^*\left(P_{x_s,y_{sk}}-P_{x_s,y^{\mathrm{i}}_{sk}}\right)\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y^{\mathrm{ii}}_{sl}}-P_{x_s,y^{\mathrm{iii}}_{sl}}\right)e_j}_{C_{ij}[sk,sl]}\Bigg]
\end{aligned}$$

(independence)

$$= 4 \cdot \mathbb{E}^p_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{sk\in S_0,\, sl\notin S_0} C_{ij}[sk,sl] + \sum_{sk\notin S_0} \mathbb{E}_{Y_{X,S_0^C},Y^{\mathrm{i}}_{X,S_0^C}|Y^{\mathrm{ii}}_X,Y^{\mathrm{iii}}_X,\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[C_{ij}[sk,sl] \right] \right.$$

$$\left. + \sum_{sk\in S_0,\, sl\in S_0} \mathbb{E}_{Y^{\mathrm{ii}}_{X,S_0},Y^{\mathrm{iii}}_{X,S_0}|Y_X,Y^{\mathrm{i}}_X,\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[C_{ij}[sk,sl] \right] \right]$$

$$\leq 4 \cdot \mathbb{E}^p_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{t,s=1}^{N} \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} C_{ij}[sk,tl] \right], \qquad \text{(independence and Jensen's inequality)} \tag{147}$$

where we added two zero expectation terms in the last equality above. Above, we also defined the block-diagonal matrix $B_{ij} \in \mathbb{R}^{N_{X,\epsilon}\times N_{X,\epsilon}}$ such that

$$C_{ij}[sk,tl] = \begin{cases} N^{-2}_{x_s,\epsilon} \cdot e_i^* \left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}} \right) \nabla f(x_s) \nabla f(x_s)^* \left(P_{x_s,y^{\mathrm{ii}}_{sl}} - P_{x_s,y^{\mathrm{iii}}_{sl}} \right) e_j, & s = t \text{ and } k \neq l, \\ 0, & s \neq t \text{ or } k = l, \end{cases} \tag{148}$$

for every $s,t \in [1:N]$, $k \in [1:N_{x_s,\epsilon}]$, $l \in [1:N_{x_t,\epsilon}]$. For every $s \in [N]$, we can also define a family of matrices $\{C_{skl}\}_{k,l\in[N_{x_s,\epsilon}]} \subset \mathbb{R}^{n\times n}$ such that

$$C_{skl}[i,j] = C_{sij}[k,l], \qquad \forall k,l \in [N_{x_s,\epsilon}]. \tag{149}$$

Note that

$$C_{skl} := \frac{1}{N^2_{x_s,\epsilon}} \begin{cases} \left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}} \right) \nabla f(x_s) \nabla f(x_s)^* \left(P_{x_s,y^{\mathrm{ii}}_{sl}} - P_{x_s,y^{\mathrm{iii}}_{sl}} \right), & k \neq l, \\ 0, & k = l. \end{cases} \tag{150}$$

The next step is to modulate the sum in the last line of (147) with a Rademacher sequence.

Modulation with Rademacher Sequences Fix $i,j \in [1:n]$, and recall the definitions of $C_{ij} \in \mathbb{R}^{N_{X,\epsilon}\times N_{X,\epsilon}}$ from (148). Let $H = \{\eta_{sk}\}_{s,k}$ (with $s \in [1:N]$ and $k \in [1:N_{x_s,\epsilon}]$) be a Rademacher sequence, that is $\{\eta_{sk}\}_{s,k}$ are independent Bernoulli random variables taking ± 1 with equal chances. Also let $H^{\mathrm{i}} = \{\eta^{\mathrm{i}}_{sk}\}_{s,k}$ be an independent copy of H. Then, we argue that

$$\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[a_{ij} \right]$$

$$\leq 4 \cdot \mathbb{E}^p_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{s,t=1}^{N} \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} C_{ij}[sk,tl] \right] \qquad \text{(see (147))}$$

$$= 4 \cdot \mathbb{E}^p_{Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon}|Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{s,k} \left(\sum_{t,l} C_{ij}[sk,tl] \right) \right] \right] \qquad \text{(see (146))}$$

$$= 4 \cdot \mathbb{E}^p_{Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},H|Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{s,k} \eta_{sk} \cdot \left(\sum_{t,l} C_{ij}[sk,tl] \right) \right] \right]$$

