A Nearly-Optimal Bound for Fast Regression with ℓ_∞ Guarantee

Zhao Song [1] **Mingquan Ye** [2] **Junze Yin** [3] **Lichen Zhang** [4]

Abstract

Given a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$, we consider the regression problem with ℓ_∞ guarantees: finding a vector $x' \in \mathbb{R}^d$ such that

$$\|x' - x^*\|_\infty \leq \frac{\epsilon}{\sqrt{d}} \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|,$$

where $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$. One popular approach for solving such ℓ_2 regression problem is via sketching: picking a structured random matrix $S \in \mathbb{R}^{m \times n}$ with $m \ll n$ and SA can be quickly computed, solve the "sketched" regression problem $\arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$. In this paper, we show that in order to obtain such ℓ_∞ guarantee for ℓ_2 regression, one has to use sketching matrices that are *dense*. To the best of our knowledge, this is the first user case in which dense sketching matrices are necessary. On the algorithmic side, we prove that there exists a distribution of dense sketching matrices with $m = \epsilon^{-2} d \log^3(n/\delta)$ such that solving the sketched regression problem gives the ℓ_∞ guarantee, with probability at least $1 - \delta$. Moreover, the matrix SA can be computed in time $O(nd \log n)$. Our row count is nearly-optimal up to logarithmic factors, and significantly improves the result in (Price et al., 2017), in which a super-linear in d rows, $m = \Omega(\epsilon^{-2} d^{1+\gamma})$ for $\gamma \in (0, 1)$ is required. Moreover, we develop a novel analytical framework for ℓ_∞ guarantee regression that utilizes the *Oblivious Coordinate-wise Embedding* (OCE) property introduced in (Song & Yu, 2021). Our analysis is much simpler and more general than that of (Price et al., 2017). Leveraging this framework, we extend the ℓ_∞ guarantee regression result to dense sketching matrices for computing the fast tensor product of vectors.

[1]Adobe Research [2]University of Illinois at Chicago [3]Boston University [4]MIT. Correspondence to: Zhao Song <zsong@adobe.com>, Junze Yin <junze@bu.edu>, Lichen Zhang <lichenz@mit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Linear regression, or ℓ_2 least-square problem is ubiquitous in numerical linear algebra, scientific computing and machine learning. Given a tall skinny matrix $A \in \mathbb{R}^{n \times d}$ and a label vector $b \in \mathbb{R}^n$, the goal is to (approximately) compute an optimal solution x' that minimizes the ℓ_2 loss $\|Ax - b\|_2$. For the regime where $n \gg d$, sketching is a popular approach to obtain an approximate solution quickly (Clarkson & Woodruff, 2013; Nelson & Nguyên, 2013; Meng & Mahoney, 2013; Cherapanamjeri et al., 2023): the idea is to pick a random matrix $S \in \mathbb{R}^{m \times n}$ from carefully-designed distributions, so that 1). S can be efficiently applied to A and 2). the row count $m \ll n$. Given these two guarantees, one can then solve the "sketched" regression problem:

$$\arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2,$$

and obtain a vector x' such that $\|Ax' - b\|_2 = (1 \pm \epsilon) \cdot \mathrm{OPT}$, where OPT denotes the optimal ℓ_2 discrepancy between vectors in column space of A and b. Recent advances in sketching (Cherapanamjeri et al., 2023) show that one can design matrix S with $m = O(\epsilon^{-2} d \log^2(n/\delta))$ and the sketched regression can be solved in time

$$O(\mathrm{nnz}(A) + d^\omega + \epsilon^{-2} d^2 \operatorname{poly} \log(n, d, 1/\epsilon, 1/\delta))$$

where $\mathrm{nnz}(A)$ denotes the number of nonzero entries of A and δ is the error probability.

Unfortunately, modern machine learning emphasizes more and more on large, complex, and nonlinear models such as deep neural networks, thus linear regression becomes less appealing as a *model*. However, it is still a very important *subroutine* in many deep learning and optimization frameworks, especially second-order method for training neural networks (Cai et al., 2019; Brand et al., 2021; Song et al., 2021b) or convex optimization (Lee & Sidford, 2014; Cohen et al., 2019; Lee et al., 2019; Brand, 2020; Song & Yu, 2021; Jiang et al., 2021; Gu & Song, 2022). In these applications, one typically seeks *forward error* guarantee (Sarlos, 2006), i.e., how close is the approximate solution x' to the optimal solution x^*. A prominent example is Newton's method: given the (possibly implicit) Hessian matrix $H \in \mathbb{R}^{d \times d}$ and the gradient $g \in \mathbb{R}^d$, one wants to compute $H^{-1} g$. A common approach is to solve the regression $\arg\min_{x \in \mathbb{R}^d} \|Hx - g\|_2$, in which one wants $\|x - H^{-1} g\|_2$

or even $\|x - H^{-1}g\|_\infty$ to be small. When the matrix S satisfies the so-called Oblivious Subspace Embedding (OSE) property (Sarlos, 2006), one can show that the approximate solution x' is close to x^* in the ℓ_2 sense:

$$\|x' - x^*\|_2 \leq O(\epsilon) \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|. \quad (1)$$

Unfortunately, ℓ_2-closeness cannot characterize how good x' approximates x^*, as x^* can have a good spread of ℓ_2 mass over all coordinates while x' concentrates its mass over a few coordinates. Formally speaking, let $a \in \mathbb{R}^d$ be a fixed vector, then one can measure how far $\langle a, x' \rangle$ deviates from $\langle a, x^* \rangle$ via Eq. (1):

$$\begin{aligned} & |\langle a, x^* \rangle - \langle a, x' \rangle| \\ = & |\langle a, x' - x^* \rangle| \\ \leq & \|a\|_2 \|x' - x^*\|_2 \\ \leq & O(\epsilon) \cdot \|a\|_2 \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|. \end{aligned}$$

This bound is clearly too loose, as one would expect the deviation on a random direction is only $\frac{1}{\sqrt{d}}$ factor of the ℓ_2 discrepancy. (Price et al., 2017) shows that this intuition is indeed true when S is picked as the subsampled randomized Hadamard transform (SRHT) (Lu et al., 2013): [1]

$$\begin{aligned} & |\langle a, x^* \rangle - \langle a, x' \rangle| \\ \lesssim & \frac{\epsilon}{\sqrt{d}} \|a\|_2 \|Ax^* - b\|_2 \|A^\dagger\|. \end{aligned} \quad (2)$$

However, their analysis is not tight as they require a row count $m = \Omega(\epsilon^{-2} d^{1+\gamma})$ for $\gamma = \Theta(\sqrt{\frac{\log \log n}{\log d}})$. Such a row count is super-linear in d as long as $n \leq \exp(d)$ and therefore is worse than the required row count for S to be a subspace embedding, in which only $m = \epsilon^{-2} d \log^2 n$ rows are required for constant success probability. In contrast, for random Gaussian matrices, the ℓ_∞ guarantee only requires nearly linear in d rows. In addition to their sub-optimal row count, the (Price et al., 2017) analysis is also complicated: let $U \in \mathbb{R}^{n \times d}$ be an orthonormal basis of A, (Price et al., 2017) has to analyze the higher moment of matrix $I_d - U^\top S^\top S U$. This makes their analysis particularly hard to generalize to other dense sketching matrices beyond SRHT.

In this work, we present a novel framework for analyzing the ℓ_∞ guarantee induced by SRHT and more generally, a large class of *dense* sketching matrices. Our analysis is arguably much simpler than (Price et al., 2017), and it exposes the fundamental structure of sketching matrices that provides ℓ_∞ guarantee: if any two columns of the sketching matrix have a small inner product with high probability, then ℓ_∞ guarantee can be preserved. We then prove that the small pairwise column inner product is also closely related to the *Oblivious Coordinate-wise Embedding* (OCE) property introduced in (Song & Yu, 2021). More concretely, for any two fixed vectors $g, h \in \mathbb{R}^n$, we say the sketching matrix is (β, δ, n)-OCE if $|\langle Sg, Sh \rangle - \langle g, h \rangle| \leq \frac{\beta}{\sqrt{m}} \cdot \|g\|_2 \|h\|_2$ holds with probability at least $1 - \delta$. This property has previously been leveraged for approximating matrix-vector product between a dynamically-changing projection matrix and an online sequence of vectors for the fast linear program and empirical risk minimization algorithms (Lee et al., 2019; Jiang et al., 2021; Song & Yu, 2021; Qin et al., 2023) and related data structures (Song et al., 2023) as these algorithms need ℓ_∞ bound on the matrix-vector product. One common theme shared by those applications and ℓ_∞ guarantee is to use *dense* sketching matrices, such as random Gaussian, the Alon-Matias-Szegedy sketch (AMS, (Alon et al., 1996)) or SRHT. This is in drastic contrast with the trending direction for using sparse matrices such as Count Sketch (Charikar et al., 2002; Clarkson & Woodruff, 2013) and OSNAP (Nelson & Nguyên, 2013; Cohen, 2016), as they can be applied in (nearly) input sparsity time.

In recent years, sketches that can be applied to the tensor product of matrices/vectors have gained popularity (Avron et al., 2014; Song et al., 2016; Diao et al., 2018; Song et al., 2019; Diao et al., 2019; Ahle et al., 2020; Woodruff & Zandieh, 2020; Song et al., 2021a; Woodruff & Zandieh, 2022; Song et al., 2022b;a; Reddy et al., 2022) as they can speed up optimization tasks and large-scale kernel learning. We show that dense sketches for degree-2 tensors also provide ℓ_∞ guarantee.

Theorem 1.1 (Nearly-optimal bound for dense sketching matrices)**.** *Suppose $n \leq \exp(d)$ and matrix $A \in \mathbb{R}^{n \times d}$ and vector $b \in \mathbb{R}^n$ are given. Let $S \in \mathbb{R}^{m \times n}$ be a subsampled randomized Hadamard transform matrix* SRHT *with*

$$m = O(\epsilon^{-2} d \log^3(n/\delta))$$

rows.

For any fixed vector $a \in \mathbb{R}^d$,

$$|\langle a, x^* \rangle - \langle a, x' \rangle| \lesssim \frac{\epsilon}{\sqrt{d}} \cdot \|a\|_2 \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|$$

with probability $1 - \delta$, where

- $x' = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$,
- $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$.

Remark 1.2. The row count $m = \epsilon^{-2} d \log^3 n$ is nearly-optimal up to logarithmic factors, as the row count for S being an OSE is $m = \epsilon^{-2} d \log^2 n$ for constant success probability. In comparison, (Price et al., 2017) requires $m = \epsilon^{-2} d^{1+\gamma}$ rows for $\gamma = \Theta(\sqrt{\frac{\log \log n}{\log d}})$ which is only

[1] We will later refer the following property as ℓ_∞ guarantee.

nearly-linear in d if $n > \exp(d)$. In most applications, we concern about $n = \mathrm{poly}(d)$, meaning that their row count is worse than ours in almost all meaningful scenarios.

