arXiv:2302.00248v1 [cs.DS] 1 Feb 2023

近乎最优界的快速回归，采用 ℓ_∞ Guarant ee

Zhao Song* Mingquan Ye† Junze Yin‡ Lichen Zhang§

摘要

给定一个矩阵 $A \in \mathbb{R}^{n \times d}$ 和一个向量 $b \in \mathbb{R}^n$，我们考虑具有 ℓ_∞保证的回归问题：寻找一个向量 $x' \in \mathbb{R}^d$，使得

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

其中 $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$。解决此类 ℓ_2 回归问题的一种常见方法是采用草图化：选取一个结构化随机矩阵 $S \in \mathbb{R}^{m \times n}$，使得 $m \ll n$ 和 SA 可以快速计算，然后求解“草图化”回归问题 $\arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$。

在本文中，我们证明，为了获得 ℓ_∞ 保证，对于 ℓ_2 回归，必须使用稠密草图矩阵。据我们所知，这是首个必须使用稠密草图矩阵的案例。在算法方面，我们证明存在一种稠密草图矩阵的分布，其 $m = \epsilon^{-2} d \log^3(n/\delta)$ 特性使得求解草图化回归问题能够以至少 ℓ_∞ 的概率提供保证 $1 - \delta$。此外，矩阵 SA 可以在 $O(nd \log n)$ 时间内计算得到。我们的行数在对数因子范围内近乎最优，并显著改进了 [PSW17], 的结果；在该结果中，d 要求超线性的行数，即对于 $\gamma = \Theta(\sqrt{\log_{\log n \log d}})$，需要 $m = \Omega(\epsilon^{-2} d^{1+\gamma})$。

此外，我们为 ℓ_∞ 保证回归开发了一种新颖的分析框架，该框架利用了无偏逐坐标嵌入（OCE）性质，该性质由 [SY21]提出。我们的分析比 [PSW17]的分析更简单、适用范围更广。借助这一框架，我们将 ℓ_∞ 保证回归的结果扩展到稠密草图矩阵，以实现向量的快速张量积计算。

*zsong@adobe.com。Adobe Research。
†mye9@uic.edu。芝加哥伊利诺伊大学。
‡junze@bu.edu。波士顿大学。
§lichenz@mit.edu。麻省理工学院。

目录

1 引言 **2**

2 预备知识 **4**
2.1 无偏子空间嵌入与逐坐标嵌入.4 2.2 草图矩阵.5 2.3 稠密草图的 OSE 性质.6 2.4 概率工具.6

3 ℓ_∞ 通过 OCE 获得保证 7
3.1 OCE 的高概率界.9 3.2 SRHT 和 SRCT 的内积界.11 3.3 SRHT 和 SRCT 的列范数界.13

4 综合起来 **13**

5 结论 **14**

A 矩阵与概率工具 15 B 带有 ℓ_∞ 保证的克罗内克积回归 15 B.1 主要结果.15 B.2 TensorSRHT 和 TensorSRCT 的无偏逐坐标嵌入 16

C SRCT 和 TensorSRCT：通过强 JL 矩性质实现子空间嵌入 **17**
C.1 符号约定.17 C.2 强 JL 矩性质.17 C.3 SRCT 和 TensorSRCT 满足强 JL 矩性质.20

D 高斯与 AMS 22
D.1 随机高斯和 AMS 的 OSE 性质.23 D.2 随机高斯和 AMS 的 OCE 性质.23

1 引言

线性回归，或 ℓ_2 最小二乘问题，在数值线性代数、科学计算和机器学习中无处不在。给定一个高瘦矩阵 $A \in \mathbb{R}^{n \times d}$ 和一个标签向量 $b \in \mathbb{R}^n$，目标是（近似地）计算一个最优解 x'，使 ℓ_2 损失 $\|Ax - b\|_2$最小。对于 $n \gg d$的情形，草图化是一种快速求得近似解的常用方法 [CW13、CSWZ23]：其思路是从精心设计的分布中选取一个随机矩阵 $S \in \mathbb{R}^{m \times n}$，使得 1）。 S 能够高效地应用于 A，且 2）。行数 $m \ll n$。有了这两项保证后，就可以求解“草图化”的回归问题：

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

并得到一个向量 x'，使得 $\|Ax' - b\|_2 = (1 \pm \epsilon)$ ·OPT，其中 OPT 表示 ℓ_2列空间中向量之间的最优差异，该列空间属于 A 和 b。草图化领域近期的进展 [CSWZ23]表明，可以设计出满足 $m = O(\epsilon^{-2} d \log^2(n/\delta))$ 的矩阵 S，并且草图化后的回归问题可以在 $O(\text{nnz}(A) + d^\omega + \epsilon^{-2} d^2 \text{ poly log}(n, d, 1/\epsilon, 1/\delta))$ 时间内求解，其中 $\text{nnz}(A)$ 表示 A 中非零元素的数量，而 δ 是失败概率。

遗憾的是，现代机器学习越来越重视大型、复杂且非线性的模型，例如深度神经网络，因此线性回归作为一种模型的吸引力逐渐减弱。然而，它仍是许多深度学习和优化框架中的重要子程序，尤其是在训练神经网络的二阶方法中 [BPSW21, SZZ21]，或在凸优化中 [LS14, LSZ19, SY21, JSWZ21]。在这些应用中，人们通常寻求前向误差保证，即近似解 x' 与最优解 x^*有多接近。一个典型的例子是牛顿法：给定（可能是隐式的）海森矩阵 $H \in \mathbb{R}^{d \times d}$和梯度 $g \in \mathbb{R}^d$，目标是计算 $H^{-1}g$。一种常见方法是求解回归问题 $\arg\min_{x \in \mathbb{R}^d} \|Hx - g\|_2$，其中希望 $\|x - H^{-1}g\|_2$ 甚至 $\|x - H^{-1}g\|_\infty$ 足够小。当矩阵 S 满足所谓的无偏子空间嵌入（OSE）性质 [Sar06], 时，可以证明近似解 x' 在 ℓ_2 意义下接近 x^*：

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

遗憾的是，ℓ_2-接近度无法刻画 x' 对 x^*的逼近效果，因为 x^* 的 ℓ_2 质量可以很好地分布在所有坐标上，而 x' 的质量却可能集中在少数几个坐标上。形式上，设 $a \in \mathbb{R}^d$ 是一个固定向量，则可以通过公式（1）来衡量 $\langle a, x' \rangle$ 偏离$\langle a$的程度：$x^* \rangle$

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

这个界显然过于宽松，因为正如预期，在随机方向上的偏差仅为$\frac{1}{\sqrt{d}}$ 这一 ℓ_2 差异的倍数。 [PSW17] 证明，当 S 取为子采样随机哈达玛变换（SRHT）时，这一直觉确实成立 [LDFU13]: [1]

$$|\langle a, x^* \rangle - \langle a, x' \rangle| \lesssim \frac{\epsilon}{\sqrt{d}} \|a\|_2 \|Ax^* - b\|_2 \|A^\dagger\|. \tag{2}$$

[1]我们之后将以下性质称为 ℓ_∞ 保证。

然而，他们的分析并不紧，因为他们要求行数 $m = \Omega(\epsilon^{-2}d^{1+\gamma})$，用于 $\gamma = \Theta(\sqrt{\log \log n / \log d})$。只要 $n \le \exp(d)$，这样的行数就会随 d 超线性增长，因此比使 S 成为子空间嵌入所需的行数更多；后者只需 $m = \epsilon^{-2}d\log^2 n$ 行即可达到常数成功概率。相比之下，对于随机高斯矩阵，ℓ_∞ 的保证仅要求行数接近 d 的线性规模。除了行数不够理想之外，[PSW17] 的分析也很复杂：设 $U \in \mathbb{R}^{n\times d}$ 是 A 的标准正交基，[PSW17] 必须分析矩阵 $I_d - U^\top S^\top SU$ 的高阶矩。这使得他们的分析很难推广到 SRHT 之外的其他稠密草图矩阵。

