Beyond Johnson-Lindenstrauss: Uniform Bounds for Sketched Bilinear Forms

Rohan Deb **Qiaobo Li** **Mayank Shrivastava** **Arindam Banerjee**
Siebel School of Computing and Data Science, University of Illinois, Urbana Champaign

Abstract

Uniform bounds on sketched inner products underpin several important computational and statistical results in machine learning and randomized algorithms, including the Johnson-Lindenstrauss (J-L) lemma, the Restricted Isometry Property (RIP), randomized sketching, etc. However, many modern analyses involve *sketched bilinear forms*, for which existing uniform bounds either do not apply or are not sharp on general sets. In this work, we develop a general framework to analyze such sketched bilinear forms, and derive uniform bounds in terms of geometric complexities of the associated sets. Our approach relies on *generic chaining* and introduces new techniques for handling suprema over pairs of sets. We further extend our results to (i) sketch matrices with conditionally independent entries, e.g., as in CountSketch and SRHT (Subsampled Randomized Hadamard Transform), and (ii) bilinear forms involving a sum of T sketch matrices, showing that the deviation scales as $\sqrt{T}$. This unified analysis recovers known results such as the J-L lemma as special cases, while extending RIP guarantees. Using our new bounds, we give tighter convergence bounds for sketched federated learning, and develop sketched bandits whose regret depends on the geometric complexity of the action and parameter sets rather than the ambient dimension.

1 INTRODUCTION

Sketching is a powerful technique for dimensionality reduction, widely used across many areas of numerical linear algebra and machine learning (ML). In linear regression, it enables solving large-scale least squares problems efficiently by reducing the size of the data matrix while preserving key geometric properties [Clarkson and Woodruff, 2013, Meng and Mahoney, 2013, Song et al., 2017, Andoni et al., 2018], while in distributed or federated learning, it facilitates communication-efficient protocols for matrix multiplication and aggregation [Song et al., 2023, Shrivastava et al., 2024, Boutsidis et al., 2016]. In recent times, sketching has also been applied to several other areas, such as reinforcement learning [Wang et al., 2020b, Shrivastava et al., 2023, Kuzborskij et al., 2019a], training deep networks [Song et al., 2021a,b, Gao et al., 2022], and convex programming [Jiang et al., 2021, Song and Yu, 2021, Qin et al., 2023] (see Appendix A for detailed related works).

A foundational result, the Johnson-Lindenstrauss (J-L) Lemma formalizes such a geometric guarantee, showing that random projections approximately preserve pairwise distances [Johnson and Lindenstrauss, 1984, Ailon and Chazelle, 2006]. More generally, RIP [Banerjee et al., 2014, Negahban et al., 2012, Vershynin, 2014] provides a framework to understand when a random matrix acts as an approximate isometry over a structured set, and underpins theoretical guarantees in compressed sensing and high-dimensional estimation. A matrix $\mathbf{S} \in \mathbb{R}^{b\times d}$ is said to satisfy RIP for a sparse set $\mathcal{U} \subseteq \mathbb{R}^d$ with constant δ_r if the following holds for all $u \in \mathcal{U}$:

$$(1-\delta_r)\|u\|_2^2 \le \|\mathbf{S}u\|_2^2 \le (1+\delta_r)\|u\|_2^2. \tag{1}$$

In the context of sketching, matrix $\mathbf{S}$ sketches the d-dimensional vector u to a b dimensional vector $\mathbf{S}u$. Condition (1) can be equivalently expressed as $\sup_{u\in\mathcal{U}} |\|\mathbf{S}u\|_2^2 - \|u\|_2^2| \le \delta_r\|u\|_2^2$.

In many applications one encounters *sketched bilinear forms* such as $(u^\top \mathbf{S}^\top \mathbf{S}\ v)$, for vectors $u \in \mathcal{U}, v \in \mathcal{V}$, and require uniform bounds for $\sup_{u\in\mathcal{U},v\in\mathcal{V}} |u^\top \mathbf{S}^\top \mathbf{S}\ v - u^\top v|$. For example in linear regression the response is given by $y_i = \beta^\top \mathbf{x}_i + \eta_i$, where $\beta \in \mathcal{B} \subseteq \mathbb{R}^d$ is the unknown parameter, $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is the input and η_i is some noise. In linear bandits [Abbasi-Yadkori et al., 2011, Chu et al., 2011] the reward is given by $r_i = \theta_*^\top a_i + \eta_i$, $\theta_* \in \Theta \subseteq \mathbb{R}^d$ is the unknown parameter, $a_i \in \mathcal{A} \subseteq \mathbb{R}^d$ is the chosen action and η_i is a sub-gaussian noise. In high dimensions, i.e., when

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

d is very large, one might sketch both the unknown parameter and the input, and would naturally want to bound $|\theta_*^\top \mathbf{S}^\top \mathbf{S} a_i - \theta_*^\top a_i|$; and similarly for linear regression. In Section 4 and 5, we develop sketched variants of linear regression and linear bandit algorithms that leverage bilinear sketching of both the unknown parameter and the input vectors and yields sharper bounds on the error and regret respectively, which scale with the *geometric complexity* of the sets, such as their Gaussian width (see Section 2 for definition and Appendix C for examples), rather than the ambient dimension d. Beyond regression and bandits, such sketched bilinear forms also arise in distributed and federated learning, where clients locally update their parameters, sketch the updates, and communicate them to the server [Song et al., 2023, Shrivastava et al., 2024]. Analyzing the finite-time behavior of these sketched algorithms similarly requires bounding deviations of the form $|g^\top \mathbf{S}^\top \mathbf{S} h - g^\top h|$, where g ranges over possible gradients and h over eigenvectors of the loss Hessian.

A general framework to analyze such deviations is the *random quadratic form* [Krahmer et al., 2014b, Banerjee et al., 2019]. Given a set of $m \times n$ matrices $\mathcal{A}$ and a sub-gaussian random vector $\boldsymbol{\xi} \in \mathbb{R}^n$ with i.i.d. entries,

$$C_{\mathcal{A}}(\boldsymbol{\xi}) = \sup_{A \in \mathcal{A}} \left| \|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2 \right| = \sup_{A \in \mathcal{A}} \left| \boldsymbol{\xi}^\top A^\top A \boldsymbol{\xi} - \mathbb{E}\, \boldsymbol{\xi}^\top A^\top A \boldsymbol{\xi} \right|. \quad (2)$$

Note that the term $\sup_{u \in \mathcal{U}} |\|\mathbf{S}u\|_2^2 - \|u\|_2^2|$ can be equivalently expressed as in (2) by converting the matrix A into a vector $u = \mathrm{vec}(A)$ and converting $\boldsymbol{\xi}$ into the matrix $\mathbf{S}$ (see Section E for details). In this work we develop deviation bounds for the following random variable with bilinear forms:

$$C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) = \sup_{M \in \mathcal{M}, N \in \mathcal{N}} \left| \boldsymbol{\xi}^\top M^\top N \boldsymbol{\xi} - \mathbb{E}\, \boldsymbol{\xi}^\top M^\top N \boldsymbol{\xi} \right| \quad (3)$$

and show that this yields sharp statistical guarantees for a variety of ML problems where such cross-terms naturally arise—particularly in settings involving bilinear sketches of the form $(u^\top \mathbf{S}^\top \mathbf{S}\, v)$.

Bilinear forms have been studied using a covering based analysis, which provides deviation bounds via Dudley's entropy integral (e.g., Sarlos [2006], Woodruff [2014]). In contrast, our approach uses *generic chaining* [Talagrand, 2005, 2014], yields bounds in terms of γ_2 functional and Gaussian-width, are *strictly sharper* on many sets than Dudley's entropy, and recover the classical finite-set rates. For a detailed comparison between Dudley's entropy integral (sometimes referred to as just 'chaining') and *generic* chaining, and a discussion on its sharpness see Appendix B. The related works section (Appendix A) also discusses various developments in ML using chaining techniques.

We outline our *main technical contributions* below:

1. **Uniform Deviation Bounds.** Our main result is a large deviation bound on $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ defined in (3) (cf. Theorem 3.1). This significantly generalizes existing result in Krahmer et al. [2014b] that provide such bounds for $C_{\mathcal{A}}(\boldsymbol{\xi})$ in (2). Our approach relies on Generic Chaining [Talagrand, 2005, 2014] and introduces new techniques for handling suprema over pairs of sets, which would be of independent interest (more details in Section 3.1). One such result we develop is a *double chaining bound* (Theorem 3.4) for stochastic processes indexed by pairs $(u, v) \in \mathcal{U} \times \mathcal{V}$. It provides a uniform deviation bound by chaining separately over $\mathcal{U}$ and $\mathcal{V}$, yielding a bound that scales with the sum of their γ_2 functional (see Remark 3.2 for more details and Section 2 for definitions).

2. **Applications.** This unified analysis recovers known results such as the J-L lemma as special cases, while also extending RIP-type guarantees to preserving inner products over arbitrary sets (see Proposition 4.1 and 4.2 respectively). We also apply our results to obtain improved finite time convergence guarantees for sketched Federated Learning algorithms (see section 4.3) and to obtain error bounds for linear regression that scale with the geometric complexity of the parameter and input sets rather than the ambient dimension (see Section I).

3. **Conditional Independence Extension.** We relax the i.i.d. assumption to conditionally independent coordinates (Assumption 4) and extend our deviation bounds for $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ (Theorem 5.1). CounSketch satisfies this condition [Banerjee et al., 2019], and we show that SRHT (Subsampled Randomized Hadamard Transform) also satisfies it.

4. **Sum of Random Quadratic Forms.** We next extend our deviation bound to $\sup_{M \in \mathcal{M}, N \in \mathcal{N}} |\sum_{t=1}^T (\boldsymbol{\xi}_t^\top M^\top N \boldsymbol{\xi}_t) - \mathbb{E}(\boldsymbol{\xi}_t^\top M^\top N \boldsymbol{\xi}_t)|$, where $\boldsymbol{\xi}_t$ are i.i.d., and show that the deviation scales by an extra $\sqrt{T}$ factor (see Theorem 5.2). This result is especially useful when the sketching matrix changes at every iteration, and is used in our bandit analysis. We note two things: (1) A standard Hoeffding type argument does not apply here because of the sup and one has to extend the generic chaining analysis. We additionally provide a simpler expected analysis in Appendix I.1 to show how the scale $\sqrt{T}$ appears. (2) Such an analysis for sum of random quadratic forms doesn't exist even for the single set case in (2).

5. **Sketched Bandits.** We design sketched linear bandit algorithms `sk-LinUCB` and `sk-LinTS` under action set assumptions, proving regret that scales

with the geometric complexity of the action and parameter sets. Our analysis splits regret into a sketched space term and a restricted isometry term, which can be bounded by our new results. Experiments show lower regret and faster runtime when the true parameter has structure.

2 PRELIMINARIES

Notation. The notation $a \lesssim b$ means that the inequality holds up to a multiplicative constants; i.e., it implies there exists a constant $C > 0$ such that $a \leq Cb$. The notations $y = \Theta(x)$ (respectively $y = O(x)$, $y = \Omega(x)$) implies there exists constants c_1, c_2, c_3, c_4 such that $c_1 \cdot x \leq y \leq c_2 \cdot x$ (respectively $y \leq c_3 \cdot x$, $y \geq c_4 \cdot x$), and $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{O}(\cdot)$ notations hide the dependence on log terms. We interchangeably use $x^\top y$ and $\langle x, y \rangle$ to denote inner the product.

Random Quadratic Forms: Quadratic forms involving random vectors arise in high-dimensional statistics, machine learning, and random matrix theory. Consider $\|A\boldsymbol{\xi}\|_2^2 = \boldsymbol{\xi}^\top (A^\top A)\boldsymbol{\xi}$ where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) \in \mathbb{R}^n$ is a random vector and $A \in \mathbb{R}^{n \times n}$ is a fixed matrix. When $\boldsymbol{\xi}$ has independent entries, this expression is referred to as a *random chaos* of order 2. A major result in this area is the *Hanson-Wright inequality* [Hanson and Wright, 1971, Vershynin, 2018], which gives exponential tail bounds for deviation of $\|A\boldsymbol{\xi}\|_2^2$ from its mean.

Theorem 2.1 (**Hanson–Wright**). *Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, sub-Gaussian entries. Let $A \in \mathbb{R}^{n \times n}$ be any fixed matrix. For any $\epsilon \geq 0$,*

$$\mathbb{P}\Big(\big| \|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}[\|A\boldsymbol{\xi}\|_2^2] \big| \geq \epsilon \Big) \lesssim \exp\left(-c \min\left(\frac{\epsilon^2}{\|A^\top A\|_F^2}, \frac{\epsilon}{\|A^\top A\|_{2\to 2}} \right) \right),$$

where $\|A\|_F$ is the Frobenius norm, and $\|A\|_{2\to 2}$ is the operator norm of A.

However, many applications in high-dimensional statistics and signal processing, such as restricted isometry property (RIP) or sketching analysis, involve the supremum of such quadratic forms over a *set* of matrices. That is, instead of bounding deviations of $\|A\boldsymbol{\xi}\|_2^2$ for fixed A, one is interested in bounding $\sup_{A \in \mathcal{A}} \left| \|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2 \right|$, where $\mathcal{A}$ is a class of matrices. Krahmer et al. [2014b] significantly generalize the Hanson-Wright result by analyzing such *suprema of chaos processes*. They show using chaining arguments, one can control this supremum in terms of geometric complexity measures of $\mathcal{A}$. The results depend on two types of *complexity measures* for the collection $\mathcal{A}$. The first type consists of the radii of $\mathcal{A}$ under different matrix norms and is defined below.

Definition 2.1 (**Radii of $\mathcal{A}$**). *The radii of $\mathcal{A}$ under the Frobenius norm and the operator norm, denoted by $d_F(\mathcal{A})$ and $d_{2\to 2}(\mathcal{A})$, respectively. The Frobenius norm is given by $\|A\|_F = \sqrt{\mathrm{Tr}(A^\top A)}$, and the operator norm is defined as $\|A\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2 \leq 1} \|A\mathbf{x}\|_2$. For a given set $\mathcal{A}$, we define:*

$$d_F(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_F, \quad d_{2\to 2}(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_{2\to 2}.$$

The second type of complexity measure involves Talagrand's γ_2 functional [Talagrand, 2005, 2014], written as $\gamma_2(\mathcal{A}, \mathbf{d})$, with distance metric $\mathbf{d}$, defined as follows.

Definition 2.2 (**Talagrand's γ_2 functional** [Talagrand, 2005, 2014]). *For a metric space $(\mathcal{A}, \mathbf{d})$, an admissible sequence of $\mathcal{A}$ is a collection of subsets of $\mathcal{A}$, $\{T_r(\mathcal{A}) : r \geq 0\}$, such that for every $s \geq 1$, $|T_r(\mathcal{A})| \leq 2^{2^r}$ and $|T_0| = 1$. For $\beta \geq 1$, define*

$$\gamma_\beta(\mathcal{A}, d) = \inf \sup_{t \in \mathcal{A}} \sum_{r=0}^{\infty} 2^{r/\beta} \mathbf{d}(t, T_r(\mathcal{A})), \tag{4}$$

where the inf is taken w.r.t. admissible sequences of $\mathcal{A}$.

The next theorem generalizes Hanson-Wright by giving deviation bound uniformly for all $A \in \mathcal{A}$.

Theorem 2.2 (Theorem 3.1, [Krahmer et al., 2014b]). *Let $\mathcal{A}$ be a set of matrices, and $\boldsymbol{\xi}$ be a random vector whose entries ξ_j are independent, mean-zero, unit variance, and L-subgaussian random variables. Set*

$$W = \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})\left(\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{A})\right) + d_F(\mathcal{A}) d_{2\to 2}(\mathcal{A}),$$
$$V = d_{2\to 2}(\mathcal{A})\left(\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{A})\right), \; U = d_{2\to 2}^2(\mathcal{A}).$$

Then, for any $\epsilon > 0$, we have $\mathbb{P}\left(\mathcal{C}_{\mathcal{A}}(\boldsymbol{\xi}) \geq c_1 W + \epsilon\right) \leq 2\exp\left(-c_2 \min(\frac{\epsilon^2}{V^2}, \frac{\epsilon}{U})\right)$, where c_1, c_2 depend on L.

Finally, we define another geometric complexity – Gaussian width. $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$ can be upper bounded by the Gaussian width up to a constants [Talagrand, 2014].

Definition 2.3 (**Gaussian Width**). *Let $\mathcal{T} \subseteq \mathbb{R}^d$ be a bounded subset. The* Gaussian width *of $\mathcal{T}$ is defined as $w(\mathcal{T}) := \mathbb{E}_g\left[\sup_{t \in \mathcal{T}} \langle g, t \rangle\right]$, where $g \sim \mathcal{N}(0, I_d)$. Further let $\mathcal{A} \subseteq \mathbb{R}^{m \times n}$ be a bounded set. The Gaussian width of $\mathcal{A}$ is defined as $w(\mathcal{A}) := \mathbb{E}_G\left[\sup_{A \in \mathcal{A}} \left|\mathrm{Tr}(G^\top A)\right|\right]$, where $G = [g_{i,j}] \in \mathbb{R}^{m \times n}$ is a random matrix with i.i.d. standard normal entries, i.e., $g_{i,j} \sim \mathcal{N}(0, 1)$.*

3 MAIN RESULTS

Let $\mathcal{M}$ be a set of $(m \times n)$ matrices and $\mathcal{N}$ be a set of $(n \times m)$ matrices. We make the following assumption on the stochastic process $\boldsymbol{\xi}$.

Assumption 1. *Suppose $\boldsymbol{\xi}$ be a random vector with independent coordinates $\boldsymbol{\xi}_i$, each of which is a mean zero L-subgaussian random variable.*

We want to develop large deviation bound for $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ defined in (2). Our analysis and results would rely on radii of $\mathcal{M},\mathcal{N}$ under the Frobenius and operator norms, denoted by $d_F(\cdot)$ and $d_{2\to2}(\cdot)$ (see Definition 2.1) and Talagrand's γ_2 functional (see Definition 2.2).

Theorem 3.1: (Uniform Deviation Bounds for Cross Inner Products)

Let $\mathcal{M}$ and $\mathcal{N}$ be set of $(m \times n)$ and $(n \times m)$ matrices respectively, and let $\boldsymbol{\xi}$ be a random vector satisfying Assumption 1. We define

$$W = \gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) + d_F(\mathcal{N})\Big] + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) + d_F(\mathcal{M})\Big]$$

$$V = \gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) d_{2\to2}(\mathcal{N}) + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) d_{2\to2}(\mathcal{M}) + \min\Big\{d_F(\mathcal{M}) \cdot d_{2\to2}(\mathcal{N}), d_F(\mathcal{N}) \cdot d_{2\to2}(\mathcal{M})\Big\}$$

$$U = d_{2\to2}(\mathcal{M})\, d_{2\to2}(\mathcal{N}).$$

Then, for $\epsilon > 0$, $\mathbb{P}\big(C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \geq c_1 W + \epsilon\big) \leq 2\exp\left(-c_2 \min\left\{\frac{\epsilon^2}{V^2}, \frac{\epsilon}{U}\right\}\right)$, *where the constants* c_1, c_2 *depend on the sub-gaussian parameter* L.

Remark 3.1. *Comparing Theorem 3.1 and 2.2 we observe that* W, V *and* U *in our case depend on product of geometric complexities of* $\mathcal{M}$ *and* $\mathcal{N}$ *and reduce to the terms in Theorem 2.2 when* $M = N$. *Interestingly the terms do not depend on the complexity of a joint set such as* $\{M^\top N : M \in \mathcal{M}, N \in \mathcal{N}\}$, *but rather on a symmetric combination of the individual complexities of* $\mathcal{M}$ *and* $\mathcal{N}$, *and is precisely because our analysis decouples them. This decoupled structure allows sharper bounds in applications where one set has small* γ_2 *functional or radii, even if the other is large.*

3.1 Proof sketch of Main Result

We outline the main steps in the proof of Theorem 3.1 along with the novel aspects of our analysis here. A detailed version and proofs of intermediate technical results are provided in Appendix D.

Our approach to getting a large deviation bound for $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ defined in (3) is based on bounding $\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$, where $\|X\|_{L_p}$ is the L_p norm of X defined as $\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p}$. Specifically, our objective is to show that $\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq a + \sqrt{p} \cdot b + p \cdot c\ , \forall p \geq 1$, where a, b, c are constants that do not depend on p, which using the moment-generating function and Markov's inequality [Williams, 1991, Vershynin, 2012] immediately implies $P(|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})| \geq a + u) \leq \exp\left\{-\min\left(\frac{u^2}{4b^2}, \frac{u}{2c}\right)\right\}$.

The proof consists of the following three main steps:

1. **Decomposing $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ into off-diagonal and diagonal terms.** We define the following terms corresponding to the off-diagonal and diagonal terms of $M^\top N$ respectively:

$$B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \triangleq \sup_{\substack{M\in\mathcal{M}\\N\in\mathcal{N}}} \Big|\sum_{\substack{j,k=1\\j\neq k}}^{n} \boldsymbol{\xi}_j\boldsymbol{\xi}_k\langle M_j, N_k\rangle\Big|,$$

$$D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \triangleq \sup_{\substack{M\in\mathcal{M}\\N\in\mathcal{N}}} \Big|\sum_{j=1}^{n} (|\boldsymbol{\xi}_j|^2 - \mathbb{E}\,|\boldsymbol{\xi}_j|^2)\langle M_j, N_j\rangle\Big|,$$

where M_i and N_i are the i-th row of M and N respectively. The next lemma relates the L_p norm of $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ to that of $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ and $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ using symmetrization [Vershynin, 2018].

Lemma 3.2. *For* $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ *and* $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ *as defined above, the following holds*

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq \|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} + \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$$

2. **Bounding the off-diagonal term.** To bound the L_p norm of $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ we consider $\boldsymbol{\xi}'$ to be an independent copy of $\boldsymbol{\xi}$ and use symmetrization (eg. Lemma 6.3 [Ledaux and Talagrand, 1991]) to get the following upper bound:

$$\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq 4\Big\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} |\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle|\ \Big\|_{L_p}.$$

Next, unlike Krahmer et al. [2014b] the inner product above contains two different matrices M and N, and therefore we consider two separate admissible sequences (cf. definition 2.2) $\{T_r(\mathcal{M})\}_{r=0}^{\infty}$ and $\{T_r(\mathcal{N})\}_{r=0}^{\infty}$ of $\mathcal{M}$ and $\mathcal{N}$ respectively. We then use a generic chaining argument by creating two separate increment sequences for $\mathcal{M}$ and $\mathcal{N}$ to show:

$$\Big\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\Big\|_{L_p} \leq \underbrace{\sup_{M\in\mathcal{M},N\in\mathcal{N}}\Big\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\Big\|_{L_p}}_{I} + \underbrace{\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) \cdot \Big\|\sup_{N\in\mathcal{N}}\|N\boldsymbol{\xi}'\|_2\Big\|_{L_p} + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) \cdot \Big\|\sup_{M\in\mathcal{M}}\|M\boldsymbol{\xi}\|_2\Big\|_{L_p}}_{II}$$

In term I, the supremum has been pulled outside the L_p-norm, which allows us to apply standard concentration bounds for sub-Gaussian random variables. Since $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are independent and sub-Gaussian, the quantity $\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\|_{L_p}$ can be bounded in terms of the operator norms of M and N. The appearance of the cross-terms in term II, arises specifically from the application of generic chaining

to $\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle$, where we create separate admissible sequences for $\mathcal{M}$ and $\mathcal{N}$. Specifically, the chaining decomposition of M contributes a sequence of approximations whose increments are paired with the worst-case realization over N, and vice versa. Finally $\|\sup_{N\in\mathcal{N}}\|N\boldsymbol{\xi}'\|_2\|_{L_p}$ can be handled using existing techniques in [Krahmer et al., 2014b] to upper bound the off-diagonal term as follows.

Theorem 3.3. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned}
&\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \\
&\lesssim \Big[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\Big(\gamma_2(\mathcal{N},\|\cdot\|_{2\to2}) + d_F(\mathcal{N}) + \\
&\qquad \sqrt{p}\; d_{2\to2}(\mathcal{N})\Big) \\
&+ \gamma_2(\mathcal{N},\|\cdot\|_{2\to2})\cdot\Big(\gamma_2(\mathcal{M},\|\cdot\|_{2\to2}) + d_F(\mathcal{M}) + \\
&\qquad \sqrt{p}\; d_{2\to2}(\mathcal{M})\Big) \\
&+ \sqrt{p}\;\min\Big\{d_F(\mathcal{M})\cdot d_{2\to2}(\mathcal{N}), d_F(\mathcal{N})\cdot d_{2\to2}(\mathcal{M})\Big\} \\
&+ p\cdot d_{2\to2}(\mathcal{M})\cdot d_{2\to2}(\mathcal{N})\Big].
\end{aligned}$$

3. **Bounding the diagonal term.** We use symmetrization ([Ledaux and Talagrand, 1991, Lemma 6.3]) and contraction ([Ledaux and Talagrand, 1991, Lemma 4.6]) to get

$$\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}))\|_{L_p} \leq \underbrace{\|D_{\mathcal{M},\mathcal{N}}(\mathbf{g})\|_{L_p}}_{I} + \underbrace{\Big\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\Big|\sum_{j=1}^{n}\varepsilon_j\langle M_j,N_j\rangle\Big|\Big\|_{L_p}}_{II}, \quad (5)$$

where $\mathbf{g}$ is a Gaussian random vector with i.i.d. entries $\mathbf{g}_j$ and $\{\varepsilon_j\}$ is a set of i.i.d. Rademacher variables independent of $\boldsymbol{\xi}$. Note that term I is the same diagonal term with the Gaussian random vector $\mathbf{g}$ instead of the sub-gaussian random vector $\boldsymbol{\xi}$, and is arguably easier to bound. However, because of the cross terms $\langle M_j, N_j\rangle$, we cannot use a standard decoupling argument as in Theorem 2.5 in Krahmer et al. [2014b]. Instead we we prove a stronger version of decoupling for Gaussian chaos that holds even for non-hermitian matrices to upper bound term I.

Bounding term II is even more challenging. In the single set case [Krahmer et al., 2014b] the term inside the sup reduces to $\sum_j \varepsilon_j\|M_j\|_2^2$ and is a sub-gaussian process relative to the metric $d(A,B) = (\sum_j \|A_j\|_2^2 - \|B_j\|_2^2)^{1/2}$ and therefore a standard chaining argument follows (see Proof of Theorem 3.5 in Krahmer et al. [2014b]). Since we cannot use such a technique, we device a new double tree argument for generic chaining in the following theorem.

Theorem 3.4. *Consider a stochastic process $\{X_{(u,v)}\}$ where $u \in \mathcal{U}$, $v \in \mathcal{V}$, and let d be the metric. Suppose for any $t > 0$, with probability at least $1 - c_0\exp\left(-\frac{t^2}{2}\right)$, $\{X_{(u,v)}\}$ satisfies*

$$\begin{aligned}
|X_{(u_1,v)} - X_{(u_2,v)}| &\leq C_u t\cdot d(u_1,u_2), \forall v\in\mathcal{V} \\
|X_{(u,v_1)} - X_{(u,v_2)}| &\leq C_v t\cdot d(v_1,v_2), \forall u\in\mathcal{U}.
\end{aligned}$$

Then, w.p. $1 - 2c_0\exp\left(-\frac{t^2}{2}\right)$, $\sup_{u\in\mathcal{U},\,v\in\mathcal{V}}|X_{u,v}| \leq 4\sqrt{2}t\left(C_u\gamma_2(\mathcal{U},d) + C_v\gamma_2(\mathcal{V},d)\right)$.

Remark 3.2 (Double Chaining Bound). *In Theorem 3.4 we consider a double indexed stochastic process X indexed by elements from sets $\mathcal{U}$ and $\mathcal{V}$. We assume that the process varies smoothly in each coordinate, i.e., for a fixed v, the process is Lipschitz in u, and vice versa, (with high probability) and provide a uniform deviation bound on the suprema of the double indexed process that scales with the sum of γ_2 functional of $\mathcal{U}$ and $\mathcal{V}$. The proof idea involves applying Generic Chaining by creating two separate chains (or trees) over $\mathcal{U}$ and $\mathcal{V}$ and relating it to the displacement in both the indices.*

The final result follows by defining the stochastic process $X_{(M,N)} = \left|\sum_{j=1}^n \mathbf{g}_j\langle M_j,N_j\rangle\right|$ and invoking Theorem 3.4 and is as given below.

Theorem 3.5. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned}
&\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \\
&\lesssim \Big[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\Big(\gamma_2(\mathcal{N},\|\cdot\|_{2\to2}) + d_F(\mathcal{N}) \\
&\qquad + \sqrt{p}\; d_{2\to2}(\mathcal{N})\Big) \\
&+ \gamma_2(\mathcal{N},\|\cdot\|_{2\to2})\cdot\Big(\gamma_2(\mathcal{M},\|\cdot\|_{2\to2}) + d_F(\mathcal{M}) \\
&\qquad + \sqrt{p}\; d_{2\to2}(\mathcal{M})\Big) \\
&+ \sqrt{p}\;\min\Big\{d_F(\mathcal{M})\cdot d_{2\to2}(\mathcal{N}), d_F(\mathcal{N})\cdot d_{2\to2}(\mathcal{M})\Big\} \\
&+ p\cdot d_{2\to2}(\mathcal{M})\cdot d_{2\to2}(\mathcal{N})\Big].
\end{aligned}$$

The proof of Theorem 3.1 now follows by combining the three steps (Lemma 3.2, Theorem 3.3 and Theorem 3.5) and collecting terms corresponding to W, V and U.

4 APPLICATIONS

4.1 John-Lindenstrauss Lemma

We show how Theorem 3.1 can be used to recover the J-L lemma [Johnson and Lindenstrauss, 1984]. The JLT is a randomized embedding that maps high-dimensional vectors into a lower-dimensional space

while approximately preserving pairwise Euclidean distances [Clarkson and Woodruff, 2013, Woodruff, 2014]. Let $X \in \mathbb{R}^{n\times p}$, with $n < p$, and let $\mathcal{A}$ be any set of N vectors in $\mathbb{R}^p$. We say that X satisfies the *Johnson–Lindenstrauss Transform (JLT)* if for every $\varepsilon > 0$,

$$(1-\varepsilon)\|u\|_2^2 \le \|Xu\|_2^2 \le (1+\varepsilon)\|u\|_2^2, \quad \text{for all } u \in \mathcal{A}.$$

In Appendix E we show that with appropriate choices of the matrices M and N we can ensure that a random matrix $\frac{1}{\sqrt{n}}\tilde{X}$ whose entries are i.i.d. standard gaussian would satisfy the JLT condition.

Proposition 4.1. *Let $X \in \mathbb{R}^{n\times p}$ be defined as $X = \frac{1}{\sqrt{n}}\tilde{X}$, where entries of $\tilde{X}$ are i.i.d. standard normal. If $n = \Omega(\varepsilon^{-2}\log N)$, then X is a Johnson–Lindenstrauss Transform (JLT) with probability at least $1 - \frac{1}{N^c}$, for some constant $c > 0$.*

4.2 Restricted Isometry Property (RIP)

A closely related property to the J-L lemma is the RIP. Let $X \in \mathbb{R}^{n\times p}$ and let $\mathcal{A}$ denote the collection of all s-sparse vectors in $\mathbb{R}^p$. We say that X satisfies the RIP with constant $\delta_s \in (0,1)$ if, for all $u \in \mathcal{A}$,

$$(1-\delta_s)\|u\|_2^2 \le \|Xu\|_2^2 \le (1+\delta_s)\|u\|_2^2. \tag{6}$$

Such matrices are of fundamental importance in high-dimensional statistics and compressed sensing, where the objective is to recover a sparse signal $\theta^* \in \mathbb{R}^p$ from a limited number of noisy linear measurements [Plan and Vershynin, 2013, Krahmer et al., 2014b]. We can extend this to preserving inner products as follows.

Let $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^p$ be two subsets. Consider the following condition: for every $\epsilon > 0$ we have

$$|\langle Xu, Xv\rangle - \langle u, v\rangle| \le \epsilon\|u\|_2\|v\|_2, \ \forall u \in \mathcal{U}, v \in \mathcal{V}. \tag{7}$$

For finite sets with $f = |\mathcal{U}| = |\mathcal{V}|$, Woodruff [2014] show that for $X \in \mathbb{R}^{n\times p}$, $X = \frac{1}{\sqrt{n}}\tilde{X}$, where the entries of $\tilde{X}$ are i.i.d. standard normal satisfies (7) for $n = \Omega(\varepsilon^{-2}\log(f/\delta))$ (see Theorem 4, Woodruff [2014]). Their analysis uses a standard concentration plus a union bound. We can significantly generalize this result using Theorem 3.1, by considering arbitrary sets $\mathcal{U}, \mathcal{V}$ in the following proposition (proof in Appendix E).

Proposition 4.2. *Consider arbitrary sets $\mathcal{U}$, $\mathcal{V}$ such that $\|u\|_2 = \|v\|_2 = 1$, for all $u \in \mathcal{U}, v \in \mathcal{V}$ and let $X \in \mathbb{R}^{n\times p}$ be defined as $X = \frac{1}{\sqrt{n}}\tilde{X}$, where the entries of $\tilde{X}$ are i.i.d. standard normal. Then for every $0 < \epsilon \le 1$, if $n = \Omega(\frac{1}{\epsilon^2}(\omega(\mathcal{U}) + \omega(\mathcal{V}))^2)$, then (7) holds with probability $1 - e^{-c(\omega(\mathcal{U})+\omega(\mathcal{V}))^2}$ for constant $c > 0$, where $\omega(\mathcal{T})$ is the Gaussian width of $\mathcal{T}$ (cf. Definition 2.3).*

For infinite sets, a standard approach is to discretize $\mathcal{U}, \mathcal{V}$ via multiscale coverings and control $\sup_{u\in\mathcal{U},\, v\in\mathcal{V}} |\langle Xu, Xv\rangle - \langle u, v\rangle|$ by concentration plus a union bound on the nets, yielding Dudley-type complexities through the entropy integral $\mathcal{D}(\mathcal{T}) = \int_0^{\operatorname{diam}(\mathcal{T})} \sqrt{\log N(\mathcal{T}, \tau)}\, \mathrm{d}\tau$ (cf. Sarlos [2006]; Woodruff [2014]). Note that, $\omega(\mathcal{T}) \asymp \gamma_2(\mathcal{T}, \|\cdot\|_2)$ (see Talagrand [2014]), whereas $\mathcal{D}(\mathcal{T})$ is an upper bound for γ_2, and there exist sets where $\mathcal{D}(\mathcal{T})$ is not sharp [Talagrand, 2005, 2014]. On such sets, our Gaussian-width sample complexity $n \gtrsim \varepsilon^{-2}(\omega(\mathcal{U}) + \omega(\mathcal{V}))^2$ is strictly sharper, while on finite sets we recover $\omega(\mathcal{T}) \asymp \sqrt{\log|\mathcal{T}|}$ and thus match the classical rates of Woodruff [2014] and the cases analyzed by Sarlos [2006].

4.3 Sketching-based Distributed Learning

We consider the setup of Sketching-based Distributed Learning (sketch-DL) as described in [Shrivastava et al., 2024, Song et al., 2023]. For a complete treatment, we refer the reader to the Appendix F.

Consider a distributed learning framework. Each client $c \in C$ has access to a local dataset $\mathcal{D}_c := \{x_{i,c}, y_{i,c}\}_{i=1}^N$. The goal is to learn $\theta \in \mathbb{R}^p$ that minimizes the joint loss function: $\mathcal{L}(\theta) = \frac{1}{C}\sum_{c=1}^C \mathcal{L}_c(\theta)$, where $\mathcal{L}_c(\theta) = \frac{1}{N}\sum_{i=1}^n \ell(f(\theta, \mathbf{x}_{i,c}), y_{i,c})$, with $\ell(\cdot,\cdot)$ being some loss function and f the output of a deep learning model.