The row count and guarantee obtained in Theorem 1.1 extend beyond SRHT; in fact, for a range of *dense* sketching matrices including random Gaussian, AMS sketch (Alon et al., 1996), SRHT and subsampled randomized Circulant Transform (see Definition 2.11) (SRCT)[2]. This is because our argument is a structural condition that can be satisfied by various dense sketches.

Our result can also be generalized to degree-2 Kronecker product regression, see Theorem B.1.

Roadmap. In Section 2, we introduce the notations that we use and explain the key definitions and properties to support the framework for ℓ_∞ guarantee regression. In Section 3, we introduce our framework by presenting a sufficient condition for a sketching matrix to give a good ℓ_∞ guarantee. In Section 4, we provide a proof for our main theorem by putting everything together. Finally, in Section 5, we summarize the main findings of this paper and through comparing with previous work. In Section A, we introduce the fundamental definitions and properties that we will use in Appendix. In Section B, we analyze and develop the ℓ_∞ guarantee of Kronecker product regressions. In Section C, we introduce the Strong JL Moment Property and prove that both Circulant Transform and Tensor Circulant Transform satisfy this. In Section D, we focus on studying AMS, random Gaussian, and SRHT and show that the inner product is bounded on any pair of different columns of AMS, random Gaussian, and SRHT–dense sketching matrices.

2. Preliminary

For a positive integer, we define $[n] := \{1, 2, \cdots, n\}$.

For a vector $x \in \mathbb{R}^n$, we define $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$ and $\|x\|_\infty := \max_{i\in[n]} |x_i|$.

For a matrix A, we define $\|A\| := \sup_x \|Ax\|_2/\|x\|_2$ to be the spectral norm of A.

We use $\|A\|_F := \sum_{i,j} A_{i,j}^2$ to be the Frobenius norm of A.

In general, we have the following property for spectral norm, $\|AB\| \leq \|A\| \cdot \|B\|$.

We use $A^\dagger$ to denote the Moore-Penrose pseudoinverse of $m \times n$ matrix A which if $A = U\Sigma V^\top$ is its SVD (where $U \in \mathbb{R}^{m\times n}$, $\Sigma \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{n\times n}$ for $m \geq n$), is given by $A^\dagger = V\Sigma^{-1}U^\top$.

We use $\mathbb{E}[\cdot]$ to denote the expectation, and $\Pr[\cdot]$ to denote the probability.

For a distribution D and a random variable x, we use $x \sim D$ to denote that we draw a random variable from the distribution D.

We use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2.

We say a random variable x is Rademacher random variables if $\Pr[x = 1] = 1/2$ and $\Pr[x = -1] = 1/2$. We also call it a sign random variable.

In addition to $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq Cg$ (resp. $\geq$) for an absolute constant C. We use $f \eqsim g$ to mean $cf \leq g \leq Cf$ for constants $c > 0$ and $C > 0$. For two matrices $A \in \mathbb{R}^{n_1\times d_1}$, $B \in \mathbb{R}^{n_2\times d_2}$, we use $A \otimes B \in \mathbb{R}^{n_1 n_2\times d_1 d_2}$ to denote the Kronecker product, i.e., the (i_1, i_2)-th row and (j_1, j_2)-th column of $A \times B$ is $A_{i_1,j_1} B_{i_2,j_2}$. For two vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, we use $x \otimes y \in \mathbb{R}^{mn}$ to denote the tensor product of vectors, in which $x \otimes y = \mathrm{vec}(xy^\top)$.

2.1. Oblivious Subspace Embedding and Coordinate-wise Embedding

Definition 2.1 (Oblivious subspace embedding (Sarlos, 2006))**.** We define (ϵ, δ, d, n)-Oolivious subspace embedding (OSE) as follows: Suppose Π is a distribution on $m\times n$ matrices S, where m is a function of n, d, ϵ, and δ. Suppose that with probability at least $1 - \delta$, for any fixed $n \times d$ orthonormal basis U, a matrix S drawn from the distribution Π has the property that the singular values of SU lie in the range $[1 - \epsilon, 1 + \epsilon]$.

The oblivious coordinate-wise embedding (OCE) is implicitly used in (Lee et al., 2019; Jiang et al., 2021) and formally introduced in (Song & Yu, 2021).

Definition 2.2 ((α, β, δ)–coordinate wise embedding (Song & Yu, 2021))**.** We say a random matrix $S \in \mathbb{R}^{m\times n}$ satisfying (α, β, δ)–coordinate wise embedding if

1. $\mathop{\mathbb{E}}_{S\sim\Pi}[g^\top S^\top Sh] = g^\top h$,
2. $\mathop{\mathbb{E}}_{S\sim\Pi}[(g^\top S^\top Sh)^2] \leq (g^\top h)^2 + \dfrac{\alpha}{m}\|g\|_2^2\|h\|_2^2$,
3. $\Pr_{S\sim\Pi}[|g^\top S^\top Sh - g^\top h| \geq \dfrac{\beta}{\sqrt{m}}\|g\|_2\|h\|_2] \leq \delta$.

In this paper, we mainly use the property 3 of Definition 2.2. For convenient, we redefine OCE as follows:

Definition 2.3 (OCE)**.** Let $\beta \geq 1$ and $\delta \in (0, 0.1)$. We say a randomized matrix $S \in \mathbb{R}^{m\times n}$ satisfy (β, δ, n)-OCE, if

$$\Pr_{S\sim\Pi}[|g^\top S^\top Sh - g^\top h| \geq \frac{\beta}{\sqrt{m}}\|g\|_2\|h\|_2] \leq \delta$$

and the distribution Π is oblivious to any fixed vectors g, h.

[2]The Toeplitz and circulant type of Johnson-Lindenstrauss transform has also been studied (Freksen & Larsen, 2020).

2.2. Sketching Matrices

In this paper, we concern a list of dense sketching matrices.

Definition 2.4 (Random Gaussian matrix, folklore)**.** We say $S \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix if all entries are sampled from $\mathcal{N}(0, 1/m)$ independently.

Definition 2.5 (AMS sketch matrix, (Alon et al., 1996))**.** Let $h_1, h_2, \cdots, h_m$ be m random hash functions picking from a 4–wise independent hash family $\mathcal{H} = \{h : [n] \to \{-\frac{1}{m}, +\frac{1}{m}\}\}$. Then $S \in \mathbb{R}^{m \times n}$ is an AMS sketch matrix if we set $S_{i,j} = h_i(j)$.

The following sketching matrices can utilize fast Fourier Transform (FFT) for efficient application to matrices.

Definition 2.6 (Subsampled randomized Hadamard transform (SRHT) (Lu et al., 2013; Song et al., 2021a))**.** The SRHT matrix $S \in \mathbb{R}^{m \times n}$ is defined as $S := \frac{1}{\sqrt{m}} PHD$, where each row of matrix $P \in \{0, 1\}^{m \times n}$ contains exactly one 1 at a random position, H is the $n \times n$ Hadamard matrix, and D is a $n \times n$ diagonal matrix with each diagonal entry being a value in $\{-1, +1\}$ with equal probability.

Remark 2.7. Using the fast Fourier transform (FFT), S can be applied to a vector in time $O(n \log n)$.

Definition 2.8 (Tensor subsampled randomized Hadamard transform (TensorSRHT) (Ahle et al., 2020; Song et al., 2021a))**.** The TensorSRHT $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ is defined as

$$S := \frac{1}{\sqrt{m}} P \cdot (HD_1 \otimes HD_2),$$

where each row of $P \in \{0, 1\}^{m \times n^2}$ contains only one 1 at a random coordinate and one can view P as a sampling matrix. H is a $n \times n$ Hadamard matrix, and D_1, D_2 are two $n \times n$ independent diagonal matrices with diagonals that are each independently set to be a Rademacher random variable (uniform in $\{-1, 1\}$).

Remark 2.9. By leveraging the FFT algorithm in the sketch space, $S(x \otimes y)$ can be computed in time

$$O(n \log n + m).$$

To store and generate a Hadamard matrix is expensive, we consider a cheaper and space-efficient way to generate an FFT matrix via circulant transform.

Definition 2.10 (Circulant matrix)**.** A circulant matrix is an $n \times n$ matrix, where $n \in \mathbb{N}$, whose row vectors consist of the same element, and compared to the preceding row vector, each row vector is rotated one element to the right.

Definition 2.11 (Subsampled randomized circulant transform (SRCT))**.** Let $x \in \mathbb{R}^n$ be a random vector, whose elements are i.i.d. Rademacher random variables.

Also, let $P \in \mathbb{R}^{m \times n}$ be a random matrix in which each row contains a 1 at a uniformly distributed coordinate and zeros elsewhere.

Let $G \in \mathbb{R}^{n \times n}$ be a circulant matrix (see Definition 2.10) generated by x and $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.

Then, the subsampled randomized circulant transform is defined as follows: $S := \frac{1}{\sqrt{m}} PGD$.

Definition 2.12 (Tensor subsampled randomized circulant transform (TensorSRCT))**.** Let $x \in \mathbb{R}^n$ be a random vector, whose elements are i.i.d. Rademacher random variables.

Also, let $P \in \mathbb{R}^{m \times n^2}$ be a random matrix in which each row contains a 1 at a uniformly distributed coordinate and zeros elsewhere.

Let $G \in \mathbb{R}^{n \times n}$ be a circulant matrix (see Definition 2.10) generated by x.

Let $D_1 \in \mathbb{R}^{n \times n}$ and $D_2 \in \mathbb{R}^{n \times n}$ be two independent diagonal matrices whose diagonal elements are i.i.d. Rademacher random variables.

Then, the tensor circulant transform $T : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ is defined as follows: $T := P \cdot (GD_1 \otimes GD_2)$.

Remark 2.13. Similar to SRHT, we can utilize the fast Fourier transform with circulant matrix. SRCT can be applied to a vector of length n in $O(n \log n)$ time, and TensorSRCT can be applied to $x \otimes y$ in $O(n \log n + m)$ time.

2.3. OSE Property of Dense Sketches

An important condition for sketch-and-solve regressions is OSE. We focus particularly on SRHT, SRCT, and their tensor variants.

Lemma 2.14 ((Ahle et al., 2020; Song et al., 2021a) , see for example, Lemma 2.11 in (Song et al., 2021a))**.** *Let S be an* SRHT *matrix defined in Definition 2.6. If*

$$m = O(\epsilon^{-2} d \log^2(nd/\delta)),$$

*then S is an (ϵ, δ, d, n)-*OSE*.*

Lemma 2.15 ((Ahle et al., 2020; Song et al., 2021a) , see for example, Lemma 2.12 in (Song et al., 2021a))**.** *Let S be a* TensorSRHT *matrix defined in Definition 2.8. If*

$$m = O(\epsilon^{-2} d \log^3(nd/\epsilon\delta)),$$

*then S is an $(\epsilon, \delta, d, n^2)$-*OSE *for degree-2 tensors.*

SRCT requires more row count than SRHT due to the Gram $G^\top G$ is only I_n in expectation.

