arXiv:2304.02261v1 [cs.DS] 5 Apr 2023

O稀疏线性回归的最优草图界限

Tung Mai* Alexander Munteanu† Cameron Musco‡ AnupB. Rao§
Chris Schwiegelshohn¶ David P. Woodruff‖

2023年4月6日

摘要

我们研究了在各种损失函数下针对 k-稀疏线性回归的无关草图化。特别地，我们关注的是一个关于草图化矩阵 $S \in \mathbb{R}^{m\times n}$ 的分布，该分布不依赖于输入 $A \in \mathbb{R}^{n\times d}$ 和 $b \in \mathbb{R}^n$，使得在能够访问 SA 和 Sb的情况下，我们可以恢复一个 k-稀疏的 $\tilde{x} \in \mathbb{R}^d$，并满足 $\|A\tilde{x} - b\|_f \leq (1+\varepsilon)\min_{k\text{-sparse } x\in\mathbb{R}^d} \|Ax - b\|_f$。其中，$\|\cdot\|_f\colon \mathbb{R}^n \to \mathbb{R}$是某种损失函数——例如某个 ℓ_p 范数，或来自一个广泛的类铰链损失函数，这些函数包括 logistic 损失和 ReLU 损失。

我们证明，对于稀疏 ℓ_2 范数回归，存在一种关于无关草图的分布，其行数为 $m = \Theta(k\log(d/k)/\varepsilon^2)$，并且这一结果在常数因子范围内是紧的。该结论可推广到 ℓ_p损失，但上界中需额外增加一项 $O(k\log(k/\varepsilon)/\varepsilon^2)$。这与相关的稀疏恢复问题形成了出人意料的区别；稀疏恢复是稀疏回归的一个重要特例，其中 A 是单位矩阵。对于这一问题，在 ℓ_2 范数下，我们得到上界 $m = O(k\log(d)/\varepsilon + k\log(k/\varepsilon)/\varepsilon^2)$，表明与稀疏回归相比，稀疏恢复的草图构造严格更容易。

对于采用类铰链损失函数的稀疏回归（包括稀疏逻辑回归和稀疏 ReLU回归），我们给出了首个已知的草图界，达到 $m = o(d)$，表明只需$m = O(\mu^2 k\log(\mu nd/\varepsilon)/\varepsilon^2)$ 行，其中 μ 是为这些损失函数获得相对误差界所需的自然复杂度参数。我们再次证明，除低阶项以及对 μ 的依赖外，该维度是紧的。

最后，我们证明 LASSO 回归也能达到类似的草图界。LASSO 回归是稀疏回归的一种常见凸松弛，其目标是最小化 $\|Ax - b\|_2^2 + \lambda\|x\|_1$，使其满足 $x \in \mathbb{R}^d$。我们证明，草图维度 $m = O(\log(d)/(\lambda\varepsilon)^2)$ 已足够，而且对 d 和 λ 的依赖是紧的。

*Adobe 研究院，tumai@adobe.com
†多特蒙德工业大学，alexander.munteanu@tu-dortmund.de
‡马萨诸塞大学阿默斯特分校，cmusco@cs.umass.edu §Adobe 研究院，anuprao@adobe.com
¶奥胡斯大学，cschwiegelshohn@gmail.com
‖卡内基梅隆大学，dwoodruf@cs.cmu.edu

目录

1 引言 11.1 相关工作.2

2 我们的技术与结果3

2.1 下界.42.2 上界.82.3 LASSO 回归.112.4 稀疏恢复与稀疏回归的区分.11**3 结论** 12**A 稀疏回归的下界** k**-稀疏回归** 17A.1 针对 ℓ_p-范数损失函数的 $p \geq 1$ .17A.2 ReLU 和类铰链损失函数的下界.23A.3 采样不适用于稀疏回归.24

B 上界 s的 k-稀疏回归 2 5

B.1 下界forthe ℓ_p-范数损失函数 for $p \in [1, 2]$ .25B.2 ReLU 损失函数的上界.28B.3 类铰链损失函数的上界.29**C LASSO 回归的无关草图** 33C.1 上界.33C.2 下界.34**D 区分稀疏恢复与稀疏回归** 35**E 信息 t**

理论基础 3 6

1 引言

我们研究针对 k-稀疏回归的无偏草图方法。给定数据矩阵 $A \in \mathbb{R}^{n\times d}$ 和目标向量 $b \in \mathbb{R}^n$，线性回归问题旨在寻找一个向量 $x \in \mathbb{R}^d$，使得 $Ax \approx b$。Ax 与 b 之间的偏差通过损失函数 $f(Ax, b)$ 来量化，常见的例子包括基于 ℓ_p 范数的损失、逻辑损失和 ReLU。这些问题的草图技术已得到广泛且成功的应用。具体而言，我们从某个分布中抽取一个草图矩阵 $S \in \mathbb{R}^{m\times n}$，并尝试在 SA 和 Sb 上求解问题。如果 S 可以从一个不依赖于 A 或 b 的分布中抽取，我们就称该草图为无偏草图。我们的目标是在保留仅使用较小草图而非原始数据提取 $(1+\varepsilon)$-近似解 $\tilde{x}$ 能力的同时，尽量减小目标维度 m，该维度称为草图复杂度或草图维度。模型参数 x 的一个特别理想的性质是稀疏性，即 x 被限制为至多包含 k 个非零元素。稀疏线性回归是处理超高维数据集的重要技术，例如生物统计学中出现的数据集 [Mba+21]。它生成的线性模型仅依赖少量参数，因此更易于解释，并且能够利用相对较小的数据样本进行准确学习。从计算角度看，施加稀疏性约束会使问题显著变难。大多数无约束回归问题都是凸问题，因此可以利用多种基于梯度的方法；但稀疏线性回归是 NP-难问题 [Nat95]。人们已开发出许多启发式方法和松弛方法来实际求解该问题，[参见 Tib96; Mil02; DK08; DK18]。这些方法的运行时间通常至少会随观测数量 n 多项式增长。因此，研究减少数据行数的方法，从而提高计算效率并保留近似最优解，是很自然的思路。

对于相关工作中涵盖的大多数所提及启发式方法和近似算法（参见第 1.1 节），我们的结果可立即带来计算节省，但潜在的加速效果需要针对具体情况进行验证，并与将草图应用于数据的成本进行对比。此外，草图化还有其他动机，例如节省内存、处理（turnstile）数据流 [CW09], 以及聚合分布式数据 [KVW14]。关于简要介绍，请参阅 [Mun23] 。需要强调的是，我们的界仅涉及缩减后的草图维度 m，而不涉及其他复杂度度量。

虽然对于无约束回归问题，草图技术的性能已有充分理解，但我们对稀疏回归问题的复杂性知之甚少。显然，无约束回归的保证同样适用于稀疏回归，但这些界是否紧确实并不清楚。特别地，稀疏恢复这一特殊情形中有许多著名结果，它们利用稀疏性，仅通过少量测量就能重构目标信号。因此，我们提出如下问题：

对于各种回归问题，无感知草图技术能否从模型稀疏性中获益？

据我们所知，以上问题此前尚未得到研究。本文对一大类损失函数肯定地回答了这个问题，其中包括 ℓ_p 损失$\|Ax-b\|_p = \sqrt[p]{\sum_{i=1}^n |A_ix-b_i|^p}$ 对于 $p \in [1,2]$、逻辑损失$\sum_{i=1}^n \ln(1+\exp(-b_i \cdot A_ix))$、ReLU$\sum_{i=1}^n \max(0, -b_i \cdot A_ix)$以及其他类铰链损失函数 f，其中我们定义 $\|Ax-b\|_f = \sum_{i=1}^n f(A_ix-b_i)$。我们还研究了用于施加稀疏性的常见启发式方法的草图复杂度，例如 LASSO 回归 [Tib96], ，该方法不是严格要求 x 为 k-稀疏，而是使用 ℓ_1 惩罚目标函数 $\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2 + \lambda\|x\|_1$ 来求得稀疏解。对于所有这些问题，我们都得到了最优或近乎最优的草图界限。关键在于，对于无约束回归，界限对 d 的依赖至少是线性的；而对于

稀疏回归则仅呈对数级依赖。因此，本文为利用稀疏性的优势来对回归问题进行草图化迈出了重要一步。

令 $\Psi_k = \{x \in \mathbb{R}^d \mid \|x\|_0 \leq k\}$ 表示 $\mathbb{R}^d$ 中所有 k-稀疏向量构成的集合。我们还研究相关的 k-稀疏仿射嵌入问题：给定草图和估计器 $E(SA,\ Sb,\ x)$，我们要求

$$\forall x \in \Psi_k : (1-\varepsilon)\|Ax-b\|_f \leq E(SA, Sb, x) \leq (1+\varepsilon)\|Ax-b\|_f.$$

k-稀疏仿射嵌入比稀疏线性回归的草图要求更强，因为它蕴含后者。它是压缩感知领域中研究的受限等距性质（RIP）的一种推广。RIP 是一种特殊情形，其中 $A = I$ 是单位矩阵，$b = 0$，并且我们希望将任意 k-稀疏 $x \in \Psi_k$ 的范数保持在 $1 \pm \varepsilon$ 失真范围内。

ℓ_2 稀疏恢复问题也可以看作稀疏 ℓ_2 回归的一种特殊情形，其中 $A = I$，且我们只能访问向量 b 的草图 Sb。目标是恢复一个 k-稀疏向量 $\tilde{x}$，使其与最小化的 x 之间的误差不超过 $(1+\varepsilon)$，即满足 $\|\tilde{x}-b\|_2 \leq (1+\varepsilon)\min_{x\in\Psi_k}\|x-b\|_2$.。稀疏回归的草图方法似乎与稀疏恢复非常相似，因为两类问题都有类似的方法可用，并能得到类似的上界。然而，我们的研究揭示了一个出人意料的分离结果。也就是说，稀疏回归问题的草图难度严格高于稀疏恢复。

除了无关线性草图之外，人们或许会疑惑，为什么我们不考虑数据相关的采样算法，因为采样技术是在非稀疏情形下近似回归问题的重要工具。这是因为可以观察到（见定理5 和6）：在稀疏回归问题中，采样并无帮助；要获得任何非平凡的界，大致必须对整个输入进行采样。

与采样相比，草图技术的关键优势似乎在于其对需要嵌入的子空间具有无关性，这使我们能够对所有 k 个坐标子集应用并集界。然而，对于每种可能情况，采样算法都需要采用不同的测度。尽管依赖数据的重要性采样技术在无约束的非稀疏回归问题中广泛奏效，但在稀疏情形下却无法给出任何非平凡的界。这凸显了无关草图技术在稀疏问题中的重要性。

1.1 相关工作

k-稀疏回归问题的一个重要特例是压缩感知 [CRT06; Don06; Bar+08]，其中矩阵 A 是 $n \times n$ 单位矩阵，目标是找到一个稀疏向量 x，使其通过线性草图（或感知）矩阵能够良好地表示非稀疏信号 b。众所周知，该问题所需的矩阵至少需要 $m = \Omega(k\log(d/k))$ 行 [Ba+10]；Price 和 Woodruff [PW11] 将其改进为 $m = \Omega(k\log(d/k)/\varepsilon + k/\varepsilon^2)$。至于上界，已知可构造一个具有 $m = O(k\log(d/k))$ 行的（高斯）RIP 矩阵，这足以在常数 ε 下求解该问题。近期出现了针对高斯构造的更新且更简洁的证明 [LXG20]，这些证明隐含地给出了 $m = \Theta(k\log(d/k)/\varepsilon^2)$。然而，要解决稀疏恢复问题，既不需要高斯矩阵，也不需要 RIP。例如，当允许稀疏度为 $2k$ 时，对于任意特定输入，都可以使用 CountSketch[CCF04] 以常数概率求解该问题，所需行数为 $m = O(k\log(d/k)/\varepsilon)$ [PW11]。

据我们所知，此前尚未研究过将无视数据的线性草图推广到稀疏线性回归。不过，关于稀疏线性列选择问题已有不少研究 ℓ_2。稀疏回归问题，即在所有 k-稀疏向量中最小化回归代价，是 NP-困难的 [Nat95],，并且在合理的复杂性理论假设下，它

甚至难以在准多项式时间内近似到比平凡界强得多的界 $\|b\|_2$ [FKT15]。鉴于此，一些作者找出了并刻画了这样一类实例：广泛使用的启发式方法在这些实例上具有性能保证 [DK08; DK18; GMS03; Tro06a; Tro06b; TGS06]。另一个更贴近所用启发式方法的研究方向，是列选择问题的在线版本。 [FKT15]证明，若某个在线算法的每次迭代都能在多项式时间内运行，则可推出 $NP \subseteq BPP$，即使允许列数增加 $O(\log d)$倍。鉴于这些不可能性结果，研究者转而关注低效算法。 [FKT15] 提出的在线算法每次迭代大约需要 $O(k\binom{d}{k})$时间。Har-Peled、Indyk 和 Mahabadi [HIM18] 通过构造一种数据结构，将查询到由输入点张成的最近的 $(k-1)$ 维平面的几何距离近似到 $(1+\varepsilon)$ 倍，从而将每次迭代的时间降至约 $\tilde{O}(d^{k-1})$；这一方法利用了稀疏线性回归的几何解释。同一文献还通过从 k-SUM 问题归约，证明任何乘法误差近似都存在约为 $\tilde{\Omega}(d^{k/2}/e^k)$ 的不可能性结果。假设满足 RIP 性质， [Kal+17]给出了一种具有保证的高效在线算法。

实践中用于求解稀疏回归的、可以说最常用的启发式方法，是Tibshirani提出的最小绝对收缩与选择算子（LASSO） [Tib96]。它引入了 ℓ_0 约束的凸松弛，将其替换为 ℓ_1 约束。对于性质良好的矩阵，已有研究表明 A，LASSO 算法能够恢复一个（近乎最优的）稀疏解 [CRT06; Don06]。所得的拉格朗日形式为$\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2+\lambda\|x\|_1$。由于 $\lambda\|x\|_1$ 正则化项非负，因此显然， ℓ_2-子空间嵌入足以在任意 $x\in\mathbb{R}^d$ 内保持（$1\pm\varepsilon$）乘法误差，其中需要 $m=\Theta(d/\varepsilon^2)$ 行 [Sar06; NN14]。据我们所知，目前尚无通过利用正则化项所诱导的稀疏性、使用更少行数对该目标进行草图化的研究结果。

对于稠密线性 ℓ_p 回归，已有大量草图研究成果。 ℓ_2、 [Sar06]首先表明， $m=O(d/\varepsilon)$行足以将最小化解保持在 $1+\varepsilon$ 误差以内。随后， [CW09]给出了与之匹配的下界。通过无关的 ℓ_p 子空间嵌入和采样，将结果扩展到 ℓ_p 回归的工作见于 [Cla05; Das+09; SW11; WZ13; CW15; WW19; LWY21]。近期工作 [MOW21; MOW23]首次针对逻辑损失和（隐含地）ReLU函数给出了无关的线性草图。 [Mun+18]开发了针对这些广义线性回归问题的重要性采样算法，随后由 [MMR21; MOP22; WY23]进一步改进和推广。

2 我们的技术与结果

我们的结果总结于表1 中，涵盖了缩减后的草图维度的界 m，而不涉及计算成本等其他复杂度度量。我们的上界针对仿射嵌入，因此使用我们的草图的算法在整个搜索空间上都享有近似界。这直接意味着相匹配的最小化上界。除类铰链损失外，我们的下界针对更具挑战性的最小化变体。对于稀疏 ℓ_2 回归，我们的界是紧的。推广到 ℓ_p 时，界与一个加性 $O(k\log(k/\varepsilon)/\varepsilon^2$项相差至多如此；具体而言，这意味着对于合理较小的 $k=O(\sqrt{\varepsilon}d$，我们的结果是紧的。ReLU的上界还包含另一个 $\tilde{O}(\mu^2)$因子，此外还存在 ℓ_1中所报告的差距。这里， μ 是一个自然参数，用于刻画将这些损失下的数据压缩至低于$\Omega(n)$所需的复杂度 [Mun+18]。对于类铰链损失函数（包括逻辑损失）同样如此，但会增加另一个加性$\log n$；相应的下界略弱，因为它针对的是 k-稀疏仿射嵌入，而不是最小化问题。