在本文中，我们提出了一种新颖的框架，用于分析 SRHT 以及更一般地说，一大类稠密草图矩阵所带来的 ℓ_∞ 保证。可以说，我们的分析比 [PSW17]，简单得多，并揭示了草图矩阵能够提供 ℓ_∞保证的基本结构：如果草图矩阵中任意两列以高概率具有较小的内积，那么就能保留 ℓ_∞ 保证。随后，我们证明，两列之间较小的成对列内积也与 [SY21] 提出的无偏坐标嵌入（OCE）性质密切相关。具体而言，对于任意两个固定向量 $g, h \in \mathbb{R}^n$，如果以至少 $1-\delta$ 的概率满足 $|\langle Sg, Sh\rangle - \langle g, h\rangle| \le \frac{\beta}{\sqrt{m}} \cdot \|g\|_2\|h\|_2$，我们就称该草图矩阵具有 (β, δ, n)-OCE 性质。此前，这一性质已用于近似动态变化的投影矩阵与在线向量序列之间的矩阵-向量乘积，应用于快速线性规划和经验风险最小化算法 [LSZ19、JSWZ21、SY21、QSZZ23]，因为这些算法需要矩阵-向量乘积的 ℓ_∞ 界。这些应用与 ℓ_∞ 保证的一个共同特点是使用稠密草图矩阵，例如随机高斯、Alon-Matias-Szegedy 草图（AMS，[AMS96]）或 SRHT。这与当前倾向于使用稀疏矩阵的趋势形成鲜明对比，例如 Count Sketch [CCFC02、CW13] 和 OSNAP [NN13、Coh16]，，因为这类矩阵能够以（近乎）输入稀疏度时间应用。

近年来，可应用于矩阵/向量张量积的草图越来越受欢迎 [ANW14，SWZ16，DSSW18，SWZ19，DJS+19，AKK+20，WZ20，SWYZ21，WZ22，SXZ22，SXYZ22，RSZ22] 因为它们能够加速优化任务和大规模核学习。我们证明，针对二阶张量的稠密草图也能提供 ℓ_∞ 保证。

定理 1.1（稠密草图矩阵的近乎最优界）。设给定 $n \le \exp(d)$、矩阵 $A \in \mathbb{R}^{n\times d}$ 和向量 $b \in \mathbb{R}^n$。令 $S \in \mathbb{R}^{m\times n}$ 为一个子采样随机哈达玛变换矩阵 SRHT，其有 $m = O(\epsilon^{-2}d\log^3(n/\delta))$ 行。

对于任意固定向量 $a \in \mathbb{R}^d$，

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

概率为 $1-\delta$，其中 $x' = \mathrm{argmin}_{x\in\mathbb{R}^d} \|SAx - Sb\|_2$，$x^* = \arg\min_{x\in\mathbb{R}^d} \|Ax - b\|_2$。

备注 1.2。行数 $m = \epsilon^{-2}d\log^3 n$ 几乎是最优的，差距仅在对数因子以内，因为要使 S 成为子空间嵌入（OSE），在常数成功概率下，所需行数为 $m = \epsilon^{-2}d\log^2 n$。相比之下，[PSW17] 对于 $\gamma = \Theta$ 需要 $m = \epsilon^{-2}d^{1+\gamma}$ 行（$\sqrt{\log_{\log n}\log d}$），只有当 $n > \exp(d)$ 时才是关于 d 的近线性。在大多数应用中，我们关注的是 $n = \mathrm{poly}(d)$，这意味着在几乎所有有实际意义的情形下，它们的行数都比我们的更多。

定理 1.1 中得到的行数和保证适用范围不止于SRHT；事实上，它适用于一系列稠密草图矩阵，包括随机高斯、AMS 草图 [AMS96]，SRHT 和子采样随机循环变换（见定义 2.11）（SRCT）。这是因为我们的论证基于一种结构条件，多种稠密草图都能满足该条件。

我们的结果可以 也可以推广到二次克罗内克积回归，参见定理 B.1 .

路线图。 在第 2 节中，我们介绍所用的符号，并解释用于支持 ℓ_∞ 保证回归框架的关键定义和性质。在第 3 节中，我们通过给出一个充分条件来介绍我们的框架，该条件确保草图矩阵能够提供良好的 ℓ_∞ 保证。在第 4 节中，我们汇总各项内容，对主要定理进行证明。最后，在第 5 节中，我们总结本文的主要发现，并通过与先前工作的比较进行讨论。

2 预备知识

对于正整数，我们定义 $[n] := \{1, 2, \cdots, n\}$。对于向量 $x \in \mathbb{R}^n$，我们定义 $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$ 和 $\|x\|_\infty := \max_{i\in[n]} |x_i|$。对于矩阵 A，我们定义 $\|A\| := \sup_x \|Ax\|_2/\|x\|_2$ 为 A 的谱范数。我们用 $\|A\|_F := \sum_{i,\ j} A_{i,\ j}^2$ 表示 A 的弗罗贝尼乌斯范数。一般而言，谱范数具有以下性质： $\|AB\| \leq \|A\| \cdot \|B\|$。我们用 $A^\dagger$ 表示 $m \times n$ 矩阵 A 的 Moore-Penrose 伪逆；如果 $A = U\Sigma V^\top$ 是其奇异值分解（其中 $U \in \mathbb{R}^{m\times n}$、$\Sigma \in \mathbb{R}^{n\times n}$ 和 $V \in \mathbb{R}^{n\times n}$ 对应 $m \geq n$），则该伪逆由 $A^\dagger = V\Sigma^{-1}U^\top$给出。

我们使用 $\mathbb{E}[\cdot]$ 表示期望，并使用 $\Pr[\cdot]$ 表示概率。对于分布D 和随机变量 x，我们使用 $x \sim D$ 表示从分布 D中抽取一个随机变量。我们使用 $\mathcal{N}(\mu, \sigma^2)$表示均值为 μ、方差为 σ^2的高斯分布。如果 $\Pr[x = 1] = 1/2$ 且 $\Pr[x = -1] = 1/2$，则称随机变量 x 为拉德马赫随机变量。我们也称其为符号随机变量。

除了 $O(\cdot)$表示法之外，对于两个函数 f和 g，我们使用简写 $f \lesssim g$ （分别为 $\gtrsim$）表示，对于某个绝对常数 C，有 $f \leq Cg$ （分别为 $\geq$）。我们用 $f \approx g$ 表示，当常数为 $c > 0$ 和 $C > 0$时的 $cf \leq g \leq Cf$。对于两个矩阵 $A \in \mathbb{R}^{n_1\times d_1}$和 $B \in \mathbb{R}^{n_2\times d_2}$，我们使用 $A \otimes B \in \mathbb{R}^{n_1 n_2 \times d_1 d_2}$ 表示克罗内克积，即， $A \times B$的第（i_1， i_2）行和第（j_1， j_2）列为 $A_{i_1,\ j_1} B_{i_2,\ j_2}$。对于两个向量 $x \in \mathbb{R}^m$ 和 $y \in \mathbb{R}^n$，我们使用 $x \otimes y \in \mathbb{R}^{mn}$ 表示向量的张量积，其中 $x \otimes y = \mathrm{vec}(xy^\top)$。

2.1 无偏子空间嵌入和逐坐标嵌入

定义 2.1（无偏子空间嵌入 [Sar06]）。我们将（ϵ, δ, d, n）-无偏子空间嵌入（OSE）定义如下：设 Π 是 $m \times n$ 矩阵 S上的一个分布，其中 m 是 n、d、ϵ和 δ的函数。假设对于任意固定的 $n \times d$ 标准正交基U，以至少 $1 - \delta$的概率，从该分布 Π 中抽取的矩阵 S 具有如下性质： SU 的奇异值均落在范围 $[1 - \epsilon, 1 + \epsilon]$内。

无偏逐坐标嵌入（OCE）在[LSZ19、JSWZ21] 中被隐式使用，并在 [SY21]中正式提出。

定义 2.2 （（α， β， δ）-逐坐标嵌入 [SY21]）。 我们称随机矩阵 $S \in \mathbb{R}^{m\times n}$满足（$\alpha$， β， δ）-逐坐标嵌入，如果

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

本文主要使用定义 2.2中的性质 3。为方便起见，我们重新定义 OCE如下：

定义 2.3（OCE）。设 $\beta \geq 1$ 和 $\delta \in$ (0,0.1)。如果随机矩阵 $S \in \mathbb{R}^{m \times n}$ 满足 (β, δ, n)-OCE，则称其满足

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

且分布 Π 对任意固定向量 g 和 h都是无关的。

2.2 草图矩阵

本文关注一类稠密草图矩阵。

定义 2.4（随机高斯矩阵，民间常识）。如果 $S \in \mathbb{R}^{m \times n}$ 的所有元素均独立地从 $\mathcal{N}(0, 1/m)$ 中采样，则称其为随机高斯矩阵。

定义 2.5（AMS 草图矩阵，[AMS96]）。令 h_1、h_2、$\cdots$、h_m 为从四路独立哈希族 $\mathcal{H} = \{h : [n] \to \{-\frac{1}{m}, +\frac{1}{m}\}\}$中选取的随机哈希函数。若令 $S_{i,\ j} = h_i(j)$，则 $S \in \mathbb{R}^{m \times n}$为 AMS 草图矩阵。

