RANDOMIZED LEARNING OF THE
SECOND-MOMENT MATRIX OF A SMOOTH FUNCTION

ARMIN EFTEKHARI∗
Institute of Electrical Engineering
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland

MICHAEL B. WAKIN
Department of Electrical Engineering
Colorado School of Mines
Denver, CO 80401, USA

PING LI
Departments of Statistics and Computer Science
Rutgers University
Piscataway, NJ 08854, USA

PAUL G. CONSTANTINE
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309, USA

Abstract. Consider an open set \( D \subseteq \mathbb{R}^n \), equipped with a probability measure \( \mu \). An important characteristic of a smooth function \( f : D \to \mathbb{R} \) 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 \( x \in D \) and \( * \) stands for transpose. For instance, the span of the leading \( r \) eigenvectors of \( \Sigma_\mu \) 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 \( \Sigma_\mu \) from random point evaluations of \( f(\cdot) \) without imposing any structural assumptions on \( \Sigma_\mu \). 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 \( D \subseteq \mathbb{R}^n \), equipped with

2010 Mathematics Subject Classification. Primary: 68W25; Secondary: 68W20.
Key words and phrases. Active subspace, second-moment matrix, covariance estimation, ridge approximation, approximation theory.

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.

∗ Corresponding author: Armin Eftekhari.
probability measure $\mu$. 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 [55, 9, 19, 31, 61, 48].

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

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

where $\nabla f(x) \in \mathbb{R}^n$ is the gradient of $f(\cdot)$ at $x \in \mathbb{D}$ and the superscript $*$ denotes vector and matrix transpose. The $[i, j]$th entry of this matrix, namely $\Sigma_{\mu}[i, j]$, measures the expected product between the $i$th and $j$th partial derivatives of $f(\cdot)$. Note that $\Sigma_\mu$ 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 $\mu$. 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 $\Sigma_\mu$ span an $r$-dimensional active subspace of $f(\cdot)$ with respect to the measure $\mu$ [11], 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, 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 $\mu$.

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 $\mu$. In particular, consider $N$ random points drawn independently from $\mu$ and stored as the columns of a matrix $X \in \mathbb{R}^{n \times N}$. It is then easy to verify [10] that

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

As suggested above, we will often suppress the dependence on $f(\cdot)$ in our notation for the sake of brevity.
is an unbiased estimator of $\Sigma_\mu$ in (1).\footnote{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 $X$.} In fact, a standard large deviation analysis reveals that $\|\hat{\Sigma}_X - \Sigma_\mu\| \propto \frac{1}{\sqrt{N}}$, 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 $\hat{\Sigma}_X$ 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 $B_{x,\epsilon}$ denote the Euclidean ball of radius $\epsilon$ about $x$, and set $B_{X,\epsilon} = \bigcup_{x \in X} B_{x,\epsilon}$. Let also $\mu_{X,\epsilon}$ be the conditional probability measure on $B_{X,\epsilon}$ induced by $\mu$. 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 $Y_{x,\epsilon} = Y_{X,\epsilon} \cap B_{x,\epsilon}$, so that $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$ contains all $\epsilon$-neighbors of $x$ in $Y_{X,\epsilon}$. This setup is illustrated in Figure 1.

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

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

(3)

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

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

(4)

In general, however, $\hat{\Sigma}_{X,Y_{X,\epsilon}}$ is a biased estimator of $\hat{\Sigma}_X$ and the importance of consistency in statistical learning motivated us to search for a better estimator, not unlike the approach taken in covariance matrix estimation \cite{42, 21}, see also Section 1.2. Algorithm 1, in fact, introduces an estimate of $\hat{\Sigma}_X$, denoted throughout by $\hat{\Sigma}_{X,Y_{X,\epsilon}}$, which has a smaller bias than $\hat{\Sigma}_{X,Y_{X,\epsilon}}$. Indeed, Theorem 3.1 in Section 3, roughly speaking,\footnote{In order to simplify this overview, we suppress the less important terms here before turning to the details later.} establishes that

$$\left\| \mathbb{E} \left[ \hat{\Sigma}_{X,Y_{X,\epsilon}} \right] - \Sigma_\mu \right\|_F \leq B_{\mu,\epsilon} + \epsilon n^{3/2},$$

(5)

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

suppressed as usual. Moreover, loosely speaking, it holds true that

$$
\left\| \hat{\Sigma}_{X,Y_{X, \epsilon}} - \Sigma_{\mu} \right\|_F \lesssim B_{\mu, \epsilon} + \epsilon n^2 + \frac{n}{\sqrt{N_{X, \epsilon}}}, \quad (6)
$$

with high probability, as described in Theorem 3.2 in Section 3. As a rule of thumb, (5) and (6) hold when \( N_{X, \epsilon} \approx N \log^2 N \). Thus, it suffices to take only \( O(\log^2 N) \) samples within the \( \epsilon \)-neighborhood of each of the \( N \) data points in \( X \). As we discuss in Remark 5, the resulting convergence rate (as a function of the total number of samples) is then nearly the same as one achieves when perfect knowledge of the gradients is available and two-stage sampling is not required.

To verify the convergence rate from (6), we consider the following numerical example. For \( x \in [-1,1]^{500} \) (i.e., \( n = 500 \)) and \( \mu \) a uniform probability distribution on the hypercube, let \( f(x) = \frac{1}{2} x^* A x + b^* x \) for a known symmetric matrix \( A \in \mathbb{R}^{500 \times 500} \) and a known vector \( b \in \mathbb{R}^{500} \). A quick calculation shows that \( \Sigma_{\mu} = \frac{1}{2} A^2 + b b^* \). For the simulation, we generate \( A \) and \( b \) randomly, and we estimate the relative error (in the Frobenius norm) in the approximation \( \hat{\Sigma}_{X,Y_{X, \epsilon}} \) from Algorithm 1. All reported results use \( \epsilon = 10^{-4} \); using \( 10^{-2} \) and \( 10^{-6} \) produced similar results. Each subfigure in Figure 2 shows results using a different value for \( N_{X, \min, \epsilon} \) from the set \{50, 200, 400, 550\}, i.e., the minimum number of samples in each ball. Note that the first three values of \( N_{X, \min, \epsilon} \) tested are significantly less than the dimension \( n \) of the space. In other words, each gradient approximation uses significantly fewer than
the $n + 1$ samples that would be needed for a finite difference approximation. The experiments in each subfigure use the number of centers $N \in \{10, 50, 100, 500, 1000\}$, and for each value of $N$ there are 10 independent replications. The slope of the line is $-1/2$ in each case, which verifies the expected convergence rate in (6), in this case to a bias of approximately zero (see Remark 1).

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

**Input:**
- Open set $\mathbb{D} \subseteq \mathbb{R}^n$, equipped with probability measure $\mu$.
- 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:**
- $\hat{\Sigma}_{X,Y_{X,\epsilon}}$, as an estimate of $\Sigma_\mu$.

**Body:**
- Draw $N$ random points independently from $\mu$ 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 $\mu$ on $B_{X,\epsilon} = \bigcup_{x \in X} B_{x,\epsilon}$. In turn, $B_{x,\epsilon} \subset \mathbb{R}^n$ is the Euclidean ball of radius $\epsilon$ about $x$. Partition $Y_{X,\epsilon}$ according to $X$ by setting $Y_{x,\epsilon} = Y_{X,\epsilon} \cap B_{x,\epsilon}$, so that $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$ contains all $\epsilon$-neighbors of $x$ in $Y_{X,\epsilon}$.
- Compute and return

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

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

$$\hat{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2} \in \mathbb{R}^n.$$
Figure 2. A simulation study using a 500-dimensional \((n = 500)\) quadratic function for the relative error in the approximation \((7)\) as a function of the number \(N_{X,\epsilon}\) of total samples. All simulations use \(\epsilon = 10^{-4}\). Each subfigure uses a different value for the minimum number \(N_{X,\min,\epsilon}\) of points in the \(\epsilon\)-neighborhood of each center point in \(X\), and varies the number \(N\) of centers. Each of the five groups of blue dots in each subfigure indicates a different number \(N\). Within each group, there are ten replications. The slope of each line is \(-1/2\), which verifies the expected convergence rate from \((6)\).

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 \([42]\). In this context, more efficient algorithms are available for estimating, for example, the covariance matrix with a sparse inverse \([21]\). 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 \( \sigma \)-algebra and probability measure \( \mu \). We assume throughout that \( f : \mathbb{D} \rightarrow \mathbb{R} \) is twice differentiable on \( \mathbb{D} \), and that

\[
L_f := \sup_{x \in \mathbb{D}} \| \nabla f(x) \|_2 < \infty, \tag{9}
\]

\[
H_f := \sup_{x \in \mathbb{D}} \| \nabla^2 f(x) \|_2 < \infty, \tag{10}
\]

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 \( \ell_2 \)-norm of vectors and the spectral norm of matrices. Moreover, for \( \epsilon > 0 \), let \( \mathbb{D}_\epsilon \subset \mathbb{D} \) denote the \( \epsilon \)-interior of \( \mathbb{D} \), namely \( \mathbb{D}_\epsilon = \{ x \in \mathbb{D} : B_{x,\epsilon} \subseteq \mathbb{D} \} \). Throughout, \( B_{x,\epsilon} \subset \mathbb{R}^n \) denotes the (open) Euclidean ball of radius \( \epsilon \) centered at \( x \).

Consider \( \Sigma_\mu \in \mathbb{R}^{n \times n} \) defined as

\[
\Sigma_\mu := \mathbb{E}_x [\nabla f(x) \cdot \nabla f(x)^*] = \int_{\mathbb{D}} \nabla f(x) \cdot \nabla f(x)^* \mu(dx), \tag{11}
\]

where \( \mathbb{E}_x \) computes the expectation with respect to \( x \sim \mu \). Our objective in this work is to estimate \( \Sigma_\mu \). To that end, consider \( N \) random points drawn independently from \( \mu \) 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

\[
\hat{\Sigma}_X := \frac{1}{N} \sum_{x \in X} \nabla f(x) \cdot \nabla f(x)^*, \tag{12}
\]

is an unbiased estimator for \( \Sigma_\mu \) in (11). To interpret (12), 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 \( \hat{\Sigma}_X \) approximates \( \Sigma_\mu \). Its proof is included in Appendix B for completeness; see [10] for related results concerning the accuracy of \( \hat{\Sigma}_X \) as an estimate of \( \Sigma_\mu \).

**Proposition 1.** Let \( X \in \mathbb{R}^{n \times N} \) contain \( N \) independent samples drawn from the probability measure \( \mu \). Then, \( \hat{\Sigma}_X \) is an unbiased estimator for \( \Sigma_\mu \in \mathbb{R}^{n \times n} \), see (11) and (12). Moreover, except for a probability of at most \( n^{-1} \), it holds that

\[
\| \hat{\Sigma}_X - \Sigma_\mu \|_F \lesssim \frac{L_f^2 \log n}{\sqrt{N}}. \tag{13}
\]

Since only point values of \( f(\cdot) \) are at our disposal, we cannot compute \( \hat{\Sigma}_X \) directly. Instead, we will systematically generate random points near the point cloud \( X \) and then estimate \( \hat{\Sigma}_X \) 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 \( 2\epsilon \)-separated point cloud that belongs to the \( \epsilon \)-interior of \( \mathbb{D} \). Formally, fix \( \epsilon \leq \epsilon_X \), where

\[
\epsilon_X := \sup \{ \epsilon' : X \subset \mathbb{D}_{\epsilon'} \text{ and } \|x - x'\|_2 \geq 2\epsilon', \forall x, x' \in X, x \neq x' \}. \tag{14}
\]
Let
\[ B_{X,\epsilon} := \bigcup_{x \in X} B_{x,\epsilon} \subseteq \mathbb{D} \]  
(15)
denote the \( \epsilon \)-neighborhood of the point cloud \( X \). Consider the conditional probability measure on \( B_{X,\epsilon} \) described as
\[ \mu_{X,\epsilon} = \begin{cases} \mu / \mu(B_{X,\epsilon}), & \text{inside } B_{X,\epsilon}, \\ 0, & \text{outside } B_{X,\epsilon}. \end{cases} \]  
(16)
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 \( \Sigma_{\mu} \) as a function of \( X, Y_{X,\epsilon} \subset \mathbb{D} \) and evaluations of \( f(\cdot) \) at these points is proposed by \( \dot{\Sigma}_{X,Y_{X,\epsilon}} \) in Algorithm 1.

3. Theoretical guarantees. Recalling (11) and (12), how well does \( \dot{\Sigma}_{X,Y_{X,\epsilon}} \) in Algorithm 1 approximate \( \dot{\Sigma}_{X} \) and in turn \( \Sigma_{\mu} \)? Parsing the answer requires introducing additional notation and imposing a certain regularity assumption on \( \mu \). 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 \( \epsilon \)-neighbors of \( x \) in \( Y_{X,\epsilon} \). In our notation, this can be written as
\[ Y_{x,\epsilon} := Y_{X,\epsilon} \cap B_{x,\epsilon}, \quad \#Y_{x,\epsilon} = N_{x,\epsilon}. \]  
(17)
Because \( \epsilon \leq \epsilon_{X} \) is small, see (14), these neighborhoods do not intersect, that is
\( Y_{x,\epsilon} \cap Y_{x',\epsilon} = \emptyset, \quad \forall x, x' \in X, \ x \neq x' \);
therefore, \( Y_{X,\epsilon} \) is simply partitioned into \( \#X = N \) subsets \( \{Y_{x,\epsilon}\}_{x \in X} \). Observe also that, conditioned on \( x \in X \) and \( N_{x,\epsilon} \), each neighbor \( y \in Y_{x,\epsilon} \) follows the conditional probability measure described as follows:
\[ y \mid x, N_{x,\epsilon} \sim \mu_{x,\epsilon} = \begin{cases} \mu / \mu(B_{x,\epsilon}), & \text{inside } B_{x,\epsilon}, \\ 0, & \text{outside } B_{x,\epsilon}. \end{cases} \]  
(18)

3.1. Regularity of \( \mu \). In order to introduce the regularity condition imposed on \( \mu \) here, consider first the special case where the domain \( \mathbb{D} \subset \mathbb{R}^n \) is bounded and \( \mu \) is the uniform probability measure on \( \mathbb{D} \). Then, for \( \epsilon > 0 \) and arbitrary \( \epsilon \)-interior point \( x \in \mathbb{D} \), the conditional measure \( \mu_{x,\epsilon} \) too is the uniform measure on \( B_{x,\epsilon} \), see (18). Draw \( y \) from \( \mu_{x,\epsilon} \), that is, \( y \mid x \sim \mu_{x,\epsilon} \) in our notation. Then it is easy to verify that \( y - x \) is an isotropic random vector, namely
\[ \mathbb{E}_{y \mid x} [(y - x)(y - x)^{\ast}] = C \cdot I_{n}, \]
for some factor \( C \).\(^4\) Above, \( I_{n} \in \mathbb{R}^{n \times n} \) is the identity matrix and \( \mathbb{E}_{y \mid x} [\cdot] = \mathbb{E}_{y} [\cdot \mid x] \) 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 \( \mu \)) Throughout this paper, we assume that there exist \( \epsilon_{\mu}, K_{\mu} > 0 \) such that for all \( \epsilon \leq \epsilon_{\mu} \), the following requirement holds for any arbitrary \( \epsilon \)-interior point \( x \in \mathbb{D}_{\epsilon} \).

\(^4\)A simple calculation shows that \( C = 1/n \). See Appendix D.
Given \( x \), draw \( y \) from the conditional measure on the \( \epsilon \)-neighborhood of \( x \), namely \( y \mid x \sim \mu_{x, \epsilon} \) defined in (18). Then, for every \( \gamma_1 \geq 0 \) and arbitrary (but fixed) \( v \in \mathbb{R}^n \), it holds that
\[
\Pr_{y \mid x} \left[ \| P_{x,y} \cdot v \|_2^2 > \gamma_1 \cdot \frac{\|v\|_2^2}{n} \right] \lesssim e^{-K_\mu \gamma_1},
\]
(19)
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, \( \Pr_{y \mid x} [\cdot] = \Pr_y [\cdot \mid x] \)
stands for conditional probability.

Roughly speaking, under Assumption 1, \( \mu \) is locally isotropic. Indeed, this assumption is met when \( \mu \) is the uniform probability measure on \( D \), as shown in Appendix C. Moreover, Assumption 1 is not too restrictive. One would expect that a probability measure \( \mu \), if dominated by the uniform measure on \( D \) 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 \( \| P_{x,y} v \|_2 \) [58, Lemma 5.5]. Finally, given the point cloud \( X \), we also conveniently set
\[
\epsilon_{\mu,X} : = \min [\epsilon_\mu, \epsilon_X]. \quad \text{(see Assumption 1 and (14))}
\]
(20)

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 3.1 and 3.2 below, for a fixed point cloud \( X \), we focus on how well the output of Algorithm 1, namely \( \hat{\Sigma}_{X,Y \mid X, \epsilon} \) defined in (7), satisfies
\[
\| \mathbb{E}_{Y \mid X} [\hat{\Sigma}_{X,Y \mid X, \epsilon} - \hat{\Sigma}_X] \|_F \lesssim B_\mu + n^2 L_f N^{-10} + \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f n^{3/2} \max(K_\mu^{-1/2}, 1) \log^{1/2} N_{X,\epsilon},
\]
(23)
where \( B_{\mu,\epsilon} \) is given explicitly in (44).

In this theorem statement and throughout the paper, we use the notation \( \log^a b \) as shorthand for \( \log \log \cdots \log b \) \( a \) times.

**Remark 1. (Discussion)** Theorem 3.1 describes how well \( \tilde{\Sigma}_{X,Y} \) approximates \( \Sigma_X \), in expectation. To form a better understanding of this result, let us first study the conditions listed in (21).

- The quantity \( \rho_{\mu,X,\epsilon} \), defined in (22), reflects the non-uniformity of \( \mu \) over the set \( D \). In particular, if \( D \subset \mathbb{R}^n \) is bounded and \( \mu \) is the uniform probability measure on \( D \), then \( \rho_{\mu,X,\epsilon} \) achieves its maximum possible value of 1. Non-uniform measures could yield \( \rho_{\mu,X,\epsilon} < 1 \).
- The requirements on \( N_{X,\epsilon} \) and \( N_{X,\min,\epsilon} \) in (21) 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 \( \mu \) is the uniform probability measure on \( D \) and \( \rho_{\mu,X,\epsilon} = 1 \), we might take \( N_{X,\min,\epsilon} \approx \log^2 N \) so that (23) holds with a total of \( N_{X,\epsilon} = O(N \log^2 N) \) samples.
- The requirement that \( N_{X,\epsilon} \geq n^2 \) in (21) 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 (23).

- The first term on the right-hand side of (23), namely \( B_{\mu,\epsilon} \), is given explicitly in (44); it depends on both the probability measure \( \mu \) and the function \( f(\cdot) \), and it can also be viewed as a measure of the non-uniformity of \( \mu \). In fact, as explained in the proof of Theorem 3.1, in the special case where \( \mu \) is the uniform probability measure on a bounded and open set \( D \) 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, (23) can in fact be sharpened by replacing the definition of \( B_{\mu,\epsilon} \) in (44) simply with \( B_{\mu,\epsilon} = 0 \). In general, the more isotropic \( \mu \) is in the sense described in Assumption 1, the smaller \( B_{\mu,\epsilon} \) will be.
- The second term on the right-hand side of (23) 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 (23) can be made arbitrarily small by choosing the neighborhood radius \( \epsilon \) appropriately small (as a function of \( L_f, H_f, n, K_{\mu} \), and \( N_{X,\epsilon} \)). In computational applications, however, choosing \( \epsilon \) too small could raise concerns about numerical precision.
- To get a sense of when the bias in (23) is small relative to the size of \( \Sigma_{\mu} \), it may be appropriate to normalize (23). A reasonable choice would be to divide both sides of (23) by \( L_f^2 \), where \( L_f \) bounds \( \| \nabla f(x) \|_2 \) on \( D \); see (9). 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 (23) can still be made arbitrarily small as described above. In the special case where \( \mu \) is uniform on \( D \) 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 (23) remains zero, as also described above. More
generally, however, $B_{\mu,\epsilon} / L_f^2$ 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 (21) 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 3.2 below.