The sketch-DL algorithm (see Appendix F) proceeds as follows: in each round, every client computes local gradients and sends their *sketched* (`sk`) versions to the server. The server aggregates these `sk`-gradients and broadcasts the result. Each client then *desketches* (`desk`) the aggregated gradient and uses it to update its local model. The `sk` and `desk` operations reduce communication by mapping p-dimensional gradients to a lower b-dimensional space (with $b \ll p$) via a shared sketching matrix. The sketch de-sketch operations are $\texttt{sk}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^b := R\mathbf{x}, \texttt{desk}(\mathbf{x}) : \mathbb{R}^b \to \mathbb{R}^p := R^\top\mathbf{x}$.

The objective is to provide provide finite time convergence guarantee by bounding the loss difference $\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*)$ after T rounds, where $\theta^* = \operatorname{argmin}\mathcal{L}(\theta)$. We make the following assumptions on the loss function $\mathcal{L}$ and the loss Hessian $\mathbf{H}$.

Assumption 2 (PL condition). *The loss function $\mathcal{L}(\cdot)$ satisfies the Polyak–Łojasiewicz (PL) condition with constant μ, i.e. $\|\nabla\mathcal{L}(\mathbf{w})\|^2 \ge 2\mu(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_*))$, where $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}}\mathcal{L}(\mathbf{w})$.*

Assumption 3 (Anistropic Loss Hessian). *For any loss Hessian $H_{i,c,t}$, assume there exists a fixed positive definite $\mathbf{H}$ such that $-\mathbf{H} \preceq H_{i,c,t} \preceq \mathbf{H}$. Let $\Lambda_1, \ldots, \Lambda_p$ be the eigenvalues of $\mathbf{H}$ and define $\Lambda_{\max} = \max_j|\Lambda_j|$. Then there exists a constant $\kappa = \mathcal{O}(1)$ such that $\sum_{j=1}^p|\Lambda_j| \le \kappa\,\Lambda_{\max}$.*

With the above assumptions, we can state the convergence guarantee of K-step sketch-DL. See Appendix F for precise statement along with the proof.

Theorem 4.1. *Suppose Assumptions 2 and 3 hold and let* $\|\nabla_\theta \ell(\theta)\| \le G$. *For suitable constant* $\varepsilon < 1$, *width of the network* m, *learning rate* η, *sketching dimension* $b = \Omega\Big(\frac{1}{\varepsilon^2}\ \text{polylog}\big(\frac{TNp^2}{\delta}\big)\Big)$, *and* $C_2(m,\kappa) := \mathcal{O}\Big(\frac{\varepsilon\kappa}{\sqrt{m}} + \frac{1}{\sqrt{m}}\Big)$, *with probability at least* $1-\delta$, *we have*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \le \big(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*)\big)\, e^{-2\mu\eta KT} + \frac{(\eta K C_2(m,\kappa) + C_3)\,(G^2+\epsilon^2)}{2\mu}. \quad (8)$$

Remark 4.1. *Our bound relies on Gaussian width of two sets: the predictor gradients* $\mathcal{G} := \{\nabla f(\theta_{c,i,t})\}$ *and the loss-Hessian eigenvectors* $\mathcal{H} := \{v_j\}$. *Using Theorem 3.1 we bound* $\sup_{g\in\mathcal{G},h\in\mathcal{H}} |g^\top R^\top R h - g^\top h| \lesssim Z(\mathcal{G},\mathcal{H},\delta)$ *in the analysis, where* $Z(\mathcal{G},\mathcal{H},\delta)$ *is a function of geometric complexities of* $\mathcal{G}$ *and* $\mathcal{H}$ *(see Appendix F for details).By combining these with the gaussian-width estimates of Banerjee et al. [2024], we obtain the stated inequality. In contrast to Shrivastava et al. [2024], who introduce a fresh sketching matrix at every iteration to avoid dependence with the gradient set, our approach allows reusing a single sketching matrix throughout. This is permissible because we bound the error directly via the Gaussian width of the predictor-gradient set, which falls outside the finite-set framework employed in their analysis.*

4.4 Linear Regression with Sketching

We observe n i.i.d. samples $(\mathbf{x}_i, y_i)$ from the linear model $y_i = \mathbf{x}_i^\top \beta^* + \varepsilon_i$, where $\mathbf{x}_i \in \mathbb{R}^d$ are sub-Gaussian covariates with covariance Σ and ε_i are independent sub-Gaussian noise terms. Further $\beta \in \mathcal{B} \subseteq \mathbb{R}^d$. We solve a least squares problem to compute $\hat{\beta}^s$ using the sketched inputs and corresponding responses, i.e., using $(\mathbf{S}\mathbf{x}_i, y_i)$. Subsequently we de-sketch the estimate using $\mathbf{S}^\top \hat{\beta}^s$ and use this to make predictions. Note that solving the least squares problem in a lower dimensional sketched space is computationally faster. Our goal is to bound the error $(\mathbf{S}^\top \hat{\beta}^s - \beta^*)^\top \mathbf{x}$, for $\mathbf{x} \in \mathcal{X}$. In Appendix G we show that one can provide error bounds that do not depend on the ambient dimension d but on geometric complexities of $\mathcal{B}$ and $\mathcal{X}$.

5 EXTENSIONS

We extend our results to two important settings. The first is the conditional independence setting, where we assume that the entries of the random vector $\boldsymbol{\xi}$ are conditionally independent. This extension lets us provide uniform bounds for Countsketch [Charikar et al., 2002] and SRHT [Ailon and Chazelle, 2009, Tropp, 2011] matrices. Next, we provide bounds for sum of T random quadratic forms and show that the bound scales as $\sqrt{T}$. A standard Hoeffding type argument does not apply here because of the sup, and such a generic chaining analysis does not exist even for the single set case.

5.1 Conditional Independence

In this section we consider a conditionally independent structure among the entries of the random vector $\boldsymbol{\xi}$. Specifically we relax Assumption 1 to the following.

Assumption 4. *Let* $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n\}$ *be a sub-Gaussian stochastic process which is decoupled when conditioned on another stochastic process* $F = \{F_i\} = \{F_1, \dots, F_n\}$.

(i) for each $i = 1, \dots, n$, $\boldsymbol{\xi}_i \mid f_{1:i}$ *is a zero-mean* L *sub-Gaussian for all realizations* $f_{1:i}$ *of* $F_{1:i}$.

(ii) for each $i = 1, \dots, n$, *there exists an index* $\varrho(i) \le i$ *which is non-decreasing, i.e.,* $\varrho(j) \le \varrho(i)$ *for* $j < i$, *such that* $\boldsymbol{\xi}_i \perp \boldsymbol{\xi}_j \mid F_{1:\varrho(i)}$ *for* $j < i$ *and* $\boldsymbol{\xi}_i \perp F_k \mid F_{1:\varrho(i)}$ *for* $k > \varrho(i)$.

Theorem 5.1. *Let* $\mathcal{M}$ *and* $\mathcal{N}$ *be set of* $(m\times n)$ *and* $(n\times m)$ *matrices respectively, and let* $\boldsymbol{\xi}$ *be a random vector satisfying Assumption 4. Let* W, V, U *be as defined in Theorem 3.1. Then, for* $\epsilon > 0$, $\mathbb{P}\big(C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \ge c_1 W + \epsilon\big) \le 2\exp\left(-c_2 \min\left\{\frac{\epsilon^2}{V^2}, \frac{\epsilon}{U}\right\}\right)$, *where the constants* c_1, c_2 *depend on the sub-gaussian parameter* L.

Remark 5.1. *Theorem 5.1 is more general than the result in Theorem 3.1, and is proved in Appendix H. Banerjee et al. [2019] show that it covers many dependence structures, including RIP for Toeplitz matrices and CountSketch. Theorem 5.1 extends those results to the bi-linear setup. Further in Appendix H.1 we show that SRHT sketch matrix also satisfies Assumption 4.*

5.2 Sum of Random Quadratic Forms

In this section we consider sum of random quadratic forms. Specifically, we want to develop large deviation bound for the following random variable.

$$C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) \triangleq \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big|\sum_{t=1}^{T} (\boldsymbol{\xi}_t^\top M^\top N \boldsymbol{\xi}_t) - \mathbb{E}(\boldsymbol{\xi}_t^\top M^\top N \boldsymbol{\xi}_t)\Big| \quad (9)$$

Theorem 5.2. *Let* $\mathcal{M}$ *and* $\mathcal{N}$ *be set of* $(m\times n)$ *and* $(n\times m)$ *matrices respectively, and* $\boldsymbol{\xi}_t$, $t \in [T]$ *be i.i.d. random vectors satisfying Assumption 1. Further let* W, V *and* U *be as defined in Theorem 3.1; then, for* $\epsilon > 0$, $\mathbb{P}\left(C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) \ge c_1 W\sqrt{T} + \epsilon\right) \le \exp\left(-c_2 \min\left\{\frac{\epsilon^2}{TV^2}, \frac{\epsilon}{\sqrt{T}\,U}\right\}\right)$, *where the constants* c_1, c_2 *depend only on the sub-gaussian parameter* L.

Algorithm 1 sk-LinUCB (Sketched Linear UCB)

1: **for** $t = 1, 2, ..., T$ **do**
2: Solve the least squares regression problem:

$$\hat{\theta}_t^s = \underset{\theta \in \mathbb{R}^b, \|\theta\|_2 \le 1}{\operatorname{argmin}} \sum_{i=1}^{t-1} (\langle \theta, \mathcal{S}_t a_i \rangle - r_i)^2 + \lambda \|\theta\|_2 \quad (10)$$

3: Construct the Confidence set $\mathcal{C}_t^s$ as in (11).
4: Compute the the optimistic estimates: $(\tilde{\theta}_t^s, a_t^s) = \operatorname{argmax}_{\theta \in \mathcal{C}_t^s, a \in \mathcal{S}_t \mathcal{A}} \langle \theta, a \rangle$
5: De-sketch and play the action $a_t = \mathcal{S}_t^\top a_t^s$; observe the reward r_t.
6: **end for**

5.3 Sketched Contextual Bandits

In this section, we present a simplified bandit setup to illustrate a concrete setting where Theorem 5.2 naturally applies. We briefly describe our results here and have pushed detailed descriptions to Appendx J. In a bandit problem Auer et al. [2002b], Lattimore and Szepesvári [2020] a learner needs to make sequential decisions over T time steps. We consider a simplified setup where at any round $t \in [T]$, the learner picks an action a_t from the action set $\mathcal{A}$, and then the associated reward of the arm $r(a_t) \in [0, 1]$ is observed. We make the following assumption on the reward.

Assumption 5 (**Linear Reward**). *The reward $r(a_t)$ is given by $r(a_t) = \langle a_t, \theta^* \rangle + \eta_t$, where $\theta^* \in \Theta^*$ is an unknown parameter and η_t is a conditionally sub-Gaussian noise, i.e., $\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda \eta_t} | a_1, \dots a_t, \eta_1, \dots, \eta_{t-1}] \le \exp(\frac{\lambda^2}{2})$. Further $\|a\|_2 = 1$ for all $a \in \mathcal{A}$.*

Definition 5.1 (**Regret**). *For actions $a_t, t \in [T]$, and $a^* = \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta^*, a \rangle$, the learner wants to minimize the regret defined as*

$$\texttt{Reg}_{\texttt{CB}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} (r(a^*) - r(a_t)) \right].$$

We develop a sketched version of the popular algorithm LinUCB (Linear Upper Confidence Bound [Abbasi-Yadkori et al., 2011]) and is summarized in Algorithm 1. At every round t the learner sketches the inputs using $\mathcal{S}_t \mathbb{R}^{b \times d}$ whose entries are drawn i.i.d. from $N(0, 1/b)$. It then solves a regularized least-squares problem (cf (10) in Algorithm 1) in the b-dimensional sketched space to obtain an estimate $\hat{\theta}_t^s$ of the unknown parameter. Subsequently, with $\bar{V}_t^s = \sum_{i=1}^{t} a_i^s a_i^{s\top} + \lambda I$, the learner constructs a confidence set $\mathcal{C}_t^s$ around the estimate:

$$\mathcal{C}_t^s = \Bigg\{ \theta \in \mathbb{R}^b : \|\hat{\theta}_t^s - \mathcal{S}_t \theta^*\|_{\bar{V}_t^s} \le \sqrt{\log\Big(\frac{\det(\bar{V}_t^s)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \Big)} + \lambda^{1/2} \|\mathcal{S}_t \theta_*\|_2 \Bigg\} \quad (11)$$

The algorithm then uses this confidence set to compute an *optimistic* parameter and action pair $(\tilde{\theta}_t^s, a_t^s)$ (Line 4) where $\tilde{\theta}_t^s \in \mathcal{C}_t^s$ and the action is in the *sketched* action space $a \in \mathcal{S}_t \mathcal{A}$. Once this sketched action is identified, it is "*de-sketched*" (Line 5) to recover the corresponding action a_t in the original space. Finally, the chosen action a_t is played, and the observed reward r_t is used in subsequent rounds to refine future estimates. Our primary result in this section is the following decomposition for the regret.

Theorem 5.3 (**(Informal) Regret Decomposition for** sk-LinUCB). *Suppose Assumption 5 holds and the de-sketched actions selected by Algorithm 1 are in $\mathcal{A}$. Then with high probability*

$$\texttt{Reg}_{\texttt{CB}}(T) = \underbrace{\tilde{O}\left(\sqrt{bT} \left[\sqrt{b} + \frac{1}{\sqrt{b}} \omega(\Theta_*) \right] \right)}_{I} + \underbrace{\sum_{t=1}^{T} \theta^{*\top} (I - \mathcal{S}_t^\top \mathcal{S}_t) a^*}_{II}$$

where $\omega(\Theta_*)$, is the Gaussian width of the set Θ_* (cf. Definition 2.3).

Remark 5.2. *Term I captures the regret in the sketched b dimensional space while term II captures the restricted isometry term due to random sketching. Term II is exactly in the bilinear form that involves a sum of T independent sketch matrices. Therefore with suitable choices of $\mathcal{M}$ and $\mathcal{N}$ we can use Theorem 5.2 to bound term II. We also develop a similar sketched version of Thompson sampling. We report the full results and associated details in Appendix J.*

6 CONCLUSION

We presented a unified theory for analyzing uniform deviation bounds of *sketched bilinear forms*, extending classical results on random quadratic forms to the setting of cross-inner products over pairs of structured sets. Our approach introduces new chaining techniques for controlling suprema over product spaces, and provides tight control in terms of the geometric complexity of the underlying sets. We also developed a uniform bound for the sum of random quadratic forms with i.i.d. random vectors, showing that the deviation scales as $\sqrt{T}$; notably, such a bound does not exist even for the single-set case considered in prior works. We apply our results on several ML problems and derive improved bounds. Our work highlights several scenarios in which sketched bilinear forms arise in modern ML algorithms and opens up several future directions, including extensions to adaptive sketches, analysis of non-linear predictors.

ACKNOWLEDGEMENTS

The work was supported by the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award. Compute support for the work was provided by the National Center for Supercomputing Applications (NCSA) and the Illinois Campus Cluster Program (ICCP).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *STOC*, 2006.

Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009. doi: 10.1137/060673096.

Alexandr Andoni, Chengyu Lin, Ying Sheng, Peilin Zhong, and Ruiqi Zhong. Subspace embedding and linear regression with orlicz norm. In *International Conference on Machine Learning (ICML)*, pages 224–233. PMLR, 2018.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002a. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002b. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Yikun Ban, Yuchen Yan, Arindam Banerjee, and Jingrui He. EE-net: Exploitation-exploration neural networks in contextual bandits. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=X_ch3VrNSRg.

A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization, 2015. URL https://arxiv.org/abs/1505.02294.

Arindam Banerjee, Qilong Gu, Vidyashankar Sivakumar, and Steven Z Wu. Random quadratic forms with dependence: Applications to restricted isometry and beyond. *Advances in Neural Information Processing Systems*, 32, 2019.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Misha Belkin. Restricted strong convexity of deep learning models with smooth activations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PINRbk7h01.

Arindam Banerjee, Qiaobo Li, and Yingxue Zhou. Loss gradient gaussian width based generalization and optimization guarantees. *arXiv preprint arXiv:2406.07712*, 2024.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *Journal of Machine Learning Research*, 22(133):1–49, 2021. URL http://jmlr.org/papers/v22/18-863.html.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Christos Boutsidis and David P. Woodruff. Optimal CUR matrix decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC, pages 353–362, 2014.

Christos Boutsidis, David P. Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 236–249, 2016.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming (ICALP)*, pages 693–703. Springer, 2002. doi: 10.1007/3-540-45465-9_59.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13*, pages 81–90, Palo Alto, CA, USA, June 1–4 2013.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference Computational Learning Theory*, 2008. URL https://api.semanticscholar.org/CorpusID:9134969.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5ep85sakT3.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SThJXvucjQ.

Yichuan Deng, Zhao Song, Yitan Wang, and Yuanyuan Yang. A nearly optimal size coreset algorithm with nearly linear time. *arXiv preprint arXiv:2210.08361*, 2022.

Hossein Esfandiari, Vahab Mirrokni, and Peilin Zhong. Almost linear time density level set estimation via dbscan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, 2021.

Yeqi Gao, Lianke Qin, Zhao Song, and Yitan Wang. A sublinear adversarial training algorithm. *arXiv preprint*, 2022.

D. L. Hanson and E. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics*, 42:1079–1083, 1971.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-efficient distributed sgd with sketching, 2020. URL https://arxiv.org/abs/1903.04488.

Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu wai Wong. An improved cutting plane method for convex optimization, convex-concave games and its applications. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2020.

Jiawei Jiang, Fangcheng Fu, Tong Yang, and Bin Cui. Sketchml: Accelerating distributed machine learning with data sketches. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1269–1284, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450347037. doi: 10.1145/3183713.3196894. URL https://doi.org/10.1145/3183713.3196894.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. Faster dynamic matrix inverse for faster lps. *arXiv preprint*, 2021. Presented at STOC.

William Johnson and Joram Lindenstrauss. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 01 1984.

F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014a.

Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014b.

Ilja Kuzborskij, Leonardo Cella, and Nicolò Cesa-Bianchi. Efficient linear bandits through matrix sketching. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 177–185. PMLR, 16–18 Apr 2019a. URL https://proceedings.mlr.press/v89/kuzborskij19a.html.

Ilja Kuzborskij, Leonardo Cella, and Nicolò Cesa-Bianchi. Efficient linear bandits through matrix sketching. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 177–185. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/kuzborskij19a.html.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

M. Ledaux and M. Talagrand. *Probability in Banach Spaces: Isometry and Processes*. Springer, 1991.

Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. In *Conference on Learning Theory*, COLT, 2019.

Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for the analysis of regularized M-estimators. *Statistical Science*, 27(4): 538–557, 2012.

Jelani Nelson and Huy L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 117–126. IEEE, 2013.

Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

Lianke Qin, Zhao Song, Lichen Zhang, and Danyang Zhuo. An online and unified algorithm for projection matrix vector multiplication with application to empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 101–156. PMLR, 2023.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2): 395–411, May 2010. ISSN 0364-765X. doi: 10.1287/moor.1100.0446. URL https://doi.org/10.1287/moor.1100.0446.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152, 2006. doi: 10.1109/FOCS.2006.37.

Anshumali Shrivastava, Zhao Song, and Zhaozhuo Xu. A tale of two efficient value iteration algorithms for solving linear mdps with large action space. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Mayank Shrivastava, Berivan Isik, Qiaobo Li, Sanmi Koyejo, and Arindam Banerjee. Sketching for distributed deep learning: A sharper analysis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0G0VpMjKyV.

Zhao Song and Zheng Yu. Oblivious sketching-based central path method for solving linear programming problems. In *38th International Conference on Machine Learning (ICML)*, 2021.

Zhao Song, David P. Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual Symposium on the Theory of Computing (STOC)*, 2017.

Zhao Song, David P. Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the 30th Annual ACM–SIAM Symposium on Discrete Algorithms*, SODA, 2019.

Zhao Song, Shuo Yang, and Ruizhe Zhang. Does preprocessing help training over-parameterized neural networks? *Advances in Neural Information Processing Systems*, 34, 2021a.

Zhao Song, Lichen Zhang, and Ruizhe Zhang. Training multi-layer over-parametrized neural network in subquadratic time. *arXiv preprint*, 2021b.

Zhao Song, Yitan Wang, Zheng Yu, and Lichen Zhang. Sketching for first order method: efficient algorithm for low-bandwidth channel and vulnerability. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

M. Talagrand. *The Generic Chaining*. Springer, 2005.

M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Applied and Computational Harmonic Analysis*, 2011. doi: 10.1142/S1793536911000787. Preprint: arXiv:1011.1595.

Joel A. Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4): 1454–1485, January 2017. ISSN 1095-7162. doi: 10.1137/17m1111590. URL http://dx.doi.org/10.1137/17M1111590.

Ramon van Handel. Chaining, interpolation, and convexity, 2016. arXiv version.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, chapter 5, pages 210–268. Cambridge University Press, 2012.

R. Vershynin. *Estimation in high dimensions: A geometric perspective*, pages 3–66. Springer International Publishing, Cham, 2014.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Ruosong Wang, Peilin Zhong, Simon S. Du, Russ R. Salakhutdinov, and Lin F. Yang. Planning with general objective functions: Going beyond total rewards. In *Advances in Neural Information Processing Systems*, NeurIPS, 2020a.

Ruosong Wang, Peilin Zhong, Simon S. Du, Russ R. Salakhutdinov, and Lin F. Yang. Planning with general objective functions: Going beyond total rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, October 2014.

David P. Woodruff and Peilin Zhong. Distributed low rank approximation of implicit functions of a matrix. In *32nd IEEE International Conference on Data Engineering*, ICDE, pages 847–858, 2016.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration.

In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Ingvar Ziemann. A vector bernstein inequality for self-normalized martingales. *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://openreview.net/forum?id=4ZJjr9YbBw. Accepted by TMLR, Published: 27 Mar 2025.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes, see Section 3
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes, see Section 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Yes, see Link for code.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes, see Section 3
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes, see Section 3, Appendix D
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes, see Section 3

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes, see Link for code.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Not Applicable
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable

Beyond Johnson-Lindenstrauss: Uniform Bounds for Sketched Bilinear Forms

A RELATED WORKS

Sketching. Sketching is a key dimensionality reduction technique with applications in machine learning, including federated learning [Ivkin et al., 2020] and low-rank approximation [Tropp et al., 2017]. The linearity property of sketching makes it an attractive tool in communication efficient training [Song et al., 2023, Jiang et al., 2018]. Matrix sketching techniques have also been applied to linear bandits to improve computational efficiency. using online sketching technique, Kuzborskij et al. [2019b] reduce the update time for OFUL and Thompson Sampling from $O(d^2)$ to $O(md)$. Randomized sketching has become a unifying primitive for fast numerical linear-algebra. Clarkson and Woodruff [2013] first showed that a sparse subspace embedding lets one obtain $(1+\varepsilon)$ low-rank approximation and over-constrained least-squares solvers in time nearly proportional to $\mathrm{nnz}(A)$. Their idea was sharpened in several directions: OSNAP matrices of Nelson and Nguyên [2013] make the embedding *ultra-sparse* (constant non-zeros per column) while guaranteeing subspace preservation; Meng and Mahoney [2013] proved that the same input-sparsity bounds extend to low-distortion embeddings for *robust* regression; and Boutsidis and Woodruff [2014] established optimal relative-error CUR decompositions within the same computational budget. Follow-up work moved beyond ℓ_2 objectives—Song et al. [2017] give the first entrywise ℓ_1 low-rank approximation—and beyond the centralized RAM model: Woodruff and Zhong [2016] develop communication-optimal sketches for distributed low-rank approximation of implicit matrix functions, while Boutsidis et al. [2016] match information-theoretic limits for distributed and streaming PCA.

Song et al. [2019] extended sketching guarantees to tensor low-rank approximation, opening the door to high-order data analysis. In data mining and clustering, sketch-based algorithms accelerate density-level-set estimation (Esfandiari et al., 2021) and yield nearly-optimal coresets for (k, z)-clustering (Deng et al., 2022). Sketching has also entered decision-making: Wang et al. [2020a] embed planning with general objective functions, and Shrivastava et al. [2023] design value-iteration routines for linear MDPs whose runtime is sublinear in the action space. On the optimization side, leverage-score maintenance lets Lee et al. [2019] solve empirical-risk-minimization problems in *matrix-multiplication* time; fast inverse maintenance from Jiang et al. [2021] accelerates interior-point and cutting-plane methods, complemented by the oblivious central-path sketch of Song and Yu [2021] and the multi-layered cutting-plane scheme of Jiang et al. [2020]. Finally, Qin et al. [2023] provide an online, unified framework for projection matrix–vector multiplication, yielding faster online ERM with either data-oblivious or adaptive sketches.

Contextual Bandits. Contextual bandits generalize multi-armed bandits by incorporating feature vectors, allowing decisions to be conditioned on context. The problem has been widely studied under linear payoffs [Chu et al., 2011, Abbasi-Yadkori et al., 2011, Bietti et al., 2021, Lattimore and Szepesvári, 2020]. Linear stochastic bandits have been extensively studied as a fundamental setting for decision-making under uncertainty. Auer et al. [2002a] introduced an early upper confidence bound (UCB) algorithm for linear bandits, with subsequent improvements in regret bounds by Dani et al. [2008] and Rusmevichientong and Tsitsiklis [2010]. The widely used OFUL (Optimism in the Face of Uncertainty Learning) algorithm (also called LinUCB) Abbasi-Yadkori et al. [2011] provides an $\mathcal{O}(d\sqrt{T})$ regret under the assumption that the action and parameter sets are d-dimensional Euclidean balls. Agrawal and Goyal [2013] extended Thompson Sampling to contextual bandits with linear payoffs, proving a regret bound of $\tilde{O}(d^{3/2}\sqrt{T})$. More recently, neural contextual bandits have been proposed to handle non-linearity in reward functions. Zhou et al. [2020] use neural networks with UCB-based exploration and achieve near-optimal $\tilde{O}(\sqrt{T})$ regret. Ban et al. [2022] introduce a novel exploration strategy using an additional neural network to estimate potential gains, outperforming traditional linear bandit baselines, while Deb et al. [2024, 2025] extend the inverse gap weighting idea with neural networks.

B COVERING ARGUMENT (DUDLEY ENTROPY INTEGRAL) VS "GENERIC" CHAINING

Specifically, consider bounding $\mathbb{E}\sup_{t\in T} X_t$ for a Gaussian process $(X_t)_{t\in T}$. There are two different arguments for bounding the expectation over the supremum of the set T.

1. **Covering Argument (Dudley entropy integral).** This goes by a few different names—"covering argument", "ϵ-net", "Dudley's entropy integral" and just "chaining". For this exposition we will refer to this as entropy integral. This approach uses the covering argument by building a finite cover over T. The sharpest way of doing this is Dudley's entropy integral, which bounds $\sup_{t\in T} X_t$ by covering the index set T with increasingly fine ϵ-nets (dyadic nets) and gives the following bound:

$$\boxed{\mathbb{E}\big[\sup_{t\in T} X_t\big] \lesssim \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon} \tag{12}$$

   where (T,d) is the metric space and $N(T,d,\varepsilon)$ is the covering number of T.

2. **Generic Chaining.** Subsequently, Generic Chaining was developed to optimize over all admissible hierarchical partitions, and not just dyadic nets. This led to a precise relationship between the "size" of the process and the "size" of the metric space, thus obtaining:

$$\boxed{\mathbb{E}\big[\sup_{t\in T} X_t\big] \lesssim \gamma_2(T,d)} \tag{13}$$

   where Talagrand's γ_2 functional is defined in Definition 2 of our paper. Specifically for the case when the metric $d = \|\cdot\|_2$, $\gamma_2(T,d) = \Theta\big(w(T)\big)$, where $w(T)$ is the Gaussian width and $\Theta(\cdot)$ hides absolute constants; see, e.g., [Vershynin, 2018, Ch. 8].

B.1 Generic Chaining is sharper than Covering using Entropy Integral

There are sets where the generic chaining bound in (12) is sharper than the entropy integral bound in (13), but there are no sets where the entropy integral is sharper than generic chaining.

Before we give references regarding the sharpness of generic chaining for various sets, we give an explicit example.

Example. Let e_i be the i-th canonical basis vector in $\mathbb{R}^n$ and suppose the metric $d = \|\cdot\|_2$. Consider the following set:

$$T_n = \Big\{ t_k := \frac{e_k}{\sqrt{1+\log k}} : k = 2,\dots,n \Big\}.$$

First note that $\gamma_2(T_n, \|\cdot\|_2) \lesssim w(T_n)$ (see, e.g., [Vershynin, 2018, Cor. 8.5.8]), where $w(T_n)$ is the Gaussian width of T_n defined as $w(T_n) = \mathbb{E}\big[\sup_{t\in T_n}\langle g,t\rangle\big]$, $g_k \sim \mathcal{N}(0,1)$.

1. **Generic Chaining Bound.** For a standard Gaussian vector $g = (g_1,\dots,g_n)$ set

$$X := \sup_{t\in T_n}\langle g,t\rangle = \max_{1\le k\le n} \frac{g_k}{\sqrt{1+\log k}}.$$

   Fix $u \ge 0$. Using $\mathbb{P}\{g_k \ge v\} \le e^{-v^2/2}$ we get

$$\mathbb{P}\big\{X \ge u\big\} = \mathbb{P}\Big\{\max_k g_k \ge u\sqrt{1+\log k}\Big\} \le \sum_{k=1}^n \exp\Big(-\frac{u^2}{2}[1+\log k]\Big) = e^{-u^2/2}\sum_{k=1}^n k^{-u^2/2}.$$

   For $u \ge \sqrt{3}$ we have $s := u^2/2 \ge 3/2 > 1$. Then

$$\sum_{k=1}^\infty k^{-s} \le 1 + \int_1^\infty x^{-s}\,dx = 1 + \frac{1}{s-1} \le 3,$$

so for every $n \geq 1$ we have $\mathbb{P}\{X \geq u\} \leq 3e^{-u^2/2}$ for $u \geq \sqrt{3}$. The identity $\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq u\}\, du$ gives

$$\begin{aligned} w(T_n) = \mathbb{E}X &= \int_0^{\sqrt{3}} \mathbb{P}\{X \geq u\}\, du + \int_{\sqrt{3}}^{\infty} \mathbb{P}\{X \geq u\}\, du \\ &\leq \sqrt{3} + 3\int_{\sqrt{3}}^{\infty} e^{-u^2/2}\, du. \end{aligned}$$

Using $\displaystyle\int_a^\infty e^{-u^2/2}\, du \leq \frac{1}{a}\, e^{-a^2/2}$ with $a = \sqrt{3}$ we obtain

$$w(T_n) \leq \sqrt{3} + \frac{3}{\sqrt{3}}\, e^{-3/2} \leq 4.$$

Therefore $\gamma_2(T_n, \|\cdot\|_2) = \Theta\big(w(T_n)\big) = \Theta(1)$.

2. **Entropy Integral.** We claim

$$I_{\text{Dudley}}(T_n) = \int_0^\infty \sqrt{\ln N(T_n, \|\cdot\|_2, \varepsilon)}\, d\varepsilon \;\gtrsim\; \frac{1}{2}\ln\ln n.$$

For $m \in [2, n]$ set

$$\varepsilon_m = \frac{1}{\sqrt{1 + \ln m}}.$$

For $2 \leq k < m$,

$$\|v_k - v_m\|_2^2 = \frac{1}{1+\ln k} + \frac{1}{1+\ln m} \geq \frac{1}{1+\ln k} \geq \frac{1}{1+\ln m} = \varepsilon_m^2,$$

so $\|v_k - v_m\|_2 \geq \varepsilon_m$. Thus the $m-1$ points $\{v_2, \ldots, v_m\}$ are pairwise ε_m-separated, giving

$$N(T_n, \|\cdot\|_2, \varepsilon_m) \geq m - 1 = O(m), \qquad \ln N(T_n, \|\cdot\|_2, \varepsilon_m) \geq \ln m - O(1).$$

Note

$$\varepsilon_m - \varepsilon_{m+1} = \frac{1}{\sqrt{1+\ln m}} - \frac{1}{\sqrt{1+\ln(m+1)}} \approx -\frac{1}{2m(1+\ln m)^{3/2}} \;\geq\; \frac{c}{m(\ln m)^{3/2}}$$

for some $c > 0$. On $[\varepsilon_{m+1}, \varepsilon_m]$ we have $\sqrt{\ln N(T_n, \varepsilon)} \gtrsim \sqrt{\ln m}$. Therefore

$$I_{\text{Dudley}}(T_n) \geq \sum_{m=2}^{n-1} \sqrt{\ln m}\,(\varepsilon_m - \varepsilon_{m+1}) \geq c \sum_{m=2}^{n-1} \frac{\sqrt{\ln m}}{m(\ln m)^{3/2}} = c \sum_{m=2}^{n-1} \frac{1}{m \ln m}.$$

Using $\sum_{m=2}^{n-1} \frac{1}{m\ln m} \sim \ln\ln n$ gives $I_{\text{Dudley}}(T_n) \gtrsim \ln\ln n$. Note that $I_{\text{Dudley}}(T_n) \to \infty$ as $n \to \infty$.

The above example shows that the Generic Chaining bound is sharper than the entropic integral bound. *In general the separation can have a* log *dependence.* For example, consider an integer m and an i.i.d. standard Gaussian sequence $(g_i)_{i \leq m}$. For $t = (t_i)_{i \leq m} \in \mathbb{R}^m$, let $X_t = \sum_{i \leq m} t_i g_i$. Consider the set

$$T = \left\{ (t_i)_{i \leq m} \in \mathbb{R}^m \;:\; t_i \geq 0,\; \sum_{i \leq m} t_i = 1 \right\},$$

the convex hull of the canonical basis. It can be shown (see [Talagrand, 2014, p. 37]) that

$$\mathbb{E} \sup_{t \in T} X_t \;\lesssim\; \sqrt{\log m},$$

but

$$\int_0^\infty \sqrt{\ln N\big(T, \|\cdot\|_2, \varepsilon\big)}\, d\varepsilon \;\gtrsim\; (\log m)^{3/2}.$$

In this case there is a $\log m$ separation. [Talagrand, 2014] discusses additional examples; see also [van Handel, 2016]. We also note that Talagrand states: ***"However useful, Dudley's bound is not optimal in a number of fundamentally important situations."*** (see [Talagrand, 2014, Ch. 1]). Talagrand [2014] then goes on to show the sub-optimality of the entropy bound for various examples (see, e.g., Eqs. (2.42), (2.49) and Section 2.13 in [Talagrand, 2014]). Likewise, Vershynin [2018, Sec. 8.5] writes: ***"Although Dudley inequality is a simple and versatile tool, it can be loose. The covering numbers just do not contain enough information to capture the magnitude of a random process."*** Since we use the same toolkit of Generic Chaining and our bounds are expressed in terms of γ_2 functionals (or Gaussian widths), they unify existing bounds and at the same time provide sharper bounds on many sets than the standard entropy integral bound [Talagrand, 2005, 2014, Vershynin, 2018].

C EXAMPLES OF GAUSSIAN WIDTH

A deviation bound in terms of the Gaussian width (or γ_2 functional) automatically captures the structure of the set, and thus generalizes bounds that apply only to specific structures. Below are examples where the Gaussian width can be far smaller than the ambient dimension d (see [Vershynin, 2018, Ch. 7] and [Banerjee et al., 2015]).