以下草图矩阵可以利用快速傅里叶变换（FFT），以便高效地应用于矩阵。

定义 2.6（子采样随机哈达玛变换（SRHT）[LDFU13，SWYZ21]）。SRHT 矩阵 $S \in \mathbb{R}^{m \times n}$定义为 $S := \frac{1}{\sqrt{m}} PHD$，其中矩阵 $P \in \{0,1\}^{m \times n}$的每一行都恰好在一个随机位置上含有一个 1，$H$ 是 $n \times n$ 哈达玛矩阵，而 D 是一个 $n \times n$对角矩阵，其每个对角元素以相同概率取 $\{-1, +1\}$ 中的一个值。

备注 2.7。利用快速傅里叶变换（FFT），S 可以在$O(n \log n)$时间内作用于一个向量。

定义 2.8（张量子采样随机哈达玛变换（TensorSRHT）[SWYZ21]）。TensorSRHT $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ 定义为 $S := \frac{1}{\sqrt{m}} P \cdot (HD_1 \otimes HD_2)$，其中 $P \in \{0,1\}^{m \times n^2}$ 的每一行仅在一个随机坐标处含有一个 1，并且可以将 P 视为采样矩阵。H 是一个 $n \times n$ 哈达玛矩阵，而 D_1、D_2 是两个 $n \times n$ 独立对角矩阵，其对角元素分别独立地设为拉德马赫随机变量（在$\{-1, 1\}$上均匀分布）。

备注 2.9。利用草图空间中的 FFT 算法，$S(x \otimes y)$ 的计算时间为 $O(n \log n + m)$。

存储和生成哈达玛矩阵的成本很高，因此我们考虑通过循环变换生成 FFT 矩阵的一种成本更低且节省空间的方法。

定义 2.10（循环矩阵）。循环矩阵是一个 $n \times n$ 矩阵，其中 $n \in \mathbb{N}$；其各行向量由相同元素组成，且与前一行向量相比，每个行向量都向右循环移位一个元素。

定义 2.11（子采样随机循环变换（SRCT））。设 $x \in \mathbb{R}^n$ 为一个随机向量，其元素是独立同分布的拉德马赫随机变量。

此外，设 $P \in \mathbb{R}^{m \times n}$ 为一个随机矩阵，其中每行在均匀随机选取的坐标处包含一个 1，其余位置均为 0。

设 $G \in \mathbb{R}^{n \times n}$ 为一个循环矩阵（见定义 2.10），由 x 生成；并设 $D \in \mathbb{R}^{n \times n}$ 为一个对角矩阵，其对角元素是独立同分布的拉德马赫随机变量。

然后，su 子采样随机循环变换定义如下：$S := \frac{1}{\sqrt{m}} PGD$.

定义 2.12（张量子采样随机循环变换（TensorSRCT））。设 $x \in \mathbb{R}^n$ 为一个随机向量，其元素为独立同分布的拉德马赫随机变量。

此外，令 $P \in \mathbb{R}^{m \times n^2}$ 为一个随机矩阵，其中每一行在均匀随机选取的坐标处为 1，其余位置均为 0。

令 $G \in \mathbb{R}^{n \times n}$ 为由 2.10 定义生成的循环矩阵（参见定义 x。

令 $D_1 \in \mathbb{R}^{n \times n}$ 和 $D_2 \in \mathbb{R}^{n \times n}$ 为两个独立对角矩阵，其对角元素为独立同分布的拉德马赫随机变量。

则张量循环变换 $T\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ 定义如下： $T := P \cdot (GD_1 \otimes GD_2)$。

备注 2.13。与 SRHT 类似，我们可以结合循环矩阵使用快速傅里叶变换。SRCT 可在 $O(n\log\ n)$ 时间内作用于长度为 n 的向量，而 TensorSRCT 可在 $O(n\log\ n + m)$ 时间内作用于 $x \otimes y$。

2.3 稠密草图的 OSE 性质

草图求解回归的一项重要条件是子空间嵌入。我们尤其关注 SRHT、SRCT 及其张量变体。

引理 2.14 （ [SWYZ21]中的引理 2.11）。 设 S 是一个根据定义 SRHT 定义的矩阵2.6。若$m = O$（$\epsilon^{-2}d$对数^2（nd/δ））， 则 S是一个 （ϵ， δ， d， n）-子空间嵌入。

引理 2.15 （[SWYZ21]中的引理 2.12）。 设 S是一个根据定义 TensorSRHT 定义的矩阵2.8。若 $m = O$（$\epsilon^{-2}d$对数^3（$nd/\epsilon\delta$））， 则 S是一个 （ϵ， δ， d， n^2）-子空间嵌入，适用于次数为 2 的张量。

由于格拉姆矩阵 $G^\top G$ 仅在 I_n 意义下成立，SRCT 所需的行数比 SRHT 更多。

引理 2.16 （推论 C.7的非正式版本）。 设 S 是一个根据 SRCT 定义 2.11 定义的矩阵。若 $m = O$（$\epsilon^{-2}d^2$ 对数^2（nd/δ））， 则 S是一个 （ϵ， δ， d， n）-子空间嵌入。

引理 2.17 (推论 C.8 的非正式版本)。 设 S 是定义 2.12 中定义的 TensorSRCT 矩阵。若 $m = O($ $\epsilon^{-2}d^2\log^3(nd/\delta))$，则 S是一个 $(\epsilon, \delta, d, n^2)$-子空间嵌入。

2.4 概率工具

引理 2.18 (Khintchine不等式 [Khi23])。 设 $\sigma_1, \cdots, \sigma_n$为独立同分布的 Rademacher 随机变量，且 z_1, $\cdots, z_n$为实数。那么，存在常数 $C, C' > 0$使得

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

引理 2.19 (Hoeffding界，[Hoe63])。 设 $Z_1, \cdots, Z_n$为独立、零均值随机变量，且满足 $Z_i \in [\alpha_i, \beta_i]$。那么，

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

引理 2.20 （Laurent和Massart第1325页的引理1 [LM00] ）。 设 $X \sim \mathcal{X}_k^2$ 为一个服从卡方分布的随机变量，其 k 自由度。每个随机变量的均值均为零，且 σ^2方差为。于是，

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

引理 2.21 （Hanson-Wright 不等式 [HW71]）。 设 $x \in \mathbb{R}^n$ 表示一个各分量相互独立的随机向量， x_i 且 $\mathbb{E}[x_i] = 0$ 并且 $|x_i| \leq K$。设 A 为一个 $n \times n$ 矩阵。那么，对任意$t \geq 0$,

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

引理 2.22（矩阵 Chernoff 界， [Tro10] 中的定理 2.2）。设 X 为一组有限的正半定矩阵，其维度为 $d \times d$。假设 $\max_{X \in \mathcal{X}} \lambda_{\max}(X) \leq B$。从 $\mathcal{X}$ 中均匀随机地无放回抽取 $\{X_1, \cdots, X_n\}$。我们将 $\mu_{\min}$ 和 $\mu_{\max}$ 定义如下： $\mu_{\min} := n \cdot \lambda_{\min}(\mathbb{E}_{X \sim \mathcal{X}}[X])$，以及 $\mu_{\max} := n \cdot \lambda_{\max}(\mathbb{E}_{X \sim \mathcal{X}}[X])$。那么，

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

对于 $\delta \in [0, 1)$,

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

对于 $\delta \geq 0$。

通过 OCE 实现 l∞ 保证

本节将给出使草图矩阵能够提供良好 ℓ_∞ 保证的充分条件：给定一对固定向量 g, h ，满足 $g^\top h = 0$，若草图矩阵以高概率近似保持内积，则它能够为回归提供良好的 ℓ_∞ 保证。

引理 3.1（核心引理）。设 $A \in \mathbb{R}^{n \times d}$ 是一个固定矩阵。令 $U \in \mathbb{R}^{n \times d}$ 表示 A的标准正交基。令 $S \in \mathbb{R}^{m \times n}$ 为满足以下两个性质的草图矩阵

- S是一个 $(0.1, \delta, d, n)$-子空间嵌入 (其中 $\delta \in (0, 0.1)$，定义 *2.1*）。
- S是一个 (β, δ, n)-OCE (其中 $\beta \geq 1$ 和 $\delta \in (0,0.1)$，定义 *2.3*）。

对于任意固定向量 $a \in \mathbb{R}^d$ 和 $b \in \mathbb{R}^n$ ，若满足 $U^\top b = 0$，则有

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

成立的概率至少为 $1 - \delta$。

证明。以概率 1，矩阵 $SA \in \mathbb{R}^{m \times d}$ 的列线性独立。因此，$(SA)^\dagger \in \mathbb{R}^{d \times m}$ 是