Lemma 2.16 (Informal version of Corollary C.7)**.** *Let S be an* SRCT *matrix defined in Definition 2.11. If*

$$m = O(\epsilon^{-2} d^2 \log^2(nd/\delta)),$$

then S is an (ϵ, δ, d, n)-OSE.

Lemma 2.17 (Informal version of Corollary C.8)**.** *Let S be an TensorSRCT matrix defined in Definition 2.12. If*

$$m = O(\epsilon^{-2} d^2 \log^3(nd/\delta)),$$

then S is an $(\epsilon, \delta, d, n^2)$-OSE.

2.4. Probability Tools

Lemma 2.18 (Khintchine's inequality (Khintchine, 1923))**.** *Let $\sigma_1, \cdots, \sigma_n$ be i.i.d. Rademacher random variables and $z_1, \cdots, z_n$ be real numbers. Then, there exists constants $C, C' > 0$ such that*

$$\Pr[|\sum_{i=1}^n z_i \sigma_i| \geq Ct\|z\|_2] \leq \exp(-C't^2).$$

Lemma 2.19 (Hoeffding bound, (Hoeffding, 1963))**.** *Let $Z_1, \cdots, Z_n$ be independent, zero-mean random variables with $Z_i \in [\alpha_i, \beta_i]$. Then,*

$$\Pr[|\sum_{i=1}^n Z_i| > t] \leq 2\exp(-\frac{t^2}{2\sum_{i=1}^n (\beta_i - \alpha_i)^2}).$$

Lemma 2.20 (Lemma 1 on page 1325 of Laurent and Massart (Laurent & Massart, 2000))**.** *Let $X \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero means and σ^2 variance. Then,*

$$\Pr[X - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t)$$
$$\Pr[k\sigma^2 - X \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t)$$

Lemma 2.21 (Hanson-Wright inequality (Hanson & Wright, 1971))**.** *Let $x \in \mathbb{R}^n$ denote a random vector with independent entries x_i with $\mathbb{E}[x_i] = 0$ and $|x_i| \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\begin{aligned} & \Pr[|x^\top A x - \mathbb{E}[x^\top A x]| > t] \\ \leq & 2 \cdot \exp(-c \min\{t^2/(K^4\|A\|_F^2), t/(K^2\|A\|)\}) \end{aligned}$$

Lemma 2.22 (Matrix Chernoff bound, Theorem 2.2 in (Tropp, 2011))**.** *Let $\mathcal{X}$ be a finite set of positive-semidefinite matrices with dimension $d \times d$. Suppose that $\max_{X \in \mathcal{X}} \lambda_{\max}(X) \leq B$. Sample $\{X_1, \cdots, X_n\}$ uniformly at random from $\mathcal{X}$ without replacement. We define $\mu_{\min}$ and $\mu_{\max}$ as follows: $\mu_{\min} := n \cdot \lambda_{\min}(\mathbb{E}[X_1])$ and $\mu_{\max} := n \cdot \lambda_{\max}(\mathbb{E}[X_1])$. Then,*

$$\begin{aligned} & \Pr[\lambda_{\min}(\sum_{i=1}^n X_i) \leq (1-\delta)\mu_{\min}] \\ \leq & d \cdot \exp(-\delta^2 \mu_{\min}/B) \end{aligned}$$

for $\delta \in [0, 1)$,

$$\begin{aligned} & \Pr[\lambda_{\max}(\sum_{i=1}^n X_i) \leq (1+\delta)\mu_{\max}] \\ \leq & d \cdot \exp(-\delta^2 \mu_{\max}/(4B)) \end{aligned}$$

for $\delta \geq 0$.

3. ℓ_∞ Guarantee via OCE

In this section, we present a sufficient condition for a sketching matrix to give good ℓ_∞ guarantee: given a pair of fixed vectors g, h such that $g^\top h = 0$, if the sketching matrix approximately preserves the inner product with high probability, then it gives good ℓ_∞ guarantee for regression.

Lemma 3.1 (Core lemma)**.** *Let $A \in \mathbb{R}^{n \times d}$ be a fixed matrix. Let $U \in \mathbb{R}^{n \times d}$ denote the orthonormal basis of A. Let $S \in \mathbb{R}^{m \times n}$ be a sketching matrix that satisfies two properties*

- *S is an $(0.1, \delta, d, n)$-OSE (with $\delta \in (0, 0.1)$, Definition 2.1)*
- *S is an (β, δ, n)-OCE (with $\beta \geq 1$ and $\delta \in (0, 0.1)$, Definition 2.3)*

For any fixed vectors $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^n$ with $U^\top b = 0$, we have

$$|a^\top (SA)^\dagger Sb| \lesssim \frac{\beta}{\sqrt{m}} \cdot \|a\|_2 \cdot \|b\|_2 \cdot \|\Sigma^{-1}\|$$

holds with probability at least $1 - \delta$.

Proof. With probability 1, the matrix $SA \in \mathbb{R}^{m \times d}$ has linearly independent columns.

Therefore, $(SA)^\dagger \in \mathbb{R}^{d \times m}$ is

$$\begin{aligned} & (SA)^\dagger \\ = & (A^\top S^\top SA)^{-1} A^\top S^\top \\ = & (V\Sigma U^\top S^\top SU\Sigma V^\top)^{-1} V\Sigma U^\top S^\top \\ = & (V^\top)^{-1}\Sigma^{-1}(U^\top S^\top SU)^{-1}\Sigma^{-1}V^{-1}V\Sigma U^\top S^\top \\ = & V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top S^\top, \end{aligned}$$

where the first step follows from $SA \in \mathbb{R}^{m \times d}$ has full rank, the second step follows from SVD on $A \in \mathbb{R}^{n \times d}$, the third step follows from $(AB)^{-1} = B^{-1}A^{-1}$, and the last step follows from the fact that V is orthogonal based on the property of SVD.

For convenience, we define x as follows:

$$x := a^\top V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top S^\top Sb.$$

In the next few paragraphs, we will explain how to upper bound $|x|$ with high probability.

Since S is a $(0.1, \delta, d, n)$-OSE (Definition 2.1), we know

$$\Pr[\|I - U^\top S^\top SU\| \leq 0.1] \geq 1 - \delta.$$

We condition on this event. It follows that

$$\begin{aligned}
&\|V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top\| \\
=& \|\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top\| \\
\leq& \|\Sigma^{-1}\|\|(U^\top S^\top SU)^{-1}\|\|U^\top\| \\
\leq& \|\Sigma^{-1}\| \cdot \frac{1}{1-0.1} \cdot 1 \\
=& O(\|\Sigma^{-1}\|),
\end{aligned}$$

where the first step follows from that V is a rotation, the second step follows from sub-multiplicativity, and the third step follows from $\|I - U^\top S^\top SU\| \leq 0.1$ and that U is a rotation.

Hence, we have

$$\begin{aligned}
&\Pr[\|a^\top V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top\|_2 \\
=& O(\|\Sigma^{-1}\| \cdot \|a\|_2)] \\
\geq& 1-\delta. \qquad (3)
\end{aligned}$$

Let us define a vector $u \in \mathbb{R}^n$

$$u := U(U^\top S^\top SU)^{-1}\Sigma^{-1}V^\top a,$$

By the definition of OCE (Definition 2.3), we have that

$$\begin{aligned}
&\Pr[|u^\top S^\top Sb - u^\top b| \leq \frac{\beta}{\sqrt{m}} \cdot \|u\|_2\|b\|_2] \\
\geq& 1-\delta,
\end{aligned}$$

where $U^\top b = 0$ gives us $u^\top b = 0$ and $u^\top S^\top Sb = x$.

Thus, the above bound translates to

$$\begin{aligned}
&\Pr[|x| \leq C \cdot \frac{\beta}{\sqrt{m}} \cdot \|\Sigma^{-1}\|\|a\|_2\|b\|_2] \\
\geq& 1-\delta \qquad (4)
\end{aligned}$$

as desired. □

Remark 3.2. We note that the above argument relies on applying the sketching in a Johnson-Lindenstrauss fashion, so that we can save a factor of $\frac{1}{\sqrt{m}}$ in the ℓ_∞ bound. An alternative approach is to adapt a Johnson-Lindenstrauss transform that is deterministic (Dadush et al., 2018). We also observe the guarantee is essentially tight up to polylogarithmic factors due to the tightness of Johnson-Lindenstrauss (Alon & Klartag, 2017; Larsen & Nelson, 2017).

We are now ready to prove the ℓ_∞ guarantee given the inner product bound of Lemma 3.1.

Theorem 3.3. *Suppose $A \in \mathbb{R}^{n\times d}$ has full column rank and $b \in \mathbb{R}^n$. Let $S \in \mathbb{R}^{m\times n}$ be a sketching matrix satisfying conditions in Lemma 3.1. For any fixed vector $a \in \mathbb{R}^d$, we have*

$$\begin{aligned}
&|\langle a, x^*\rangle - \langle a, x'\rangle| \\
\lesssim& \frac{\epsilon}{\sqrt{d}} \cdot \|a\|_2 \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|,
\end{aligned}$$

holds with probability at least $1-\delta$, where

- $x^* = \arg\min_{x\in\mathbb{R}^d} \|Ax - b\|_2$,
- $x' = \arg\min_{x\in\mathbb{R}^d} \|SAx - Sb\|_2$.

Proof. Since A has full column rank, we have that $x^* = A^\dagger b$. Similarly, SA has full column rank with probability 1, therefore $x' = (SA)^\dagger Sb$ and $(SA)^\dagger SA = I$. Thus, we have

$$\begin{aligned}
&|\langle a, x^*\rangle - \langle a, x'\rangle| \\
=& |\langle a, x^* - (SA)^\dagger Sb\rangle| \\
=& |\langle a, (SA)^\dagger S(Ax^* - b)\rangle| \\
=& |\langle a, (SA)^\dagger S(AA^\dagger b - b)| \\
=& |\langle ((SA)^\dagger S)^\top a, (I - UU^\top)b\rangle| \qquad (5)
\end{aligned}$$

where $U \in \mathbb{R}^{n\times d}$ is an orthonormal basis for A. It is well-known that $I - UU^\top = U_\perp U_\perp^\top$ where $U_\perp \in \mathbb{R}^{n\times(n-d)}$ is the orthonormal basis for the orthogonal component of $\mathrm{span}(A)$. To maximize the above expression, we shall let $b \in \mathrm{span}(U_\perp)$ or equivalently, $U^\top b = 0$. Thus, bounding Eq. (5) is equivalent to consider

$$|a^\top (SA)^\dagger Sb| \lesssim \frac{\beta}{\sqrt{m}} \cdot \|a\|_2 \cdot \|b\|_2 \cdot \|A^\dagger\|,$$

the inequality holds with probability at least $1-2\delta$ by Lemma 3.1. Finally, note that since $U^\top b = 0$, we have that $Ax^* = 0$ and we have proved

$$\begin{aligned}
&|\langle a, x^*\rangle - \langle a, x'\rangle| \\
\lesssim& \frac{\epsilon}{\sqrt{d}} \cdot \|a\|_2 \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|.
\end{aligned}$$

Thus, we complete the proof. □

Note that we only require the OSE with $\epsilon = O(1)$ and the ϵ dependence follows from the row count of OCE.