1. **Finite Point Set.** Any finite $T \subset \mathbb{R}^d$ with $\|x\| \le 1$ for all $x \in T$ and $|T| < \infty$. Then
$$w(T) \lesssim \sqrt{\log |T|}\,.$$

2. **Cross-Polytope.** $B_1^d = \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$. Then
$$w(B_1^d) \lesssim \sqrt{\log d}\,.$$

3. **k-sparse Unit Vectors.** $T = \{x \in \mathbb{R}^d : \|x\|_2 = 1,\ \|x\|_0 \le k\}$. Then
$$w(T) \lesssim \sqrt{2k \log(d/k)}\,.$$

4. **Group-Sparse Ball.** Partition $\{1, \dots, p\}$ into disjoint groups $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_M\}$. For exponents $\nu = (\nu_1, \dots, \nu_M)$ define $\|\alpha\|_{\mathcal{G},\nu} = \sum_{t=1}^M \|\alpha_{\mathcal{G}_t}\|_{\nu_t}$ and $\Omega_{\mathcal{G},\nu} = \{\alpha : \|\alpha\|_{\mathcal{G},\nu} \le 1\}$. Let $m := \max_{1 \le t \le M} |\mathcal{G}_t|$. Then
$$w(\Omega_{\mathcal{G},\nu}) \lesssim \sqrt{m + \log M}\,.$$

We provide an explicit computation of Gaussian width and discuss how it informs the sketching dimension. Consider the *cross-polytope* (the ℓ_1 unit ball)

$$B_1^n \;=\; \{x \in \mathbb{R}^n :\ \|x\|_1 \le 1\}.$$

The Gaussian width of a set $T \subset \mathbb{R}^n$ is

$$w(T) \;=\; \mathbb{E} \sup_{x \in T} \langle g, x \rangle, \qquad g \sim \mathcal{N}(0, I_n).$$

By ℓ_1–ℓ_∞ duality,

$$\sup_{x \in B_1^n} \langle g, x \rangle \;=\; \|g\|_\infty \quad\Longrightarrow\quad w(B_1^n) \;=\; \mathbb{E}\|g\|_\infty \;=\; \mathbb{E} \max_{1 \le i \le n} |g_i|.$$

Explicit upper bound. Let $g_1, \ldots, g_n$ be mean-zero, variance-one Gaussian random variables (not necessarily independent). For any real numbers $x_1, \ldots, x_n$ and any $t > 0$,

$$\max_{i \le n} x_i \;\le\; \frac{1}{t} \log\Big(\sum_{i=1}^{n} e^{t x_i}\Big).$$

Apply this with $x_i = g_i$ and take expectations. By Jensen's inequality (since log is concave),

$$\mathbb{E} \max_{i \le n} g_i \;\le\; \frac{1}{t}\, \mathbb{E} \log\Big(\sum_{i=1}^{n} e^{t g_i}\Big) \;\le\; \frac{1}{t} \log\Big(\sum_{i=1}^{n} \mathbb{E} e^{t g_i}\Big).$$

Each $g_i \sim \mathcal{N}(0,1)$ has moment generating function $\mathbb{E} e^{t g_i} = e^{t^2/2}$, hence

$$\mathbb{E} \max_{i \le n} g_i \;\le\; \frac{1}{t} \log\big(n e^{t^2/2}\big) \;=\; \frac{\log n}{t} + \frac{t}{2}.$$

Optimizing the right-hand side over $t > 0$ gives $t^\star = \sqrt{2 \log n}$ and thus

$$\mathbb{E} \max_{i \le n} g_i \;\le\; \sqrt{2 \log n}.$$

For absolute values,

$$\max_{i \le n} |g_i| \;=\; \max\{g_1, \ldots, g_n, -g_1, \ldots, -g_n\},$$

so the same argument applied to the $2n$ variables $\{g_1, \ldots, g_n, -g_1, \ldots, -g_n\}$ (using $\mathbb{E} e^{t(-g_i)} = e^{t^2/2}$) yields

$$\mathbb{E} \max_{i \le n} |g_i| \;\le\; \frac{1}{t} \log\big(2n e^{t^2/2}\big) \;=\; \frac{\log(2n)}{t} + \frac{t}{2},$$

optimized at $t^\star = \sqrt{2 \log(2n)}$ to give

$$\mathbb{E} \max_{i \le n} |g_i| \;\le\; \sqrt{2 \log(2n)}.$$

Therefore,

$$\boxed{w(B_1^n) = \mathbb{E} \|g\|_\infty \;\le\; \sqrt{2 \log(2n)}}.$$

Implication for sketching dimension. A natural choice is to set the sketching dimension b at the scale

$$b \;=\; w(B_1^n)^2 \;=\; 2 \log(2n).$$

Furthermore, one has the lower bound $w(B_1^n) \ge c \sqrt{\log(2n)}$ (see, e.g., [Vershynin, 2018, Ex. 7.5.8]). Since practitioners typically tune around principled scales rather than compute global constants exactly, this suggests setting b on the order of $w(B_1^n)^2 \asymp \log(2n)$ and adjusting empirically for the task at hand.

D DETAILED PROOF OF MAIN RESULT (THEOREM 3.1)

To develop large deviation bounds on $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$, we decompose the quadratic form into terms depending on the off-diagonal and the diagonal elements of $M^\top N$ respectively as follows.

$$B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \;\triangleq\; \sup_{M \in \mathcal{M}, N \in \mathcal{N}} \Big| \sum_{\substack{j,k=1 \\ j \ne k}}^{n} \boldsymbol{\xi}_j \boldsymbol{\xi}_k \langle M_j, N_k \rangle \Big|,$$

$$D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) \;\triangleq\; \sup_{M \in \mathcal{M}, N \in \mathcal{N}} \Big| \sum_{j=1}^{n} (|\boldsymbol{\xi}_j|^2 - \mathbb{E}\,|\boldsymbol{\xi}_j|^2) \langle M_j, N_j \rangle \Big|,$$

where M_i and N_i are the i-th row of M and N respectively.

Our approach to getting a large deviation bound for $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ is based on bounding $\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$, where where for a random variable X, the L_p norm is defined as:

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} .$$

This in turn is based on bounding $\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$ and $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$. Such bounds lead to a bound on $\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$ of the form

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq W + \sqrt{p}\cdot V + p\cdot U \ , \quad \forall p \geq 1 \ , \tag{14}$$

where a, b, c are constants which do not depend on p. Note that by using the moment-generating function and Markov's inequality [Williams, 1991, Vershynin, 2012], these L_p-norm bounds imply, for all $\epsilon > 0$

$$P(|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})| \geq W + V\cdot\sqrt{\epsilon} + U\cdot\epsilon) \leq e^{-\epsilon} \ , \tag{15}$$

or, equivalently

$$P(|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})| \geq W + \epsilon) \leq \exp\left\{-\min\left(\frac{\epsilon^2}{4V^2}, \frac{u}{2U}\right)\right\} \ , \tag{16}$$

which yields the desired large deviation bound.

The analysis for bounding the L_p norms of $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ for any $p \geq 1$ will thus be based on bounding the L_p norms of $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$, a term based on the off-diagonal elements of $M^\top N$, and that of $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$, a term based on the diagonal elements of $M^\top N$.

Lemma 3.2. *For $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ and $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ as defined above, the following holds*

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq \|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} + \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$$

Proof. Note that the contributions from the off-diagonal terms of $M^\top N$ to $\mathbb{E}[(\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi})]$ is 0. To see this, by linearity of expectation we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\xi}}\left[\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_j\boldsymbol{\xi}_k\langle M_j, N_k\rangle\right] = \sum_{\substack{j,k=1\\ j\neq k}}^{n} \mathop{\mathbb{E}}_{\boldsymbol{\xi}_j,\boldsymbol{\xi}_k}[\boldsymbol{\xi}_j\boldsymbol{\xi}_k]\langle M_j, N_k\rangle = 0 \ ,$$

where the last equality follows by Assumption 1.

Now, using Jensen's inequality, we have

$$\begin{aligned}
C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) &= \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|(\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}) - \mathbb{E}[(\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi})]\right| \ . \\
&= \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_j\boldsymbol{\xi}_k\langle M_j, N_k\rangle + \sum_{j=1}^{n}(|\boldsymbol{\xi}_j|^2 - \mathbb{E}\,|\boldsymbol{\xi}_j|^2)\langle M_j, N_j\rangle\right| \\
&\leq \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_j\boldsymbol{\xi}_k\langle M_j, N_k\rangle\right| + \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{j=1}^{n}(|\boldsymbol{\xi}_j|^2 - \mathbb{E}\,|\boldsymbol{\xi}_j|^2)\langle M_j, N_j\rangle\right| \\
&= B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}) + D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})
\end{aligned}$$

Therefore, for any $p \in [1, \infty)$, we have

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \leq \|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} + \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \ . \tag{17}$$

□

We bound $\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$ and $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$ using Theorem 3.3 and Theorem 3.5 respectively. Note that collecting terms corresponding to W, V and U and combining them with the observation in (16) completes the proof of Theorem 3.1. The rest of the proof is devoted to bounding the off-diagonal term $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ and the diagonal term $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ in section D.1 and D.2 respectively.

D.1 The Off-diagonal Term $B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$

The main result for the off-diagonal term is the following:

Theorem 3.3. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \lesssim \Bigg[& \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big(\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p}\, d_{2\to 2}(\mathcal{N})\Big) \\ & + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p}\, d_{2\to 2}(\mathcal{M})\Big) \\ & + \sqrt{p}\, \min\Big\{d_F(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) \cdot d_{2\to 2}(\mathcal{M})\Big\} + p \cdot d_{2\to 2}(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}) \Bigg].\end{aligned}$$

D.1.1 Proof of Theorem 3.3

We start by using a decoupling inequality from [Krahmer et al., 2014b].

Lemma D.1 (Theorem 2.4, Krahmer et al. [2014b]). *. Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$ be a sequence of independent, centered random variables, and let F be a convex function. If $\mathcal{B}$ is a collection of matrices and $\boldsymbol{\xi}'$ is an independent copy of $\boldsymbol{\xi}$, then*

$$\mathbb{E} \sup_{B\in\mathcal{B}} F\left(\sum_{\substack{j,k=1\\ j\neq k}}^{n} \xi_j \xi_k B_{j,k}\right) \leq \mathbb{E} \sup_{B\in\mathcal{B}} F\left(4\sum_{j,k=1}^{n} \xi_j \xi'_k B_{j,k}\right).$$

We use Lemma D.1 with $F(x) = |x|^p, p \geq 1$ as the convex function and set $B_{j,k} = \langle M_j, N_k\rangle$. Then

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} &= \mathbb{E} \sup_{M\in\mathcal{M},N\in\mathcal{N}} F\left(\sum_{\substack{j,k=1\\ j\neq k}}^{n} \xi_j \xi_k \langle M_j, N_k\rangle\right) \\ &\leq E \sup_{M\in\mathcal{M},N\in\mathcal{N}} F\left(4\sum_{\substack{j,k=1\\ j\neq k}}^{n} \xi_j \xi'_k \langle M_j, N_k\rangle\right) \\ &\lesssim \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} |\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle| \right\|_{L_p}. \end{aligned} \tag{18}$$

Note that for fixed M, N, the term $|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle|$ conditioned on $\boldsymbol{\xi}'$ is sub-gaussian and therefore its L_p norm can be bounded. However, the $\sup_{M\in\mathcal{M},N\in\mathcal{N}}$ inside the $\|\cdot\|_{L_p}$ does not let us use this approach. The next Theorem therefore upper bounds $\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}} |\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle|\right\|_{L_p}$ by $\sup_{M\in\mathcal{M},N\in\mathcal{N}} \|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\|_{L_p}$ plus some additional complexity terms. Unlike Krahmer et al. [2014b] the inner product contains two different matrices M and N, and therefore we consider two separate admissible sequences (cf. definition 2.2) $\{T_r(\mathcal{M})\}_{r=0}^{\infty}$ and $\{T_r(\mathcal{N})\}_{r=0}^{\infty}$ of $\mathcal{M}$ and $\mathcal{N}$ respectively. We then use a generic chaining argument by creating two separate increment sequences for $\mathcal{M}$ and $\mathcal{N}$ and is detailed in the following theorem.

Lemma D.2. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1, and $\boldsymbol{\xi}'$ be a decoupled tangent sequence to $\boldsymbol{\xi}$. Then, for every $p \geq 1$,*

$$\begin{aligned}\left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \right\|_{L_p} \lesssim\ & \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p} d_{2\to 2}(\mathcal{N})\Big] \\ & + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M})\Big] \\ & + \sup_{M\in\mathcal{M},N\in\mathcal{N}} \|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\|_{L_p}, \end{aligned} \tag{19}$$

Proof of Lemma D.2. Without loss of generality, assume $\mathcal{M}$ and $\mathcal{N}$ are finite Talagrand [2014]. Let $\{T_r(\mathcal{M})\}_{r=0}^{\infty}$ and $\{T_r(\mathcal{N})\}_{r=0}^{\infty}$ be admissible sequences for $\mathcal{M}$ and $\mathcal{N}$ for which the minimum in the definition of $\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})$ and $\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2})$ are attained respectively. Let

$$\pi_r M = d_{2\to 2}(M, T_r(\mathcal{M})) = \operatorname*{argmin}_{B\in T_r(\mathcal{M})} \|B - A\|_{2\to 2} \qquad \text{and} \qquad \Delta_r M = \pi_r M - \pi_{r-1} M \ .$$
$$\pi_r N = d_{2\to 2}(N, T_r(\mathcal{N})) = \operatorname*{argmin}_{B\in T_r(\mathcal{N})} \|B - A\|_{2\to 2} \qquad \text{and} \qquad \Delta_r N = \pi_r N - \pi_{r-1} N \ .$$

For any given $p \geq 1$, let ℓ be the largest integer for which $2^\ell \leq 2p$. Then,

$$\begin{aligned}
&\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle - \langle(\pi_\ell M)\boldsymbol{\xi}, (\pi_\ell N)\boldsymbol{\xi}'\rangle = \sum_{r=\ell}^{\infty} \langle(\pi_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle - \langle(\pi_r M)\boldsymbol{\xi}, (\pi_r N)\boldsymbol{\xi}'\rangle \\
&= \sum_{r=\ell}^{\infty} \langle(\pi_r M + \Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle - \langle(\pi_r M)\boldsymbol{\xi}, (\pi_{r+1}N - \Delta_{r+1}N)\boldsymbol{\xi}'\rangle \\
&= \sum_{r=\ell}^{\infty} \langle(\pi_r M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle + \langle(\Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle - \langle(\pi_r M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle + \langle(\pi_r M)\boldsymbol{\xi}, (\Delta_{r+1}N)\boldsymbol{\xi}'\rangle \\
&= \sum_{r=\ell}^{\infty} \langle(\Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle + \langle(\pi_r M)\boldsymbol{\xi}, (\Delta_{r+1}N)\boldsymbol{\xi}'\rangle
\end{aligned}$$

Now by applying triangle inequality, we have

$$|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle - \langle(\pi_\ell M)\boldsymbol{\xi}, (\pi_\ell N)\boldsymbol{\xi}'\rangle| \leq \underbrace{\left|\sum_{r=\ell}^{\infty} \langle(\Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle\right|}_{S_1} + \underbrace{\left|\sum_{r=\ell}^{\infty} \langle(\pi_r M)\xi, (\Delta_{r+1}N)\xi'\rangle\right|}_{S_2} \ . \tag{20}$$

We first consider S_1. Let us define

$$X_r(M, N) = \langle(\Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle \ .$$

Conditioning $X_r(M, N)$ on $\boldsymbol{\xi}'$, we note

$$X_r(M, N) \ \big|\ \boldsymbol{\xi}' = \langle(\Delta_{r+1}M)\boldsymbol{\xi}, (\pi_{r+1}N)\boldsymbol{\xi}'\rangle \ \big|\ \boldsymbol{\xi}' = \langle\boldsymbol{\xi}, (\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'\rangle \ \big|\ \boldsymbol{\xi}'$$

is a sub-Gaussian random variable and therefore using Azuma-Hoeffding bound [Boucheron et al., 2013, Vershynin, 2018] gives

$$P\left(|X_r(M, N)| > u\|(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'\|_2 \ \middle|\ \boldsymbol{\xi}'\right) \leq 2\exp(-u^2/2) \ .$$

Using $u = t2^{r/2}$, we get

$$P\left(|X_r(M, N)| > t2^{r/2}\|(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'\|_2 \ \middle|\ \boldsymbol{\xi}'\right) \leq 2\exp(-t^2 2^r/2) \ .$$

Since

$$\left|(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'\right| \leq \|\Delta_{r+1}M\|_{2\to 2} \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2 \ .$$

we have

$$P\left(|X_r(M, N)| > t2^{r/2}\|\Delta_{r+1}M\|_{2\to 2} \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2 \ \middle|\ \boldsymbol{\xi}'\right) \leq 2\exp(-t^2 2^r/2) \ .$$

Now, since $|\{\pi_r M : M \in \mathcal{M}\}| = |T_r(\mathcal{M})| \leq 2^{2^r}$ and $|\{\pi_r N : N \in \mathcal{M}\}| = |T_r(\mathcal{N})| \leq 2^{2^r}$, by union bound, we get

$$P\left(\sup_{M\in\mathcal{M}, N\in\mathcal{N}} \sum_{r=\ell}^{\infty} |X_r(M, N)| > t\left(\sup_{M\in\mathcal{M}} \sum_{r=\ell}^{\infty} 2^{r/2}\|\Delta_{r+1}M\|_{2\to 2}\right) \cdot \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2 \ \middle|\ \boldsymbol{\xi}'\right)$$

$$\leq 2\sum_{r=\ell}^{\infty} |T_r(\mathcal{M})| \cdot |T_{r+1}(\mathcal{M})| \cdot |T_{r+1}(\mathcal{N})| \cdot \exp(-t^2 2^r/2)$$
$$\leq 2\sum_{r=\ell}^{\infty} 2^{2^{r+2}} \cdot \exp(-t^2 2^r/2)$$
$$\leq 2\exp(-2^\ell t^2) \ ,$$

for all $t \geq t_0$, a constant. Next, note that

$$\sup_{M\in\mathcal{M}} \sum_{r=\ell}^{\infty} 2^{r/2} \|\Delta_{r+1} M\|_{2\to 2} = \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}).$$

Therefore we have

$$P\Big(\sup_{M\in\mathcal{M}, N\in\mathcal{N}} \sum_{r=\ell}^{\infty} |X_r(M,N)| > t\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2 \ \Big| \ \boldsymbol{\xi}' \Big) \leq 2\exp(-pt^2) \ ,$$

since $p \leq 2^\ell$ by construction which implies with $V(\boldsymbol{\xi}') = \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2$, for $t \geq t_0$ we have

$$P\left(S_1 \geq tV(\boldsymbol{\xi}') \mid \boldsymbol{\xi}'\right) \leq 2\exp(-pt^2) \ .$$

Note that

$$\|S_1\|_{L_p}^p = \mathop{\mathbb{E}}_{\boldsymbol{\xi},\boldsymbol{\xi}'} S_1^p = E_{\boldsymbol{\xi}'} \int_0^\infty pt^{p-1} P(S_1 > t \mid \boldsymbol{\xi}') dt \ ,$$

and

$$\int_0^\infty pt^{p-1} P(S_1 > t \mid \boldsymbol{\xi}') dt \leq c^p V(\boldsymbol{\xi}')^p + \int_{cV(\boldsymbol{\xi}')}^\infty pt^{p-1} P(S_1 > t \mid \boldsymbol{\xi}') dt$$
$$\leq c^p V(\boldsymbol{\xi}')^p + V(\boldsymbol{\xi}')^p \int_c^\infty p\tau^{p-1} P(S_1 > \tau V(\boldsymbol{\xi}') | \boldsymbol{\xi}') d\tau$$
$$\leq c_1^p V(\boldsymbol{\xi}')^p \ ,$$

where $c \geq t_0, c_1$ are suitable constants that depend on L. As a result, $\|S_1\|_{L_p} \leq c_1 V(\boldsymbol{\xi}') = c_1 \|V(\boldsymbol{\xi})\|_{L_p}$, i.e., we have the following bound on S_1.

$$\|S_1\|_{L_p} \lesssim \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \Big\| \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}\|_2 \Big\|_{L_p}$$

Note that a similar analysis follows for S_2, and we can bound $\|S_2\|_{L_p}$. As a result

$$\|S_1 + S_2\|_{L_p} \lesssim \Big(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big\| \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}\|_2 \Big\|_{L_p} + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big\| \sup_{M\in\mathcal{M}} \|M\boldsymbol{\xi}\|_2 \Big\|_{L_p} \Big) \tag{21}$$

Further, since $|\{\pi_\ell M : M \in \mathcal{M}\}| \leq 2^{2^\ell} \leq \exp(2p)$, and $|\{\pi_\ell N : N \in \mathcal{N}\}| \leq 2^{2^\ell} \leq \exp(2p)$ we have

$$\mathbb{E} \sup_{M\in\mathcal{M}, N\in\mathcal{N}} |\langle (\pi_\ell M)\boldsymbol{\xi}, (\pi_\ell N)\boldsymbol{\xi}\rangle|^p \leq \sum_{M\in T_\ell(M), N\in T_\ell(N)} \mathbb{E}\,|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}\rangle|^p$$
$$\leq 2^{2p} 2^{2p} \sup_{M\in\mathcal{M}, N\in\mathcal{N}} E|\langle A\boldsymbol{\xi}, A\boldsymbol{\xi}\rangle|^p = 2^{4p} \sup_{M\in\mathcal{M}, N\in\mathcal{N}} |\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}\rangle|^p \ ,$$

so that

$$\Big\| \sup_{M\in\mathcal{M}, N\in\mathcal{N}} |\langle (\pi_\ell M)\boldsymbol{\xi}, (\pi_\ell N)\boldsymbol{\xi}\rangle \Big\|_{L_p} \leq 16 \sup_{M\in\mathcal{M}, N\in\mathcal{N}} \Big\| \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}\rangle \Big\|_{L_p} \ . \tag{22}$$

Combining (20), (21), and (22) and using triangle inequality we have

$$\Big\| \sup_{M\in\mathcal{M}, N\in\mathcal{N}} \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p} \lesssim \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big\| \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}'\|_2 \Big\|_{L_p} + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big\| \sup_{M\in\mathcal{M}} \|M\boldsymbol{\xi}\|_2 \Big\|_{L_p}$$

$$+ \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p} .$$

Using Theorem 3.5 from [Krahmer et al., 2014a] we have that

$$\left\| \sup_{M\in\mathcal{M}} \|M\boldsymbol{\xi}\|_2 \right\|_{L_p} \lesssim \cdot\Big(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M})\Big)$$
$$\left\| \sup_{N\in\mathcal{N}} \|N\boldsymbol{\xi}\|_2 \right\|_{L_p} \lesssim \cdot\Big(\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p} d_{2\to 2}(\mathcal{N})\Big)$$

Combining all these completes the proof of Lemma D.2. □

Therefore using (18) and Lemma D.2 we get

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \lesssim \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p} d_{2\to 2}(\mathcal{N})\Big] \\ + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M})\Big] \\ + \sup_{M\in\mathcal{M},N\in\mathcal{N}} \|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\|_{L_p} \end{aligned} \tag{23}$$

Next we consider $\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p}$ and bound it using the following Lemma.

Lemma D.3. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1, and let $\boldsymbol{\xi}'$ be a decoupled tangent sequence. Then, for every $p \geq 1$,*

$$\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p} \lesssim \min\Big\{\sqrt{p}\cdot d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}), \sqrt{p}\cdot d_F(\mathcal{N})\cdot d_{2\to 2}(\mathcal{M})\Big\} + p\cdot d_{2\to 2}(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}) . \tag{24}$$

Proof of Lemma D.3. Fix $M \in \mathcal{M}, N \in \mathcal{N}$ and let $S = \{M^\top N x : x \in B_2^n, M \in \mathcal{M}, N \in \mathcal{N}\}$, where $B_2^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. Conditioned on $\boldsymbol{\xi}$, the random variable $\langle \boldsymbol{\xi}, M^\top N\boldsymbol{\xi}'\rangle$ is sub-gaussian. Therefore for some global constant $\tilde{C} > 0$, we have

$$\begin{aligned}\left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p} &\leq \tilde{C}\left(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}\left((\mathop{\mathbb{E}}_{\boldsymbol{\xi}}|\langle \boldsymbol{\xi}, M^\top N\boldsymbol{\xi}'\rangle|^p)^{1/p}\right)^p\right)^{1/p} \leq \tilde{C}(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}(L\sqrt{p})^p\|M^\top N\boldsymbol{\xi}'\|_2^p)^{1/p} \\ &\lesssim L\sqrt{p}\left(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}\sup_{y\in S}|\langle y, \boldsymbol{\xi}'\rangle|^p\right)^{1/p}\end{aligned}$$

Now using Theorem 2.3 from Krahmer et al. [2014a] we have for every $p \geq 1$

$$\left(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}\sup_{y\in S}|\langle y, \boldsymbol{\xi}'\rangle|^p\right)^{1/p} \lesssim \left(\mathop{\mathbb{E}}_{\mathbf{g}}\sup_{y\in S}|\langle \mathbf{g}, y\rangle| + \sup_{y\in S}(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}|\langle \boldsymbol{\xi}', y\rangle|^p)^{1/p}\right)$$

where $\mathbf{g}$ is a standard Gaussian vector. The first term in the rhs can be bounded as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{g}}\sup_{y\in S}|\langle \mathbf{g}, y\rangle| = \mathop{\mathbb{E}}_{\mathbf{g}}\|M^\top N g\|_2 \leq (\mathop{\mathbb{E}}_{\mathbf{g}}\|M^\top N g\|_2^2)^{1/2} = \|M^\top N\|_F \\ \leq \min\Big\{\|M\|_{2\to 2}\|N\|_F, \|N\|_{2\to 2}\|M\|_F\Big\}\end{aligned}$$

Next the second term in the rhs can be bounded as follows:

$$\sup_{y\in S}(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'}|\langle \boldsymbol{\xi}', y\rangle|^p)^{1/p} \leq L \sup_{z\in B_2^p}\sqrt{p}\|M^\top N z\|_2 \leq L\sqrt{p}\|M\|_{2\to 2}\|N\|_{2\to 2}.$$

Combining all the above bounds we get:

$$\left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p} \lesssim \sqrt{p}\min\left\{\|M\|_{2\to2}\|N\|_F, \|N\|_{2\to2}\|M\|_F\right\} + p\|M\|_{2\to2}\|N\|_{2\to2} \tag{25}$$

Now recall that for the set $\mathcal{M}$, we have $d_F(\mathcal{M}) = \sup_{M\in\mathcal{M}}\|M\|_F$, and $d_{2\to2}(\mathcal{M}) = \sup_{A\in\mathcal{M}}\|A\|_{2\to2}$, which implies

$$\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left\|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\right\|_{L_p} \lesssim \left(\sqrt{p}\min\left\{d_{2\to2}(\mathcal{M})d_F(\mathcal{N}), d_{2\to2}(\mathcal{N})d_F(\mathcal{M})\right\} + pd_{2\to2}(\mathcal{M})d_{2\to2}(\mathcal{N})\right).$$

□

Combining Lemma 24 with (23) completes the proof of Theorem 3.5.

D.2 The Diagonal Term $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$

For the diagonal terms, we have the following main result: **Theorem 3.5** *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned}\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p} \lesssim \Big[&\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2})\cdot\left(\gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) + d_F(\mathcal{N}) + \sqrt{p}\; d_{2\to2}(\mathcal{N})\right)\\ &+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2})\cdot\left(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) + d_F(\mathcal{M}) + \sqrt{p}\; d_{2\to2}(\mathcal{M})\right)\\ &+ \sqrt{p}\;\min\left\{d_F(\mathcal{M})\cdot d_{2\to2}(\mathcal{N}), d_F(\mathcal{N})\cdot d_{2\to2}(\mathcal{M})\right\} + p\cdot d_{2\to2}(\mathcal{M})\cdot d_{2\to2}(\mathcal{N})\;.\Big]\end{aligned}$$

D.2.1 Proof of Theorem 3.5

By definition of $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ and from Lemma 9 in Banerjee et al. [2019], we have

$$\begin{aligned}\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}))\|_{L_p} &= \left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n(|\boldsymbol{\xi}_j|^2 - \mathbb{E}\,|\boldsymbol{\xi}_j|^2)\langle M_j, N_j\rangle\right|\right\|_{L_p}\\ &\leq 2\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n\varepsilon_j|\boldsymbol{\xi}_j|^2\langle M_j, N_j\rangle\right|\right\|_{L_p},\end{aligned}$$

where $\{\varepsilon_j\}$ is a set of independent Rademacher variables independent of $\boldsymbol{\xi}$. Let $\{g_j\}$ be a sequence of independent Gaussian random variables. Since $\boldsymbol{\xi}_j$ is a L-sub-Gaussian random variable [Vershynin, 2018], there is an absolute constant c such that for all $t > 0$

$$\mathbb{P}\left(|\xi_j|^2 \geq tL^2\right) \leq c\mathbb{P}(g_j^2 \geq t)\;.$$

Then, from contraction of stochastic processes ([Ledaux and Talagrand, 1991, Lemma 4.6]), we have

$$\begin{aligned}\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}))\|_{L_p} &\leq 2\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n\varepsilon_j|\boldsymbol{\xi}_j|^2\langle M_j, N_j\rangle\right|\right\|_{L_p}\\ &\leq 2cL^2\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n\varepsilon_j|g_j|^2\langle M_j, N_j\rangle\right|\right\|_{L_p}\\ &\overset{(a)}{\leq} 2cL^2\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n\varepsilon_j(|g_j|^2-1)\langle M_j, N_j\rangle\right|\right\|_{L_p} + 2cL^2\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^n\varepsilon_j\langle M_j, N_j\rangle\right|\right\|_{L_p}\end{aligned}$$

$$\overset{(b)}{\leq} 4cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{j=1}^{n} (|g_j|^2 - 1)\langle M_j, N_j\rangle \right| \right\|_{L_p} + 2cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j\rangle \right| \right\|_{L_p}$$

$$\leq 4cL^2 \left\| D_{\mathcal{M},\mathcal{N}}(\mathbf{g}) \right\|_{L_p} + 2cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j\rangle \right| \right\|_{L_p}, \tag{26}$$

where (a) follows from Jensen's inequality and since $E|g_j|^2 = 1$, and (b) follows by de-symmetrization following [Banerjee et al., 2019, Lemma 11] and since the convex function here is 1-Lipschitz.

By triangle inequality, we have

$$\left\| D_{\mathcal{M},\mathcal{N}}(\mathbf{g}) \right\|_{L_p} \leq \left\| C_{\mathcal{M},\mathcal{N}}(\mathbf{g}) \right\|_{L_p} + \left\| B_{\mathcal{M},\mathcal{N}}(\mathbf{g}) \right\|_{L_p} \tag{27}$$

In order to handle $\left\| C_{\mathcal{M},\mathcal{N}}(\mathbf{g}) \right\|_{L_p}$, we require a stronger decoupling inequality for an order 2 Gaussian chaos than the one used in Theorem 2.5 in Krahmer et al. [2014b].

Theorem D.4. *There exists an absolute constant C such that the following holds for all $p \geq 1$. Let $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_n)$ be a sequence of independent standard normal random variables. If $\mathcal{B}$ is a collection of matrices and $\mathbf{g}'$ is an independent copy of $\mathbf{g}$, then*

$$\mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{g}_j \mathbf{g}_k B_{j,k} + \sum_{j=1}^{n} \left(\mathbf{g}_j^2 - 1\right) B_{j,j} \right|^p \leq C^p \, \mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{j,k=1}^{n} \mathbf{g}_j \, \mathbf{g}'_k \, B_{j,k} \right|^p.$$

Proof. For each $B \in \mathcal{B}$, denote $C^B = (C^B_{j,k})_{n\times n}$ such that $C^B_{j,k} = \frac{B_{j,k}+B_{k,j}}{2}, \forall j,k \in [n]\times[n]$, and denote $\mathcal{C}^{\mathcal{B}}$ is the collection of C^B for all $B \in \mathcal{B}$. Then $\mathcal{C}^{\mathcal{B}}$ is a collection of Hermitian matrices. According to Theorem 2.5 in Krahmer et al. [2014b], we have

$$\mathbb{E} \sup_{C^B\in\mathcal{C}^{\mathcal{B}}} \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{g}_j \mathbf{g}_k C^B_{j,k} + \sum_{j=1}^{n} \left(\mathbf{g}_j^2 - 1\right) C^B_{j,j} \right|^p \leq c^p \, \mathbb{E} \sup_{C^B\in\mathcal{C}^{\mathcal{B}}} \left| \sum_{j,k=1}^{n} \mathbf{g}_j \, \mathbf{g}'_k \, C^B_{j,k} \right|^p. \tag{28}$$

In addition, the left-hand side is

$$\begin{aligned}
&\mathbb{E} \sup_{C^B\in\mathcal{C}^{\mathcal{B}}} \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{g}_j \mathbf{g}_k C^B_{j,k} + \sum_{j=1}^{n} \left(\mathbf{g}_j^2 - 1\right) C^B_{j,j} \right|^p \\
=&\mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{g}_j \mathbf{g}_k \frac{B_{j,k}+B_{k,j}}{2} + \sum_{j=1}^{n} \left(\mathbf{g}_j^2 - 1\right) B_{j,j} \right|^p \\
=&\mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{g}_j \mathbf{g}_k B_{j,k} + \sum_{j=1}^{n} \left(\mathbf{g}_j^2 - 1\right) B_{j,j} \right|^p.
\end{aligned} \tag{29}$$

Further the right-hand side can be upper bounded as follow.

$$\begin{aligned}
c^p \, \mathbb{E} \sup_{C^B\in\mathcal{C}^{\mathcal{B}}} \left| \sum_{j,k=1}^{n} \mathbf{g}_j \, \mathbf{g}'_k \, C^B_{j,k} \right|^p &= c^p \, \mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{j,k=1}^{n} \mathbf{g}_j \, \mathbf{g}'_k \, \frac{B_{j,k}+B_{k,j}}{2} \right|^p \\
&= \frac{c^p}{2^p} \, \mathbb{E} \sup_{B\in\mathcal{B}} \left| \sum_{j,k=1}^{n} \mathbf{g}_j \, \mathbf{g}'_k \, (B_{j,k}+B_{k,j}) \right|^p
\end{aligned}$$

$$
\begin{aligned}
&= \frac{c^p}{2^p}\,\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{j,k=1}^{n}\mathbf{g}_j\,\mathbf{g}'_k\,B_{j,k}+\sum_{j,k=1}^{n}\mathbf{g}_j\,\mathbf{g}'_k\,B_{k,j}\right|^p\\
&\overset{(a)}{\leq}\frac{c^p}{2^p}\mathbb{E}\sup_{B\in\mathcal{B}}2^p\left(\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}'_k B_{j,k}\right|^p+\left|\sum_{j,k=1}^{n}\mathbf{g}_j\,\mathbf{g}'_k\,B_{k,j}\right|^p\right)
\end{aligned}
\tag{30}
$$

where (a) utilizes the inequality $|a+b|^p \leq 2^p(|a|^p+|b|^p)$. Therefore

$$
\begin{aligned}
c^p\,\mathbb{E}\sup_{C^{\mathcal{B}}\in\mathcal{C}^{\mathcal{B}}} &\leq c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}'_k B_{j,k}\right|^p+c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{j,k=1}^{n}\mathbf{g}_j\,\mathbf{g}'_k\,B_{k,j}\right|^p\\
&= c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}'_k B_{j,k}\right|^p+c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{j,k=1}^{n}\mathbf{g}'_j\,\mathbf{g}_k\,B_{j,k}\right|^p\\
&\overset{(b)}{=} 2c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}'_k B_{j,k}\right|^p
\end{aligned}
\tag{31}
$$

where (b) holds since $\mathbf{g}'$ is an independent copy of $\mathbf{g}$. Substituting (29) and (31) into (28), we can get that

$$
\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}_k B_{j,k}+\sum_{j=1}^{n}\left(\mathbf{g}_j^2-1\right)\,B_{j,j}\right|^p\leq 2c^p\mathbb{E}\sup_{B\in\mathcal{B}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}'_k B_{j,k}\right|^p
$$

By setting $C=2^{\frac{1}{p}}c$, we finish the proof. □

Now

$$
\begin{aligned}
\left\|C_{\mathcal{M},\mathcal{N}}(\mathbf{g})\right\|_{L_p} &= \sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathbf{g}_j\mathbf{g}_k\langle M_j,N_k\rangle+\sum_{j=1}^{n}(|\mathbf{g}_j|^2-\mathbb{E}\,|\mathbf{g}_j|^2)\langle M_j,N_j\rangle\right|\\
&\overset{(a)}{\leq} C\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j,k=1}^{n}\mathbf{g}_j\mathbf{g'}_k\langle M_j,N_k\rangle\right|\right\|_{L_p}=\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}|\langle M\mathbf{g},N\mathbf{g}'\rangle|\right\|_{L_p}\\
&\overset{(b)}{\lesssim}\ \gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\left[\gamma_2(\mathcal{N},\|\cdot\|_{2\to2})+d_F(\mathcal{N})+\sqrt{p}d_{2\to2}(\mathcal{N})\ \right]\\
&\quad+\ \gamma_2(\mathcal{N},\|\cdot\|_{2\to2})\cdot\left[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})+d_F(\mathcal{M})+\sqrt{p}d_{2\to2}(\mathcal{M})\ \right]\\
&\quad+\min\left\{\sqrt{p}\cdot d_F(\mathcal{M})\cdot d_{2\to2}(\mathcal{N}),\sqrt{p}\cdot d_F(\mathcal{N})\cdot d_{2\to2}(\mathcal{M})\right\}\\
&\qquad+p\cdot d_{2\to2}(\mathcal{M})\cdot d_{2\to2}(\mathcal{N})\ .
\end{aligned}
$$

where (a) uses Theorem D.4, and (b) holds because of Lemma D.2 and Lemma D.3. Term $\|B_{\mathcal{M},\mathcal{N}}(\mathbf{g})\|_{L_p}$ can be bounded using Theorem 3.3 thus giving

$$
\left\|D_{\mathcal{M},\mathcal{N}}(\mathbf{g})\right\|_{L_p}\lesssim\left[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\left[\gamma_2(\mathcal{N},\|\cdot\|_{2\to2})+d_F(\mathcal{N})+\sqrt{p}d_{2\to2}(\mathcal{N})\ \right]\right.
$$

$$
\begin{aligned}
&+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M}) \Big] \\
&+ \min\Big\{ \sqrt{p} \cdot d_F(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}), \sqrt{p} \cdot d_F(\mathcal{N}) \cdot d_{2\to 2}(\mathcal{M}) \Big\} \\
&\quad + p \cdot d_{2\to 2}(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}) \Big] \qquad (32)
\end{aligned}
$$

Next we bound the second term $\left\| \sup_{M\in\mathcal{M}, N\in\mathcal{N}} \left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j \rangle \right| \right\|_{L_p}$. We use the following theorem to bound this term.