任何子采样方法对于任意损失函数都有相应的 $\Theta(n)$ 界。下界为

表 1：结果总结。此处 μ 是一个数据依赖的复杂度参数 [Mun+18]，且 λ是 LASSO 的正则化参数，详情请参阅正文。

损失函数 / 类型	**界限**	**参考文献**
ℓ_2	$\Theta(k$ 对数$(d/k)/\varepsilon^2)$	定理1 和7
$\ell_p\ p \in [1,2)$		
下界	$\Omega(k\log(d/k)/\varepsilon^2)$	定理 2
上界	$O(k\log(d/k)/\varepsilon^2 + k\log(k/\varepsilon)/\varepsilon^2)$	定理 8
ReLU & 铰链型		
下界	$\Omega(k\log(d/k)/\varepsilon^2)$	推论3 和 4
ReLU 的上界	$O(\mu^2 k\log(\mu d/\varepsilon)/\varepsilon^2)$	定理 9
铰链型的上界	$O(\mu^2 k\log(\mu n d/\varepsilon)/\varepsilon^2)$	定理 10
通过采样处理任意损失	$\Theta(n)$	定理5 和6
LASSO 与 $\lambda\|x\|_1$		
下界	$\Omega(\log(\lambda d)/\lambda^2)$	定理13
上界	$O(\log(d)/(\lambda\varepsilon)^2)$	定理 12
ℓ_2 稀疏恢复		
下界	$\Omega(k\log(d/k)/\varepsilon + k/\varepsilon^2)$	[PW11]
上界	$O(k\log(d)/\varepsilon + k\log(k/\varepsilon)/\varepsilon^2)$	定理 14

该下界是在最小化设定下，以及对数据具有不同程度无关性的子空间嵌入情形中给出的。对于正则化参数为 λ的 LASSO，我们的上界为 $m = O(\log(d)/(\lambda\varepsilon)^2)$，并给出了相应的下界 $m = \Omega(\log(\lambda d)/\lambda^2)$。最后，对于 ℓ_2稀疏恢复，我们的上界与已知最佳下界之间仅存在一个很小的加性差距 $O(k\log(k/\varepsilon))$。更有意思的是，该界足以将其与严格更难的 稀疏回归问题区分开来。总之，对于一般参数化，我们的界在低阶（加性）项意义下是紧的；当 k/ε取值相当小时，这些界也是紧的。

在接下来的部分中，我们将介绍主要结果，以及这些结果证明背后的主要思路和技术挑战。为使呈现更加简洁，正式证明和细节分别放在相应的附录中。

2.1 下界

我们通过给出一系列归约来获得下界。我们针对稀疏ℓ_2 回归的主要下界（后续的其他下界均由此推导）是通过从稀疏 PCA（sPCA）的近似（常数比例）支撑集恢复问题进行归约而得到的。

我们注意到， [BPP18], 中给出了一种相关的归约，而近似 sPCA 的困难性则在 [CMW13]中讨论。然而，先前这些工作的结合并不能为我们的问题得出所需的困难性结果，原因如下。它们的归约要求 d 精确求解稀疏回归，其中 每个 列都要对所有其余列进行回归。通过将投影到最优稀疏子空间上的范数与某个固定阈值进行比较，来判断第 i 次迭代中的列是否属于稀疏支撑集。主要问题在于，从 sPCA 归约到稀疏回归时，它们只能得到 k^2 级依赖，而不是 k，而其分析需要后者才能区分列是否属于支撑集。目前尚不清楚如何用随机化决策替代这些步骤，同时又不进一步增大对k 或其他参数的依赖。此外，在每次迭代中引入随机化只会得到

针对 $1/d$ 误差 概率的下界，而非我们针对常数概率给出的下界。

为了证明针对常数误差的最优下界，我们通过仅求解一个单独的回归问题给出一种归约。先前工作中的论证和支撑集构造 [AW09；CMW13] 并不直接适用，原因如下：我们将未知支撑集上的信息植入一个额外的列中，并对标准列进行回归。该额外信息的权重足够高，因此我们可以恢复支撑集的（常数比例）部分。

关键在于，权重也足够低，因此支持恢复问题仍然难以求解。但这需要使用我们的方法重新证明，进而说明稀疏回归问题的求解难度。我们注意到，完全恢复具有$\Omega(k\log(d-k))$ 下界[AW09], ，这大于我们对稀疏回归问题给出的上界。因此，我们借助纠错码进行一种松弛，以近似恢复固定比例的支持集；这要求我们为简化的问题证明一个新的下界，其中额外的植入信息被提供给算法。

我们以 [AW09] 为起点。作者研究了 尖峰协方差模型下的稀疏 PCA（sPCA）：我们从协方差矩阵为（$I_d + vv^T$）的高斯分布中获取测量值，并且 v是一个 k-稀疏向量。在这里，寻找最大特征向量意味着：给定从高斯分布中抽取的向量，我们需要找到向量 v。由于所有非零元素的取值都相同，这就转化为寻找 k-稀疏支撑集 v。作者通过信息论论证表明，基于少量测量值，该问题以较高概率无法求解。这里，我们借鉴其总体思路，并为调整后的稀疏 PCA 变体证明新的困难性界。

稀疏 ℓ_2 回归 我们概述主定理的证明。为使表述简洁，省略若干技术细节。详细的技术推导见附录。

定理 1。 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。假设 $S \in \mathbb{R}^{m\times n}$ 是用于 k-稀疏 ℓ_2回归的无信息线性草图，配有任意估计器 $E(SA, Sb, x)$，使得 $\tilde{x} \in \operatorname{argmin}_{x\in\Psi_k} E(SA, Sb, x)$ 以常数概率满足 $\|A\tilde{x} - b\|_2 \le (1+\varepsilon)\min_{x\in\Psi_k} \|Ax - b\|_2$。那么 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

我们的证明结构如下：首先，我们在 k- 稀疏支撑集上构造一个合适的（困难）分布，并用它来定义输入分布。其次，我们证明，对于来自该输入分布的样本，若测量（行）的数量低于信息论下界，则不可能恢复支撑集中的一个常数比例。第三，我们构造一个 ℓ_2- 回归实例，使得任何由无偏草图得到的 $1 + \Theta(\varepsilon)$ 近似，若与任意估计器配合使用，都会揭示支撑集中的一个常数比例。因此，该困难性结果也适用于回归问题。

为构造困难分布（证明的第一步），我们构造一个纠错码$\mathcal{C}$，其规模约为 $|\mathcal{C}| = \binom{d}{k}^k$，由 k- 稀疏二进制向量组成，且任意两个向量的重叠至多为 ck 个索引。该码还满足如下意义下的精确平衡性：每个索引 i 出现在码字中的次数都与其他任意索引 i' 完全相同，即约为 $\binom{d}{k}^{k-1} = \frac{k}{d}|\mathcal{C}|$ 次；并且对于 $i \ne j$，每一对 (i, j) 出现的次数都与其他任意一对 (i', j')（其中 $i' \ne j'$）完全相同，即约为 $\binom{d}{k}^{k-2} = \frac{k^2}{d^2}|\mathcal{C}|$ 次。我们再为每个码字添加另一个坐标（不失一般性地取第一个坐标），并将其固定为 1。

我们均匀随机地选取一个码字 $c \in \mathcal{C}$，并令输入 $n \times d$ 的分布为 $Z = [b, A] = G(I_d + vv^T)^{1/2}$，其中 G 是一个随机高斯矩阵，即每个 $G_{ij} \overset{i.i.d.}{\sim} N(0, 1)$，并且

对于所有 $j \in [d]\backslash\{1\}$，我们设$v_j = \frac{\varepsilon}{\sqrt{k}} c_j$；而对于第一个增广坐标，我们设 $v_1 = c_1 = 1$。至此，困难输入分布以及草图处理之前的情形介绍完毕。

现在，由于我们证明的是针对任意估计器的下界，而这类估计器例如可以任意但适当地改变基底并进行缩放，因此不失一般性，我们可以假设草图矩阵 $S \in \mathbb{R}^{m\times n}$ 的行是正交归一的。于是，草图具有如下形式。

$$SZ = SG(I_d + vv^T)^{1/2} = H(I_d + vv^T)^{1/2},$$

其中， H 仍然是一个高斯矩阵。因此， SZ 与输入矩阵 Z 具有相同的高斯分布，但行数从 n 减少到 $m < n$。

我们进入第二个证明步骤，即证明：对于上述输入分布，仅使用少量测量（行）不可能恢复支持集中的一个常数比例。为此，令 X 为一个随机变量，其分布与我们的草图的一行相同。我们使用 Fano不等式，根据编码 $\mathcal{C}$ 的大小以及互信息来界定失败概率，其中互信息量化了草图的各行（记为 X^m）揭示未知支持集 U 中关于 $c \in \mathcal{C}$的信息量：

$$\mathbb{P}[\texttt{error}] \geq 1 - \frac{I(U; X^m) + \log 2}{\log(|\mathcal{C}| - 1)}.$$

为了界定互信息，我们根据熵的链式法则和高斯分布的最大熵性质 [CT06] 注意到，互信息可以用行列式的对数来界定：

$$\begin{aligned} I(U, X^m) &= H(X^m) - H(X^m|U) \\ &\leq \frac{m}{2}\log\det\mathbb{E}[xx^T] - \frac{m}{2}\log\det\mathbb{E}[xx^T|U] \end{aligned}$$

利用我们代码构造的平衡结构，所涉及的矩阵具有良好的分块结构，我们利用这一点将互信息界定为 $O(\varepsilon^2 m)$。进一步注意到，对数代码大小满足 $\log(|\mathcal{C}|) = O(k\log(d/k))$。将其代回 Fano 不等式，我们得到一个下界 $m = \Omega(k\log(d/k)/\varepsilon^2)$，适用于任何具有常数误差概率的支持恢复算法，这就完成了证明的第二步。形式化细节请参见附录。

最后，我们介绍如何归约到稀疏 ℓ_2 回归问题，也就是证明的第三步，也是最后一步。给定如上所述的 $Z = [b, A]$，我们考虑以下实例：

$$\min_{x\in\Psi_k} \left\| \begin{bmatrix} M\, M\, \ldots\, M \\ A \end{bmatrix} x - \begin{bmatrix} \sqrt{k}M \\ b \end{bmatrix} \right\|_2,$$

其中， M 足够大，从而确保$\sum_i x_i$ 接近 $\sqrt{k}$。特别是，这对于避免平凡解 $x = 0$是必要的；更准确地说，是为了施加 $\|x\|_2 \geq \frac{1}{\sqrt{k}}\|x\|_1 \geq \frac{1}{\sqrt{k}}\sum_i x_i \approx 1$。由于我们证明的是针对任意估计器的下界，因此可以假设该估计器已知任意草图矩阵 $S' = [s_1, S]$ 的第一列，以及附加行 $r_1 = [M, M, \ldots, \sqrt{k}M]$的结构，包括 M的值。这样，估计器便能消除张量积 $s_1 \cdot r_1$的影响，从而只需对 $SZ = S[b, A]$ 进行估计。最后，我们证明，如果将上述问题（在草图上）求解到因子 $1 + \Theta(\varepsilon)$ 以内，那么所得解 $\tilde{x}$ 的支撑集与随机码字 $c \in \mathcal{C}$的实际支撑集 U 有一个常数比例的重合。由于纠错码的构造，这可以唯一确定完整的 未知支撑集 U，从而完成证明的第三步。由此，我们得到了针对常数误差概率的稀疏ℓ_2 回归无偏草图算法的$m = \Omega(k\log(d/k)/\varepsilon^2)$下界。

稀疏 ℓ_p 回归 我们的下一个目标是将 $m = \Omega(k\log(d/k)/\varepsilon^2)$结果推广到任意 $p \geq 1$下的 ℓ_p 范数回归最小化问题。

定理2. 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。假设 $S \in \mathbb{R}^{m\times n}$ 是针对任意 $p \geq 1$ 的 k-稀疏 ℓ_p 回归的无偏线性草图，并具有估计器 $E_p(SA, Sb, x)$，使得极小化解$\tilde{x} \in \operatorname{argmin}_{x\in\Psi_k} E_p(SA, Sb, x)$ 满足 $\|A\tilde{x} - b\|_p \leq (1+\varepsilon)\min_{x\in\Psi_k}\|Ax - b\|_p$。那么 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

我们先从 ℓ_1 开始，并在下文讨论一般情形 $p \in [1,\infty)$。根据 Dvoretzky 定理，我们可以使用随机高斯映射 G 将 ℓ_2 嵌入 ℓ_1，失真为 $1\pm\varepsilon$。更准确地说，对于所有 k-稀疏的 x，都有 $\|Ax-b\|_2 = (1\pm\varepsilon)\|GAx - Gb\|_1$，其中 G 有 $O(n\log(1/\varepsilon)/\varepsilon^2)$ 行，且由独立同分布的高斯随机变量组成。特别地，它是一种无偏嵌入。现在假设我们有一个用于 k-稀疏 ℓ_1 回归问题的草图 S。那么可以证明，$\|SGAx - SGb\|_1$ 的极小化解 $\tilde{x}$ 也是 $\|Ax-b\|_2$ 的一个 $1+O(\varepsilon)$ 近似解。这意味着 SG 是用于 ℓ_2-范数问题的无偏草图（配有一个 ℓ_1-范数估计器），因此 S 至少需要 $m = \Omega(k\log(d/k)/\varepsilon^2)$ 行。

类似地，ℓ_2 可通过随机高斯矩阵 G，对所有 $p \geq 1$ 进行无偏嵌入，并将误差控制在 $1\pm\varepsilon$ 以内，其中行数取决于 $O(n)$ ——对于 $1 \leq p \leq 2$ 而言——以及 $n^{O(p)}$ ——对于 $p > 2$而言；[参见 Mat13，第 30]，页。不过，在我们的情形中，G 的行数并不重要，因为它会被草图 S压缩。因此，对于每个 $p \geq 1$，ℓ_p都满足 $\Omega(k\log(d/k)/\varepsilon^2)$ 下界。

稀疏 ReLU 与铰链型损失回归 我们进一步将下界扩展到 ReLU 损失函数。这次，我们通过设计从 ℓ_1 到 $\|\cdot\|_{\mathrm{ReLU}}$的精确嵌入，将问题归约为稀疏 ℓ_1 回归问题。

推论 3. 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。假设 $S \in \mathbb{R}^{m\times n}$ 是用于 k-稀疏ReLU 回归的无偏线性草图，其估计器为 $E_{\mathrm{ReLU}}(SA, Sb, x)$，且 $\tilde{x} \in \operatorname{argmin}_{x\in\Psi_k} E_{\mathrm{ReLU}}(SA, Sb, x)$ 满足 $\|A\tilde{x} - b\|_{\mathrm{ReLU}} \leq (1+\varepsilon)$ 最小值 $_{x\in\Psi_k}\|Ax - b\|_{\mathrm{ReLU}}$。那么 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

更准确地说，对所有 $x \in \mathbb{R}^d$，都有 $\|x\|_1 = \|x\|_{\mathrm{ReLU}} + \|-x\|_{\mathrm{ReLU}}$。与高斯 $\ell_2 \to \ell_p$ 嵌入情形中的论证类似，将 G 替换为嵌入矩阵 $P = [I, -I]^T$（该矩阵会复制并取反输入向量），即可得到关于草图化 $\|\cdot\|_{\mathrm{ReLU}}$的下界。由此可知，$SP$ 是针对 ℓ_1 的无偏草图，因此具有相同的下界 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

最后，在这一系列归约中，我们推导出如下类别的 k稀疏仿射嵌入的下界：类铰链损失函数 [MMR21]，与 ReLU 函数的差异至多为加性偏差 a (例如，逻辑损失，其中 $a \leq \ln(2)$)，参见定义 1。

定义 1. 若 $f(\cdot)$ 是一个(L, a_1, a_2) 类铰链损失函数，则称其为类铰链型损失函数，当且仅当 f 是 L *Lipschitz* 连续的，$\forall x \geq 0$: $f(x) \geq a_2 > 0$，且 $\forall x$: $|f(x) - \mathrm{ReLU}(x)| \leq a_1$.