(independence and symmetry)

$$= 4 \cdot \mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},H|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\mathbb{E}^p_{Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon}|Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{t,l} \left(\sum_{s,k} \eta_{sk} C_{ij}[sk,tl] \right) \right] \right]$$

$$= 4 \cdot \mathbb{E}^p_{Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},H|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\mathbb{E}^p_{Y^{\mathrm{ii}}_{X,\epsilon},Y^{\mathrm{iii}}_{X,\epsilon},H^{\mathrm{i}}|Y_{X,\epsilon},Y^{\mathrm{i}}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X} \left[\sum_{t,l} \eta^{\mathrm{i}}_{tl} \cdot \left(\sum_{s,k} \eta_{sk} C_{ij}[sk,tl] \right) \right] \right]$$

(independence and symmetry)

$$= 4 \cdot \mathbb{E}^p_{\widetilde{Y}_{X,\epsilon},H,H^{\mathrm{i}}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_l,\epsilon}} \eta_{sk}\eta^{\mathrm{i}}_{tl}\cdot C_{ij}[sk,tl]\right]$$

$$=: 4 \cdot \mathbb{E}^p_{\widetilde{Y}_{X,\epsilon},H,H^{\mathrm{i}}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right], \tag{151}$$

where we set

$$c_{ij} := \left|\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_l,\epsilon}} \eta_{sk}\eta^{\mathrm{i}}_{tl}\cdot C_{ij}[sk,tl]\right|. \tag{152}$$

Conditioned on everything but H and H^{i}, c_{ij} is a *decoupled chaos*: decoupled because $H = \{\eta_{sk}\}_{s,k}$ and $H^{\mathrm{i}} = \{\eta^{\mathrm{i}}_{sk}\}_{s,k}$ are independent (Rademacher) sequences. The behavior of the moments of a chaos random variable is well-understood.

Moments of a Decoupled Chaos The fist moment of c_{ij}, namely its expectation, can be estimated as follows. First observe that

$$\begin{aligned}
\mathbb{E}_{H,H^{\mathrm{i}}|\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}[c_{ij}] &\le \sqrt{\mathbb{E}_{H,H^{\mathrm{i}}|\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}[c_{ij}^2]} \qquad \text{(Jensen's inequality)}\\
&= \|C_{ij}\|_F \qquad \big(H \text{ and } H^{\mathrm{i}} \text{ are independent Rademacher sequences}\big)\\
&\le \sqrt{N}\cdot \max_{s\in[1:N]} \|C_{sij}\|_F . \qquad (C_{ij} \text{ is block-diagonal})
\end{aligned} \tag{153}$$

Let us therefore focus on $\|C_{sij}\|_F$ for fixed $s \in [1:N]$:

$$\begin{aligned}
\|C_{sij}\|_F^2 &= \sum_{k,l=1}^{N_{x_s,\epsilon}} |C_{sij}[k,l]|^2\\
&= \sum_{k,l=1}^{N_{x_s,\epsilon}} |C_{skl}[i,j]|^2 \qquad \text{(see (149))}\\
&\le N_{x_s,\epsilon}^2 \cdot \max_{k,l\in[1:N_{x_s,\epsilon}]} |C_{skl}[i,j]|^2\\
&\le N_{x_s,\epsilon}^2 \cdot \max_{k,l\in[1:N_{x_s,\epsilon}]} \|C_{skl}\|_\infty^2 ,
\end{aligned} \tag{154}$$

where $\|A\|_\infty$ is the largest entry of A in magnitude. With $e_i \in \mathbb{R}^n$ denoting the ith canonical vector, we continue by noting that

$$\begin{aligned}
&\|C_{skl}\|_\infty\\
&= \max_{i,j\in[1:n]} |C_{skl}[i,j]|\\
&= \max_{i,j\in[1:n]} |e_i^* C_{skl} e_j|\\
&= N_{x_s,\epsilon}^{-2} \cdot \max_{i,j\in[1:n]} \left| e_i^* \left(P_{x_s,y_{sk}} - P_{x_s,y^{\mathrm{i}}_{sk}}\right) \nabla f(x_s) \nabla f(x_s)^* \left(P_{x_s,y^{\mathrm{i}\,\mathrm{i}}_{sl}} - P_{x_s,y^{\mathrm{i}\,\mathrm{i}\,\mathrm{i}}_{sl}}\right) e_j \right| \qquad \text{(see (148))}\\
&\le 4N_{x_s,\epsilon}^{-2} \cdot \max_{s\in[1:N]} \max_{i\in[1:n]} \max_{y_s\in\widetilde{Y}_{x_s,\epsilon}} |e_i^* P_{x_s,y_s} \nabla f(x_s)|^2 \qquad \text{(see (146))}\\
&\le 4N_{x_s,\epsilon}^{-2} \cdot \max_{s\in[1:N]} \max_{i\in[1:n]} \max_{y_s\in\widetilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} e_i\|_2^2 \cdot \|P_{x_s,y_s} \nabla f(x_s)\|_2^2 \qquad \text{(Cauchy-Schwarz's inequality)}\\
&\le 4N_{x_s,\epsilon}^{-2} \cdot \frac{Q_{X,\epsilon}}{n} \cdot \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}, \qquad \text{(conditioned on the event } \mathcal{E}_3)
\end{aligned} \tag{155}$$