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

$$Y_{X,\epsilon} \i.i.d. \sim \mu_{X,\epsilon}. \quad \text{(see (16))}$$

(24)

This sampling strategy appears to best fit our fixed budget of $N_{X,\epsilon}$ samples, as it “prioritizes” the areas of $D$ with larger “mass.” For example, suppose that $\mu(dx) \gg \mu(dx')$ and $\nabla f(x) \approx \nabla f(x')$ for a pair $x, x' \in D$. 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 $\Sigma_\mu$ (see (11)). In the same scenario, assume naturally that $\mu(B_{x,\epsilon}) \gg \mu(B_{x',\epsilon})$, 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 $\epsilon$-neighbors in $Y_{X,\epsilon}$ compared to $x'$. Loosely speaking then, the contribution of $x$ to $\hat{\Sigma}_{X,Y_{X,\epsilon}}$ 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 $D$ 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 $\mu$ on $D$ and those falling outside $\bigcup_{x \in X} B_{x,\epsilon}$ 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 $B_{x,\epsilon}$ with $x \in X$ is selected with probability proportional to $\mu(B_{x,\epsilon})$, 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 $\epsilon$ 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 $\mu$ and higher order terms in $f(\cdot)$ by introducing quantities that are simpler to work with but are similar to $\hat{\Sigma}_{X,Y_{X,\epsilon}}$. 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 $\hat{\Sigma}_{X,Y_{X,\epsilon}}$ in (4) to arrive at $\hat{\Sigma}_{X,Y_{X,\epsilon}}$.

Our second result, proved in Section 5.2, is a finite-sample bound for $\hat{\Sigma}_{X,Y_{X,\epsilon}}$.

Theorem 3.2. (Finite-sample bound) Under the same setup as in Theorem 3.1 including the conditions in (21), and under the mild assumptions that $\log(n) \geq 1$,
\( N \geq \log(n) \), and \( N_{X,\epsilon} \geq n \), it holds that
\[
\| \Sigma_{X,Y,\epsilon} - \Sigma_X \|_F
\leq \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f n^2 + B_{\mu,\epsilon}
+ \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] L_f^2 + n^2 L_f^2 N_{X,\epsilon}^{-3}
\]
(25)
except with a probability of \( O \left( n^{-3} + N^{-3} \right) \). Here, \( O(.) \) 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 \( B_{\mu,\epsilon} \) is given explicitly in (44).

**Remark 4. (Discussion)** Theorem 3.2 states that \( \Sigma_{X,Y,\epsilon} \) can reliably estimate \( \Sigma_X \) with high probability. We offer several remarks to help interpret this result.

- The conditions in (21) were discussed in Remark 1. The requirement that \( N_{X,\epsilon} \geq n \frac{1}{20} \) in (21) has been strengthened to \( N_{X,\epsilon} \geq n \) in the statement of Theorem 3.2. 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 (25). As discussed in Remark 1, \( B_{\mu,\epsilon} \) in effect captures the non-uniformity of measure \( \mu \). In particular, the right-hand side of (25) can be sharpened by setting \( B_{\mu,\epsilon} = 0 \) in the setting described in that remark.
- Similar to Remark 1, the terms involving \( \epsilon \) on the right hand side of (25) can be made negligible by choosing \( \epsilon \) to be suitably small. We omit these terms in the discussion below.
- The fourth term on the right hand side of (25) 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 (25) is negligible compared to the fourth, and we omit this in our discussion below.

**Remark 5. (Estimating \( \Sigma_{\mu} \))** Let
\[
\rho_{\mu,\epsilon} := \min_{x \in B} \mu(B_{x,\epsilon}) \leq \max_{x' \in B} \mu(B_{x',\epsilon}),
\]
and note that \( \rho_{\mu,X,\epsilon} \geq \rho_{\mu,\epsilon} \) for any \( X \); see (22). Now combining Theorem 3.2 with Proposition 1, removing the conditioning on \( X \), and omitting the negligible terms yields
\[
\left\| \Sigma_{X,Y,\epsilon} - \Sigma_{\mu} \right\|_F
\leq B_{\mu,\epsilon} + \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] L_f^2 + L_f^2 \log n, \quad \text{(26)}
\]
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 \( f(\cdot) \). This conclusion holds for any value of \( \epsilon \) small enough that (i) \( \epsilon_X \geq \epsilon \) (see (14)) with high probability over the random draw of \( X \) and (ii) the terms involving \( \epsilon \) on the right hand side of (25) are made negligible. As suggested in Remark 1, we can normalize this bound by dividing both sides by \( L_f^2 \):
\[
\frac{\left\| \Sigma_{X,Y,\epsilon} - \Sigma_{\mu} \right\|_F}{L_f^2} \leq \frac{B_{\mu,\epsilon}}{L_f^2} + \log^4(N_{X,\epsilon}) \cdot \frac{n\sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] + \frac{\log n}{\sqrt{N}}, \quad \text{(27)}
\]
We discuss the terms appearing on the right hand side of (27):

• As described in Remark 1, in some settings $B_{\mu,\epsilon} / L^2 f$ will be zero, while in other settings controlling $B_{\mu,\epsilon} / L^2 f$ will require choosing $N_{X,\epsilon} \gtrsim \sqrt{n} \log^2 N$.

• The second and third terms in (27) 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 $B_{\mu,\epsilon} / L^2 f$) 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 (27) will dominate the third. Controlling the second term in (27) 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 [4], (26) might be used to control the distance between the spectrum of $\Sigma_{X,Y,\epsilon}$ and that of $\Sigma_{\mu}$. Likewise, given an integer $r \leq n$, standard perturbation results [60] might be deployed to measure the principal angle between the span of the leading $r$ eigenvectors of $\Sigma_{X,Y,\epsilon}$ and an $r$-dimensional active subspace of $f(\cdot)$. To obtain the sharpest bounds, both these improvements would require controlling the spectral norm of $\Sigma_{X,Y,\epsilon} - \Sigma_{\mu}$ rather than its Frobenius norm (which is bounded in Theorem 3.2 above). A detailed argument favoring the spectral norm in the context of active subspaces is also given in [38]. Controlling the spectral norm of the error however appears to be considerably more difficult. As an aside, let us point out that the spectrum of $\Sigma_{\mu}$ in relation to $f(\cdot)$ has been studied in [19, 57]. Another interesting question for future work is the extension of these results to the case where $f(\cdot)$ is vector-valued, rather than scalar-valued; see [64] and the references therein.

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 [46, 15, 9, 54, 24, 34, 26, 43, 7, 35]. 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 $\Sigma_{\mu}$, which will have rank $r$. This illuminates the connection between ridge approximation and second-moment matrices.
In [19], 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 [57, 5] 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 3.2 which, in contrast, provides a guarantee on learning the entire second-moment matrix \( \Sigma_\mu \) and holds without any assumption (such as low rank) on \( \Sigma_\mu \). 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 [57] 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 [57], however, the sampling complexity in this general case will scale with \( n^5 \) (ignoring log factors). In contrast, our bound (26) requires only that the total number of function samples \( N + N_{X,\varepsilon} \) 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 [10, 11, 12]. 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 \( \Sigma_\mu \).

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 [30], the aim is to construct \( f(\cdot) \) using functions of the form \( \sum_i h_i(a_i^*x) \) [22, 33, 16]. Further connections with neural networks are studied in [45], Chapter 11. See also [59, 56] for connections with Gaussian process regression and uncertainty quantification. Sufficient dimension reduction and related topics [39, 63, 13, 62, 23, 52, 32, 25] 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 [6, 8, 14, 36, 49]. In this context, it appears that [3, 37, 2, 47] are the most relevant to the present work, in part because of their lack of any structural assumptions. For the sake of brevity, we focus on [3], which offers an unbiased estimator for the covariance matrix of a random vector \( x \) given few measurements of multiple realizations of \( x \) in the form of \( \{\Phi_t x_t\}_t \), for low-dimensional (and uniformly random) orthogonal projection matrices \( \{\Phi_t\}_t \). It is important to point out that, by design, the estimator in [3] is not applicable to our setup.\(^5\) Our framework might be interpreted as sum of rank-1 projections. To further complicate matters, the probability

\(^5\)The use of finite differences will effectively replace \( \Phi_t x_t \) in \( \hat{\Sigma}_1 \) in [3, Section 3] with a sum of rank-1 projections of \( x_t \).
Moreover, consider the event \( E \) for 5.1.

Proof of Theorem 3.1.

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

5.1. Proof of Theorem 3.1. Let us begin by outlining the proof strategy.

- First, we introduce a new quantity: \( \hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} \in \mathbb{R}^{n \times n} \). Conditioned on a certain “good” event \( \mathcal{E}_1 \), \( \hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} \) is easier to work with than \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \).
- 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 5.2 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 5.3 below shows that \( \hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} \) is a nearly unbiased estimator of \( \hat{\Sigma}_X \):

\[
\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X}[\hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon}] \approx \hat{\Sigma}_X. \tag{28}
\]

- Lastly, we remove the conditioning on \( \mathcal{E}_1 \cap \mathcal{E}_2 \) to complete the proof of Theorem 3.1.

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

\[
\hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} := \frac{1}{N} \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{\text{min},\epsilon}^{-1} \right)^{-1} \sum_{N_{x,\epsilon} \geq N_{\text{min},\epsilon}} \hat{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \hat{\nabla}_{Y_{x,\epsilon}} f(x)^* - \frac{\sum_{N_{x,\epsilon} \geq N_{\text{min},\epsilon}} \left\| \hat{\nabla}_{Y_{x,\epsilon}} f(x) \right\|^2}{(1 + \frac{2}{n}) N_{\text{min},\epsilon} + n + 1 - \frac{2}{n} I_n},
\]

(29)

Here,

\[
\hat{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n,
\]

(30)

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

Lemma 5.1. Fix \( X \) and \( \epsilon \in (0, \epsilon_\mu, X] \). It holds that

\[
\left\| \hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} - \hat{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \leq \frac{1}{2} \epsilon^2 H_f^2 n^2 + 2\epsilon L_f H_f n^2.
\]

(31)

Moreover, consider the event

\[
\mathcal{E}_1 := \left\{ \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \left\| P_{x,y} \cdot \nabla f(x) \right\|^2_2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\},
\]

(32)

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

\[
\left\| \hat{\Sigma}_{X,Y_{X,\epsilon},\epsilon} - \hat{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \leq \frac{1}{2} \epsilon^2 H_f^2 n^2 + 2\epsilon L_f H_f Q_{X,\epsilon}^{1/2} n^{3/2}.
\]

(33)
Thanks to Assumption 1, the event $E_1$ 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

$$\Pr_{Y_{X,\epsilon}|X} \left[ E_1^C \right] \lesssim N_{X,\epsilon}^{1-K_{\gamma_2}},$$

(34)

which follows from (19) and an application of the union bound (similar to the slightly more general result in Lemma I.3).

Roughly speaking, in light of Lemma 5.1, $\Sigma_{X,Y_{X,\epsilon}} \approx \Sigma_{X,Y_{X,\epsilon}}$. It therefore suffices to study the bias of $\Sigma_{X,Y_{X,\epsilon}}$ 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 5.2.** Fix $X$ and $\epsilon \in (0, \epsilon_X]$. With $\gamma_3 \geq 1$, assume that

$$N_{X,\epsilon} \gtrsim \gamma_3^2 \log^2 N \cdot \min_{x \in X} \frac{\mu(B_{X,\epsilon})}{\mu(B_{x,\epsilon})}.$$  

(35)

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

$$\frac{1}{2} \cdot \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})} N_{x,\epsilon} \leq N_{x,\epsilon} \leq \frac{3}{2} \cdot \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})} N_{x,\epsilon}, \quad \forall x \in X.$$  

(36)

To use Lemma 5.2 here, we proceed as follows. For $\gamma_3 \geq 1$, suppose that

$$N_{X,\epsilon} \gtrsim N_{X,min,\epsilon} \cdot \frac{\mu(B_{X,\epsilon})}{\mu(B_{x,\epsilon})},$$

(37)

and consider the event

$$E_2 := \bigcap_{x \in X} \left\{ N_{x,\epsilon} \geq \frac{1}{2} \cdot \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})} N_{X,\epsilon} \gtrsim N_{X,min,\epsilon} \right\},$$

(38)

where, in particular, each $x \in X$ has at least $N_{X,min,\epsilon}$ neighbors in $Y_{X,\epsilon}$. In light of Lemma 5.2, $E_2$ is very likely to happen. To be specific,

$$\Pr_{Y_{X,\epsilon}|X} \left[ E_2^C \right] \leq N^{1-\gamma_3},$$

(39)

provided that

$$N_{X,\epsilon} \gtrsim \frac{N_{X,min,\epsilon} \cdot \mu(B_{X,\epsilon})}{\min_{x \in X} \mu(B_{x,\epsilon})} = \frac{N_{X,min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, \quad \text{(see (37))}$$

(40)

where we conveniently defined

$$\rho_{\mu,X,\epsilon} = N \cdot \min_{x \in X} \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})}.$$  

(41)

Conditioned on the event $E_2$, $\Sigma_{X,Y_{X,\epsilon}}$ in (29) takes the following simplified form:

$$\Sigma_{X,Y_{X,\epsilon}} = \frac{1}{N} \left( 1 + \frac{1}{1 + 2/n} \cdot N_{X,min,\epsilon}^{-1} \right)^{-1}$$

$$\cdot \left( \sum_{x \in X} \hat{Y}_{Y_{X,\epsilon}} f(x) \hat{Y}_{Y_{X,\epsilon}} f(x)^* - \frac{\sum_{x \in X} \| \hat{Y}_{Y_{X,\epsilon}} f(x) \|_2^2}{\left( 1 + \frac{2}{n} \right) N_{X,min,\epsilon} + n + 1 - \frac{2}{n} \cdot I_n} \right).$$

(42)

Using the above simplified form, we will prove the following result in Appendix G. Roughly speaking it states that, conditioned on the event $E_2$, $\Sigma_{X,Y_{X,\epsilon}}$ is a nearly-unbiased estimator of $\Sigma_X$. 


Lemma 5.3. Fix $X$ and $\epsilon \in (0, \epsilon_{\mu,X})$. Then, it holds that

$$\left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] - \hat{\Sigma}_{X} \right\|_F \leq B_{\mu, \epsilon},$$

(43)

where

$$B_{\mu, \epsilon} := \frac{2 B''_{\mu, \epsilon}}{N_{X,\min, \epsilon}} + 4 B'_{\mu, \epsilon} \left( B'_{\mu, \epsilon} + 1 \right) L_f^2 + \frac{2 L_f^2 (1 + \sqrt{n})}{N_{X,\min, \epsilon}},$$

(44)

Moreover, suppose that $\mu$ 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 $\xi_2$, one can replace $B_{\mu, \epsilon}$ with 0, and thus $\hat{\Sigma}_{X,Y,X,}$ is an unbiased estimator of $\Sigma_X$.

Next, we remove the conditioning on the event $\xi_2$, with the aid of the following bounds:

$$\left\| \hat{\Sigma}_{X} \right\|_F \leq L_f^2,$$  \hspace{1cm} (see (12) and (9))

$$\left\| \nabla_{Y,X,} f(x) \right\|_2 \leq n L_f, \hspace{1cm} \forall x \in X, \hspace{1cm} (\text{see (30) and (9)})$$

Then, we write that

$$\left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] - \hat{\Sigma}_{X} \right\|_F$$

$$= \left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] \cdot \mathbb{P}_{Y,X,|\xi_2} [\xi_2] + \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] \cdot \mathbb{P}_{Y,X,|\xi_2} [\xi_2^C] - \hat{\Sigma}_{X} \right\|_F$$

$$\leq \left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] - \hat{\Sigma}_{X} \right\|_F \cdot \mathbb{P}_{Y,X,|\xi_2} [\xi_2]$$

$$+ \left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] - \hat{\Sigma}_{X} \right\|_F \cdot \mathbb{P}_{Y,X,|\xi_2} [\xi_2^C]$$

$$\leq \left\| \mathbb{E}_{Y,X,|\xi_2, X} \left[ \hat{\Sigma}_{X,Y,X,}\right] - \hat{\Sigma}_{X} \right\|_F \cdot \mathbb{P}_{Y,X,|\xi_2} [\xi_2]$$

$$+ \left( \sup_{Y,X,|\xi_2} \left\| \hat{\Sigma}_{X,Y,X,}\right\|_F + n^2 L_f^2 \cdot N^{1-\gamma_3} \right) \left( \text{see (45) and (39)} \right)$$

$$\leq B_{\mu, \epsilon} + n^2 L_f^2 \cdot N^{1-\gamma_3}, \hspace{1cm} \text{(see Lemma 5.3 and (12))}$$

(46)
which, to reiterate, holds with $N_{X,\min} \gtrsim \gamma_2^2 \log^2 N$ and under (40). Lastly, we reintroduce $\Sigma_{X,Y_{X,\epsilon}}$ by invoking Lemma 5.1 as follows:

\[
\|E_{X,\epsilon|X} \left[ \Sigma_{X,Y_{X,\epsilon}} - \Sigma_{X,Y_{X,\epsilon}} \right] \|_F \\
\leq \|E_{X,\epsilon|X} \left[ \Sigma_{X,Y_{X,\epsilon}} - \Sigma_{X,Y_{X,\epsilon}} \right] \|_F \cdot \Pr_{Y_{X,\epsilon}|X} [E_1] \\
+ \left( \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F + \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F \right) \cdot \Pr_{Y_{X,\epsilon}|X} [E_1^C] \quad \text{(similar to (46))}
\]

\[
\lesssim \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f Q_{X,\epsilon}^{1/2} n^{3/2} \\
+ \left( \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F + \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F \right) \cdot \Pr_{Y_{X,\epsilon}|X} [E_1^C] \quad \text{(see (33))}
\]

\[
\lesssim \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{1/2} n^{3/2} \\
+ \left( \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F + \sup \| \Sigma_{X,Y_{X,\epsilon}} \|_F \right) \cdot N_{X,\epsilon}^{1-K_\mu} \gamma_2 \quad \text{(Q_{X,\epsilon} in (34))}
\]

\[
\lesssim \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{1/2} n^{3/2} \\
+ \frac{1}{2} \left( \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f n^2 + L_f n^2 \right) \cdot N_{X,\epsilon}^{1-K_\mu} \gamma_2 \quad \text{(see (31) and (45))}
\]

Combining the above bound with (46) yields that

\[
\|E_{X,\epsilon|X} \left[ \Sigma_{X,Y_{X,\epsilon}} - \Sigma_X \right] \|_F \\
\leq \|E_{X,\epsilon|X} \left[ \Sigma_{X,Y_{X,\epsilon}} - \Sigma_X \right] \|_F + \|E_{X,\epsilon|X} \left[ \Sigma_{X,Y_{X,\epsilon}} - \Sigma_{X,Y_{X,\epsilon}} \right] \|_F \quad \text{(triangle ineq.)}
\]

\[
\lesssim \frac{B_{\mu,\epsilon} n^2 L_f}{2} \cdot N_{X,\epsilon}^{1-K_\mu} + \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{1/2} n^{3/2} \\
+ \left( \epsilon^2 H_f^2 n^2 + \epsilon L_f H_f n^2 + L_f n^2 \right) \cdot N_{X,\epsilon}^{1-K_\mu} \gamma_2 \quad \text{(see (46) and (47))}
\]

\[
= B_{\mu,\epsilon} + n^2 L_f \left( N_{X,\epsilon}^{1-K_\mu} + N_{X,\epsilon}^{1-K_\mu} \gamma_2 \right) + \epsilon^2 H_f^2 n^2 \left( 1 + N_{X,\epsilon}^{1-K_\mu} \gamma_2 \right) \\
+ \epsilon L_f H_f n^2 \left( \frac{(\gamma_2 \log N_{X,\epsilon})}{n} \right)^{1/2} + N_{X,\epsilon}^{1-K_\mu} \gamma_2
\]

\[
\lesssim B_{\mu,\epsilon} + n^2 L_f \left( N_{X,\epsilon}^{10} + N_{X,\epsilon}^{10} \right) + \epsilon^2 H_f^2 n^2 \left( 1 + N_{X,\epsilon}^{10} \right) \\
+ \epsilon L_f H_f n^2 \left( \left( \frac{\max(K^{-1}_{\mu}, 1) \log N_{X,\epsilon}}{n} \right)^{1/2} + N_{X,\epsilon}^{10} \right)
\]

(setting $\gamma_2 = 11 \max(K^{-1}_{\mu}, 1)$ and $\gamma_3 = 11$)

\[
\lesssim B_{\mu,\epsilon} + n^2 L_f n^{-10} + \epsilon^2 H_f^2 n^2 \\
+ \epsilon L_f H_f n^2 \left( \left( \frac{\max(K^{-1}_{\mu}, 1) \log N_{X,\epsilon}}{n} \right)^{1/2} + N_{X,\epsilon}^{10} \right) \quad (N_{X,\epsilon} \geq N \geq 1)
\]

\[
\lesssim B_{\mu,\epsilon} + n^2 L_f n^{-10} + \epsilon^2 H_f^2 n^2
\]
\[ + \epsilon L_f H_f n^{3/2} \max(K_{\mu}^{-1/2}, 1) \log^{1/2} N_{X,\epsilon}. \quad (N_{X,\epsilon} \geq n^{2/7}) \]  

This completes the proof of Theorem 3.1.