Theorem D.5. *Consider a stochastic process $\{X_{(u,v)}\}$ where $u \in \mathcal{U}$, $v \in \mathcal{V}$, and let $d(\cdot,\cdot)$ is the metric. Suppose for any $t > 0$, with probability at least $1 - c_0 \exp\left(-\frac{t^2}{2}\right)$, $\{X_{(u,v)}\}$ satisfies the following:*

$$
\begin{aligned}
\left|X_{(u_1,v)} - X_{(u_2,v)}\right| &\le C_u t \cdot d(u_1, u_2), \\
\left|X_{(u,v_1)} - X_{(u,v_2)}\right| &\le C_v t \cdot d(v_1, v_2).
\end{aligned}
$$

Then, with probability $1 - 2c_0 \exp\left(-\frac{t^2}{2}\right)$

$$
\sup_{u\in\mathcal{U},\, v\in\mathcal{V}} |X_{u,v}| \le 4\sqrt{2} t \left(C_u \gamma_2(\mathcal{U}, d) + C_v \gamma_2(\mathcal{V}, d) \right).
$$

Proof. Let $(\mathcal{U}_k)$ be an admissible sequence of subsets of $\mathcal{U}$, and denote $\mathcal{U}_0 = u_0$. Let $(\mathcal{V}_k)$ be an admissible sequence of subsets of $\mathcal{V}$, and denote $\mathcal{V}_0 = v_0$. We now walk from u_0 to a general point $u \in \mathcal{U}$ along the chain

$$
u_0 \;=\; \pi_0(u) \;\to\; \pi_1(u) \;\to\; \cdots \;\to\; \pi_{K_1}(u) \;=\; u,
$$

of points $\pi_k(u) \in \mathcal{U}_k$ that are chosen as best approximations to u in $\mathcal{U}_k$, i.e.

$$
d\big(u, \pi_k(u)\big) \;=\; d\big(u, \mathcal{U}_k\big).
$$

Similarly, We walk from v_0 to a general point $v \in \mathcal{V}$ along the chain

$$
v_0 \;=\; \pi_0(v) \;\to\; \pi_1(v) \;\to\; \cdots \;\to\; \pi_{K_2}(v) \;=\; v,
$$

of points $\pi_k(v) \in \mathcal{V}_k$ that are chosen as best approximations to v in $\mathcal{V}_k$, i.e.

$$
d\big(v, \pi_k(v)\big) \;=\; d\big(v, \mathcal{V}_k\big).
$$

The displacement $X_{(u,v)} - X_{(u_0,v_0)}$ can be expressed as a telescoping sum:

$$
\begin{aligned}
\left|X_{(u,v)} - X_{(u_0,v_0)}\right| &= \left| \sum_{k=1}^{K_2} \big(X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(t))} \big) + \sum_{k=1}^{K_1} \big(X_{(\pi_k(u),v_0)} - X_{(\pi_{k-1}(u),v_0)} \big) \right| \\
&\le \left| \sum_{k=1}^{K_2} \big(X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(t))} \big) \right| + \left| \sum_{k=1}^{K_1} \big(X_{(\pi_k(u),v_0)} - X_{(\pi_{k-1}(u),v_0)} \big) \right|. \qquad (33)
\end{aligned}
$$

According to the assumption, with probability $1 - c_0 \exp\left(-4t^2 2^k d\left(\pi_k(v), \pi_{k-1}(v)\right)\right)$, we have

$$
\left|X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(v))}\right| \le 2\sqrt{2} C_v t 2^{k/2} d\left(\pi_k(v), \pi_{k-1}(v)\right).
$$

We can now unfix $v \in \mathcal{V}$ by taking a union bound over

$$
|\mathcal{V}_k|\,|\mathcal{V}_{k-1}| \le |\mathcal{V}_k|^2 = 2^{2^{k+1}}
$$

possible pairs $(\pi_k(v), \pi_{k-1}(v))$. Similarly, we can unfix k by a union bound over all $k \in \mathbb{N}$. Then with probability at least

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} \cdot c_0 \exp\left(-4t^2 2^k\right) \geq 1 - c_0 \exp\left(-\frac{t^2}{2}\right),$$

for all $v \in \mathcal{V}$ and $k \in \mathbb{N}$,

$$\begin{aligned}
\left|\sum_{k=1}^{K_2}\left(X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(t))}\right)\right| &\leq \sum_{k=1}^{K_2}\left|X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(t))}\right| \\
&\leq 2\sqrt{2}C_v t \sum_{k=1}^{\infty} 2^{k/2} d\left(\pi_k(v), \pi_{k-1}(v)\right) \\
&\leq 2\sqrt{2}C_v t \sum_{k=1}^{\infty} 2^{k/2}\left(d\left(v, \pi_k(v)\right) + d\left(v, \pi_{k-1}(v)\right)\right) \\
&= 2\sqrt{2}C_v t \sum_{k=1}^{\infty} 2^{k/2} d\left(v, \mathcal{V}_k\right) + 4C_u a \sum_{k=0}^{\infty} 2^{k/2} d\left(v, \mathcal{V}_k\right) \\
&\leq 4\sqrt{2}C_v t \gamma_2\left(\mathcal{V}, d\right). \qquad (34)
\end{aligned}$$

Following a similar analysis, with probability at least $1 - c_0 \exp\left(-\frac{t^2}{2}\right)$,

$$\left|\sum_{k=1}^{K_1}\left(X_{(\pi_k(u),v_0)} - X_{(\pi_{k-1}(u),v_0)}\right)\right| \leq 4\sqrt{2}C_u t \gamma_2\left(\mathcal{U}, d\right). \qquad (35)$$

Therefore, substituting (34) and (35) into (33), with probability at least $1 - 2c_0 \exp\left(-\frac{t^2}{2}\right)$,

$$\begin{aligned}
\left|X_{(u,v)} - X_{(u_0,v_0)}\right| &\leq \left|\sum_{k=1}^{K_2}\left(X_{(u,\pi_k(v))} - X_{(u,\pi_{k-1}(t))}\right)\right| + \left|\sum_{k=1}^{K_1}\left(X_{(\pi_k(u),v_0)} - X_{(\pi_{k-1}(u),v_0)}\right)\right| \\
&\leq 4\sqrt{2}C_u t \gamma_2\left(\mathcal{U}, d\right) + 4\sqrt{2}C_v t \gamma_2\left(\mathcal{V}, d\right)
\end{aligned}$$

Then we finish the proof. □

According to [Banerjee et al., 2019, Lemma 7],

$$\begin{aligned}
&\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^{n}\varepsilon_j\langle M_j, N_j\rangle\right|\right\|_{L_p} \\
&= \left(\mathbb{E}_\varepsilon\left[\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^{n}\varepsilon_j\langle M_j, N_j\rangle\right|\right]^p\right)^{1/p} \\
&= \left(\mathbb{E}_\varepsilon\left[\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^{n}\varepsilon_j\langle M_j, N_j\rangle\right|^p\right]\right)^{1/p} \\
&\lesssim \mathbb{E}_{\mathbf{g}}\left[\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^{n}\mathbf{g}_j\langle M_j, N_j\rangle\right|\right] + \sup_{M\in\mathcal{M},N\in\mathcal{N}}\left(\mathbb{E}_\varepsilon\left[\left|\sum_{j=1}^{n}\varepsilon_j\langle M_j, N_j\rangle\right|^p\right]\right)^{1/p} \qquad (36)
\end{aligned}$$

in which $\mathbf{g}_j \sim \mathcal{N}(0,1)$ are independent. For the first term, define $X_{(M,N)} = \left|\sum_{j=1}^{n}\mathbf{g}_j\langle M_j, N_j\rangle\right|$. Then

$$\left|X_{(M^1,N)} - X_{(M^2,N)}\right| = \left|\left|\sum_{j=1}^{n}\mathbf{g}_j\left\langle M_j^1, N_j\right\rangle\right| - \left|\sum_{j=1}^{n}\mathbf{g}_j\left\langle M_j^2, N_j\right\rangle\right|\right|$$

$$
\begin{aligned}
&= \left| \sum_{j=1}^{n} \mathbf{g}_j \left\langle M_j^1 - M_j^2, N_j \right\rangle \right| \\
&\leq \sum_{j=1}^{n} \varepsilon_j \left\langle M_j^1 - M_j^2, N_j \right\rangle
\end{aligned}
$$

Since $\mathbf{g}_j$ are standard normal random variables, therefore with probability $1 - 2e^{-u^2/2}$,

$$
\begin{aligned}
\left| X_{(M^1,N)} - X_{(M^2,N)} \right| &\leq u \left(\sum_{j=1}^{n} \langle M_j^1 - M_j^2, N_j \rangle^2 \right)^{\frac{1}{2}} \\
&\leq u \left(\sum_{j=1}^{n} \left\| M_j^1 - M_j^2 \right\|_2^2 \|N_j\|_2^2 \right)^{\frac{1}{2}} \\
&\leq u d_F(\mathcal{N}) \left\| M^1 - M^2 \right\|_{2\to 2}
\end{aligned}
$$

Similarly with probability $1 - 2e^{-u^2/2}$,

$$
\left| X_{(M,N^1)} - X_{(M,N^2)} \right| \leq u d_F(\mathcal{M}) \left\| N^1 - N^2 \right\|_{2\to 2}
$$

Using Theorem D.5, with probability $1 - 4e^{-u^2/2}$,

$$
\sup_{M\in\mathcal{M}, N\in\mathcal{N}} \left| \sum_{j=1}^{n} \mathbf{g}_j \langle M_j, N_j \rangle \right| \lesssim u \left(d_F(\mathcal{N}) \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \right).
$$

so we have

$$
\mathbb{E}_{\mathbf{g}} \left[\sup_{M\in\mathcal{M}, N\in\mathcal{N}} \left| \sum_{j=1}^{n} \mathbf{g}_j \langle M_j, N_j \rangle \right| \right] \lesssim d_F(\mathcal{N}) \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}). \tag{37}
$$

For the second term, for each $M \in \mathcal{M}, N \in \mathcal{N}$, since ε_j are Rademacher random variables, therefore with probability at least $1 - 2e^{-u^2}$,

$$
\begin{aligned}
\left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j \rangle \right| &\leq u \left(\sum_{j=1}^{n} \langle M_j, N_j \rangle^2 \right)^{\frac{1}{2}} \leq u \left(\sum_{j=1}^{n} \|M_j\|_2^2 \|N_j\|_2^2 \right)^{\frac{1}{2}} \\
&\leq \min \left\{ d_F(\mathcal{M}) d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) d_{2\to 2}(\mathcal{M}) \right\}.
\end{aligned}
$$

According to [Vershynin, 2018, Proposition 2.5.2],

$$
\left(\mathbb{E}_{\varepsilon} \left[\left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j \rangle \right|^p \right] \right)^{1/p} \lesssim \sqrt{p} \min \left\{ d_F(\mathcal{M}) d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) d_{2\to 2}(\mathcal{M}) \right\},
$$

so

$$
\sup_{M\in\mathcal{M}, N\in\mathcal{N}} \left(\mathbb{E}_{\varepsilon} \left[\left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j \rangle \right|^p \right] \right)^{1/p} \lesssim \sqrt{p} \min \left\{ d_F(\mathcal{M}) d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) d_{2\to 2}(\mathcal{M}) \right\} \tag{38}
$$

Substituting (37) and (38) into (36), we have

$$
\left\| \sup_{M\in\mathcal{M}, N\in\mathcal{N}} \left| \sum_{j=1}^{n} \varepsilon_j \langle M_j, N_j \rangle \right| \right\|_{L_p}
$$

$$\lesssim \mathbb{E}_{\mathbf{g}}\left[\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{j=1}^{n}\mathbf{g}_j\langle M_j,N_j\rangle\right|\right]+\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left(\mathbb{E}_{\varepsilon}\left[\left|\sum_{j=1}^{n}\varepsilon_j\langle M_j,N_j\rangle\right|^{p}\right]\right)^{1/p}$$
$$\lesssim d_F(\mathcal{N})\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})+d_F(\mathcal{M})\gamma_2(\mathcal{N},\|\cdot\|_{2\to2})+\sqrt{p}\min\left\{d_F\left(\mathcal{M}\right)d_{2\to2}\left(\mathcal{N}\right),d_F\left(\mathcal{N}\right)d_{2\to2}\left(\mathcal{M}\right)\right\} \tag{39}$$

Substituting (32) and (39) into (26), we get

$$\begin{aligned}\left\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\right\|_{L_p}\lesssim\Big[&\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\Big[\gamma_2(\mathcal{N},\|\cdot\|_{2\to2})+d_F(\mathcal{N})+\sqrt{p}d_{2\to2}(\mathcal{N})\Big]\\&+\gamma_2(\mathcal{N},\|\cdot\|_{2\to2})\cdot\Big[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})+d_F(\mathcal{M})+\sqrt{p}d_{2\to2}(\mathcal{M})\Big]\\&+\sqrt{p}\min\Big\{d_F(\mathcal{M})\cdot d_{2\to2}(\mathcal{N}),d_F(\mathcal{N})\cdot d_{2\to2}(\mathcal{M})\Big\}\\&+p\cdot d_{2\to2}(\mathcal{M})\cdot d_{2\to2}(\mathcal{N})\Big].\end{aligned}$$

That completes the proof of Theorem 3.5.

E JOHNSON LINDENSTRAUSS AND RESTRICTED ISOMETRY PROPERTY

E.1 Proof of Proposition 4.1

Proof. Suppose the entries of $\tilde{X}$ be i.i.d. standard normal. For any $u\in\mathcal{U}$ we define both $M\in\mathcal{M}$ and $N\in\mathcal{N}$ as follows:

$$M=N=\frac{1}{\sqrt{n}}\begin{bmatrix}u^\top & 0 & \cdots & 0\\ 0 & u^\top & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & u^\top\end{bmatrix},$$

□

Further we define the random vector $\boldsymbol{\xi}=[\tilde{X}_{1,:},\tilde{X}_{2,:},\ldots,\tilde{X}_{n,:}]^\top$ where $\tilde{X}_{j,:}$ is the j-th row of $\tilde{X}$. Observe that

$$\begin{aligned}\sup_{u\in\mathcal{U}}\left|\|Xu\|_2^2-\mathbb{E}\|Xu\|_2^2\right|&=\sup_{M\in\mathcal{M},N\in\mathcal{N}}|\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}-\mathbb{E}\,\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}|.\\&=\sup_{M\in\mathcal{M}}\left|\,\|M\boldsymbol{\xi}\|_2^2-\mathbb{E}\|M\boldsymbol{\xi}\|_2^2\,\right|\end{aligned}$$

where $\mathcal{U}$ is a set of N vectors. Then,

$$d_F(\mathcal{M})=\sup_{M\in\mathcal{M}}\|M\|_F=\sup_{u\in\mathcal{U}}\frac{1}{\sqrt{n}}\left(\sqrt{\sum_{i=1}^{n}\|u\|_2^2}\right)=\sup_{u\in\mathcal{U}}\|u\|_2$$
$$d_{2\to2}(\mathcal{M})=\sup_{M\in\mathcal{M}}\|M\|_{2\to2}=\sup_{M\in\mathcal{M}}\sup_{w\in S^d}Mw=\sup_{u\in\mathcal{U}}\frac{1}{\sqrt{n}}\|u\|_2$$

Further using the fact that the γ_2 functional can be upper bounded by the Gaussian width up to constants [Talagrand, 2005, 2014] we get

$$\gamma_2(\mathcal{M},\|\cdot\|_2)\lesssim\omega(\mathcal{M})\lesssim\frac{1}{\sqrt{n}}\omega(\mathcal{U})$$

Now the Gaussian width of $\mathcal{U}$ is $\omega(\mathcal{U})\lesssim\sup_{u\in\mathcal{U}}\|u\|_2\sqrt{\log N}$ [Vershynin, 2018]. Therefore,

$$W=2\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})\cdot\Big[\gamma_2(\mathcal{M},\|\cdot\|_{2\to2})+d_F(\mathcal{M})\Big]$$

$$\lesssim \frac{2}{\sqrt{n}}\omega(\mathcal{U})\left[\frac{1}{\sqrt{n}}\omega(\mathcal{U}) + \sup_{u\in\mathcal{U}}\|u\|_2\right]$$
$$\lesssim 2\sup_{u\in\mathcal{U}}\|u\|_2^2\left[\frac{1}{n}\log N + \frac{1}{\sqrt{n}}\sqrt{\log N}\right]$$

$$V = 2\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})d_{2\to 2}(\mathcal{N}) + d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N})$$
$$\lesssim \frac{1}{\sqrt{n}}\omega(\mathcal{U})\ \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2 \ + \ \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2$$
$$\lesssim \frac{1}{n}\sup_{u\in\mathcal{U}}\|u\|_2^2\sqrt{\log N} \ + \ \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2$$

$$U = d_{2\to 2}(\mathcal{M})\ d_{2\to 2}(\mathcal{N}). = \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2\ \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2 = \frac{1}{n}\sup_{u\in\mathcal{U}}\|u\|_2^2.$$

Note that $\boldsymbol{\xi}$ satisfies Assumption 1 and therefore invoking Theorem 3.1 we have

$$P\left\{\sup_{u\in\mathcal{U}}|\|Xu\|_2^2 - \mathbb{E}\|Xu\|_2^2| \gtrsim \sup_{u\in\mathcal{U}}\|u\|_2^2\left[\frac{1}{n}\log N + \frac{1}{\sqrt{n}}\sqrt{\log N}\right] + \epsilon\sup_{u\in\mathcal{U}}\|u\|_2^2\right\}$$
$$\lesssim \exp\left(-\min\left\{\frac{\epsilon^2\sup_{u\in\mathcal{U}}\|u\|_2^4}{\left(\frac{1}{n}\sup_{u\in\mathcal{U}}\|u\|_2^2\sqrt{\log N} + \frac{1}{\sqrt{n}}\sup_{u\in\mathcal{U}}\|u\|_2\right)^2}, \frac{\epsilon\sup_{u\in\mathcal{U}}\|u\|_2^2}{\frac{1}{n}\sup_{u\in\mathcal{U}}\|u\|_2^2}\right\}\right)$$

Now choosing $n = \Omega(\varepsilon^{-2}\log N)$ immediately proves the theorem.

E.2 Restricted Isometric Property

E.2.1 Proof of Proposition 4.2

Suppose the entries of $\tilde{X}$ be i.i.d. standard normal. For any $u \in \mathcal{U}$ and $v \in \mathcal{V}$ we define the corresponding $M \in \mathcal{M}$ and $N \in \mathcal{N}$ as follows:

$$M = \frac{1}{\sqrt{n}}\begin{bmatrix} u^\top & 0 & \cdots & 0 \\ 0 & u^\top & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & u^\top \end{bmatrix}, \quad N = \frac{1}{\sqrt{n}}\begin{bmatrix} v & 0 & \cdots & 0 \\ 0 & v & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & v \end{bmatrix}$$

Further we define the random vector $\boldsymbol{\xi} = [\tilde{X}_{1,:}, \tilde{X}_{2,:}, \ldots, \tilde{X}_{n,:}]^\top$ where $\tilde{X}_{j,:}$ is the j-th row of $\tilde{X}$. Observe that

$$\sup_{u\in\mathcal{U},v\in\mathcal{V}}|\langle Xu, Xv\rangle - \mathbb{E}\langle Xu, Xv\rangle| = \sup_{M\in\mathcal{M},N\in\mathcal{N}}|\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}|.$$

We can now compute the various complexity measures required in Theorem 3.1 as follows:

$$d_F(\mathcal{M}) = \sup_{M\in\mathcal{M}}\|M\|_F = \sup_{u\in\mathcal{U}}\frac{1}{\sqrt{n}}\left(\sqrt{\sum_{i=1}^n\|u\|_2^2}\right) = \sup_{u\in\mathcal{U}}\|u\|_2 = 1$$
$$d_{2\to 2}(\mathcal{M}) = \sup_{M\in\mathcal{M}}\|M\|_{2\to 2} = \sup_{M\in\mathcal{M}}\sup_{w\in S^d}Mw = \sup_{u\in\mathcal{U}}\frac{1}{\sqrt{n}}\|u\|_2 = \frac{1}{\sqrt{n}}$$

Further using the fact that the γ_2 functional can be upper bounded by the Gaussian width up to constants [Talagrand, 2005, 2014] we get

$$\gamma_2(\mathcal{M}, \|\cdot\|_2) \lesssim \omega(\mathcal{M}) \lesssim \frac{1}{\sqrt{n}}\omega(\mathcal{U})$$

Similarly, $d_F(\mathcal{N}) = 1$, $d_2(\mathcal{N}) = \frac{1}{\sqrt{n}}$ and $\gamma_2(\mathcal{N}, \|\cdot\|_2) \lesssim \frac{1}{\sqrt{n}}\omega(\mathcal{V})$. We can now compute W, V and U as follows:

$$\begin{aligned}
W &= \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N})\Big] \\
&\qquad + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M})\Big], \\
&\lesssim \frac{1}{\sqrt{n}}\omega(\mathcal{U})\left[\frac{1}{\sqrt{n}}\omega(\mathcal{V}) + 1\right] + \frac{1}{\sqrt{n}}\omega(\mathcal{U})\left[\frac{1}{\sqrt{n}}\omega(\mathcal{V}) + 1\right] \\
&= \frac{2}{n}\omega(\mathcal{U})\omega(\mathcal{V}) + \frac{1}{\sqrt{n}}(\omega(\mathcal{U}) + \omega(\mathcal{V}))
\end{aligned}$$

$$\begin{aligned}
V &= \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) d_{2\to 2}(\mathcal{N}) + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) d_{2\to 2}(\mathcal{M}) \\
&\qquad + \min\Big\{d_F(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) \cdot d_{2\to 2}(\mathcal{M})\Big\} \\
&= \frac{1}{\sqrt{n}}\omega(\mathcal{U})\ \frac{1}{\sqrt{n}} \ + \ \frac{1}{\sqrt{n}}\omega(\mathcal{V})\ \frac{1}{\sqrt{n}} \ + \ \min\left\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\} \\
&= \frac{1}{n}(\omega(\mathcal{U}) + \omega(\mathcal{V})) \ + \ \frac{1}{\sqrt{n}}
\end{aligned}$$

$$\begin{aligned}
U &= d_{2\to 2}(\mathcal{M})\ d_{2\to 2}(\mathcal{N}). \\
&= \frac{1}{\sqrt{n}}\ \frac{1}{\sqrt{n}} = \frac{1}{n}.
\end{aligned}$$

Note that $\boldsymbol{\xi}$ satisfies Assumption 1 and therefore invoking Theorem 3.1 we have

$$\begin{aligned}
P\left\{\sup_{u\in\mathcal{U}, v\in\mathcal{V}} |\langle Xu, Xv\rangle - \mathbb{E}\langle Xu, Xv\rangle| \geq \frac{1}{n}\omega(\mathcal{U})\omega(\mathcal{V}) + \frac{1}{\sqrt{n}}(\omega(\mathcal{U}) + \omega(\mathcal{V})) + \epsilon\right\} \\
\lesssim \exp\left(-\min\left\{\frac{\epsilon^2}{\left(\frac{1}{n}(\omega(\mathcal{U}) + \omega(\mathcal{V})) + \frac{1}{\sqrt{n}}\right)^2}, \frac{\epsilon}{\frac{1}{n}}\right\}\right)
\end{aligned}$$

Choosing $n = \Omega\left(\frac{1}{\epsilon^2}\big(\omega(\mathcal{U}) + \omega(\mathcal{V})\big)^2\right)$ in the above we get

$$\begin{aligned}
\mathbb{P}\left\{\sup_{u\in\mathcal{U}, v\in\mathcal{V}} |\langle Xu, Xv\rangle - \mathbb{E}\langle Xu, Xv\rangle| \gtrsim \frac{\epsilon^2}{\Big(\omega(\mathcal{U}) + \omega(\mathcal{V})\Big)^2}\omega(\mathcal{U})\omega(\mathcal{V}) + \frac{\epsilon(\omega(\mathcal{U}) + \omega(\mathcal{V}))}{(\omega(\mathcal{U}) + \omega(\mathcal{V}))} + \epsilon\right\} \\
\lesssim \exp\left(-\min\left\{\frac{\epsilon^2}{\left(\frac{1}{n}(\omega(\mathcal{U}) + \omega(\mathcal{V})) + \frac{1}{\sqrt{n}}\right)^2}, \frac{\epsilon}{\frac{1}{n}}\right\}\right) \\
\lesssim \exp\left(-\min\left\{\frac{\epsilon^2}{\left(\epsilon^2 + \frac{\epsilon}{\omega(\mathcal{U})+\omega(\mathcal{V})}\right)^2}, \frac{\epsilon}{\frac{\epsilon^2}{(\omega(\mathcal{U})+\omega(\mathcal{V}))^2}}\right\}\right) \\
\lesssim \exp\left(-(\omega(\mathcal{U}) + \omega(\mathcal{V}))^2\right)
\end{aligned}$$

Now note that

$$\frac{\epsilon^2}{\Big(\omega(\mathcal{U}) + \omega(\mathcal{V})\Big)^2}\omega(\mathcal{U})\omega(\mathcal{V}) + \frac{\epsilon(\omega(\mathcal{U}) + \omega(\mathcal{V}))}{(\omega(\mathcal{U}) + \omega(\mathcal{V}))} + \epsilon \geq \epsilon^2 + 2\epsilon \gtrsim \epsilon$$

where the first inequality uses the fact that $\big(\omega(\mathcal{U}) + \omega(\mathcal{V})\big)^2 \geq \omega(\mathcal{U})\omega(\mathcal{V})$. Therefore with probability $1 - \exp\left(-c(\omega(\mathcal{U}) + \omega(\mathcal{V}))^2\right)$

$$|\langle Xu, Xv\rangle - \mathbb{E}\langle Xu, Xv\rangle| \leq \epsilon\|u\|_2\|v\|_2 \quad \text{holds for all } u \in \mathcal{U}, v \in \mathcal{V}.$$

E.2.2 Experimental Validation

We empirically validate Proposition 4.2 when U and V are *different* structured subsets of the sphere in $\mathbb{R}^p$. For each configuration we fix $p = 80$, choose structured sets $U, V \subset S^{p-1}$, and approximate

$$Z_n = \sup_{u \in U, v \in V} \left| u^\top X^\top X v - u^\top v \right|,$$

where $X \in \mathbb{R}^{n \times p}$ has i.i.d. entries $X_{ij} \sim \mathcal{N}(0, 1/n)$ so that $\mathbb{E}[X^\top X] = I_p$, over several independent draws of X, and report the Monte Carlo mean $\mathbb{E}[\widehat{Z}_n]$. Proposition 4.2 predicts that, for universal constants,

$$\mathbb{E}[Z_n] \lesssim \frac{\omega(U) + \omega(V)}{\sqrt{n}},$$

so we expect the normalized quantity

$$\frac{\sqrt{n}\,\mathbb{E}[Z_n]}{\omega(U) + \omega(V)}$$

to be constant in n. Below we consider two families of sets.

(1) s-sparse sets

Here U and V consist of s_U-sparse and s_V-sparse unit vectors in $\mathbb{R}^{80}$, respectively. The Gaussian widths are approximated analytically as

$$\omega_{\text{sparse}}(s) \approx \sqrt{2s \log(ep/s)}.$$

We report normalized deviation

$$\frac{\sqrt{n}\,\mathbb{E}[Z_n]}{\omega(U) + \omega(V)}.$$

Here we use $\omega(U) = \omega_{\text{sparse}}(s_U)$ and $\omega(V) = \omega_{\text{sparse}}(s_V)$.

$n \downarrow (s_U, s_V) \rightarrow$	$(2, 8)$	$(2, 16)$	$(4, 8)$	$(4, 16)$	$(8, 16)$
200	0.336	0.294	0.301	0.271	0.255
400	0.328	0.309	0.296	0.281	0.229
600	0.338	0.295	0.323	0.269	0.243
800	0.355	0.286	0.302	0.261	0.238
1000	0.337	0.287	0.294	0.261	0.234

For each pair (s_U, s_V), the normalized quantity stays in a narrow band as n varies, confirming that the $1/\sqrt{n}$ dependence is fully captured by $\sqrt{n}$ and the scale is determined by $\omega(U) + \omega(V)$.

(2) ℓ_1-constrained sets

We now let U and V be ℓ_1-constrained unit vectors:

$$U = \{u \in \mathbb{R}^d : \|u\|_2 = 1, \|u\|_1 \le L_u\}, \qquad V = \{v \in \mathbb{R}^{80} : \|v\|_2 = 1, \|v\|_1 \le L_v\}.$$

The Gaussian width is approximated by the standard bound

$$\omega_{\ell_1}(L) \lesssim L\sqrt{2 \log p},$$

and we take $\omega(U) = \omega_{\ell_1}(L_u)$, $\omega(V) = \omega_{\ell_1}(L_v)$, and report the normalized deviation

$$\frac{\sqrt{n}\,\mathbb{E}[Z_n]}{\omega(U) + \omega(V)}.$$

$n \downarrow (L_u, L_v) \rightarrow$	$(4, 4)$	$(4, 6)$	$(4, 8)$	$(6, 8)$	$(6, 10)$
200	0.172	0.141	0.112	0.097	0.085
400	0.175	0.127	0.111	0.094	0.088
600	0.160	0.128	0.111	0.095	0.083
800	0.161	0.135	0.108	0.094	0.092
1000	0.166	0.133	0.107	0.093	0.082

For each (L_u, L_v) the normalized curves are nearly flat across n, with magnitudes decreasing as (L_u, L_v) (and hence $\omega(U) + \omega(V)$) increase, which is consistent with the predicted linear dependence on the Gaussian widths.

In both families, the values from Tables 1 and 2 are small (on the order of 10^{-1}), indicating that the theoretical scaling in $(\omega(U) + \omega(V))^2/\varepsilon^2$ is not only qualitatively correct but also quantitatively informative for picking n in practice: one can achieve a target accuracy ε with a sample size proportional to this quantity and a moderate constant factor.

F APPLICATION IN DISTRIBUTED LEARNING

F.1 Sketching-based Distributed Learning

We outline the sketching-based distributed learning framework in Algorithm 2. Each client receives a random seed from the server to initialize the local parameters $\theta_{c,1}$, and generate a sketching matrix R. At each local step $k \in [1, \ldots, K]$, each client performs local gradient descent (GD) over their local dataset $\mathcal{D}_c$. At each communication round, the client accumulates the changes over K-local steps, sketches the local updates, and sends the sketched update to the server. The server then aggregates the sketched changes and sends the aggregated sketched updates back to the clients. To update the local parameters, each client needs to recover an unbiased estimate of the true vector from the aggregated sketched update. We call this the desk (desketching) operation (Line 9), for which we use the transpose of the sketching matrix R. Each client then desketches the received aggregated sketched updates by applying desk and updates their local parameters. We refer to the sketching and desketching operations using the sk and desk operators defined as:

$$\texttt{sk} := R \in \mathbb{R}^{b\times p} \quad \text{(Sketching)},$$
$$\texttt{desk} := R^\top \in \mathbb{R}^{p\times b} \quad \text{(Desketching)}.$$

Denote $R_t \in \mathbb{R}^{b\times p}$ the sketching matrix at round $t \in [T]$.