where we defined the event $\mathcal{E}_3$ as

$$\mathcal{E}_3 = \left\{ \max_{s\in[1:N]} \max_{i\in[1:n]} \max_{y_s\in\widetilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} e_i\|_2^2 \le \frac{Q_{X,\epsilon}}{n} \right\} \bigcap \left\{ \max_{s\in[1:N]} \max_{y_s\in\widetilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} \nabla f(x_s)\|_2^2 \le \frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n} \right\}, \tag{156}$$

for $Q_{X,\epsilon} > 0$ to be set later. For $p \geq 1$ to be assigned later, we also assume that $\mathcal{E}_3$ is very likely to happen:

$$\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] \lesssim \left(\frac{p}{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}\right)^{\frac{p}{2}}. \qquad \text{(see (21))} \tag{157}$$

We now complete our calculation of the first moment of c_{ij}:

$$\begin{aligned}
\mathbb{E}_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}[c_{ij}] &\leq \|C_{ij}\|_F \\
&\leq \sqrt{N}\cdot \max_{s\in[1:N]} \|C_{sij}\|_F \qquad \text{(see (153))} \\
&\leq \sqrt{N}\cdot \max_{s\in[1:N]} \max_{k,l\in[1:N_{x_s,\epsilon}]} N_{x_s,\epsilon}\cdot \|C_{skl}\|_\infty \qquad \text{(see (154))} \\
&\leq \sqrt{N}\cdot \max_{s\in[1:N]} N_{x_s,\epsilon}\cdot \frac{4Q_{X,\epsilon}^2\,\mathrm{L}_f^2}{n^2 N_{x_s,\epsilon}^2} \qquad \text{(see (155))} \\
&\leq \frac{4\sqrt{N}Q_{X,\epsilon}^2\,\mathrm{L}_f^2}{n^2\cdot \min_{x\in X} N_{x,\epsilon}}.
\end{aligned} \tag{158}$$

To control the higher order moments of c_{ij}, we invoke the following result [62, Corollary 2].

Proposition 4. (Moments of a decoupled chaos) *For a square matrix C, a Rademacher sequence $H = \{\eta_k\}_k$, and an independent copy $H^{\mathrm{i}} = \{\eta_l^{\mathrm{i}}\}_l$, consider the decoupled (second-order) chaos*

$$c = \left|\sum_{k,l} \eta_k \eta_l^{\mathrm{i}} \cdot C[k,l]\right|.$$

Then, it holds that

$$\mathbb{E}^p[c - \mathbb{E}[c]] \lesssim p\cdot b + \sqrt{p}\cdot\sigma, \qquad \forall p \geq 1, \tag{159}$$

where

$$b := \|C\|, \tag{160}$$

$$\sigma := \sqrt{\mathbb{E}_H\left[\|C\eta\|_2^2\right]} = \|C\|_F, \tag{161}$$

and η is the vector formed from the Rademacher sequence H.

We now appeal to Proposition 4 in order to bound the moments of the chaos random variable c_{ij} in (152) (conditioned on $X, \widetilde{Y}_{X,\epsilon}$ and the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$). To that end, note that

$$\begin{aligned}
b &= \|C_{ij}\| \qquad \text{(see (160))} \\
&= \max_{s\in[1:N]} \|C_{sij}\|. \qquad (C_{ij} \text{ is block-diagonal})
\end{aligned} \tag{162}$$

Let us then focus on $\|C_{sij}\|$ for fixed $s \in [1:N]$. Observe that

$$\begin{aligned}
\|C_{sij}\| &\leq N_{x_s,\epsilon}\cdot \|C_{sij}\|_\infty \qquad \left(\|A\| \leq a\cdot\|A\|_\infty,\ \forall A\in\mathbb{R}^{a\times a}\right) \\
&\leq N_{x_s,\epsilon}\cdot \max_{k,l\in[1:N_{x_s,\epsilon}]} \|C_{skl}\|_\infty \qquad \text{(see (149))} \\
&\leq N_{x_s,\epsilon}\cdot \frac{4Q_{X,\epsilon}^2\,\mathrm{L}_f^2}{n^2 N_{x_s,\epsilon}^2} \qquad \text{(see (155))} \\
&\leq \frac{4Q_{X,\epsilon}^2\,\mathrm{L}_f^2}{n^2\cdot\min_{x\in X} N_{x,\epsilon}}. \qquad \text{(see (37))}
\end{aligned} \tag{163}$$