3.1. High Probability Bound for OCE

In this section, we provide a unified framework for proving the high probability bound of OCE. Our analysis utilizes the three dense sketching matrices that can all be designed as first picking a set of fresh random signs, then picking the sketching matrix according to the distribution.

We state the key assumptions on dense sketching matrices that are sufficient for OCE property.

Assumption 3.4. *Let $S \in \mathbb{R}^{m \times n}$ be a dense sketching matrix satisfying the following two assumptions:*

- *Pairwise inner product bound:*

$$\Pr[\max_{i \neq j} |\langle S_{*,i}, S_{*,j} \rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \geq 1 - \delta.$$

- *Column norm bound:*

$$\Pr[|\|S_{*,i}\|_2^2 - 1| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \geq 1 - \delta,$$

for all $i \in [n]$.

Lemma 3.5. *Let $S \in \mathbb{R}^{m \times n}$ be a dense sketching matrix meets Assumption 3.4. Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ be two fixed vectors. Then, the following properties hold:*

$$|(g^\top S^\top S h) - (g^\top h)| \leq \frac{\log^{1.5}(n/\delta)}{\sqrt{m}} \|g\|_2 \|h\|_2$$

holds with probability at least $1 - \delta$.

Proof. We can rewrite $(g^\top S^\top S h) - (g^\top h)$ as follows:

$$\begin{aligned}
& (g^\top S^\top S h) - (g^\top h) \\
= & \sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle S_{*,i}, S_{*,j} \rangle + \sum_{i=1}^n g_i h_i (\|S_{*,i}\|_2^2 - 1) \\
= & \underbrace{\sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle \sigma_i \overline{S}_{*,i}, \sigma_j \overline{S}_{*,j} \rangle}_{\texttt{off-diag}} \\
+ & \underbrace{\sum_{i=1}^n g_i h_i (\|S_{*,i}\|_2^2 - 1)}_{\texttt{diag}},
\end{aligned}$$

where the first step follows from the fact that σ_i's are independent Rademacher random variables and $S_{*,i} = \sigma_i \overline{S}_{*,i}$, $\forall i \in [n]$, the second step follows from separating diagonal and off-diagonal terms.

We will focus on bounding the quantity `off-diag`, as `diag` can be handled in a rather simple fashion.

We define matrix $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ as follows:

$$\begin{aligned}
A_{i,j} &:= g_i h_j \cdot \langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle, && \forall i \in [n], j \in [n] \\
B_{i,j} &:= g_i h_j \cdot \max_{i' \neq j'} |\langle \overline{S}_{*,i'}, \overline{S}_{*,j'} \rangle|, && \forall i \in [n], j \in [n].
\end{aligned}$$

We define $A^\circ \in \mathbb{R}^{n \times n}$ to be the matrix $A \in \mathbb{R}^{n \times n}$ with removing diagonal entries.

By applying Hanson-Wright inequality (Lemma 2.21), we have

$$\begin{aligned}
& \Pr[|\sigma^\top A^\circ \sigma| > \tau] \\
\leq & 2 \cdot \exp\left(-c \cdot \min\{\tau^2 / \|A^\circ\|_F^2, \tau / \|A^\circ\|\}\right)
\end{aligned}$$

We can upper bound $\|A^\circ\|$ and $\|A^\circ\|_F$.

$$\begin{aligned}
& \|A^\circ\| \\
\leq & \|A^\circ\|_F \\
\leq & \|A\|_F \\
\leq & \|B\|_F \\
\leq & \|g\|_2 \cdot \|h\|_2 \cdot \max_{i \neq j} |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle|,
\end{aligned}$$

where the first step follows from $\|\cdot\| \leq \|\cdot\|_F$, the second step follows from the definition of A°, the third step follows from the definition of A and B, and the fourth step follows from B is rank 1 as $B = \max_{i \neq j} |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \cdot g h^\top$.

It remains to obtain a bound on $\max_{i \neq j} |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle|$. Note that for any column i, j,

$$\begin{aligned}
& |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \\
= & |\langle \sigma_i \overline{S}_{*,i}, \sigma_j \overline{S}_{*,j} \rangle| \\
= & |\langle S_{*,i}, S_{*,j} \rangle|,
\end{aligned}$$

where the first step follows from the fact that random signs do not change the magnitude of the inner product and the second step follows from the definition of $S_{*,i}$ and $S_{*,j}$.

Since S meets Assumption 3.4, we have that with probability at least $1 - \delta$,

$$\max_{i \neq j} |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}.$$

Conditioning on the above event holds, we have that

$$\begin{aligned}
& \Pr[|\texttt{off-diag}| > \tau] \\
\leq & 2 \cdot \exp(-c \cdot \frac{\tau}{\|g\|_2 \cdot \|h\|_2 \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}}),
\end{aligned}$$

choosing

$$\tau = \|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(n/\delta) / \sqrt{m},$$

we can show that

$$\Pr[|\texttt{off-diag}| \geq \|g\|_2 \cdot \|h\|_2 \frac{\log^{1.5}(n/\delta)}{\sqrt{m}}] \leq \Theta(\delta).$$

To bound the term `diag`, note that due to Assumption 3.4, we have with probability at least $1 - \delta$,

$$|\|S_{*,i}\|_2^2 - 1| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}.$$

Conditioning on this event, we have

$$\begin{aligned} & |\texttt{diag}| \\ \leq & \max_{i \in [n]} |\|S_{*,i}\|_2^2 - 1| \cdot |g^\top h| \\ \leq & \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}} \cdot \|g\|_2 \cdot \|h\|_2, \end{aligned}$$

where the last step is by Cauchy-Schwartz. Note that $|\texttt{diag}|$ is subsumed by $|\texttt{off-diag}|$.

Union bounding over all events, we have that

$$\Pr[|g^\top S^\top S h - g^\top h| \geq \frac{\log^{1.5}(n/\delta)}{\sqrt{m}} \cdot \|g\|_2 \cdot \|h\|_2] \leq \Theta(\delta).$$

This completes the proof. □

3.2. Inner Product Bound for SRHT and SRCT

We will show that SRHT and SRCT satisfy Assumption 3.4. Before proving the pairwise inner product bound, we state a general property to characterize these sketching matrices. This key property will be used in the later proof.

Definition 3.6 (Sign structure)**.** For any sketching matrix, we say it has "Sign structure" if the following properties hold

- $S_{k,i} \in \{\pm \frac{1}{\sqrt{m}}\}$, for all $k \in [m], i \in [n]$.
- $S_{k,i}$ and $S_{k,j}$ are independent for any $i \neq j$.
- $\mathbb{E}[S_{k,i}] = 0$ for all $k \in [m]$ and $i \in [n]$.

Lemma 3.7. *Both* SRHT *and* SRCT *satisfy Definition 3.6.*

Proof. It follows from the definitions of two sketching matrices directly. □

Lemma 3.8 (SRHT and SRCT)**.** *Let $S \in \mathbb{R}^{m \times n}$ be any sketching matrices that satisfy the Definition 3.6. Then, we have*

$$\Pr[\max_{i \neq j} |\langle S_{*,i}, S_{*,j} \rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \Theta(\delta).$$

Proof. Fix a pair of indices $i \neq j$ and we define $X \in \mathbb{R}^{m \times 2}$ as follows:

$$X := \begin{bmatrix} S_{*,i} & S_{*,j} \end{bmatrix}$$

The Gram matrix is

$$X^\top X = \sum_{k=1}^m G_k,$$

where

$$\begin{aligned} & G_k \\ = & \begin{bmatrix} S_{k,i} & S_{k,j} \end{bmatrix}^\top \begin{bmatrix} S_{k,i} & S_{k,j} \end{bmatrix} \\ = & \begin{bmatrix} S_{k,i} \\ S_{k,j} \end{bmatrix} \begin{bmatrix} S_{k,i} & S_{k,j} \end{bmatrix} \\ = & \begin{bmatrix} S_{k,i}^2 & S_{k,i} S_{k,j} \\ S_{k,i} S_{k,j} & S_{k,j}^2 \end{bmatrix} \\ = & \begin{bmatrix} \frac{1}{m} & S_{k,i} S_{k,j} \\ S_{k,i} S_{k,j} & \frac{1}{m} \end{bmatrix}. \end{aligned}$$

where the first step follows from the definition of G_k, the second step follows from rewriting $\begin{bmatrix} S_{k,i} & S_{k,j} \end{bmatrix}^\top$, the third step follows from the definition of matrix multiplication, and the last step follows from $S_{k,i}^2 = 1/m$ and $S_{k,j}^2 = 1/m$.

Note that G_k has eigenvalues 0 and $\frac{2}{m}$, i.e.,

$$\lambda_1(G_k) = 2/m, \quad \lambda_2(G_k) = 0.$$

Since $S_{k,i}$ and $S_{k,j}$ are independent Rademacher random variables, we have

$$\begin{aligned} & \mathbb{E}[S_{k,i} S_{k,j}] \\ = & \mathbb{E}[S_{k,i}] \cdot \mathbb{E}[S_{k,j}] \\ = & 0. \end{aligned}$$

Thus, we know

$$\mathbb{E}[G_k] = \begin{bmatrix} 1/m & 0 \\ 0 & 1/m \end{bmatrix}. \tag{6}$$

Consequently, we have

$$\begin{aligned} & \mathbb{E}[X^\top X] \\ = & \mathbb{E}[\sum_{k=1}^m G_k] \\ = & m \cdot \mathbb{E}[G_k] \\ = & m \cdot \begin{bmatrix} 1/m & 0 \\ 0 & 1/m \end{bmatrix} \\ = & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned}$$

where the first step follows from the definition of $X^\top X$, the second step follows from the fact that $\mathbb{E}[ca] = c\,\mathbb{E}[a]$ for a constant c, the third step follows from Eq. (6), and the last step follows from simple algebra.