Choice of sketching matrix: We use a $(1/\sqrt{b})$-sub-Gaussian matrix as the choice of sketching matrix. We say $R \in \mathbb{R}^{b\times p}$ is a $(1/\sqrt{b})$-sub-Gaussian matrix if each row R_i is an independent mean-zero, sub-Gaussian isotropic random-vector such that $\|R_i\|_{\psi_2} \le 1/\sqrt{b}$. We assume $\mathbb{E}[R^\top R] = I_{p\times p}$. From the above definition, we can see that for $g_1, g_2 \in \mathbb{R}^p$

$$R(g_1 + g_2) = Rg_1 + Rg_2 \quad \text{(Linearity)},$$
$$\mathbb{E}_{R\sim\Pi}\left[R^\top R\, g\right] = g \quad \text{(Unbiasedness)}.$$

In line with standard works Shrivastava et al. [2024], Banerjee et al. [2023], we study a fully-connected feedforward neural network f of depth L, where each layer $l \in [L] := 1, \ldots, L$ has associated activations $\alpha^{(l)}$, defined recursively as:

$$\alpha^{(0)}(x) = x,$$
$$\alpha^{(l)}(x) = \phi\left(\tfrac{1}{\sqrt{m_{l-1}}}\, W_t^{(l)}\, \alpha^{(l-1)}(x)\right), \quad \forall\, l \in [L],$$
$$f(\theta; x) = \alpha^{(L+1)}(x) = \frac{1}{\sqrt{m_L}}\, v_t^\top\, \alpha^{(L)}(x),$$

where $W_t^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ denotes the weight matrix at the l^{th} layer, $v_t \in \mathbb{R}^{m_L}$ is the output-layer vector at time t, and ϕ is a smooth pointwise activation function. The input dimension is $m_0 = \dim(x) = d$. The complete parameter vector at iteration t is given by

$$\theta_t = \left(\left(\operatorname{vec} W_t^{(1)}\right)^\top, \ldots, \left(\operatorname{vec} W_t^{(L)}\right)^\top, v_t^\top\right)^\top \in \mathbb{R}^p.$$

For notational simplicity, we assume all hidden layers have uniform width m, i.e., $m_l = m$ for every $l \in [L]$, so that the total number of parameters is $p = (L-1)m^2 + md + m$. We restrict our focus to scalar-valued networks $f(\theta; x) \in \mathbb{R}$, though the analysis extends naturally to multi-output settings.

Algorithm 2 Sketching-Based Distributed Learning.

Hyperparameters: server learning rate η_{global}, local learning rate η_{local}.
Inputs: local datasets $\mathcal{D}_c$ of size n_c for clients $c = 1, \ldots, C$, number of communication rounds T.
Output: final model θ_T.

Broadcast a random SEED to the clients.
for $t = 0, \ldots, T$ **do**
  On Client Nodes:
  for $c = 1, \ldots, C$ **do**
    if $t = 0$ **then**
      Initialize the local model $\theta_0 = \theta_{c,0,0} \in \mathbb{R}^p$.
    else
      Receive $\text{sk}(\bar{\Delta}_{t-1})$ from the server.
      Desketch and update the model parameters $\theta_t \leftarrow \theta_{t-1} + \texttt{desk}(\texttt{sk}(\bar{\Delta}_{t-1}))$.
      Assign the local model's parameters $\theta_{c,t,0} \leftarrow \theta_t$ to be updated locally.
    end if
    for $k = 1, \ldots, K$ **do**
      $\theta_{c,t,k} \leftarrow \theta_{c,t,k-1} - \eta_{\text{local}} \cdot \nabla \mathcal{L}_c(\theta_{c,t,k})$
    end for
    $\Delta_{c,t} \leftarrow \theta_t - \theta_{c,t,K}$
    Send $\texttt{sk}(\Delta_{c,t})$ to the server.
  end for
  On the Server Node:
  Receive $\texttt{sk}(\Delta_{c,t})$ from clients $c = 1, \ldots, C$.
  Aggregate: $\texttt{sk}(\bar{\Delta}_t) \leftarrow \eta_{\text{global}} \cdot \frac{1}{C} \sum_{c=1}^{C} \texttt{sk}(\Delta_{c,t})$
  Broadcast $\texttt{sk}(\bar{\Delta}_t)$ to the clients.
end for

We now state the standard assumptions on the activation function, loss function, and initialization, which are satisfied by most commonly used choices in practice.

Assumption 6 (Activation function)**.** *The activation ϕ is 1-Lipschitz and β_ϕ-smooth, that is, $|\phi'| \leq 1$ and $|\phi''| \leq \beta_\phi$.*

Assumption 7 (Initialization)**.** *The initial parameters are drawn as $w_{0,ij}^{(l)} \sim \mathcal{N}(0, \sigma_0^2)$ for each layer $l \in [L]$, where $\sigma_0 = \frac{\sigma_1}{2\left(1+\frac{\sqrt{\log m}}{\sqrt{2m}}\right)}$ with $\sigma_1 > 0$. The final layer vector v_0 is a random unit vector with $|v_0|_2 = 1$.*

Assumption 8 (Loss function)**.** *Let $\ell_{i,c} = \ell(y_{i,c}, \hat{y}i, c)$ denote the per-example loss, with first and second derivatives $\ell' i, c = \frac{d\ell_{i,c}}{d\hat{y}i,c}$ and $\ell'' i, c = \frac{d^2\ell_{i,c}}{d\hat{y}i,c^2}$. The loss satisfies: (i) Lipschitz continuity: $|\ell' i, c| \leq c_\ell$, and (ii) Smoothness: $\ell'' i, c \leq c_s$, for some constants $c\ell, c_s > 0$.*

F.2 Convergence Analysis for Sketching-based Distributed Learning

We begin by presenting a set of standard assumptions that are widely adopted in the literature on first-order stochastic methods.

Assumption 9 (Bounded Loss Gradients)**.** *There exists a constant $G \geq 0$, such that for every $(\mathbf{x}, y)$, $\|\nabla \ell(\theta; (x, y))\|_2 \leq G$.*

Assumption 10 (**Anistropic Loss Hessian**)**.** *For any loss Hessian $H_{i,c,t}$, assume there exists a fixed positive definite $\mathbf{H}$ such that $-\mathbf{H} \preceq H_{i,c,t} \preceq \mathbf{H}$. Let $\Lambda_1, \ldots, \Lambda_p$ be the eigenvalues of $\mathbf{H}$ and define $\Lambda_{\max} = \max_j |\Lambda_j|$. Then there exists a constant $\kappa = \mathcal{O}(1)$ such that $\sum_{j=1}^{p} |\Lambda_j| \leq \kappa \Lambda_{\max}$.*

From Theorem 3.1, we can derive the following sketching guarantee.

Theorem F.1. *Let $\mathcal{G}$ and $\mathcal{H}$ be set of d-dimensional vectors respectively, and let $R \in \mathbb{R}^{b \times d}$ denote a random*

$\frac{1}{\sqrt{b}}$-*sub-Gaussian matrix. Denote* $d_2(\mathcal{G}) = \sup_{g\in\mathcal{G}} \|g\|_2$ *and* $d_2(\mathcal{H}) = \sup_{h\in\mathcal{H}} \|h\|_2$. *We define*

$$A = \frac{w(\mathcal{G})w(\mathcal{H})}{b} + \frac{w(\mathcal{G})d_2(\mathcal{H}) + w(\mathcal{H})d_2(\mathcal{G})}{\sqrt{b}},$$
$$B = \frac{w(\mathcal{G})d_2(\mathcal{H}) + w(\mathcal{H})d_2(\mathcal{G})}{b} + \frac{d_2(\mathcal{G})d_2(\mathcal{H})}{\sqrt{b}}$$
$$C = \frac{d_2(\mathcal{G})d_2(\mathcal{H})}{b}.$$

Then, for $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G},h\in\mathcal{H}} \left|g^\top R^\top Rh - g^\top h\right| \geq c_1 A + \epsilon\right) \leq 2\exp\left(-c_2 \min\left\{\frac{\epsilon^2}{B^2}, \frac{\epsilon}{C}\right\}\right).$$

where the constants c_1, c_2 *are absolute constants. Equivalently, with probability at least* $1-2\delta$,

$$\sup_{g\in\mathcal{G},h\in\mathcal{H}} \left|g^\top R^\top Rh - g^\top h\right| \lesssim Z(\mathcal{G},\mathcal{H},\delta) := A + \sqrt{\log(1/\delta)}B + \log(1/\delta)C \tag{40}$$

Proof. Denote $r = \left[\sqrt{b}R_{1,:}, \sqrt{b}R_{2:}, \ldots, \sqrt{b}R_{b:}\right]^\top$ be a 1-sub-Gaussian random vector of length bd by concatenating the rows of R and rescaling by $\sqrt{b}$. For any $g \in \mathcal{G}$ and $h \in \mathcal{H}$, define

$$M_g := \frac{1}{\sqrt{b}}\begin{bmatrix} g^T & 0 & \cdots & 0 \\ 0 & g^T & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & g^T \end{bmatrix}, N_h = \frac{1}{\sqrt{b}}\begin{bmatrix} h^T & 0 & \cdots & 0 \\ 0 & h^T & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & h^T \end{bmatrix}$$

so $M_g \in \mathbb{R}^{b\times bd}$ and $N_h \in \mathbb{R}^{b\times bd}$ are block diagonal matrices, and $g^\top R^\top Rh = r^\top M_g^\top N_h r$. Denote $\mathcal{M}_\mathcal{G}$ and $\mathcal{N}_\mathcal{H}$ be the sets of such M_g and N_h respectively. According to results in Talagrand [2014], we can see that $\gamma_2(\mathcal{M}_\mathcal{G}, \|\cdot\|_{2\to2}) \leq C\frac{w(\mathcal{G})}{\sqrt{b}}$, $\gamma_2(\mathcal{N}_\mathcal{H}, \|\cdot\|_{2\to2}) \leq C\frac{w(\mathcal{H})}{\sqrt{b}}$. In addition, we can get that $d_F(\mathcal{M}_\mathcal{G}) = d_2(\mathcal{G})$, $d_F(\mathcal{N}_\mathcal{H}) = d_2(\mathcal{H})$, $d_{2\to2}(\mathcal{M}_\mathcal{G}) = \frac{1}{\sqrt{b}}d_2(\mathcal{G})$, $d_{2\to2}(\mathcal{N}_\mathcal{H}) = \frac{1}{\sqrt{b}}d_2(\mathcal{H})$. Therefore,

$$\begin{aligned}
W &= \gamma_2(\mathcal{M}_\mathcal{G}, \|\cdot\|_{2\to2})\cdot\left[\gamma_2(\mathcal{N}_\mathcal{H}, \|\cdot\|_{2\to2}) + d_F(\mathcal{N}_\mathcal{H})\right] + \gamma_2(\mathcal{N}_\mathcal{H}, \|\cdot\|_{2\to2})\cdot\left[\gamma_2(\mathcal{M}_\mathcal{G}, \|\cdot\|_{2\to2}) + d_F(\mathcal{M}_\mathcal{G})\right] \\
&\lesssim \frac{w(\mathcal{G})}{\sqrt{b}}\cdot\left(\frac{w(\mathcal{H})}{\sqrt{b}} + d_2(\mathcal{H})\right) + \frac{w(\mathcal{H})}{\sqrt{b}}\left(\frac{w(\mathcal{G})}{\sqrt{b}} + d_2(\mathcal{G})\right) \\
&\lesssim \frac{w(\mathcal{G})w(\mathcal{H})}{b} + \frac{w(\mathcal{G})d_2(\mathcal{H}) + w(\mathcal{H})d_2(\mathcal{G})}{\sqrt{b}} = A, \\
V &= \gamma_2(\mathcal{M}_\mathcal{G}, \|\cdot\|_{2\to2})d_{2\to2}(\mathcal{N}_\mathcal{H}) + \gamma_2(\mathcal{N}_\mathcal{H}, \|\cdot\|_{2\to2})d_{2\to2}(\mathcal{M}_\mathcal{G}) \\
&\qquad + \min\left\{d_F(\mathcal{M}_\mathcal{G})\cdot d_{2\to2}(\mathcal{N}_\mathcal{H}), d_F(\mathcal{N}_\mathcal{H})\cdot d_{2\to2}(\mathcal{M}_\mathcal{G})\right\} \\
&= \frac{w(\mathcal{G})}{\sqrt{b}}\cdot\frac{d_2(\mathcal{H})}{\sqrt{b}} + \frac{w(\mathcal{H})}{\sqrt{b}}\cdot\frac{d_2(\mathcal{G})}{\sqrt{b}} + \min\left\{d_2(\mathcal{G})\cdot\frac{d_2(\mathcal{H})}{\sqrt{b}}, d_2(\mathcal{H})\cdot\frac{d_2(\mathcal{G})}{\sqrt{b}}\right\} \\
&= \frac{w(\mathcal{G})d_2(\mathcal{H}) + w(\mathcal{H})d_2(\mathcal{G})}{b} + \frac{d_2(\mathcal{G})d_2(\mathcal{H})}{\sqrt{b}} = B, \\
U &= d_{2\to2}(\mathcal{M}_\mathcal{G})\, d_{2\to2}(\mathcal{N}_\mathcal{H}) = \frac{d_2(\mathcal{G})}{\sqrt{b}}\cdot\frac{d_2(\mathcal{H})}{\sqrt{b}} = \frac{d_2(\mathcal{G})d_2(\mathcal{H})}{b} = C.
\end{aligned}$$

According to Theorem 3.1, we can see that

$$\mathbb{P}\left(\sup_{g\in\mathcal{G},h\in\mathcal{H}} \left|g^\top R^\top Rh - g^\top h\right| \geq c_1 A + \epsilon\right)$$

$$
\begin{aligned}
=&\mathbb{P}\left(\sup_{M_g\in\mathcal{M}_\mathcal{G},N_h\in\mathcal{N}_\mathcal{H}}\left|r^\top M_g^\top N_h r-M_g^\top N_h^\top\right|>c_1A+\epsilon\right)\\
\leq&\mathbb{P}\left(\sup_{M_g\in\mathcal{M}_\mathcal{G},N_h\in\mathcal{N}_\mathcal{H}}\left|r^\top M_g^\top N_h r-M_g^\top N_h^\top\right|>c_1W+\epsilon\right)\\
\leq&2\exp\left(-c_2\min\left\{\frac{\epsilon^2}{V^2},\frac{\epsilon}{U}\right\}\right)\\
\leq&2\exp\left(-c_2\min\left\{\frac{\epsilon^2}{B^2},\frac{\epsilon}{C}\right\}\right)
\end{aligned}
$$

then we finish the proof. □

Theorem 4.1. *Suppose Assumptions 2 and 3 hold and let* $\|\nabla_\theta\ell(\theta)\|\leq G$. *For suitable constant* $\varepsilon<1$, *width of the network* m, *learning rate* η, *sketching dimension* $b=\Omega\Big(\frac{1}{\varepsilon^2}\operatorname{polylog}\big(\frac{TNp^2}{\delta}\big)\Big)$, *and* $C_2(m,\kappa):=\mathcal{O}\Big(\frac{\varepsilon\kappa}{\sqrt{m}}+\frac{1}{\sqrt{m}}\Big)$, *with probability at least* $1-\delta$, *we have*

$$
\begin{aligned}
\mathcal{L}(\theta_T)-\mathcal{L}(\theta^*)\leq&\left(\mathcal{L}(\theta_0)-\mathcal{L}(\theta^*)\right)e^{-2\mu\eta KT}\\
&+\frac{\left(\eta KC_2(m,\kappa)+C_3\right)\left(G^2+\epsilon^2\right)}{2\mu}.
\end{aligned}
\tag{8}
$$

Proof. Denote $\mathcal{G}$ and $\mathcal{H}$ the sets of all possible loss gradients and the eigenvectors of $\mathbf{H}$ respectively. According to the algorithm, we can write the update in the sync step as:

$$
\begin{aligned}
\theta_{t+1}-\theta_t&=-\eta_{\text{global}}\texttt{desk}\left(\texttt{sk}\left(\bar{\Delta}_{t-1}\right)\right)\\
&=-\eta_{\text{global}}\texttt{desk}\left(\frac{1}{C}\sum_{c=1}^{C}\Delta_{c,t}\right)\\
&=-\eta_{\text{global}}\texttt{desk}\left(\frac{1}{C}\sum_{c=1}^{C}\texttt{sk}\left(\eta_{\text{local}}\sum_{k=1}^{K}\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right)\\
&=-\eta_{\text{global}}R_t^\top\left(\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}R_t\sum_{k=1}^{K}\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\\
&=-\eta\frac{1}{C}\sum_{c=1}^{C}R_t^\top R_t\sum_{k=1}^{K}\nabla\mathcal{L}_c(\theta_{c,t,k})
\end{aligned}
$$

in which $\eta=\eta_{\text{global}}\eta_{\text{local}}$. By Taylor expansion, we have

$$
\mathcal{L}(\theta_{t+1})=\mathcal{L}(\theta_t)+\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)+\frac{1}{2}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)
\tag{41}
$$

Bounding T_1:

For each term in T_1, we have

$$
\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)\\
=&-\nabla\mathcal{L}(\theta_t)^\top\left(\frac{\eta}{C}\sum_{c=1}^{C}R_t^\top R_t\sum_{k=1}^{K}\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\\
=&-\frac{\eta}{C}\nabla\mathcal{L}\left(\theta_t\right)^\top\sum_{c=1}^{C}R_t^\top R_t\sum_{k=1}^{K}\nabla\mathcal{L}_c(\theta_{c,t,k})\\
=&-\eta K\left\langle\nabla\mathcal{L}\left(\theta_t\right),\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c\left(\theta_t\right)\right\rangle-\eta\left\langle\nabla\mathcal{L}\left(\theta_t\right),\frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c\left(\theta_{c,t,k}\right)-\nabla\mathcal{L}_c\left(\theta_t\right)\right)\right\rangle
\end{aligned}
$$

$$-\eta\left\langle \nabla\mathcal{L}\left(\theta_t\right), \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(R_t^\top R_t \nabla\mathcal{L}_c\left(\theta_{c,t,k}\right) - \nabla\mathcal{L}_c\left(\theta_{c,t,k}\right)\right)\right\rangle \tag{42}$$

For the first term, according to the definition,

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle = \|\nabla\mathcal{L}(\theta_t)\|_2^2$$

so

$$\sum_{t=0}^{T-1}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle = \sum_{t=0}^{T-1}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \tag{43}$$

For the second term, we have

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \\
=&\frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), \hat{H}_{\mathcal{L}}^{c,t,k}\left(\theta_{c,t,k} - \theta_t\right)\right\rangle \\
=&\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), \hat{H}_{\mathcal{L}}^{c,t,k}\sum_{\kappa=1}^{k} g_{c,t,\kappa}\right\rangle \\
\geq&-\frac{\eta_{\text{local}}}{N}\sum_{c=1}^{C}\sum_{k=1}^{K}\|\nabla\mathcal{L}(\theta_t)\|_2\cdot\Lambda_{\max}\sum_{\kappa=1}^{k}\|g_{c,t,\kappa}\|_2 \\
=&-\frac{\eta_{\text{local}}}{N}\cdot N\cdot G^2\cdot\Lambda_{\max}\sum_{k=1}^{K}k \\
\geq&-\frac{\eta_{\text{local}}\Lambda_{\max}K^2G^2}{2};
\end{aligned} \tag{44}$$

For the third term, according to Theorem F.1, we have that with probability at least $1-\frac{2\delta}{T}$,

$$\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}\left(\theta_t\right), \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\right| \\
=&\frac{1}{C}\left|\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\nabla\mathcal{L}(\theta_t)^\top R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}(\theta_t)^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right| \\
\leq&K\sup_{g\in\mathcal{G},h\in\mathcal{G}}\left|g^\top R_t^\top R_t h - g^\top h\right| \\
\leq&KZ\left(\mathcal{G},\mathcal{G},\frac{\delta}{T}\right) \\
=&K\left(\frac{w^2(\mathcal{G})}{b} + \frac{Gw(\mathcal{G})}{\sqrt{b}} + \frac{Gw(\mathcal{G})\sqrt{\log(T/\delta)}}{b} + \frac{G^2\sqrt{\log(T/\delta)}}{\sqrt{b}} + \frac{G^2\log(T/\delta)}{b}\right) \\
=&\frac{Kw^2(\mathcal{G})}{b} + \frac{KG}{\sqrt{b}}\left(w(\mathcal{G}) + G\sqrt{\log(T/\delta)}\right)\left(1+\frac{\sqrt{\log(T/\delta)}}{\sqrt{b}}\right)
\end{aligned} \tag{45}$$

Substituting (43)-(45) into (42), we have that with probability at least $1-\frac{2\delta}{T}$,

$$\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t) \\
=&-\eta K\left\langle \nabla\mathcal{L}\left(\theta_t\right), \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c\left(\theta_t\right)\right\rangle - \eta\left\langle \nabla\mathcal{L}\left(\theta_t\right), \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c\left(\theta_{c,t,k}\right) - \nabla\mathcal{L}_c\left(\theta_t\right)\right)\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&-\eta\left\langle\nabla\mathcal{L}\left(\theta_t\right), \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle \\
\leq&-\eta K\left\|\nabla\mathcal{L}(\theta_t)\right\|_2^2+\frac{\eta\eta_{\text{local}}\Lambda_{\max}K^2G^2}{2}+\eta KZ\left(\mathcal{G},\mathcal{G},\frac{\delta}{T}\right)
\end{aligned}
\tag{46}
$$

in which

$$
Z\left(\mathcal{G},\mathcal{G},\frac{\delta}{T}\right)=\frac{w^2(\mathcal{G})}{b}+\frac{G}{\sqrt{b}}\left(w(\mathcal{G})+G\sqrt{\log(T/\delta)}\right)\left(1+\frac{\sqrt{\log(T/\delta)}}{\sqrt{b}}\right)
\tag{47}
$$

Bounding T_2:

According to Assumption 10, we can get that

$$
\begin{aligned}
\left|\left(\theta_{t+1}-\theta_t\right)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)\right|&\leq\left(\theta_{t+1}-\theta_t\right)^\top\mathbf{H}(\theta_{t+1}-\theta_t) \\
&=\eta_{\text{global}}^2\left(\texttt{desk}\left(\bar{\tilde{\Delta}}_t\right)\right)^\top\left(\sum_{i=1}^{p}\Lambda_i v_i v_i^\top\right)\left(\texttt{desk}\left(\bar{\Delta}_t\right)\right) \\
&=\eta_{\text{global}}^2\sum_{i=1}^{p}\Lambda_i\left|\left(\texttt{desk}\left(\bar{\Delta}_t\right)\right)^\top v_i\right|^2
\end{aligned}
\tag{48}
$$

For each $i\in[d]$, we have

$$
\begin{aligned}
\left|\left(\texttt{desk}(\bar{\Delta}_t)\right)^\top v_i\right|&=\left|\left\langle\texttt{desk}(\bar{\tilde{\Delta}}_t),v_i\right\rangle\right| \\
&=\frac{\eta_{\text{local}}}{C}\left|\left\langle\texttt{desk}\left(\sum_{c=1}^{C}\texttt{sk}\left(\Delta_{c,t}\right)\right),v_i\right\rangle\right| \\
&=\frac{\eta_{\text{local}}}{C}\left|\left\langle\sum_{c=1}^{C}R_t^\top R_t\Delta_{c,t},v_i\right\rangle\right| \\
&=\frac{\eta_{\text{local}}}{C}\left|\left\langle\sum_{c=1}^{C}\sum_{k=1}^{K}R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k}),v_i\right\rangle\right|
\end{aligned}
$$

According to Theorem F.1, we have that with probability at least $1-\frac{2\delta}{dT}$,

$$
\begin{aligned}
&\left|\left(\texttt{desk}(\bar{\Delta}_t)\right)^\top v_i\right| \\
=&\left|\frac{\eta_{\text{local}}}{C}\left\langle\sum_{c=1}^{C}\sum_{k=1}^{K}R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k}),v_i\right\rangle\right| \\
=&\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(v_i^\top R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k})-v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right)+\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right| \\
\leq&\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(v_i^\top R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k})-v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right|+\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right| \\
\leq&\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(v_i^\top R_t^\top R_t\nabla\mathcal{L}_c(\theta_{c,t,k})-v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right|+\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right| \\
\leq&\eta_{\text{local}}K\sup_{g\in\mathcal{G},h\in\mathcal{H}}\left|g^\top R_t^\top R_t h-g^\top h\right|+\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right| \\
\leq&\eta_{\text{local}}KZ\left(\mathcal{G},\mathcal{H},\frac{\delta}{T}\right)+\left|\frac{\eta_{\text{local}}}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}v_i^\top\nabla\mathcal{L}_c(\theta_{c,t,k})\right|
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{\eta_{\text{local}} K w(\mathcal{G}) w(\mathcal{H})}{b}+\frac{\eta_{\text{local}} K w(\mathcal{G})+\eta_{\text{local}} K G w(\mathcal{H})}{\sqrt{b}}+\frac{\eta_{\text{local}} K \sqrt{\log (T / \delta)}\left(w(\mathcal{G})+G w(\mathcal{H})\right)}{b} \\
&+\frac{\eta_{\text{local}} K G \sqrt{\log (T / \delta)}}{\sqrt{b}}+\frac{\eta_{\text{local}} K G \log (T / \delta)}{b}+\left|\frac{\eta_{\text{local}}}{C} \sum_{c=1}^{C} \sum_{k=1}^{K} v_i^{\top} \nabla \mathcal{L}_c\left(\theta_{c, t, k}\right)\right| \qquad (49)
\end{aligned}
$$

Substituting (49) into (48), we have that with probability at least $1-\frac{2\delta}{T}$,

$$
\begin{aligned}
\left(\theta_{t+1}-\theta_t\right)^{\top} \hat{H}_{\mathcal{L}, t}\left(\theta_{t+1}-\theta_t\right) & \leq \eta_{\text{global}}^2 \sum_{i=1}^{p} \Lambda_i\left|\left(\texttt{desk}\left(\bar{\Delta}_t\right)\right)^{\top} v_i\right|^2 \\
&=\eta_{\text{global}}^2 \sum_{i=1}^{p} \Lambda_i\left(\eta_{\text{local}} K Z\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)+\left|\frac{\eta_{\text{local}}}{C} \sum_{c=1}^{C} \sum_{k=1}^{K} v_i^{\top} \nabla \mathcal{L}_c\left(\theta_{c, t, k}\right)\right|\right)^2 \\
&\leq 2 \eta^2 K^2 Z^2\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right) \sum_{i=1}^{p}\left|\Lambda_i\right|+2 \eta^2 \sum_{i=1}^{p} \Lambda_i\left(\frac{1}{C} \sum_{c=1}^{C} \sum_{k=1}^{K} v_i^{\top} \nabla \mathcal{L}_c\left(\theta_{c, t, k}\right)\right)^2 \\
&\leq 2 \eta^2 K^2 \kappa \Lambda_{\max } Z^2\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)+2 \Lambda_{\max } \eta^2 K^2 G^2 \qquad (50)
\end{aligned}
$$

in which

$$
\begin{aligned}
Z\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)=&\frac{w(\mathcal{G}) w(\mathcal{H})}{b}+\frac{w(\mathcal{G})+G w(\mathcal{H})}{\sqrt{b}}+\frac{\sqrt{\log (T / \delta)}\left(w(\mathcal{G})+G w(\mathcal{H})\right)}{b} \\
&+\frac{G \sqrt{\log (T / \delta)}}{\sqrt{b}}+\frac{G \log (T / \delta)}{b} \qquad (51)
\end{aligned}
$$

Substituting (46) and (50) into (41), and according to Assumption 2, we have that with probability at least $1-\frac{4\delta}{T}$,

$$
\begin{aligned}
&\mathcal{L}\left(\theta_{t+1}\right)-\mathcal{L}\left(\theta_t\right) \\
=&\nabla \mathcal{L}\left(\theta_t\right)^{\top}\left(\theta_{t+1}-\theta_t\right)+\frac{1}{2}\left(\theta_{t+1}-\theta_t\right)^{\top} \hat{H}_{\mathcal{L}, t}\left(\theta_{t+1}-\theta_t\right) \\
\leq&-\eta K\left\|\nabla \mathcal{L}\left(\theta_t\right)\right\|_2^2+\frac{\eta \eta_{\text{local}} \Lambda_{\max } K^2 G^2}{2}+\eta K Z\left(\mathcal{G}, \mathcal{G}, \frac{\delta}{T}\right) \\
&+2 \eta^2 K^2 \kappa \Lambda_{\max }\left(Z^2\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)+G^2\right) \\
\leq&-\eta K \cdot 2 \mu\left(\nabla \mathcal{L}\left(\theta_t\right)-\mathcal{L}^*\right)+\frac{\eta \eta_{\text{local}} \Lambda_{\max } K^2 G^2}{2}+\eta K Z\left(\mathcal{G}, \mathcal{G}, \frac{\delta}{T}\right) \\
&+2 \eta^2 K^2 \kappa \Lambda_{\max }\left(Z^2\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)+G^2\right)
\end{aligned}
$$

According to [Banerjee et al., 2024, Theorem 4], $w(\mathcal{G}) \leq O(1)$ when $L=\Omega(\log m)$. Since $\mathcal{H}$ is the finite set of d eigenvectors of $\mathbf{H}$, $w(\mathcal{H}) \leq \log p$ Vershynin [2018]. Substituting these adn $b=\frac{\text{polylog}\left(\frac{TNp^2}{\delta}\right)}{\varepsilon^2}$ into (47) and (51), we have

$$
\begin{aligned}
Z\left(\mathcal{G}, \mathcal{G}, \frac{\delta}{T}\right)=&\frac{w^2(\mathcal{G})}{b}+\frac{G}{\sqrt{b}}\left(w(\mathcal{G})+G \sqrt{\log (T / \delta)}\right)\left(1+\frac{\sqrt{\log (T / \delta)}}{\sqrt{b}}\right) \\
\lesssim&\frac{1}{b}+\frac{G}{\sqrt{b}}\left(1+G \sqrt{\log (T / \delta)}\right)\left(1+\frac{\sqrt{\log (T / \delta)}}{\sqrt{b}}\right) \\
\lesssim&G^2+\varepsilon^2; \\
Z\left(\mathcal{G}, \mathcal{H}, \frac{\delta}{T}\right)=&\frac{w(\mathcal{G}) w(\mathcal{H})}{b}+\frac{w(\mathcal{G})+G w(\mathcal{H})}{\sqrt{b}}+\frac{\sqrt{\log (T / \delta)}\left(w(\mathcal{G})+G w(\mathcal{H})\right)}{b}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{G\sqrt{\log(T/\delta)}}{\sqrt{b}} + \frac{G\log(T/\delta)}{b} \\
&\lesssim \frac{\log d}{b} + \frac{1+G\log d}{\sqrt{b}} + \frac{\sqrt{\log(T/\delta)}\,(1+G\log d)}{b} + \frac{G\sqrt{\log(T/\delta)}}{\sqrt{b}} + \frac{G\log(T/\delta)}{b} \\
&\lesssim \varepsilon\,(1+G)\,.
\end{aligned}
$$

In addition, according to Banerjee et al. [2023], $\Lambda_{\max} \le \frac{c_H c_l}{\sqrt{m}} + c_s\varrho^2$. Combining these with $\eta = \eta_{\text{local}}$, we can get that

$$
\begin{aligned}
&\mathcal{L}(\theta_{t+1}) - \mathcal{L}^* \\
&\le (1-2\mu\eta K)\cdot(\nabla\mathcal{L}(\theta_t) - \mathcal{L}^*) + \frac{\eta\eta_{\text{local}}\Lambda_{\max}K^2G^2}{2} + \eta K Z\left(\mathcal{G},\mathcal{G},\frac{\delta}{T}\right) + 2\eta^2K^2\kappa\Lambda_{\max}Z^2\left(\mathcal{G},\mathcal{H},\frac{\delta}{T}\right) + 2\eta^2 \\
&\le (1-2\mu\eta K)\cdot(\nabla\mathcal{L}(\theta_t) - \mathcal{L}^*) + \frac{\eta^2K^2G^2}{2}\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right) + c_1\eta K(G^2+\varepsilon^2) \\
&\quad + 2\eta^2K^2\kappa\cdot\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right)\cdot c_2\varepsilon^2\left(G^2+1\right) + 2\eta^2K^2G^2\cdot\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right) \\
&= (1-2\mu\eta K)\cdot(\nabla\mathcal{L}(\theta_t) - \mathcal{L}^*) + \frac{5\eta^2K^2G^2}{2}\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right) + c_1\eta K(G^2+\varepsilon^2) \\
&\quad + 2\eta^2K^2\kappa\cdot\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right)\cdot c_2\varepsilon^2\left(G^2+1\right)
\end{aligned}
$$

Iterating from 0 to $T-1$, we can get that

$$
\begin{aligned}
&\mathcal{L}(\theta_T) - \mathcal{L}^* \\
&\le (1-2\mu\eta K)^T\left(\mathcal{L}(\theta_0) - \mathcal{L}^*\right) + \frac{1}{2\mu\eta K}\Big(\frac{5\eta^2K^2G^2}{2}\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right) + c_1\eta K(G^2+\varepsilon^2) \\
&\quad + 2\eta^2K^2\kappa\cdot\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right)\cdot c_2\varepsilon^2\left(G^2+1\right)\Big) \\
&\le \left(\mathcal{L}(\theta_0) - \mathcal{L}^*\right)e^{-2\mu\eta KT} + \frac{5\eta KG^2}{4\mu}\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right) + \frac{c_1\left(G^2+\varepsilon^2\right)}{2\mu} + \frac{\eta K\kappa\varepsilon^2\left(G^2+1\right)}{\mu}\left(\frac{c_Hc_l}{\sqrt{m}} + c_s\varrho^2\right)
\end{aligned}
$$

then we finish the proof. □

G LINEAR REGRESSION WITH SKETCHING

We observe n i.i.d. samples $(\mathbf{x}_i, y_i)$ from the linear model $y_i = \mathbf{x}_i^\top\beta^* + \varepsilon_i$, where $\mathbf{x}_i \in \mathbb{R}^d$ are sub-Gaussian covariates with covariance Σ and ε_i are independent sub-Gaussian noise terms. Further, $\beta^\star \in \mathcal{B} \subseteq \mathbb{R}^d$. We assume throughout that

$$
\sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_2 \le 1, \qquad \sup_{\beta\in\mathcal{B}}\|\beta\|_2 \le 1.
$$

We solve a least-squares problem to compute $\hat{\beta}^s$ using the sketched inputs and corresponding responses, i.e., using $(\mathbf{S}\mathbf{x}_i, y_i)$. Subsequently, we de-sketch the estimate using $\mathbf{S}^\top\hat{\beta}^s$ and use this to make predictions. Note that solving the least-squares problem in a lower-dimensional sketched space is computationally faster. Our goal is to bound the error $(\mathbf{S}^\top\hat{\beta}^s - \beta^\star)^\top\mathbf{x}$ for $\mathbf{x}\in\mathcal{X}$.

We assume that the sketching matrix $\mathbf{S}$ has all its entries drawn i.i.d. from $N(0, 1/b)$. Now consider $\mathbf{x}\in\mathcal{X}$. We have

$$
\begin{aligned}
(\mathbf{S}^\top\hat{\beta}^s - \beta^\star)^\top\mathbf{x} &= \langle\mathbf{S}^\top\hat{\beta}^s, \mathbf{x}\rangle - \langle\mathbf{S}\beta^\star, \mathbf{S}\mathbf{x}\rangle + \langle\mathbf{S}\beta^\star, \mathbf{S}\mathbf{x}\rangle - \langle\beta^\star, \mathbf{x}\rangle \\
&= \langle\hat{\beta}^s - \mathbf{S}\beta^\star, \mathbf{S}\mathbf{x}\rangle + \beta^{\star\top}(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}.
\end{aligned}
$$

Let $X \in \mathbb{R}^{n\times d}$ be the design matrix whose i-th row is $\mathbf{x}_i^\top$, and let $X^s \in \mathbb{R}^{n\times b}$ be the sketched design matrix whose i-th row is $(\mathbf{S}\mathbf{x}_i)^\top$. Thus $X^s = X\mathbf{S}^\top$. The sketched ordinary least-squares estimator is

$$
\hat{\beta}^s = (X^{s\top}X^s)^{-1}X^{s\top}y,
$$

where $y = (y_1, \ldots, y_n)^\top$. Further, let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.