In light of (162), it follows that

$$b \leq \max_{s\in[1:N]} \|C_{sij}\| \qquad \text{(see (162))}$$

$$\leq \frac{4Q_{X,\epsilon}^2 \mathrm{L}_f^2}{n^2 \cdot \min_{x\in X} N_{x,\epsilon}}. \qquad \text{(see (163))} \tag{164}$$

We argue likewise to find σ:

$$\begin{aligned}\sigma &= \|C_{ij}\|_F \qquad \text{(see (161))}\\ &\leq \frac{4\sqrt{N}Q_{X,\epsilon}^2 \mathrm{L}_f^2}{n^2 \cdot \min_{x\in X} N_{x,\epsilon}}. \qquad \text{(see (158))}\end{aligned} \tag{165}$$

With b and σ at hand, we now invoke Proposition 4 to write that

$$\begin{aligned}&\mathbb{E}^p_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right]\\
&= \mathbb{E}^p_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}}\eta_{sk}\eta^{\mathrm{i}}_{tl}\cdot C_{ij}[sk,tl]\right] \qquad \text{(see (152))}\\
&\leq \mathbb{E}^p_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}-\mathbb{E}_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right]\right]+\mathbb{E}_{H,H^{\mathrm{i}}|\mathcal{E}_3,\widetilde{Y}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right] \qquad \text{(triangle inequality)}\\
&\lesssim (p\cdot b+\sqrt{p}\cdot\sigma)+\frac{\sqrt{N}Q_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\cdot\min_x N_{x,\epsilon}} \qquad \text{(see Proposition 4 and (158))}\\
&\lesssim \left(p\cdot\frac{Q_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\cdot\min_x N_{x,\epsilon}}+\sqrt{p}\cdot\frac{\sqrt{N}Q_{X,f,\epsilon}^2\mathrm{L}_f^2}{n^2\cdot\min_x N_{x,\epsilon}}\right)+\frac{\sqrt{N}Q_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\cdot\min_x N_{x,\epsilon}} \qquad \text{(see (164) and (165))}\\
&\lesssim \sqrt{p}\cdot\frac{\sqrt{N}Q_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\cdot\min_x N_{x,\epsilon}} \qquad \text{(if } 1\leq p\leq N\text{)}\\
&\lesssim \sqrt{p}\cdot\frac{NQ_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\cdot\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}. \qquad \text{(see (37) and (21))}\end{aligned} \tag{166}$$

Conditioned on $\overline{N}_{X,\epsilon}$, the bound above is independent of $\widetilde{Y}_{X,\epsilon}$, which allows us to remove the conditioning and find that

$$\mathbb{E}^p_{\widetilde{Y}_{X,\epsilon},H,H^{\mathrm{i}}|\mathcal{E}_3,\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right]\lesssim\sqrt{p}\cdot\frac{NQ_{X,\epsilon}^2\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\cdot\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}. \tag{167}$$

As a useful aside, we also record a uniform bound on c_{ij} for every $i,j\in[1:n]$:

$$\begin{aligned}|c_{ij}| &= \left|\sum_{s,t=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}}\eta_{sk}\eta^{\mathrm{i}}_{tl}\cdot C_{ij}[sk,tl]\right| \qquad \text{(see (152))}\\
&\leq \sum_{s,t=1}^{N}\left|\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_t,\epsilon}}\eta_{sk}\eta^{\mathrm{i}}_{tl}\cdot C_{ij}[sk,tl]\right| \qquad \text{(triangle inequality)}\\
&= \sum_{s=1}^{N}\left|\sum_{k=1}^{N_{x_s,\epsilon}}\sum_{l=1}^{N_{x_s,\epsilon}}\eta_{sk}\eta^{\mathrm{i}}_{sl}\cdot C_{sij}[k,l]\right| \qquad \big(C_{ij}\text{ is block-diagonal with blocks } C_{sij}\in\mathbb{R}^{N_{x_s,\epsilon}\times N_{x_s,\epsilon}}\text{, see (149)}\big)\\
&\leq \sum_{s=1}^{N}N_{x_s,\epsilon}\cdot\|C_{sij}\| \qquad \big(H,H^{\mathrm{i}}\text{ are Rademacher sequences}\big)\\
&\leq \sum_{s=1}^{N}N_{x_s,\epsilon}^2\cdot\frac{4Q_{X,\epsilon}^2\mathrm{L}_f^2}{n^2N_{x_s,\epsilon}^2} \qquad \text{(see (163))}\\
&= \frac{4NQ_{X,\epsilon}^2\mathrm{L}_f^2}{n^2}.\end{aligned} \tag{168}$$