Let $\lambda_i(X^\top X)$ be the i-th eigenvalue of $X^\top X \in \mathbb{R}^{2 \times 2}$. By matrix Chernoff bound (Lemma 2.22 with $B = 2/m$), for any $t > 0$, we have

$$\Pr[\forall i \in [2], |\lambda_i(X^\top X) - 1| \geq t]$$

$$\leq 4\exp(-t^2 m/2).$$

This means with probability at least $1 - 4\exp(-t^2 m/2)$, the eigenvalues of $X^\top X$ are between $[1-t, 1+t]$ and consequently, the eigenvalues of $X^\top X - I_2$ are between $[-t, t]$.

Let us choose

$$t = O(\frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}),$$

we have

$$\Pr[\|X^\top X - I_2\| \geq C \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \frac{\delta}{n^2}.$$

The proof can be wrapped up by noting that

$$X^\top X - I_2 = \begin{bmatrix} 0 & \langle S_{*,i}, S_{*,j} \rangle \\ \langle S_{*,i}, S_{*,j} \rangle & 0 \end{bmatrix},$$

the spectral norm of this matrix is $|\langle S_{*,i}, S_{*,j} \rangle|$ and union bound over all n^2 pairs of columns, we have

$$\Pr[\max_{i \neq j} |\langle S_{*,i}, S_{*,j}| \geq C \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \delta.$$

This completes the proof. □

3.3. Column Norm Bound for SRHT and SRCT

In this section, we prove the column norm bound for SRHT and SRCT. In particular, their columns are unit vectors. In Appendix D, we prove for random Gaussian matrix, the squared column norm is χ^2_m random vriable that concentrates around 1 with high probability.

Lemma 3.9 (SRHT and SRCT)**.** *Let $S \in \mathbb{R}^{m \times n}$ be an* SRHT *matrix or* SRCT *matrix.*

Then, for any $i \in [n]$, we have

$$\|S_{*,i}\|_2^2 = 1.$$

Proof. The proof directly follows from the definition.

For SRHT, recall $S = PHD$, the column norm of H is $\sqrt{n}$, and D is a random sign that does not change the norm. The matrix P subsamples m rows and rescale each entry by $\sqrt{\frac{1}{m}}$. The (squared) column norm is then 1.

For SRCT, the column norm of G is $\sqrt{n}$ as well. Thus, by the same argument, SRCT has unit column vectors. □

4. Put Things Together

Now, we're ready to present the proof for Theorem 1.1.

Proof of Theorem 1.1. Using Lemma 2.14 (it shows SRHT gives OSE), we know if

$$m \geq d\log^2(n/\delta),$$

it gives $(O(1), \delta, n, d)$-OSE.

Using Lemma 3.8 (it shows SRHT gives OCE), we know

$$\beta = O(\log^{1.5}(n/\delta)).$$

Using Lemma 3.1 (it shows OSE + OCE implies our result), we need to choose

$$m \geq \epsilon^{-2} d\beta^2 \geq \epsilon^{-2} d\log^3(n/\delta).$$

Combining the above relations together, we have

$$\begin{aligned} m &\geq d\log^2(n/\delta) + \epsilon^{-2} d\log^3(n/\delta) \\ &\geq \epsilon^{-2} d\log^3(n/\delta). \end{aligned}$$

This completes the proof. □

5. Conclusion

In this paper, we study the sketching-based regression algorithm with an ℓ_∞ guarantee. We show that SRHT with $m = \epsilon^{-2} d\log^3(n/\delta)$ rows provides the desired ℓ_∞ guarantee solution, improving upon the $\epsilon^{-2} d^{1+\gamma}$ rows for $\gamma = \sqrt{\frac{\log\log n}{\log d}}$ of (Price et al., 2017). This is nearly-optimal up to logarithmic factors. Our proof adapts the oblivious coordinate-wise embedding property introduced in (Song & Yu, 2021) in a novel way. We also greatly extends the reach of ℓ_∞ guarantee to degree-2 Kronecker product regression via TensorSRHT matrix.

In addition, we introduce the SRCT and TensorSRCT matrices. These matrices can be applied in a fashion similar to SRHT, and they have similar OCE behaviors as SRHT.

Our result provides an elegant way to integrate fast, sketching-based regression solver for optimization process, in particular second-order methods. The regression problem per iteration can be solved in time nearly-linear in the input size, and the ℓ_∞ guarantee comes in handy when analyzing convergence with approximate step. It also gives improved generalization bound on approximate regression via SRHT (Price et al., 2017).

Acknowledgement

Lichen Zhang is supported by NSF grant No. 1955217 and NSF grant No. 2022448.

References

Ahle, T. D., Kapralov, M., Knudsen, J. B., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 141–160. SIAM, 2020.

Alon, N. and Klartag, B. Optimal compression of approximate inner products and dimension reduction. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, 2017.

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 20–29, 1996.

Avron, H., Nguyen, H., and Woodruff, D. Subspace embeddings for the polynomial kernel. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2258–2266, 2014.

Brand, J. v. d. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 259–278. SIAM, 2020.

Brand, J. v. d., Peng, B., Song, Z., and Weinstein, O. Training (overparametrized) neural networks in near-linear time. In *ITCS*, 2021.

Cai, T., Gao, R., Hou, J., Chen, S., Wang, D., He, D., Zhang, Z., and Wang, L. Gram-gauss-newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pp. 693–703. Springer, 2002.

Cherapanamjeri, Y., Silwal, S., Woodruff, D. P., and Zhou, S. Optimal algorithms for linear algebra in the current matrix multiplication time. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2023.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference (STOC)*, pp. 81–90, 2013.

Cohen, M. B. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 278–287. SIAM, 2016.

Cohen, M. B., Lee, Y. T., and Song, Z. Solving linear programs in the current matrix multiplication time. In *STOC*, 2019.

Dadush, D., Guzmán, C., and Olver, N. Fast, deterministic and sparse dimensionality reduction. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, 2018.

Diao, H., Song, Z., Sun, W., and Woodruff, D. Sketching for kronecker product regression and p-splines. In *International Conference on Artificial Intelligence and Statistics*, pp. 1299–1308. PMLR, 2018.

Diao, H., Jayaram, R., Song, Z., Sun, W., and Woodruff, D. Optimal sketching for kronecker product regression and low rank approximation. *Advances in neural information processing systems*, 32, 2019.

Fahrbach, M., Fu, T., and Ghadiri, M. Subquadratic kronecker regression with applications to tensor decomposition. In *Thirty-Sixth Conference on Neural Information Processing Systems*, NeurIPS'22, 2022.

Foss, S., Korshunov, D., and Zachary, S. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer series in operations research. Springer, New York, NY, 1st ed edition, 2011.

Freksen, C. B. and Larsen, K. G. On using Toeplitz and circulant matrices for Johnson-Lindenstrauss transforms. *Algorithmica*, 2020.

Gu, Y. and Song, Z. A faster small treewidth sdp solver. *arXiv preprint arXiv:2211.06033*, 2022.

Hanson, D. L. and Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *Journal of the American Statistical*, pp. 13–30, 1963.

Jiang, S., Song, Z., Weinstein, O., and Zhang, H. A faster algorithm for solving general lps. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 823–832, 2021.

Khintchine, A. Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

Larsen, K. G. and Nelson, J. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 633–638. IEEE, 2017.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Lee, Y. T. and Sidford, A. Path finding methods for linear programming: Solving linear programs in $O(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 424–433, 2014.

Lee, Y. T., Song, Z., and Zhang, Q. Solving empirical risk minimization in the current matrix multiplication time. In *Conference on Learning Theory*, pp. 2140–2157. PMLR, 2019.

Li, X., Haupt, J., and Woodruff, D. Near optimal sketching of low-rank tensor regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Lu, Y., Dhillon, P., Foster, D. P., and Ungar, L. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems (NIPS)*, pp. 369–377, 2013.

Meng, X. and Mahoney, M. W. 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)*, pp. 91–100, 2013.

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

Price, E., Song, Z., and Woodruff, D. P. Fast regression with an ℓ_∞ guarantee. In *ICALP*, 2017.

Qin, L., Song, Z., Zhang, L., and Zhuo, D. An online and unified algorithm for projection matrix vector multiplication with application to empirical risk minimization. In *AISTATS*, 2023.

Reddy, A., Song, Z., and Zhang, L. Dynamic tensor product regression. In *Thirty-Sixth Conference on Neural Information Processing Systems*, NeurIPS'22, 2022.

Sarlos, T. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS)*, pp. 143–152. IEEE, 2006.

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

Song, Z., Woodruff, D., and Zhang, H. Sublinear time orthogonal tensor decomposition. *Advances in Neural Information Processing Systems*, 29, 2016.

Song, Z., Woodruff, D. P., and Zhong, P. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2772–2789. SIAM, 2019.

Song, Z., Woodruff, D., Yu, Z., and Zhang, L. Fast sketching of polynomial kernels of polynomial degree. In *International Conference on Machine Learning*, pp. 9812–9823. PMLR, 2021a.

Song, Z., Zhang, L., and Zhang, R. Training multi-layer over-parametrized neural network in subquadratic time. *arXiv preprint arXiv:2112.07628*, 2021b.

Song, Z., Xu, Z., Yang, Y., and Zhang, L. Accelerating frank-wolfe algorithm using low-dimensional and adaptive data structures. *arXiv preprint arXiv:2207.09002*, 2022a.

Song, Z., Xu, Z., and Zhang, L. Speeding up sparsification using inner product search data structures. *arXiv preprint arXiv:2204.03209*, 2022b.

Song, Z., Yang, X., Yang, Y., and Zhang, L. Sketching meets differential privacy: fast algorithm for dynamic kronecker projection maintenance. In *International Conference on Machine Learning*, 2023.

Tropp, J. A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 2011.

Woodruff, D. and Zandieh, A. Leverage score sampling for tensor product matrices in input sparsity time. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Woodruff, D. P. and Zandieh, A. Near input sparsity time kernel embeddings via adaptive sampling. In *ICML*, 2020.

Appendix

Roadmap. In Section A, we introduce the fundamental definitions and properties that we will use in Appendix. In Section B, we analyze and develop the ℓ_∞ guarantee of Kronecker product regressions. In Section C, we introduce the Strong JL Moment Property and prove that both Circulant Transform and Tensor Circulant Transform satisfy this. In Section D, we focus on studying AMS, random Gaussian, and SRHT and show that the inner product is bounded on any pair of different columns of AMS, random Gaussian, and SRHT–dense sketching matrices.

A. Tools for Matrices and Probability

For matrix $A_1 \in \mathbb{R}^{n_1 \times d_1}$ and $A_2 \in \mathbb{R}^{n_2 \times d_2}$, we use $A_1 \otimes A_2 \in \mathbb{R}^{n_1 n_2 \times d_1 d_2}$ to denote the matrix that $(i_1 - 1) \cdot (n_2) + i_2, (j_1 - 1)d_2 + j_2$ -th entry is $(A_1)_{i_1,j_1} \cdot (A_2)_{i_2,j_2}$.