推论 4. 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。假设 $S \in \mathbb{R}^{m\times n}$ 是针对某个类铰链损失函数 f的无偏子空间嵌入，其估计量为 $E_f(SA, Sb, x)$，且满足 $\forall x \in \Psi_k: (1-\varepsilon)\|Ax - b\|_f \leq E_f(SA, Sb, x) \leq (1+\varepsilon)\|Ax - b\|_f$. 那么 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

由于我们关于 ReLU 的最小化下界直接推出仿射嵌入下界，因此只需证明，任意用于 f 的 k 稀疏仿射嵌入都能导出一个用于 ReLU 的 k 稀疏仿射嵌入。难点在于，我们仅对正部获得了将 ReLU 嵌入到 f 中的近似乘法嵌入。因此，需要单独处理 ReLU 取值为零的其余部分，最小化下界也就无法像我们先前的论证那样直接推出。

采样不适用于稀疏回归 基于采样的算法是非稀疏回归草图中的重要工具 [DMM06；Das+09；Mun+18；MMR21]。然而，它们在稀疏情形下并未取得任何非平凡结果。

为说明为什么我们的上界都基于无须预先知道输入的线性草图，而非采样，我们针对对输入矩阵的行进行子采样（并重新加权）的矩阵 S 证明下界。这里我们证明，若 A 是单位矩阵，且 b 是一个随机标准基向量，那么任何只能访问 SA 和 Sb的算法，在 $k = 1$这一情形下就必然以 $> 1/2$ 的概率失败。

定理 5. 考虑任意有界近似因子 $\alpha \geq 1$ 以及任意满足 $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ 的函数；该函数对全零向量的取值为0，对任何其他向量的取值均为正数。对于任意$n > 9$，都存在某个输入矩阵 $A \in \mathbb{R}^{n\times n}$ 及向量分布 $b \in \mathbb{R}^n$，使得对于任意满足 $m < n/3$的采样矩阵 $S \in \mathbb{R}^{m\times n}$，任何仅能访问 SA 和 Sb 的算法，都无法以至少 $1/2$ 的概率（对 b 的选择以及算法中任何可能的随机性取概率）输出一个k-稀疏的 $\tilde{x} \in \Psi_k$，使其满足 $\|A\tilde{x} - b\| \leq \alpha \cdot$最小值$_{x\in\Psi_k} \|Ax - b\|$。

原因在于，随机选择 b时，如果 $m < n/3$，那么草图中只包含满足 $b = 0$ 的行；但要获得有界的近似误差，保留非零行至关重要。根据构造， A 无法帮助找到这个坐标，因此，样本中未包含的所有可能情况成功的概率都大致相同，约为 $1/n$。由此可得下界 $m = \Omega(n)$。

对于这个下界，至关重要的是算法在构造 S 时无法访问 b，这也支持了这样一个事实：正是无关性 这一性质使草图与稀疏回归中的采样有所区别。然而，如果我们的目标是通过采样获得一个 k-稀疏仿射嵌入，那么即使算法能够完全访问数据，且即使$k = 1$， $A = I$ 和 $b = 0$，仍可得出下界 $m = \Omega(n)$；这也意味着无法通过采样获得 RIP 性质。

定理 6. 考虑任意有界近似因子 $\alpha \geq 1$ 以及任意 $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}^{\geq 0}$，它在全零向量上的取值为0，在任何其他向量上的取值均为正数。对于任意$n > 9$，都存在某个输入矩阵 $A \in \mathbb{R}^{n\times n}$，使得不存在满足 $m < n$的采样矩阵 $S \in \mathbb{R}^{m\times n}$，能够对所有 k-稀疏的 $x \in \Psi_k$ 满足 $\alpha^{-1}\|Ax\| \leq \|SAx\| \leq \alpha\|Ax\|$。

值得注意的是，与采样不可能性结果不同，草图方法通过相对简单的并集界论证即可成功，例如，[Bar+08] 对于 ℓ_2 情形，正如我们将在下一节看到的那样。

2.2 上界

稀疏 ℓ_2 回归 同样，我们从 ℓ_2 开始。这个上界与通过 Johnson-Lindenstrauss 嵌入构造 RIP矩阵的已知方法类似 [CRT06；Don06；Bar+08] [JL84], , 即高斯矩阵。与这些工作的主要区别在于，任意固定的k稀疏解空间所形成的子空间不一定彼此正交，也不一定与标准基向量对齐。

定理7. 令 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。存在一个随机矩阵 $S \in \mathbb{R}^{m\times n}$上的分布，使得以常数概率有 $m = O(k \log(d/k)/\varepsilon^2)$ $\forall x \in \Psi_k : (1-\varepsilon)\|Ax - b\|_2 \leq \|S(Ax - b)\|_2 \leq (1+\varepsilon)\|Ax - b\|_2$.

其思路是，至多有 $\binom{d}{k} \leq (ed/k)^k$ 种不同的 k-稀疏支撑集，每一种都对应于从 k 中选取 A 列的一种方式。每种选取所张成的都是 k 维线性子空间 $\mathbb{R}^n$。根据 [Sar06], 的子空间嵌入构造，对每一种列选取所形成的子空间，都可以通过嵌入该子空间内覆盖单位球的一个网中的点来处理，该网的大小为 $(3/\varepsilon)^k$。其余向量可通过三角不等式与网点关联起来；而通过

线性性。事实上，通过更精巧的论证（例如，参见 [Woo14]），可以用一个绝对常数 $\varepsilon_0 := 1/2$ 取代 ε。因此，为了达到 $(1 \pm \varepsilon)$ 失真，所需嵌入的点的总数至多为 $|\mathcal{N}| \leq (d/k)^k \cdot c^k$，其中 c 为绝对常数。通过先应用 Johnson-Lindenstrauss引理，再应用并集界，使用一个具有 $m = O(\log(|\mathcal{N}|)/\varepsilon^2) = O(k \log(d/k)/\varepsilon^2)$ 行的高斯矩阵即可实现该嵌入，与下界相匹配。

稀疏 ℓ_p 回归。为了推广到 ℓ_p，我们首先聚焦于 $p = 1$。与 ℓ_2 情形不同，已知将 ℓ_1^n 直接嵌入到 ℓ_1^m中，即使对于固定常数 ε，其规模也要么是指数级的，即 $m \in \Omega(2^{\sqrt{d}})$，要么必然会产生 $\Omega(d/\text{polylog}(d))$ 的失真 [WW19; LWY21],，因此 $(1 \pm \varepsilon)$-近似似乎难以实现。另一种方法是采用 [Ind06] 的非线性中位数估计器，但通常会避免这种方法，因为它会导致草图空间中的优化问题非凸且通常难以求解。不过，我们注意到，k-稀疏回归本身就是非凸且 NP-难的。因此，这种方法适合我们的设定。[Ind06] 的草图矩阵是一个线性草图 $C \in \mathbb{R}^{m \times n}$，其元素是独立同分布的柯西随机变量。已知这些变量具有 1-稳定性，即它们与向量 x 的点积仍是尺度参数为 $\|x\|_1$ 的柯西随机变量，因此每一行都能给出一个合理的估计器。不过，为了实现集中性，整体估计器取所有行估计器的中位数，而不是它们的 ℓ_1-范数。在 [Bac+16] 中，这种草图与 ℓ_1 的网论证相结合，为一个 $O(k\log(k/\varepsilon\delta)/\varepsilon^2)$ 维子空间得到了草图维度。然而，若仅对 $\binom{d}{k}$ 个 k 列的选择应用并集界，则会得到 $m = O((k^2 \log(d/k) + k \log(k/\varepsilon))/\varepsilon^2)$，远高于下界。我们展开证明，将其改进为 $m = O((k\log(d/k) + k\log(k/\varepsilon))/\varepsilon^2)$；除非 k 相对较大（量级为 $k = \omega(\sqrt{\varepsilon d})$），否则这与我们的下界相匹配。 —

定理8。令 $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n, p \in [1, 2)$。存在一个关于随机矩阵的分布$S \in \mathbb{R}^{m \times n}$，且 $m = O(k(\log(d/k) + \log(k/(\varepsilon\delta)))/\varepsilon^2)$，使得以至少$1 - \delta$ 的概率满足 $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_p \leq \|S(Ax - b)\|_{\text{med}} \leq (1 + \varepsilon)\|Ax - b\|_p$，其中对于任意$y \in \mathbb{R}^d, \|y\|_{\text{med}} :=$ 中位数$\{|y_i| \mid i \in [d]\}$。

将这一结果推广到 $\ell_p, p \in [1, 2)$ 后，我们得到了类似的上界。草图矩阵仍是一个线性草图 $C \in \mathbb{R}^{m \times n}$，其条目是从 p-稳定分布中独立同分布抽取的随机变量；这种分布是对 1-稳定柯西分布的推广。[Ind06] 曾提出将这种推广用于单向量草图，但据我们所知，除 $p \in \{1, 2\}$ 外，由于缺乏累积分布函数（cdf）和概率密度函数（pdf）的闭式表达式，这一方法尚未得到具体推导。本文展示了如何直接为所有 k-稀疏向量构造子空间嵌入。为此，我们利用了关于 p-稳定分布尾部的界 [B-LM18]。需要指出的是，p-稳定分布具有高峰厚尾特性。更具体地说，除 $p = 2$ 外，其尾部较重，衰减规律为 $\Pr[|X| > \tau] \leq 1/\tau^p$。因此，要构造并应用网论证，我们需要依赖非线性分位数估计量，从而对任意 $p \in [1, 2)$ 实现集中性，正如上文所述情形一样。另一个更有意思的问题是，如何在没有闭式表达式的情况下分析 p-稳定分布的累积分布函数。我们解决这一问题的思路基于以下事实：任意 p-稳定分布的特征函数都有已知的闭式表达式，且除归一化常数外，它等于一个 p-广义正态分布的概率密度函数 [Dyt+18]。也就是说，它的表达式为 $\phi(t) = \exp(-|\gamma_p t|^p)$，其中 γ_p 是依赖于 p 的常数尺度参数。利用 L´evy 的反演定理，我们可以通过一个涉及特征函数的积分来分析累积分布函数及其导数。作为一个附带结果，我们证实了 [Ind06] 的猜想：对所有 $p \in [1, 2]$[1]，都可以通过中位数估计量得到 $(1 \pm \varepsilon)$ 近似。

[1]参考文献 [Ind06] 给出了一个非构造性的证明，表明存在某个（未知的）分位数，该分位数可能依赖于 p 和 ε，并且能够产生一个良好的估计量。

ReLU 和类铰链损失函数 众所周知 [Mun+18]，在不作任何假设的情况下，这类函数无法通过具有 $o(n)$ 行的草图实现相对误差保证。为了解决这一问题，[Mun+18] 引入了一个自然的概念，用于衡量对矩阵 A 进行草图化的复杂度；我们也用它来刻画结果。直观而言，当存在某个 x，使所有正项的 ℓ_1-范数与所有负项的 ℓ_1-范数之间出现显著失衡时，参数 $\mu := \sup_{x\in\mathbb{R}^d\setminus\{0\}} \|(Ax)^+\|_1/\|(Ax)^-\|_1$ 会很大。例如，数据具有完美线性可分性时，就可能出现这种情况。然而，正如 [Mun+18] 所指出的，我们通常预期 μ 较小。近期有人利用这些假设，提出了首个用于逻辑回归的无偏线性草图 [MOW21; MOW23], ，并据此为非稀疏情形下的最小化问题设计了高效算法。 [MMR21] 的另一项近期工作，则在采样和核心集算法适用的情形下，使对 d、μ 和 ε 的依赖较小。不过，我们已经指出，在稀疏回归的设置中，采样并不可行。

因此，我们基于针对 ℓ_1的中位数草图，结合并扩展这些结果，使草图覆盖更广泛的损失函数类别，并改进对近似参数的依赖关系。我们注意到，这牺牲了在草图空间中进行优化的效率；但正如我们之前所论证的，在稀疏回归的背景下，找到正确的支持集本身就是一个难题，这促使我们专注于尽可能好的参数化。

定理 9。令 $A \in \mathbb{R}^{n\times d}$、$b \in \mathbb{R}^n$。存在一个无关草图 S，包含 $O(\frac{\mu^2 k}{\varepsilon^2}$对数$(\frac{\mu d}{\varepsilon\delta}))$ 行，以及一个估计量 $g_{\mathrm{ReLU}}(SA, Sb, x)$，使得以至少 $1-\delta$ 的概率，有 $\forall x \in \Psi_k$：$(1-\varepsilon)\|Ax-b\|_{\mathrm{ReLU}} \le g_{\mathrm{ReLU}}(SA, Sb, x) \le (1+\varepsilon)\|Ax-b\|_{\mathrm{ReLU}}$.

我们的 k 稀疏仿射嵌入草图利用了这样一个事实：$\mathrm{ReLU}(x) = (\sum_i x_i + \|x\|_1)/2$，因为负值在求和时为负，在范数中又为正，因此会相互抵消。正值在这两部分中均为正，因此会被计算两次，所以除以 2 即可得到 $\mathrm{ReLU}(x)$ 的精确值。只需使用一个行向量即可精确地对各项之和进行草图化，而 ℓ_1 范数则通过带中位数估计器的（1-稳定）柯西草图进行草图化，详见上一段。现在，该估计的误差为 $\varepsilon\|x\|_1$，但根据 μ-复杂度假设，ℓ_1 范数与正值相差大约一个 μ因子，因此将 μ 并入 ε，即可得到 $\varepsilon\mathrm{ReLU}(x)$ 的误差。我们将这一结果扩展到更丰富的一类类铰链损失函数，其中包括逻辑回归。这些函数与 ReLU 函数加性接近，见定义 1。例如，逻辑回归损失 $\zeta(x) = \ln(1+\exp(x))$ 在 $\pm\infty$ 的极限下，其渐近线与 ReLU 相同。然而，在零附近，这两个函数的差异更为显著，其最大偏差有界，为 $\zeta(0) - \mathrm{ReLU}(0) = \ln(2)$。

定理 *10.* 设 $A \in \mathbb{R}^{n\times d}$、$b \in \mathbb{R}^n$，并设 f 为一个 (L, a_1, a_2) 类铰链损失函数。存在一个无偏草图 S，其行数为 $m = O(\frac{c^{10}\mu^2 k}{\varepsilon^2} log(\frac{cn\mu d}{\varepsilon\delta}))$，其中 $c = max(1, L, a_1, 1/a_2)$，以及一个估计器 $g_f(SA, Sb, x)$，使得以至少 $1-\delta$ 的概率，有 $\forall x \in \Psi_k$: $(1-\varepsilon)\|Ax-b\|_f \le g_f(SA, Sb, x) \le (1+\varepsilon)\|Ax-b\|_f$.

现在的思路是将损失函数拆分为两个部分 $\sum_i(\zeta(x_i) - \mathrm{ReLU}(x_i)) + \sum_i \mathrm{ReLU}(x_i)$。ReLU函数可以按照上一段所述的方法处理，其余部分则是有界项的和。这样，我们再次能够实现集中，并对 k-稀疏向量的网进行并集界估计，误差约为 $\varepsilon\frac{n}{\mu}$。这一误差可以通过利用 μ-复杂度假设得到的补充下界来抵消，这与 [MMR21类似； MOW21], 最终给出相对误差保证。 –

为完整起见，我们给出以下简单结果，它建立了在 k-稀疏仿射嵌入的草图空间中进行最小化（如上述所有上界所示）与原始最小化问题之间的联系。

推论 *11*。设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。设 $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ 为任意损失函数。设 S 为 $[A,b]$ 的一种无偏线性草图，并设 $E(SA,Sb,x)$ 为一个满足 $\forall x \in \Psi_k: (1-\varepsilon)\|Ax-b\| \leq E(SA,Sb,x) \leq (1+\varepsilon)\|Ax-b\|$. 的估计量。那么， $\tilde{x} \in \mathit{argmin}_{x\in\Psi_k} E(SA,Sb,x)$ 满足 $\|A\tilde{x}-b\| \leq (1+O(\varepsilon))\ \mathit{min}_{x\in\Psi_k}\|Ax-b\|$.