Putting everything back together, we finally argue that

$$\begin{aligned}
\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right] &\leq 4\cdot\mathbb{E}^p_{\widetilde{Y}_{X,\epsilon},H,H^{\mathrm{i}}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right] \qquad \text{(see (151))}\\
&\leq 4\cdot\mathbb{E}^p_{\widetilde{Y}_{X,\epsilon},H,H^{\mathrm{i}}|\mathcal{E}_3,\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[c_{ij}\right] + 4\cdot\sup|c_{ij}|\cdot\left(\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right]\right)^{\frac{1}{p}}\\
&\qquad \text{(see (53))}\\
&\lesssim \sqrt{p}\cdot\frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + \frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2}\cdot\sqrt{\frac{p}{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\\
&\qquad \text{(see (167), (168), and (157))}\\
&\lesssim \sqrt{p}\cdot\frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}, \qquad (169)
\end{aligned}$$

when $1\leq p\leq N$ (see (166)). At last, (169) describes the moments of the random variable a_{ij} for fixed i,j (and conditioned on $\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X$).

K.1.3 Applying Markov's Inequality

Given the estimates of the moments of a_{ij} in (169), we can simply apply Markov's inequality to translate this information into a tail bound for a_{ij}. Indeed, for arbitrary $1\leq p\leq N$ and $\gamma_8>0$, it holds that

$$\begin{aligned}
\Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[|a_{ij}|>\gamma_8\right] &= \Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[|a_{ij}|^p>\gamma_8^p\right]\\
&\leq \left(\frac{\mathbb{E}^p_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[a_{ij}\right]}{\gamma_8}\right)^p \qquad \text{(Markov's inequality)}\\
&\leq \left(\frac{\mathrm{C}_1\sqrt{p}NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{\gamma_8 n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\right)^p, \qquad \text{(see (169))} \qquad (170)
\end{aligned}$$

for an absolute constant C_1. In particular, the choice of

$$\gamma_8 = \mathrm{C}_1\gamma_7\cdot\sqrt{\log n}\cdot\frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}, \qquad p=\max\left[\log n,1\right]\leq N, \qquad \gamma_7\geq 1,$$

yields

$$\begin{aligned}
&\Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[|a_{ij}|\gtrsim\gamma_7\cdot\sqrt{\log n}\cdot\frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{x,\epsilon}}}\right]\\
&\leq \gamma_7^{-\log n} = n^{-\log\gamma_7}. \qquad (171)
\end{aligned}$$

With the tail bound of a_{ij} finally available above (for fixed $i,j\in[1:n]$ and conditioned on $\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X$), we next quantify how $\ddot{\Sigma}^{\circ}_{X,Y_{X,\epsilon}}$ concentrates about its expectation.

K.1.4 Applying the Union Bound

In light of (171) and by applying the union bound to $\{a_{ij}\}_{i,j}$, we arrive at the following statement.

$$\begin{aligned}
&\Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\max_{i,j\in[1:n]}|a_{ij}|\lesssim\gamma_7\cdot\sqrt{\log n}\cdot\frac{NQ^2_{X,\epsilon}\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{x,\epsilon}}}\right]\\
&\geq 1-n^2\cdot n^{-\log\gamma_7}. \qquad \text{(union bound and (171))} \qquad (172)
\end{aligned}$$

K.2 Second and third norms

In this section, we bound the second and third norms in the last line of (131) using the Bernstein inequality. Let us bound the second norm as

$$\begin{aligned}
&\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\sum_{l\neq k}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\leq\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\frac{1}{N_{x_s,\epsilon}}\left(P_{x_s,y_{sk}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&=:\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}A_{x_s,y_{sk}}\right\|_F, \qquad (173)
\end{aligned}$$

where $\{A_{x_s,y_{sk}}\}_{sk}\subset\mathbb{R}^{n\times n}$ is a sequence of zero-mean and independent random matrices. To apply the Bernstein inequality (Proposition 2) conditioned on the event $\mathcal{E}_1$, we write that