Lemma A.1 (Markov's inequality)**.** *If X is a non-negative random variable and $a > 0$. Then we have*

$$\Pr[X \geq a] \leq \mathbb{E}[X]/a.$$

Definition A.2 (Sub-exponential distribution ((Foss et al., 2011)))**.** We say $X \in \mathsf{SubExp}(\sigma^2, \alpha)$ with parameters $\sigma > 0$, $\alpha > 0$ if:

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\lambda^2 \sigma^2 / 2), \forall |\lambda| < 1/\alpha.$$

Lemma A.3 (Tail bound for sub-exponential distribution ((Foss et al., 2011)))**.** *Let $X \in \mathsf{SubExp}(\sigma^2, \alpha)$ and $\mathbb{E}[X] = \mu$. Then,*

$$\Pr[|X - \mu| \geq t] \leq \exp(-0.5 \min\{t^2/\sigma^2, t/\alpha\}).$$

Claim A.4. *For every matrix $A \in \mathbb{R}^{n_1 \times n_2}, B \in \mathbb{R}^{n_2 \times n_3}, C \in \mathbb{R}^{d_1 \times d_2}, D \in \mathbb{R}^{d_2 \times d_3}$*

$$(A \cdot B) \otimes (C \cdot D) = (A \otimes C) \cdot (B \otimes D).$$

B. Kronecker Product Regression with ℓ_∞ Guarantee

In this section, we study the ℓ_∞ guarantee of Kronecker product regressions. Given two matrices $A_1, A_2 \in \mathbb{R}^{n \times d}$ and a label vector $b \in \mathbb{R}^{n^2}$, the goal is to solve the regression $\arg\min_{x \in \mathbb{R}^{d^2}} \|(A_1 \otimes A_2)x - b\|_2^2$. This problem can be easily generalized to product of q matrices and fast, input-sparsity time algorithms have been studied in a line of works (Li et al., 2017; Diao et al., 2018; Song et al., 2019; Diao et al., 2019; Fahrbach et al., 2022; Reddy et al., 2022).

B.1. Main result

Theorem B.1 (Tensor version of Theorem 1.1)**.** *Suppose $n \leq \exp(d)$ and matrix $A \in \mathbb{R}^{n^2 \times d^2}$ and vector $b \in \mathbb{R}^{n^2}$ are given, where $A = A_1 \otimes A_2$ for matrices $A_1, A_2 \in \mathbb{R}^{n \times d}$ and $b = b_1 \otimes b_2$ for vectors $b_1, b_2 \in \mathbb{R}^n$. Let $S \in \mathbb{R}^{m \times n^2}$ be a*

- *tensor subsampled randomized Hadamard transform matrix (*$\mathsf{TensorSRHT}$*) with $m = \Theta(\epsilon^{-2} d^2 \log^3(n/\delta))$ rows or*
- *tensor subsampled randomized circulant transform matrix (*$\mathsf{TensorSRCT}$*) with $m = \Theta(\epsilon^{-2} d^4 \log^3(n/\delta))$ rows.*

For

$$x' = \arg\min_{x \in \mathbb{R}^{d^2}} \|SAx - Sb\|_2$$

and

$$x^* = \arg\min_{x \in \mathbb{R}^{d^2}} \|Ax - b\|_2,$$

and any fixed $a \in \mathbb{R}^{d^2}$,

$$|\langle a, x^* \rangle - \langle a, x' \rangle| \leq \frac{\epsilon}{d} \cdot \|a\|_2 \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|$$

with probability $1 - 1/\operatorname{poly}(d)$.

Proof. Recall that we require $(O(1), \delta, d, n)$-OSE and $\beta = O(\log^{1.5}(n/\delta))$-OCE for it to give ℓ_∞ guarantee.

For OCE, it follows from Lemma B.3.

For TensorSRHT's OSE, it follows from Lemma 2.15 and for TensorSRCT, it follows from Corollary C.8. □

Remark B.2. The slightly different guarantee follows from the small dimension becomes d^2 instead of d. Let us discuss the utility of using these sketching matrices for solving the regression. As discussed in Def. 2.8 and 2.12, each column of $A_1 \otimes A_2$ can be computed in $O(n \log n + m)$ time instead of n^2, thus the total running time of applying S to A is $O(nd^2 \log n + \mathrm{poly}(d))$. Similarly, Sb can be applied in time $O(n \log n + \mathrm{poly}(d))$. The regression can then be solved in $\widetilde{O}(nd^2 + \mathrm{poly}(d))$ time. Prior works mainly focus on input-sparsity sketches (Diao et al., 2018), importance sampling (Diao et al., 2019), iterative method (Fahrbach et al., 2022) or more complicated sketches that scale well to q products and in dynamic setting (Reddy et al., 2022). To the best of our knowledge, this is the first ℓ_∞ guarantee for Kronecker product regression (with two matrices).

B.2. Oblivious Coordinate-wise Embedding for TensorSRHT and TensorSRCT

Lemma B.3 (TensorSRHT and TensorSRCT, Tensor version of Lemma 3.8)**.** *Let $S \in \mathbb{R}^{m \times n}$ be* TensorSRHT *or* TensorSRCT*. Then, S is an* OCE *with parameter $\beta = \log^{1.5}(n/\delta)$.*

Proof. To prove this result, we show that TensorSRHT and TensorSRCT satisfy Definition 3.6.

For TensorSRHT, recall $S = \frac{1}{\sqrt{m}} P(HD_1 \times HD_2)$, since H is Hadamard matrix and D_1, D_2 are just diagonal matrices with random signs. Thus, all entries of $HD_1 \times HD_2$ are also in $\{\pm 1\}$. As P is a row sampling matrix and we rescale each entry by $\frac{1}{\sqrt{m}}$. Thus, each entry of S is in $\{\pm \frac{1}{\sqrt{m}}\}$. For entries at the same row but two different columns i, j, if i is generated from two columns disjoint from j, then it's clear then are independent. Otherwise, suppose i is generated from columns a, b and j is generated from columns a, c with $b \neq c$. Then it is again independent, as the sign is completely determined by signs of b and c. Finally, we need to verify $\mathbb{E}[S_{k,i}] = 0$, this is trivially true since product of two random signs is still a random sign. For TensorSRCT, the argument is exactly the same.

Now that both of these matrices satisfy Definition 3.6, we can use Lemma 3.8 to give a bound on pairwise inner product. The column norm bound is automatically satisfied by definition. Thus, we can invoke Lemma 3.5 to wrap up the proof. □

C. SRCT and TensorSRCT: OSE via Strong JL Moment Property

In this section, we prove that both SRCT and TensorSRCT are OSE's. We prove this property via the strong JL moment property (Ahle et al., 2020). This gives a worse row count compared to that of SRHT and TensorSRHT. We believe that these two family of distributions should have similar row count for an OSE and leave it as a major open problem to close the gap between these two distributions.

C.1. Notations

To make the notation less heavy, we will use $\|X\|_{L^t}$ for the t-th moment of a random variable X. This is formally defined below.

Definition C.1 (t-th moment)**.** For every integer $t \geq 1$ and any random variable $X \in \mathbb{R}$, we write

$$\|X\|_{L^t} = \left(\mathbb{E}\left[|X|^t\right]\right)^{1/t}$$

Note that

$$\|X + Y\|_{L^t} \leq \|X\|_{L^t} + \|Y\|_{L^t}$$

for any random variables X, Y by the Minkowski inequality.

C.2. Strong JL Moment Property

We show that both SRCT (see Definition 2.11) and TensorSRCT (see Definition 2.12) satisfy the so-called *strong JL moment property*. Strong JL moment property is one of the core properties that can show the sketching matrix has subspace

embedding property (Sarlos, 2006).

Definition C.2 (Strong JL moment property (Ahle et al., 2020))**.** For every $\epsilon, \delta \in [0, 1]$, we say a distribution over random matrices $S \in \mathbb{R}^{m \times n}$ has the Strong (ϵ, δ)-JL Moment Property when

$$\| \|Sx\|_2^2 - 1 \|_{L^t} \leq \epsilon \sqrt{t / \log(1/\delta)}$$

and

$$\mathbb{E}\left[\|Sx\|_2^2 \right] = 1$$

for all $x \in \mathbb{R}^n$, $\|x\|_2 = 1$ and every integer $t \leq \log(1/\delta)$.

Given a distribution with strong JL moment property, it is well-known that such distribution provides OSE.

Lemma C.3 (Lemma 11 of (Ahle et al., 2020))**.** *Let $S \in \mathbb{R}^{m \times n}$ be a random matrix with $(\epsilon/d, \delta)$-strong JL moment property (Def. C.2). Then, S is also an (ϵ, δ, d, n)-OSE (Def. 2.1).*

To prove that SRCT (see Definition 2.11) and TensorSRCT (see Definition 2.12) satisfy the strong JL moment property, we will do this by proving that a more general class of matrices satisfies the strong JL moment property.

More precisely, let $k \in \mathbb{Z}_{>0}$ be a positive integer and

$$(D^{(i)})_{i \in [k]} \in \prod_{i \in [k]} \mathbb{R}^{n_i \times n_i}$$

be independent matrices, each with diagonal entries given by independent Rademacher variables.

Let $n = \prod_{i \in [k]} n_i$ and $P \in \{0, 1\}^{m \times n}$ be a random sampling matrix in which each row contains exactly one uniformly distributed nonzero element which has value one.

Then, we prove that the matrix

$$S = \frac{1}{\sqrt{m}} PG \cdot (D_1 \otimes \cdots \otimes D_k)$$

satisfies the strong JL moment property, where G is $n \times n$ circulant matrix (see Definition 2.10) generated by a random vector whose elements are Rademacher variables.

If $k = 1$, then S is just a SRCT (see Definition 2.11). If $k = 2$, then M is a TensorSRCT (see Definition 2.12).

In order to prove this result we need a couple of lemmas. The first lemma can be seen as a version of Khintchine's Inequality (see Lemma 2.18) for higher order chaos.

Lemma C.4 (Lemma 19 in (Ahle et al., 2020))**.** *Let $t \geq 1$ and $k \in \mathbb{Z}_{>0}$. Let $(\sigma^{(i)})_{i \in [k]} \in \prod_{i \in [k]} \mathbb{R}^{n_i}$ be independent vectors each satisfying the Khintchine's inequality (see Lemma 2.18):*

$$\| \langle \sigma^{(i)}, x \rangle \|_{L^t} \leq C_t \|x\|_2$$

for $t \geq 1$ and any vector $x \in \mathbb{R}^{d_i}$.