2.3 LASSO 回归

LASSO回归 [Tib96] 是 k-稀疏 ℓ_2 回归的凸松弛，并作为实现稀疏性和特征选择的启发式方法而广受欢迎。LASSO回归是我们研究的一个特殊对象，因为在这里我们不假设任何解都是 k-稀疏的。在这种情况下，可以针对稠密 ℓ_2 问题进行子空间嵌入，将其草图压缩到 $m = \Theta(d/\varepsilon^2)$ 维 [Sar06;NN14]。此外，在正则化参数非常小、因而解的范数实际上不受约束的情况下，该问题等价于最小二乘回归；此时， $\Theta(d/\varepsilon)$ 是充分且必要的 [Sar06;CW09]。然而，通常会施加 ℓ_1正则化，以得到一个稀疏极小化解；对于这种解，我们可以再次希望利用其诱导的稀疏性，并以正则化参数的取值 λ为参数，从而将行数减少到poly$(1/\lambda, \log d)$。这里我们给出一个取决于 ℓ_1 正则化参数 λ的草图上界。

定理12. 考虑 $A \in \mathbb{R}^{n\times d}$、 $b \in \mathbb{R}^n$和 $\lambda \in (0,\ 1)$。假设 $\|A\|_2 \leq 1$ 与 $\|b\|_2 \leq 1$，若 $S \in \mathbb{R}^{m\times n}$是随机高斯矩阵（即每一项均独立同分布地采样自 $N(0,\ 1/m)$），则对任意 $\varepsilon\ \delta \in\ 0,1 \qquad m = O \qquad 1-\delta \qquad \tilde{x} = \qquad _{x\in\mathbb{R}^d}\|SAx-$，() 与($^{\log d/\delta}_{\lambda^2\cdot\varepsilon^2}$)，以至少 $Sb\|_2^2+\lambda\|x\|_1 \qquad \|A\tilde{x}-b\|_2^2+\lambda\|\tilde{x}\|_1 \leq \overline{1+\varepsilon}\cdot \qquad _{x\in\mathbb{R}^d}\|Ax-b\|_2^2+\lambda\|x\|_1.$ 的概率，若argmin则() min

请注意，我们对 $\|A\|_2$、 $\|b\|_2$ 施加的约束是必要的，因为 LASSO 回归不具备尺度不变性。该结果首先通过证明优化器的范数必须根据最优目标值 $\|x\|_1 \leq {}^{2\cdot OPT}_{\lambda}$有界来得到。这使我们能够将注意力集中于集合 $\mathcal{T} = \{y = Ax-b \quad \|x\|_1 \leq {}^{2\cdot OPT}_{\lambda}\}$ ——— $2d+1:$ ，该集合根据有界范数可以表示为单位球中各点的凸包。对于该集合，我们可以利用 [NN19]的嵌入结果，为集合中的所有向量 $\varepsilon\lambda$ 取得加性误差 $\mathcal{T}$，从而将草图误差与 $O(\varepsilon\cdot OPT)$联系起来，并最终得到我们的 $1+\varepsilon$ 相对误差近似结果。

我们通过以下下界补充上界，该下界与d 和 λ的依赖关系相匹配。证明基于我们为稀疏 ℓ_2 回归开发的新技术。然而，对于有界范数输入的条件 A， b，我们无法植入额外的行构造。因此，我们需要将 λ 缩小 ε倍，但遗憾的是，这抵消了先前 ℓ_2 下界中对 ε 的依赖。尽管如此，我们的结果表明，对于任何带有估计器的草图，若要将 LASSO 的求解误差控制在 $1+\varepsilon$ 近似范围内，都至少需要 $\log(\lambda d)/\lambda^2$ 行。

定理13。 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$ 具有有界的 $\|A\|_2 \leq 1$ 和 $\|b\|_2 \leq 1$，并令 $\lambda \in (0,1)$。假设 $S \in \mathbb{R}^{m\times n}$ 是用于 *LASSO* 回归的一个无偏线性草图，其估计器为 $E_\lambda(SA,Sb,x)$，且以常数概率 $\tilde{x} \in \mathrm{argmin}_{x\in\Psi_k} E_\lambda(SA,Sb,x)$ 满足 $\|A\tilde{x}-b\|_2^2+\lambda\|\tilde{x}\|_1 \leq (1+\varepsilon)\cdot\min_{x\in\mathbb{R}^d}\|Ax-b\|_2^2+\lambda\|x\|_1$. 那么 $m = \Omega(\frac{\log(\lambda d)}{\lambda^2})$。

2.4 从稀疏回归中区分稀疏恢复

这里我们给出了一个上界 $m = O(k\log(d)/\varepsilon + k\log(k/\varepsilon))/\varepsilon^2)$，它非常接近下界 $m = \Omega(k\log(d/k)/\varepsilon + k/\varepsilon^2)$，对应 [PW11]。令人惊讶的是，这揭示了 k-稀疏恢复问题与 k-稀疏回归问题之间的差异。结合我们的主要下界 $\Omega(k\log(d/k)/\varepsilon^2)$，可知 k-稀疏回归的草图构造严格更难，相比于

稀疏恢复。为了得到新的上界，问题在于我们需要确定使用哪个子集，但无法承担估算每个子集代价的开销，尽管我们可以对每个子集都得到良好的估计。为了解决这个问题，我们并行运行一个两阶段估计流程：一个 CountSketch 用于粗略估计各项，从而得到一个包含 I k 个感兴趣坐标的超集，即大小为 $|I| = O(k/\varepsilon)$ 的超集；另一个 CountSketch 精度更高，但只需恢复集合 I 中数量相对较少的 k-子集的估计值。现在，如果我们输出一个仅在 I 中非零、且取值为前 k 项及其通过高精度草图获得的估计值的向量，就能以比求解k-稀疏回归问题所需更少的行数，为 k-稀疏恢复提供所需的 $1+\varepsilon$ 近似。

定理 14. 对于输入 $x \in \mathbb{R}^d$，上述稀疏恢复方案使用 $O(k\log(k/\varepsilon)/\varepsilon^2 + k\log(d)/\varepsilon)$次测量，并以至少 $1 - 1/\operatorname{poly}(d) - 1/\operatorname{poly}(k/\varepsilon)$的概率返回一个 k-稀疏向量$\hat{}x \in \Psi_k$，满足 $\|x - \hat{}x\|_2^2 \le (1+\varepsilon)\min_{x_k \in \Psi_k} \|x - x_k\|_2^2$.

3 结论

本文研究了在多种回归损失函数下，稀疏回归问题中无偏线性草图的复杂度，例如 ℓ_p 回归、逻辑回归、ReLU 损失和类铰链损失函数。对于所有这些问题，我们得到的界基本上是[2] 紧确的，为 $\Theta(k\log(d/k)/\varepsilon^2)$。我们进一步研究了 LASSO 的草图复杂度。LASSO 是一种常用的凸松弛方法，通常作为求解稀疏线性回归的启发式方法。我们给出了第一个低于关于 d线性依赖的界，即 $O(\log(d)/(\lambda\varepsilon)^2)$，其中 λ 是正则化参数。此外，我们证明了它与压缩感知中研究的稀疏恢复问题之间存在分离。令人意外的是，我们发现稀疏回归问题需要 $m = \Omega(k\log(d/k)/\varepsilon^2)$，因此其草图复杂度严格高于稀疏恢复；对于后者，我们证明了一个新的 $m = O(k\log(d)/\varepsilon + k\log(k/\varepsilon)/\varepsilon^2)$ 上界。我们还表明，尽管数据依赖的重要性采样技术在无约束的非稀疏回归问题中广泛有效，但在稀疏情形下无法给出任何非平凡的界。这凸显了无偏草图技术在稀疏问题背景下的重要性。未来的研究方向包括缩小尚存的差距，尤其是针对类铰链损失函数。另一个值得探索的方向是结合我们的草图技术，开发更具可扩展性、速度更快的启发式方法，并评估其实际表现。最后，由于我们的草图针对尽可能小的目标维度进行了优化，研究将其应用于数据时的速度与增加目标维度之间的权衡也很有意义。

致谢

Alexander Munteanu 得到了德国研究联合会（DFG）、SFB 876 协作研究中心 C4 项目以及多特蒙德数据科学中心（DoDSc）的支持。D. P. Woodruff 得到了美国海军研究局（ONR）N00014-18-1-2562 号资助和西蒙斯研究员奖的部分资助。Cameron Musco 得到了 Adobe Research 资助、Google Research 学者奖以及美国国家科学基金会（NSF）编号 2046235 和编号 1763618 的资助

[2]至多相差低阶多对数项以及问题特定参数，例如 μ；具体项见表 1 。

参考文献

[Aue30] 赫尔曼·奥尔巴赫。《关于具有共轭直径的凸曲线面积》。博士论文。利沃夫大学，1930 年。[AW09] Arash A. Amini 和 Martin J. Wainwright。《稀疏主成分半定松弛的高维分析》。载于：《统计年刊》37.5B（2009），第 2877–2921 页。 DOI：10 . 1214/ 08- AOS664。 URL：https :/ / doi . org / 10 . 1214/ 08 - AOS664。[Ba+10] Khanh Do Ba 等。《稀疏恢复的下界》。载于：第二十一届 ACM-SIAM 离散算法年度研讨会论文集（SODA）。2010 年，第 1190–1197 页。[Bac+16] Arturs Backurs 等。《通用范数下稀疏恢复的近最优界及其在 k-中位数草图中的应用》。载于：第二十七届 ACM-SIAM 离散算法年度研讨会论文集（SODA）。2016 年，第 318–337 页。[Bar+08] Richard Baraniuk 等。《随机矩阵受限等距性质的简单证明》。载于：《构造性逼近》28.3（2008），第 253–263 页。[BHW05] Szymon Borak、Wolfgang H¨ardle 和 Rafa-l Weron。《稳定分布》。载于：《金融与保险统计工具》。Springer，2005 年，第 21–44 页。[B-LM18] W. Bednorz、R. -Lochowski 和 R. Martynek。《关于对称及完全非对称 α-稳定分布的尾部》。载于：arXiv:Probability（2018）。https : / / arxiv . org/abs/1802.00612。[BPP18] Guy Bresler、Sung Min Park 和 Madalina Persu。《从稀疏线性回归中进行稀疏 PCA》。载于：《神经信息处理系统进展》第 31 卷（NeurIPS）。2018 年，第 10965–10975 页。[CCF04] Moses Charikar、Kevin C. Chen 和 Martin Farach-Colton。《在数据流中查找高频项》。载于：《理论计算机科学》312.1（2004），第 3–15 页。[Cla05] Kenneth L Clarkson。《回归的次梯度与采样算法》。载于：第十六届 ACM-SIAM 离散算法年度研讨会论文集（SODA）。2005 年，第 257–266 页。[CMW13] T. Tony Cai、Zongming Ma 和 Yihong Wu。《稀疏 PCA：最优速率与自适应估计》。载于：《统计年刊》41.6（2013），第 3074–3110 页。 URL：https : //doi.org/10.1214/13-AOS1178。[CP19] Xue Chen 和 Eric Price。《通过线性样本稀疏化实现主动回归》。载于：第三十二届计算学习理论年度会议论文集（COLT）。2019 年。[CRT06] Emmanuel J Candes、Justin K Romberg 和 Terence Tao。《从不完整且不准确的测量中稳定恢复信号》。载于：《纯粹与应用数学通讯》59.8（2006），第 1207–1223 页。[CT06] Thomas M. Cover 和 Joy A. Thomas。《信息论基础》（Wiley 电信与信号处理丛书），第二版。Wiley-Interscience，2006 年。ISBN：0471241954。[CW09] Kenneth L. Clarkson 和 David P. Woodruff。《流式模型中的数值线性代数》。载于：第四十一届 ACM 计算理论年度研讨会论文集（STOC）。2009 年，第 205–214 页。

[CW15] Kenneth L. Clarkson 和 David P. Woodruff。“用于 M-估计量的草图：稳健回归的统一方法”。载于：第 26 届 ACM-SIAM 离散算法年度研讨会（SODA）论文集。2015 年，第 921–939 页。[CW79] J Lawrence Carter 和 Mark N Wegman。“通用哈希函数类”。载于：《计算机与系统科学杂志》18.2（1979 年），第 143–154 页。[Das+09] Anirban Dasgupta 等。“回归的采样算法与核心集”。载于：《SIAM 计算杂志》38.5（2009 年），第 2060–2078 页。 DOI: 10.1137/070696507。[DK08] Abhimanyu Das 和 David Kempe。“线性回归子集选择算法”。载于：第 40 届 ACM 理论计算年度研讨会（STOC）论文集。2008 年，第 45–54 页。 DOI: 10.1145/1374376.1374384。 URL: https://doi.org/10.1145/ 1374376.1374384。[DK18] Abhimanyu Das 和 David Kempe。“近似次模性及其应用：子集选择、稀疏近似与字典选择”。载于：《机器学习研究杂志》19（2018 年），3:1–3:34。 URL: http://jmlr.org/papers/v19/16-534.html。[DMM06] Petros Drineas、Michael W. Mahoney 和 S. Muthukrishnan。“ 回归采样算法及其应用 ”。载于：第十七届 ACM-SIAM 离散算法年度研讨会（SODA）论文集。2006 年，第 1127–1136 页。[Don06] David L. Donoho。“压缩感知”。载于：《IEEE 信息论汇刊》52.4（2006 年），第 1289–1306 页。[Dyt+18] Alex Dytso 等。“广义高斯分布的解析性质”。载于：《统计分布与应用杂志》5（2018 年 12 月），第 1–40 页。[FKT15] Dean P. Foster、Howard J. Karloff 和 Justin Thaler。“变量选择很难”。载于：第 28 届学习理论会议（COLT）论文集。2015 年，第 696–709 页。[GMS03] Anna C. Gilbert、S. Muthukrishnan 和 Martin Strauss。“利用相干性近似冗余字典上的函数”。载于：第十四届 ACM-SIAM 离散算法年度研讨会（SODA）论文集。2003 年，第 243–252 页。 URL: http: // dl.acm.org/citation.cfm?id=644108.644149。[HIM18] Sariel Har-Peled、Piotr Indyk 和 Sepideh Mahabadi。“近似稀疏线性回归”。载于：第 45 届国际自动机、语言与程序设计研讨会（ICALP）。第 107 卷。2018 年，77:1–77:14。[Ind06] Piotr Indyk。“稳定分布、伪随机生成器、嵌入与数据流计算”。载于：《ACM 杂志》53.3（2006 年），第 307–323 页。[JL84] William B. Johnson 和 Joram Lindenstrauss。“将 Lipschitz 映射扩展到希尔伯特空间”。载于：《当代数学》26.189-206（1984 年），第 1–1 页。[Kal+17] Satyen Kale 等。“自适应特征选择：RIP 条件下计算高效的在线稀疏线性回归”。载于：第 34 届国际机器学习会议（ICML）论文集。2017 年，第 1780–1788 页。 URL: http://proceedings.mlr.press/ v70/kale17a.html。[KVW14] Ravi Kannan、Santosh Vempala 和 David P. Woodruff。“分布式数据的主成分分析与高阶相关性”。载于：第 27 届学习理论会议（COLT 2014）论文集，西班牙巴塞罗那，2014 年 6 月 13–15 日。2014 年，第 1040–1057 页。[L´ev25] Paul L´evy。《概率论计算》。Gauthier-Villars, 1925 年。

[LWY21] Yi Li、David P. Woodruff 和 Taisuke Yasuda。“用于 ℓ_1 的指数级改进降维：子空间嵌入与独立性检验”。载于：学习理论会议（COLT）。2021 年，第 3111–3195 页。[LXG20] Gen Li、Xingyu Xu 和 Yuantao Gu。“RIP 常数的下界与最高阶次序统计量之和的集中性”。载于：《IEEE 信号处理汇刊》68（2020 年），第 3169–3178 页。[Mas77] Pesi Rustom Masani。“P. L´evy 反演定理的推广”。载于：《多元分析杂志》7（1977 年），第 292–335 页。[Mat13] Jiˇr´ı Matouˇsek。《度量嵌入讲义》。https : / / kam. mff .cuni .cz /~matousek/ba-a4.pdf。访问日期：2021-10-20。2013 年。[Mba+21] Joelle Mbatchou 等。“用于定量和二元性状的计算高效全基因组回归”。载于：《自然遗传学》53（2021 年），第 1097–1103 页。 DOI：10.1038/s41588-021-00870-7。 URL：https://doi.org/10.1038/s41588-021-00870-7。[Mil02] A. Miller。《回归中的子集选择》。Chapman & Hall/CRC 统计与应用概率专著。CRC 出版社，2002 年。 ISBN：9781420035933。 URL：https:// books.google.de/books?id=7p59iir822sC。[MMR21] Tung Mai、Cameron Musco 和 Anup Rao。“用于分类的核心集——简化并强化”。载于：神经信息处理系统进展 34（NeurIPS）。2021 年，第 11643–11654 页。[MOP22] Alexander Munteanu、Simon Omlor 和 Christian Peters。“p-广义 Probit 回归：通过草图和核心集实现可扩展的极大似然估计”。载于：国际人工智能与统计会议（AISTATS）。2022 年，第 2073–2100 页。 URL：https://proceedings.mlr.press/v151/munteanu22a.html。[MOW21] Alexander Munteanu、Simon Omlor 和 David P. Woodruff。“逻辑回归的无偏草图”。载于：第 38 届国际机器学习会议（ICML）论文集。2021 年，第 7861–7871 页。[MOW23] Alexander Munteanu、Simon Omlor 和 David P. Woodruff。“用于 ℓ_1 和逻辑回归的近线性常数因子草图”。载于：第 11 届国际学习表征会议（ICLR）论文集。2023 年。[Mun+18] Alexander Munteanu 等。“关于逻辑回归的核心集”。载于：神经信息处理系统进展 31（NeurIPS）。2018 年，第 6562–6571 页。 URL：https://proceedings.neurips.cc/paper/2018/hash/63bfd6e8f26d1d3537f4c5038264ef36-Abstract.html。[Mun23] Alexander Munteanu。“数据流和分布式数据回归问题的核心集与草图”。载于：《资源受限条件下的机器学习》，第 1 卷——基础。De Gruyter，2023 年，第 85–98 页。 ISBN：9783110785944。 DOI：doi:10.1515/ 9783110785944-003。 URL：https://doi.org/10.1515/9783110785944-003。[Nat95] B. K. Natarajan。“线性系统的稀疏近似解”。载于：《SIAM 计算杂志》24.2（1995 年），第 227–234 页。[NN14] Jelani Nelson 和 Huy L. Nguyˆen。“无偏子空间嵌入的下界”。载于：《自动机、语言与程序设计——第 41 届国际研讨会（ICALP）》。2014 年，第 883–894 页。