$$\begin{aligned}
\|A_{x_s,y_{sk}}\|_F&=\frac{1}{N_{x_s,\epsilon}}\left\|\left(P_{x_s,y_{sk}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\leq\frac{1}{N_{x_s,\epsilon}}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\quad+\frac{1}{N_{x_s,\epsilon}}\left\|\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\leq\frac{1}{N_{x_s,\epsilon}}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\right\|_2\left\|\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\right\|_2+\frac{1}{N_{x_s,\epsilon}}\left\|\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\right\|_2^2\\
&\leq\frac{1}{N_{x_s,\epsilon}}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\right\|_2\cdot\sqrt{\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left\|P_{x_s,y_s}\nabla f(x_s)\right\|_2^2}+\frac{1}{N_{x_s,\epsilon}}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left\|P_{x_s,y_s}\nabla f(x_s)\right\|_2^2\\
&\qquad\text{(Jensen's inequality)}\\
&\leq\frac{1}{N_{x_s,\epsilon}}\sqrt{\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}}\sqrt{\frac{\mathrm{L}_f^2}{\mathrm{K}_\mu\,n}}+\frac{\mathrm{L}_f^2}{\mathrm{K}_\mu\,N_{x_s,\epsilon}n}\qquad\text{(see (121), (8), and [11, Lemma 5.5])}\\
&\leq\frac{2Q_{X,\epsilon}\,\mathrm{L}_f^2}{\min_{s\in[N]}N_{x_s,\epsilon}n}=:b.\qquad\left(\text{if }Q_{X,\epsilon}\geq\mathrm{K}_\mu^{-1}\right) \qquad (174)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left\|A_{x_s,y_{sk}}\right\|_F^2\\
&\leq\sum_{s=1}^{N}\frac{1}{N_{x_s,\epsilon}^2}\sum_{k=1}^{N_{x_s,\epsilon}}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left\|P_{x_s,y_s}\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F^2\\
&\leq\sum_{s=1}^{N}\frac{1}{N_{x_s,\epsilon}^2}\sum_{k=1}^{N_{x_s,\epsilon}}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\right\|_2^2\left\|\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\right\|_2^2\\
&\leq\sum_{s=1}^{N}\frac{1}{N_{x_s,\epsilon}^2}\sum_{k=1}^{N_{x_s,\epsilon}}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\right\|_2^4\qquad\text{(Jensen's inequality)}\\
&\leq\sum_{s=1}^{N}\frac{\mathrm{L}_f^4}{\mathrm{K}_\mu^2\,N_{x_s,\epsilon}n^2}\qquad\text{(see (121), (8), and [11, Lemma 5.5])}\\
&\leq\frac{NQ_{X,\epsilon}^2\,\mathrm{L}_f^4}{\min_{s\in[N]}N_{x_s,\epsilon}n^2}=:\sigma^2.\qquad\left(\text{if }Q_{X,\epsilon}\geq\mathrm{K}_\mu^{-1}\right) \qquad (175)
\end{aligned}$$

The second line above uses the fact that $\mathbb{E}\left\|Z-\mathbb{E}[Z]\right\|_F^2 \leq \mathbb{E}\left\|Z\right\|_F^2$ for a random matrix Z. It follows that

$$\max[b,\sigma] \leq 2\sqrt{\frac{N}{\min_{s\in[N]} N_{x_s,\epsilon}}}\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}. \tag{176}$$

An application of the Bernstein inequality now yields that conditioned on $\mathcal{E}_1, \mathcal{E}_2, \overline{N}_{X,\epsilon}, X$,

$$\begin{aligned}
&\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}\frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\sum_{l\neq k}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\leq \frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k=1}^{N_{x_s,\epsilon}}A_{x_s,y_{sk}}\right\|_F\\
&\lesssim \frac{n^2}{N}\cdot\gamma\max[b,\sigma]\\
&\lesssim \frac{\gamma n^2}{N}\sqrt{\frac{N}{\min_{s\in[N]} N_{x_s,\epsilon}}}\frac{Q_{X,\epsilon}\,\mathrm{L}_f^2}{n}\\
&= \frac{\gamma n Q_{X,\epsilon}\,\mathrm{L}_f^2}{\sqrt{N\min_{s\in[N]} N_{x_s,\epsilon}}}\\
&= \frac{\gamma n Q_{X,\epsilon}\,\mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}
\end{aligned} \tag{177}$$

for $\gamma \geq 1$ and except with a probability of at most $e^{-\gamma}$. An identical bound holds for the third norm in the last line of (131).

K.3 Bound on (131)

We now combine the bounds for the terms in (131) obtained in Sections K.1 and K.2. Applying (138), we have that conditioned on $\mathcal{E}_1, \mathcal{E}_2, \overline{N}_{X,\epsilon}, X$,