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

其中，第一步由 $SA \in \mathbb{R}^{m \times d}$ 满秩得出，第二步由对$A \in \mathbb{R}^{n \times d}$进行奇异值分解得出，第三步由 $(AB)^{-1} = B^{-1}A^{-1}$得出，最后一步则根据奇异值分解的性质，由 V 是正交的这一事实得出。

为方便起见，我们定义 x 如下：

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

在接下来的几段中，我们将说明如何以高 $|x|$ 概率给出其上界。由于 S是一个 $(0.1, \delta, d, n)$-子空间嵌入 (定义 2.1)，我们知道

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

我们以该事件为条件。由此可得

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

其中，第一步是因为V是一个旋转，第二步由次乘性得出，第三步则由 $\|I - U^\top S^\top S U\| \leq 0.1$ 以及 U是一个旋转得出。

因此，我们有

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

定义一个向量 $u \in \mathbb{R}^n$

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

根据 OCE (定义 2.3，我们有

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

其中 $U^\top b = 0$ 给出 $u^\top b = 0$ 和 $u^\top S^\top S b = x$。

因此，上述界可转化为

$$\Pr[|x| \leq C \cdot \frac{\beta}{\sqrt{m}} \cdot \|\Sigma^{-1}\| \|a\|_2 \|b\|_2] \geq 1 - \delta \tag{4}$$

如所需。 □

现在我们 准备证明 ℓ_∞ 在引理 3.1的内积界条件下给出的保证 .

定理 3.2。 假设 $A \in \mathbb{R}^{n \times d}$ 具有满列秩，且 $b \in \mathbb{R}^n$。令 $S \in \mathbb{R}^{m \times n}$ 为满足引理 *3.1*中条件的草图矩阵。对于任意固定向量 $a \in \mathbb{R}^d$，有

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

以至少 $1 - \delta$ 的概率成立，其中 $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$和 $x' = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$。

证明。由于 A 满列秩，我们有 $x^* = A^\dagger b$。类似地， SA 以概率 1 满列秩，因此 $x' = (SA)^\dagger Sb$ 和 $(SA)^\dagger SA = I$。因此，我们有

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

其中 $U \in \mathbb{R}^{n \times d}$是 A的标准正交基。众所周知， $I - UU^\top = U_\perp U_\perp^\top$其中$U_\perp \in \mathbb{R}^{n \times (n-d)}$是 span($A$) 的正交分量的标准正交基。为使上述表达式最大化，我们令 $b \in \mathrm{span}(U_\perp)$，等价地，$U^\top b = 0$。因此，对式（5）求界等价于考虑

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

根据引理，以下不等式以至少 $1 - 2\delta$ 的概率成立 3.1。最后，注意到由于$U^\top b = 0$，我们有 $Ax^* = 0$ ，从而证明了

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

□

注意，我们只需要满足 $\epsilon = O(1)$ 的 OSE，而对 ϵ 的依赖关系可由 OCE 的行数得到。

3.1 OCE 的高概率界

本节给出一个用于证明 OCE 高概率界的统一框架。我们的分析使用三个稠密草图矩阵，它们都可以通过以下方式构造：先选取一组新的随机符号，再根据分布选取草图矩阵。

我们陈述 k 于稠密草图矩阵的关键假设，足以保证 OCE 性质 .

假设 3.3. 设 $S \in \mathbb{R}^{m \times n}$ 为满足以下两个假设的稠密草图矩阵：

- 两两内积界：

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

- 列范数界：

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

对于所有 $i \in [n]$。

引理 3.4。 设 $S \in \mathbb{R}^{m \times n}$ 是满足假设 3.3 的稠密草图矩阵。设 $h \in \mathbb{R}^n$ 和$g \in \mathbb{R}^n$ 是两个固定向量。那么，以下性质成立：

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

以至少为 $1 - \delta$ 的概率成立。

证明。我们可以将 $(g^\top S^\top Sh)$ –$(g^\top h)$ 重写如下：

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

其中，第一步基于以下事实： σ_i's 是独立的拉德马赫随机变量，且 $S_{*,\,i} = \sigma_i \overline{S}_{*,\,i}$、 $\forall i \in [n]$；第二步则是将对角项与非对角项分开。

我们将重点对 off-diag 这一量进行界定，因为 diag 可以用相当简单的方法处理。

我们定义矩阵 $A \in \mathbb{R}^{n \times n}$ 和 $B \in \mathbb{R}^{n \times n}$ 如下：

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

我们将去掉对角线元素后的矩阵 $A \in \mathbb{R}^{n \times n}$ 定义为 $A^\circ \in \mathbb{R}^{n \times n}$ 。

应用 Hanson-Wright 不等式（引理 2.21），可得

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

我们可以给出 $\|A^\circ\|$ 和 $\|A^\circ\|_F$的上界。

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

其中，第一步由 $\|\cdot\| \leq \|\cdot\|_F$ 得出，第二步由 A° 的定义得出，第三步由 A 和 B 的定义得出，第四步则由 B 的秩为 1 得出，因为 $B = \max_{i \neq j} |\langle \overline{S}_{*,\,i}, \overline{S}_{*,\,j} \rangle| \cdot g h^\top$。

接下来需要对 $\max_{i\neq j}|\langle \overline{S}_{*,\ i}, \overline{S}_{*,\ j}\rangle|$ 求一个界。注意，对于任意列 i, j,

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

其中，第一步是因为随机符号不会改变内积的大小，第二步则由 $S_{*,i}$ 和 $S_{*,j}$ 的定义得出。

由于 S 满足假设 3.3，因此我们以至少 $1-\delta$ 的概率得到,

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

在上述事件成立的条件下，我们有

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

取 $\tau = \|g\|_2\cdot\|h\|_2\cdot \log^{1.5}(n/\delta)/\sqrt{m}$，我们可以证明

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

为了界定 diag 项，注意到根据假设 3.3，我们以至少$1-\delta$的概率有 $|\|S_{*,i}\|_2^2 - 1| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{m}}$。

在此事件成立的条件下，我们有

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

其中 t 最后一步利用了柯西-施瓦茨不等式。注意， |diag| 可并入 |off-diag |.

对所有事件应用并集界，可得

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

□

3.2 SRHT 和 SRCT 的内积界

我们将证明 SRHT和 SRCT 满足假设 3.3。在证明两两内积界之前，我们先陈述一个刻画这些草图矩阵的一般性质。这一关键性质将在后续证明中使用。

定义 3.5（符号结构）。对于任意草图矩阵，若其满足以下性质，我们就称其具有“符号结构”

- $S_{k,\ i} \in \{\pm\frac{1}{\sqrt{m}}\}$，对所有 $k\in[m], i\in[n]$。
- $S_{k,\ i}$ 和 $S_{k,\ j}$ 对于任意 $i\neq j$都是独立的。

- $\mathbb{E}[S_{k,i}] = 0$ 对于所有 $k \in [m]$ 和 $i \in [n]$。

引理 3.6。 SRHT 和 SRCT均满足定义 *3.5*。

证明。直接由这两个草图矩阵的定义可得。 □

引理 3.7 (SRHT 和 SRCT)。 设 $S \in \mathbb{R}^{m \times n}$ 为任意满足定义 *3.5*的草图矩阵。那么，我们有

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

证明。固定一对索引 $i \neq j$，并如下定义 $X \in \mathbb{R}^{m \times 2}$：

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

格拉姆矩阵为 $X^\top X = \sum_{k=1}^m G_k$，其中

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

其中，第一步由 G_k 的定义得出，第二步由改写 $\begin{bmatrix} S_{k,i} & S_{k,j} \end{bmatrix}^\top$ 得出，第三步由矩阵乘法的定义得出，最后一步由 $S_{k,i}^2 = 1/m$ 和 $S_{k,j}^2 = 1/m$ 得出。

注意， G_k 的特征值为 0 和 $\frac{2}{m}$，即

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

由于 $S_{k,\ i}$ 和 $S_{k,\ j}$ 是相互独立的拉德马赫随机变量，因此我们有

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

因此，我们知道

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

因此，我们得到

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

其中，第一步根据 $X^\top X$ 的定义得出，第二步根据以下事实得出：$\mathbb{E}[ca] = c\,\mathbb{E}[a]$ 对于常数 c，第三步根据式（6）得出，最后一步则由简单代数运算得出。

令 $\lambda_i(X^\top X)$ 为 i-th 个特征值，属于 $X^\top X \in \mathbb{R}^{2\times 2}$。根据矩阵 Chernoff 界（引理 2.22，其中 $B = 2/m$），对于任意 $t > 0$，我们有

$$\Pr[\forall i \in [2], |\lambda_i(X^\top X) - 1| \geq t] \leq 4\exp(-t^2 m/2)$$