Since

$$y = X\beta^\star + \varepsilon = X\mathbf{S}^\top\mathbf{S}\beta^\star + X(I - \mathbf{S}^\top\mathbf{S})\beta^\star + \varepsilon = X^s(\mathbf{S}\beta^\star) + X(I - \mathbf{S}^\top\mathbf{S})\beta^\star + \varepsilon,$$

we obtain

$$\hat{\beta}^s - \mathbf{S}\beta^\star = (X^{s\top}X^s)^{-1}X^{s\top}\big(X(I - \mathbf{S}^\top\mathbf{S})\beta^\star + \varepsilon\big).$$

Therefore,

$$\|\hat{\beta}^s - \mathbf{S}\beta^\star\|_2 \le \underbrace{\|(X^{s\top}X^s)^{-1}X^{s\top}\varepsilon\|_2}_{A_{\text{noise}}} + \underbrace{\|(X^{s\top}X^s)^{-1}X^{s\top}X(I - \mathbf{S}^\top\mathbf{S})\beta^\star\|_2}_{A_{\text{bias}}}.$$

Following a standard analysis, we have with probability $1 - \delta$,

$$A_{\text{noise}} \lesssim \frac{\sigma}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\Big(\sqrt{b} + \sqrt{2\log(1/\delta)}\Big),$$

where σ is the sub-Gaussian parameter of the noise process ε_i.

For the second term, using

$$\|(X^{s\top}X^s)^{-1}X^{s\top}\|_{\text{op}} = \frac{1}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}},$$

we get

$$\begin{aligned}
A_{\text{bias}} &\le \frac{1}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\|X(I - \mathbf{S}^\top\mathbf{S})\beta^\star\|_2 \\
&= \frac{1}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\left(\sum_{i=1}^n \big(\mathbf{x}_i^\top(I - \mathbf{S}^\top\mathbf{S})\beta^\star\big)^2\right)^{1/2} \\
&\le \frac{\sqrt{n}}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\sup_{\mathbf{x}\in\mathcal{X}}\big|\beta^{\star\top}(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}\big|.
\end{aligned}$$

Hence, with probability $1 - \delta$,

$$\|\hat{\beta}^s - \mathbf{S}\beta^\star\|_2 \lesssim \frac{\sigma}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\Big(\sqrt{b} + \sqrt{2\log(1/\delta)}\Big) + \frac{\sqrt{n}}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\sup_{\mathbf{x}\in\mathcal{X}}\big|\beta^{\star\top}(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}\big|.$$

Therefore,

$$\sup_{\mathbf{x}\in\mathcal{X}}\big|(\mathbf{S}^\top\hat{\beta}^s - \beta^\star)^\top\mathbf{x}\big| \le \|\hat{\beta}^s - \mathbf{S}\beta^\star\|_2 \underbrace{\sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{S}\mathbf{x}\|_2}_{I} + \underbrace{\sup_{\beta\in\mathcal{B},\,\mathbf{x}\in\mathcal{X}}\big|\beta^\top(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}\big|}_{II}.$$

Term I can be handled using the one-set result from [Krahmer et al., 2014b]. Using a construction as in Section 4.1 and thereafter invoking Theorem 3.1 from [Krahmer et al., 2014b], we have

$$\begin{aligned}
&P\left\{\sup_{\mathbf{x}\in\mathcal{X}}\big|\|\mathbf{S}\mathbf{x}\|_2^2 - \mathbb{E}\|\mathbf{S}\mathbf{x}\|_2^2\big| \gtrsim \left[\frac{1}{b}\omega^2(\mathcal{X}) + \frac{1}{\sqrt{b}}\omega(\mathcal{X})\right] + \epsilon\right\} \\
&\lesssim \exp\left(-\min\left\{\frac{\epsilon^2}{\left(\frac{1}{b}\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\right)^2}, \frac{\epsilon}{\frac{1}{b}}\right\}\right).
\end{aligned}$$

Since $\mathbb{E}\|\mathbf{S}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 \le 1$, this implies that on the above event,

$$I = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{S}\mathbf{x}\|_2 \le \left(1 + \sup_{\mathbf{x}\in\mathcal{X}} \left|\|\mathbf{S}\mathbf{x}\|_2^2 - \mathbb{E}\|\mathbf{S}\mathbf{x}\|_2^2\right|\right)^{1/2}.$$

Next, using Theorem 3.1, we have

$$P\left\{\sup_{\beta\in\mathcal{B},\,\mathbf{x}\in\mathcal{X}} \left|\beta^\top(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}\right| \ge \frac{1}{b}\omega(\mathcal{B})\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\big(\omega(\mathcal{B}) + \omega(\mathcal{X})\big) + \epsilon\right\}$$
$$\lesssim \exp\left(-\min\left\{\frac{\epsilon^2}{\left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)^2}, \frac{\epsilon}{\frac{1}{b}}\right\}\right).$$

Choose

$$\epsilon^2 = \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)^2 \bar{\epsilon}^2,$$

and take a union bound over both events to get, with probability at least $1 - \exp(-\bar{\epsilon}^2)$,

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\|\mathbf{S}\mathbf{x}\|_2^2 - \mathbb{E}\|\mathbf{S}\mathbf{x}\|_2^2\right| \lesssim \left[\frac{1}{b}\omega^2(\mathcal{X}) + \frac{1}{\sqrt{b}}\omega(\mathcal{X})\right] + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\bar{\epsilon},$$
$$\sup_{\beta\in\mathcal{B},\,\mathbf{x}\in\mathcal{X}} \left|\beta^\top(\mathbf{S}^\top\mathbf{S} - I)\mathbf{x}\right| \lesssim \frac{1}{b}\omega(\mathcal{B})\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\big(\omega(\mathcal{B}) + \omega(\mathcal{X})\big) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\bar{\epsilon}.$$

Define

$$\Delta_{\mathcal{X}} := \frac{1}{b}\omega^2(\mathcal{X}) + \frac{1}{\sqrt{b}}\omega(\mathcal{X}) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\bar{\epsilon},$$
$$\Gamma_{\mathcal{B},\mathcal{X}} := \frac{1}{b}\omega(\mathcal{B})\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\big(\omega(\mathcal{B}) + \omega(\mathcal{X})\big) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\bar{\epsilon}.$$

Then on the same event,

$$I \le (1 + \Delta_{\mathcal{X}})^{1/2}, \qquad II \lesssim \Gamma_{\mathcal{B},\mathcal{X}}.$$

Taking a union bound over all the high-probability events, we obtain that with probability at least $1 - \delta$, for all $\mathbf{x} \in \mathcal{X}$,

$$\left|(\mathbf{S}^\top\hat{\beta}^s - \beta^\star)^\top\mathbf{x}\right| \lesssim \left[\frac{\sigma}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\left(\sqrt{b} + \sqrt{2\log(1/\delta)}\right) + \frac{\sqrt{n}}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\Gamma_{\mathcal{B},\mathcal{X}}\right](1 + \Delta_{\mathcal{X}})^{1/2} + \Gamma_{\mathcal{B},\mathcal{X}}.$$

Equivalently, after substituting the definitions of $\Delta_{\mathcal{X}}$ and $\Gamma_{\mathcal{B},\mathcal{X}}$ and taking $\bar{\epsilon} \asymp \sqrt{\log(\delta^{-1})}$, we get

$$\left|(\mathbf{S}^\top\hat{\beta}^s - \beta^\star)^\top\mathbf{x}\right| \lesssim \left[\frac{\sigma}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\left(\sqrt{b} + \sqrt{2\log(1/\delta)}\right)\right.$$
$$\left. + \frac{\sqrt{n}}{\sqrt{\lambda_{\min}(X^{s\top}X^s)}}\left(\frac{1}{b}\omega(\mathcal{B})\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\big(\omega(\mathcal{B}) + \omega(\mathcal{X})\big) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\sqrt{\log(\delta^{-1})}\right)\right]$$
$$\cdot\left[1 + \frac{1}{b}\omega^2(\mathcal{X}) + \frac{1}{\sqrt{b}}\omega(\mathcal{X}) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\sqrt{\log(\delta^{-1})}\right]^{1/2}$$
$$+ \frac{1}{b}\omega(\mathcal{B})\omega(\mathcal{X}) + \frac{1}{\sqrt{b}}\big(\omega(\mathcal{B}) + \omega(\mathcal{X})\big) + \left(\frac{1}{b}(\omega(\mathcal{B}) + \omega(\mathcal{X})) + \frac{1}{\sqrt{b}}\right)\sqrt{\log(\delta^{-1})}.$$

The above bound depends on the Gaussian widths $\omega(\mathcal{X})$ and $\omega(\mathcal{B})$, instead of depending on the ambient dimension d. In cases where $\omega(\mathcal{X})$ and $\omega(\mathcal{B})$ are small, one can choose the sketching dimension b to balance the terms and obtain a tighter bound. The additional term proportional to $\sqrt{n}\,\Gamma_{\mathcal{B},\mathcal{X}}/\sqrt{\lambda_{\min}(X^{s\top}X^s)}$ is the sketch-induced bias term arising from the fact that the sketched regression model is misspecified through the residual $X(I-\mathbf{S}^\top\mathbf{S})\beta^\star$.

H EXTENSION TO CONDITIONALLY INDEPENDENT PROCESS

Let $\xi=\{\xi_1,\ldots,\xi_n\}$ be a stochastic process satisfying Assumptio 4. Then the following holds.

Theorem 5.1. *Let $\mathcal{M}$ and $\mathcal{N}$ be set of $(m\times n)$ and $(n\times m)$ matrices respectively, and let $\boldsymbol{\xi}$ be a random vector satisfying Assumption 4. Let W, V, U be as defined in Theorem 3.1. Then, for $\epsilon>0$, $\mathbb{P}\big(C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\geq c_1W+\epsilon\big)\leq 2\exp\left(-c_2\min\left\{\frac{\epsilon^2}{V^2},\frac{\epsilon}{U}\right\}\right)$, where the constants c_1,c_2 depend on the sub-gaussian parameter L.*

Proof. The proof structurally follows the same approach as in the proof of Theorem 3.1, and specify the parts that need to be modified.

1. **Decomposing off-diagonal and diagonal terms.** In the proof of Lemma 3.2 in our paper, the first line uses independence of ξ. To extend this, using the analysis in Proposition 7 of [Banerjee et al., 2019] and since $\{\xi_i\}$ is a stochastic process adapted to $\{F_i\}$ satisfying Assumption 4 (i) and (ii), with $F=F_{1:n}$ we have
$$\mathbb{E}_{\xi_j,\xi_k}[\xi_j\xi_k]=\mathbb{E}_F\big[\mathbb{E}_{\xi_j,\xi_k|F}[\xi_j\xi_k]\big]=\mathbb{E}_F\big[\mathbb{E}_{\xi_j|F}[\xi_j]\,\mathbb{E}_{\xi_k|F}[\xi_k]\big]=0,$$
since $\xi_j\perp\xi_k\mid F$ by (SP-2) and $\mathbb{E}_{\xi_j|F}[\xi_j]=0=\mathbb{E}_{\xi_k|F}[\xi_k]$ by Assumption 4 (i). Therefore
$$\begin{aligned}C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})&=\sup_{M\in\mathcal{M},\,N\in\mathcal{N}}\big|\big(\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}\big)-\mathbb{E}\big[\boldsymbol{\xi}^\top M^\top N\boldsymbol{\xi}\big]\big|\\&=\sup_{M\in\mathcal{M},\,N\in\mathcal{N}}\left|\sum_{\substack{j,k=1\\ j\neq k}}^{n}\xi_j\xi_k\,\langle M_j,N_k\rangle\;+\;\sum_{j=1}^{n}\big(|\xi_j|^2-\mathbb{E}|\xi_j|^2\big)\,\langle M_j,N_j\rangle\right|,\end{aligned}$$
and the rest of the analysis in Lemma 3.2 of our paper follows.

2. **Off-diagonal term.** We next describe how our results for bounding the off-diagonal change.
   - The decoupling result (Lemma D.1) is the classical decoupling with independent random variables. Replace this with the "conditioned on F" version as in Theorem 6 of [Banerjee et al., 2019]. The proof technique is the same; the only difference is that every expectation is first conditioned on F, and then we take the expectation with respect to F. Conditioning on F lets us use conditional independence from Assumption 4 (i) and (ii).
   - In the proof of Lemma D.2, we apply generic chaining. When bounding each term, the argument should be conditioned on F, following the approach in the last two lines of p. 27 in [Banerjee et al., 2019], which uses Assumption 4 (i) and (ii). All subsequent inequalities should likewise be conditioned on F. Corresponding to $\|S_1\|_{L_p}^p$, where a direct expectation is taken, we instead integrate with respect to F, as in the final equation on p. 28 of [Banerjee et al., 2019]. Later, while bounding $\|\sup_{M\in\mathcal{M}}\|M\boldsymbol{\xi}\|_2\|_{L_p}$, where the independent version is applied, replace it by [Banerjee et al., 2019] Lemma 7 for Assumption 4 (i) and (ii).
   - In the proof of Lemma D.3, we use Theorem 2.3 from Krahmer et al. [2014a]. In contrast, the corresponding step in [Banerjee et al., 2019] establishes the Assumption 4 (i) and (ii) version (see Lemma 7 in [Banerjee et al., 2019]). Moreover, in the analysis of $\|\langle M\boldsymbol{\xi},N\boldsymbol{\xi}'\rangle\|$, the argument should be conditioned on F, as shown at the top of p. 32 in [Banerjee et al., 2019].

3. **Diagonal term.** We finally describe how our results for bounding the diagonal change.
   - In the symmetrization step, to bound $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}))\|_{L_p}$, Lemma 9 of [Banerjee et al., 2019] is applied. This lemma remains valid under Assumption 4 (i) and (ii).

- Next in obtaining the contraction condition, we must first condition on F and then take the expectation with respect to F, as done at the end of p. 37 in [Banerjee et al., 2019], before applying the contraction inequality. The de-symmetrization step in [Banerjee et al., 2019] is already incorporated in equation (26) of our paper, and it also holds under Assumption 4 (i) and (ii).

□

H.1 Subsampled Randomized Hadamard Transform (SRHT)

We say $R \in \mathbb{R}^{b\times n}$ is a subsampled randomized Hadamard transform matrix if it has the form

$$R \;=\; \sqrt{\frac{n}{b}}\, S\, H\, D,$$

where $S \in \mathbb{R}^{b\times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, $H \in \mathbb{R}^{n\times n}$ is a normalized Walsh–Hadamard matrix, and $D \in \mathbb{R}^{n\times n}$ is diagonal with i.i.d. Rademacher entries. We show that SRHT matrix satisfies Assumption 4.

First note that H is deterministic and the random signs of the entries of H are determined by the diagonal matrix D. Since the diagonal entries of D are i.i.d., the columns of HD are independent. Next, left multiplication by S projects each of the columns of HD along b different random standard-basis directions. The randomness of S is independent of HD; therefore, with $F_{i,j} = R_{1:(i-1),\cdot}$ (all previous rows) for $i, j \in [n]$, $R_{i,j}$ follows a uniform distribution over the remaining entries in column j of HD, excluding the entries that have already appeared in $R_{1,j}, R_{2,j}, \ldots, R_{i-1,j}$. This implies the process satisfies Assumption 4 (ii); see also Figure 3 (GM3) of [Banerjee et al., 2019]. Assumption Assumption 4 (i) is immediately satisfied since the entries of D are i.i.d. Rademacher and S selects rows uniformly at random without replacement.

I EXTENSION TO SUM OF RANDOM QUADRATIC FORMS

I.1 Expected Analysis

In this section we provide an expected deviation bound on the random quadratic form. For simplicity we consider the single set version.

Theorem I.1. *Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$ be a set of matrices and let ϵ be a Rademacher vector of length n. Then*

$$\mathbb{E}\sup_{A\in\mathcal{A}} \|A\boldsymbol{\xi}_t\|_2^2 - \mathbb{E}\,\|A\boldsymbol{\xi}_t\|_2^2 \le C_1\left(\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})^2 + d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{A})d_{2\to 2}(\mathcal{A})\right).$$

Proof. By the decoupling inequality of Theorem 2.4 [Krahmer et al., 2014b], we have

$$\begin{aligned}
\mathbb{E}\, C_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) &= \mathbb{E}\sup_{A\in\mathcal{A}}\sum_{t=1}^{T}\|A\boldsymbol{\xi}_t\|_2^2 - \mathbb{E}\,\|A\boldsymbol{\xi}_t\|_2^2 = \mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{t=1}^{T}\sum_{j\neq k}\boldsymbol{\xi}_{t,j}\boldsymbol{\xi}_{t,k}(A^\top A)_{j,k}\right| \\
&\le 4\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{t=1}^{T}\sum_{j,k}\boldsymbol{\xi}'_{t,j}\boldsymbol{\xi}_{t,k}(A^\top A)_{j,k}\right| \\
&\lesssim \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})\mathbb{E}N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) + \sup_{A\in\mathcal{A}}\mathbb{E}|\sum_{t=1}^{T}\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle|. \qquad (52)
\end{aligned}$$

where $N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) := \sup_{A\in\mathcal{A}}\|\sum_{t=1}^{T}A\boldsymbol{\xi}_t\|_2$, and the inequality follows by Lemma I.3 by setting $M = N = A$. Since

$$\mathbb{E}|\sum_{t=1}^{T}\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle| \overset{(a)}{\le} \mathbb{E}|\sum_{t=1}^{T}\varepsilon_t\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle| \overset{(b)}{\le} \mathbb{E}_{\boldsymbol{\xi}_t,\boldsymbol{\xi}'_t}\sqrt{\sum_{t=1}^{T}|\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle|^2}$$

$$\overset{(c)}{\leq} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{\xi}_t, \boldsymbol{\xi}'_t} |\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle|^2} = \sqrt{T}\|A^\top A\|_F \leq \sqrt{T}\|A\|_{2\to 2}\|A\|_F,$$

we have

$$\sup_{A\in\mathcal{A}} \mathbb{E}|\sum_{t=1}^{T}\langle A\boldsymbol{\xi}_t, A\boldsymbol{\xi}'_t\rangle| \leq \sqrt{T} d_{2\to 2}(\mathcal{A}) d_F(\mathcal{A}) \leq \sqrt{T} d_F(\mathcal{A})^2. \tag{53}$$

Here (a) follows by symmetrization [Ledaux and Talagrand, 1991], (b) follows by Kinchinte [Vershynin, 2018] and (c) follows by Jensen's inequality. We conclude that

$$\begin{aligned}(\mathbb{E}N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}))^2 \leq \mathbb{E}N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T})^2 &\leq \mathbb{E}\,\mathcal{C}_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) + \sqrt{T} d_F(\mathcal{A}) \\ &\lesssim \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})\mathbb{E}\, N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) + \sqrt{T} d_F(\mathcal{A})^2,\end{aligned}$$

so that

$$\mathbb{E}N_{\mathcal{A}}(\boldsymbol{\xi}^{1:T}) \lesssim \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + \sqrt{T} d_F(\mathcal{A}).$$

Plugging this into (52) yields the claim.

I.2 High Probability Analysis

Our objective is to develop large deviation bound for the following random variable.

$$C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) \triangleq \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{t=1}^{T}(\boldsymbol{\xi}_t^\top M^\top N\boldsymbol{\xi}_t) - \mathbb{E}(\boldsymbol{\xi}_t^\top M^\top N\boldsymbol{\xi}_t)\right| \tag{54}$$

To develop large deviation bounds on $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})$, as in the proof of Theorem 3.1 we decompose the quadratic form into terms depending on the off-diagonal and the diagonal elements of $M^\top N$ respectively as follows.

$$\begin{aligned}B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) &\triangleq \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{t=1}^{T}\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_{t,j}\boldsymbol{\xi}_{t,k}\langle M_j, N_k\rangle\right|, \\ D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) &\triangleq \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{t=1}^{T}\sum_{j=1}^{n}(|\boldsymbol{\xi}_{t,j}|^2 - \mathbb{E}\,|\boldsymbol{\xi}_{t,j}|^2)\langle M_j, N_j\rangle\right|,\end{aligned}$$

where M_i and N_i are the i-th row of M and N respectively.

Our large deviation bound for $C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})$ is based on bounding $\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$ via $\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$ and $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})\|_{L_p}$, yielding

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} \leq W\cdot\sqrt{T} + \sqrt{p}\cdot V\cdot\sqrt{T} + p\cdot U\cdot\sqrt{T}, \quad \forall p\geq 1\ . \tag{55}$$

Using standard moment bounds and Markov's inequality [Williams, 1991, Vershynin, 2012], this gives for all $\epsilon > 0$:

$$\mathbb{P}\left(|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})| \geq W\cdot\sqrt{T} + \epsilon\right) \leq \exp\left\{-\min\left(\frac{\epsilon^2}{4\,V\cdot T^2}, \frac{\epsilon}{2\,U\cdot\sqrt{T}}\right)\right\}. \tag{56}$$

Following the proof of Lemma 3.2 we can show that:

$$\|C_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} \leq \|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} + \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}\ . \tag{57}$$

Next we bound $\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$ and $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$ using Theorem I.2 and Theorem 3.5 respectively. Note that collecting terms corresponding to W, V and U and combining them with the observation in (56) completes the proof of Theorem 5.2. Next we bound $\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$ and $\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p}$.

We first control the off-diagonal term in the following theorem.

Theorem I.2. *Let $\boldsymbol{\xi}_t$ for all $t \in [T]$ be stochastic processes satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} \lesssim \sqrt{T}\Big[&\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big(\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p}\; d_{2\to 2}(\mathcal{N})\Big) \\ &+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p}\; d_{2\to 2}(\mathcal{M})\Big) \\ &+ \sqrt{p}\; \min\Big\{d_F(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N}) \cdot d_{2\to 2}(\mathcal{M})\Big\} + p \cdot d_{2\to 2}(\mathcal{M}) \cdot d_{2\to 2}(\mathcal{N})\Big].\end{aligned}$$

I.3 Proof of Theorem I.2

We use Lemma D.1 with $F(x) = |x|^p, p \geq 1$ as the convex function and set $B_{j,k} = \langle M_j, N_k\rangle$. Then

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} &= \mathbb{E} \sup_{M\in\mathcal{M},N\in\mathcal{N}} F\left(\sum_{t=1}^{T}\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_{t,j}\boldsymbol{\xi}_{t,k}\langle M_j, N_k\rangle\right) \\ &\leq E \sup_{M\in\mathcal{M},N\in\mathcal{N}} F\left(4\sum_{t=1}^{T}\sum_{\substack{j,k=1\\ j\neq k}}^{n} \boldsymbol{\xi}_{t,j}\boldsymbol{\xi}_{t,k}\langle M_j, N_k\rangle\right) \\ &\lesssim \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left|\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\right|\right\|_{L_p}. \end{aligned} \tag{58}$$

Note that for fixed M, N, the term $\left|\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\right|$ conditioned on $\boldsymbol{\xi}_t'$ is sub-gaussian and therefore its L_p norm can be bounded. However, the $\sup_{M\in\mathcal{M},N\in\mathcal{N}}$ inside the $\|\cdot\|_{L_p}$ does not let us use this approach. The next theorem therefore upper bounds $\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\right|\right\|_{L_p}$ by $\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left\|\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\right\|_{L_p}$ plus some additional complexity terms. Unlike Krahmer et al. [2014b] the inner product contains two different matrices M and N, and therefore we consider two separate admissible sequences (cf. definition 2.2) $\{T_r(\mathcal{M})\}_{r=0}^{\infty}$ and $\{T_r(\mathcal{N})\}_{r=0}^{\infty}$ of $\mathcal{M}$ and $\mathcal{N}$ respectively. We then use a generic chaining argument by creating two separate increment sequences for $\mathcal{M}$ and $\mathcal{N}$ and is detailed in the following theorem.

Lemma I.3. *Let $\boldsymbol{\xi}_t$ for all $t \in [T]$ be stochastic processes satisfying Assumption 1, and $\boldsymbol{\xi}_t'$ be a decoupled tangent sequence to $\boldsymbol{\xi}_t$. Then, for every $p \geq 1$,*

$$\begin{aligned}\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\right\|_{L_p} \lesssim\; &\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p}d_{2\to 2}(\mathcal{N})\Big] \\ &+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p}d_{2\to 2}(\mathcal{M})\Big] \\ &+ \sup_{M\in\mathcal{M},N\in\mathcal{N}} \|\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}_t'\rangle\|_{L_p}, \end{aligned} \tag{59}$$

Proof of Lemma I.3. Without loss of generality, assume $\mathcal{M}$ and $\mathcal{N}$ are finite [Talagrand, 2014]. Let $\{T_r(\mathcal{M})\}_{r=0}^{\infty}$ and $\{T_r(\mathcal{N})\}_{r=0}^{\infty}$ be admissible sequences for $\mathcal{M}$ and $\mathcal{N}$ for which the minimum in the definition of $\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2})$ and $\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2})$ are attained respectively. Let

$$\begin{aligned}\pi_r M = d_{2\to 2}(M, T_r(\mathcal{M})) &= \operatorname*{argmin}_{B\in T_r(\mathcal{M})} \|B - A\|_{2\to 2} \quad \text{and} \quad \Delta_r M = \pi_r M - \pi_{r-1} M. \\ \pi_r N = d_{2\to 2}(N, T_r(\mathcal{N})) &= \operatorname*{argmin}_{B\in T_r(\mathcal{N})} \|B - A\|_{2\to 2} \quad \text{and} \quad \Delta_r N = \pi_r N - \pi_{r-1} N.\end{aligned}$$

For any given $p \geq 1$, let ℓ be the largest integer for which $2^\ell \leq 2p$. Then,

$$\begin{aligned}
&\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle - \langle(\pi_\ell M)\boldsymbol{\xi}_t, (\pi_\ell N)\boldsymbol{\xi}'_t\rangle = \sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\pi_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle - \langle(\pi_r M)\boldsymbol{\xi}_t, (\pi_r N)\boldsymbol{\xi}'_t\rangle \\
&= \sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\pi_r M + \Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle - \langle(\pi_r M)\boldsymbol{\xi}_t, (\pi_{r+1}N - \Delta_{r+1}N)\boldsymbol{\xi}'_t\rangle \\
&= \sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\pi_r M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle + \langle(\Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle \\
&\qquad\qquad - \langle(\pi_r M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle + \langle(\pi_r M)\boldsymbol{\xi}_t, (\Delta_{r+1}N)\boldsymbol{\xi}'_t\rangle \\
&= \sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle + \langle(\pi_r M)\boldsymbol{\xi}_t, (\Delta_{r+1}N)\boldsymbol{\xi}'_t\rangle
\end{aligned}$$

Now by applying triangle inequality, we have

$$\begin{aligned}
\left|\sum_{t=1}^{T}\langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle - \langle(\pi_\ell M)\boldsymbol{\xi}_t, (\pi_\ell N)\boldsymbol{\xi}'_t\rangle\right| \leq \underbrace{\left|\sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle\right|}_{S_1} \\
+ \underbrace{\left|\sum_{t=1}^{T}\sum_{r=\ell}^{\infty}\langle(\pi_r M)\boldsymbol{\xi}_t, (\Delta_{r+1}N)\boldsymbol{\xi}'_t\rangle\right|}_{S_2} . \qquad (60)
\end{aligned}$$

We first consider S_1. Let us define

$$X_r(M,N) = \sum_{t=1}^{T}\langle(\Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle \ .$$

Conditioning $X_r(M,N)$ on $\boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T$, we note

$$\begin{aligned}
X_r(M,N) \ \Big|\ \boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T &= \sum_{t=1}^{T}\langle(\Delta_{r+1}M)\boldsymbol{\xi}_t, (\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle \ \Big|\ \boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T \\
&= \sum_{t=1}^{T}\langle\boldsymbol{\xi}_t, (\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'_t\rangle \ \Big|\ \boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T
\end{aligned}$$

is a sub-Gaussian random variable and therefore using Azuma-Hoeffding bound [Boucheron et al., 2013, Vershynin, 2018] gives

$$P\left(|X_r(M,N)| > u\left\|\sum_{t=1}^{T}(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'_t\right\|_2 \ \Big|\ \boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T\right) \leq 2\exp(-u^2/2) \ .$$

Using $u = t2^{r/2}$, we get

$$P\left(|X_r(M,N)| > t2^{r/2}\left\|\sum_{t=1}^{T}(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'_t\right\|_2 \ \Big|\ \boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_T\right) \leq 2\exp(-t^2 2^r/2) \ .$$

Since

$$\left|\sum_{t=1}^{T}(\Delta_{r+1}M)^T(\pi_{r+1}N)\boldsymbol{\xi}'_t\right| \leq \|\Delta_{r+1}M\|_{2\to 2}\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}'_t\right\|_2 .$$

we have

$$P\left(|X_r(M,N)| > t2^{r/2}\|\Delta_{r+1}M\|_{2\to 2}\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2 \;\middle|\; \boldsymbol{\xi}_1',\ldots,\boldsymbol{\xi}_T'\right) \leq 2\exp(-t^2 2^r/2)\ .$$

Now, since $|\{\pi_r M : M\in\mathcal{M}\}| = |T_r(\mathcal{M})| \leq 2^{2^r}$ and $|\{\pi_r N : N\in\mathcal{M}\}| = |T_r(\mathcal{N})| \leq 2^{2^r}$, by union bound, we get

$$\begin{aligned}
P\Bigg(\sup_{M\in\mathcal{M},N\in\mathcal{N}}\sum_{r=\ell}^{\infty}|X_r(M,N)| > t\left(\sup_{M\in\mathcal{M}}\sum_{r=\ell}^{\infty}2^{r/2}\|\Delta_{r+1}M\|_{2\to 2}\right)\cdot\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2 \;\middle|\; \boldsymbol{\xi}_1',\ldots,\boldsymbol{\xi}_T'\Bigg) \\
&\leq 2\sum_{r=\ell}^{\infty}|T_r(\mathcal{M})|\cdot|T_{r+1}(\mathcal{M})|\cdot|T_{r+1}(\mathcal{N})|\cdot\exp(-t^2 2^r/2) \\
&\leq 2\sum_{r=\ell}^{\infty}2^{2^{r+2}}\cdot\exp(-t^2 2^r/2) \\
&\leq 2\exp(-2^\ell t^2)\ ,
\end{aligned}$$

for all $t\geq t_0$, a constant. Next, note that

$$\sup_{M\in\mathcal{M}}\sum_{r=\ell}^{\infty}2^{r/2}\|\Delta_{r+1}M\|_{2\to 2} = \gamma_2(\mathcal{M},\|\cdot\|_{2\to 2}).$$

Therefore we have

$$P\Bigg(\sup_{M\in\mathcal{M},N\in\mathcal{N}}\sum_{r=\ell}^{\infty}|X_r(M,N)| > t\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2})\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2 \;\middle|\; \boldsymbol{\xi}_1',\ldots,\boldsymbol{\xi}_T'\Bigg) \leq 2\exp(-pt^2)\ ,$$

since $p\leq 2^\ell$ by construction which implies with $V(\boldsymbol{\xi}') = \gamma_2(\mathcal{M},\|\cdot\|_{2\to 2})\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2$, for $t\geq t_0$ we have

$$P\left(S_1\geq tV(\boldsymbol{\xi}') \;\middle|\; \boldsymbol{\xi}'\right) \leq 2\exp(-pt^2)\ .$$

Note that

$$\|S_1\|_{L_p}^p = \mathop{\mathbb{E}}_{\boldsymbol{\xi},\boldsymbol{\xi}'}S_1^p = E_{\boldsymbol{\xi}'}\int_0^\infty pt^{p-1}P(S_1>t \mid \boldsymbol{\xi}')dt\ ,$$

and

$$\begin{aligned}
\int_0^\infty pt^{p-1}P(S_1>t\mid\boldsymbol{\xi}')dt &\leq c^pV(\boldsymbol{\xi}')^p + \int_{cV(\boldsymbol{\xi}')}^\infty pt^{p-1}P(S_1>t\mid\boldsymbol{\xi}')dt \\
&\leq c^pV(\boldsymbol{\xi}')^p + V(\boldsymbol{\xi}')^p\int_c^\infty p\tau^{p-1}P(S_1>\tau V(\boldsymbol{\xi}')|\boldsymbol{\xi}')d\tau \\
&\leq c_1^pV(\boldsymbol{\xi}')^p\ ,
\end{aligned}$$

where $c\geq t_0, c_1$ are suitable constants that depend on L. As a result, $\|S_1\|_{L_p}\leq c_1V(\boldsymbol{\xi}') = c_1\|V(\boldsymbol{\xi})\|_{L_p}$, i.e., we have the following bound on S_1.

$$\|S_1\|_{L_p}\lesssim\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2})\left\|\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2\right\|_{L_p}$$

Note that a similar analysis follows for S_2, and we can bound $\|S_2\|_{L_p}$. As a result

$$\|S_1+S_2\|_{L_p}\lesssim\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2})\cdot\left\|\sup_{N\in\mathcal{N}}\left\|\sum_{t=1}^{T}N\boldsymbol{\xi}_t'\right\|_2\right\|_{L_p} + \gamma_2(\mathcal{N},\|\cdot\|_{2\to 2})\cdot\left\|\sup_{M\in\mathcal{M}}\left\|\sum_{t=1}^{T}M\boldsymbol{\xi}_t'\right\|_2\right\|_{L_p} \tag{61}$$

Further, since $|\{\pi_\ell M : M \in \mathcal{M}\}| \le 2^{2^\ell} \le \exp(2p)$, and $|\{\pi_\ell N : N \in \mathcal{N}\}| \le 2^{2^\ell} \le \exp(2p)$ we have

$$\mathbb{E} \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^T \langle (\pi_\ell M)\boldsymbol{\xi}_t, (\pi_\ell N)\boldsymbol{\xi}'_t\rangle \Big|^p \le \sum_{M\in T_\ell(M),N\in T_\ell(N)} \mathbb{E} \Big| \sum_{t=1}^T \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big|^p$$
$$\le 2^{2p}2^{2p} \sup_{M\in\mathcal{M},N\in\mathcal{N}} E \Big| \sum_{t=1}^T \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big|^p = 2^{4p} \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^T \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big|^p ,$$

so that

$$\Big\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^T \langle (\pi_\ell M)\boldsymbol{\xi}_t, (\pi_\ell N)\boldsymbol{\xi}'_t\rangle \Big| \Big\|_{L_p} \le 16 \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big\| \sum_{t=1}^T \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big\|_{L_p} . \tag{62}$$

Combining (60), (61), and (62) and using triangle inequality we have

$$\Big\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \sum_{t=1}^T \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p} \lesssim \gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) \cdot \Big\| \sup_{N\in\mathcal{N}} \Big\| \sum_{t=1}^T N\boldsymbol{\xi}'_t \Big\|_2 \Big\|_{L_p}$$
$$+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) \cdot \Big\| \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\boldsymbol{\xi}'_t \Big\|_2 \Big\|_{L_p}$$
$$+ \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big\| \sum_{t=1}^T \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big\|_{L_p} . \tag{63}$$

Note that unlike in the proof of Theorem 3.1, we cannot use Theorem 3.5 from [Krahmer et al., 2014a] to control $\Big\| \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\boldsymbol{\xi}'_t \Big\|_2 \Big\|_{L_p}$. Below we derive new results to handle them.