5.2. **Proof of Theorem 3.2.** At a high level, the proof strategy here matches that of Theorem 3.1. First, we replace \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \) with the simpler quantity \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \) defined in (29). More specifically, in light of Lemma 5.1, it suffices to study \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \) in the sequel.

Next, for \( N_{X,\text{min},\epsilon} > 0 \) to be set later, recall the “good” event \( \mathcal{E}_2 \) in (38) whereby every \( x \in X \) has at least \( N_{X,\text{min},\epsilon} \) neighbors in \( Y_{X,\epsilon} \). Conditioned on the event \( \mathcal{E}_2 \), \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \) takes the simpler form of (42), using which we prove the following result in Appendix H.

**Lemma 5.4.** 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

\[
\left\| \hat{\Sigma}_{X,Y_{X,\epsilon}} - \hat{\Sigma}_X \right\|_F \lesssim \mathbb{B}_{\mu,\epsilon} + \gamma_2^2 \log^4(N_{X,\epsilon}) \cdot \frac{n \sqrt{\log n}}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] L_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))},
\]

for \( \gamma_2 \geq 1 \) and \( \gamma_2 \geq 3 \), except with a probability \( \lesssim e^{-n^2 \log \gamma_2} + n^{1-\gamma_2 \log(N_{X,\epsilon})} + N_{X,\epsilon}^{1-\gamma_2} \), (Lemma 5.4 and (39))

under (37). Lastly, we reintroduce \( \hat{\Sigma}_{X,Y_{X,\epsilon}} \) by invoking Lemma 5.1: it holds that

\[
\left\| \hat{\Sigma}_{X,Y_{X,\epsilon}} - \hat{\Sigma}_X \right\|_F \lesssim \left\| \hat{\Sigma}_{X,Y_{X,\epsilon}} - \hat{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \left\| \hat{\Sigma}_{X,Y_{X,\epsilon}} - \hat{\Sigma}_X \right\|_F \quad \text{(triangle inequality)}
\]

\[
\lesssim \frac{1}{2} \epsilon^2 H_f^2 n^2 + 2\epsilon L_f H_f n^2 + R \quad \text{(see Lemma 5.1)}
\]

with a failure probability of the order of

\[ e^{-n^2 \log \gamma_2} + n^{1-\gamma_2 \log(N_{X,\epsilon})} + N_{X,\epsilon}^{1-\gamma_2} \quad \text{(see (50))}, \]

assuming (37) 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 3.2.

6. **Acknowledgments.** The authors thank Rachel Ward for her helpful discussions during the preparation of this work. At the time, author AE was a Peter O’Donnell, Jr. Postdoctoral Fellow at UT Austin, mentored by Rachel Ward. AE would also like to thank Hemant Tyagi for many interesting conversations regarding ridge approximation.
REFERENCES

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

[2] 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)*, 2014, 1341–1349.

[3] M. Azizyan, A. Krishnamurthy and A. Singh, Extreme compressive sampling for covariance estimation, *IEEE Trans. Inform. Theory*, 64 (2018), 7613–7635, arXiv:1506.00898.

[4] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*, Princeton reference, Princeton University Press, 2009, https://books.google.co.uk/books?id=x7isojLkD7c.

[5] 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)*, IEEE, 2015, 2189–2193.

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

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

[8] 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 (2015), 4034–4059.

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

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

[11] P. G. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*, SIAM, Philadelphia, 2015, https://books.google.com/books?id=TOJ9BwAAQBAJ.

[12] P. G. Constantine, A. Eftekhari and R. Ward, A near-stationary subspace for ridge approximation, *Computer Methods in Applied Mechanics and Engineering*, 326 (2017), 402–421, arXiv:1606.01929.

[13] 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*, American Statistical Association Alexandria, VA, 1994, 18–25.

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

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

[16] D. L. Donoho and I. M. Johnstone, Projection-based approximation and a duality with kernel methods, *The Annals of Statistics*, 17 (1989), 58–106.

[17] A. Eftekhari, L. Balzano and M. B. Wakin, What to expect when you are expecting on the Grassmannian, *IEEE Signal Processing Letters*, 24 (2017), 872–876, arXiv:1611.07216.

[18] A. Eftekhari, M. B. Wakin and R. A. Ward, MC²: A two-phase algorithm for leveraged matrix completion, *Inf. Inference*, 7 (2018), 581–604, arXiv:1609.01795.

[19] M. Fornasier, K. Schnass and J. Vybiral, Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *The Journal of Machine Learning Research*, 5 (2004), 1373–1388.

[20] A. T. Glaws, P. G. Constantine and R. D. Cook, Inverse regression for ridge recovery, arXiv preprint, arXiv:1702.02227.
[26] G. K. Golubev, Asymptotic minimax estimation of regression in the additive model, *Problem Peredachi Informatsii*, **28** (1992), 3–15.

[27] A. Gonen, D. Rosenbaum, Y. Eldar and S. Shalev-Shwartz, The sample complexity of subspace learning with partial information, *J. Mach. Learn. Res.*, **17** (2016), Paper No. 52, 21 pp. [arXiv:1402.4844](https://arxiv.org/abs/1402.4844).

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

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

[30] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Taylor and Francis, 1990, [https://books.google.com/books?id=qa29r1Ze1coC](https://books.google.com/books?id=qa29r1Ze1coC).

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

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

[33] P. J. Huber, *Projection pursuit*, The annals of Statistics, **13** (1985), 435–475.

[34] A. B. Juditsky, O. V. Lepski and A. B. Tsybakov, Nonparametric estimation of composite functions, *The Annals of Statistics*, **37** (2009), 1360–1404.

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

[36] 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)*, 2012, 551–558.

[37] A. Krishnamurthy, M. Azizyan and A. Singh, Subspace learning from extremely compressed measurements, arXiv preprint, [arXiv:1404.0751](https://arxiv.org/abs/1404.0751).

[38] R. Lam, O. Zahm, Y. Marzouk and K. Willcox, Multifidelity dimension reduction via active subspaces, arXiv preprint, [arXiv:1809.05567](https://arxiv.org/abs/1809.05567).

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

[40] 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** (2007), 20167–20172.

[41] P. L. Loh and M. J. Wainwright, High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity, *Ann. Statist.*, **40** (2012), 1637–1664.

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

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

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

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

[46] A. Pinkus, *Ridge Functions*, Cambridge Tracts in Mathematics, Cambridge University Press, 2015, [https://books.google.com/books?id=dtMcGqAACBAJ](https://books.google.com/books?id=dtMcGqAACBAJ).

[47] F. Pourkamali-Anaraki, Estimation of the sample covariance matrix from compressive measurements, *IET Signal Processing*, **10** (2016), p1089, [arXiv:1512.08887](https://arxiv.org/abs/1512.08887).

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

[49] 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** (2011), 935–980.

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

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

[52] A. M. Samorov, Exploring regression structure using nonparametric functional estimation, *Journal of the American Statistical Association*, **88** (1993), 836–847.
Appendix 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 $\mathcal{A}^c$):\footnote{To see why the first inequality holds, note that
\begin{align*}
\mathbb{E}_z^p [z] & = \mathbb{E}_z^p [z \cdot 1_{\mathcal{A}}(z) + z \cdot 1_{\mathcal{A}^c}(z)] \\
 & \leq \mathbb{E}_z^p [z \cdot 1_{\mathcal{A}}(z)] + \mathbb{E}_z^p [z \cdot 1_{\mathcal{A}^c}(z)],
\end{align*}
where $1_{\mathcal{A}}(\cdot)$ is the indicator function for the event $\mathcal{A}$. It is easily verified that
\begin{align*}
\mathbb{E}_z^p [z \cdot 1_{\mathcal{A}}(z)] & \leq \mathbb{E}_{z|\mathcal{A}} [z], \\
\mathbb{E}_z^p [z \cdot 1_{\mathcal{A}^c}(z)] & \leq \mathbb{E}_{z|\mathcal{A}^c} [z],
\end{align*}
from which (54) follows immediately.}

\[ \mathbb{E}_z^p [z] \leq \mathbb{E}_{z|\mathcal{A}} [z] + \sup |z| \cdot \left( \Pr_z [\mathcal{A}^c] \right)^{\frac{1}{2}}, \quad \text{(if } \sup |z| < \infty), \]
\[ \Pr_z [z > z_0] \leq \Pr_{z|\mathcal{A}} [z > z_0] + \Pr_z [\mathcal{A}^c], \quad \forall z_0. \tag{54} \]

We also recall the Bernstein inequality [28].

\textbf{Proposition 2. (Bernstein inequality)} Let $\{A_i\}$ be a finite sequence of zero-mean independent random matrices, and set
\begin{align*}
b & := \max_i \|A_i\|_F, \tag{55} \\
\sigma^2 & := \sum_i \mathbb{E} \|A_i\|^2_F. \tag{56}
\end{align*}
Then, for $\gamma \geq 1$ and except with a probability of at most $e^{-\gamma}$, it holds that
\[
\left\| \sum_i A_i \right\|_F \lesssim \gamma \cdot \max[b, \sigma]. \quad (57)
\]

**Appendix B. Proof of Proposition 1.** Recalling the definition of $\dot{\Sigma}_X$ from (12), we write that
\[
E_X \left[ \dot{\Sigma}_X \right] = \frac{1}{N} \sum_{x \in X} E_X \left[ \nabla f(x) \nabla f(x)^* \right] \quad (\text{see (12)})
\]
\[
= E_x \left[ \nabla f(x) \nabla f(x)^* \right] \quad (#X = N)
\]
\[
= \Sigma_{\mu}, \quad (\text{see (11)}) \quad (58)
\]
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
\[
\dot{\Sigma}_X - \Sigma_{\mu} = \dot{\Sigma}_X - E_X \left[ \dot{\Sigma}_X \right] = \frac{1}{N} \sum_{x \in X} \nabla f(x) \nabla f(x)^* - E_x \left[ \nabla f(x) \nabla f(x)^* \right]
\]
\[
= : \sum_{x \in X} A_x, \quad (59)
\]
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
\[
b = \max_{x \in X} \left\| A_x \right\|_F
\]
\[
= \frac{1}{N} \max_{x \in X} \left\| \nabla f(x) \nabla f(x)^* - E_x \left[ \nabla f(x) \nabla f(x)^* \right] \right\|_F \quad (\text{see (59)})
\]
\[
\leq \frac{1}{N} \max_{x \in X} \left\| \nabla f(x) \nabla f(x)^* \right\|_F + \frac{1}{N} E_x \left\| \nabla f(x) \nabla f(x)^* \right\|_F \quad (\text{triangle, Jensen’s ineqs.})
\]
\[
\leq \frac{2}{N} \sup_{x \in D} \left\| \nabla f(x) \nabla f(x)^* \right\|_F
\]
\[
= \frac{2}{N} \sup_{x \in D} \left\| \nabla f(x) \right\|_2^2
\]
\[
= \frac{2 L_f^4}{N} \quad (\text{see (9)})
\]
and
\[
\sigma^2 = \sum_{x \in X} E_x \left\| A_x \right\|_F^2
\]
\[
= \frac{1}{N} E_x \left\| \nabla f(x) \nabla f(x)^* - E_x \left[ \nabla f(x) \nabla f(x)^* \right] \right\|_F^2 \quad (\text{see (59) and #X = N})
\]
\[
\leq \frac{1}{N} E_x \left\| \nabla f(x) \nabla f(x)^* \right\|_F^2 \quad (E\left\| Z - E[Z] \right\|_2^2 \leq E\left\| Z \right\|_2^4 \text{ for a random matrix } Z)
\]
\[
= \frac{1}{N} E_x \left\| \nabla f(x) \right\|_2^4
\]
\[
\leq \frac{L_f^4}{N}, \quad (\text{see (9)})
\]
and thus
\[ \max[b, \sigma] \leq \frac{2L_f^2}{\sqrt{N}}. \] (60)

Therefore, for \( \gamma_4 \geq 1 \) and except with a probability of at most \( e^{-\gamma_4} \), Proposition 2 dictates that
\[
\| \hat{\Sigma}_X - \Sigma_\mu \|_F = \left\| \sum_{x \in X} A_x \right\|_F \quad \text{(see (59))} \\
\leq \gamma_4 \cdot \max[b, \sigma] \\
\leq \gamma_4 \cdot \frac{L_f^2}{\sqrt{N}},
\]
which completes the proof of Proposition 1 when we take \( \gamma_4 = \log n \).

Appendix C. Uniform measure satisfies Assumption 1. We verify in this appendix that the uniform probability measure on \( D \) satisfies Assumption 1. Fix arbitrary \( \epsilon > 0 \) and \( x \) in the \( \epsilon \)-interior of \( D \subseteq \mathbb{R}^n \), namely \( x \in D_\epsilon \), assuming that \( D_\epsilon \neq \emptyset \). The conditional measure in the neighborhood \( B_{x, \epsilon} \) too is uniform, so that \( y|x \sim \text{uniform}(B_{x, \epsilon}) \). Then, for fixed \( v \in \mathbb{R}^n \) with \( \|v\|_2 = 1 \), observe that
\[ \|P_{x,y}v\|_2^2 \sim \text{beta}\left(\frac{1}{2}, n-\frac{1}{2}\right). \] (61)

To study the tail bound of the random variable \( \|P_{x,y}v\|_2^2 \), we proceed as follows. We note that (19) 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
\[
\Pr_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] \\
= \Pr_{y|x}\left[\|P_{x,y}v\|_2^{2\lambda} > \left(\frac{\gamma_1}{n}\right)^\lambda\right] \quad (\lambda > 0) \\
\leq \left(\frac{\gamma_1}{n}\right)^{-\lambda} \mathbb{E}\left[\|P_{x,y}v\|_2^{2\lambda}\right] \quad \text{(Markov’s inequality)} \\
= \left(\frac{\gamma_1}{n}\right)^{-\lambda} B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right) B\left(\frac{1}{2}, \frac{n-1}{2}\right), \quad (62)
\]
where
\[
B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\]
(63)
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 \( \lambda \) above, we rewrite (62) as
\[ \Pr_{y|x}\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] \leq e^{-\lambda \log\left(\frac{\gamma_1}{n}\right) + \log(B(\lambda + \frac{1}{2}, \frac{n-1}{2})) - \log(B(\frac{1}{2}, \frac{n-1}{2}))} \\
= e^{l(\lambda)}. \] (64)

In order to minimize \( l(\cdot) \), we compute its derivative:
\[
l'(\lambda) = -\log\left(\frac{\gamma_1}{n}\right) + \frac{d}{d\lambda} \log\left(B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)\right) - \frac{d}{d\lambda} \log\left(B\left(\frac{1}{2}, \frac{n-1}{2}\right)\right)
\]
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, [44]). To guide our choice of \( \lambda \), note that if \( n \) is sufficiently large and we take \( \lambda \) such that \( 1 \ll \lambda \ll n \), we have that

\[
I'(\lambda) = -\log \left( \frac{\gamma_1}{n} \right) + \psi \left( \lambda + \frac{1}{2} \right) - \psi \left( \lambda + \frac{n}{2} \right) \quad \text{(see (65))}
\]

\[
\approx -\log \left( \frac{\gamma_1}{n} \right) + \log \lambda - \log \left( \frac{n}{2} \right)
\]

\[
= -\log \left( \frac{2\lambda}{\gamma_1} \right),
\]

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

\[
\text{Pr}_{y|x} \left[ \|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n} \right] \leq \left( \frac{\gamma_1}{n} \right)^{-\frac{1}{2}} \frac{B \left( \frac{n+1}{2}, \frac{n-1}{2} \right)}{B \left( \frac{1}{2}, \frac{n-1}{2} \right)} \quad \text{(see (62))}
\]

\[
= \left( \frac{\gamma_1}{n} \right)^{-\frac{1}{2}} \frac{\Gamma \left( \frac{n+1}{2} \right) \Gamma \left( \frac{n}{2} \right)}{\Gamma \left( \frac{n}{2} \right) \Gamma \left( \frac{n+1}{2} \right)} \quad \text{(see (63))}
\]

\[
\leq \left( \frac{n}{n+\gamma_1} \right)^{-\frac{n+1}{2}} \quad \text{(see (63))}
\]

\[
\leq \left( \frac{n}{n+\gamma_1} \right)^{-\frac{n-1}{2}} \quad (\gamma_1 > 0)
\]

\[
= \left( 1 + \frac{\gamma_1}{n} \right)^{-\frac{n-1}{2}} \quad \text{(67)}
\]

Therefore, Assumption 1 holds for the uniform probability measure with \( \epsilon_\mu = \infty \) and \( K_\mu = 1/2 \).