这意味着，以至少 $1 - 4\exp(-t^2 m/2)$ 的概率，$X^\top X$ 的特征值介于 $[1-t, 1+t]$ 之间，因此，$X^\top X - I_2$ 的特征值介于 $[-t, t]$ 之间。令 $t = O(\frac{\sqrt{\log(n/\delta)}}{\sqrt{m}})$，则有

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

证明可以通过注意到以下事实来完成

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

该矩阵的谱范数为 $|\langle S_{*,\,i}, S_{*,\,j}\rangle|$，并对所有 n^2 对列应用并集界，可得

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

□

3.3 SRHT 和 SRCT 的列范数界

本节将证明 SRHT 和 SRCT 的列范数界。特别地，它们的列都是单位向量。在附录D中，我们将证明，对于随机高斯矩阵，其列范数平方是一个 χ^2_m 随机变量，以高概率集中于 1 附近。

Lemm命题 3.8（SRHT 和 SRCT）。令 $S \in \mathbb{R}^{m\times n}$ 为 SRHT 矩阵或 SRCT 矩阵 $ix.$

那么，对于任意 $i \in [n]$，我们有 $\|S_{*,i}\|_2^2 = 1.$

证明。该证明可直接由定义得出。

对于 SRHT，回忆一下，$S = PHD$，H 的列范数为 $\sqrt{n}$，而 D 是一个不会改变范数的随机符号。矩阵 P对 m 行进行子采样，并将每个元素缩放 $\sqrt{\frac{1}{m}}$。因此，（平方）列范数为 1。

对于 SRCT，G 的列范数同样为 $\sqrt{n}$ 。因此，按照相同的论证，SRCT 的列向量均为单位向量。

□

4 综上所述

现在，我们可以给出定理 1.1 的证明了。

定理 *1.1* 的证明。 根据引理 2.14 （该引理表明 SRHT 可生成 子空间嵌入），我们知道，如果 $m \geq d\log^2(n/\delta)$，它将生成 $(O(1)\ \delta, n, d)$-子空间嵌入。

利用引理3.7 (它表明SRHT 具有OCE)，我们知道 $\beta = O(\log^{1.5}(n/\delta))$.

利用引理3.1 (它表明子空间嵌入 + OCE 蕴含我们的结果)，我们需要选择

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

将上述方程合并起来，我们得到

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

□

5 结论

本文研究了具有 ℓ_∞ 保证的基于草图的回归算法。我们证明，SRHT 使用 $m = \epsilon^{-2} d\log^3(n/\delta$行可提供所需的 ℓ_∞ 保证解，优于 $\epsilon^{-2} d^{1+\gamma}$行 (对于 $\gamma = \sqrt{\log_{\log n \log d}}$个 [PSW17]) 。这一结果除对数因子外近乎最优。我们的证明以新颖的方式改进了 [SY21]提出的无偏坐标嵌入性质。我们还大幅拓展了 ℓ_∞ 保证的适用范围，通过TensorSRHT 矩阵将其扩展至二次克罗内克积回归。

此外，我们还引入了 SRCT 和 TensorSRCT 矩阵。这些矩阵的应用方式与 SRHT 类似，并且具有与 SRHT 相似的 OCE 特性。

我们的结果提供了一种优雅的方法，将快速的、基于草图的回归求解器整合到优化过程中，尤其是二阶方法中。每次迭代中的回归问题都可以在几乎与输入规模呈线性关系的时间内求解，而 ℓ_∞ 这一保证在使用近似步长分析收敛性时很有帮助。它还通过SRHT [PSW17]改进了近似回归的泛化界。

附录

路线图。 在第 A 节中，我们介绍将在附录中使用的基本定义和性质。在第 B 节中，我们分析并推导克罗内克积回归的 ℓ_∞ 保证。在第 C 节中，我们介绍强 JL 矩性质，并证明循环变换和张量循环变换都满足该性质。在第 D 节中，我们着重研究 AMS、随机高斯和 SRHT，并证明对于 AMS、随机高斯和 SRHT 稠密草图矩阵中任意一对不同的列，其内积都有界。

A 矩阵与概率工具

对于矩阵 $A_1 \in \mathbb{R}^{n_1 \times d_1}$ 和 $A_2 \in \mathbb{R}^{n_2 \times d_2}$，我们用 $A_1 \otimes A_2 \in \mathbb{R}^{n_1 n_2 \times d_1 d_2}$ 表示满足 $(i_1 - 1) \cdot (n_2) + i_2$，$(j_1 - 1)d_2 + j_2$ 阶元素为 $(A_1)_{i_1,\ j_1} \cdot (A_2)_{i_2,\ j_2}$ 的矩阵。

引理 A.1（马尔可夫不等式）。若 X 是一个非负随机变量，且 $a > 0$。则有

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

定义 A.2 (次指数分布（[FKZ11]）)。 我们称 $X \in$ 次指数分布$(\sigma^2,\ \alpha)$，参数为 $\sigma > 0,\ \alpha > 0$，若满足:

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

引理 A.3 (次指数分布的尾界（[FKZ11]）)。 设 $X \in$ 次指数分布$(\sigma^2,\ \alpha)$ 且$\mathbb{E}[X] = \mu$。 则,

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

断言 A.4. 对于每个矩阵 $A \in \mathbb{R}^{n_1 \times n_2}, B \in \mathbb{R}^{n_2 \times n_3}, C \in \mathbb{R}^{d_1 \times d_2}, D \in \mathbb{R}^{d_2 \times d}$ 3

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

B 带有 ℓ_∞ 保证的克罗内克积回归

在本节中，我们研究 ℓ_∞ 克罗内克积回归的保证。给定两个矩阵A_1、$A_2 \in \mathbb{R}^{n \times d}$ 以及标签向量 $b \in \mathbb{R}^{n^2}$，目标是求解回归问题 $\arg\min_{x \in \mathbb{R}^{d^2}} \|(A_1 \otimes A_2)x - b\|_2^2$。这个问题可以轻松推广到由 q 个矩阵组成的乘积；多项研究已探讨快速的输入稀疏时间算法[DSSW18、SWZ19、DJS+19、FFG22、RSZ22]。

B.1 主要结果

定理 B.1 (定理的张量版本 1.1)。 假设 $n \leq \exp(d)$ 以及矩阵 $A \in \mathbb{R}^{n^2 \times d^2}$ 和向量 $b \in \mathbb{R}^{n^2}$ 已给定，其中 $A = A_1 \otimes A_2$ 对于矩阵 A_1、$A_2 \in \mathbb{R}^{n \times d}$ 以及 $b = b_1 \otimes b_2$ 对于向量b_1、$b_2 \in \mathbb{R}^n$。令 $S \in \mathbb{R}^{m \times n^2}$ 为一个

- 张量子采样随机哈达玛变换矩阵 (TensorSRHT)，其中 $m = \Theta(\epsilon^{-2} d^2 \log^3(n/\delta)$ 行，或)
- 张量子采样随机循环变换矩阵 (TensorSRCT)，具有 $m = \Theta(\epsilon^{-2} d^4 \log^3(n/\delta))$ 行。

For

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

and

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

以及任意固定的 $a \in \mathbb{R}^{d^2}$,

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

以 $1 - 1/\mathrm{poly}(d)$ 的概率。

证明。回顾一下，我们要求 $(O(1)\ \delta, d, n)$-子空间嵌入（OSE）和 $\beta = O(\log^{1.5}(n/\delta))$-OCE，才能为其提供 ℓ_∞保证。

对于 OCE，由引理 B.3可得。

对于 TensorSRHT 的 子空间嵌入（OSE），由引理 2.15 可得；对于 TensorSRCT，由推论 C.8可得。 □

注 B.2。略有不同的保证源于较小的维度变为 d^2 而不是 d。下面讨论使用这些草图矩阵求解回归的实用性。如定义 2.8 和 2.12 中所述， $A_1 \otimes A_2$ 的每一列都可以在 $O(n \log n + m)$ 时间内计算，而不是 n^2，因此将 S 应用于 A 的总运行时间为 $O(nd^2 \log n + \mathrm{poly}(d))$。类似地， Sb 可以在 $O(n \log n + \mathrm{poly}(d))$ 时间内应用。随后可以在 $\widetilde{O}(nd^2 + \mathrm{poly}(d))$ 时间内求解该回归问题。以往的工作主要关注输入稀疏草图 [DSSW18], 、重要性采样 [DJS+19]、迭代方法 [FFG22] ，或更复杂的草图，这些草图能够很好地扩展到 q 积，并适用于动态设置 [RSZ22]。据我们所知，这是首个针对克罗内克积回归（包含两个矩阵）的 ℓ_∞ 保证。

B.2 TensorSRHT 和 TensorSRCT 的无偏逐坐标嵌入

引理 B.3 (TensorSRHT 和 TensorSRCT，引理 3.7 的张量版本)。设 $S \in \mathbb{R}^{m \times n}$ 是 TensorSRHT 或 TensorSRCT。则 S 是一个参数为 $\beta = \log^{1.5}(n/\delta)$ 的OCE 。

证明。为证明 这一结果，我们证明 TensorSRHT 和 TensorSRCT 满足定义 3。 5.