[NN19] Shyam Narayanan 和 Jelani Nelson。《欧几里得空间中的最优终端维数约简》。载于：第 51 届 ACM SIGACT 计算理论年度研讨会论文集（STOC）。2019 年，第 1064–1069 页。 DOI: 10.1145/3313276.3316307.URL: https://doi.org/10.1145/3313276.3316307.[PW11] Eric Price 和 David P Woodruff。《 $(1+\varepsilon)$- 近似稀疏恢复》。载于：2011 年 IEEE 第 52 届计算机科学基础年度研讨会（FOCS）。2011 年，第 295–304 页。[Sar06] Tam´as Sarl´os。《通过随机投影改进大矩阵的近似算法》。载于：第 47 届 IEEE 计算机科学基础年度研讨会论文集（FOCS）。2006 年，第 143–152 页。[SW11] Christian Sohler 和 David P. Woodruff。《用于 ℓ_1-范数的子空间嵌入及其应用》。载于：第 43 届 ACM 计算理论研讨会论文集（STOC）。2011 年，第 755–764 页。[TGS06] Joel A. Tropp、Anna C. Gilbert 和 Martin J. Strauss。《同时稀疏近似算法。第一部分：贪心追踪》。载于：Signal Process. 86.3 (2006)，第 572–588 页。 DOI: 10.1016/j.sigpro.2005.05.030. URL: https://doi.org/10.1016/j.sigpro.2005.05.030.[Tib96] Robert Tibshirani。《通过 LASSO 进行回归收缩与变量选择》。载于：Journal of the Royal Statistical Society. Series B (Methodological) 58.1 (1996)，第 267–288 页。 ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.[Tro06a] Joel A. Tropp。《同时稀疏近似算法。第二部分：凸松弛》。载于：Signal Process. 86.3 (2006)，第 589–602 页。 DOI: 10.1016/j.sigpro. 2005.05.031. URL: https://doi.org/10.1016/j.sigpro.2005.05.031.[Tro06b] Joel A. Tropp。《尽管放松：用于在噪声中识别稀疏信号的凸规划方法》。载于：IEEE Trans. Inf. Theory 52.3 (2006)，第 1030–1051 页。 DOI: 10.1109/TIT. 2005.864420. URL: https://doi.org/10.1109/TIT.2005.864420.[Ver18] Roman Vershynin。《高维概率：数据科学应用导论》。第 47 卷。剑桥大学出版社，2018 年。[Woo14] David P. Woodruff。《草图：数值线性代数的工具》。载于：Foundations and Trends in Theoretical Computer Science 10.1-2 (2014)，第 1–157 页。 DOI: 10.1561/0400000060. URL: https://doi.org/10.1561/0400000060.[WW19] Ruosong Wang 和 David P. Woodruff。《 ℓ_p 无偏子空间嵌入的紧确界》。载于：第 30 届 ACM-SIAM 离散算法研讨会论文集（SODA）。2019 年，第 1825–1843 页。[WW22] Ruosong Wang 和 David P. Woodruff。《 ℓ_1 无偏子空间嵌入的紧确界》。载于：ACM Trans. Algorithms 18.1 (2022)，8:1–8:32。 DOI: 10.1145/3477537. URL: https://doi.org/10.1145/3477537.[WY23] David P. Woodruff 和 Taisuke Yasuda。《在线 Lewis 权重采样》。载于：2023 年 ACM-SIAM 离散算法研讨会论文集（SODA）。2023 年，第 4622–4666 页。[WZ13] David P. Woodruff 和 Qin Zhang。《使用指数随机变量进行子空间嵌入与 ℓ_p-回归》。载于：第 26 届学习理论年度会议论文集（COLT）。2013 年，第 546–567 页。

A 稀疏回归的下界

A.1 $p \geq 1$ 的 ℓ_p-范数损失函数下界

本节将证明我们的主要定理1 关于 ℓ_2 ，随后将其扩展至 ℓ_p， $p \geq 1$，详见下文定理2 。

定理 1. 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。假设 $S \in \mathbb{R}^{m\times n}$ 是用于 k-稀疏 ℓ_2 回归的无偏线性草图，并配有任意估计器 $E(SA, Sb, x)$，使得 $\tilde{x} \in \operatorname{argmin}_{x\in\Psi_k} E(SA, Sb, x)$ 以常数概率满足 $\|A\tilde{x} - b\|_2 \leq (1+\varepsilon)\min_{x\in\Psi_k} \|Ax - b\|_2$。则 $m = \Omega(k\log(d/k)/\varepsilon^2)$。

证明思路如下：

1. 我们构造一个合适的（困难）分布，定义在 k-稀疏支撑集上，并据此定义输入分布。2. 我们证明：当测量（行）数量低于信息论下界时，无法从下述输入分布中恢复支撑集的常数比例。3. 我们构造一个 ℓ_2-回归实例，使得由无偏草图与任意估计器结合得到的任何 $1 + \Theta(\varepsilon)$ 近似都能揭示支撑集的常数比例。因此，该困难性结果也适用于回归问题。

对于第一个证明步骤，我们从构造一个误差纠正码开始，该码将用于主要论证中，以构造一个困难的输入分布。

定义2 （平衡且可校正的支持集）。 考虑一个由 $\mathcal{U} \subset [d]^{k+1}$ 个 $k+1$指标集合组成的集合，使得对于所有 $U \in \mathcal{U}$、 $1 \in U$， $\mathcal{U}$被称为平衡的，如果令 $c_i = |\{U \in \mathcal{U}: \ i \in U\}|$，则对于所有 i、 $j \in [d] \setminus \{1\}$都有 $c_i = c_j$，此外，令 $c_{ij} = |\{U \in \mathcal{U}: \ i,$ 对于所有 $i \neq j$ 以及满足 i、 j、 k、 $l \in [d] \setminus \{1\}$的 $k \neq l$ 和 $j \in U\}|$、 $c_{ij} = c_{kl}$ 都有。若对于所有 U_1、 $U_2 \in \mathcal{U}$、$|U_1 \cap U_2| \leq 9/10 \cdot k$均成立，则称该集合是可校正的。

我们证明如下：存在一个足够大的平衡且可校正的支持集。假设我们随机、均匀地选择 t 个集合 S_1、.....、 S_t，每个集合的大小为 k，并且每个集合均从 $[n] = \{1, 2, \ldots, n\}$中选取。设 $\mathcal{H}$ 是一个两两独立的 $n \cdot (n-1)$ 个哈希函数 $h: \ [n] \to [n]$；众所周知，当 n 为素数时，这样的函数族存在， [CW79], 对此我们不妨作此假设。对于每个 S_i 和 $h \in \mathcal{H}$，令 $h(S_i)$表示 S_i 在 h下的像。

引理1。 对于任意常数 $0 < c < 1$，存在常数 $C > 0$ 以及 $t = \exp(Ck\log(n/k))$个子集 S_1、.....、 S_t，每个子集的大小为 k，使得 $|h(S_i) \cap h'(S_j)| < ck$ 对于所有 $1 \leq i < j \leq t$ 以及所有 $h \neq h' \in \mathcal{H}$均成立。

证明。 注意，对于 $i < j$， $h(S_i)$ 和 $h'(S_j)$ 都是大小为 k 的随机且相互独立的子集。为了计算它们的交集大小，我们可以固定 $h(S_i)$。那么， $|h(S_i) \cap h'(S_j)|$ 至少为 ck 的概率，至多等于 $h(S_i)$ 中某个大小为 ck 的子集同时也是 $h'(S_j)$ 的子集的概率。根据以下结果，该概率进一步有界为

$$\begin{aligned}\binom{k}{ck} \cdot \frac{\binom{n-ck}{k-ck}}{\binom{n}{k}} &\leq \binom{k}{ck} \cdot \frac{(n-ck)! \cdot k!}{(k-ck)! \cdot n!} \\ &\leq 2^k \cdot \left(\frac{k}{n}\right)^{ck} \leq \exp(-Ck\log(n/k)),\end{aligned}$$

其中 $C > 0$ 是一个适当的常数。因此，通过对所有的 $1 \le i < j \le t$ 对以及 $h \neq h' \in \mathcal{H}$ 的所有取值应用并集界，我们可以得出：存在一个选择，使得有 $t = \exp(Ck\log(n/k))$ 个这样的集合，其中常数取另一个值 $C > 0$。

□

上述引理意味着，如果我们有一个大小为 k 的集合 T，它与某个 $h(S_i)$ 至少在 ck 个位置上相交，那么 S_i 就是唯一确定的。

此外，根据 $\mathcal{H}$的两两独立性，对于任意 $a \neq b \in [n]$，并集中的集合数量$\cup_{h\in\mathcal{H},\ 1\le i\le t} h(S_i)$中满足 a 和 b 同时发生的集合数量是相同的。此外，$\cup_{h\in\mathcal{H}\text{中的集合数量},\ 1\le i\le t} h(S_i)$中包含任意特定值 $a \in [n]$ 的集合数量，与包含其他任意特定值 $b \in [n]$的集合数量相同。利用这一论证，我们只需将引理1 应用于k和 $n = d$，以便对 $\{2, \ldots, d+1\}$ 上的集合族进行索引，并将元素 1 添加到每个集合中。之后将从该集合中均匀选取一个元素来构造困难实例。至此，我们证明的第一部分完成。

第二步，我们证明：当测量次数（行数）少于信息论下界时，无法从下面的输入分布中恢复出支持集的一个常数比例。为此，我们需要先界定互信息，然后将其代入 Fano 不等式。

引理 2（互信息界）。设 U 从平衡且可纠正的支持集 $\mathcal{U}$（定义 2）中均匀随机选取，并且对某个 $\varepsilon > 0$，令 $z \in \mathbb{R}^d$ 满足 $z(1) = 1$，$z(i) = \varepsilon/\sqrt{k}$对所有 $i \in U \setminus \{1\}$ 成立，且 $z(i) = 0$ 对所有 $i \notin U$成立。令 $X \in \mathbb{R}^{n\times d}$ 的各行独立地取自均值为零、协方差为 $I + zz^T$的多元高斯分布。那么：

$$I(U; X) \le 7n \cdot \varepsilon^2. \tag{1}$$

证明。从 [AW09], 的高层次概述出发，由于 X 的各行相互独立，我们可以写出：

$$\begin{aligned} I(U; X) &= H(X) - H(X|U) \\ &\le n \cdot [H(x) - H(x|U)] = \frac{n}{2} \cdot \left[\log\det(\mathbb{E}[xx^T]) - \log\det(\mathbb{E}[xx^T|U])\right]. \end{aligned} \tag{2}$$

现在我们计算所需的对数行列式。首先注意到 $\mathbb{E}[xx^T|U] = \mathbb{E}[I + zz^T|U] = I + \mathbb{E}[zz^T|U]$。我们可以观察到 $\mathbb{E}[zz^T|U] = DMD$其中 $M_{ij} = \varepsilon^2/k$对于 $i,\ j \in U$且 $M_{ij} = 0$在其他情况下，并且其中 D 是对角矩阵，其对角元素均为 $D_{11} = \sqrt{k}/\varepsilon$，且对所有 $i \neq 1$都有 $D_{ii} = 1$。注意到 M是秩为 1 的半正定矩阵。因此，DMD也如此。因此，DMD 只有一个非零特征值，等于其迹，即$1 + \varepsilon^2/k \cdot k = 1 + \varepsilon^2$。因此，$\mathbb{E}[xx^T|U] = I + DMD$有一个特征值等于 $2 + \varepsilon^2$，另有 $n - 1$ 个特征值等于 1，因此

$$\log\det(\mathbb{E}[xx^T|U]) = \log(2 + \varepsilon^2) \ge \log(2). \tag{3}$$

接下来考虑 $\mathbb{E}[xx^T]$。同样，我们有 $\mathbb{E}[xx^T] = \mathbb{E}[I + zz^T] = I + \mathbb{E}[zz^T]$。由于 $\mathcal{U}$是平衡的，对所有 $i \in [d] \setminus \{1\}$，$z(i) = \varepsilon/\sqrt{k}$ 以概率 $k/(d-1)$ 发生；而对于 $i,\ j \in [d] \setminus \{1\}$ 满足 $i \neq j$，$z(i) = z(j) = \varepsilon/\sqrt{k}$ 以概率 $\frac{k(k-1)}{(d-1)(d-2)}$ 发生。我们可以写作 $\mathbb{E}[zz^T] = D + E$。这里，$E_{11} = 0$、$E_{ij} = \frac{\varepsilon^2}{k} \cdot \frac{k(k-1)}{(d-1)(d-2)} = \frac{\varepsilon^2(k-1)}{(d-1)(d-2)}$对于$i,\ j \neq 1$成立，而对于 $i \neq 1$，$E_{i1} = E_{1i} = \frac{\varepsilon^2}{k} \cdot \frac{k}{d-1} = \frac{\varepsilon\sqrt{k}}{d-1}$。$D_{11} = 1$、$D_{ii} = \frac{\varepsilon^2}{d-1} - \frac{\varepsilon^2}{k} \cdot \frac{k(k-1)}{(d-1)(d-2)} \le \frac{\varepsilon^2}{d-1}$。

注意，$\|E\|_F \le \sqrt{2(d-1) \cdot \frac{\varepsilon^2 k}{(d-1)^2} + (d-1)^2 \cdot \frac{\varepsilon^4}{(k-1)^2(d-1)^2}} \le \sqrt{3} \cdot \varepsilon^2$。此外，$E$的秩为 3，因此仅有 3 个非零特征值。根据外尔不等式，

$$\lambda_1(\mathbb{E}[zz^T]) = \lambda_1(D + E) \le \lambda_1(D) + \lambda_1(E) \le 1 + \sqrt{3}\varepsilon^2.$$

对于 $i = 2, 3, 4, 5$,

$$\lambda_i(\mathbb{E}[zz^T]) = \lambda_i(D+E) \le \lambda_2(D) + \lambda_1(E) \le \frac{\varepsilon^2}{d-1} + \sqrt{3}\varepsilon^2 \le (\sqrt{3}+1)\varepsilon^2.$$

最后，对于 $i \ge 5$,

$$\lambda_i(\mathbb{E}[zz^T]) = \lambda_i(D+E) \le \lambda_2(D) + \lambda_4(E) = \frac{\varepsilon^2}{d-1} + 0 = \frac{\varepsilon^2}{d-1}.$$

因此,

$$\begin{aligned}
\log\det(\mathbb{E}[xx^T]) &= \log\det(I + \mathbb{E}[zz^T]) \\
&\le \log(2+\sqrt{3}\varepsilon^2) + 4\log(1+(\sqrt{3}+1)\varepsilon^2) + (d-5)\log\left(1+\frac{\varepsilon^2}{d-1}\right) \\
&\le \log(2) + (5\sqrt{3}+4+1)\varepsilon^2 \\
&\le \log(2) + 14\varepsilon^2. \qquad (4)
\end{aligned}$$

结合（3）可得 $\log\det(\mathbb{E}[xx^T]) - \log\det(\mathbb{E}[xx^T|U]) \le 14\varepsilon^2$，代回（2），得到 $I(U;\ X) \le 7n \cdot \varepsilon^2$ 即得所需结果。 □