**Appendix 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 \( \epsilon\)-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 (18)). 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

$$\left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon},x} \left[ \nabla Y_{x,\epsilon}f(x) \right] - \nabla f(x) \right\|_2 \leq B'_{\mu,\epsilon} L_f + \frac{\epsilon H_f n}{2}. \quad (68)$$

where

$$\nabla Y_{x,\epsilon}f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2} \in \mathbb{R}^n, \quad (69)$$

$$B'_{\mu,\epsilon} := n \cdot \sup_{x \in \mathbb{B}_{x,\epsilon}} \left\| \mathbb{E}_y | x \left[ P_{x,y} \right| - \frac{I_n}{n} \right\| \cdot \left| y|x \sim \mu_{x,\epsilon} \right). \quad (70)$$

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

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

$$\bar{\nabla} Y_{x,\epsilon}f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n, \quad (71)$$

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:

$$\left\| \nabla Y_{x,\epsilon}f(x) - \bar{\nabla} Y_{x,\epsilon}f(x) \right\|_2$$

$$= \frac{n}{N_{x,\epsilon}} \left| \sum_{y \in Y_{x,\epsilon}} \frac{y - x}{\|y - x\|_2} \cdot (f(y) - f(x) - (y - x)^* \nabla f(x)) \right|_2 \quad \left( P_{x,y} = \frac{(y-x)(y-x)^*}{\|y-x\|_2^2} \right)$$

$$\leq \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \left| \frac{y - x}{\|y - x\|_2} \cdot (f(y) - f(x) - (y - x)^* \nabla f(x)) \right|_2 \quad \text{(triangle inequality)}$$

$$= \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \left| f(y) - f(x) - (y - x)^* \nabla f(x) \right| \quad \left( P_{x,y} = \frac{(y-x)(y-x)^*}{\|y-x\|_2^2} \right)$$

$$\leq n \cdot \sup_{y \in \mathbb{B}_{x,\epsilon}} \left| f(y) - f(x) - (y - x)^* \nabla f(x) \right| \quad \left( \#Y_{x,\epsilon} = N_{x,\epsilon} \right)$$

$$\leq n \cdot \frac{H_f \|y-x\|_2}{2} \quad \text{(Taylor’s expansion and (10))}$$

$$\leq n \cdot \frac{H f \cdot \epsilon}{2}. \quad (y \in \mathbb{B}_{x,\epsilon}) \quad (72)$$

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

$$\left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon},x} \left[ \bar{\nabla} Y_{x,\epsilon}f(x) \right] - \nabla f(x) \right\|_2$$

$$= \left\| n \cdot \mathbb{E}_y | x \left[ P_{x,y} \nabla f(x) \right] - \nabla f(x) \right\|_2 \quad \left| y|x \sim \mu_{x,\epsilon} \right).$$

$$= \left\| n \cdot \mathbb{E}_y | x \left[ P_{x,y} \nabla f(x) - \nabla f(x) \right] \right\|_2$$

$$\leq n \cdot \sup_{x \in \mathbb{D}} \left\| \mathbb{E}_y | x \left[ P_{x,y} \right| - \frac{I_n}{n} \right\| \cdot \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2$$

$$=: B'_{\mu,\epsilon} L_f, \quad \text{(see (9))} \quad (73)$$
which, in turn, implies that

\[
\left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon}} \left[ \nabla Y_{x,\epsilon} f(x) \right] - \nabla f(x) \right\|_2 \\
\leq \left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon}} \left[ \nabla Y_{x,\epsilon} f(x) - \nabla Y_{x,\epsilon} f(x) \right] \right\|_2 \\
+ \left\| \mathbb{E}_{Y_{x,\epsilon}|N_{x,\epsilon}} \left[ \nabla Y_{x,\epsilon} f(x) \right] - \nabla f(x) \right\|_2 \\
\leq \frac{n H_f \epsilon}{2} + B'_\mu,\epsilon L_f. \quad \text{(see (72) and (73))} \tag{74}
\]

In particular, when \( \mu \) 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

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

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

\( \square \)

**Appendix E. Proof of Lemma 5.1.** We only verify the second claim, as the other proof is similar. Conditioned on the event \( \mathcal{E}_1 \), note that

\[
\left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 \leq \frac{n}{N_{x,\epsilon} \cdot \max_{y \in Y_{x,\epsilon}} \left\| P_{x,y} \cdot \nabla f(x) \right\|_2 \quad \text{(see (30))}
\]

\[
\leq n \cdot \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \left\| P_{x,y} \cdot \nabla f(x) \right\|_2 \quad \text{(\#} Y_{x,\epsilon} = N_{x,\epsilon})
\]

\[
\leq n \cdot \sqrt{\frac{Q_{x,\epsilon} L_f^2}{n}}. \quad \text{(see (32))} \tag{75}
\]

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

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

\[
\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \nabla Y_{x,\epsilon} f(x) \cdot \nabla Y_{x,\epsilon} f(x)^* \right\|_F \quad \text{(triangle inequality)}
\]

\[
\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 \left( \left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 + \left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 \right) \quad \text{(\#} X = N)
\]

\[
\leq \max_{x \in X} \left\| \nabla Y_{x,\epsilon} f(x) - \nabla Y_{x,\epsilon} f(x) \right\|_2 \\
\cdot \left( \max_{x \in X} \left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 + \max_{x \in X} \left\| \nabla Y_{x,\epsilon} f(x) \right\|_2 \right) \quad \text{(triangle ineq.)}
\]
where the third line above uses the fact that $|\text{trace}(A)| \leq \sqrt{n}\|A\|_F$ for any $A \in \mathbb{R}^{n \times n}$. Recall the definitions of $\tilde{\Sigma}_{X,Y_{X,\epsilon}}$ and $\tilde{\Sigma}_{X,Y_{X,\epsilon}}$ in (7) and (29). Then, by combining (76) and (77), it follows that

$$1 \leq \epsilon \frac{\mu_{Y_{X,\epsilon}}(B_{x,\epsilon})}{\mu(B_{x,\epsilon})}.$$

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

$$1_{y \in B_{x,\epsilon}} \sim \text{Bernoulli} \left( \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})} \right).$$

Then,

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

which, in turn, immediately implies that

$$\frac{1}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \tilde{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \tilde{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right\|_F$$

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

$$\leq \frac{1}{4} \epsilon^2 H_f^2 \bar{n}^{5/2} + \epsilon L_f H_f Q_{X,\epsilon}^{1/2} \bar{n}^{3/2},$$

(76)

where the third line above uses the fact that $|\text{trace}(A)| \leq \sqrt{n}\|A\|_F$ for any $A \in \mathbb{R}^{n \times n}$. Recall the definitions of $\tilde{\Sigma}_{X,Y_{X,\epsilon}}$ and $\tilde{\Sigma}_{X,Y_{X,\epsilon}}$ in (7) and (29). Then, by combining (76) and (77), it follows that

$$\left\| \tilde{\Sigma}_{X,Y_{X,\epsilon}} - \tilde{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F$$

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

$$\leq \frac{1}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \tilde{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \tilde{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right\|_F \cdot \frac{\|L_n\|_F}{n}$$

$$\leq \frac{1}{2} \epsilon^2 H_f^2 \bar{n}^2 + 2 \epsilon L_f H_f Q_{X,\epsilon}^{1/2} \bar{n}^{3/2}.$$

(80)

This completes the proof of Lemma 5.1.

Appendix F. Proof of Lemma 5.2. Our objective is to establish that, given $X$ and neighborhood radius $\epsilon$, 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 $\epsilon$-neighborhood of the point cloud $X$ (see (16)). With $y \sim \mu_{X,\epsilon}$ and for fixed $x \in X$, observe that $y$ belongs to the $\epsilon$-neighborhood of $x$ (namely, $y \in B_{x,\epsilon}$) with the following probability:

$$\Pr_{y \mid x} [y \in B_{x,\epsilon}] = \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})}.$$

(79)

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

$$1_{y \in B_{x,\epsilon}} \sim \text{Bernoulli} \left( \frac{\mu(B_{x,\epsilon})}{\mu(B_{X,\epsilon})} \right).$$

(80)

Then,

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

(81)
and, to investigate the concentration of $N_{x,\epsilon}$ about its expectation, we write that
\[
N_{x,\epsilon} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} \cdot N_{X,\epsilon} = N_{x,\epsilon} - \mathbb{E}_{Y_{x,\epsilon}|X}[N_{x,\epsilon}] \quad \text{(see (81))}
\]
\[
= \sum_{y \in Y_{X,\epsilon}} (1_{y \in B_{x,\epsilon}} - \mathbb{E}_{Y_{x,\epsilon}|X}[1_{y \in B_{x,\epsilon}}])
\]
\[
= \sum_{y \in Y_{X,\epsilon}} \left(1_{y \in B_{x,\epsilon}} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}\right)
\]
\[
= \sum_{y \in Y_{X,\epsilon}} a_y,
\]
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 (82), we write that
\[
b = \max_y |a_y|
\]
\[
= \max_y \left|1_{y \in B_{x,\epsilon}} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}\right| \quad \text{(see (82))}
\]
\[
\leq 1,
\]
\[
\sigma^2 = \sum_{y \in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|X}[a_y^2]
\]
\[
= \sum_{y \in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|X}\left[\left(1_{y \in B_{x,\epsilon}} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}\right)^2\right] \quad \text{(see (82))}
\]
\[
= \sum_{y \in Y_{X,\epsilon}} \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} \left(1 - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}\right) \quad \text{(see (80))}
\]
\[
\leq \sum_{y \in Y_{X,\epsilon}} \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} = \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} \cdot N_{X,\epsilon}, \quad \text{if } N_{X,\epsilon} = N_{x,\epsilon}
\]
\[
\max[b, \sigma] = \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} N_{X,\epsilon}, \quad \text{if } N_{X,\epsilon} \geq \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}
\]
(85)

From Proposition 2, then, it follows that
\[
\left|N_{x,\epsilon} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} N_{X,\epsilon}\right| \leq \gamma_5 \cdot \max[b, \sigma]
\]
\[
= \gamma_5 \cdot \sqrt{\frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}} N_{X,\epsilon}, \quad \text{(see (85))}
\]
(86)

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
\[
\max_{x \in X} \left|N_{x,\epsilon} - \frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)} N_{X,\epsilon}\right| \leq \gamma_3 \log N \cdot \sqrt{\frac{\mu(B_{x,\epsilon})}{\mu(B_X,\epsilon)}} N_{X,\epsilon},
\]
(87)
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 (85) hold
then (87) readily yields that

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

(88)

then (87) readily yields that

$$\frac{1}{2} \cdot \frac{\mu(\mathbb{B}_{X,\epsilon})}{\mu(\mathbb{B}_{x,\epsilon})} N_{X,\epsilon} \leq N_{x,\epsilon} \leq \frac{3}{2} \cdot \frac{\mu(\mathbb{B}_{X,\epsilon})}{\mu(\mathbb{B}_{x,\epsilon})} N_{X,\epsilon}, \quad \forall x \in X,$$

(89)

except with a probability of at most $N^{1^{-\eta}}$. This completes the proof of Lemma 5.2.

Appendix G. Proof of Lemma 5.3. Throughout, $X$ and $\epsilon \in (0, \epsilon_{\mu,X}]$ are fixed, and we further assume that the event $\mathcal{E}_2$ holds (see (38)). For now, suppose in addition that the neighborhood structure $\mathcal{N}_{X,\epsilon} := \{N_{x,\epsilon}\}_{x \in X}$ is fixed too. Recalling the definition of $\nabla Y_{x,\epsilon} f(\cdot)$ from (30), we first set

$$\mathbb{R}^{n \times n} \ni \sum_{x \in X} \nabla Y_{x,\epsilon} f(x) \nabla Y_{x,\epsilon} f(x)^*,$$

(90)

for short, and then separate the “diagonal” and “off-diagonal” components of the expectation of $\sum_{x \in X} \nabla Y_{x,\epsilon}$, as follows:

$$\mathbb{E}_{Y_{x,\epsilon}|N_{X,\epsilon},X} \left[ \sum_{x \in X} \nabla Y_{x,\epsilon} f(x) \cdot \nabla Y_{x,\epsilon} f(x)^* \right]$$

(see (90))

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

(see (30))

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

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

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

$$+ \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}} \sum_{y \neq y' \in Y_{x,\epsilon}} \mathbb{E}_{y,y'|x} \left[ 1_{y \neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \right] \quad (y, y' \sim \mu_{x,\epsilon})$$

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

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

(91)

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,
To replace the off-diagonal term in the last line of (91), first recall the inequality
\[ \| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,e}} \sum_{y \neq y'} \mathbb{E}_{y,x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \]
\[ - \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,e}} \nabla f(x) \nabla f(x)^* \]
\[ \leq \frac{n^2}{\min_{x \in X} N_{x,e}} \cdot \sup_{x \in \mathbb{D}} \left\| \mathbb{E}_{y,x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \right\|_F \]
\[ - \left( \frac{2 \nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\| \nabla f(x) \|^2}{n(n+2)} \cdot I_n \right) \]
\[ \leq \frac{B''_{\mu,e}}{N_{X,\min,e}}. \quad \text{(see (38))} \] (92)

To replace the off-diagonal term in the last line of (91), first recall the inequality
\[ \|ab^* - cd^*\|_F \]
\[ \leq 2 \max \{ ||a - c||_2, ||b - d||_2 \} \cdot \max \{ ||b||_2, ||c||_2 \}, \quad a, b, c, d \in \mathbb{R}^n, \] (93)

and then note that
\[ \left\| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,e}} \sum_{y \neq y'} \mathbb{E}_{y,x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y',x} [\nabla f(x)^* P_{x,y}] \right\|_F \]
\[ - \frac{1}{N} \sum_{x \in X} \frac{N_{x,e} - 1}{N_{x,e}} \nabla f(x) \nabla f(x)^* \]
\[ = \left\| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,e}} \sum_{y \neq y'} 1_{y \neq y'} \cdot \left( \mathbb{E}_{y,x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y',x} [\nabla f(x)^* P_{x,y}] - \frac{\nabla f(x) \nabla f(x)^*}{n^2} \right) \right\|_F \] (\#Y_{x,e} = N_{x,e})
\[ \leq n^2 \max_{x \in X} \max_{y, y' \in Y_{x,e}} \left\| \mathbb{E}_{y,x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y',x} [\nabla f(x)^* P_{x,y}] \right\|_F \]
\[ - \frac{\nabla f(x) \nabla f(x)^*}{n^2} \] (\#X = N, \#Y_{X,e} = N_{x,e})
\[ \leq 2n^2 \max_{x \in X} \left[ \left\| \mathbb{E}_{y,x} [P_{x,y} \nabla f(x)] \right\|_2 \cdot \frac{\| \nabla f(x) \|_2}{n} \right] \]
\[ \cdot \max \left[ \left\| \mathbb{E}_{y,x} [P_{x,y} \nabla f(x)] \right\|_2, \frac{\| \nabla f(x) \|_2}{n} \right] \] (see (93))
\[ \leq 2n^2 \max_{x \in X} \left[ \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] - \frac{\nabla f(x)}{n} \right\|_2 \cdot \left( \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] - \frac{\nabla f(x)}{n} \right\|_2 + \frac{\| \nabla f(x) \|_2}{n} \right) \right] \] (triangle ineq.)

\[ \leq 2n^2 \left( \frac{B'_{\mu,\epsilon}}{n} \cdot L_f \right) \left( \frac{B'_{\mu,\epsilon}}{n} \cdot L_f + \frac{L_f}{n} \right) \quad \text{(see (9)} \text{ and (70))} \]

\[ = 2 B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 . \quad \text{(94)} \]

We may now replace the diagonal and off diagonal components in the last line of (91) with simpler expressions while incurring a typically small error. More specifically, in light of (92) and (94), (91) now implies that

\[ \left\| \mathbb{E}_{Y|x} |N_{x,e}, X [\bar{X}^T X, Y, X, \epsilon] - \frac{1}{N} \sum_{x \in X} \left( 1 + \frac{n-2}{N_{x,\epsilon} (n+2)} \right) \nabla f(x) \nabla f(x)^* \right\|_F \]

\[ - \frac{n}{N_{x,\epsilon} (n+2)} \sum_{x \in X} \| \nabla f(x) \|_2^2 \cdot I_n \cdot \right\|_F \]

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

\[ \leq \frac{B''_{\mu,\epsilon}}{N_{x,\epsilon, \min}} + 2 B'_\mu,\epsilon (B'_{\mu,\epsilon} + 1) L_f^2 . \quad \text{(see (92) and (94))} \quad \text{(95)} \]

We can further simplify the first line of (95) by replacing \( N_{x,\epsilon} \) with \( N_{X,\epsilon, \min} \) as follows. By invoking (12) in the second line below, we note that

\[ \left\| \mathbb{E}_{Y|x} |N_{x,e}, X [\bar{X}^T X, Y, X, \epsilon] - \left( 1 + \frac{n-2}{N_{X,\epsilon, \min} (n+2)} \right) \bar{X} \right\|_F \]

\[ - \frac{n}{N_{X,\epsilon, \min} (n+2)} \cdot \text{trace} \left[ \bar{X} \right] \cdot I_n \right\|_F \]

\[ = \left\| \mathbb{E}_{Y|x} |N_{x,e}, X [\bar{X}^T X, Y, X, \epsilon] - \left( 1 + \frac{n-2}{N_{X,\epsilon, \min} (n+2)} \right) \frac{1}{N} \sum_{x \in X} \nabla f(x) \nabla f(x)^* \right\|_F \]

\[ - \frac{n}{N_{X,\epsilon, \min} (n+2)} \cdot \frac{1}{N} \sum_{x \in X} \| \nabla f(x) \|_2^2 \cdot I_n \right\|_F \]

\[ \leq \left( \frac{B''_{\mu,\epsilon}}{N_{X,\epsilon, \min}} + 2 B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 \right) \]

\[ + \max_{x \in X} \left[ \frac{1}{N_{x,\epsilon}} - \frac{1}{N_{X,\epsilon, \min}} \right] \cdot \max_{x \in X} \| \nabla f(x) \|_2^2 \cdot (1 + \| I_n \|_F) \quad \text{(see (95))} \]

\[ \leq \left( \frac{B''_{\mu,\epsilon}}{N_{X,\epsilon, \min}} + 2 B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 \right) \]
\[
+ \max_{x \in X} \frac{1}{N_{x,e}} - \frac{1}{N_{X,\min,e}} \cdot L^2_f (1 + \sqrt{n}) \quad \text{(see (9))}
\]
\[
\leq \left( \frac{B''_{\mu,e}}{N_{X,\min,e}} + 2B'_{\mu,e} (B'_{\mu,e} + 1) L^2_f \right) + \frac{L^2_f (1 + \sqrt{n})}{N_{X,\min,e}} \quad \text{(see (38))}
\]
\[
= : \frac{1}{2} B_{\mu,e}. \quad (96)
\]

Next, we replace \( \text{trace}[\dot{\Sigma}_X] \) in the first line of (96) with \( \text{trace}[\ddot{\Sigma}_{X,Y_{X,e}}] \). To that end, we first notice the following consequence of (96):

\[
\left| \text{trace} \left[ \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \ddot{\Sigma}_{X,Y_{X,e}} ] \right] - \left( 1 + \frac{n - 2}{N_{X,\min,e}(n + 2)} \right) \text{trace} [\dot{\Sigma}_X] \right|
\]
\[
= \left| \text{trace} \left[ \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \ddot{\Sigma}_{X,Y_{X,e}} ] - \left( 1 + \frac{n - 2}{N_{X,\min,e}(n + 2)} \right) \dot{\Sigma}_X \right] - \frac{n}{N_{X,\min,e}(n + 2)} \cdot \text{trace} [\dot{\Sigma}_X] \cdot I_n \right|
\]
\[
\leq \sqrt{n} \left\| \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \ddot{\Sigma}_{X,Y_{X,e}} ] - \left( 1 + \frac{n - 2}{N_{X,\min,e}(n + 2)} \right) \dot{\Sigma}_X \right\|_F
\]
\[
\leq \frac{\sqrt{n}}{2} B_{\mu,e}, \quad \text{(see (96))} \quad (97)
\]

where the second line uses the fact that \( \text{trace}[I_n] = n \). Also, the third line follows from the inequality \( \text{trace}[A] \leq \sqrt{n} \| A \|_F \) for an arbitrary matrix \( A \in \mathbb{R}^{n \times n} \). After rearranging, (97) immediately implies that

\[
\left| \left( 1 + \frac{n^2 + n - 2}{N_{X,\min,e}(n + 2)} \right)^{-1} \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \text{trace} [\ddot{\Sigma}_{X,Y_{X,e}}] ] - \text{trace} [\dot{\Sigma}_X] \right|
\]
\[
\leq \left( 1 + \frac{n^2 + n - 2}{N_{X,\min,e}(n + 2)} \right)^{-1} \frac{\sqrt{n}}{2} B_{\mu,e}. \quad (98)
\]