对于 TensorSRHT，回顾 $S = \frac{1}{\sqrt{m}} P(HD_1 \times HD_2)$，因为 H是哈达玛矩阵，而 D_1、D_2只是带有随机符号的对角矩阵。因此， $HD_1 \times HD_2$ 的所有元素也都属于 $\{\pm 1\}$。由于 P是行采样矩阵，并且我们将每个元素按$\frac{1}{\sqrt{m}}$进行缩放。因此， S的每个元素都属于 $\{\pm \frac{1}{\sqrt{m}}\}$。对于同一行但不同列的元素 i、j，如果 i 由与 j不相交的两列生成，则显然它们是独立的。否则，假设 i 由列 a、b生成，且 j 由列 a、c 生成，其中 $b \neq c$。那么它们再次独立，因为该符号完全由 b 和 c的符号决定。最后，我们需要验证 $\mathbb{E}[S_{k,i}] = 0$；这显然成立，因为两个随机符号的乘积仍然是随机符号。对于 TensorSRCT，论证完全相同。

既然这两个矩阵都满足定义 3.5，我们就可以利用引理3.7 来给出成对内积的界。根据定义，列范数界自动成立。因此，我们可以调用引理 3.4 来完成证明。

□

C SRCT 和 TensorSRCT：通过强 JL 矩性质实现子空间嵌入

在本节中，我们证明 SRCT和 TensorSRCT 都是 子空间嵌入。我们通过强 JL 矩性质 [AKK+20]来证明这一性质。与 SRHT 和 TensorSRHT相比，这会导致更大的行数。我们认为，这两类分布对于一个子空间嵌入 应该具有相近的行数，并将缩小这两类分布之间的差距留作一个重要的开放问题。

C.1 符号约定

为使符号表示更简洁，我们将用$\|X\|_{L^t}$ 表示随机变量 X 的第 t 阶矩。其正式定义如下。

定义 C.1（t 阶矩）。对于每个整数 $t \geq 1$ 和任意随机变量 $X \in \mathbb{R}$，我们记作 e

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

注意，

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

根据闵可夫斯基不等式，对任意随机变量 X, Y。

C.2 强 JL 矩性质

我们证明 SRCT(见定义 2.11) 和 TensorSRCT (见定义 2.12) 都满足所谓的 强 JL 矩性质。强 JL 矩性质是核心性质之一，它可以表明草图矩阵具有子空间嵌入性质 [Sar06]。

定义 C.2 (强 JL 矩性质 [AKK+20])。 对于每个 $\epsilon, \delta \in [0, 1]$，我们称随机矩阵上的一个分布 $S \in \mathbb{R}^{m \times n}$ 具有强（ϵ, δ）-JL 矩性质，当

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

and

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

对于所有 $x \in \mathbb{R}^n$, $\|x\|_2 = 1$ 以及每个整数 $t \leq \log(1/\delta)$。

众所周知，具有强 JL 矩性质的分布可提供子空间嵌入。

引理 C.3 ([AKK+20]中的引理 11。 设 $S \in \mathbb{R}^{m \times n}$ 是一个具有 $(\epsilon/d, \delta)$-强 JL 矩性质（定义 *C.2*）的随机矩阵。那么， S 也是一个 (ϵ, δ, d, n)-子空间嵌入 （定义 *2.1*）。

为了证明 SRCT（见定义 2.11）和 TensorSRCT （见定义 2.12）满足强 JL 矩性质，我们将证明一个更一般的矩阵类满足强 JL 矩性质。

更准确地说，令 $k \in \mathbb{Z}_{>0}$ 为正整数，并且

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

是独立矩阵，每个矩阵的对角线元素由独立的拉德马赫变量构成。

令 $n = \prod_{i\in[k]} n_i$ 和 $P \in \{0,1\}^{m\times n}$ 为随机采样矩阵，其中每一行恰好包含一个均匀分布的非零元素，其值为 1。

接下来，我们证明矩阵

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

满足强 JL 矩性质，其中 G 是 $n \times n$ 循环矩阵（见定义 2.10），由一个元素为拉德马赫变量的随机向量生成。

若 $k = 1$，则 S 就是一个 SRCT （见定义 2.11）。若 $k = 2$，则 M 是一个 TensorSRCT （见定义 2.12）。

为了证明这一结果，我们需要几个引理。第一个引理可以看作是 Khintchine 不等式（见引理2.18）在高阶混沌情形下的一个版本。

引理 C.4 （[AKK+20]中的引理 19）。设 $t \geq 1$ 且 $k \in \mathbb{Z}_{>0}$。设 $(\sigma^{(i)})_{i\in[k]} \in \prod_{i\in[k]} \mathbb{R}^{n_i}$ 为独立向量，且每个向量都满足 Khintchine 不等式（见引理 *2.18*）：

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

对于 $t \geq 1$ 和任意向量 $x \in \mathbb{R}^{d_i}$。

设 $(a_{i_1,\ldots,i_k})_{i_1\in[n_j],\ldots,i_k\in[n_k]}$ 是 $\mathbb{R}^{n_1\times\cdots\times n_k}$ 中的一个张量。那么，

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

对于 $t \geq 1$。

将 $a \in \mathbb{R}^{n_1\times\ldots\times n_k}$ 视为向量，则

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

对于 $t \geq 1$。

证明。下面将对 k 使用归纳法进行证明。

基本情形：对于 $k = 1$，根据向量满足 Khintchine不等式这一假设，结论成立。

归纳情形：假设对于不超过 $k-1$ 的所有值，结论均成立。

Let

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

然后，我们将其从定理左侧的项中提取出来：

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

$$\leq C_t^{k-1}(\sum_{i_1\in[n_1],\ldots,i_{k-1}\in[n_{k-1}]} \|B_{i_1,\ldots,i_{k-1}}^2\|_{L^{t/2}})^{\frac{1}{2}}$$
$$= C_t^{k-1}(\sum_{i_1\in[n_1],\ldots,i_{k-1}\in[n_{k-1}]} \|B_{i_1,\ldots,i_{k-1}}\|_{L^t}^2)^{\frac{1}{2}},$$

其中，第一步由式（7）得出，第二步由归纳假设得出，第三步由 $\|\cdot\|$ 的定义得出，第四步由三角不等式得出，第五步由 $\|\cdot\|$ 的定义得出。

剩下的任务是给出其界

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

by 应用 Khintchine不等式，即可完成归纳步骤，从而完成证明。 □

接下来要用的引理是一种基于基本原理的 Rosenthal 不等式。它以一种精巧的方式结合了随机变量的大阶矩和小阶矩。为求完整起见，我们在此给出证明。

引理 C.5 (具有 t 阶矩的随机变量的性质，引理 20，见 [AKK+20])。 存在一个普适常数 L，使得对于 $t \geq 1$，若 $X_1,\ldots,X_k$ 是具有 t 阶矩的独立非负随机变量，则

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

证明。在以下计算中， L_1、 L_2 和 L_3 均表示普适常数。

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

第一步由对 X_i 进行对称化得到，第二步由 Khintchine不等式（见引理 2.18）得到，第三步由 X_i 的非负性得到，第四步由柯西 - 施瓦茨不等式得到，最后一步由三角不等式得到。

现在，令 A、 B、 C 定义如下:

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

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

and

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

然后，通过改写方程（8），我们得到

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

这意味着 C 小于该二次方程的较大根。
解这个二次不等式可得

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

至此证明完成。 □

C.3 SRCT 和 TensorSRCT 满足强 JL 矩性质

现在我们可以证明 SRCT（参见定义 2.11）和 TensorSRCT （参见定义 2.12）满足强 JL 矩性质。

Th定理 C.6. 存在一个普适常数 L，使得以下结论成立 $s.$
令 $k \in \mathbb{Z}_{>0}$。令 $(D^{(i)})_{i\in[k]} \in \prod_{i\in[k]} \mathbb{R}^{n_i\times n_i}$ 为由独立拉德马赫变量构成的独立对角矩阵。