Theorem I.4. *Let $\boldsymbol{\xi}_t$ for all $t \in [T]$ be stochastic processes satisfying Assumption 1. Then*

$$\Big\| \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\boldsymbol{\xi}_t \Big\|_2 \Big\|_{L_p} \le \gamma_2(\mathcal{M}, \|\cdot\|) + \sqrt{T} d_F(\mathcal{M}) + \sqrt{pT} d_{2\to2}(\mathcal{M})$$

Proof. We apply Theorem 2.3 from Krahmer et al. [2014b] with the set $S = \{M^\top x : x \in B_2^n, M \in \mathcal{M}\}$. Since $\boldsymbol{\xi}_t$ is L-subgaussian for all $t \in [T]$ we obtain

$$\Big\| \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\boldsymbol{\xi}_t \Big\|_2 \Big\|_{L_p} = \left(\mathbb{E} \sup_{M\in\mathcal{M},\, x\in B_2^n} |\sum_{t=1}^T \langle M\boldsymbol{\xi}_t, x\rangle|^p \right)^{1/p} = \left(\mathbb{E} \sup_{u\in S} |\sum_{t=1}^T \langle \boldsymbol{\xi}_t, u\rangle|^p \right)^{1/p}$$
$$\lesssim_L \mathbb{E} \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\mathbf{g}_t \Big\|_2 + \sup_{u\in S} \left(\mathbb{E} |\sum_{t=1}^T \langle \boldsymbol{\xi}_t, u\rangle|^p \right)^{1/p}$$
$$\lesssim_L \mathbb{E} \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\mathbf{g}_t \Big\|_2 + \sqrt{pT} \sup_{M\in\mathcal{M},\, x\in B_2^n} \|M^\top x\|_2$$
$$\lesssim_L \mathbb{E} \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\mathbf{g}_t \Big\|_2 + \sqrt{pT} d_{2\to2}(\mathcal{M})$$

Next we handle $\mathbb{E} \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^T M\mathbf{g}_t \Big\|_2$. Using Theorem 2.5 in Krahmer et al. [2014b] we have

$$\|C_{\mathcal{M}}(\mathbf{g}^{1:T})\|_{L_p} = \Big\| \sup_{M\in\mathcal{M}} \Big| \sum_{t=1}^T \sum_{\substack{j,k \\ j\neq k}} g_{t,j} g_{t,k} \langle M^j, N^k\rangle + \sum_{t=1}^T \sum_j (g_{t,j}^2 - 1) \|M^j\|_2^2 \Big| \Big\|_{L_p}$$

$$\lesssim \left\| \sup_{M\in\mathcal{M}} \left| \sum_{t=1}^{T} \sum_{j,k} g_{t,j} g'_{t,k} \langle M^j, M^k \rangle \right| \right\|_{L_p} = \left\| \sup_{M\in\mathcal{M}} \sum_{t=1}^{T} \langle M\mathbf{g}_t, M\mathbf{g}'_t \rangle \right\|_{L_p}$$

$$\lesssim_L \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \left\| \sup_{M\in\mathcal{M}} \left\| \sum_{t=1}^{T} M\boldsymbol{\xi}_t \right\|_2 \right\|_{L_p} + \left\| \sup_{M\in\mathcal{M}} \left| \sum_{t=1}^{T} \langle M\mathbf{g}_t, M\mathbf{g}'_t \rangle \right| \right\|_{L_p} .$$

where in the last inequality we have used (63) with a single set $\mathcal{M}$.

Fix $M \in \mathcal{M}$ and set $S = \{M^\top Mx : x \in B_2^n\}$. Since the random vectors $\mathbf{g}_t, t \in [T]$ are subgaussian, the random variable $\sum_{t=1}^{T} \langle \mathbf{g}_t, M^\top M\mathbf{g}'_t \rangle$ is subgaussian conditionally on $\mathbf{g}'_t, t \in [T]$. Therefore,

$$\begin{aligned}
\left\| \sum_{t=1}^{T} \langle M\mathbf{g}_t, M\mathbf{g}'_t \rangle \right\|_{L_p} &= \left(\mathbb{E}_{\mathbf{g}'_t} \left(\left(\mathbb{E}_{\mathbf{g}_t} \left(\sum_{t=1}^{T} \langle \mathbf{g}_t, M^\top M\mathbf{g}'_t \rangle \right)^p \right)^{1/p} \right)^p \right)^{1/p} \\
&\lesssim \left(\mathbb{E}_{\mathbf{g}'_t} \left(L\sqrt{p}^p \left\| \sum_{t=1}^{T} M^\top M\mathbf{g}'_t \right\|_2^p \right) \right)^{1/p} \\
&= L\sqrt{p} \left(\mathbb{E}_{\mathbf{g}'_t} \left\| \sum_{t=1}^{T} M^\top M\mathbf{g}'_t \right\|_2^p \right)^{1/p} \\
&= L\sqrt{p} \left(\mathbb{E}_{\mathbf{g}'_t} \sup_{y\in S} \left| \sum_{t=1}^{T} \langle y, \mathbf{g}'_t \rangle \right|^p \right)^{1/p} .
\end{aligned}$$

Now we use Theorem 2.3 from [Krahmer et al., 2014b]. The first term in the rhs can be bounded as

$$\begin{aligned}
\mathbb{E}_{\mathbf{g}'_t} \sup_{y\in S} \left| \sum_{t=1}^{T} \langle y, \mathbf{g}'_t \rangle \right|^p = \mathbb{E} \left\| \sum_{t=1}^{T} M^\top M\mathbf{g}'_t \right\|_2 &\le \left(\mathbb{E} \left\| \sum_{t=1}^{T} M^\top M\mathbf{g}'_t \right\|_2^2 \right)^{1/2} \\
&\le \sqrt{T} \left\| M^\top M \right\|_F = \sqrt{T} \left\| M \right\|_F \left\| M \right\|_{2\to 2} .
\end{aligned}$$

For the second term,

$$\begin{aligned}
\sup_{y\in S} \left(\mathbb{E} \left| \sum_{t=1}^{T} \langle y, \boldsymbol{\xi}'_t \rangle \right|^p \right)^{1/p} &= \sup_{z\in B_2^n} \left(\mathbb{E} \left| \sum_{t=1}^{T} \langle M^\top Mz, \boldsymbol{\xi}'_t \rangle \right|^p \right)^{1/p} \\
&\lesssim L \sup_{z\in B_2^n} \sqrt{pT} \left\| M^\top Mz \right\|_2 = L\sqrt{pT} \left\| M \right\|_{2\to 2}^2 .
\end{aligned}$$

Hence applying Theorem 2.3 from Krahmer et al. [2014b] and taking the supremum over $M \in \mathcal{M}$ we get

$$\left\| \sum_{t=1}^{T} \langle M\mathbf{g}_t, M\mathbf{g}'_t \rangle \right\|_{L_p} \lesssim \sqrt{T} \left(\sqrt{p} d_F(\mathcal{M}) d_{2\to 2}(\mathcal{M}) + p d_{2\to 2}(\mathcal{M}) \right)$$

Therefore,

$$\mathbb{E}(\sup_{M\in\mathcal{M}} \left\| \sum_{t=1}^{T} M\mathbf{g}_t \right\|_2)^2 \le C_{\mathcal{M}}(\mathbf{g}^{1:T}) + d_F^2(\mathcal{M}) \lesssim \gamma_2(\mathcal{M}, \|\cdot\|) \, \mathbb{E}(\sup_{M\in\mathcal{M}} \left\| \sum_{t=1}^{T} M\mathbf{g}_t \right\|_2) + \sqrt{T} \left(d_F^2(\mathcal{M}) \right)$$

which implies

$$\mathbb{E}(\sup_{M\in\mathcal{M}} \left\| \sum_{t=1}^{T} M\mathbf{g}_t \right\|_2) \le \gamma_2(\mathcal{M}, \|\cdot\|) + \sqrt{T} d_F(\mathcal{M})$$

Combining all these we get

$$\Big\| \sup_{M\in\mathcal{M}} \Big\| \sum_{t=1}^{T} M\boldsymbol{\xi}_t \Big\|_2 \Big\|_{L_p} \leq \gamma_2(\mathcal{M}, \|\cdot\|) + \sqrt{T} d_F(\mathcal{M}) + \sqrt{pT} d_{2\to 2}(\mathcal{M})$$

which completes the proof of Theorem I.4. □

Therefore

$$\begin{aligned}\|B_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} \lesssim\ & \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + \sqrt{T} d_F(\mathcal{N}) + \sqrt{pT} d_{2\to 2}(\mathcal{N}) \Big] \\ &+ \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \Big[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + \sqrt{T} d_F(\mathcal{M}) + \sqrt{pT} d_{2\to 2}(\mathcal{M}) \Big] \\ &+ \sup_{M\in\mathcal{M}, N\in\mathcal{N}} \|\langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle\|_{L_p} \end{aligned} \tag{64}$$

Next we consider $\sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big\| \sum_{t=1}^{T} \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p}$ and bound it using the following Lemma.

Lemma I.5. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1, and let $\boldsymbol{\xi}'$ be a decoupled tangent sequence. Then, for every $p \geq 1$,*

$$\begin{aligned}\sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big\| \sum_{t=1}^{T} \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p} \lesssim\ & \min\Big\{\sqrt{p}\cdot d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}), \sqrt{p}\cdot d_F(\mathcal{N})\cdot d_{2\to 2}(\mathcal{M})\Big\} \\ &+ p\cdot d_{2\to 2}(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N})\ . \end{aligned} \tag{65}$$

Proof of Lemma I.5. Fix $M \in \mathcal{M}, N \in \mathcal{N}$ and let $S = \{M^\top N x : x \in B_2^n, M \in \mathcal{M}, N \in \mathcal{N}\}$, where $B_2^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. Conditioned on $\boldsymbol{\xi}_t$, the random variable $\sum_{t=1}^{T}\langle \boldsymbol{\xi}_t, M^\top N\boldsymbol{\xi}_t'\rangle$ is sub-gaussian. Therefore for some global constant $\tilde{C} > 0$, we have

$$\Big\| \sum_{t=1}^{T} \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p} \lesssim L\sqrt{p} \left(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'} \sup_{y\in S} |\sum_{t=1}^{T} \langle y, \boldsymbol{\xi}_t'\rangle|^p \right)^{1/p}$$

Now using Theorem 2.3 from Krahmer et al. [2014a] we have for every $p \geq 1$

$$\left(\mathop{\mathbb{E}}_{\boldsymbol{\xi}'} \sup_{y\in S} |\sum_{t=1}^{T} \langle y, \boldsymbol{\xi}_t'\rangle|^p \right)^{1/p} \lesssim \left(\mathop{\mathbb{E}}_{\mathbf{g}_t} \sup_{y\in S} |\sum_{t=1}^{T} \langle \mathbf{g}_t, y\rangle| + \sup_{y\in S} (\mathop{\mathbb{E}}_{\boldsymbol{\xi}_t'} |\sum_{t=1}^{T} \langle \boldsymbol{\xi}_t', y\rangle|^p)^{1/p} \right)$$

where $\mathbf{g}_t, t \in [T]$ are standard Gaussian vector. The first term in the rhs can be bounded as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{g}_t} \sup_{y\in S} |\sum_{t=1}^{T} \langle \mathbf{g}_t, y\rangle| = \mathop{\mathbb{E}}_{\mathbf{g}} \|\sum_{t=1}^{T} M^\top N\mathbf{g}_t\|_2 \leq (\mathop{\mathbb{E}}_{\mathbf{g}} \|\sum_{t=1}^{T} M^\top N\mathbf{g}_t\|_2^2)^{1/2} \leq \sqrt{T}\|M^\top N\|_F \\ \leq \sqrt{T}\min\Big\{\|M\|_{2\to 2}\|N\|_F, \|N\|_{2\to 2}\|M\|_F\Big\}\end{aligned}$$

Next the second term in the rhs can be bounded as follows:

$$\sup_{y\in S} (\mathop{\mathbb{E}}_{\boldsymbol{\xi}'} |\sum_{t=1}^{T} \langle \boldsymbol{\xi}_t', y\rangle|^p)^{1/p} \leq L \sup_{z\in B_2^p} \sqrt{pT}\|M^\top Nz\|_2 \leq L\sqrt{pT}\|M\|_{2\to 2}\|N\|_{2\to 2}.$$

Combining all the above bounds we get:

$$\Big\| \langle M\boldsymbol{\xi}, N\boldsymbol{\xi}'\rangle \Big\|_{L_p} \lesssim \sqrt{pT}\min\Big\{\|M\|_{2\to 2}\|N\|_F, \|N\|_{2\to 2}\|M\|_F\Big\} + p\sqrt{T}\|M\|_{2\to 2}\|N\|_{2\to 2} \tag{66}$$

Now recall that for the set $\mathcal{M}$, we have $d_F(\mathcal{M}) = \sup_{M\in\mathcal{M}} \|M\|_F$, and $d_{2\to2}(\mathcal{M}) = \sup_{A\in\mathcal{M}} \|A\|_{2\to2}$, which implies

$$\sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big\| \sum_{t=1}^{T} \langle M\boldsymbol{\xi}_t, N\boldsymbol{\xi}'_t\rangle \Big\|_{L_p} \lesssim \Bigg(\sqrt{pT} \min\Big\{ d_{2\to2}(\mathcal{M}) d_F(\mathcal{N}), d_{2\to2}(\mathcal{N}) d_F(\mathcal{M}) \Big\} + p\sqrt{T} d_{2\to2}(\mathcal{M}) d_{2\to2}(\mathcal{N}) \Bigg).$$

□

Combining Lemma I.5 with (64) completes the proof of Theorem I.2. □

I.4 The Diagonal Term $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})$

For the diagonal term, we have the following main result:

Theorem I.6. *Let $\boldsymbol{\xi}$ be a stochastic process satisfying Assumption 1. Then, for all $p \geq 1$, we have*

$$\begin{aligned} \big\| D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}) \big\|_{L_p} \lesssim \sqrt{T} \Bigg[& \gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) \cdot \Big(\gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) + d_F(\mathcal{N}) + \sqrt{p}\; d_{2\to2}(\mathcal{N}) \Big) \\ & + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to2}) \cdot \Big(\gamma_2(\mathcal{M}, \|\cdot\|_{2\to2}) + d_F(\mathcal{M}) + \sqrt{p}\; d_{2\to2}(\mathcal{M}) \Big) \\ & + \sqrt{p}\; \min\Big\{ d_F(\mathcal{M}) \cdot d_{2\to2}(\mathcal{N}), d_F(\mathcal{N}) \cdot d_{2\to2}(\mathcal{M}) \Big\} + p \cdot d_{2\to2}(\mathcal{M}) \cdot d_{2\to2}(\mathcal{N}) \Bigg]. \end{aligned}$$

I.4.1 Proof of Theorem 3.5

By definition of $D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi})$ and from Lemma 9 in Banerjee et al. [2019], we have

$$\begin{aligned} \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\|_{L_p} &= \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} (|\boldsymbol{\xi}_{t,j}|^2 - \mathbb{E}\,|\boldsymbol{\xi}_{t,j}|^2) \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p} \\ &\leq 2 \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} |\boldsymbol{\xi}_{t,j}|^2 \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p}, \end{aligned}$$

where $\{\varepsilon_{t,j}\}$ is a set of independent Rademacher variables independent of $\boldsymbol{\xi}$. Let $\{g_j\}$ be a sequence of independent Gaussian random variables. Since $\boldsymbol{\xi}_{t,j}$ is a L-sub-Gaussian random variable [Vershynin, 2018], there is an absolute constant c such that for all $t > 0$

$$\mathbb{P}\left(|\xi_j|^2 \geq tL^2\right) \leq c\mathbb{P}(g_j^2 \geq t)\ .$$

Then, from contraction of stochastic processes ([Ledaux and Talagrand, 1991, Lemma 4.6]), we have

$$\begin{aligned} \|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T}))\|_{L_p} &\leq 2 \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} |\boldsymbol{\xi}_{t,j}|^2 \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p} \\ &\leq 2cL^2 \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} |g_{t,j}|^2 \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p} \\ &\overset{(a)}{\leq} 2cL^2 \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} (|g_{t,j}|^2 - 1) \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p} + 2cL^2 \Bigg\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \Big| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} \langle M_j, N_j\rangle \Big| \Bigg\|_{L_p} \end{aligned}$$

$$\overset{(b)}{\leq} 4cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} (|g_{t,j}|^2 - 1)\langle M_j, N_j\rangle \right| \right\|_{L_p} + 2cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j}\langle M_j, N_j\rangle \right| \right\|_{L_p}$$

$$\leq 4cL^2 \left\| D_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} + 2cL^2 \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j}\langle M_j, N_j\rangle \right| \right\|_{L_p}, \tag{67}$$

where (a) follows from Jensen's inequality and since $E|g_j|^2 = 1$, and (b) follows by de-symmetrization following [Banerjee et al., 2019, Lemma 11] and since the convex function here is 1-Lipschitz.

By triangle inequality, we have

$$\left\| D_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} \leq \left\| C_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} + \left\| B_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} \tag{68}$$

In order to handle $\left\| C_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p}$, we use our new Theorem D.4. We have

$$\begin{aligned}
\left\| C_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} &= \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{\substack{j,k=1\\ j\neq k}}^{n} \mathbf{g}_{t,j}\mathbf{g}_{t,k}\langle M_j, N_k\rangle + \sum_{t=1}^{T}\sum_{j=1}^{n} (|\mathbf{g}_{t,j}|^2 - \mathbb{E}\,|\mathbf{g}_{t,j}|^2)\langle M_j, N_j\rangle \right| \\
&\overset{(a)}{\leq} C \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j,k=1}^{n} \mathbf{g}_{t,j}\mathbf{g'}_{t,k}\langle M_j, N_k\rangle \right| \right\|_{L_p} = \left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T} \langle M\mathbf{g}_t, N\mathbf{g'}_t\rangle \right| \right\|_{L_p} \\
&\overset{(b)}{\lesssim} \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \left[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p} d_{2\to 2}(\mathcal{N}) \right] \\
&\quad + \gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \left[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M}) \right] \\
&\quad + \min\left\{ \sqrt{p}\cdot d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}), \sqrt{p}\cdot d_F(\mathcal{N})\cdot d_{2\to 2}(\mathcal{M}) \right\} \\
&\qquad + p\cdot d_{2\to 2}(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}) \,.
\end{aligned}$$

where (a) uses Theorem D.4, and (b) holds because of Lemma I.3 and Lemma I.5. Term $\left\| B_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p}$ can be bounded using Theorem I.6 thus giving

$$\begin{aligned}
\left\| D_{\mathcal{M},\mathcal{N}}(\mathbf{g}^{1:T}) \right\|_{L_p} \lesssim \sqrt{T} \Bigg[& \gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) \cdot \left[\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p} d_{2\to 2}(\mathcal{N}) \right] \\
& + \sqrt{T}\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \cdot \left[\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p} d_{2\to 2}(\mathcal{M}) \right] \\
& + \min\left\{ \sqrt{p}\cdot d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}), \sqrt{p}\cdot d_F(\mathcal{N})\cdot d_{2\to 2}(\mathcal{M}) \right\} \\
& + p\cdot d_{2\to 2}(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}) \,. \Bigg]
\end{aligned} \tag{69}$$

Next we bound the second term $\left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j}\langle M_j, N_j\rangle \right| \right\|_{L_p}$ using our new Theorem D.5.

According to [Banerjee et al., 2019, Lemma 7],

$$\begin{aligned}
&\left\| \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j}\langle M_j, N_j\rangle \right| \right\|_{L_p} \\
&= \left(\mathbb{E}_\varepsilon \left[\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j}\langle M_j, N_j\rangle \right|^p \right] \right)^{1/p}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbb{E}_\varepsilon \left[\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} \langle M_j, N_j\rangle \right|^p \right] \right)^{1/p} \\
&\lesssim \mathbb{E}_{\mathbf{g}} \left[\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \langle M_j, N_j\rangle \right| \right] + \sup_{M\in\mathcal{M},N\in\mathcal{N}} \left(\mathbb{E}_\varepsilon \left[\left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} \langle M_j, N_j\rangle \right|^p \right] \right)^{1/p} \qquad (70)
\end{aligned}
$$

in which $\mathbf{g}_{t,j} \sim \mathcal{N}(0,1)$ are independent. For the first term, define $X_{(M,N)} = \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \langle M_j, N_j\rangle \right|$. Then

$$
\begin{aligned}
\left| X_{(M^1,N)} - X_{(M^2,N)} \right| &= \left| \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \left\langle M_j^1, N_j \right\rangle \right| - \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \left\langle M_j^2, N_j \right\rangle \right| \right| \\
&= \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \left\langle M_j^1 - M_j^2, N_j \right\rangle \right| \\
&\le \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \left\langle M_j^1 - M_j^2, N_j \right\rangle
\end{aligned}
$$

Since $\mathbf{g}_j$ are standard normal random variables, therefore with probability $1 - 2e^{-u^2/2}$,

$$
\begin{aligned}
\left| X_{(M^1,N)} - X_{(M^2,N)} \right| &\le u\sqrt{T} \left(\sum_{j=1}^{n} \langle M_j^1 - M_j^2, N_j\rangle^2 \right)^{\frac{1}{2}} \\
&\le u\sqrt{T} \left(\sum_{j=1}^{n} \left\| M_j^1 - M_j^2 \right\|_2^2 \|N_j\|_2^2 \right)^{\frac{1}{2}} \\
&\le u\sqrt{T} d_F\left(\mathcal{N}\right) \left\| M^1 - M^2 \right\|_{2\to 2}
\end{aligned}
$$

Similarly with probability $1 - 2e^{-u^2/2}$,

$$
\left| X_{(M,N^1)} - X_{(M,N^2)} \right| \le u\sqrt{T} d_F\left(\mathcal{M}\right) \left\| N^1 - N^2 \right\|_{2\to 2}
$$

Using Theorem D.5, with probability $1 - 4e^{-u^2/2}$,

$$
\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{t,j} \langle M_j, N_j\rangle \right| \lesssim u\sqrt{T} \left(d_F(\mathcal{N})\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M})\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2}) \right).
$$

so we have

$$
\mathbb{E}_{\mathbf{g}} \left[\sup_{M\in\mathcal{M},N\in\mathcal{N}} \left| \sum_{t=1}^{T}\sum_{j=1}^{n} \mathbf{g}_{j} \langle M_j, N_j\rangle \right| \right] \lesssim \sqrt{T}(d_F(\mathcal{N})\gamma_2(\mathcal{M}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{M})\gamma_2(\mathcal{N}, \|\cdot\|_{2\to 2})). \qquad (71)
$$

For the second term, for each $M \in \mathcal{M}, N \in \mathcal{N}$, since ε_j are Rademacher random variables, therefore with probability at least $1 - 2e^{-u^2}$,

$$
\begin{aligned}
\left| \sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_{t,j} \langle M_j, N_j\rangle \right| &\le u\sqrt{T} \left(\sum_{j=1}^{n} \langle M_j, N_j\rangle^2 \right)^{\frac{1}{2}} \le u\sqrt{T} \left(\sum_{j=1}^{n} \|M_j\|_2^2 \|N_j\|_2^2 \right)^{\frac{1}{2}} \\
&\le \sqrt{T} \min\left\{ d_F\left(\mathcal{M}\right) d_{2\to 2}\left(\mathcal{N}\right), d_F\left(\mathcal{N}\right) d_{2\to 2}\left(\mathcal{M}\right) \right\}.
\end{aligned}
$$

According to [Vershynin, 2018, Proposition 2.5.2],

$$\left(\mathbb{E}_{\varepsilon}\left[\left|\sum_{t=1}^{T}\sum_{j=1}^{n}\varepsilon_{t,j}\langle M_j, N_j\rangle\right|^{p}\right]\right)^{1/p} \lesssim \sqrt{pT}\min\left\{d_F\left(\mathcal{M}\right)d_{2\to 2}\left(\mathcal{N}\right), d_F\left(\mathcal{N}\right)d_{2\to 2}\left(\mathcal{M}\right)\right\},$$

Using (71) and (70), we have

$$\begin{aligned}
&\left\|\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{t=1}^{T}\sum_{j=1}^{n}\varepsilon_{t,j}\langle M_j, N_j\rangle\right|\right\|_{L_p}\\
\lesssim&\mathbb{E}_{\mathbf{g}_t}\left[\sup_{M\in\mathcal{M},N\in\mathcal{N}}\left|\sum_{t=1}^{T}\sum_{j=1}^{n}\mathbf{g}_{t,j}\langle M_j, N_j\rangle\right|\right] + \sup_{M\in\mathcal{M},N\in\mathcal{N}}\left(\mathbb{E}_{\varepsilon_t}\left[\left|\sum_{t=1}^{T}\sum_{j=1}^{n}\varepsilon_{t,j}\langle M_j, N_j\rangle\right|^{p}\right]\right)^{1/p}\\
\lesssim&\sqrt{T}\Big[d_F(\mathcal{N})\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2}) + d_F(\mathcal{M})\gamma_2(\mathcal{N},\|\cdot\|_{2\to 2})\Big]\\
&+\sqrt{pT}\min\left\{d_F\left(\mathcal{M}\right)d_{2\to 2}\left(\mathcal{N}\right), d_F\left(\mathcal{N}\right)d_{2\to 2}\left(\mathcal{M}\right)\right\}
\end{aligned} \tag{72}$$

Substituting (69) and (72) into (67), we get

$$\begin{aligned}
\left\|D_{\mathcal{M},\mathcal{N}}(\boldsymbol{\xi}^{1:T})\right\|_{L_p} \lesssim \sqrt{T}\Bigg[&\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2})\cdot\Big[\gamma_2(\mathcal{N},\|\cdot\|_{2\to 2}) + d_F(\mathcal{N}) + \sqrt{p}d_{2\to 2}(\mathcal{N})\Big]\\
&+\gamma_2(\mathcal{N},\|\cdot\|_{2\to 2})\cdot\Big[\gamma_2(\mathcal{M},\|\cdot\|_{2\to 2}) + d_F(\mathcal{M}) + \sqrt{p}d_{2\to 2}(\mathcal{M})\Big]\\
&+\sqrt{p}\min\Big\{d_F(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N}), d_F(\mathcal{N})\cdot d_{2\to 2}(\mathcal{M})\Big\}\\
&+p\cdot d_{2\to 2}(\mathcal{M})\cdot d_{2\to 2}(\mathcal{N})\Bigg].
\end{aligned}$$

That completes the proof of Theorem I.2.

J SKETCHED CONTEXTUAL BANDITS

In a bandit problem, a learner needs to make sequential decisions over T time steps. At any round $t \in [T]$, the learner observes the action set $\mathcal{A} \subseteq \mathbb{R}^d$. The learner chooses an arm a_t and then the associated reward of the arm $r(a_t) \in [0,1]$ is observed. We make the following assumption on the reward.

Assumption 11 (Linear Reward). *The reward $r(a_t)$ is given by $r(a_t) = \langle a_t, \theta^*\rangle + \eta_t$, where $\theta^* \in \Theta^*$ is an unknown parameter vector and η_t is a conditionally sub-Gaussian noise, i.e., $\forall\lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda\eta_t}|a_1,\dots a_t,\eta_1,\dots,\eta_{t-1}] \le \exp(\frac{\lambda^2}{2})$. Further we assume that $\|a\|_2 = 1$ for all $a \in \mathcal{A}, t \in [T]$.*

Definition J.1 (Regret). *The learner's goal is to minimize the regret for the selected actions $a_t, t \in [T]$, defined as*

$$\texttt{Reg}_{\texttt{CB}}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(r(a^*) - r(a_t))\right] = \sum_{t=1}^{T}\langle\theta^*, a^*\rangle - \langle\theta^*, a_t\rangle\ ,$$

where $a^* = \text{argmax}_{a_t\in\mathcal{A}}\langle\theta^*, a_t\rangle$, maximizes the expected reward in round t.

J.1 Sketched Linear UCB

We develop a sketched version of the popular algorithm LinUCB (Linear Upper Confidence Bound [Abbasi-Yadkori et al., 2011]) and is summarized in Algorithm 3. At every round t the learner sketches the inputs using $\mathcal{S}_t \subseteq \mathbb{R}^{b\times d}$ whose entries are drawn i.i.d. from $N(0, 1/b)$. It then solves a regularized least-squares problem (cf

Algorithm 3 `sk-LinUCB` (Sketched Linear UCB)

1: **for** $t = 1, 2, ..., T$ **do**
2: Solve the least squares regression problem given by

$$\hat{\theta}_t^s = \text{argmin}_{\theta \in \mathbb{R}^b, \|\theta\|_2 \leq 1} \sum_{i=1}^{t-1} (\langle \theta, \mathcal{S}_t a_i \rangle - r_i)^2 + \lambda \|\theta\|_2^2 \tag{73}$$

3: Construct the Confidence Set $\mathcal{C}_t^s$ according to (75).
4: Compute the the optimistic estimates of the parameter and the action:

$$(\tilde{\theta}_t^s, a_t^s) = \text{argmax}_{\theta \in \mathcal{C}_{t-1}^s, a \in \mathcal{S}_t \mathcal{A}} \langle \theta, a \rangle \tag{74}$$

5: De-sketch the action $a_t = \mathcal{S}_t^\top a_t^s$
6: Play the action a_t and observe the reward r_t.
7: **end for**

(73) in Algorithm 3) in the b-dimensional sketched space to obtain an estimate $\hat{\theta}_t^s$ of the unknown parameter. Subsequently, with $\bar{V}_t^s = \sum_{i=1}^t \mathcal{S}_t a_i (\mathcal{S}_t a_i)^\top + \lambda I$, the learner constructs a confidence set $\mathcal{C}_t^s$ around the estimate:

$$\mathcal{C}_t^s = \left\{ \theta \in \mathbb{R}^b : \|\hat{\theta}_t^s - \mathcal{S}_t \theta\|_{\bar{V}_t^s} \leq \sqrt{\log\Big(\frac{\det(\bar{V}_t^s)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \Big)} + \lambda^{1/2} \|\mathcal{S}_t \theta_*\|_2 \right\}. \tag{75}$$

The algorithm then uses this confidence set to compute an *optimistic* parameter and action pair $(\tilde{\theta}_t^s, a_t^s)$ (Line 4) where $\tilde{\theta}_t^s \in \mathcal{C}_t^s$ and the action is in the *sketched* action space $a \in \mathcal{S}_t \mathcal{A}$. Once this sketched action is identified, it is "*de-sketched*" (Line 5) to recover the corresponding action a_t in the original space. Finally, the chosen action a_t is played, and the observed reward r_t is used in subsequent rounds to refine future estimates. Our primary result in this section is the following decomposition for the regret.

Theorem J.1 (Regret Decomposition for `Sk-LinUCB`). *Suppose Assumption 11 holds, the actions are selected according to Algorithm 3, and the de-sketched actions selected by Algorithm 3 are in $\mathcal{A}$. Then after some burn-in period t_{burn}, for some constant $C > 0$, with probability $1 - \delta$*

$$\begin{aligned} \texttt{Reg}_{\texttt{CB}}(T) \leq C \Bigg[& \sqrt{b \log(T/\delta)} + \frac{1}{\sqrt{b}} \Big(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})} \Big) \Bigg] \\ & \sqrt{Tb \log(1 + T \max_t \omega^2(\mathcal{A})/b^2) \log\log(2/\delta)} \\ & + \sum_{t=1}^T {\theta^*}^\top (I - \mathcal{S}_t^\top \mathcal{S}_t) a^* \end{aligned}$$

where $\omega(\mathcal{M}) = \mathbb{E}[\sup_{\theta \in \mathcal{M}} \langle g, \theta \rangle]$, $g \sim N(0, I_d)$ is the Gaussian width of the set $\mathcal{M}$.

Note that term I captures the regret in the sketched b dimensional space while term II captures the restricted isometry term due to random sketching.

Before we move to the proof of Theorem J.1, we discuss what it entails. First note that we can bound term II following the analysis in Section G, Section 4.2 and invoking Theorem 5.2, we get

$$\begin{aligned} P\Bigg\{ \sup_{\theta^* \in \Theta, a \in \mathcal{A}} |\sum_{t=1}^T {\theta^*}^\top (\mathcal{S}_t^\top \mathcal{S}_t - I) a| \geq \sqrt{T} \Big(\frac{1}{b} \omega(\Theta_*) \omega(\mathcal{A}) + \frac{1}{\sqrt{b}} (\omega(\Theta_*) + \omega(\mathcal{A})) \Big) + \epsilon \Bigg\} \\ \lesssim \exp\left(- \min\left\{ \frac{\epsilon^2}{T \big(\frac{1}{b} (\omega(\Theta_*) + \omega(\mathcal{A})) + \frac{1}{\sqrt{b}} \big)^2}, \frac{\epsilon}{\sqrt{T} \frac{1}{b}} \right\} \right) \end{aligned}$$

Choose $\epsilon^2 = T(\frac{1}{b}(\omega(\Theta_*)+\omega(\mathcal{A}))+\frac{1}{\sqrt{b}})^2\bar{\epsilon}^2$, and take a union bound over both the events to get with probability $1-\delta$

$$\sup_{\theta^*\in\Theta, a\in\mathcal{A}} |\sum_{t=1}^{T} {\theta^*}^\top(\mathcal{S}_t^\top\mathcal{S}_t - I)a| \lesssim \sqrt{T}(\frac{1}{b}\omega(\Theta_*)\omega(\mathcal{A}) + \frac{1}{\sqrt{b}}(\omega(\Theta_*)+\omega(\mathcal{A})) + (\frac{1}{b}(\omega(\Theta_*)+\omega(\mathcal{A}))+\frac{1}{\sqrt{b}})\sqrt{\log(\delta^{-1})})$$

Combining with Theorem J.1 we have, with high probability

$$\texttt{Reg}_{\texttt{CB}}(T) = \underbrace{\tilde{O}\left(\sqrt{bT}\left[\sqrt{b}+\frac{1}{\sqrt{b}}\omega(\Theta_*)\right]\right)}_{I} + \underbrace{\sqrt{T}\frac{1}{b}\omega(\Theta_*)\omega(\mathcal{A}) + \frac{1}{\sqrt{b}}(\omega(\Theta_*)+\omega(\mathcal{A}))}_{II}$$

Note that the above bound depends on the Gaussian widths of the parameter and action set instead of the ambient dimension d. In cases when the Gaussian widths are small (eg., sparse) one can choose the sketching dimension b to obtain a tighter regret bound.

Proof of Theorem J.1 Recall that the regret is given by

$$\begin{aligned}\texttt{Reg}_{\texttt{CB}}(T) &= \sum_{t=1}^{T}\langle\theta^*, a^*\rangle - \langle\theta^*, a_t\rangle \\ &= \sum_{t=1}^{T}\langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle + \langle\theta^*, a^*\rangle - \langle\theta^*, a_t\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle\end{aligned}$$

To be able to use the optimistic estimates $(\tilde{\theta}_t^s, a_t^s)$ computed in line 5 of Algorithm 3, we first show that the sketched unknown parameter belongs to the confidence set.

Lemma J.2 (Confidence Ellipsoid). *Suppose $\mathcal{C}_t^s$ be the confidence set defined as in (75). Then for all $t\in[T]$, with probability $1-\delta$, we have $\mathcal{S}_t\theta^*\in\mathcal{C}_t^s$.*

Further note that $\mathcal{S}_t a_t^* \in \mathcal{S}_t\mathcal{A}$ and therefore from the definition of the optimistic estimates, $\langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle \le \langle\tilde{\theta}_t^s, a_t^s\rangle$. Using this we have

$$\begin{aligned}\texttt{Reg}_{\texttt{CB}}(T) &\le \sum_{t=1}^{T}\langle\tilde{\theta}_t^s, a_t^s\rangle + \langle\theta^*, a^*\rangle - \langle\theta^*, a_t\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle \\ &\overset{(a)}{=} \sum_{t=1}^{T}\langle\tilde{\theta}_t^s, a_t^s\rangle - \langle\mathcal{S}_t\theta^*, a_t^s\rangle + \langle\mathcal{S}_t\theta^*, a_t^s\rangle + \langle\theta^*, a^*\rangle - \langle\theta^*, a_t\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle \\ &= \sum_{t=1}^{T}\langle\tilde{\theta}_t^s - \mathcal{S}_t\theta^*, a_t^s\rangle + \langle\mathcal{S}_t\theta^*, a_t^s\rangle - \langle\theta^*, a_t\rangle + \langle\theta^*, a^*\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle \\ &\overset{(b)}{=} \sum_{t=1}^{T}\langle\tilde{\theta}_t^s - \mathcal{S}_t\theta^*, a_t^s\rangle + \langle\mathcal{S}_t\theta^*, a_t^s\rangle - \langle\theta^*, \mathcal{S}_t^\top a_t^s\rangle + \langle\theta^*, a^*\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle\end{aligned}$$

where (a) follows by adding and subtracting $\langle\mathcal{S}_t\theta^*, a_t^s\rangle$ and (b) follows by noting that $a_t = \mathcal{S}_t^\top a_t^s$. Therefore

$$\begin{aligned}\texttt{Reg}_{\texttt{CB}}(T) &\le \sum_{t=1}^{T}\langle\tilde{\theta}_t^s - \mathcal{S}_t\theta^*, a_t^s\rangle + \langle\mathcal{S}_t\theta^*, a_t^s\rangle - \langle\mathcal{S}_t\theta^*, a_t^s\rangle + \langle\theta^*, a^*\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t a^*\rangle \\ &= \sum_{t=1}^{T}\langle\tilde{\theta}_t^s - \mathcal{S}_t\theta^*, a_t^s\rangle + {\theta^*}^\top(I - \mathcal{S}_t^\top\mathcal{S}_t)a^* \\ &\le \underbrace{\sum_{t=1}^{T}\left\|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\right\|_{\bar{V}_{t-1}^s}\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}}_{(A)} + \underbrace{\sum_{t=1}^{T}{\theta^*}^\top(I-\mathcal{S}_t^\top\mathcal{S}_t)a^*}_{(B)}\end{aligned}$$

where the last inequality follows by Cauchy Schwartz. Note that the inner products in term (A) are b-dimensional. The price we paid for reducing the original d-dimensional inner product into this b-dimensional inner product is term (B). Using the Confidence Ellipsoid in (75) and elliptical potential lemma we bound term (A) in the following lemma.