The above inequality enables us to remove \( \text{trace}[\dot{\Sigma}_X] \) from the first line of (96):

\[
\left\| \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \ddot{\Sigma}_{X,Y_{X,e}} ] - \left( 1 + \frac{n - 2}{N_{X,\min,e}(n + 2)} \right) \dot{\Sigma}_X \right. \]
\[
- \frac{n}{N_{X,\min,e}(n + 2)} \cdot \left( 1 + \frac{n^2 + n - 2}{N_{X,\min,e}(n + 2)} \right)^{-1} \mathbb{E}_{Y_{X,e} \mid N_{X,e},X} [ \text{trace} [\ddot{\Sigma}_{X,Y_{X,e}}] ] \cdot I_n \right\|_F
\]
\[
\leq \frac{1}{2} B_{\mu,e} + \frac{n}{N_{X,\min,e}(n + 2)} \left( 1 + \frac{n^2 + n - 2}{N_{X,\min,e}(n + 2)} \right)^{-1} \frac{\sqrt{n}}{2} B_{\mu,e} \cdot \| I_n \|_F \quad (96,98)
\]
Lastly, (99) can be rewritten as follows by introducing $\tilde{\Sigma}_{X,Y_{x_{*}}}$ \in \mathbb{R}^{n \times n}:

\[
\left\| \mathbb{E}_{Y_{x_{*}},|E_{2}} [\tilde{\Sigma}_{X,Y_{x_{*}}}] - \tilde{\Sigma}_{X} \right\|_{F} \\
\leq \frac{1}{2} \left( 1 + \frac{n-2}{N_{X_{min},\epsilon}(n+2)} \right)^{-1} \\
\cdot \left( 1 + \frac{n^2}{N_{X_{min},\epsilon}(n+2)} \right)^{-1} \cdot \text{trace} \left[ \tilde{\Sigma}_{X,Y_{x_{*}}} \right] \cdot I_n \\
\leq B_{\mu,\epsilon}, \quad \text{(the factor in front of } B_{\mu,\epsilon} \text{ does not exceed 1)}
\]

(100)

where, above, we set

\[
\tilde{\Sigma}_{X,Y_{x_{*}}}
\]

:= \left( 1 + \frac{n-2}{N_{X_{min},\epsilon}(n+2)} \right)^{-1} \\
\cdot \left( \tilde{\Sigma}_{X,Y_{x_{*}}} - \frac{n}{N_{X_{min},\epsilon}(n+2)} \right)^{-1} \\
\cdot \left( \tilde{\Sigma}_{X,Y_{x_{*}}} - \left( 1 + \frac{2}{n} \right) N_{X_{min},\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \cdot \text{trace} \left[ \tilde{\Sigma}_{X,Y_{x_{*}}} \right] \cdot I_n \\

= \left( 1 + \frac{1-\frac{2}{n}}{1+\frac{2}{n}} \cdot N_{X_{min},\epsilon}^{-1} \right)^{-1} \\
\cdot \left( \frac{1}{N} \sum_{x \in X} \tilde{\nabla}_{Y_{x_{*}},f(x)} \tilde{\nabla}_{Y_{x_{*}},f(x)}^{*} \\
- \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\| \tilde{\nabla}_{Y_{x_{*}},f(x)} \right\|_{2}^{2} \cdot I_n \\

= \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_{*}} \geq N_{X_{min},\epsilon}} \tilde{\nabla}_{Y_{x_{*}},f(x)} \tilde{\nabla}_{Y_{x_{*}},f(x)}^{*} \\
- \left( 1 + \frac{2}{n} \right) N_{X_{min},\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \sum_{N_{x_{*}} \geq N_{X_{min},\epsilon}} \left\| \tilde{\nabla}_{Y_{x_{*}},f(x)} \right\|_{2}^{2} \cdot I_n,
\]

(101)

where the third identity uses (90) and the last line above follows from (38). Because $B_{\mu,\epsilon}$ does not depend on $N_{X_{*}}$, it is easy to remove the conditioning on $N_{X_{*}}$ in (100):

\[
\left\| \mathbb{E}_{Y_{x_{*}},|E_{2}} [\tilde{\Sigma}_{X,Y_{x_{*}}}] - \tilde{\Sigma}_{X} \right\|_{F} \\
= \left\| \mathbb{E} \left[ \mathbb{E}_{Y_{x_{*}},|E_{2}} [\tilde{\Sigma}_{X,Y_{x_{*}}}] - \tilde{\Sigma}_{X} \right] \right\|_{F}
\]
Finally, to the fourth line below, we write that

\[ \mathbf{B}_{\mu,\epsilon} \quad \text{(see (100))} \]

\[ = B_{\mu,\epsilon} \quad \text{(see (96))} \quad \text{(102)} \]

Consider also the following special case. Let \( \mu \) be the uniform probability measure on \( \mathbb{D} \) and fix \( x \) within the \( \epsilon \)-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 (18)). Then, as stated in Proposition 3, \( B'_{\mu,\epsilon} = 0 \). Furthermore, it is known [3] that

\[ P_{x,y} \cdot \nabla f(x) \overset{\text{dist}}{=} \mathbf{\omega} \cdot \nabla f(x) + \sqrt{\mathbf{\omega} \cdot \nabla f(x)^2} \cdot \alpha, \quad \text{(103)} \]

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

\[ \omega \sim \text{beta} \left( \frac{1}{2}, \frac{n-1}{2} \right), \quad \alpha \sim \text{uniform} \left( \mathbb{S}^{n-2} \right), \quad \omega \perp \alpha. \]

Finally, \( A \in \mathbb{R}^{n \times (n-1)} \) in (103) is an orthonormal basis for the directions orthogonal to \( \nabla f(x) \in \mathbb{R}^n \), namely

\[ A^* \nabla f(x) = 0, \quad A^* A = I_{n-1}. \quad \text{(104)} \]

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

\[ E_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \]

\[ = E \left[ (\mathbf{\omega} \nabla f(x) + \sqrt{\mathbf{\omega} \cdot \nabla f(x)^2}) \cdot A \alpha \right] \cdot \left( \mathbf{\omega} \nabla f(x) + \sqrt{\mathbf{\omega} \cdot \nabla f(x)^2} \cdot A \alpha \right)^* \quad \text{(see (103))} \]

\[ = E \left[ \mathbf{\omega}^2 \cdot \nabla f(x) \nabla f(x)^* \right. \]

\[ + E \left[ \mathbf{\omega} \cdot \nabla f(x) \right] \cdot \mathbb{E} [\alpha \alpha^*] \cdot A^* \quad (\mathbf{\omega} \perp \alpha, \quad \mathbb{E} \alpha = 0) \]

\[ = \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* \]

\[ + \frac{n-1}{n(n+2)} \cdot \| \nabla f(x) \|^2 \cdot A \cdot \frac{I_{n-1}}{n-1} \cdot A^* \quad \left( \mathbb{E} [\alpha \alpha^*] = \frac{I_{n-1}}{n-1} \right) \]

\[ = \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* \]

\[ + \frac{1}{n(n+2)} \cdot \| \nabla f(x) \|^2 \cdot \left( \| \nabla f(x) \|^2 \cdot \| \nabla f(x) \|^2 + A A^* \right) \]

\[ = \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* \]

\[ + \frac{1}{n(n+2)} \cdot \| \nabla f(x) \|^2 \cdot \left( \| \nabla f(x) \|^2 \cdot \| \nabla f(x) \|^2 + A A^* \right) \]

\[ = \frac{2}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \| \nabla f(x) \|^2 \cdot I_n, \quad \text{(see (104))} \quad \text{(105)} \]

and, consequently, \( B''_{\mu,\epsilon} = 0 \). Furthermore, assume that \( N_{x,\epsilon} = N_{x',\epsilon} \) for every pair \( x, x' \in X \). Then, we observe that the upper bound in (96) can be improved to \( B'_{\mu,\epsilon} = 0 \), namely \( \hat{\Sigma}_{X,Y_{x,\epsilon}} \) is an unbiased estimator of \( \hat{\Sigma}_X \), conditioned on the event \( \mathcal{E}_2 \). This completes the proof of Lemma 5.3.
Appendix H. Proof of Lemma 5.4. Throughout, $X$ is fixed and we assume that the event $E_2$ holds (see (38)). We also consider $\mathcal{N}_{X,\epsilon} = \{N_x,\epsilon\}_{x\in X}$ (see (17)) to be any fixed neighborhood structure consistent with $E_2$.

To bound the estimation error, we write that
\[
\left\| \tilde{\Sigma}^{X,Y,e} - \hat{\Sigma}^X \right\|_F \\
\leq \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F + \left\| \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] - \Sigma^X \right\|_F \\
\leq \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F + B_{p,e} \cdot \text{(see (100))} \\
\] (106)

It therefore suffices to study the concentration of $\tilde{\Sigma}^{X,Y,e}$ about its expectation. In fact, as we show next, it is more convenient to first study the concentration of $\tilde{\Sigma}^{X,Y,e} \in \mathbb{R}^{n\times n}$ instead, where
\[
\tilde{\Sigma}^{X,Y,e} := \frac{1}{N} \sum_{x \in X} \hat{\nabla}_{Y,e} f(x) \cdot \hat{\nabla}_{Y,e} f(x)^*, \\
\hat{\nabla}_{Y,e} f(x) := \frac{n}{N_x,\epsilon} \sum_{y \in Y,e} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n, \quad \forall x \in X.
\]

Indeed, conditioned on $E_2$, $\mathcal{N}_{X,e}$, $X$, the expression for $\tilde{\Sigma}^{X,Y,e}$ in (29) simplifies to
\[
\tilde{\Sigma}^{X,Y,e} = \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,e}^{-1} \right) \\
\cdot \left( \tilde{\Sigma}^{X,Y,e} - \frac{\text{trace}\left[ \tilde{\Sigma}^{X,Y,e} \right]}{\left( 1 + \frac{2}{n} \right) N_{X,\min,e} + n - \frac{2}{n} \cdot I_n} \cdot I_n \right). \\
\] (108)

Consequently, the deviation of $\tilde{\Sigma}^{X,Y,e}$ about its expectation can be bounded as:
\[
\left\| \tilde{\Sigma}^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \\
\leq \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \\
+ \left( \left( 1 + \frac{2}{n} \right) N_{X,\min,e} + n - \frac{2}{n} \right)^{-1} \\
\cdot \text{trace}\left[ \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right] \cdot \|I_n\|_F \\
\leq \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \\
+ \left( \left( 1 + \frac{2}{n} \right) N_{X,\min,e} + n - \frac{2}{n} \right)^{-1} \\
\cdot \sqrt{n} \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \cdot \sqrt{n} \quad (\|I_n\|_F = \sqrt{n}) \\
= \left( 1 + \frac{2}{n} \right) N_{X,\min,e} + n - \frac{2}{n} \right) \\
\left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \\
\leq 2 \left\| \Sigma^{X,Y,e} - \mathbb{E}_{X,e}[\Sigma^{X,Y,e}] \left[ \Sigma^{X,Y,e} \right] \right\|_F \cdot \text{(factor above \leq 2)} \\
\] (109)

Above, the first inequality uses (108). We also used the linearity of trace and the inequality $\left|\text{trace}[A]\right| \leq \sqrt{n} \|A\|_F$ for arbitrary $A \in \mathbb{R}^{n\times n}$. Thanks to (109), it
suffices to study the concentration of $\tilde{\Sigma}_{X,Y,X_\epsilon}$ about its expectation. The following result is proved in Appendix I.

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

$$
\left\| \tilde{\Sigma}_{X,Y,X_\epsilon} - \mathbb{E}_{Y,X_\epsilon}|\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X \left[ \tilde{\Sigma}_{X,Y,X_\epsilon} \right] \right\|_F \lesssim \gamma_{12}^2 \log^4(N_{x,\epsilon}) \cdot \frac{n \sqrt{\log N}}{\sqrt{\rho_{\mu,X_\epsilon}N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] L_f^2 \\
+ 4n^2 L_f^2 N_{X,\epsilon}^{2(1 - \gamma_{12} \log(N_{x,\epsilon}))},
$$

(110)

for $\gamma_1 \geq 1$ and $\gamma_2 \geq 3$, except with a probability of at most

$$
e^{-\gamma_1} + n^2 \log \gamma_1 + N_{X,\epsilon}^{(1 - \gamma_1 \log(N_{x,\epsilon}))},
$$

(111)

Combining (106), (109), and Lemma H.1 tells us that if $\log(n) \geq 1$, $N \geq \log(n)$, and $\log(N_{x,\epsilon}) \geq \log(n)$, then conditioned on $\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X$,

$$
\left\| \tilde{\Sigma}_{X,Y,X_\epsilon} - \tilde{\Sigma}_X \right\|_F \lesssim B_{\mu_\epsilon} + \gamma_{12}^2 \log^4(N_{x,\epsilon}) \cdot \frac{n \sqrt{\log N}}{\sqrt{\rho_{\mu,X_\epsilon}N_{X,\epsilon}}} \cdot \max[K_{\mu}^{-1}, K_{\mu}^{-2}] L_f^2 \\
+ 4n^2 L_f^2 N_{X,\epsilon}^{2(1 - \gamma_{12} \log(N_{x,\epsilon}))},
$$

except with the probability appearing in (111). We observe that this expression and probability do not depend on $\mathcal{N}_{X,\epsilon}$, 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 5.4.

**Appendix I. Proof of Lemma H.1.** Throughout, $X$ is fixed and the event $\mathcal{E}_2$ holds. We will also use $\mathcal{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$ to summarize the neighborhood structure of data (see (17)). As in Appendix G, we again decompose $\tilde{\Sigma}_{X,Y,X_\epsilon}$ into “diagonal” and “off-diagonal” components:

$$
\tilde{\Sigma}_{X,Y,X_\epsilon} = \frac{1}{N} \sum_{x \in X} \tilde{\Sigma}_{Y_\epsilon,x}(x) \cdot \tilde{\Sigma}_{Y_\epsilon,x}(f(x)^* \text{ (see (107))})
$$

$$
= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y' \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \text{ (see (30))}
$$

$$
= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}
$$

$$
+ \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \sum_{y' \in Y_{x,\epsilon}} 1_{y \neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'}
$$

$$
=: \tilde{\Sigma}^{d}_{X,Y,X_\epsilon} + \tilde{\Sigma}^{o}_{X,Y,X_\epsilon}.
$$

(112)

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

$$
\left\| \tilde{\Sigma}_{X,Y,X_\epsilon} - \mathbb{E}_{Y,X_\epsilon}|\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X \left[ \tilde{\Sigma}_{X,Y,X_\epsilon} \right] \right\|_F \leq \left\| \tilde{\Sigma}^{d}_{X,Y,X_\epsilon} - \mathbb{E}_{Y,X_\epsilon}|\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X \left[ \tilde{\Sigma}^{d}_{X,Y,X_\epsilon} \right] \right\|_F \\
+ \left\| \tilde{\Sigma}^{o}_{X,Y,X_\epsilon} - \mathbb{E}_{Y,X_\epsilon}|\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X \left[ \tilde{\Sigma}^{o}_{X,Y,X_\epsilon} \right] \right\|_F.
$$

(113)
We bound the norms on the right-hand side above separately in Appendices J and K, respectively, and report the results below.

**Lemma I.1.** Fix $X$ and $\epsilon \in (0, \epsilon_{\mu,X}]$. Consider the event

$$\mathcal{E}_1 := \left\{ \max_{x \in X} \max_{y \in Y_{X, \epsilon}} \| P_{x,y} \nabla f(x) \|^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\}, \quad (114)$$

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

$$\left\| \sum_{x \in X, y \in Y_{X, \epsilon}, i \in [1:n]} \nabla f_i \right\|_F \leq \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} + 2n^2 \frac{L_f^2}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}} \Pr_{X, \epsilon} \left[ \mathcal{E}_1^C \right], \quad (115)$$

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

**Lemma I.2.** Fix $X$ and $\epsilon \in (0, \epsilon_{\mu,X}]$. Let $\tilde{Y}_{X,\epsilon}$ contain $Y_{X,\epsilon}$ and three independent copies of it. That is, $\tilde{Y}_{X,\epsilon} = \cup_{x \in X} \tilde{Y}_{x,\epsilon}$, where each $\tilde{Y}_{x,\epsilon}$ contains $Y_{x,\epsilon}$ and three independent copies of it. Consider the event $\mathcal{E}_3$ defined in (114) for $Q_{X,\epsilon} > K_{\mu}^{-1}$ to be set later. Consider also the event

$$\mathcal{E}_3 := \left\{ \max_{x \in X} \max_{y \in \tilde{Y}_{x,\epsilon}, i \in [1:n]} \| P_{x,y} e_i \|^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\} \cap \left\{ \max_{x \in X} \| P_{x,y} \nabla f(x) \|^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\}, \quad (116)$$

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

$$\Pr_{\tilde{Y}_{X,\epsilon} \mid \mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X} \left[ \mathcal{E}_3^C \right] \leq \left( \frac{\log n}{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}} \right)^{\frac{\log n}{n}}, \quad (117)$$

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

$$\left\| \sum_{x \in X, y \in \tilde{Y}_{x,\epsilon}, i \in [1:n]} \nabla f_i \right\|_F \leq \gamma_7 \sqrt{\log n} \cdot \frac{n \cdot \max \left[ Q_{X,\epsilon}, Q_{X,\epsilon}^2 \right] \cdot L_f^2}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} + 2n^2 \frac{L_f^2}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}} \Pr_{Y_{X,\epsilon} \mid \mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X} \left[ \mathcal{E}_3^C \right], \quad (118)$$

for $\gamma_7 \geq 1$ and except with a probability of at most

$$e^{-\gamma_7} + n^2 \cdot n^{-\log \gamma_7} + \Pr_{Y_{X,\epsilon} \mid \mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X} \left[ \mathcal{E}_3^C \right].$$

Before we can apply Lemmas I.1 and I.2 to the right-hand side of (113), 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 I.3.** Fix $\epsilon \in (0, \epsilon_{\mu}]$ and $X$. Suppose that $Q_{X,\epsilon} = \gamma_2 K_{\mu}^{-1} \log^2 (N_{X,\epsilon})$ for $\gamma_2 \geq 3$. Then, conditioned on $\mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X$, it holds that

$$\Pr_{Y_{X,\epsilon} \mid \mathcal{E}_2, \mathcal{N}_{X,\epsilon}, X} \left[ \mathcal{E}_3^C \right] \leq N_{X,\epsilon}^{1 - \gamma_2 \log (N_{X,\epsilon})}. \quad (119)$$
Moreover, if $1 \leq \log(n) \leq \log(N_{X,\epsilon})$ and if $N_{X,\epsilon}$ is large enough such that
\[ \Pr_{Y_{X,\epsilon}|N_{X,\epsilon},E_2,X} [\mathcal{C}^c] \leq \frac{1}{2}, \]
then conditioned on $\mathcal{E}_1, \mathcal{E}_2, N_{X,\epsilon}, X$, the requirement in (117) is satisfied.