Let $(a_{i_1, \ldots, i_k})_{i_1 \in [n_j], \ldots, i_k \in [n_k]}$ be a tensor in $\mathbb{R}^{n_1 \times \cdots \times n_k}$. Then,

$$\left\| \sum_{i_1 \in [n_1], \ldots, i_k \in [n_k]} \left(\prod_{j \in [k]} \sigma_{i_j}^{(j)} \right) a_{i_1, \ldots, i_k} \right\|_{L^t} \leq C_t^k \left(\sum_{i_1 \in [n_1], \ldots, i_k \in [n_k]} a_{i_1, \ldots, i_k}^2 \right)^{\frac{1}{2}},$$

for $t \geq 1$.

Viewing $a \in \mathbb{R}^{n_1 \times \cdots \times n_k}$ as a vector, then

$$\| \langle \sigma^{(1)} \otimes \cdots \otimes \sigma^{(k)}, a \rangle \|_{L^t} \leq C_t^k \|a\|_2,$$

for $t \geq 1$.

Proof. The proof will be by induction on k.

Base case: For $k = 1$, the result is by the assumption that the vectors satisfy Khintchine's inequality.

Inductive case: Assume that the result is true for every value up to $k - 1$.

Let

$$B_{i_1,\ldots,i_{k-1}} = \sum_{i_k \in [d_k]} \sigma_{i_k}^{(k)} a_{i_1,\ldots,i_k}. \tag{7}$$

We then pull it out of the left hand term in the theorem:

$$\begin{aligned}
\| \sum_{i_1 \in [n_1],\ldots,i_k \in [n_k]} (\prod_{j \in [k]} \sigma_{i_j}^{(j)}) a_{i_1,\ldots,i_k} \|_{L^t} &= \| \sum_{i_1 \in [n_1],\ldots,i_{k-1} \in [n_{k-1}]} (\prod_{j \in [k-1]} \sigma_{i_j}^{(j)}) B_{i_1,\ldots,i_{k-1}} \|_{L^t} \\
&\leq C_t^{k-1} \| (\sum_{i_1 \in [d_1],\ldots,i_{k-1} \in [n_{k-1}]} B_{i_1,\ldots,i_{k-1}}^2)^{\frac{1}{2}} \|_{L^t} \\
&= C_t^{k-1} \| \sum_{i_1 \in [n_1],\ldots,i_{k-1} \in [n_{k-1}]} B_{i_1,\ldots,i_{k-1}}^2 \|_{L^{t/2}}^{\frac{1}{2}} \\
&\leq C_t^{k-1} (\sum_{i_1 \in [n_1],\ldots,i_{k-1} \in [n_{k-1}]} \| B_{i_1,\ldots,i_{k-1}}^2 \|_{L^{t/2}})^{\frac{1}{2}} \\
&= C_t^{k-1} (\sum_{i_1 \in [n_1],\ldots,i_{k-1} \in [n_{k-1}]} \| B_{i_1,\ldots,i_{k-1}} \|_{L^t}^2)^{\frac{1}{2}},
\end{aligned}$$

where the first step follows from Eq. (7), the second step follows from the inductive hypothesis, the third step follows from the definition of $\| \cdot \|$, the fourth step follows from the triangle inequality, the fifth step follows from the definition of $\| \cdot \|$.

It remains to bound

$$\| B_{i_1,\ldots,i_{k-1}} \|_{L^t}^2 \leq C_t^2 \sum_{i_k \in [n_k]} a_{i_1,\ldots,i_k}^2$$

by Khintchine's inequality, which finishes the induction step and hence the proof. □

The next lemma we will be using is a type of Rosenthal inequality based on first principles. It mixes large and small moments of random variables in an intricate way. For completeness, we include a proof here.

Lemma C.5 (Properties of random variables with t-moment, Lemma 20 in (Ahle et al., 2020))**.** *There exists a universal constant L, such that, for $t \geq 1$ if $X_1, \ldots, X_k$ are independent non-negative random variables with t-moment, then*

$$\| \sum_{i \in [k]} (X_i - \mathbb{E}[X_i]) \|_{L^t} \leq L \cdot \big(\sqrt{t} \cdot \| \max_{i \in [k]} X_i \|_{L^t}^{1/2} \cdot (\sum_{i \in [k]} \mathbb{E}[X_i])^{1/2} + t \cdot \| \max_{i \in [k]} X_i \|_{L^t} \big).$$

Proof. Throughout these calculations L_1, L_2 and L_3 will be universal constants.

$$\begin{aligned}
\| \sum_{i \in [k]} (X_i - \mathbb{E}[X_i]) \|_{L^t} &\leq L_1 \| \sum_{i \in [k]} \sigma_i X_i \|_{L^t} \\
&\leq L_2 \sqrt{t} \cdot \| \sum_{i \in [k]} X_i^2 \|_{L^{t/2}}^{1/2} \\
&\leq L_2 \sqrt{t} \cdot \| \max_{i \in [k]} X_i \cdot \sum_{i \in [k]} X_i \|_{L^{t/2}}^{1/2} \\
&\leq L_2 \sqrt{t} \cdot \| \max_{i \in [k]} X_i \|_{L^t}^{1/2} \cdot \| \sum_{i \in [k]} X_i \|_{L^t}^{1/2} \\
&\leq L_2 \sqrt{t} \cdot \| \max_{i \in [k]} X_i \|_{L^t}^{1/2} \cdot \Big((\sum_{i \in [k]} \mathbb{E}[X_i])^{1/2} + L_2 \| \sum_{i \in [k]} (X_i - \mathbb{E}[X_i]) \|_{L^t}^{1/2} \Big)
\end{aligned} \tag{8}$$

where the first step follows from symmetrization of X_i, the second step follows from Khintchine's inequality (see Lemma 2.18), the third step follows from Non-negativity of X_i, the fourth step follows from Cauchy-Schwartz inequality, and the last step follows from the triangle inequality.

Now, let A, B, C be defined as follows:

$$C := \| \sum_{i \in [k]} (X_i - \mathbb{E}[X_i]) \|_{L^t}^{1/2},$$

$$B := L_2 (\sum_{i \in [k]} \mathbb{E}[X_i])^{1/2},$$

and

$$A := \sqrt{t} \| \max_{i \in [k]} X_i \|_{L^t}^{1/2}.$$

Then, by rewriting Eq. (8), we have

$$C^2 \leq A(B + C).$$

This implies C is smaller than the largest of the roots of the quadratic.

Solving this quadratic inequality gives

$$C^2 \leq L_3 (AB + A^2),$$

which completes the proof. □

C.3. SRCT and TensorSRCT Satisfying Strong JL Moment Property

We can now prove that SRCT (see Definition 2.11) and TensorSRCT (see Definition 2.12) have the strong JL moment property.

Theorem C.6. *There exists a universal constant L, such that, the following holds.*

Let $k \in \mathbb{Z}_{>0}$. Let $(D^{(i)})_{i \in [k]} \in \prod_{i \in [k]} \mathbb{R}^{n_i \times n_i}$ be independent diagonal matrices with independent Rademacher variables.

We define $n := \prod_{i \in [k]} n_i$, $D := D_1 \otimes D_2 \otimes \cdots \otimes D_k \in \mathbb{R}^{n \times n}$ and $G := G_1 \otimes \ldots \otimes G_k \in \mathbb{R}^{n \times n}$, where each $G_i \in \mathbb{R}^{n_i \times n_i}$ is a circulant matrix generated by an independent Rademacher random vector. Let $P \in \mathbb{R}^{m \times n}$ be a row sampling matrix that has exactly one nonzero per row. Let $S := PGD$.

Let $x \in \mathbb{R}^n$ be any vector with $\|x\|_2 = 1$ and $t \geq 1$.

Then,

$$\left\| \frac{1}{m} \|PGDx\|_2^2 - 1 \right\|_{L^t} \leq L(\sqrt{\frac{tr^k}{m}} + \frac{tr^k}{m}),$$

where $r = \max\{t, \log m\}$.

There exists a universal constant $C_0 > 1$, such that, by setting

$$m = \Omega(\epsilon^{-2} \log(1/\delta) \cdot (C_0 \log(1/(\epsilon\delta)))^k),$$

we get that $\frac{1}{\sqrt{m}} PGD$ has (ϵ, δ)-strong JL moment property.

Proof. Throughout the proof C_1, C_2 and C_3 will denote universal constants.

For every $i \in [m]$, we let P_i be the random variable that says which coordinates the i-th row of P samples.

We define the random variable

$$Z_i := M_i x = G_{P_i} D x.$$

We note that since the variables $(P_i)_{i \in [m]}$ are independent, so the variables $(Z_i)_{i \in [m]}$ are conditionally independent given D, that is, if we fix D, then $(Z_i)_{i \in [m]}$ are independent.

Then, we could get the following inequality:

$$\begin{aligned}
& \| \frac{1}{m} \sum_{i \in [m]} Z_i^2 - 1 \|_{L^t} \\
= & \| (\mathbb{E}[(\frac{1}{m} \sum_{i \in [m]} Z_i^2 - 1) \mid D])^{1/t} \|_{L^t} \\
\leq & C_1 \| \frac{\sqrt{t}}{m} \cdot (\mathbb{E}[(\max_{i \in [m]} Z_i^2) \mid D])^{1/(2t)} \cdot (\sum_{i \in [m]} \mathbb{E}[Z_i^2 \mid D])^{1/2} + \frac{t}{m} \cdot (\mathbb{E}[(\max_{i \in [m]} Z_i^2)^t \mid D])^{1/t} \|_{L^t} \\
\leq & C_1 \frac{\sqrt{t}}{m} \cdot \| (\mathbb{E}[(\max_{i \in [m]} Z_i^2) \mid D])^{1/(2t)} \cdot (\sum_{i \in [m]} \mathbb{E}[Z_i^2 \mid D])^{1/2} \|_{L^t} + C_1 \frac{t}{m} \cdot \| \max_{i \in [m]} Z_i^2 \|_{L^t} \\
\leq & C_1 \frac{\sqrt{t}}{m} \cdot \| \max_{i \in [m]} Z_i^2 \|_{L^t}^{1/2} \cdot \| \sum_{i \in [m]} \mathbb{E}[Z_i^2 \mid D] \|_{L^t}^{1/2} + C_1 \frac{t}{m} \cdot \| \max_{i \in [m]} Z_i^2 \|_{L^t}
\end{aligned}$$

where the first step follows from Definition C.1, the second step follows from Lemma C.5, the third step follows from triangle inequality, and the last step follows from Cauchy-Schwartz inequality.