$$\begin{aligned}
&\left\|\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}-\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F\\
&\leq \frac{n^2}{N}\sqrt{\sum_{i,j=1}^{n}a_{ij}^2}+\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x_s,\epsilon}^2}\left(P_{x_s,y_{sk}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\nabla f(x_s)\nabla f(x_s)^*\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right\|_F\\
&\quad+\frac{n^2}{N}\left\|\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x_s,\epsilon}^2}\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\nabla f(x_s)\nabla f(x_s)^*\left(P_{x_s,y_{sl}}-\mathbb{E}_{y_s|\mathcal{E}_1,x_s}[P_{x_s,y_s}]\right)\right\|_F\\
&\leq \frac{n^3}{N}\cdot\max_{i,j\in[1:n]}|a_{ij}|+2\frac{\gamma n Q_{X,\epsilon}\,\mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\quad\text{(see (177))}\\
&\lesssim \frac{n^3}{N}\cdot\gamma_7\cdot\sqrt{\log n}\cdot\frac{NQ_{X,\epsilon}^2\,\mathrm{L}_f^2}{n^2\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{x,\epsilon}}}+2\frac{\gamma n Q_{X,\epsilon}\,\mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\quad\text{(see (172))}\\
&= \gamma_7\cdot\sqrt{\log n}\cdot\frac{nQ_{X,\epsilon}^2\,\mathrm{L}_f^2}{\sqrt{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{x,\epsilon}}}+2\frac{\gamma n Q_{X,\epsilon}\,\mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\\
&\lesssim \max[\gamma_7,\gamma]\cdot\sqrt{\log n}\cdot\frac{n\cdot\max[Q_{X,\epsilon},Q_{X,\epsilon}^2]\cdot\mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}
\end{aligned} \tag{178}$$

for $\gamma, \gamma_7 \geq 1$ and except with a probability of at most $e^{-\gamma}+n^2\cdot n^{-\log\gamma_7}$. This holds under (157) (with $p=\max[\log n,1]\leq N$).

Finally, we proceed to remove the conditioning on $\mathcal{E}_1$. Similar to (127), we have

$$\begin{aligned}
&\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right] \\
&= \Pr_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right]\left(\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1^C,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right). \qquad (179)
\end{aligned}$$

Since for any X, $Y_{X,\epsilon}$, we have

$$\begin{aligned}
\left\|\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right\|_F &\le \frac{n^2}{N}\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x_s,\epsilon}^2}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\nabla f(x_s)^* P_{x_s,y_{sl}}\right\|_F \qquad \text{(see (111))} \\
&\le \frac{n^2}{N}\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x_s,\epsilon}^2}\left\|P_{x_s,y_{sk}}\nabla f(x_s)\right\|_2\left\|\nabla f(x_s)^* P_{x_s,y_{sl}}\right\|_2 \\
&\le \frac{n^2}{N}\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x,\epsilon}^2}\left\|\nabla f(x_s)\right\|_2^2 \\
&\le \frac{n^2}{N}\sum_{s=1}^{N}\sum_{k\neq l}\frac{1}{N_{x,\epsilon}^2}L_f^2 \qquad \text{(see (8))} \\
&\le \frac{n^2}{N}\sum_{s=1}^{N}L_f^2 \\
&= n^2 L_f^2, \qquad (180)
\end{aligned}$$

we conclude that

$$\begin{aligned}
&\left\|\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F \\
&\le \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\left\|\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1^C,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F \qquad \text{(see (179))} \\
&\le 2n^2 L_f^2 \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]. \qquad \text{(triangle inequality and (180))} \qquad (181)
\end{aligned}$$

Lastly, we remove the conditioning on the event $\mathcal{E}_1$ as follows:

$$\begin{aligned}
&\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F\right. \\
&\qquad \left.\gtrsim \max[\gamma_7,\gamma]\cdot\sqrt{\log n}\cdot\frac{n\cdot\max[Q_{X,\epsilon},Q_{X,\epsilon}^2]\cdot \mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 2n^2L_f^2\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\right] \\
\le &\Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F\right. \\
&\qquad \left.\gtrsim \max[\gamma_7,\gamma]\cdot\sqrt{\log n}\cdot\frac{n\cdot\max[Q_{X,\epsilon},Q_{X,\epsilon}^2]\cdot \mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}} + 2n^2L_f^2\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]\right] \\
&\quad + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right] \qquad \text{(see (53))} \\
\le &\Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\left\|\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\dddot{\Sigma}^{o}_{X,Y_{X,\epsilon}}\right]\right\|_F \gtrsim \max[\gamma_7,\gamma]\cdot\sqrt{\log n}\cdot\frac{n\cdot\max[Q_{X,\epsilon},Q_{X,\epsilon}^2]\cdot \mathrm{L}_f^2}{\sqrt{\rho_{\mu,X,\epsilon}N_{X,\epsilon}}}\right] \\
&\quad + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right] \qquad \text{(see (181))} \\
\le &\, e^{-\gamma} + n^2\cdot n^{-\log\gamma_7} + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,\overline{N}_{X,\epsilon},X}\left[\mathcal{E}_1^C\right]. \qquad \text{(see (178))} \qquad (182)
\end{aligned}$$

This holds for $\gamma, \gamma_7 \geq 1$ and under (157) (with $p = \max[\log n, 1] \leq N$). Setting $\gamma = \gamma_7$ completes the proof of Lemma 7.