Revisiting (113), 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 $\mathcal{E}_2, N_{X,\epsilon}, X$,
\[ \left\| \sum_{x,y,X_{\epsilon}} - E_{Y_{X,\epsilon}|E_2,N_{X,\epsilon},X} \left[ \sum_{x,y} \right] \right\|_F \]
\[ \leq \left\| \sum_{x,y,X_{\epsilon}} - E_{Y_{X,\epsilon}|E_2,N_{X,\epsilon},X} \left[ \sum_{x,y} \right] \right\|_F + \left\| \sum_{x,y,X_{\epsilon}} - E_{Y_{X,\epsilon}|E_2,N_{X,\epsilon},X} \left[ \sum_{x,y} \right] \right\|_F \]
\[ \leq \gamma_6 \cdot \frac{Q_{X,\epsilon}L_f^2 n}{\sqrt{\rho_{\mu,X_{\epsilon}N_{X,\epsilon}}} + \gamma_7 \cdot \sqrt{\log n} \cdot n \cdot \max[Q_{X,\epsilon}, Q_{X,\epsilon}^2] \cdot L_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} \]
\[ \leq \gamma_7 \gamma_2^2 \log^2(N_{X,\epsilon}) \cdot \frac{n \log n}{\sqrt{\rho_{\mu,X_{\epsilon}N_{X,\epsilon}}} \cdot \max[Q_{X,\epsilon}, Q_{X,\epsilon}^2] \cdot L_f^2 + 4n^2 L_f^2 N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))} \]
\[ \leq e^{-\gamma_6} + e^{-\gamma_7} + n^2 \cdot n^{-\log \gamma_7 + 2} \cdot \Pr_{Y_{X,\epsilon}|E_2,N_{X,\epsilon},X} [\mathcal{C}^c] \]
\[ \leq e^{-\gamma_6} + e^{-\gamma_7} + n^{2-\log \gamma_7 + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}} \]
\[ \leq e^{-\gamma_7} + n^{2-\log \gamma_7 + N_{X,\epsilon}^{(1-\gamma_2 \log(N_{X,\epsilon}))}}. \]

This completes the proof of Lemma H.1.

**Appendix J. Proof of Lemma I.1.** Throughout, $X$ and the neighborhood structure $N_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$ (see (17)) are fixed. Moreover, we assume that the event $\mathcal{E}_2$ holds (see (38)). In addition, for $Q_{X,\epsilon} \geq K_{\mu}^{-1}$ to be set later, we condition on the following event:
\[ \mathcal{E}_1 := \left\{ \max_{x \in X} \max_{y \in Y_{X,\epsilon}} \|P_{x,y} \nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon}L_f^2}{n} \right\}. \]

By the definition of $\sum_{x,y,X_{\epsilon}}$ in (112), we observe that
\[ \left\| \sum_{x,y,X_{\epsilon}} - E_{Y_{X,\epsilon}|N_{X,\epsilon},E_1,E_2,X} \left[ \sum_{x,y} \right] \right\|_F \]
On the other hand, note that

\[ \frac{n^2}{N} \left| \sum_{x \in X} \sum_{y \in Y_{x, \epsilon}} \frac{1}{N_{x, \epsilon}} \left( P_{x, y} \nabla f(x) \nabla f(x)^* P_{x, y} - \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \right) \right|_F, \] (see (112))

\[ =: \frac{n^2}{N} \left\| \sum_{x \in X} \sum_{y \in Y_{x, \epsilon}} A_{x, y} \right\|_F. \] (123)

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 \( \delta \) and \( \sigma \) parameters below. For arbitrary \( x \in X \) and \( y \in Y_{x, \epsilon} \), note that

\[
\|A_{x, y}\|_F \\
= \frac{1}{N_{x, \epsilon}} \left\| P_{x, y} \nabla f(x) \nabla f(x)^* P_{x, y} - \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \right\|_F \quad \text{(see (123))}
\leq \frac{1}{N_{x, \epsilon}} \left\| P_{x, y} \nabla f(x) \nabla f(x)^* P_{x, y} \right\|_F \\
+ \frac{1}{N_{x, \epsilon}} \cdot \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \|_F \quad \text{(Jensen’s inequality)}
\leq \frac{1}{N_{x, \epsilon}} \left\| P_{x, y} \nabla f(x) \right\|_2^2 + \frac{1}{N_{x, \epsilon}} \cdot \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^2 \|_2 \quad \text{(see (122))}
\leq \frac{1}{N_{x, \epsilon}} \min_{x \in X} N_{x, \epsilon} \cdot \max_{x \in X} \|P_{x, y} \nabla f(x)\|_2^2 \\
+ \frac{1}{N_{x, \epsilon}} \cdot \max_{x \in X} \|\nabla f(x)\|_2 \quad \text{(see [58, Lemma 5.5])}
\leq \frac{1}{N_{x, \epsilon}} \min_{x \in X} N_{x, \epsilon} \cdot \max_{x \in X} \|P_{x, y} \nabla f(x)\|_2^2 + \frac{1}{N_{x, \epsilon}} \cdot \frac{L_f^2}{K_n} \quad \text{(see (9))}
=: \frac{1}{N_{x, \epsilon}} \min_{x \in X} N_{x, \epsilon} \cdot \frac{Q_X L_f^2}{n} + \frac{1}{N_{x, \epsilon}} \cdot \frac{L_f^2}{K_n} \quad \text{(see (122))}
\leq \frac{1}{N_{x, \epsilon}} \min_{x \in X} N_{x, \epsilon} \cdot \frac{Q_X L_f^2}{n} \quad \text{(when } Q_X \geq K_n^{-1} \text{)}
\] (124)

On the other hand, note that

\[
\sum_{x \in X} \sum_{y \in Y_{x, \epsilon}} \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \\
= \sum_{x \in X} \sum_{y \in Y_{x, \epsilon}} \frac{1}{N_{x, \epsilon}} \cdot \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \\
\leq \sum_{x \in X} \sum_{y \in Y_{x, \epsilon}} \frac{1}{N_{x, \epsilon}} \cdot \mathbb{E}_{y \mid \mathbb{N}_{x, \epsilon}} \nabla f(x) \nabla f(x)^* P_{x, y} \|_F^2
\]
\[
\sum_{x \in X} \sum_{y \in Y_{x,e}} \frac{1}{N_{x,e}^2} \cdot \mathbb{E}_{y \mid \mathcal{N}_{x,e}, \mathcal{E}_{1,x}, \mathcal{E}_{2,x}} \left\| P_{x,y} \nabla f(x) \right\|_2^4
\leq \sum_{x \in X} \sum_{y \in Y_{x,e}} \frac{1}{N_{x,e}^2} \cdot \mathbb{E}_{y \mid \mathcal{N}_{x,e}, \mathcal{E}_{2,x}} \left\| P_{x,y} \nabla f(x) \right\|_2^4 \quad \text{(see (122))}
\]
\[
\leq \sum_{x \in X} \sum_{y \in Y_{x,e}} \frac{1}{K_{\mu} n^2} \cdot \max_{x \in X} \left\| \nabla f(x) \right\|_2^4 \quad \text{(see [58, Lemma 5.5])}
\]
\[
\leq \frac{N}{\min_{x \in X} N_{x,e}} \cdot \frac{L_j^4}{K_{\mu} n^2} \quad \text{(see (126))}
\]
where the second line uses (123). 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
\[
\max[b, \sigma] \lesssim \sqrt{\frac{N}{\min_{x \in X} N_{x,e}}} \cdot \frac{Q_{X,e} L_j^2}{n}, \quad \text{if } Q_{X,e} \geq K_{\mu}^{-1}.
\]
Thus, in light of Proposition 2, and conditioned on \(\mathcal{E}_1, \mathcal{E}_2, \mathcal{N}_{x,e}, X\), it follows that
\[
\left\| \tilde{\Sigma}^d_{X,Y_{X,e}} - \mathbb{E}_{Y_{X,e} \mid \mathcal{E}_1, \mathcal{N}_{X,e}, \mathcal{E}_2, X} \left[ \tilde{\Sigma}^d_{X,Y_{X,e}} \right] \right\|_F
\]
\[
= \frac{n^2}{N} \left\| \sum_{x \in X} \sum_{y \in Y_{x,e}} A_{x,y} \right\|_F \quad \text{(see (123))}
\]
\[
\lesssim \frac{n^2}{N} \cdot \gamma_6 \cdot \max[b, \sigma] \quad \text{(see (126))}
\]
\[
\lesssim \frac{n^2}{N} \cdot \gamma_6 \frac{N}{\min_{x \in X} N_{x,e}} \cdot \frac{Q_{X,e} L_j^2}{n} \quad \text{(see (126))}
\]
\[
= \gamma_6 \frac{n}{\sqrt{N \cdot \min_{x \in X} N_{x,e}}} \cdot Q_{X,e} L_j^2 \quad \text{(see (38))}
\]
\[
\lesssim \gamma_6 \frac{n}{\sqrt{N \cdot \min_{x \in X} \mu(\mathcal{B}_{x,e})}} \cdot Q_{X,e} L_j^2 \quad \text{(see (22))}
\]
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
\[
\mathbb{E}_{Y_{X,e} \mid \mathcal{N}_{X,e}, \mathcal{E}_2, X} \left[ \tilde{\Sigma}^d_{X,Y_{X,e}} \right]
\]
\[
= \mathbb{E}_{Y_{X,e} \mid \mathcal{E}_1, \mathcal{N}_{X,e}, \mathcal{E}_2, X} \left[ \tilde{\Sigma}^d_{X,Y_{X,e}} \right] \mathbb{P}_{Y_{X,e} \mid \mathcal{E}_2, \mathcal{N}_{X,e}, X} [\mathcal{E}_1]
\]
\[
+ \mathbb{E}_{Y_{X,e} \mid \mathcal{E}_1^c, \mathcal{N}_{X,e}, \mathcal{E}_2, X} \left[ \tilde{\Sigma}^d_{X,Y_{X,e}} \right] \mathbb{P}_{Y_{X,e} \mid \mathcal{E}_2, \mathcal{N}_{X,e}, X} [\mathcal{E}_1^c],
\]

from which it follows that

\[
\mathbb{E}_{Y_{x,e},X} \left[ \sum_{i} x X_{y,e} \right] - \mathbb{E}_{Y_{x,e},[e_{1},X_{e},z_{2},x]} \left[ \sum_{i} x X_{y,e} \right] = \mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right] \left( \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] - \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right). 
\]

Since for any \( X, Y_{x,e} \), we have

\[
\left\| \sum_{i} x X_{y,e} \right\|_{F} \leq \frac{n_{2}}{N} \sum_{x \in X} \sum_{y \in Y_{X_{e},e}} \frac{1}{N_{x,e}^{2}} \left\| P_{x,y} \nabla f(x) \nabla f(x)^{*} P_{x,y} \right\|_{F} \quad \text{(see (112))}
\]

\[
\leq \frac{n_{2}}{N} \sum_{x \in X} \sum_{y \in Y_{X_{e},e}} \frac{1}{N_{x,e}^{2}} \left\| P_{x,y} \nabla f(x) \right\|^{2}_{2}
\]

\[
\leq \frac{n_{2}}{N} \sum_{x \in X} \sum_{y \in Y_{X_{e},e}} \frac{1}{N_{x,e}^{2}} \left\| \nabla f(x) \right\|^{2}_{2}
\]

\[
\leq \frac{n_{2}}{N} \sum_{x \in X} \sum_{y \in Y_{X_{e},e}} \frac{1}{N_{x,e}^{2}} L_{j}^{2} \quad \text{(see (9))}
\]

\[
= \frac{n_{2}}{N} \sum_{x \in X} \frac{1}{N_{x,e}^{2}} L_{j}^{2}
\]

\[
\leq n_{2} L_{j}^{2}, 
\]

we conclude that

\[
\left\| \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] - \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right\|_{F} \leq \mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right] \left( \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] - \mathbb{E}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right) \quad \text{(see (128))}
\]

\[
\leq 2n_{2} L_{j}^{2} \mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right]. \quad \text{(triangle inequality and (129))}
\]

Lastly, we remove the conditioning on the event \( \varepsilon_{1} \) as follows:

\[
\mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \left\| \sum_{i} x X_{y,e} - \mathbb{E}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right\|_{F} \right] \geq \gamma_{6} \cdot \frac{Q_{x,e} L_{j}^{2} n}{\sqrt{\rho_{x,e} N_{e,x}}} + 2n_{2} L_{j}^{2} \mathbb{P}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right] \]

\[
\leq \mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \left\| \sum_{i} x X_{y,e} - \mathbb{E}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right\|_{F} \right] \geq \gamma_{6} \cdot \frac{Q_{x,e} L_{j}^{2} n}{\sqrt{\rho_{x,e} N_{e,x}}} + 2n_{2} L_{j}^{2} \mathbb{P}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right] \]

\[
+ \mathbb{P}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \varepsilon_{1}^{C} \right] \quad \text{(see (54))}
\]

\[
\leq \mathbb{P}_{Y_{x,e},[e_{1},N_{e},X_{e},X]} \left[ \left\| \sum_{i} x X_{y,e} - \mathbb{E}_{Y_{x,e},[e_{2},N_{e},X_{e},X]} \left[ \sum_{i} x X_{y,e} \right] \right\|_{F} \right].
\]
The proof of Lemma I.1 is now complete.

Appendix K. Proof of Lemma I.2. Throughout, $X$ and the neighborhood structure $\mathcal{N}_{X, \epsilon} = \{N_{x, \epsilon}\}_{x \in X}$ (see (17)) are fixed. Moreover, we assume that the event $\mathcal{E}_2$ holds (see (38)). In addition, for $Q_{X, \epsilon} \geq K_0^{-1}$ to be set later, we condition on $\mathcal{E}_1$ as defined in (122).

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 $\epsilon$). Recalling the definition of $\mathcal{E}_{1, x_s}$ from (112), we aim to find an upper bound for

$$\sum_{x_s, y_{sk}} \mathbf{E}_{\mathcal{N}_{x_s, \epsilon}, \mathcal{E}_{1, x_s}} \left[ \mathbf{E}_{y_{sk}, x_s} \mathbf{E}_{y_{sk}, x_s} \mathbf{E}_{x_s, y_{sk}} \right]$$

after which we will remove the conditioning on $\mathcal{E}_1$. Above, $y_{sk} | \mathcal{E}_1, x_s$ 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 (132) by writing it as a chaos random variable. Let us first write that

$$\mathbf{E}_{\mathcal{N}_{x_s, \epsilon}} \left[ \mathbf{E}_{y_{sk}, x_s} \mathbf{E}_{y_{sk}, x_s} \mathbf{E}_{x_s, y_{sk}} \right]$$

The proof of Lemma I.1 is now complete.
\[
\begin{aligned}
&\left. \left( P_{x_s,y_{sk}} - E_{y_s|x_s,z_s}[P_{x_s,y_s}] \right) \nabla f(x_s) \nabla f(x_s)^* \left( P_{x_s,y_{sl}} - E_{y_s|x_s,z_s}[P_{x_s,y_s}] \right) \right|_F \\
= &\frac{n^2}{N} \left| \sum_{s=1}^{N} \sum_{k,l=1}^{N_{x_s,\epsilon}} A_{skl} \right|_F \\
= &\frac{n^2}{N} \sqrt{\sum_{i,j=1}^{n} \sum_{s=1}^{N} \sum_{k,l=1}^{N_{x_s,\epsilon}} A_{skl[i,j]}}. 
\end{aligned}
\]
product of $s$ and $k$). With this new notation, we revisit (133) to write that

$$\begin{align*}
\frac{n^2}{N} \left\| \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N_s^2} \left( P_{x_s,y_{sk}} - E_{x_s}[P_{x_s,y_{sk}}] \right) \nabla f(x_s) \nabla f(x_s)^* \right.
& \cdot \left( P_{x_s,y_{sl}} - E_{x_s}[P_{x_s,y_{sl}}] \right) \left\|_F \frac{n}{2} \sum_{i,j=1}^{\frac{N}{2}} \left| \sum_{t=1}^{N_s} \sum_{k=1}^{N_{x_s,t}} \sum_{l=1}^{N_{x_s,l}} A_{ij}[sk,tl] \right|^2 \\
= \frac{n^2}{N} \sum_{i,j=1}^{n} \left| \sum_{t=1}^{n} \sum_{k=1}^{N_{x_s,t}} \sum_{l=1}^{N_{x_s,l}} A_{ij}[sk,tl] \right|^2 \right.
& \left. \frac{n}{2} \forall p \geq 1. \right)
\end{align*}$$

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

- First control its moments, namely

$$\mathbb{E}^p_{Y_{X,s} \mid N_{X,s}, E_1, E_2, X} [a_{ij}] = \mathbb{E}^p_{Y_{X,s} \mid N_{X,s}, E_1, E_2, X} \left[ \sum_{s,t=1}^{N} \sum_{k=1}^{N_{x_s,t}} \sum_{l=1}^{N_{x_s,l}} A_{ij}[sk,tl] \right], \quad \forall p \geq 1. \quad (140)$$

- 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 $E_1$, let $Y_{x_s}^t \in \mathbb{R}^{N_{x_s,t}}$ be an independent copy of $Y_{x_s,t}$. Then note that

$$\mathbb{E}^p_{Y_{X,s} \mid N_{X,s}, E_1, E_2, X} [a_{ij}] = \mathbb{E}^p_{Y_{X,s} \mid N_{X,s}, E_1, E_2, X} \left[ \sum_{s,t=1}^{N} \sum_{k=1}^{N_{x_s,t}} \sum_{l=1}^{N_{x_s,l}} A_{ij}[sk,tl] \right].$$
Our next step is to decouple the sum in the last line of (141).

**Observation:**

One and zero with equal probabilities. We will shortly use the following simple

of independent standard Bernoulli random variables: each

where we defined the block-diagonal matrix $B_i$ (in the sense that its distribution is symmetric about the origin).