将我们的互信息界代入 Fano 不等式，可得到近似恢复（即部分恢复）支撑集所需样本复杂度的下界。稍后，这将转化为草图所需的行数。

推论 15（样本复杂度下界）。 设 U, X 服从引理2 所述的分布，且满足 $n \le ck\log(d/k)\ \varepsilon^2$，其中常数 c ~~足够小~~。则不存在任何仅以 X 为输入、能够~~输出一个集合 U，使其满足 $|U| = k$ 且 $|U \cap U| > 19k/20$ 的算法，其成功概率为 $\ge 2/3$。

~~证明。 假设存在这样的算法。由于所有 U，$U' \in \mathcal{U}$ 都满足 $|U \cap U'| \le k \cdot 9/10$，若 $|U \cap U| > 19k/20$，则 U 必定包含 $> k/20$ 个不属于 U' 的元素，对任意 $U' \in \mathcal{U}$ 满足 $U' \ne U$ 的情形。因此，我们~~必须有 $|U \cap U'| < 19k/20$。于是 U 可用于唯一确定 U。也就是说，该算法以概率 U 识别出 $\ge 2/3$。然而，根据 Fano 不等式（引理 6），该算法失败的概率至少为

$$1 - \frac{I(U;X) + \log 2}{\log|\mathcal{U}|} \ge 1 - \frac{7\varepsilon^2 n + \log 2}{\log|\mathcal{U}|},$$

其中，$I(U;\ X)$ 的界由引理 2 得出。由于 $\log\ |\mathcal{U}| = \Theta(k\log(d/k))$，当 $n = ck\log(d/k)\,\varepsilon^2$ 时，该失败概率为 $> 1/3$；对于足够~~小的 c 和~~ d, k 大于足够大的常数时，情况亦然。这与算法以概率 $\ge 2/3$ 成功的假设相矛盾，因此推论得证。 □

这就结束了我们关于支持恢复困难性的证明的第二部分。对于第三部分，即将这个困难问题归约到稀疏线性回归问题，我们首先需要几个技术引理，然后才能最终证明定理 1。

第一个技术性结果建立了这样一种联系：将损失函数 L 近似到误差不超过 $(1 + O(\varepsilon))$，与揭示支撑集中的一个常数比例之间的联系。我们指出，在后续的归约中，L 将表示回归代价。

引理3。 设 $v \in \mathbb{R}^d$ 是一个 k-稀疏向量，其中有 k 个非零项等于 $1/\sqrt{k}$。设 x 是另一个k-稀疏向量。设 $M > \sqrt{n/(\varepsilon k)}$， $\alpha = |\mathrm{supp}(v) \cap \mathrm{supp}(x)|$，并且

$$L = 1 + \|x\|_2^2 + (1 - \varepsilon x^T v)^2 + \frac{M^2}{n}\left(\sum_i x(i) - \sqrt{k}\right)^2.$$

存在一个常数 c，使得任何满足 $\alpha < 19k/20$ 的 x 都不是 L 的 $1 + c\varepsilon$ 近似解。

证明。令 $S = \mathrm{supp}(x) \cap \mathrm{supp}(v)$ 和 $M' = M/\sqrt{n}$。令 $\beta = \sum_i x(i)$和 γ使得$\sum_{i\in S} x(i) = \gamma\beta$。我们将对 $\mathbb{R}$中 β 和 γ 的所有可能取值优化 L。注意，任何使 L 最小化的 x都必须具有以下形式

$$x(i) = \begin{cases} \frac{\gamma\beta}{\alpha} & \text{for } i \in S, \\ \frac{(1-\gamma)\beta}{k-\alpha} & \text{for } i \in \mathrm{supp}(x) \setminus S. \end{cases}$$

这是因为对于固定的 S, β, γ，令 x 具有上述形式，可在不影响 $x^T v$ 的情况下最小化 $\|x\|$。因此，

$$\begin{aligned} L &= 1 + \frac{\gamma^2\beta^2}{\alpha} + \frac{(1-\gamma)^2\beta^2}{k-\alpha} + \left(1 - \varepsilon\alpha\frac{1}{\sqrt{k}}\frac{\gamma\beta}{\alpha}\right)^2 + M'^2\left(\beta - \sqrt{k}\right)^2 \\ &= \gamma^2\left(\frac{\beta^2}{\alpha} + \frac{\beta^2}{k-\alpha} + \varepsilon^2\frac{\beta^2}{k}\right) - 2\gamma\left(\frac{\beta^2}{k-\alpha} + \varepsilon\frac{\beta}{\sqrt{k}}\right) + 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2. \end{aligned}$$

对所有 $\gamma \in \mathbb{R}$ 进行最小化得到

$$\begin{aligned} \min_\gamma L &= 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2 - \frac{\left(\frac{\beta^2}{k-\alpha} + \varepsilon\frac{\beta}{\sqrt{k}}\right)^2}{\frac{\beta^2}{\alpha} + \frac{\beta^2}{k-\alpha} + \varepsilon^2\frac{\beta^2}{k}} \\ &= 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2 - \frac{\frac{\beta^2}{(k-\alpha)^2} + \frac{2\beta\varepsilon}{(k-\alpha)\sqrt{k}} + \frac{\varepsilon^2}{k}}{\frac{k}{\alpha(k-\alpha)} + \frac{\varepsilon^2}{k}} \\ &= 2 + M'^2(\beta - \sqrt{k})^2 - \frac{\frac{\beta^2}{(k-\alpha)^2}\left(1 - \frac{k}{\alpha}\right) + \frac{\beta^2\varepsilon^2}{k(k-\alpha)} + \frac{2\beta\varepsilon}{(k-\alpha)\sqrt{k}} + \frac{\varepsilon^2}{k}}{\frac{k}{\alpha(k-\alpha)} + \frac{\varepsilon^2}{k}} \\ &= 2 + M'^2(\beta - \sqrt{k})^2 + \frac{\beta^2 k - \beta^2\varepsilon^2\alpha - 2\beta\sqrt{k}\alpha\varepsilon - \varepsilon^2(k-\alpha)\alpha}{k^2 + \varepsilon^2(k-\alpha)\alpha} \\ &\approx_\varepsilon 2 + M'^2(\beta - \sqrt{k})^2 + \frac{\beta^2 k - 2\beta\sqrt{k}\alpha\varepsilon}{k^2} \\ &= \beta^2\left(M'^2 + \frac{1}{k}\right) - \beta\left(2M'^2\sqrt{k} + \frac{2\alpha\varepsilon}{k\sqrt{k}}\right) + 2 + M'^2 k. \end{aligned}$$

对 β 进行最小化得到

$$\min_\beta \min_\gamma L \approx_\varepsilon 2 + M'^2 k - \frac{\left(M'^2\sqrt{k} + \frac{\alpha\varepsilon}{k\sqrt{k}}\right)^2}{M'^2 + \frac{1}{k}}$$

$$= 2 + \frac{M'^2 - \frac{2M'^2\alpha\varepsilon}{k} - \frac{\alpha^2\varepsilon^2}{k^3}}{M'^2 + \frac{1}{k}}$$
$$\approx_\varepsilon 2 + \frac{M'^2 - \frac{2M'^2\alpha\varepsilon}{k}}{M'^2 + \frac{1}{k}}.$$

由 $M' = M/\sqrt{n} \geq 1/\sqrt{\varepsilon k}$ 可得

$$\min_\beta \min_\gamma L \approx_\varepsilon 3 - 2\frac{\alpha}{k}\varepsilon.$$

可以看出，当 $\alpha = k$ 在 $3 - 2\varepsilon$ 时，右侧达到最小值。此外，如果 $\alpha < c_1 k$ 对于 $c_1 < 1$，则其至少为

$$3 - 2c_1\varepsilon \geq (3 - 2\varepsilon)\left(1 + \frac{2(1 - c_1)}{3}\varepsilon\right).$$

因此，存在一个足够小的常数 c，使得如果 $\alpha < 19k/20$，则无法在 $1 + c\varepsilon$ 因子范围内最小化 L。□

下一个工具将帮助我们通过消除输入分布中随机高斯矩阵的影响，确定性地分析回归代价，误差上界为 $1 \pm \varepsilon$。根据高斯矩阵的标准尾界（例如 [Ver18]中的练习 4.7.3），我们有以下引理。

引理 4。设 X 是一个协方差为 Σ. 的 $n \times d$ 高斯矩阵。那么存在一个常数C，使得以概率 $\geq 1 - \delta$

$$(1 - \varepsilon)\Sigma \preceq \frac{1}{n}X^T X \preceq (1 + \varepsilon)\Sigma,$$

当 $n \geq C\frac{d + \log(1/\delta)}{\varepsilon^2}$.

将上述引理应用于一个固定的 k维子空间，并对所有 $\binom{d}{k}$ k维子空间应用并集界，可得其成立的概率至少为 $1 - \delta$

$$\forall k\text{-sparse vectors } v \in \Psi_k \colon (1 - \varepsilon)v^T\Sigma v \preceq \frac{1}{n}v^T X^T X v \preceq (1 + \varepsilon)v^T\Sigma v,$$

当 $n \geq C'^k \log(d/(k\delta))\,\varepsilon^2$ 对于某个 $C' > C$。

利用前面的结果，我们现在已经准备好给出定理 1的主要证明，并由此完成我们证明概要的第三个也是最后一个部分。

我们构造了一个 ℓ_2-回归实例，对于该实例，任何从一个 oblivious 草图中得到的、与任意估计器配对的 $1 + \Theta(\varepsilon)$ 近似都会揭示出其支撑集中的一个常数比例。因此，该困难性结果也适用于回归问题。

证明。（定理 1）设 $X = [b\ A]$ 服从 $G(I + zz^T)^{1/2}$ 分布，其中 z 的分布如引理 2 所述。我们将构造一个 k-稀疏 ℓ_2 回归问题，基于 X，使得一个 $1 + \varepsilon$ 近似（针对所构造的问题）能够恢复足够大的比例，即大于 $19/20$ 的比例，这一比例取自 z 的支撑集。我们可以不失一般性地假设草图矩阵具有标准正交的行，并且由于 X 是高斯矩阵，该草图的行采样自与 X 的行相同的分布。之所以如此，是因为我们是在针对草图上的任意估计量证明下界。

根据推论 15，若样本数小于 $ck\log(k/d)/\varepsilon^2$，则没有任何算法能够恢复超过 $19/20$ 比例的支撑集部分，该支撑集属于 z，且概率大于 $2/3$。因此，我们有下界 $\Omega(k\log(d/k)/\varepsilon^2)$，该下界针对 $1+\varepsilon$ 近似，其对象为 k-稀疏 ℓ_2 回归。

考虑以下 ℓ_2 稀疏回归问题

$$\min_{x\in\Psi_k}\left\|\begin{bmatrix} M\,M\,\dots\,M \\ A \end{bmatrix}x-\begin{bmatrix}\sqrt{k}M\\ b\end{bmatrix}\right\|_2.$$

我们令 M 为一个非常大的数，这会强制 $\sum_i x(i)$接近$\sqrt{k}$。上述回归问题的平方损失为

$$L=\|Ax-b\|_2^2+M^2\left(\sum_i x(i)-\sqrt{k}\right)^2=\|X\tilde{x}\|_2^2+M^2\left(\sum_i x(i)-\sqrt{k}\right)^2,$$

其中 $\tilde{x}=(-1,\ x)$。根据引理 4 中的矩阵集中性，对于 $n=\Omega(k\log(d/(k\delta))/\varepsilon^2$，以至少 $1-\delta$ 的概率，

$$X^TX\approx_\varepsilon n(I+zz^T).$$

回顾一下，z 对于所有 $i\in U\setminus\{1\}$ 都有 $z(1)=1$，$z(i)=\varepsilon/\sqrt{k}$，并且对于所有 $i\notin U$都有 $z(i)=0$。我们有

$$\frac{1}{n}\|X\tilde{x}\|_2^2\approx_\varepsilon\|\tilde{x}\|_2^2+(\tilde{x}^Tz)^2=(1+\|x\|_2^2)+(1-\varepsilon x^Tv)^2,$$

其中 v 满足 $z=(1,\ v)$。

因此，

$$\frac{L}{n}\approx_\varepsilon 1+\|x\|_2^2+(1-\varepsilon x^Tv)^2+\frac{M^2}{n}\left(\sum_i x(i)-\sqrt{k}\right)^2.$$

注意，x 是对 ℓ_2 损失的 $1+\varepsilon$ 近似，当且仅当 x 是对 ℓ_2^2 损失的 $1+\Theta(\varepsilon)$近似。

令 $\alpha=|\mathrm{supp}(x)\cap\mathrm{supp}(v)|$。根据引理 3，存在常数 c，使得如果我们能在因子 $1+c\varepsilon$以内近似 L，则必有 $\alpha>19k/20$。对 ε 重新缩放，并结合推论 15 以及并集界（将 δ 设为一个较小的常数），即可证明该定理。 □

接下来，我们将下界扩展到所有 $p\geq 1$的 ℓ_p。

定理2。 令 $A\in\mathbb{R}^{n\times d},b\in\mathbb{R}^n$。假设 $S\in\mathbb{R}^{m\times n}$ 是针对任意 $p\geq 1$ 的 k-稀疏 ℓ_p 回归的无偏线性草图，并配有估计器 $E_p(SA,Sb,x)$，使得极小化解$\tilde{x}\in\mathrm{argmin}_{x\in\Psi_k}E_p(SA,Sb,x)$ 满足 $\|A\tilde{x}-b\|_p\leq(1+\varepsilon)\min_{x\in\Psi_k}\|Ax-b\|_p$。那么 $m=\Omega(k\log(d/k)/\varepsilon^2)$。

证明。（定理2的证明）我们利用 Dvoretzky 定理所保证的事实，从 ℓ_2 情形进行归约：对于所有 $p\geq 1$，ℓ_2 都可以以至多 $(1\pm\varepsilon)$ 的失真无视具体输入地嵌入到 ℓ_p 中。事实上，可以通过随机映射 $G\in\mathbb{R}^{r\times n}$ 构造这样的嵌入，其中各项是经过适当缩放的独立同分布高斯随机变量。特别地，G 是一个无视具体输入的线性映射。当 $1\leq p\leq 2$ 时，行数为 $r=O(n\log(1/\varepsilon)/\varepsilon^2)$；当 $p>2$时，行数为 $r=n^{O(p)}$；[参见 Mat13，第 30] 页。不过，我们注意到，在本文的语境中，G 的行数并不重要，因为在下文中会通过应用 S 将其缩减。

更精确地说，对于所有 $x\in\mathbb{R}^d$，都有 $(1-\varepsilon)\|Ax-b\|_2\leq\|GAx-Gb\|_p\leq(1+\varepsilon)\|Ax-b\|_2$。现在假设 S 是针对 ℓ_p 中 k-稀疏回归问题的无关数据的草图矩阵。那么我们可以求得 $\tilde{x}\in\mathrm{argmin}_{x\in\Psi_k}E_p(SGA,SGb,x)$。根据定义，有 $\|G(A\tilde{x}-b)\|_p\leq$

$(1+\varepsilon)\|G(Ax_G^*-b)\|_p$，其中 $x_G^*\in\operatorname{argmin}_{x\in\Psi_k}\|G(Ax-b)\|_p$。另外，令 $x^*\in\operatorname{argmin}_{x\in\Psi_k}\|Ax-b\|_2$ 为 ℓ_2 问题的极小化解。现在可得

$$\begin{aligned}\|A\tilde{x}-b\|_2 &\le \|G(A\tilde{x}-b)\|_p/(1-\varepsilon)\le \|G(Ax_G^*-b)\|_p(1+\varepsilon)/(1-\varepsilon)\\ &\le \|G(Ax^*-b)\|_p(1+\varepsilon)/(1-\varepsilon)\le \|Ax^*-b\|_2(1+\varepsilon)^2/(1-\varepsilon)\\ &\le (1+7\varepsilon)\|Ax^*-b\|_2,\end{aligned}$$

通过重新缩放 ε，这意味着 SG是针对 ℓ_2-范数问题的无先验线性草图，其 ℓ_p-范数最小化估计器的误差至多为 $1+\varepsilon$。利用定理 ℓ_2 中给出的下界 1，可得 SG，因此 S 也有 $m=\Omega(k\log(d/k)/\varepsilon^2)$ 行。□