L Proof of Lemma 8

First, to prove (118), suppose X and the neighborhood structure $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$ are fixed. Then, for every $x \in X$, the columns of the matrix $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$ are random vectors drawn from the conditional probability measure $\mu_{x,\epsilon}$ (see (17)). For fixed $x \in X$ and with $y \sim \mu_{x,\epsilon}$, recall from Assumption 1 that

$$\Pr{}_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] \lesssim e^{-\,\mathrm{K}_\mu\,\gamma_1}, \tag{183}$$

for arbitrary (but fixed) $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$ and $\gamma_1 \geq 0$. The inequality (118) readily follows with an application of the union bound: For all possible choices of x, y, it holds that

$$\|P_{x,y}\nabla f(x)\|_2^2 \leq \frac{\gamma_1 \,\|\nabla f(x)\|_2^2}{n} \leq \frac{\gamma_1 \,\mathrm{L}_f^2}{n}, \qquad \text{(see (8))} \tag{184}$$

except with a probability $\lesssim N_{X,\epsilon} e^{-\,\mathrm{K}_\mu\,\gamma_1}$. With the choice of $\gamma_1 = Q_{X,\epsilon} = \gamma_2\,\mathrm{K}_\mu^{-1}\log^2(N_{X,\epsilon})$ for $\gamma_2 \geq 3$, we establish (118).

Our next goal is to prove that (116) is satisfied. Note that the probability in (116) is conditioned on $\mathcal{E}_1$. We can remove this conditioning using the law of total probability:

$$\begin{aligned}
\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] &= \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1\right] + \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1^C,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right] \\
&\geq \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1\right].
\end{aligned}$$

Rearranging terms, we have that

$$\begin{aligned}
\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_1,\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] &\leq \frac{\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right]}{\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1\right]} \\
&= \frac{\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right]}{1 - \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right]} \\
&\lesssim \Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right],
\end{aligned} \tag{185}$$

where the last line follows under the assumption that $N_{X,\epsilon}$ large enough that, under (118), $\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_1^C\right]$ is bounded above by a constant smaller than 1. To bound the right hand side in (185), suppose X and the neighborhood structure $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$ are fixed. Then, for every $x \in X$, the columns of the matrix $\widetilde{Y}_{x,\epsilon} \in \mathbb{R}^{n \times (4N_{x,\epsilon})}$ are random vectors drawn from the conditional probability measure $\mu_{x,\epsilon}$ (see (17)). For fixed $x \in X$ and with $y \sim \mu_{x,\epsilon}$, recall from Assumption 1 that (183) holds for arbitrary (but fixed) $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$ and $\gamma_1 \geq 0$. For all possible choices of x, y, i, it follows that

$$\|P_{x,y}e_i\|_2^2 \leq \frac{\gamma_1}{n}, \qquad \|P_{x,y}\nabla f(x)\|_2^2 \leq \frac{\gamma_1\,\|\nabla f(x)\|_2^2}{n} \leq \frac{\gamma_1\,\mathrm{L}_f^2}{n}, \qquad \text{(see (8))} \tag{186}$$

except with a probability $\lesssim n N_{X,\epsilon} e^{-\,\mathrm{K}_\mu\,\gamma_1}$. With the choice of $\gamma_1 = Q_{X,\epsilon} = \gamma_2\,\mathrm{K}_\mu^{-1}\log^2(N_{X,\epsilon})$ for $\gamma_2 \geq 3$, we find that

$$\begin{aligned}
\Pr_{\widetilde{Y}_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}\left[\mathcal{E}_3^C\right] &\lesssim n N_{X,\epsilon}^{(1-\gamma_2\log(N_{X,\epsilon}))} \\
&\lesssim n^{(2-\gamma_2\log(N_{X,\epsilon}))} \quad (N_{X,\epsilon} \geq n)
\end{aligned}$$

$$
\begin{aligned}
&\lesssim n^{(2-\gamma_2)\log(N_{X,\epsilon})} \quad (\log(N_{X,\epsilon}) \geq 1) \\
&\lesssim n^{-\log(N_{X,\epsilon})} \quad (\gamma_2 \geq 3) \\
&= N_{X,\epsilon}^{-\log(n)} \\
&= \left(\frac{1}{N_{X,\epsilon}^2}\right)^{\frac{1}{2}\log(n)} \\
&\leq \left(\frac{\log n}{N_{X,\min,\epsilon}\rho_{\mu,X,\epsilon}N_{X,\epsilon}}\right)^{\frac{\log n}{2}}, \qquad (187)
\end{aligned}
$$

where the last line follows since $N_{X,\epsilon} \geq N_{X,\min,\epsilon}$, $\log(n) \geq 1$, and $\rho_{\mu,X,\epsilon} \leq 1$. Combining (185) and (187) proves that (116) is satisfied and thus completes the proof of Lemma 8.