我们定义 $n := \prod_{i\in[k]} n_i$、$D := D_1 \otimes D_2 \otimes \cdots \otimes D_k \in \mathbb{R}^{n\times n}$ 和 $G := G_1 \otimes \ldots \otimes G_k \in \mathbb{R}^{n\times n}$，其中每个 $G_i \in \mathbb{R}^{n_i\times n_i}$ 都是由独立拉德马赫随机向量生成的循环矩阵。令 $P \in \mathbb{R}^{m\times n}$ 为每行恰有一个非零元素的行采样矩阵。令 $S := PGD$。
令 $x \in \mathbb{R}^n$ 为任意向量，满足 $\|x\|_2 = 1$ 和 $t \ge 1$。
那么，

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

其中 $r = \max\{t, \log\, m\}$。
存在一个普适常数 $C_0 > 1$，使得令

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

我们得到，$\frac{1}{\sqrt{m}}PGD$ 具有 (ϵ,δ)-强 JL 矩性质。

证明。整个证明过程中，C_1、C_2 和 C_3 均表示普适常数。
对于每个 $i \in [m]$，我们令 P_i 为一个随机变量，用来表示P 的第 i行采样了哪些坐标。

我们定义随机变量

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

注意，由于变量 $(P_i)_{i\in[m]}$ 相互独立，因此给定 D后，变量 $(Z_i)_{i\in[m]}$ 条件独立；也就是说，若固定 D，则 $(Z_i)_{i\in[m]}$ 相互独立。
于是，我们可以得到以下不等式：

$$\|\frac{1}{m}\sum_{i\in[m]} Z_i^2 - 1\|_{L^t}$$

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

其中，第一步根据定义C.1，第二步根据引理C.5，第三步根据三角不等式，最后一步根据柯西-施瓦茨不等式。

注意到， G 的每一行都是由独立的拉德马赫随机向量的张量积生成的，因此我们可以将行向量本身视为长度为 n 的拉德马赫随机向量。因此，

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

其中，第一步根据期望值的定义得出， $\mathbb{E}[Z_i^2|D]$；第二步通过展开所有 2^n 种可能性得出；第三步根据简单代数运算得出；最后一步根据 $\|\cdot\|_2^2$ 的定义得出。

我们可以得到

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

其中，第一步根据式（9）得出，第二步根据 $\|x\|_2^2 = 1$ 得出。

为了给 $\|\max_{i\in[m]} Z_i^2\|_{L^t}$ 求界，我们可以证明

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

其中，第一步根据 Z_i 的定义得出，第二步根据 G 的每一行都是
i独立的 Rademacher 向量，因此 $\mathbb{E}[G^\top G] = I$，最后一步由引理 C.4 得出。接下来，我们使用足够高阶的和来界定最大值：

$$
\begin{aligned}
\|\max_{i\in[m]} Z_i^2\|_{L^t} &\le \|\max_{i\in[m]} Z_i^2\|_{L^r} \\
&\le (\sum_{i\in[m]}\|Z_i^2\|_{L^r}^r)^{1/r}
\end{aligned}
$$

$$\leq m^{1/r} r^k \|x\|_2^2 \leq e r^k,$$

第一步由定义 C.1 得出，第二步由 Z_i^2 非负得出，最后一步由 $r \geq \log m$ 得出。

由此可得

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

这就完成了证明的第一部分。

我们希望如下选择 m

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

A根据上述对 m的选择，我们知道 r 满足以下条件

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

因此，

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

对于所有 $1 \leq t \leq \log(1/\delta)$，我们便得到

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

其中第一步由式（10）得出，第二步由 m 的选择得出，第三步由简单代数运算得出，最后一步由 $\epsilon^2 \leq \epsilon$ 和 $t/\log(1/\delta)\sqrt{t/\log(1/\delta)}$ 得出（因为 $t/\log(1/\delta) \in (0, 1)$）。

至此，证明完成。 □

作为两个推论 ，我们得到 SRCT 和 TensorSRCT 是具有 d^2 行而非 d 行的子空间嵌入。

推论 C.7（SRCT 是子空间嵌入）。设 $S \in \mathbb{R}^{m\times n}$ 是一个 SRCT 矩阵，具有 $m = \Theta(\epsilon^{-2}d^2\log^2(n/\epsilon\delta))$，则 S 是一个 (ϵ, δ, d, n)- 子空间嵌入。

证明。 证明结合引理 C.3 和定理 C.6 与 $k = 1$。 □

推论 C.8（TensorSRCT 是一种子空间嵌入）。设 $S \in \mathbb{R}^{m\times n}$ 是一个 TensorSRCT 矩阵，且 $m = \Theta(\epsilon^{-2}d^2\log^3(n/\epsilon\delta))$，则 S是一个$(\epsilon, \delta, d, n^2)$-子空间嵌入。

证明。证明可由引理C.3 和定理C.6 结合 $k = 2$得出。 □

D 高斯与 AMS

在本节中，我们证明随机高斯矩阵和 AMS 矩阵都满足 OCE，且参数 β 良好。结合它们是子空间嵌入这一事实，可以推导出它们的 ℓ_∞ 保证。

D.1 随机高斯和 AMS 的 OSE 性质

OSE 性质　　这两种分布的相关结论早已广为人知。它们的证明可参见，例如 [Woo14]　　.

引理 D.1。 设 S 是定义 2.4 中的随机高斯矩阵。 2.4。若 $m = \Theta(\epsilon^{-2}(d+\log(d/\delta)))$，则 S 是一个 $(\epsilon,\ \delta,\ d,\ n)$-子空间嵌入。

引理 D.2。 设 S 是定义 2.5 中的 AMS 矩阵。 2.5。若 $m = \Theta(\epsilon^{-2}d\log^2(n/\delta))$，则 S 是一个 $(\epsilon,\ \delta,\ d,\ n)$-子空间嵌入。

D.2 随机高斯和 AMS 的 OCE 性质

本节证明随机高斯和 AMS 的 OCE 性质。我们首先给出这两种分布的成对内积界。对于列范数界，AMS 的列是单位列，而我们将证明随机高斯的情形。

引理 D.3 （高斯两两内积界， [SY21]中的引理 B.18）。设 $S \in \mathbb{R}^{m\times n}$ 为一个随机高斯矩阵（定义 2.4）。

于是，我们有：

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

证明。注意，对于 $i \neq j$, $S_{*,\ i}, S_{*,\ j} \sim \mathcal{N}(0, \frac{1}{m}I_m)$ 是两个独立的高斯向量。令 $z_k = S_{k,\ i}S_{k,\ j}$ 和 $z = \langle S_{*,\ i}, S_{*,\ j}\rangle$。

于是，对于任意 $|\lambda| \leq m/2$，我们有

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

其中第一步由 $z_k = \frac{1}{4}(S_{k,\ i} + S_{k,\ j})^2 + \frac{1}{4}(S_{k,\ i} - S_{k,\ j})^2 = \frac{m}{2}(Q_1 - Q_2)$ 得出，其中 Q_1、$Q_2 \sim \chi_1^2$ 和 $\mathbb{E}[e^{\lambda Q}] = \frac{1}{\sqrt{1-2\lambda}}$ 对任意 $Q \sim \chi_1^2$ 均成立。

这意味着 $z_k \in \mathsf{SubExp}(2/m^2, 2/m)$ 是一个次指数随机变量。这里，SubExp 是次指数随机变量的简称。

因此，我们有

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

根据次指数集中引理A.3，我们有

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

对于 $0 < t < 1$。取 $t = \sqrt{\log(n^2/\delta)/m,}$ 我们有

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

Ta对所有 $(i,j) \in [n]\times[n]$ 和 $i \neq j$应用并集界，我们便完成了证明　　f. □

引理 D.4 (AMS 成对内积界)。 设 $S \in \mathbb{R}^{m \times n}$ 是一个 AMS 矩阵（定义 2.5。令 $\{\sigma_i\}_{i \in [n]}$ 为独立拉德马赫随机变量，并且 $\overline{S} \in \mathbb{R}^{m \times n}$ 满足$\overline{S}_{*,\ i} = \sigma_i S_{*,\ i}, \forall i \in [n]$。

于是，我们有:

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

证明。 注意，对于任意固定的 $i \neq j$， $\overline{S}_{*,\ i}$和$\overline{S}_{*,\ j}$都是独立的。根据 Hoeffding 不等式（引理 2.19），我们有

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

其中第二步由简单代数运算（$m \cdot 1/m^2 = 1/m$）可得。

取 $t = \sqrt{2\log(2n^2/\delta)/\sqrt{m}}$，我们有

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

对所有 n^2 列对应用并集界，即可得到所需结果。 □

引理 D.5（高斯列范数界）。设 $S \in \mathbb{R}^{m \times n}$ 为一个随机高斯矩阵。那么，对于任意 $i \in [n]$, 我们有

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

证明。 对于任意一列 $S_{*,\ i}$，注意到$\|S_{*,\ i}\|_2^2 \sim \chi_m^2$，它们的均值均为零，方差为 $\frac{1}{m}$。由引理 2.20，可得

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

令 $t = \log(n/\delta)$，可得

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

对所有 n 列应用并集界，即可完成证明。 □

由此可得，随机高斯和 AMS 均为 OCE。

引理 D.6（高斯 OCE）。设 $S \in \mathbb{R}^{m \times n}$ 为一个随机高斯矩阵，则 S 是一个（$\log^{1.5}(n/\delta)\ \delta, n$）- OCE。

证明。 由引理 D.3 和引理 D.5，我们知道两两内积界和列范数界均成立，因此根据引理 3.4， S 满足所需的 OCE 性质。 □

引理 D.7（AM S OCE）。设 $S \in \mathbb{R}^{m \times n}$ 是一个 AMS 矩阵，则 S 是一个 $(\log^{1.5}(n/\delta)\ \delta, n)$-OCE .