Moreover, similar to (138), it holds that

$$E_{y_{ix_1}, \xi_{x_1}} \left[ \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N^2_{x_i, \xi}} e_s^T \left( P_{x_s, y_{sk}} - E_{y_{sk}|x_1, \xi} \left[ P_{y_{sk}|x_1, \xi} \right] \right) \nabla f(x_s) \nabla f(x_s)^T \right]$$

$$= E_{y_{ix_1}, \xi_{x_1}} \left[ \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N^2_{x_i, \xi}} e_s^T \left( P_{x_s, y_{sk}} - E_{y_{sk}|x_1, \xi} \left[ P_{y_{sk}|x_1, \xi} \right] \right) \nabla f(x_s) \nabla f(x_s)^T \right]$$

$$= E_{y_{ix_1}, \xi_{x_1}} \left[ \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N^2_{x_i, \xi}} e_s^T \left( P_{x_s, y_{sk}} - E_{y_{sk}|x_1, \xi} \left[ P_{y_{sk}|x_1, \xi} \right] \right) \nabla f(x_s) \nabla f(x_s)^T \right]$$

$$= E_{y_{ix_1}, \xi_{x_1}} \left[ \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N^2_{x_i, \xi}} e_s^T \left( P_{x_s, y_{sk}} - E_{y_{sk}|x_1, \xi} \left[ P_{y_{sk}|x_1, \xi} \right] \right) \nabla f(x_s) \nabla f(x_s)^T \right]$$

$$\leq E_{y_{ix_1}, \xi_{x_1}} \left[ \sum_{s=1}^{N} \sum_{k \neq l} \frac{1}{N^2_{x_i, \xi}} e_s^T \left( P_{x_s, y_{sk}} - E_{y_{sk}|x_1, \xi} \left[ P_{y_{sk}|x_1, \xi} \right] \right) \nabla f(x_s) \nabla f(x_s)^T \right]$$

$$(142)$$

$$= \sum_{s, t=1}^{N_{x_i, \xi}} \sum_{k=1}^{N_{y_{sk}, \xi}} \sum_{l=1}^{N_{x_i, \xi}} B_{ij}[sk, tl],$$

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

$$B_{ij}[sk, tl]$$

$$= \begin{cases} \frac{N-2}{N} e_s^T \left( P_{x_s, y_{sk}} - P_{x_s, y_{sl}} \right) \nabla f(x_s) \nabla f(x_s)^T, & s = t \text{ and } k \neq l, \\ 0, & s \neq t \text{ or } k = l, \\ \end{cases}$$

$$= \begin{cases} \frac{N-2}{N} e_s^T \left( P_{x_s, y_{sk}} - P_{x_s, y_{sl}} \right) \nabla f(x_s) \nabla f(x_s)^T, & s = t \text{ and } k \neq l, \\ 0, & s \neq t \text{ or } k = l, \\ \end{cases}$$

for every $s, t \in [1 : N], k \in [1 : N_{y_{sk}, \xi}], l \in [1 : N_{x_i, \xi}]$. 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 (138), it holds that

$$B_{ij}[sk, sk] = 0,$$

$$E_{y_{x_i}, \xi_{x_i}} \left[ B_{ij}[sk, tl] \right] = 0, \quad s, t \in [1 : N], k \in [1 : N_{y_{sk}, \xi}], l \in [1 : N_{x_i, \xi}].$$

(143)

Our next step is to decouple the sum in the last line of (141). Decoupling. Let $\Xi = \{\xi_{sk}\}_{s,k}$ (with $s \in [1 : N]$ and $k \in [1 : N_{y_{sk}, \xi}]$) be a sequence of independent standard Bernoulli random variables: each $\xi_{sk}$ independently takes one and zero with equal probabilities. We will shortly use the following simple observation:

$$E_{\Xi} \left[ \xi_{sk} (1 - \xi_{tl}) \right] = \frac{1}{4}, \quad sk \neq tl.$$  

(144)
We now revisit (141) and write that
\[
\mathbb{E}^p_{Y_{x,\epsilon},X_{x,\epsilon},\ell_1,\ell_2} \left[ a_{ij} \right]
\leq \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{s,t=1}^N N_{x,s} N_{x,t} \sum_{k=1}^s \sum_{l=1}^t B_{ij}[sk,tl] \right]
\quad \text{ (see (141))}
\]
\[
= 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{s,t=1}^N N_{x,s} N_{x,t} \sum_{k=1}^s \sum_{l=1}^t \mathbb{E} \left[ \xi_{sk} (1 - \xi_{tl}) \right] \cdot B_{ij}[sk,tl] \right]
\quad \text{ (B_{ij}[sk,sl] = 0, and (144))}
\]
\[
\leq 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{s,t=1}^N N_{x,s} N_{x,t} \sum_{k=1}^s \sum_{l=1}^t \xi_{sk} (1 - \xi_{tl}) \cdot B_{ij}[sk,tl] \right],
\quad \text{(145)}
\]
where the last line above uses the Jensen's inequality. In particular, there must exist $\mathbb{E}_0 = \{\xi_{sk}\}_{s,k}$ that exceeds the expectation in the last line above, so that
\[
\mathbb{E}^p_{Y_{x,\epsilon},X_{x,\epsilon},\ell_1,\ell_2} \left[ a_{ij} \right]
\leq 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{s,t=1}^N N_{x,s} N_{x,t} \sum_{k=1}^s \sum_{l=1}^t \xi_{sk} (1 - \xi_{tl}) \cdot B_{ij}[sk,tl] \right]
\]
\[
= 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{\xi_{sk} = 1, \xi_{tl} = 0} B_{ij}[sk,tl] \right]
\]
\[
= 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{sk \in S_0, tl \notin S_0} B_{ij}[sk,tl] \right].
\quad \text{(S_0 := \{sk : \xi_{sk} = 1\} \subseteq [1 : N_{x,\epsilon}])}
\quad \text{(146)}
\]
Let $\{Y_{x,\epsilon}^{i}, Y_{x,\epsilon}^{iii}\} \subset \mathbb{R}^{n \times \#N_{x,\epsilon}}$ be an independent copy of $\{Y_{x,\epsilon}, Y_{X_{x,\epsilon}}^{i}\}$. For the sake of brevity, we will use the following short hand:
\[
\tilde{Y}_{x,\epsilon} := Y_{x,\epsilon} \cup Y_{x,\epsilon}^{i} \cup Y_{X_{x,\epsilon}}^{i} \cup Y_{X_{x,\epsilon}}^{iii},
\]
\[
\tilde{Y}_{x,\epsilon} := Y_{x,\epsilon} \cup Y_{x,\epsilon}^{i} \cup Y_{x,\epsilon}^{ii} \cup Y_{x,\epsilon}^{iii}, \quad \forall x_s \in X.
\quad \text{(147)}
\]
Equipped with the construction above, we revisit (146) and write that
\[
\mathbb{E}^p_{Y_{x,\epsilon},X_{x,\epsilon},\ell_1,\ell_2} \left[ a_{ij} \right]
\leq 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{sk \in S_0, tl \notin S_0} B_{ij}[sk,tl] \right]
\quad \text{ (see (146))}
\]
\[
= 4 \cdot \mathbb{E}^p_{Y_{x,\epsilon},Y_{X_{x,\epsilon}}^{\ell_1,\ell_2}|X_{x,\epsilon},\ell_2} \left[ \sum_{sk \in S_0, tl \notin S_0} N_{x,\epsilon}^{-2} e_s^i \left( P_{x,y_{sk}} - P_{x,y^s_{tl}} \right) e_j \right]
\quad \text{ (see (142))}
\]
\[
= 4 \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},e_1,e_2,X}\left[ \sum_{sk \in S_0, sl \notin S_0} N_{x,s}^{-2} \cdot e_i^*(P_{x,y,sk} - P_{x,y,sk}^*) \nabla f(x_s) \nabla f(x_s)^* \left( P_{x,y_{sl}} - P_{x,y_{sl}}^* \right) e_j \right]_{C_{ij}[sk,sl]}
\]
\]

\[
= 4 \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},e_1,e_2,X}\left[ \sum_{sk \in S_0, sl \notin S_0} C_{ij}[sk,sl] \right]
+ \sum_{sk \notin S_0} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},e_1,e_2,X}\left[ C_{ij}[sk,sl] \right]
+ \sum_{sk \in S_0, sl \in S_0} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},e_1,e_2,X}\left[ C_{ij}[sk,sl] \right]
\]
\[
\leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},e_1,e_2,X}\left[ \sum_{l,s=1}^{N_{x,s}} \sum_{k=1}^{N_{x,s}} \sum_{l=1}^{N_{x,s}} C_{ij}[sk,tl] \right],
\]

where we added two zero expectation terms in the last equality above. The last line above uses independence and Jensen’s inequality. Above, we also defined the block-diagonal matrix \(B_{ij} \in \mathbb{R}^{N_{x,s} \times N_{x,s}}\) such that
\[
C_{ij}[sk,tl] = \begin{cases} N_{x,s}^{-2} \cdot e_i^*(P_{x,y,sk} - P_{x,y,sk}^*) \nabla f(x_s) \nabla f(x_s)^* \left( P_{x,y_{sl}} - P_{x,y_{sl}}^* \right) e_j, & s = t \text{ and } k \neq l, \\ 0, & s \neq t \text{ or } k = l, \end{cases}
\]
for every \(s, t \in [1 : N], k \in [1 : N_{x,s}], l \in [1 : N_{x,s}].\) For every \(s \in [N],\) we can also define a family of matrices \(\{C_{skl}\}_{k,l \in [N_{x,s}]} \subset \mathbb{R}^{n \times n}\) such that
\[
C_{skl}[i,j] = C_{skl}[k,l], \quad \forall k, l \in [N_{x,s}].
\]
Note that
\[
C_{skl} := \frac{1}{N_{x,s}^2} \begin{pmatrix} P_{x,y,sk} - P_{x,y,sk}^* \nabla f(x_s) \nabla f(x_s)^* \left( P_{x,y_{sl}} - P_{x,y_{sl}}^* \right) & , k \neq l, \\ 0, & k = l. \end{pmatrix}
\]

The next step is to modulate the sum in the last line of (148) with a Rademacher sequence.

Modulation with Rademacher Sequences. Fix \(i, j \in [1 : n],\) and recall the definitions of \(C_{ij} \in \mathbb{R}^{N_{x,s} \times N_{x,s}}\) from (149). Let \(H = \{\eta_{sk}\}_{s,k}\) (with \(s \in [1 : N]\) and \(k \in [1 : N_{x,s}]\)) be a Rademacher sequence, that is \(\{\eta_{sk}\}_{s,k}\) are independent Bernoulli random variables taking \(\pm 1\) with equal chances. Also let \(H^1 = \{\eta_{sk}\}_{s,k}\) be an
independent copy of $H$. Then, we argue that

$$E_{Y_{x,e} | N_{x,e}, e_1, e_2, X} [c_{ij}]$$

$$\leq 4 \cdot E_{Y_{x,e} | N_{x,e}, e_1, e_2, X} \left[ \sum_{s,t=1}^{N} \sum_{k=1}^{N_{e_1,e}} \sum_{l=1}^{N_{e_2,e}} C_{ij}[s,k,t,l] \right]$$

(see (148))

$$= 4 \cdot E_{Y_{x,e} | Y_{X,e}^{(1)}, N_{x,e}, e_1, e_2, X} \left[ \sum_{s,k} \left( \sum_{t,l} C_{ij}[s,k,t,l] \right) \right]$$

(see (147))

$$= 4 \cdot E_{Y_{x,e} | Y_{X,e}^{(1)}, H | Y_{X,e}^{(1)}, Y_{X,e}^{(1)}, e_1, e_2, X} \left[ \sum_{s,k} \left( \sum_{t,l} \eta_{sk} \cdot \left( \sum_{t,l} C_{ij}[s,k,t,l] \right) \right) \right]$$

(independence and symmetry)

$$= 4 \cdot E_{Y_{x,e} | Y_{X,e}^{(1)}, H | N_{x,e}, e_1, e_2, X} \left[ \sum_{s,k} \left( \sum_{t,l} \eta_{sk} C_{ij}[s,k,t,l] \right) \right]$$

$$= 4 \cdot E_{Y_{x,e} | Y_{X,e}^{(1)}, H | Y_{X,e}^{(1)}, Y_{X,e}^{(1)}, e_1, e_2, X} \left[ \sum_{s,k} \left( \sum_{t,l} \eta_{sk} \cdot \left( \sum_{t,l} C_{ij}[s,k,t,l] \right) \right) \right]$$

(independence and symmetry)

$$= 4 \cdot E_{Y_{x,e} | H, H^1 | N_{x,e}, e_1, e_2, X} \left[ \sum_{s,k=1}^{N_{e_1,e}} \sum_{l=1}^{N_{e_2,e}} \eta_{sk} \eta_{tl} \cdot C_{ij}[s,k,t,l] \right]$$

$$= 4 \cdot E_{Y_{x,e} | H, H^1 | N_{x,e}, e_1, e_2, X} [c_{ij}] ,$$

where we set

$$c_{ij} := \left| \sum_{s,k=1}^{N_{e_1,e}} \sum_{l=1}^{N_{e_2,e}} \eta_{sk} \eta_{tl} \cdot C_{ij}[s,k,t,l] \right| .$$

Conditioned on everything but $H$ and $H^1$, $c_{ij}$ is a decoupled chaos: decoupled because $H = \{ \eta_{sk} \}_{s,k}$ and $H^1 = \{ \eta_{tl} \}_{s,k}$ are independent (Rademacher) sequences. The behavior of the moments of a chaos random variable is well-understood.

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

$$E_{H, H^1 | Y_{X,e} | e_1, e_2, X} [C_{ij}]$$

$$\leq \sqrt{E_{H, H^1 | Y_{X,e} | e_1, e_2, X} [C_{ij}^2]}$$

(Jensen’s inequality)

$$= ||C_{ij}||_F$$

($H$ and $H^1$ are independent Rademacher sequences)
\[ \leq \sqrt{N} \cdot \max_{s \in [1:N]} \| C_{stij} \|_F. \quad (C_{ij} \text{ is block-diagonal}) \] (154)

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

\[ \| C_{stij} \|_F^2 = \sum_{k,l=1}^{N_{xx,\epsilon}} |C_{stij}[k,l]|^2 \]
\[ = \sum_{k,l=1}^{N_{xx,\epsilon}} |C_{skl}[i,j]|^2 \quad (\text{see } (150)) \]
\[ \leq N_{xx,\epsilon}^2 \cdot \max_{k,l \in [1:N_{xx,\epsilon}]} |C_{skl}[i,j]|^2 \]
\[ \leq N_{xx,\epsilon}^2 \cdot \max_{k,l \in [1:N_{xx,\epsilon}]} \| C_{skl} \|_\infty^2, \quad (155) \]

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

\[ \| C_{skl} \|_\infty \]
\[ = \max_{i,j \in [1:n]} |C_{skl}[i,j]| \]
\[ = \max_{i,j \in [1:n]} |e_i^* C_{skl} e_j| \]
\[ = N_{xx,\epsilon}^2 \]
\[ \cdot \max_{i,j \in [1:n]} \left| e_i^* \left( P_{x,\epsilon,y,k} - P_{x,\epsilon,y,k} \right) \nabla f(x_s) \nabla f(x_s)^* \left( P_{x,\epsilon,y,k} - P_{x,\epsilon,y,k} \right) e_j \right| \quad (\text{see } (149)) \]
\[ \leq 4N_{xx,\epsilon}^2 \cdot \max_{i,j \in [1:n]} \max_{s \in [1:N]} \max_{y \in Y_{xx,\epsilon}} \| e_i^* P_{x,\epsilon,y} \nabla f(x_s) \|^2 \quad (\text{see } (147)) \]
\[ \leq 4N_{xx,\epsilon}^2 \cdot \max_{i,j \in [1:n]} \max_{s \in [1:N]} \max_{y \in Y_{xx,\epsilon}} \| P_{x,\epsilon,y} e_i \|_2 \cdot \| P_{x,\epsilon,y} \nabla f(x_s) \|_2^2 \quad (\text{Cauchy-Schwarz’s}) \]
\[ \leq 4N_{xx,\epsilon}^2 \cdot \frac{Q_{x,\epsilon} L_j^2}{n}, \quad (\text{conditioned on the event } \mathcal{E}_3) \] (156)

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

\[ \mathcal{E}_3 = \left\{ \max_{s \in [1:N]} \max_{i \in [1:n]} \max_{y \in Y_{xx,\epsilon}} \| P_{x,\epsilon,y} e_i \|_2^2 \leq \frac{Q_{x,\epsilon}}{n} \right\} \]
\[ \cap \left\{ \max_{s \in [1:N]} \max_{y \in Y_{xx,\epsilon}} \| P_{x,\epsilon,y} \nabla f(x_s) \|_2^2 \leq \frac{Q_{x,\epsilon} L_j^2}{n} \right\}, \quad (157) \]

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:

\[ \Pr_{\tilde{Y}_{x,\epsilon} \in \mathbb{X}_{x,\epsilon}, \tilde{E}_1, \tilde{E}_2, X} \left[ \mathcal{E}_3^C \right] \leq \left( \frac{p}{N_{XX,\epsilon,\epsilon} \rho_{H,\epsilon} N_{XX,\epsilon}} \right)^\frac{q}{2}. \quad (\text{see } (22)) \] (158)

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

\[ \mathbb{E}_{H,H'}[\tilde{Y}_{x,\epsilon}, \tilde{E}_1, \tilde{E}_2, X] [c_{ij}] \]
\[ \leq \| C_{ij} \|_F \]
\[ \leq \sqrt{N} \cdot \max_{s \in [1:N]} \| C_{stij} \|_F \quad (\text{see } (154)) \]
\[ \leq \sqrt{N} \cdot \max_{s \in [1:N]} \max_{k,l \in [1:N_{x^2}\epsilon]} N_{x^2}\epsilon \cdot \|C_{skl}\|_{\infty} \quad \text{(see (155))} \]

\[ \leq \sqrt{N} \cdot \max_{s \in [1:N]} \frac{4Q_{X,e}^2 L_f^2}{n^2 N_{x^2}\epsilon^2} \quad \text{(see (156))} \]

\[ \leq \frac{4\sqrt{N}Q_{X,e}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x^2}\epsilon}. \quad \text{(159)} \]

To control the higher order moments of \(c_{ij}\), we invoke the following result [1, 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 \(H^1 = \{\eta^1_k\}_k\), consider the decoupled (second-order) chaos

\[ c = \left| \sum_{k,l} \eta_k \eta^1_l \cdot C[k,l] \right|. \]

Then, it holds that

\[ \mathbf{E}[c - \mathbf{E}[c]] \preceq p \cdot b + \sqrt{p} \cdot \sigma, \quad \forall p \geq 1, \quad \text{(160)} \]

where

\[ b := \|C\|, \quad \text{(161)} \]

\[ \sigma := \sqrt{\mathbf{E}_H \left[ \|C\eta\|_2^2 \right]} = \|C\|_F, \quad \text{(162)} \]

and \(\eta\) 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 (153) (conditioned on \(X, \tilde{Y}_{X^2}\epsilon\) and the event \(E_1 \cap E_2 \cap E_3\)). To that end, note that

\[ b = \|C_{ij}\| \quad \text{(see (161))} \]

\[ = \max_{s \in [1:N]} \|C_{sij}\|. \quad \text{\(C_{ij}\) is block-diagonal} \quad \text{(163)} \]

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

\[ \|C_{sij}\| \leq N_{x^2}\epsilon \cdot \|C_{sij}\|_\infty \quad \text{(\(\|A\| \leq a \cdot \|A\|_\infty, \:\forall A \in \mathbb{R}^{a \times a}\))} \]

\[ \leq N_{x^2}\epsilon \cdot \max_{k,l \in [1:N_{x^2}\epsilon]} \|C_{skl}\|_\infty \quad \text{(see (150))} \]

\[ \leq N_{x^2}\epsilon \cdot \frac{4Q_{X,e}^2 L_f^2}{n^2 N_{x^2}\epsilon^2} \quad \text{(see (156))} \]

\[ \leq \frac{4Q_{X,e}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x^2}\epsilon}. \quad \text{(38)} \]

In light of (163), it follows that

\[ b \leq \max_{s \in [1:N]} \|C_{sij}\| \quad \text{(see (163))} \]

\[ \leq \frac{4Q_{X,e}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x^2}\epsilon}. \quad \text{(164)} \]
We argue likewise to find $\sigma$:

$$\sigma = \|C_{ij}\|_F \quad \text{(see (162))}$$

$$\leq \frac{4\sqrt{N}Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min x \in X N_{x,\epsilon}}. \quad \text{(see (159))} \tag{166}$$