A.2 ReLU 和类铰链损失函数的下界

这里我们给出与定理 2类似的进一步归约，以便将主要结果推广到 ReLU 和类铰链损失函数。

推论 3. 设 $A\in\mathbb{R}^{n\times d}, b\in\mathbb{R}^n$。假设 $S\in\mathbb{R}^{m\times n}$ 是针对 k-稀疏ReLU 回归的无先验线性草图，且具有估计器 $E_{\mathrm{ReLU}}(SA,Sb,x)$，使得 $\tilde{x}\in\operatorname{argmin}_{x\in\Psi_k}E_{\mathrm{ReLU}}(SA,Sb,x)$ 满足 $\|A\tilde{x}-b\|_{\mathrm{ReLU}}\le(1+\varepsilon)\min_{x\in\Psi_k}\|Ax-b\|_{\mathrm{ReLU}}$。那么 $m=\Omega(k\log(d/k)/\varepsilon^2)$。

证明。注意 $\|Ax-b\|_1=\|Ax-b\|_{\mathrm{ReLU}}+\|-(Ax-b)\|_{\mathrm{ReLU}}$，$\forall x$. 因此，给定 A，b我们有 $\|Ax-b\|_1=\|PAx-Pb\|_{\mathrm{ReLU}}$，其中

$$P=\begin{bmatrix} I_n\\ -I_n\end{bmatrix}$$

是一个 $2n\times n$ 矩阵。假设 $S\in\mathbb{R}^{m\times 2n}$ 是一个用于 k-稀疏 ReLU 回归的无关线性草图，其估计器为 $E_{\mathrm{ReLU}}(SA,Sb,x)$。这意味着 $\tilde{x}\in\operatorname{argmin}_{x\in\Psi_k}E_{\mathrm{ReLU}}(SPA,SPb,x)$ 满足

$$\|A\tilde{x}-b\|_1=\|P(A\tilde{x}-b)\|_{\mathrm{ReLU}}\le(1+\varepsilon)\min_{x\in\Psi_k}\|P(Ax-b)\|_{\mathrm{ReLU}}=(1+\varepsilon)\min_{x\in\Psi_k}\|Ax-b\|_1.$$

因此，$S\cdot P$ 是用于 ℓ_1 回归的无关输入线性草图。根据定理 2，我们有$m=\Omega(k\log(d/k)/\varepsilon^2)$。□

定义 1. 我们称 $f(\cdot)$ 为一个 (L,a_1,a_2) 类铰链损失函数，当且仅当 f 是 L *Lipschitz* 连续的，$\forall x\ge 0$: $f(x)\ge a_2>0$，且 $\forall x$: $|f(x)-\mathrm{ReLU}(x)|\le a_1$.

逻辑损失函数 $\log(1+e^{-x})$ 和铰链损失函数 $\max(0,1-x)$ 分别是 (1，ln(2)，ln(2)) 和 (1，1，1) 类铰链损失函数。

我们将使用记号 $a=(1\pm\varepsilon)b$ 来表示 $(1-\varepsilon)b\le a\le(1+\varepsilon)b$. 我们注意到，对于 $\varepsilon>0$，若 $\|x\|_{\mathrm{ReLU}}>0$，则存在一个常数 $c>0$ 使得

$$c\|x\|_{\mathrm{ReLU}}=\|cx\|_{\mathrm{ReLU}}=(1\pm\varepsilon)\|cx\|_f.$$

推论 4. 设 $A\in\mathbb{R}^{n\times d}, b\in\mathbb{R}^n$。假设 $S\in\mathbb{R}^{m\times n}$ 是针对某个类铰链损失函数 f的无关子空间嵌入，并配有估计量 $E_f(SA,Sb,x)$，满足 $\forall x\in\Psi_k$: $(1-\varepsilon)\|Ax-b\|_f\le E_f(SA,Sb,x)\le(1+\varepsilon)\|Ax-b\|_f$. 那么 $m=\Omega(k\log(d/k)/\varepsilon^2)$。

证明。假设 $S\in\mathbb{R}^{m\times n}$ 是针对 f的无感知子空间嵌入，其估计器为 $E_f(SA,Sb,x)$.

设 $x \in \Psi_k$ 是某个 k-稀疏向量。我们将证明， $S \in \mathbb{R}^{m\times n}$ 是一个适用于 ReLU 的无偏子空间嵌入，其估计器

$$E_{\mathrm{ReLU}}(SA, Sb, x) = \lim_{c\to\infty} \frac{E_f(ScA, Scb, x)}{c}.$$

首先，考虑 $\|Ax-b\|_{\mathrm{ReLU}} > 0$的情形。当 c 足够大时，我们有

$$\begin{aligned}\|Ax-b\|_{\mathrm{ReLU}} &= \frac{\|c(Ax-b)\|_{\mathrm{ReLU}}}{c} \\ &= (1\pm\varepsilon)\frac{\|c(Ax-b)\|_f}{c} \\ &= (1\pm 3\varepsilon)\frac{E_f(ScA, Scb, x)}{c}.\end{aligned}$$

这里，我们假设 ε 足够小，从而$(1+\varepsilon)^2 \le 1+3\varepsilon$。

接下来，我们考虑 $\|Ax-b\|_{\mathrm{ReLU}} = 0$的情形。

我们有

$$\begin{aligned}E_f(ScA, Scb, x) &\le (1+\varepsilon)\|c(Ax-b)\|_f \\ &\le (1+\varepsilon)na_1.\end{aligned}$$

因此，

$$\lim_{c\to\infty} \frac{E_f(ScA, Scb, x)}{c} \le \lim_{c\to\infty} \frac{2na_1}{c} = 0$$

如所欲证。 □

A.3 稀疏回归中采样失效

本节论证，基于采样的算法虽然是非稀疏回归草图中的重要工具 [DMM06; Das+09; Mun+18; MMR21], 但在稀疏情形下无法得出任何非平凡的结果。具体而言，这些算法无法将数据压缩到非稀疏情形所能达到的程度以下，也就是无法压缩到输入矩阵的秩以下。这正是我们的所有上界都基于一般线性草图，而非采样的原因。

定理 5。 考虑任意有界的近似因子 $\alpha \ge 1$，以及任意 $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}^{\ge 0}$，它在全零向量上取值为$0$，在任何其他向量上取某个正值。对于任意$n > 9$，存在某个输入矩阵 $A \in \mathbb{R}^{n\times n}$ 以及向量上的某个分布 $b \in \mathbb{R}^n$，使得对于任意满足 $m < n/3$的采样矩阵 $S \in \mathbb{R}^{m\times n}$，任何仅访问 SA 和 Sb 的算法都无法输出一个k-稀疏的 $\tilde{x} \in \Psi_k$，使其满足 $\|A\tilde{x}-b\| \le \alpha \cdot$最小值$_{x\in\Psi_k}\|Ax-b\|$，且该算法以至少 $1/2$ 的概率做到这一点（概率取自 b 的选择以及算法可能采用的任何随机性）。

证明。 令 $k=1$ 和 A 为 $n\times n$ 单位矩阵。令 b 以概率 $1/n$设为 i^{th} 标准基向量。注意到 $\min_{1\text{-sparse } x\in\mathbb{R}^n}\|Ax-b\| = 0$，因此若要达到任何有界近似因子，算法必须以零成本输出 x，即 $x=b$。任何访问 $m < n/3$ 行 A 和 b 的算法，以至少 $2/3$的概率只能在 Sb 中看到零元素。令 $\mathcal{X}_0$ 为给定 $Sb=0$时算法输出的分布。只有当 $x \sim \mathcal{X}_0$ 满足 $x=b$时，算法才能达到有界近似因子。由于在观察到 $Sb=0$后， b 剩余的 $2/3\cdot n$ 种可能情况都同样可能，因此这种情况发生的概率至多为 $\frac{3}{2n}$。因此，对于 $n > 9$，算法成功的概率至多为 $1/3 + 3/(2n) < 1/2$。

□

注意，定理 5 中的采样矩阵 S 可以依赖于 A，但不能依赖于 b。这是必要的。如果采样矩阵可以依赖于 b，那么在不受计算能力限制的情况下，采样算法理论上可以将问题压缩为 $O(k/\varepsilon)$ 行。具体而言，它只需求出最优的 $x^* = \text{argmin}_{k\text{-sparse } x \in \Psi_k} \|Ax - b\|$，并且只考虑 A 中位于 x^* 支撑集内的 k 列。例如，在 ℓ_2 情形下，只需对这些列应用标准杠杆分数采样的变体，[Woo14; CP19]，即可输出采样矩阵 $S \in \mathbb{R}^{m \times n}$，用于 $m = O(k/\varepsilon)$，使得$\tilde{x} = \text{argmin}_{k\text{-sparse } x \in \Psi_k} \|SAx - Sb\|$ 满足 $\|A\tilde{x} - b\| \le (1+\varepsilon) \min_{k\text{-sparse } x \in \Psi_k} \|Ax - b\|$。

如果要求采样矩阵对每个 k-稀疏的 $x \in \Psi_k$ 都保留 $\|Ax - b\|$，那么即使采样算法可以读取 b，也很容易看出它至少必须采样 n 行。即使在 $b = 0$ 的情况下，这一点也成立。

定理 6. 考虑任意有界近似因子 $\alpha \ge 1$ 以及任意 $\|\cdot\|$：$\mathbb{R}^n \to \mathbb{R}^{\ge 0}$，其在全零向量上取值为0，而在任何其他向量上取某个正数。对于任意$n > 9$，存在某个输入矩阵 $A \in \mathbb{R}^{n \times n}$，使得不存在采样矩阵 $S \in \mathbb{R}^{m \times n}$ 满足$m < n$，对所有 k-稀疏 $x \in \Psi_k$均满足 $\alpha^{-1}\|Ax\| \le \|SAx\| \le \alpha\|Ax\|$。

证明。令 A 为 $n \times n$单位矩阵。那么，如果我们采样 $m < n$行，当 x 至少是一个标准基向量时，我们将有 $\|SAx\| = 0$。这违反了近似界。 □

B 稀疏回归的 k上界

B.1 $p \in [1, 2]$的 ℓ_p-范数损失函数的上界

我们证明了 k-稀疏仿射嵌入的上界，并指出作为推论，我们可以得到最小化的相同上界。我们从 ℓ_2开始。

定理7。 令 $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$。存在一个关于随机矩阵 $S \in \mathbb{R}^{m \times n}$且具有 $m = O(k \log(d/k)/\varepsilon^2)$ 的分布，使得以常数概率有 $\forall x \in \Psi_k : (1-\varepsilon)\|Ax - b\|_2 \le \|S(Ax - b)\|_2 \le (1+\varepsilon)\|Ax - b\|_2$。

证明。（定理7的证明）上界类似于通过Johnson-Lindenstrauss嵌入 [JL84], 构造RIP矩阵的已知构造 [Bar+08]，即适当重新缩放的高斯矩阵。主要区别在于，由 x 的任意固定 k-稀疏支撑所形成的子空间不必正交，也不必与标准基向量对齐。第一个想法是，至多存在 $\binom{d}{k} \le (ed/k)^k$ 个不同的 k-稀疏支撑，并且其中每一个对应于 A的一种 k列选择。每一种这样的选择都张成一个维数为 $\le k$的 k维线性子空间。根据 [Sar06], 中的子空间嵌入构造，由一种 k 列选择所形成的每个子空间，都可以通过嵌入覆盖该子空间中单位球、大小为$(3/\varepsilon)^k$ 的网中的点来处理。其余向量可通过三角不等式与网点相关联，并且由线性性，该嵌入可延拓到单位球面之外任意范数的向量。 [Woo14第 13]页中一个稍更精细的论证指出，该网可以用绝对常数 $\varepsilon_0 := 1/2$替换 ε 来构造。因此，需要嵌入至多$(1 \pm \varepsilon)$失真的点的总数由 $|\mathcal{N}| \le (d/k)^k \cdot c^k$ 界定，其中 $c = 3e/\varepsilon_0$为绝对常数。该嵌入可通过Johnson-Lindenstrauss引理，再对 $\mathcal{N}$使用并集界来完成，从而得到矩阵$S = \frac{1}{\sqrt{m}} G$，其元素为缩放后的独立同分布标准高斯 $G_{ij} \sim N(0, 1)$，行数为$m = O(\log(|\mathcal{N}|)/\varepsilon^2) = O(k \log(d/k)/\varepsilon^2)$。[3]

—=

□

[3]我们注意到，使用随机符号（Rademacher）矩阵也可以获得相同的结果 [CW09] 这在流式计算和其他空间受限的场景中更加方便。

我们继续讨论 ℓ_p，$p \in [1,2)$。我们注意到，其思路类似于 [Bac+16，附录F。1]但我们的结果经过非平凡的改写以适应稀疏设定，并推广到 ℓ_p。

定理8. 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n, p \in [1,2)$。存在一个随机矩阵$S \in \mathbb{R}^{m\times n}$ 上的分布，使得 $m = O(k(\log(d/k) + \log(k/(\varepsilon\delta)))/\varepsilon^2)$ 以至少$1-\delta$ 的概率满足 $\forall x \in \Psi_k : (1-\varepsilon)\|Ax-b\|_p \le \|S(Ax-b)\|_{\text{med}} \le (1+\varepsilon)\|Ax-b\|_p$；其中对于任意的$y \in \mathbb{R}^d$, $\|y\|_{\text{med}} :=$ 中位数$\{|y_i| \mid i \in [d]\}$。

证明。(定理 8的证明) 我们选取 $S \in \mathbb{R}^{m\times n}$ 为一个矩阵，其元素是独立同分布的 p-稳定随机变量，尺度参数为 $\gamma_p = c^{1-1/p}$，其中 $c \approx 1.099055$；详见下文。我们将证明该矩阵具有所需性质。首先注意到 $Ax-b = [A,b][x^T,-1]^T$，因此我们可以直接假设输入仅由 $A' = [A,b]$ 构成，只需证明 $\|SA'x\|_{\text{med}} = (1\pm\varepsilon)\|A'x\|_p$对所有 $x \in \Psi_{k'} \subseteq \mathbb{R}^{d+1}$ 在 $k' = k+1$的情况下均成立。下文为便于表述，我们将 A 重新代入 A'，并将 k 重新代入 k'。

固定任意一个由 $I \subseteq [d]$ 索引的 k-稀疏支撑集，其中 $|I| = k$。令 A_I 为矩阵，其 k 列是 A 中满足 $i \in I$ 的列 A_{*i}。根据 Auerbach 的一个经典结果 [（参见 Aue30；Das+09]），可知由这些列张成的 p-范数子空间存在一组基 $L \in \mathbb{R}^{n\times k}$ WW22，引理 2.22]，满足以下性质：每一列 $i \in [k]$ 的 ℓ_p 范数恰好为 $\|L_{*i}\|_p = k^{1/q}$，并且对于所有 $x \in \mathbb{R}^k$，均有 $\|Lx\|_p \ge \|x\|_q$，其中，对于 $p = 1$，$q = \infty$ 表示 $p\in$ 的对偶范数，而对于 $p \in$，$q = \frac{p}{p-1}$ (1, 2) 表示 p 的对偶范数。

此外，根据 p-稳定随机变量的一个性质，我们有任意条 y

$$(SL)_{ij} = \sum_{h=1}^{n} S_{ih}L_{hj} \sim C\cdot\|L_{*j}\|_p = C\cdot k^{1/q}$$

目也是一个缩放后的 p-稳定随机变量；[参见Ind06；Dyt+18]。因此，对于任意阈值 τ，SL 中任意元素的绝对值大于 τ 的概率至多为 $O((k^{1/q}/\tau)^p)$ $=O(k/\tau^p)$，[参见B·LM18]。

令 $\tau = O((mk^2/\delta)^{1/p})$ $= \tilde{O}((k^3\log(d)/\delta)^{1/p})$，则 SL 的所有元素同时以概率 $1-\delta/2$有界于 τ。假设此事件成立。那么，对于所有具有由 I索引的固定支撑集的 $x \in \Psi_k$，根据 Hölder 不等式（仅应用于 k-稀疏支撑集，因为其他项均为零）以及上述讨论的性质，我们有

$$\begin{aligned}\|SLx\|_\infty &= \max_{i\in[m]}|S_{i*}Lx| \le \max_{i\in[m]}\|(SL)_{i*}\|_\infty\|x\|_1 \le \tau k^{1-\frac{1}{q}}\|x\|_q \le \tau k^{\frac{1}{p}}\|Lx\|_p \\ &\le \tilde{O}((k^4\log(d)/\delta)^{1/p})\|Lx\|_p.\end{aligned}$$