证明。该 p 上界与引理D.6类似。列范数界由定义直接得到 . □

参考文献

[AKK+20] Thomas D Ahle、Michael Kapralov、Jakob BT Knudsen、Rasmus Pagh、Ameya Vel- ingker、David P Woodruff 和 Amir Zandieh。高阶多项式核的无偏草图化。载于第十四届ACM-SIAM离散算法研讨会（SODA）论文集，第141–160页。SIAM，2020年。[AMS96] Noga Alon、Yossi Matias 和 Mario Szegedy。逼近频率矩的空间复杂度。载于第二十八届ACM计算理论研讨会论文集，第20–29页，1996年。[ANW14] Haim Avron、Huy Nguyen 和 David Woodruff。多项式核的子空间嵌入。载于 Z. Ghahramani、M. Welling、C. Cortes、N. D. Lawrence 和 K. Q. Weinberger（编），《神经信息处理系统进展》第27卷，第2258–2266页。2014年。[BPSW21] Jan van den Brand、Binghui Peng、Zhao Song 和 Omri Weinstein。以近线性时间训练（过度参数化）神经网络。载于 ITCS，2021年。[CCFC02] Moses Charikar、Kevin Chen 和 Martin Farach-Colton。寻找数据流中的高频项。载于自动机、语言与编程国际研讨会（ICALP），第693–703页。施普林格，2002年。[Coh16] Michael B Cohen。通过迹不等式构造近乎紧的无偏子空间嵌入。载于第二十七届ACM-SIAM离散算法研讨会论文集，第278–287页。SIAM，2016年。[CSWZ23] Yeshwanth Cherapanamjeri、Sandeep Silwal、David P. Woodruff 和 Samson Zhou。在当前矩阵乘法时间内实现线性代数的最优算法。载于2023年ACM-SIAM离散算法研讨会（SODA）论文集，2023年。[CW13] Kenneth L. Clarkson 和 David P. Woodruff。输入稀疏度时间内的低秩近似与回归。载于计算理论研讨会会议（STOC），第81–90页，2013年。[DJS+19] Huaian Diao、Rajesh Jayaram、Zhao Song、Wen Sun 和 David Woodruff。克罗内克积回归与低秩近似的最优草图化。《神经信息处理系统进展》，第32卷，2019年。[DSSW18] Huaian Diao、Zhao Song、Wen Sun 和 David Woodruff。克罗内克积回归与p样条的草图化。载于人工智能与统计国际会议，第1299–1308页。PMLR，2018年。[FFG22] Matthew Fahrbach、Thomas Fu 和 Mehrdad Ghadiri。带有张量分解应用的次二次克罗内克回归。载于第三十六届神经信息处理系统大会，NeurIPS’22，2022年。[FKZ11] Sergey Foss、Dmitry Korshunov 和 Stan Zachary。《重尾分布与次指数分布引论》。施普林格运筹学丛书。施普林格，纽约州纽约市，第一版，2011年。

[Hoe63] Wassily Hoeffding. 有界随机变量之和的概率不等式。载于《美国统计学杂志》，第 13–30 页。1963年。[HW71] David Lee Hanson 和 Farroll Tim Wright. 独立随机变量二次型尾概率的一个界。《数学统计年刊》，42(3):1079–1083，1971年。[JSWZ21] Shunhua Jiang、Zhao Song、Omri Weinstein 和 Hengjie Zhang. 求解一般线性规划的一种更快算法。载于第 53 届 ACM SIGACT计算理论研讨会论文集，第 823–832 页，2021年。.[Khi23] Aleksandr Khintchine. 关于二进分数。《数学杂志》，18(1):109–116，1923年。[LDFU13] Yichao Lu、Paramveer Dhillon、Dean P Foster 和 Lyle Ungar. 通过子采样随机哈达玛变换实现更快的岭回归。载于《神经信息处理系统进展》（NIPS），第 369–377 页，2013年。[LM00] Beatrice Laurent 和 Pascal Massart. 通过模型选择对二次泛函进行自适应估计。《统计年刊》，第 1302–1338 页，2000年。[LS14] Yin Tat Lee 和 Aaron Sidford. 线性规划的路径搜索方法：在 $O(\sqrt{rank})$次迭代内求解线性规划以及更快的最大流算法。载于第 55 届 IEEE计算机科学基础研讨会（FOCS），第 424–433 页，2014年。[LSZ19] Yin Tat Lee、Zhao Song 和 Qiuyi Zhang. 在当前矩阵乘法时间内求解经验风险最小化问题。载于学习理论会议，第 2140–2157 页。PMLR，2019年。[NN13] Jelani Nelson 和 Huy L Nguyˆen. OSNAP：通过更稀疏的子空间嵌入实现更快的数值线性代数算法。载于 2013年 IEEE 第 54 届计算机科学基础年度研讨会（FOCS），第 117–126 页。IEEE，2013年。[PSW17] Eric Price、Zhao Song 和 David P Woodruff. 具有 ℓ_∞ 保证的快速回归。载于 ICALP，2017 年。[QSZZ23] Lianke Qin 、 Zhao Song 、 Lichen Zhang 和 Danyang Zhuo. 一种用于投影矩阵向量乘法的在线统一算法及其在经验风险最小化中的应用。载于 AISTATS，2023年。[RSZ22] Aravind Reddy、Zhao Song 和 Lichen Zhang. 动态张量积回归。载于第三十六届神经信息处理系统大会（NeurIPS'22），2022年。[Sar06] Tamas Sarlos. 通过随机投影改进大型矩阵的近似算法。载于 2006年 IEEE 第 47 届计算机科学基础年度研讨会（FOCS），第 143–152 页。IEEE，2006年。[SWYZ21] Zhao Song、David Woodruff、Zheng Yu 和 Lichen Zhang. 多项式次数多项式核的快速草图化。载于国际机器学习会议，第 9812–9823 页。PMLR，2021年。

[SWZ16] Zhao Song、David Woodruff 和 Huan Zhang. 次线性时间正交张量分解。《神经信息处理系统进展》，29，2016年。[SWZ19] Zhao Song、David P Woodruff 和 Peilin Zhong. 张量低秩近似的相对误差。载于第三十届 ACM-SIAM 离散算法年度研讨会论文集，第 2772–2789 页。SIAM，2019年。[SXYZ22] Zhao Song、Zhaozhuo Xu、Yuanyuan Yang 和 Lichen Zhang. 利用低维自适应数据结构加速 Frank-Wolfe 算法。arXiv 预印本 arXiv:2207.09002，2022年。[SXZ22] Zhao Song、Zhaozhuo Xu 和 Lichen Zhang. 利用内积搜索数据结构加速稀疏化。arXiv 预印本 arXiv:2204.03209，2022年。[SY21] Zhao Song 和 Zheng Yu. 用于求解线性规划问题的基于 oblivious sketching 的中心路径方法。载于第 38 届国际机器学习会议（ICML），2021年。[SZZ21] Zhao Song、Lichen Zhang 和 Ruizhe Zhang. 在次二次时间内训练多层过参数化神经网络。arXiv 预印本 arXiv:2112.07628，2021年。[Tro10] Joel A. Tropp. 子采样随机哈达玛变换的改进分析。2010年。[Woo14] David P Woodruff. 草图化作为数值线性代数工具。《理论计算机科学基础与趋势》®，10(1–2):1–157，2014年。[WZ20] David P Woodruff 和 Amir Zandieh. 通过自适应采样实现近输入稀疏时间的核嵌入。载于 ICML，2020年。[WZ22] David Woodruff 和 Amir Zandieh. 输入稀疏度时间内张量积矩阵的杠杆分数采样。载于第 39 届国际机器学习会议论文集，《机器学习研究论文集》。PMLR，2022年。