Note that each row of G is generated by taking the tensor product of independent Rademacher random vectors, we thus can view the row vector itself as a length n Rademacher random vector. Thus,

$$\begin{aligned}
\mathbb{E}[Z_i^2 | D] = & \sum_{\sigma \in \{-1,+1\}^n} p_\sigma \cdot (\langle x, \sigma \rangle)^2 \\
= & \frac{(x_1 + x_2 + \cdots + x_n)^2}{2^n} + \frac{(x_1 + x_2 + \cdots - x_n)^2}{2^n} + \cdots + \frac{(-x_1 - x_2 - \cdots - x_n)^2}{2^n} \\
= & x_1^2 + x_2^2 + \cdots + x_n^2 \\
= & \|x\|_2^2,
\end{aligned} \tag{9}$$

where the first step follows from the definition of the expected value, $\mathbb{E}[Z_i^2 | D]$, the second step follows from expanding all the 2^n possibilities, the third step follows from simple algebra, and the last step follows from the definition of $\| \cdot \|_2^2$.

We could get that

$$\begin{aligned}
\sum_{i \in [m]} \mathbb{E}[Z_i^2 \mid D] = & \sum_{i \in [m]} \|x\|_2^2 \\
= & m,
\end{aligned}$$

where the first step follows from Eq. (9) and the second step follows from $\|x\|_2^2 = 1$.

To bound $\| \max_{i \in [m]} Z_i^2 \|_{L^t}$, we could show

$$\begin{aligned}
\|Z_i^2\|_{L^r} = & \|G_{P^i} D x\|_{L^{2r}}^2 \\
= & \|Dx\|_{L^{2r}}^2 \\
\leq & r^k \|x\|_2^2.
\end{aligned}$$

where the first step follows from the definition of Z_i, the second step follows from each row of G is independent Rademacher vector, therefore $\mathbb{E}[G^\top G] = I$, and the last step follows from Lemma C.4.

We then bound the maximum using a sufficiently high-order sum:

$$\begin{aligned}\|\max_{i\in[m]} Z_i^2\|_{L^t} &\leq \|\max_{i\in[m]} Z_i^2\|_{L^r} \\ &\leq (\sum_{i\in[m]} \|Z_i^2\|_{L^r}^r)^{1/r} \\ &\leq m^{1/r} r^k \|x\|_2^2 \leq e r^k,\end{aligned}$$

where the first step follows from Definition C.1, the second step follows from Z_i^2 is non-negative, and the last step follows from $r \geq \log m$.

This gives us that

$$\|\frac{1}{m}\sum_{i\in[m]} Z_i^2 - \|x\|_2^2\|_{L^t} \leq C_2\sqrt{\frac{tr^k}{m}} + C_2\frac{tr^k}{m} \tag{10}$$

which finishes the first part of the proof.

We want to choose m as follows

$$m = 16C_2^2\epsilon^{-2}\cdot\log(1/\delta)\cdot(C_3\log(1/(\delta\epsilon)))^k.$$

According to the above choice of m, we know following condition for r is holding

$$r \leq C_3\log(1/(\delta\epsilon)).$$

Hence,

$$m \geq 16C_2^2\epsilon^{-2}\cdot\log(1/\delta)\cdot r^k.$$

For all $1 \leq t \leq \log(1/\delta)$, we then get that

$$\begin{aligned}\|\|PGDx\|_2^2 - 1\|_{L^t} &\leq C_2\sqrt{\frac{tr^k}{m}} + C_2\frac{tr^k}{m} \\ &\leq C_2(\frac{tr^k}{16C_2^2\epsilon^{-2}\log(1/\delta)r^k})^{1/2} + C_2\frac{tr^k}{16C_2^2\epsilon^{-2}\log(1/\delta)r^k} \\ &\leq 0.5\epsilon\sqrt{t/\log(1/\delta)} + 0.5\epsilon^2 t/\log(1/\delta) \\ &\leq \epsilon\sqrt{t/\log(1/\delta)}.\end{aligned}$$

where the first step follows from Eq. (10), and the second step follows from choice of m, the third step follows from simple algebra, and the last step follows from $\epsilon^2 \leq \epsilon$ and $t/\log(1/\delta)\sqrt{t/\log(1/\delta)}$ (since $t/\log(1/\delta) \in (0,1)$) .

This finishes the proof. □

As two corollaries, we have SRCT and TensorSRCT are OSE's with d^2 rows, instead of d rows.

Corollary C.7 (SRCT is an OSE)**.** *Let $S \in \mathbb{R}^{m\times n}$ be an* SRCT *matrix with $m = \Theta(\epsilon^{-2}d^2\log^2(n/\epsilon\delta))$, then S is an* (ϵ,δ,d,n)-OSE.

Proof. The proof follows from combining Lemma C.3 and Theorem C.6 with $k = 1$. □

Corollary C.8 (TensorSRCT is an OSE)**.** *Let $S \in \mathbb{R}^{m\times n}$ be a* TensorSRCT *matrix with $m = \Theta(\epsilon^{-2}d^2\log^3(n/\epsilon\delta))$, then S is an* (ϵ,δ,d,n^2)-OSE.

Proof. The proof follows from combining Lemma C.3 and Theorem C.6 with $k = 2$. □

D. Gaussian and AMS

In this section, we prove that both random Gaussian matrices and AMS matrices satisfy OCE with good parameter β. Combining with the fact that they are OSE's, one can derive ℓ_∞ guarantee for them.

D.1. OSE Property of Random Gaussian and AMS

The OSE property for these two distributions are folklore. For a proof for them, see, e.g., (Woodruff, 2014).

Lemma D.1. *Let S be a random Gaussian matrix defined in Def. 2.4. If $m = \Theta(\epsilon^{-2}(d + \log(d/\delta)))$, then S is an (ϵ, δ, d, n)-OSE.*

Lemma D.2. *Let S be an AMS matrix defined in Def. 2.5. If $m = \Theta(\epsilon^{-2} d \log^2(n/\delta))$, then S is an (ϵ, δ, d, n)-OSE.*

D.2. OCE Property of Random Gaussian and AMS

In this section, we prove the OCE property of random Gaussian and AMS. We start with the pairwise inner product bound for these two distributions. For column norm bound, AMS has unit columns and we will prove for random Gaussian.

Lemma D.3 (Gaussian pairwise inner product bound, Lemma B.18 in (Song & Yu, 2021))**.** *Let $S \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix (Definition 2.4).*

Then, we have:

$$\Pr[\max_{i \neq j} |\langle S_{*,i}, S_{*,j} \rangle| \geq C \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \Theta(\delta).$$

Proof. Note for $i \neq j$, $S_{*,i}, S_{*,j} \sim \mathcal{N}(0, \frac{1}{m} I_m)$ are two independent Gaussian vectors. Let $z_k = S_{k,i} S_{k,j}$ and $z = \langle S_{*,i}, S_{*,j} \rangle$.

Then, we have for any $|\lambda| \leq m/2$,

$$\mathbb{E}[e^{\lambda z_k}] = \frac{1}{\sqrt{1 - \lambda^2/m^2}} \leq \exp(\lambda^2/m^2),$$

where the first step follows from $z_k = \frac{1}{4}(S_{k,i} + S_{k,j})^2 + \frac{1}{4}(S_{k,i} - S_{k,j})^2 = \frac{m}{2}(Q_1 - Q_2)$ where $Q_1, Q_2 \sim \chi_1^2$, and $\mathbb{E}[e^{\lambda Q}] = \frac{1}{\sqrt{1-2\lambda}}$ for any $Q \sim \chi_1^2$.

This implies $z_k \in \mathsf{SubExp}(2/m^2, 2/m)$ is a sub-exponential random variable. Here SubExp is the shorthand of sub-exponential random variable.

Thus, we have

$$z = \sum_{k=1}^{m} z_k \in \mathsf{SubExp}(2/m, 2/m),$$

by sub-exponential concentration Lemma A.3, we have

$$\Pr[|z| \geq t] \leq 2 \exp(-mt^2/4)$$

for $0 < t < 1$. Picking $t = \sqrt{\log(n^2/\delta)/m}$, we have

$$\Pr[|\langle S_{*,i}, S_{*,j} \rangle| \geq C \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \delta/n^2.$$

Taking the union bound over all $(i, j) \in [n] \times [n]$ and $i \neq j$, we complete the proof. □

Lemma D.4 (AMS pairwise inner product bound)**.** *Let $S \in \mathbb{R}^{m \times n}$ be an AMS matrix (Definition 2.5. Let $\{\sigma_i\}_{i \in [n]}$ be independent Rademacher random variables and $\overline{S} \in \mathbb{R}^{m \times n}$ with $\overline{S}_{*,i} = \sigma_i S_{*,i}$, $\forall i \in [n]$.*

Then, we have:

$$\Pr[\max_{i \neq j} |\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \Theta(\delta)$$

Proof. Note for any fixed $i \neq j$, $\overline{S}_{*,i}$ and $\overline{S}_{*,j}$ are independent. By Hoeffding inequality (Lemma 2.19), we have

$$\begin{aligned} & \Pr[|\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \geq t] \\ \leq & 2\exp(-\frac{2t^2}{\sum_{i=1}^m (\frac{1}{m} - (-\frac{1}{m}))^2}) \\ \leq & 2e^{-t^2 m/2}, \end{aligned}$$

where the second step follows from simple algebra ($m \cdot 1/m^2 = 1/m$).

Choosing $t = \sqrt{2\log(2n^2/\delta)}/\sqrt{m}$, we have

$$\Pr[|\langle \overline{S}_{*,i}, \overline{S}_{*,j} \rangle| \geq \sqrt{2\log(2n^2/\delta)}/\sqrt{m}] \leq \frac{\delta}{n^2},$$

union bound over all n^2 pairs of columns gives the desired result. □

Lemma D.5 (Gaussian column norm bound)**.** *Let $S \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix.*

Then, for any $i \in [n]$, we have

$$\Pr[|\|S_{*,i}\|_2^2 - 1| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \Theta(\delta).$$

Proof. For any column $S_{*,i}$, note that $\|S_{*,i}\|_2^2 \sim \chi_m^2$, each one with zero mean and variance $\frac{1}{m}$.

By Lemma 2.20, we have

$$\Pr[|\|S_{*,i}\|_2^2 - 1| \geq 2\frac{\sqrt{t}}{\sqrt{m}}] \leq 2\exp(-t).$$

Setting $t = \log(n/\delta)$, we have

$$\Pr[|\|S_{*,i}\|_2^2 - 1| \geq C \cdot \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}] \leq \delta/n,$$

the proof is concluded by union bounding over all n columns. □

We conclude random Gaussian and AMS are OCE's.

Lemma D.6 (Gaussian OCE)**.** *Let $S \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix, then S is a $(\log^{1.5}(n/\delta), \delta, n)$-OCE.*

Proof. By Lemma D.3 and Lemma D.5, we know both pairwise inner product bound and column norm bound hold and thus, by Lemma 3.5, S satisfies the desired OCE property. □

Lemma D.7 (AMS OCE)**.** *Let $S \in \mathbb{R}^{m \times n}$ be an AMS matrix, then S is a $(\log^{1.5}(n/\delta), \delta, n)$-OCE.*

Proof. The proof is similar to Lemma D.6. The column norm bound follows from definition. □