Lemma J.3. *Suppose $(\tilde{\theta}_t^s, a_t^s)$ be as computed in (74) and suppose $\lambda = 1$. Then for some constant $C > 0$, with probability $1 - \delta$*

$$\sum_{t=1}^{T} \|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\|_{\bar{V}_{t-1}^s} \|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}$$
$$\leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\sqrt{Tb\log(1+T\omega^2(\mathcal{A})/b^2)\log\log(2/\delta)}\right)$$

Finally using Lemma J.3 we have for some constant $C > 0$ with probability $1 - \delta$

$$\mathtt{Reg}_{\mathtt{CB}}(T) \leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\right.$$
$$\left.\sqrt{Tb\log(1+T\max_t \omega^2(\mathcal{A})/b^2)\log\log(2/\delta)}\right)$$
$$+ \sum_{t=1}^{T} {\theta^*}^\top (I - \mathcal{S}_t^\top \mathcal{S}_t) a^* \qquad \square$$

J.1.1 Proof of Auxiliary Lemmas

Proof of Lemma J.2. Suppose $\bar{V}_t^s = \sum_{i=1}^{t-1} \mathcal{S}_t a_i (\mathcal{S}_t a_i)^\top + \lambda I$. Then we can express $\hat{\theta}_t^s$ as follows:

$$\hat{\theta}_t^s = (\bar{V}_t^s)^{-1}\Big(\sum_{i=1}^{t-1}(\mathcal{S}_t a_i) r_i\Big)$$
$$\overset{(a)}{=} (\bar{V}_t^s)^{-1}\Big(\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\langle\theta^*, \mathcal{S}_t^\top(\mathcal{S}_t a_i)\rangle + \eta_i + \varepsilon_i\Big)$$

where in (a), $\varepsilon_i = \langle\theta^*, a_i - \mathcal{S}_t^\top \mathcal{S}_t a_i\rangle$ and we have used the fact that $r_i = \langle\theta^*, a_i\rangle + \eta_i$. Therefore

$$\hat{\theta}_t^s = (\bar{V}_t^s)^{-1}\Big(\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)(\mathcal{S}_t a_i)^\top \mathcal{S}_t\theta^*\Big) + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\eta_i + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\varepsilon_i$$
$$= (\bar{V}_t^s)^{-1}\Big(\underbrace{\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)(\mathcal{S}_t a_i)^\top + \lambda I}_{\bar{V}_t^s}\Big)\mathcal{S}_t\theta^* - \lambda(\bar{V}_t^s)^{-1}\mathcal{S}_t\theta^* + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\eta_i + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\varepsilon_i$$
$$= \mathcal{S}_t\theta^* + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\eta_i - \lambda(\bar{V}_t^s)^{-1}\mathcal{S}_t\theta^* + (\bar{V}_t^s)^{-1}\sum_{i=1}^{t-1}(\mathcal{S}_t a_i)\varepsilon_i$$

Therefore

$$\|\hat{\theta}_t^s - \mathcal{S}_t\theta^*\| \leq \underbrace{\Big\|\sum_{i=1}^{t-1} a_i^s \eta_i\Big\|_{(\bar{V}_t^s)^{-1}}}_{I} + \underbrace{\Big\|\sum_{i=1}^{t-1} a_i^s \varepsilon_i\Big\|_{(\bar{V}_t^s)^{-1}}}_{II} + \underbrace{\lambda^{1/2}\|\mathcal{S}_t\theta^*\|}_{III}$$

Consider term I first. Using the Self-Normalized Bound for Vector-Valued Martingales (Theorem 1, Abbasi-Yadkori et al. [2011]) we have, with probability $1-\delta$

$$\Big\|\sum_{i=1}^{t-1} a_i^s \eta_i\Big\|_{(\bar{V}_t^s)^{-1}} \leq 2\sqrt{\log\left(\frac{\det(\bar{V}_t^s)^{1/2}\det(\lambda I_b)^{-1/2}}{\delta}\right)}$$

Now using the fact that $\|a\| \leq 1$ and $\lambda = 1$ we have

$$\Big\|\sum_{i=1}^{t-1} a_i^s \eta_i\Big\|_{(\bar{V}_t^s)^{-1}} \leq 2\sqrt{b\log\left(\frac{1+t}{\delta}\right)}$$

Next we handle term II. Note that ε_i is not sub-Gaussian and therefore we cannot use the above. It rather follows a Bernstein type condition with the moment generating function given by: for every λ with $|\lambda| < 1/b_{\text{Bern}}$,

$$\mathbb{E}\exp\{\lambda(X-\mu)\} \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b_{\text{Bern}}|\lambda|}\right).$$

where $b_{\text{Bern}} = \Theta(\frac{1}{\sqrt{b}})$ and $\sigma^2 = \Theta(1/b)$. We use a recent result [Ziemann, 2025] that provides self-normalized bounds for such Bernstein type processes. We invoke Theorem 1 from Ziemann [2025] to conclude that

$$\alpha \triangleq \left(\frac{\sqrt{e(1+\nu)}\,\|S_\tau\|_{(V_\tau+\Gamma)^{-1}V(V_\tau+\Gamma)^{-1}}}{\nu\sqrt{d+2}} - 1\right)\vee 0.$$

Then as long as $V_\tau + \Gamma \succeq e(1+\nu)^2 V \succeq (1+\nu)^2 e^{-1}(d+2)B_W^2 B_X^2$, we have that with probability at least $1-\delta$:

$$\|S_\tau\|^2_{(V_\tau+\Gamma)^{-1}} \leq \left(\frac{(1+\alpha)^2}{1+2\alpha}\times\frac{1}{1-\epsilon}\right)\times\sigma^2\times\left[\log\left(\frac{\det(V_\tau+\Gamma)}{\det(V)}\right)+2\log\left(\frac{1}{\delta}\right)\right].$$

where $V_\tau = \bar{V}_\tau^s, V = \lambda I$ and $S_\tau = \Big\|\sum_{i=1}^{\tau-1} a_i^s\epsilon_i\Big\|_{(\bar{V}_t^s)^{-1}}$.

We can set $\nu = \epsilon = 1/2$ and after a burn-in period t_{burn} such that $\bar{V}_\tau^s + \Gamma \succeq (1+\nu)^2 e^{-1}(d+2)B_W^2 B_X^2$ and $\sqrt{e(1+\nu)}\,\|S_\tau\|_{(V_\tau+\Gamma)^{-1}V(V_\tau+\Gamma)^{-1}} \leq \nu\sqrt{d+2}$ we have that with probability at least $1-\delta$:

$$\|S_\tau\|^2_{(V_\tau+\Gamma)^{-1}} \lesssim \sigma^2\times\left[\log\left(\frac{\det(V_\tau+\Gamma)}{\det(V)}\right)+2\log\left(\frac{1}{\delta}\right)\right].\lesssim\sqrt{b\log\left(\frac{1+t}{\delta}\right)}$$

Finally, to control term III, since $\mathcal{S}_t \sim N(0,1/b)$ we have with probability $1-2\exp(c\omega^2(\Theta_*)-p^2)$ for some $c>0$, (see Banerjee et al. [2014], Theorem 5)

$$\|\mathcal{S}_t\theta_*\| \leq 1 + \frac{\omega(\Theta_*)+p}{\sqrt{b}}$$

Therefore with probability $(1-\delta)$, for some constant $C>0$

$$\|\mathcal{S}_t\theta_*\| \leq 1 + \frac{C}{\sqrt{b}}(\omega(\Theta_*)+\sqrt{\log(2/\delta)})$$

Combining all these it is immediate that $\mathcal{S}_t\theta^* \in \mathcal{C}_t^s$. Further, plugging all the bounds back we get, with probability $(1-\delta)$, for some constant $C>0$

$$\|\hat{\theta}_t^s - \mathcal{S}_t\theta^*\| \leq 2\sqrt{b\log\left(\frac{1+t}{\delta}\right)} + 1 + \frac{C}{\sqrt{b}}(\omega(\Theta_*)+\sqrt{\log(2/\delta)})$$

□

Lemma J.3. *Suppose* $(\tilde{\theta}_t^s, a_t^s)$ *be as computed in* (74) *and suppose* $\lambda = 1$. *Then for some constant* $C > 0$, *with probability* $1-\delta$

$$\sum_{t=1}^{T} \left\|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\right\|_{\bar{V}_{t-1}^s} \|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}$$
$$\leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\sqrt{Tb\log(1+T\omega^2(\mathcal{A})/b^2)\log\log(2/\delta)}\right)$$

Proof. Note that $\tilde{\theta}_t^s \in \mathcal{C}_t^s = \left\{\theta \in \mathbb{R}^b : \|\hat{\theta}_t^s - \mathcal{S}_t\theta^*\|_{\bar{V}_{t-1}^s} \leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+t}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\right)\right\}$ which along with $t \leq T$ immediately implies that with probability $1-\delta$

$$\sum_{t=1}^{T}\left\|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\right\|_{\bar{V}_{t-1}^s}\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}} \leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\right)\sum_{t=1}^{T}\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}$$

Using Cauchy Schwarz

$$\sum_{t=1}^{T}\left\|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\right\|_{\bar{V}_{t-1}^s}\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}} \leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\right)$$
$$\sqrt{T\sum_{t=1}^{T}\min\left(\|a_t^s\|^2_{(\bar{V}_{t-1}^s)^{-1}}, 1\right)}$$

Using Lemma 11 from Abbasi-Yadkori et al. [2011] we have for $\|a_t^s\| \leq L$,

$$\sum_{t=1}^{T}\min(\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}, 1) \leq b\log\left(1 + \frac{TL^2}{b\lambda}\right)$$

Now since for any $a \in \mathcal{A}$, $\|a\| \leq 1$, using the same proof technique to bound $\|\mathcal{S}_t\theta_*\|$ in Lemma J.2 we have that with probability $1-\delta$, $\|a_t^s\| \leq 1 + C\frac{\omega(\mathcal{A})+\sqrt{\log(2/\delta)}}{\sqrt{b}}$, for some $C > 0$. Taking a union bound for all $t \in [T]$ and with the event that $\tilde{\theta}_t^s \in \mathcal{C}_t^s$ we have, with probability $1-\delta$

$$\sum_{t=1}^{T}\left\|\tilde{\theta}_t^s - \mathcal{S}_t\theta^*\right\|_{\bar{V}_{t-1}^s}\|a_t^s\|_{(\bar{V}_{t-1}^s)^{-1}}$$
$$\leq \mathcal{O}\left(\sqrt{b\log\left(\frac{1+T}{\delta}\right)} + \frac{1}{\sqrt{b}}\left(\omega(\Theta_*) + \sqrt{\log(\frac{2}{\delta})}\right)\right)\sqrt{Tb\log(1+T\omega^2(\mathcal{A})/b^2)\log\log(2/\delta)}$$

□

J.2 Sketched Thompson Sampling

Next we show that a similar decomposition can also be obtained for Thompson Sampling algorithm ([Agrawal and Goyal, 2013, 2012]). We consider a finite action set as in Agrawal and Goyal [2013], i.e., $\mathcal{A} = \{\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \ldots, \mathbf{x}_{t,K}\}$, where K is the number of arms. The learner chooses an action $i_t \in [K]$ and observes the reward $\langle \mathbf{x}_{t,i_t}, \theta^*\rangle + \eta_t$, i.e., $a_t = \mathbf{x}_{t,i_t}$ in Assumption 11.

We take a Bayesian approach by placing a Gaussian prior $\boldsymbol{\theta_t^s} \sim \mathcal{N}(\boldsymbol{\mu}_0^s, \boldsymbol{\Sigma}_0^s)$, where $\boldsymbol{\mu}_0^s = \mathbf{0} \in \mathbb{R}^b$ and $\boldsymbol{\Sigma}_0^s = \boldsymbol{I}_{b\times b} \in \mathbb{R}^{b\times b}$ are the prior mean and covariance matrix, respectively. Note that the mean vector and covariance matrix are in the b-dimensional space. After t observations, the posterior distribution of θ_t^s is given by $\theta_t^s \mid \{\mathbf{x}_{\tau,i_\tau}, r(\mathbf{x}_{\tau,i_\tau})\}_{\tau=1}^t \sim \mathcal{N}(\boldsymbol{\mu}_{t-1}^s, v^2(\boldsymbol{\Sigma}_{t-1}^s)^{-1})$, where the posterior mean $\boldsymbol{\mu}_t^s$ and covariance $\boldsymbol{\Sigma}_t^s$ are given by standard Bayesian linear regression update equations:

$$\boldsymbol{\Sigma_t^s} = \left(\boldsymbol{\Sigma_0^s} + \sum_{\tau=1}^{t-1}\mathbf{x}_{\tau,a_\tau}^s\mathbf{x}_{\tau,a_\tau}^{s\top}\right)$$

Algorithm 4 `sk-LinTS` (Sketched Linear TS)

Input: variance parameter $v = \sqrt{9b\log(t/\delta)}$
for $t = 1, 2, \ldots$ **do**
  Sample θ_t^s from $\mathcal{N}(\boldsymbol{\mu}_t^s,\ v^2(\boldsymbol{\Sigma}_t^s)^{-1})$.
  Play arm $i_t := \arg\max_{i\in[K]} \mathbf{x}_{t,i_t}^\top \theta_t^s$, and observe reward $r_t = r_{\mathbf{x}_{t,i_t}}(t)$.
  Update $\boldsymbol{\Sigma}_t^s$ and $\boldsymbol{\mu}_t^s$ according to (76).
end for

$$\boldsymbol{\mu}_t^s = (\boldsymbol{\Sigma}_t^s)^{-1}\left(\sum_{\tau=1}^{t-1} \mathbf{x}_{\tau,i_\tau}^s r(\mathbf{x}_{\tau,i_\tau})\right) \tag{76}$$

where $\mathbf{x}_{t,i}^s = \mathcal{S}_t \mathbf{x}_{t,i}$ is the sketched context vector.

The regret decomposition for `sk-LinTS` is given by the following Theorem.

Theorem J.4 (Regret Decomposition for `Sk-LinTS`). *Suppose Assumption 11 holds and the actions are selected according to Algorithm 3. Then for some constant $C > 0$, with probability $1-\delta$*

$$\mathtt{Reg}_{\mathtt{CB}}(T) \le C(\underbrace{6\sqrt{b\log(KT)\log(T/\delta)} + \sqrt{b\log(T^3/\delta)})(\sqrt{bT\log T} + Z\sqrt{T\log(1/\delta)})}_{I}$$
$$\underbrace{+\,2\sup_{\theta\in\Theta_*}\sup_{a_t\in\mathcal{A},t\in[T]}\sum_{t=1}^{T}\theta^\top(S^\top S - I)a_t}_{II},$$
$$\text{where } Z = \tilde{\mathcal{O}}\left(\frac{1}{b}\omega(\Theta_*)\right),$$

and $\omega(\mathcal{M}) = \mathbb{E}[\sup_{\theta\in\mathcal{M}}\langle g, \theta\rangle]$, $g \sim N(0, I_d)$ is the Gaussian width of the set $\mathcal{M}$.

As in the case of `sk-LinUCB` regret decomposition, term I captures the regret in the sketched b dimensional space while term II captures the restricted isometry term.

Proof. The regret is given by

$$\begin{aligned}
\mathtt{Reg}_{\mathtt{CB}}(T) &= \langle\theta^*, \mathbf{x}_{t,a^*}\rangle - \langle\theta^*, \mathbf{x}_{t,a_t}\rangle = \langle\theta^*, \mathbf{x}_{t,a^*}\rangle - \langle\theta^*, \mathbf{x}_{t,a_t}\rangle \\
&\quad + \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}^s\rangle + \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}^s\rangle \\
&= \langle\theta^*, \mathbf{x}_{t,a^*}\rangle - \langle\theta^*, \mathbf{x}_{t,a_t}\rangle + \langle\mathcal{S}_t\theta^*, \mathcal{S}_t\mathbf{x}_{t,a^*}\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t\mathbf{x}_{t,a_t}\rangle - \langle\mathcal{S}_t\theta^*, \mathcal{S}_t\mathbf{x}_{t,a^*}\rangle + \langle\mathcal{S}_t\theta^*, \mathcal{S}_t\mathbf{x}_{t,a_t}\rangle \\
&= \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}^s\rangle + 2\sup_{\theta\in\Theta_*}\sup_{a_t\in\mathcal{A},t\in[T]}\sum_{t=1}^{T}\theta^\top(S^\top S - I)a_t
\end{aligned}$$

Suppose $\tilde{a}_t = \arg\min_{i\in[K]} \sqrt{{\mathbf{x}_{t,i}^s}^\top \boldsymbol{\Sigma}_t^{s-1}\mathbf{x}_{t,i}^s}$. Then

$$\begin{aligned}
\langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}\rangle &= \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}^s\rangle + \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,\tilde{a}_t}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,\tilde{a}_t}^s\rangle \\
&\quad + 2\sup_{\theta\in\Theta_*}\sup_{a_t\in\mathcal{A},t\in[T]}\sum_{t=1}^{T}\theta^\top(S^\top S - I)a_t \\
&= \underbrace{\langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a^*}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,\tilde{a}_t}^s\rangle + \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,\tilde{a}_t}^s\rangle - \langle\mathcal{S}_t\theta^*, \mathbf{x}_{t,a_t}^s\rangle}_{I}
\end{aligned}$$

$$+ 2 \sup_{\theta \in \Theta_*} \sup_{a_t \in \mathcal{A}, t \in [T]} \sum_{t=1}^{T} \theta^\top (S^\top S - I) a_t \tag{77}$$

We define the following events

$$E_1 = = \left\{ \forall i : \left| \, {\mathbf{x}^s_{t,i}}^\top \hat{\theta}_t - {\mathbf{x}^s_{t,i}}^\top \mathcal{S}_t \theta_* \, \right| \leq (\sqrt{b \ln(t^3/\delta)} + 1) \sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} \right\} \tag{78}$$

$$E_2 = \left\{ \forall i : \left| \, \zeta_{t,i} - {\mathbf{x}^s_{t,i}}^\top \hat{\theta}_t \, \right| \leq 6 \sqrt{b \log(Kt) \ln(2t/\delta)} \, {\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i} \right\} \tag{79}$$

We bound term I in the following lemma.

Lemma J.5. *Suppose E_1 and E_2 be as defined in Equation (78) and Equation (79) and let $g_t = 6\sqrt{b \log(Kt) \log(t/\delta)} + \sqrt{b \log(t^3/\delta)} + 1$ and $p = \frac{1}{4e\sqrt{\pi}}$. Then for any filtration $\mathcal{F}_{t-1}$ such that E_1 is true, we have*

$$\mathbb{E}[\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle | \mathcal{F}_{t-1}] \leq \frac{3}{p} g_t \left(\mathbb{E}[\sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} | \mathcal{F}_t] + \frac{1}{t^2} \right)$$

Proof. Note that

$$\begin{aligned} &\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle \\ &\leq 2 g_t \sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + g_t \sqrt{{\mathbf{x}^s_{t,\tilde{a}_t}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,a_t}} \end{aligned}$$

follows when events E_1 and E_2 are assumed to be true. Therefore

$$\begin{aligned} &\mathbb{E}[\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle] \\ &\qquad \leq \mathbb{E}[2 g_t \sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + g_t \sqrt{{\mathbf{x}^s_{t,\tilde{a}_t}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,a_t}} | \mathcal{F}_t] + P(E_2^C) \\ &\qquad \leq \frac{3}{p} g_t \, \mathbb{E} \left[\sqrt{{\mathbf{x}^s_{t,\tilde{a}_t}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,a_t}} | \mathcal{F}_t \right] + \frac{2 g_t}{p t^2} \end{aligned}$$

where we have used Lemma 1 and 4 from Agrawal and Goyal [2013]. □

Now note that using Lemma J.5, we have $Y_s = \sum_{t=1}^{s} (\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle) \mathbb{1}\{E_1\} - \frac{3}{p} g_t \left(\sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + \frac{1}{t^2} \right)$ is super-martingale with respect to the filtration $\mathcal{F}_t = \sigma\{(r_i, a_i), 1 \leq i \leq t-1, a_t\}$. The following lemma provides a bound on the absolute value of super-martingale process for every t with high probability.

Lemma J.6. *Suppose $X_t = (\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle) \mathbb{1}\{E_1\} - \frac{3}{p} g_t \left(\sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + \frac{1}{t^2} \right)$, where $g_t = 6\sqrt{b \log(tK) \ln(t/\delta)} + \sqrt{b \ln(t^3/\delta)} + 1$. Then with probability $(1-\delta)$ for some constant $C > 0$*

$$|X_t| \leq C \left(\frac{1}{b} \big(\omega(\Theta_*) + \sqrt{\log(2/\delta)} \big) \big(\sqrt{\log(KT)} + \sqrt{\log(2/\delta)} \big) + \frac{3}{p} g_t \frac{\sqrt{\log(KT)} + \sqrt{\log(2/\delta)}}{\sqrt{b}} \right)$$

Proof. Note that

$$\begin{aligned} X_t &= (\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle + \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,\tilde{a}_t} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle) \mathbb{1}\{E_1\} - \frac{3}{p} g_t \left(\sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + \frac{1}{t^2} \right) \\ &= (\langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a^*} \rangle - \langle \mathcal{S}_t \theta^*, \mathbf{x}^s_{t,a_t} \rangle) \mathbb{1}\{E_1\} - \frac{3}{p} g_t \left(\sqrt{{\mathbf{x}^s_{t,i}}^\top \boldsymbol{\Sigma}_t^{\boldsymbol{s}-1} \mathbf{x}^s_{t,i}} + \frac{1}{t^2} \right) \end{aligned}$$

Now consider $\sqrt{{\mathbf{x}^s_{t,i}}^\top {\mathbf{\Sigma}^s_t}^{-1} \mathbf{x}^s_{t,i}}$. We know that $\mathbf{\Sigma}^s_t \succeq \lambda I_b$ implying $\lambda_{\min}({\mathbf{\Sigma}^s_t}^{-1}) \leq \lambda$. Therefore

$$\sqrt{{\mathbf{x}^s_{t,i}}^\top {\mathbf{\Sigma}^s_t}^{-1} \mathbf{x}^s_{t,i}} \leq \sqrt{\lambda} \|\mathbf{x}^s_{t,i}\|_2 \leq \|\mathbf{x}^s_{t,i}\|_2$$

Since $\mathcal{S}_t \sim N(0, 1/b)$ we have with probability $1 - 2\exp(c\omega^2(\mathcal{A}) - q^2)$ for some $c > 0$, (see Banerjee et al. [2014], Theorem 5)

$$\|\mathcal{S}_t \mathbf{x}_{t,a}\| \leq 1 + \frac{\omega(\mathcal{A}) + q}{\sqrt{b}}$$

where $\mathcal{A} = \{\mathbf{x}_{t,k}, t \in [T], k \in [K]\}$. Therefore with probability $(1-\delta)$, for all$t \in [T], k \in [K]$, for some constant $C > 0$

$$\begin{aligned}\|\mathbf{x}^s_{t,k}\| &\leq 1 + \frac{C}{\sqrt{b}}(\omega(\mathcal{A}) + \sqrt{\log(2/\delta)}) \\ &\leq 1 + \frac{C}{\sqrt{b}}(\sqrt{\log(KT)} + \sqrt{\log(2/\delta)})\end{aligned}$$

Next, to control $|\langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a^*}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a_t}\rangle|$, note that by Cauchy Schwarz

$$|\langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a^*}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a_t}\rangle| \leq 2\|\mathcal{S}_t\theta^*\|\|\mathbf{x}^s_{t,a_t}\|$$

Using the same technique as above, we have with probability $(1-\delta)$

$$\|\mathcal{S}_t\theta^*\| \leq 1 + \frac{C}{\sqrt{b}}(\omega(\Theta_*) + \sqrt{\log(2/\delta)})$$

Taking a union bound over all the events we have, with probability $(1-\delta)$ for some constant $C > 0$

$$|X_t| \leq C\left(\frac{1}{b}\big(\omega(\Theta_*) + \sqrt{\log(2/\delta)}\big)\big(\sqrt{\log(KT)} + \sqrt{\log(2/\delta)}\big) + \frac{3}{p} g_t \frac{\sqrt{\log(KT)} + \sqrt{\log(2/\delta)}}{\sqrt{b}}\right)$$

□

Therefore taking a union bound over the event in Lemma J.6 and using using Azuma-Hoeffding inequality we have with probability $1-\delta$ we have

$$\begin{aligned}&\sum_{t=1}^T \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a^*}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,\tilde{a}_t}\rangle + \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,\tilde{a}_t}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a_t}\rangle \\ &\quad\leq \frac{3}{p} g_T \sum_{t=1}^T \left(\sqrt{{\mathbf{x}^s_{t,i}}^\top {\mathbf{\Sigma}^s_t}^{-1} \mathbf{x}^s_{t,i}} + \frac{1}{t^2}\right) + C\sqrt{2\sum_t \frac{9}{p^2} g_t^2 Z^2 \log(4/\delta)} \\ &\quad\leq \frac{3}{p} g_T 5\sqrt{bT \ln T} + \frac{6}{p} g_T C \sqrt{\sum_t Z^2 \log(1/\delta)}\end{aligned}$$

where $Z = \frac{1}{b}\big(\omega(\Theta_*) + \sqrt{\log(4/\delta)}\big)\big(\sqrt{\log(KT)} + \sqrt{\log(2/\delta)}\big) + \frac{3}{p}\frac{\sqrt{\log(KT)} + \sqrt{\log(2/\delta)}}{\sqrt{b}}$ and the last step follows from Lemma 11 of Chu et al. [2011]. Therefore with probability $1-\delta$ we have

$$\begin{aligned}&\sum_{t=1}^T \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a^*}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,\tilde{a}_t}\rangle + \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,\tilde{a}_t}\rangle - \langle \mathcal{S}_t\theta^*, \mathbf{x}^s_{t,a_t}\rangle \\ &\quad\leq C g_T(\sqrt{bT\log T} + Z\sqrt{T\log(1/\delta)}) \\ &\quad\leq C(6\sqrt{b\log(KT)\log(T/\delta)} + \sqrt{b\log(T^3/\delta)} + 1)(\sqrt{bT\log T} + Z\sqrt{T\log(1/\delta)})\end{aligned}$$

Combining with Equation (77) we get with probability $1-\delta$

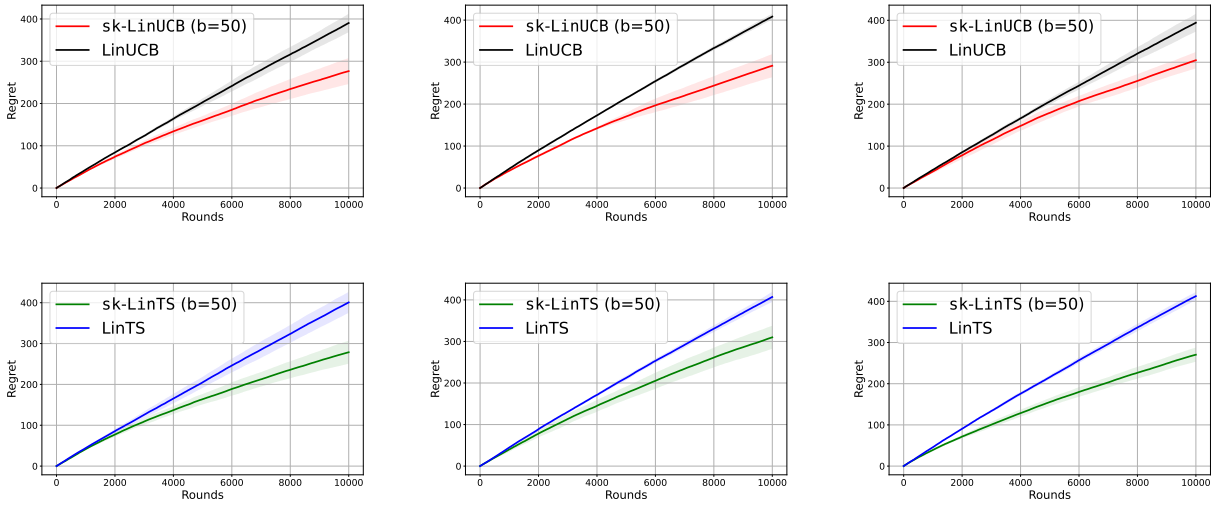


Figure 1: Comparison of cumulative regret of sk-LinUCB and sk-LinTS with the baselines LinUCB and LinTS on a synthetic dataset, averaged over 5 runs. The results are shown for three sparsity cases: (a) context sparsity, (b) parameter sparsity, and (c) both context and parameter sparsity. The sketching dimension $b = 50$.

$$\begin{aligned}\texttt{Reg}_{\texttt{CB}}(T) \leq C(6\sqrt{b\log(KT)\log(T/\delta)} + \sqrt{b\log(T^3/\delta)} + 1)(\sqrt{bT\log T} + Z\sqrt{T\log(1/\delta)}) \\ + 2\sup_{\theta\in\Theta_*}\sup_{a_t\in\mathcal{A},t\in[T]}\sum_{t=1}^{T}\theta^\top(S^\top S - I)a_t,\end{aligned}$$

completing the proof. □

J.3 Experiments

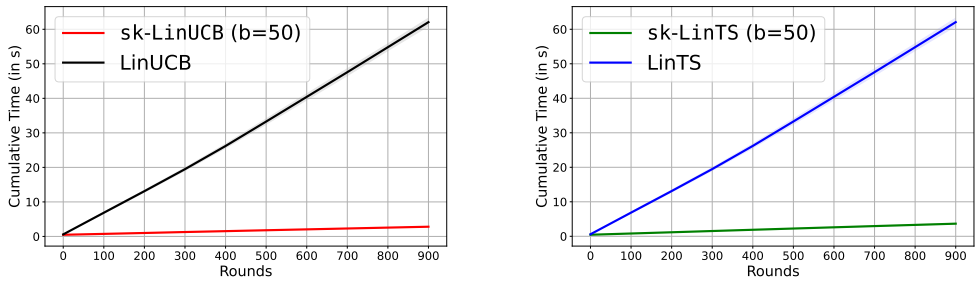
In this section, we evaluate the performance of sk-LinUCB and sk-LinTS in comparison to their unsketched counterparts, LinUCB and LinTS, on synthetic datasets. We consider a contextual bandit setting with a context dimension of $d = 500$ and $K = 4$ possible actions. The number of rounds is set to $T = 10,000$.

Dataset: To illustrate the effects of sketching, we construct a dataset with low metric entropy through truncation. The context vectors $a_t \in \mathcal{A}_t$ and the unknown parameter θ_* are sampled uniformly from the unit ball $\mathcal{B}(0, 1)$. We investigate three scenarios: (a) only the context vectors are sparse, (b) only the unknown parameter θ_* is sparse, and (c) both the context vectors and the unknown parameter θ_* are sparse. For sparsity, we apply truncation to both the context vectors and θ_*. Specifically, for a vector of dimension $d = 500$, we retain only the first $s = 50$ coordinates (0.9 sparsity).

Sketching: In all experiments, we use a random Gaussian matrix $S \in \mathbb{R}^{b\times d}$, where each entry $S_{ij} \sim \mathcal{N}(0, 1/b)$. The sketching dimension is set to $b = 50$. The reward function is given by: $h(a_t) = a_t^T\theta_* + \eta_t$, where $\eta_t \sim \mathcal{N}(0, 1)$.

Results: The results for both the baseline and sketched algorithms are shown in Figure 1. In all three scenarios, the sketched algorithms achieve lower regret than their unsketched counterparts, even with a sketching dimension of $b = 50$. To compare computational cost, we also plot the cumulative runtime for each algorithm in Figure 2, highlighting significant computational savings for the sketched methods.

We additionally consider linear bandit with a convex action set, where the action space is ℓ_2 unit ball in $\mathbb{R}^d$ with $d = 1000$ and the unknown parameter $\theta^\star$ lies in an ℓ_1 ball. At each round we select an action via projected



(a) `sk-LinUCB` (b) `sk-LinTS`

Figure 2: Comparison of run time (in seconds) for `sk-LinUCB` and `sk-LinTS` with the baselines LinUCB and LinTS on a synthetic dataset, averaged over 5 runs. The sketching dimension is denoted by b.

gradient descent (100 steps) and report the cumulative regret (mean with standard error in parentheses across 4 runs) in Table 1.

Table 1: Cumulative regret for linear bandits with an ℓ_2 action set and an ℓ_1-constrained unknown parameter. Entries report the mean regret with standard error in parentheses across 4 runs.

Algorithm	Step 50	Step 100	Step 200	Step 500	Step 800	Step 999
LinearBandits	49.30 (0.40)	96.23 (0.47)	186.30 (1.19)	436.22 (3.91)	658.79 (11.34)	793.60 (15.98)
SketchLinearBandits	48.43 (0.53)	92.38 (1.28)	172.04 (3.01)	356.80 (5.07)	503.76 (5.19)	600.12 (5.23)

Finally we evaluate the performance on non-gaussian sketches for our bandit setup, and report the regret in Table 2. For this, we set the sketching dimension to be 100. Similar to Appendix G, the contexts $x_{i,t}$ and parameter $\theta^\star$ are drawn from an ℓ_2 ball and we set sparsity to 95%.

Table 2: Cumulative regret for different sketching distributions in the linear bandit setup with sketch dimension 100. Entries report the mean regret with standard error in parentheses across 4 runs.

Algorithm	Step 5000	Step 6000	Step 7000	Step 8000	Step 9000	Step 10000
LinUCB	57.51 (6.25)	68.52 (7.71)	78.62 (9.61)	89.13 (11.54)	99.68 (13.38)	110.20 (15.58)
SkLinUCB (Gaussian)	16.10 (0.77)	18.10 (0.85)	20.13 (0.84)	21.90 (1.05)	23.67 (1.04)	25.30 (1.11)
SkLinUCB (SRHT)	52.89 (0.04)	63.32 (0.02)	73.60 (0.06)	83.99 (0.04)	94.26 (0.03)	104.84 (0.02)
SkLinUCB (CountSketch)	54.28 (2.06)	65.10 (2.77)	75.96 (3.27)	87.05 (3.92)	97.68 (4.50)	108.39 (5.13)