With $b$ and $\sigma$ at hand, we now invoke Proposition 4 to write that

$$\mathbb{E}^p_{H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} [c_{ij}]$$

$$= \mathbb{E}^p_{H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x,\epsilon}} \sum_{l=1}^{N_{x,\epsilon}} \eta_{sk} \eta_{tl} \cdot C_{ij}[sk, tl] \right] \quad \text{(see (153))}$$

$$\leq \mathbb{E}^p_{H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} \left[ c_{ij} - \mathbb{E}^p_{H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} [c_{ij}] \right] + \mathbb{E}^p_{H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} [c_{ij}]$$

$$\lesssim (p \cdot b + \sqrt{p} \cdot \sigma) + \frac{\sqrt{N}Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min x \in X N_{x,\epsilon}} \quad \text{(see Proposition 4 and (159))}$$

$$\lesssim \left( p \cdot Q_{X,\epsilon}^2 L_f^2 + \frac{\sqrt{p} \cdot \sqrt{N}Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min x \in X N_{x,\epsilon}} \right) \quad \text{(see (165,166))}$$

$$\lesssim \sqrt{p} \cdot \frac{Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min x \in X N_{x,\epsilon}} \quad \text{(if $1 \leq p \leq N$)}$$

$$\lesssim \sqrt{p} \cdot \frac{Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min x_{\text{min},\epsilon} \cdot \rho_{\mu,X,\epsilon} N_{X,\epsilon}} \quad \text{(see (38) and (22))} \tag{167}$$

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

$$\mathbb{E}^p_{\tilde{Y}_{X,\epsilon},H,H^i|\mathcal{F}_3,\tilde{Y}_{X,\epsilon},\tilde{y}_{1,\epsilon},\tilde{y}_{2,\epsilon}} [c_{ij}] \lesssim \sqrt{p} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\text{min},\epsilon} \cdot \rho_{\mu,X,\epsilon} N_{X,\epsilon}}}. \tag{168}$$

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

$$|c_{ij}| = \left| \sum_{s,t=1}^N \sum_{k=1}^{N_{x,\epsilon}} \sum_{l=1}^{N_{x,\epsilon}} \eta_{sk} \eta_{tl} \cdot C_{ij}[sk, tl] \right| \quad \text{(see (153))}$$

$$\leq \sum_{s,t=1}^N \sum_{k=1}^{N_{x,\epsilon}} \sum_{l=1}^{N_{x,\epsilon}} \eta_{sk} \eta_{tl} \cdot C_{ij}[sk, tl] \quad \text{(triangle inequality)}$$

$$= \sum_{s=1}^N \sum_{k=1}^{N_{x,\epsilon}} \sum_{l=1}^{N_{x,\epsilon}} \eta_{sk} \eta_{tl} \cdot C_{sij}[k, l] \quad \text{(see (164))}$$

$$\leq \sum_{s=1}^N N_{x,\epsilon} \cdot \|C_{sij}\| \quad \text{($H, H^i$ are Rademacher sequences)}$$

$$\leq \sum_{s=1}^N N_{x,\epsilon}^2 \cdot \frac{4Q_{X,\epsilon}^2 L_f^2}{n^2 N_{x,\epsilon}^2} \quad \text{(see (164))}$$

$$= \frac{4NQ_{X,\epsilon}^2 L_f^2}{n^2}. \tag{169}$$
where the third line uses the fact that $C_{i,j}$ is block-diagonal with blocks $C_{s_{i,j}} \in \mathbb{R}^{N_{X,\epsilon},x,\epsilon \times N_{X,\epsilon}}$ and also uses (150). Putting everything back together, we finally argue that
\[
\mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[a_{ij}] 
\leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon}|H,H'|N_{X,\epsilon},E_{1,2},X}[c_{ij}] \quad \text{(see (152))}
\]
\[
\leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon}|H,H'|E_{3},N_{X,\epsilon},E_{1,2},X}[c_{ij}] + 4 \sup |c_{ij}| \cdot \left( \mathbb{E}_{\tilde{Y}_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[E_{3}^{C}] \right)^{\frac{1}{p}}
\]
\[
(\text{see (154))}
\]
\[
\leq \sqrt{p} \cdot \frac{NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}} + \frac{NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}}
\]
\[
(\text{see (168), (169), and (158))}
\]
\[
\leq \sqrt{p} \cdot \frac{NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}},
\]
(170)

when $1 \leq p \leq N$ (see (167)). At last, (170) describes the moments of the random variable $a_{ij}$ for fixed $i,j$ (and conditioned on $N_{X,\epsilon},E_{1,2},X$).

K.1.3. **Applying Markov’s inequality.** Given the estimates of the moments of $a_{ij}$ in (170), 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
\[
\Pr_{Y_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[|a_{ij}| > \gamma_{8}] = \Pr_{Y_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[|a_{ij}|^{p} > \gamma_{8}^{p}]
\]
\[
\leq \left( \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[|a_{ij}|] \right)^{\frac{p}{\gamma_{8}}}
\]
\[
(\text{Markov’s inequality})
\]
\[
\leq \left( \frac{C_{1}\sqrt{p}NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{\gamma_{8}^{p}n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}} \right)^{p},
\]
(171)

for an absolute constant $C_{1}$. In particular, the choice of
\[
\gamma_{8} = C_{1}\gamma_{7} \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}}, \quad p = \max \{\log n, 1\} \leq N, \quad \gamma_{7} \geq 1,
\]

yields
\[
\Pr_{Y_{X,\epsilon}|N_{X,\epsilon},E_{1,2},X}[|a_{ij}| \geq \gamma_{7} \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^{2}L_{f}^{2}_{j}}{n^{2} \sqrt{N_{X,\min,\epsilon}^{2}p_{\mu,X,\epsilon}N_{X,\epsilon}}}]
\]
\[
\leq \gamma_{7}^{-\log n} \cdot n^{-\log \gamma_{7}},
\]
(172)

With the tail bound of $a_{ij}$ finally available above (for fixed $i,j \in [1 : n]$ and conditioned on $N_{X,\epsilon},E_{1,2},X$), we next quantify how $\sum_{X,Y_{X,\epsilon}}^{o}$ concentrates about its expectation.
K.1.4. Applying the union bound. In light of (172) and by applying the union bound to \( \{a_{ij}\}_{i,j} \), we arrive at the following statement.

\[
\Pr_{Y_X,\{N_{X,x}\},\mathcal{E}_1,\mathcal{E}_2,X} \left[ \max_{1 \leq i,j \leq n} |a_{ij}| \lesssim \gamma \cdot \sqrt{\log n} \cdot \frac{nQ_{X,e}^2 L_f^2}{n^2 \sqrt{N_{X,min,e} P_{\mu,X,e} N_{X,e}}} \right] \\
\geq 1 - n^2 \cdot n^{-\log \gamma} \quad \text{(union bound and (172))} \tag{173}
\]

K.2. Second and third norms. In this section, we bound the second and third norms in the last line of (132) using the Bernstein inequality. Let us bound the second norm as

\[
n^2 \frac{1}{N} \left( \sum_{s=1}^N \sum_{k=1}^{N_{x,e}} (P_{s,y,k} - E_{y|x}|\mathcal{E}_1,\mathcal{E}_2,P_{s,y}) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y|x|\mathcal{E}_1,x}[P_{s,y}] \right) \|_F \\
\leq n^2 \frac{1}{N} \left( \sum_{s=1}^N \sum_{k=1}^{N_{x,e}} (P_{s,y,k} - E_{y|x}|\mathcal{E}_1,\mathcal{E}_2,P_{s,y}) \nabla f(x_s) \nabla f(x_s)^* E_{y|x|\mathcal{E}_1,x}[P_{s,y}] \right) \|_F \\
= n^2 \frac{1}{N} \left( \sum_{s=1}^N \sum_{k=1}^{N_{x,e}} A_{x,y,k} \right) \|_F \tag{174}
\]

where \( \{A_{x,y,k}\}_{s,k} \subset \mathbb{R}^{n \times n} \) 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

\[
\|A_{x,y,k}\|_F = \frac{1}{N_{x,e}} \left\| (P_{x,y,k} - E_{y|x}|\mathcal{E}_1,\mathcal{E}_2,P_{x,y}) \nabla f(x_s) \nabla f(x_s)^* E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \right\|_F \\
\leq \frac{1}{N_{x,e}} \left\| P_{x,y,k} \nabla f(x_s) \nabla f(x_s)^* E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \right\|_F \\
+ \frac{1}{N_{x,e}} \left\| E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \nabla f(x_s) \nabla f(x_s)^* E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \right\|_F \\
\leq \frac{1}{N_{x,e}} \left\| P_{x,y,k} \nabla f(x_s) \right\|_2 \left\| E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \nabla f(x_s) \right\|_2 \\
+ \frac{1}{N_{x,e}} \left\| E_{y|x|\mathcal{E}_1,x}[P_{x,y}] \nabla f(x_s) \right\|_2^2 \quad \text{(Jensen’s inequality)} \\
\leq \frac{1}{N_{x,e}} \sqrt{Q_{X,e} L_f^2} \sqrt{\frac{L_f^2}{n}} + \frac{L_f^2}{K_{\mu} N_{x,e} n} \quad \text{(see (122,9), [58, Lemma 5.5])} \\
\leq \frac{2Q_{X,e} L_f^2}{\min_{s \in [N]} N_{x,e}} =: b \quad \text{(if } Q_{X,e} \geq K_{\mu}^{-1} \text{)} \tag{175}
\]

On the other hand,

\[
\sum_{s=1}^N \sum_{k=1}^{N_{x,e}} E_{Y_X,E_{X,e}|\mathcal{E}_1,\mathcal{E}_2,X} \|A_{x,y,k}\|_F^2
\]
An application of the Bernstein inequality now yields that conditioned on $E$,

$$\sum_{s=1}^{N} \frac{1}{N_{x,s}} \sum_{k=1}^{N_{x,s}} E_{y_s} \left[ \sum_{k=1}^{N_{x,s}} \left| P_{x,s,y} \nabla f(x_s) \right|^2 \right] \leq \sum_{s=1}^{N} \frac{1}{N_{x,s}} \sum_{k=1}^{N_{x,s}} E_{y_s} \left[ P_{x,s,y} \nabla f(x_s) \right]^2 \leq \sum_{s=1}^{N} \frac{1}{N_{x,s}} \sum_{k=1}^{N_{x,s}} E_{y_s} \left[ P_{x,s,y} \nabla f(x_s) \right]^2 \leq \sum_{s=1}^{N} \frac{1}{N_{x,s}} \sum_{k=1}^{N_{x,s}} E_{y_s} \left[ P_{x,s,y} \nabla f(x_s) \right]^4 \quad \text{(Jensen’s inequality)}$$

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

$$\max[b, \sigma] \leq 2 \sqrt{\frac{N \cdot Q_{x,c} L_f^4}{\min_{s \in [N]} N_{x,s}}}.$$  

(177)

An application of the Bernstein inequality now yields that conditioned on $E_1, E_2, N_{x,s}$, $X$,

$$\frac{n^2}{N} \sum_{s=1}^{N} \sum_{k=1}^{N_{x,s}} \left( P_{x,s,y} - E_{y_s} P_{x_s,y} \right) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y_s} \left[ P_{x_s,y} \right] \right] \leq \frac{n^2}{N} \sum_{s=1}^{N} \sum_{k=1}^{N_{x,s}} \left( P_{x,s,y} - E_{y_s} P_{x_s,y} \right) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y_s} \left[ P_{x_s,y} \right] \right] \leq \frac{n^2}{N} \cdot \gamma \max[b, \sigma] \leq \frac{n^2}{N} \cdot \gamma \max[b, \sigma]$$

$$\leq \frac{\gamma n^2}{N} \sum_{s=1}^{N} \sum_{k=1}^{N_{x,s}} \left( P_{x,s,y} - E_{y_s} P_{x_s,y} \right) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y_s} \left[ P_{x_s,y} \right] \right] \leq \frac{\gamma n^2}{N} \sum_{s=1}^{N} \sum_{k=1}^{N_{x,s}} \left( P_{x,s,y} - E_{y_s} P_{x_s,y} \right) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y_s} \left[ P_{x_s,y} \right] \right] \leq \frac{\gamma n^2}{N} \sum_{s=1}^{N} \sum_{k=1}^{N_{x,s}} \left( P_{x,s,y} - E_{y_s} P_{x_s,y} \right) \nabla f(x_s) \nabla f(x_s)^* \sum_{l \neq k} E_{y_s} \left[ P_{x_s,y} \right] \right]$$

(178)

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 (132).

K.3. Bound on (132). We now combine the bounds for the terms in (132) obtained in Sections K.1 and K.2. Applying (139), we have that conditioned on $E_1, E_2, N_{x,s}, X$,

$$\left\| \sum_{i,j=1}^{n} a_{ij} \right\|_F \leq \frac{n^2}{N} \sum_{i,j=1}^{n} a_{ij}^2.$$
we conclude that 

\[ p \text{ under (158) (with } 384 \text{ A. EFTEKHARI, M. B. WAKIN, P. LI AND P. G. CONSTANTINE) } \]

Since for any \( \gamma, \gamma \)

\[ \leq n^3 \cdot \max_{i,j \in [1:n]} |a_{ij}| + 2 \cdot \frac{\gamma nQ_X e L_f^2}{\sqrt{\rho_{\mu,X,e}N_{X,e}}} (\text{see (178)}) \]

\[ \leq n^3 \cdot \gamma \cdot \sqrt{\log n} \cdot \frac{N Q_X e L_f^2}{n^2 \sqrt{N_{\min,e} \rho_{\mu,X,e} N_{X,e}}} + 2 \cdot \frac{\gamma nQ_X e L_f^2}{\sqrt{\rho_{\mu,X,e}N_{X,e}}} (\text{see (173)}) \]

\[ = \gamma \cdot \sqrt{\log n} \cdot \frac{n Q_X e L_f^2}{\sqrt{N_{\min,e} \rho_{\mu,X,e} N_{X,e}}} + 2 \cdot \frac{\gamma nQ_X e L_f^2}{\sqrt{\rho_{\mu,X,e}N_{X,e}}} \]

\[ \leq \max[\gamma, \gamma] \cdot \sqrt{\log n} \cdot \frac{n \cdot \max[Q_X e, Q_{X,e}^2] \cdot L_f^2}{\sqrt{\rho_{\mu,X,e}N_{X,e}}} \]  

(179)

for \( \gamma, \gamma \geq 1 \) and except with a probability of at most \( e^{-\gamma} + n^2 \cdot n^{-\log \gamma} \). This holds under (158) (with \( p = \max[\log n, 1] \leq N \)).

Finally, we proceed to remove the conditioning on \( \mathcal{E}_1 \). Similar to (128), we have

\[ \mathbb{E}_{X,Y,X_0} \| \sum_{i=1}^{n} \sum_{k \neq j} \frac{1}{N_{x,e}} \| P_{x,y_{sk}} \nabla f(x_s) \nabla f(x_s) \| \| P_{x,y_{sk}} \|_F \]  

(see (112))

\[ \leq n^2 \sum_{s=1}^{N} \sum_{k \neq j} \frac{1}{N_{x,e}} \| P_{x,y_{sk}} \nabla f(x_s) \|_2 \| \nabla f(x_s) \|_2 \]

\[ \leq n^2 \sum_{s=1}^{N} \sum_{k \neq j} \frac{1}{N_{x,e}} \| \nabla f(x_s) \|_2^2 \]

\[ \leq \frac{n^2}{N} \sum_{s=1}^{N} \sum_{k \neq j} \frac{L_f^2}{N_{x,e}} \]  

(see (9))

\[ \leq \frac{n^2}{N} \sum_{s=1}^{N} L_f^2 \]

\[ = n^2 L_f^2, \]  

(181)

we conclude that

\[ \| \mathbb{E}_{X,Y,X_0} \| \sum_{i=1}^{n} \sum_{k \neq j} \frac{1}{N_{x,e}} \| \sum_{i=1}^{n} \sum_{k \neq j} \frac{1}{N_{x,e}} \| X_{Y,X_0}^a \| \sum_{i=1}^{n} \sum_{k \neq j} \frac{1}{N_{x,e}} \| X_{Y,X_0}^a \| \| X_{Y,X_0}^a \| \| X_{Y,X_0}^a \| \]  

\[ = n^2 L_f^2, \]  

(181)
\[ \text{Proof of Lemma I.3.} \] First, to prove (119), suppose \( X \) and the neighborhood structure \( \mathcal{N}_{X,e} = \{N_{x,e} \}_{x \in X} \) are fixed. Then, for every \( x \in X \), the columns of the matrix \( Y_{x,e} \in \mathbb{R}^{n \times N_{x,e}} \) are random vectors drawn from the conditional probability measure \( \mu_{x,e} \) (see (18)). For fixed \( x \in X \) and with \( y \sim \mu_{x,e} \), recall from Assumption 1 that

\[ \text{Pr}_{y|x} \left[ \left\| P_{x,y}v \right\|_2^2 > \frac{\gamma_1}{n} \right] \leq e^{-K_{\mu} \gamma_1}, \quad (184) \]

for arbitrary (but fixed) \( v \in \mathbb{R}^n \) with \( \|v\|_2 = 1 \) and \( \gamma_1 \geq 0 \). The inequality (119) readily follows with an application of the union bound: For all possible choices of \( x, y \), it holds that

\[ \|P_{x,y} \nabla f(x)\|_2^2 \leq \frac{\gamma_1}{n} \left\| \nabla f(x) \right\|_2^2 \leq \frac{\gamma_1 L_2^2}{n}, \quad (185) \]

except with a probability \( \leq N_{X,e} e^{-K_{\mu} \gamma_1} \). With the choice of

\[ \gamma_1 = Q_{X,e} = \gamma_2 K_{\mu}^{-1} \log^2(N_{X,e}) \]

This holds for \( \gamma, \gamma_7 \geq 1 \) and under (158) (with \( p = \max[\log n, 1] \leq N \)). Setting \( \gamma = \gamma_7 \) completes the proof of Lemma I.2.
for $\gamma_2 \geq 3$, we establish (119).

Our next goal is to prove that (117) is satisfied. Note that the probability in (117) is conditioned on $\mathcal{E}_1$. We can remove this conditioning using the law of total probability:

$$
\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C] = \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C | \mathcal{E}_1] \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_1] + \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C | \mathcal{E}_1^C] \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_1^C]
$$

Rearranging terms, we have that

$$
\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C] \leq \frac{\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C | \mathcal{E}_1]}{\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_1]} = \frac{\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C | \mathcal{E}_1^C]}{1 - \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_1^C]} \lesssim \Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C],
$$

where the last line follows under the assumption that $N_{X,e}$ large enough that, under (119), $\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C | \mathcal{E}_1^C]$ is bounded above by a constant smaller than 1.

To bound the right hand side in (186), suppose $X$ and the neighborhood structure $\overline{N}_{X,e} = \{N_{x,e} \}_{x \in X}$ are fixed. Then, for every $x \in X$, the columns of the matrix $\tilde{Y}_{x,e} \in \mathbb{R}^{n \times (4N_{x,e})}$ are random vectors drawn from the conditional probability measure $\mu_{x,e}$ (see (18)). For fixed $x \in X$ and with $y \sim \mu_{x,e}$, recall from Assumption 1 that (184) 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

$$
\|P_{x,y}^i e_i\|^2 \leq \frac{\gamma_1}{n}, \quad \|P_{x,y} \nabla f(x)\|^2 \leq \frac{\gamma_1 \|\nabla f(x)\|^2}{n} \leq \frac{\gamma_1 L_f^2}{n}, \quad \text{see (9)}
$$

(187)

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

$$
\Pr_{\tilde{Y}_{X,e}} [\mathcal{E}_3^C] \lesssim nN_{X,e}^{-1} \gamma_2 \log(N_{X,e})
$$

(188)
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 (186) and (188) proves that (117) is satisfied and thus completes the proof of Lemma I.3.

E-mail address: armin.eftekhari@gmail.com
E-mail address: mwakin@mines.edu
E-mail address: pingli@stat.rutgers.edu
E-mail address: paul.constantine@colorado.edu