令 $\tau' = \tilde{O}((k^4\log(d)/\delta)^{1/p})$ $= \tilde{O}(k^4\log(d)/\delta)$。我们在 ℓ_p范数下，为单位 ℓ_p 球与由 L张成的子空间的交集构造一个 $\frac{\varepsilon}{\tau'}$-网 $\mathcal{N}_I^k$。根据线性性质，将限制在单位球上不失一般性。我们对每个 k-稀疏支撑集重复这一构造，并将所有支撑集对应的网的并集定义为我们的网，即 $\mathcal{N} = \bigcup_{I\subseteq[d],|I|=k}\mathcal{N}_I^k$。不同子空间至多有 $\binom{d}{k} \le (ed/k)^k$ 个，根据标准的体积论证，每个子空间都可由大小至多为 $|\mathcal{N}_I^k| \le (3\tau'/\varepsilon)^k$的网覆盖。因此，对于某个绝对常数 c_1，我们有 $|\mathcal{N}| \le (ed/k)^k\cdot(c_1k^4\log(d)/(\varepsilon\delta))^k$ $=\exp(O(k\log(d/k) + k\log(k/(\varepsilon\delta))))$。

接下来，我们研究随机变量 $F_p(x)$ 的累积分布函数 $|X|$，其中 X 服从 p-稳定分布。除 $p \in \{1,2\}$ 这些对应柯西分布和正态分布的情形外，累积分布函数和概率密度函数均不存在已知的解析／闭式表达式。因此，我们转而利用 L´evy [L´évy； Mas77]基于特征函数的反演定理，对于

已知其闭式表达式 [BHW05；Dyt+18]: $\phi_p(t) = \exp -|\gamma_p t|^p$，其中 γ_p 是上文定义的常数尺度参数。更确切地说，有

$$\begin{aligned} F_p(x) &= P(|X| \le x) = P(X \le x) - P(X \le -x) \\ &= \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{itx} - e^{-itx}}{it} \phi_p(t)\, dt \\ &= \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{itx} - e^{-itx}}{it} e^{-|\gamma_p t|^p}\, dt \\ &= \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{2i\sin(tx)}{it} e^{-|\gamma_p t|^p}\, dt \\ &= \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{\sin(tx)}{t} e^{-|\gamma_p t|^p}\, dt. \end{aligned}$$

由此可得

$$F_p(1) = \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{\sin t}{t} e^{-|\gamma_p t|^p}\, dt = \frac{1}{2} \tag{5}$$

还需证明 F的导数在 $F(1) = \frac{1}{2}$处有界。为此我们注意到

$$\begin{aligned} F_p'(x) &= \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{\cos(tx)\cdot t}{t}\, e^{-|\gamma_p t|^p}\, dt \\ &= \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \cos(tx)\, e^{-|\gamma_p t|^p}\, dt. \end{aligned}$$

因此，

$$F_p'(1) = \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \cos(t)\, e^{-|\gamma_p t|^p}\, dt.$$

现在，利用被积函数的对称性以及特征函数关于指数 p 的单调性，我们有

$$\begin{aligned} \frac{1}{\pi} = \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \cos(t)\, e^{-|\gamma_1 t|}\, dt \le F_p'(1) &\le \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \cos(t)\, e^{-|\gamma_2 t|^2}\, dt \\ &\le \frac{1}{\pi} \lim_{T\to\infty} \int_{-T}^{T} \cos(t)\, e^{-|t|^2}\, dt = \frac{1}{\pi^{1/2} e^{1/4}} \end{aligned} \tag{6}$$

其中特别注意， $\gamma_1 = 1$用于下界，而 $\gamma_2 \ge 1$ 用于上界。

推广 [Ind06，引理 2], 由方程（5）和（6）可知，若 $F_p(z) \in [1/2 - c\varepsilon,\ 1/2 + c\varepsilon]$ 对于某个绝对常数 c 则 $z \in [1-\varepsilon, 1+\varepsilon]$，下文将用到这一点。对于任意 $x \in \Psi_k$，我们称 SAx 是良好的，如果草图空间中只有 $\frac{1}{2} - c_2\varepsilon$ 比例的坐标过大或过小，即

$$\begin{aligned} |\{i : |(SAx)_i| < (1-\varepsilon)\|Ax\|_p\}| &\le \left(\frac{1}{2} - c_2\varepsilon\right) m \\ |\{i : |(SAx)_i| > (1+\varepsilon)\|Ax\|_p\}| &\le \left(\frac{1}{2} - c_2\varepsilon\right) m \end{aligned}$$

对于某个较小的常数 c_2。如果 SAx 是良好的，那么对于任何 y 至多有 $c_2\varepsilon m$ 个坐标大于 $\varepsilon\|Ax\|_p$的

$$(1-2\varepsilon)\|Ax\|_p \le \|SAx+y\|_{\text{med}} \le (1+2\varepsilon)\|Ax\|_p. \tag{7}$$

根据 p-稳定性，$(SAx)_i$ 是一个尺度为 $\|Ax\|_p$ 的 p-稳定随机变量，因此我们有。

$$\Pr[|(SAx)_i| < (1-\varepsilon)\|Ax\|_p] < 1/2 - \Omega(\varepsilon)$$
$$\Pr[|(SAx)_i| > (1+\varepsilon)\|Ax\|_p] < 1/2 - \Omega(\varepsilon).$$

根据Chernoff 界，当 c_2 足够小时，我们有 SAx 以至少 $1-\exp(-\Omega(\varepsilon^2 m))$的概率是良好的。对于我们对 m的选择，我们可以使用并集界，以至少 $1-\exp(-\Omega(\varepsilon^2 m))\cdot|\mathcal{N}| \ge 1-\delta^{\Omega(k)}$ 的概率得到：对于所有 $x\in\mathcal{N}$， SAx同时是 良好的 。假设该事件为真。

那么每个 $y=Ax$ 对于 $x\in\Psi_k$ 具有 $\|y\|_p=1$ 可表示为 $y=z+\eta$ 其中 $z\in\mathcal{N}$ 且 $\|\eta\|_p \le \varepsilon/\tau'$。我们有 Sz是 良好的 并且 $\|S\eta\|_\infty \le \tau'\|\eta\|_p \le \varepsilon$。因此根据（7），

$$(1-2\varepsilon)\|z\|_p \le \|S(z+\eta)\|_{\text{med}} \le (1+2\varepsilon)\|z\|_p.$$

这意味着

$$(1-3\varepsilon)\|y\|_p \le \|Sy\|_{\text{med}} \le (1+3\varepsilon)\|y\|_p.$$

由于 S 是线性的，因此无需限制为 $\|y\|_p=1$ 。对 ε 进行缩放即可完成证明。 □

B.2 ReLU损失函数的上界

记号 对于函数 $f:\ \mathbb{R}\to\mathbb{R}$ 和向量 $y\in\mathbb{R}^n$，我们用 $f(y)\in\mathbb{R}^n$ 表示将 f 逐元素作用于 y。令 y_i 表示 y的第 i^{th} 个元素。因此 $f(y)_i = f(y_i)$。此外，令$\|y\|_f = \sum_{i=1}^n f(y_i)$。令 y^+和 y^- 分别表示限制在正元素和负元素集合上的 y。最后，对于 $A\in\mathbb{R}^{n\times d}$ 和 $x\in\mathbb{R}^d$，令 $\mu(A) = \sup_{x\neq0}(Ax)^+(Ax)^-$。当 A 的含义可由上下文确定时，我们在 μ的记号中省略 A 。

定理 9. 设 $A\in\mathbb{R}^{n\times d}, b\in\mathbb{R}^n$。存在一个无偏草图 S，其行数为 $O(\frac{\mu^2 k}{\varepsilon^2}\log(\frac{\mu d}{\varepsilon\delta}))$，以及一个估计量 $g_{\text{ReLU}}(SA,Sb,x)$，使得其概率至少为 $1-\delta$，我们有 $\forall x\in\Psi_k$:$(1-\varepsilon)\|Ax-b\|_{\text{ReLU}} \le g_{\text{ReLU}}(SA,Sb,x) \le (1+\varepsilon)\|Ax-b\|_{\text{ReLU}}$.

证明。 将 b 附加到 A 并将 k 增加 1 后，只需证明定理在 $b=0$ 情况下成立。注意，

$$\|Ax\|_{\text{ReLU}} = \|(Ax)^+\|_1 = \frac{\|Ax\|_1 + \mathbf{1}^T Ax}{2}$$

由于 $\|Ax\|_1 = \|(Ax)^+\|_1 + \|(Ax)^-\|_1$和 $\mathbf{1}^T Ax = \|(Ax)^+\|_1 - \|(Ax)^-\|_1$。根据定理8，存在一个草图 S_{ℓ_1}，其具有 $O\left((k/\varepsilon^2)\log(d/\varepsilon\delta)\right)$ 行，以及一个估计量 $g_{\ell_1}(S_{\ell_1}A,x) = \|S_{\ell_1}Ax\|_{\text{med}}$，使得，以至少 $1-\delta$ 的概率，

$$\forall x\in\Psi_k : (1-\varepsilon)\|Ax\|_1 \le g_{\ell_1}(S_{\ell_1}A,x) \le (1+\varepsilon)\|Ax\|_1. \tag{8}$$

此外， $\mathbf{1}^T A$ 可以仅用一行精确计算。令 S 为将 S_{ℓ_1} 与单行 $\mathbf{1}^T$组合后得到的草图，并令

$$g_{\text{ReLU}}(SA,x) \overset{\text{def}}{=} \frac{g_{\ell_1}(S_{\ell_1}A,x) + \mathbf{1}^T Ax}{2} = \frac{\|S_{\ell_1}Ax\|_{\text{med}} + \mathbf{1}^T Ax}{2}.$$

现在，（8）意味着

$$\begin{aligned}\left|g_{\mathrm{ReLU}}(SA,x)-\|Ax\|_{\mathrm{ReLU}}\right| &= \left|\frac{g_{\ell_1}(S_{\ell_1}A,x)+\mathbf{1}^TAx}{2}-\frac{\|Ax\|_1+\mathbf{1}^TAx}{2}\right| \\ &= \left|\frac{g_{\ell_1}(S_{\ell_1}A,x)-\|Ax\|_1}{2}\right| \\ &\le \frac{\varepsilon}{2}\|Ax\|_1. \end{aligned} \tag{9}$$

根据 μ的定义，

$$\mu+1 \ge \frac{\|(Ax)^-\|_1}{\|(Ax)^+\|_1}+1 = \frac{\|Ax\|_1}{\|Ax\|_{\mathrm{ReLU}}}. \tag{10}$$

由（9）和（10），可得

$$\left|g_{\mathrm{ReLU}}(SA,x)-\|Ax\|_{\mathrm{ReLU}}\right| \le \frac{\varepsilon(\mu+1)}{2}\|Ax\|_{\mathrm{ReLU}} \le \varepsilon\mu\|Ax\|_{\mathrm{ReLU}},$$

其中，最后一个不等式成立是因为 $\mu \ge 1$。将 ε 缩放$1/\mu$倍即可得到该定理。

□

B.3 类铰链损失函数的上界

首先，我们给出一个关于类铰链函数的引理，它对于ReLU函数具有与（10）类似的作用，并已在 [MMR21]的推论9中得到证明。

引理5。设 f是一个（L、 a_1、 a_2）类铰链损失函数，并设 $C = 16\max(1、\ L、\ a_1、\ 1/a_2)$ [4]。设 $A \in \mathbb{R}^{n\times d}$。则对于任意 $x \in \mathbb{R}^d$，

$$\|Ax\|_f \ge \frac{n+\|Ax\|_1}{C\mu}.$$

证明。我们有

$$\begin{aligned}\|Ax\|_f &\ge \sum_{i:[Ax]_i\in[0,2a_1]} f(Ax)_i + \sum_{i:[Ax]_i\ge 2a_1} f(Ax)_i \\ &\ge \sum_{i:[Ax]_i\in[0,2a_1]} a_2 + \sum_{i:[Ax]_i\ge 2a_1} \mathrm{ReLU}(Ax)_i - a_1 \\ &\ge \min\left(\frac{a_2}{2a_1},\frac{1}{2}\right)\cdot\|(Ax)^+\|_1 \\ &\ge \min\left(\frac{a_2}{2a_1},\frac{1}{2}\right)\cdot\frac{\|Ax\|_1}{\mu+1}, \end{aligned} \tag{11}$$

其中第二个不等式成立是因为 f 是（L， a_1， a_2）铰链型，最后一个不等式由定理10中的9式得出。

令 $\gamma \overset{\text{定义}}{=} \min\left(\frac{a_2}{2a_1},\frac{1}{2}\right)$。现在我们声称，

$$\|Ax\|_f = \sum_{i=1}^n f(Ax)_i \ge \frac{na_2\gamma}{4\mu\cdot\max(1,L)}.$$

如果 $\sum_{i=1}^{n} f(Ax)_i \geq \frac{na_2}{4}$，则结论立即成立，因为 $\mu(X) \geq 1$，$\max(1, L) \geq 1$ 且 $\gamma \leq 1$。否则，假设 $\sum_{i=1}^{n} f(Ax)_i \leq \frac{na_2}{4}$。由于对所有 $z \geq 0$，$f(z) \geq a_2 z$，且 f 是 L-Lipschitz 连续的，因此对所有 $z \geq -\frac{a_2}{2L}$，$f(z) \geq \frac{a_2}{2}$。这意味着 Ax 至多有 $\frac{na_2/4}{a_2/2} = \frac{n}{2}$ 个非零项 $\geq -\frac{a_2}{2L}$。因此，$-Ax$ 至少有 $\frac{n}{2}$ 个非零项 $\leq -\frac{a_2}{2L}$，所以 $\|(Ax)^-\|_1 \geq \frac{na_2}{4L}$。由 μ 的定义以及式 (11)，可得

$$\|Ax\|_f \geq \gamma \cdot \|(Ax)^+\|_1 \geq \frac{na_2\gamma}{4\mu L} \geq \frac{na_2\gamma}{4\mu \cdot \max(1, L)}. \tag{12}$$

将（11）与（12）结合可得

$$\begin{aligned}
\|Ax\|_f &\geq \frac{\gamma \cdot \|Ax\|_1}{2\mu + 2} + \frac{na_2\gamma}{8\mu \cdot \max(1, L)} \\
&\geq (\|Ax\|_1 + n) \cdot \frac{\gamma \cdot \min(1, a_2)}{8\mu \cdot \max(1, L)} \\
&\geq (\|Ax\|_1 + n) \frac{1}{8\mu \cdot \max(1, L) \cdot \max(1, 1/a_2) \cdot \max(2, 2a_1/a_2)} \\
&\geq (\|Ax\|_1 + n) \frac{1}{16\mu \cdot \max(1, L, a_1, 1/a_2)^4}.
\end{aligned}$$

代入 $C = 16\max(1, L, a_1, 1/a_2)^4$ 即完成证明。 □

由引理 5可知，只需将 $\|Ax\|_f$ 近似到 $O((\varepsilon/\mu)(n+\|Ax\|_1))$ 以内，即可获得相对误差保证。定理 9 提供了一种将 $\|Ax\|_{\mathrm{ReLU}}$ 近似到 $O((\varepsilon/\mu)\|Ax\|_1)$ 以内的方法。在定理 10中，我们将证明均匀采样可以将 $\|Ax\|_f$与 $\|Ax\|_{\mathrm{ReLU}}$ 之差近似到 $O((\varepsilon/\mu)n)$ 以内。

定理 10。设 $A \in \mathbb{R}^{n\times d}$、$b \in \mathbb{R}^n$，并设 f 为一个(L, a_1, a_2)类铰链损失函数。存在一个无关草图 S，其具有 $m = O(\frac{c^{10}\mu^2 k}{\varepsilon^2}\log(\frac{cn\mu d}{\varepsilon\delta}))$行，其中 $c = \max(1, L, a_1, 1/a_2)$，以及一个估计量$g_f(SA, Sb, x)$，使得以至少 $1-\delta$的概率，我们有 $\forall x \in \Psi_k$：$(1-\varepsilon)\|Ax-b\|_f \leq g_f(SA, Sb, x) \leq (1+\varepsilon)\|Ax-b\|_f$.

证明。同样地，我们可以假设 $b = 0$。我们有

$$\|Ax\|_f = \sum_{i=1}^{n} f(Ax)_i = \sum_{i=1}^{n} (f(Ax)_i - \mathrm{ReLU}(Ax)_i) + \|Ax\|_{\mathrm{ReLU}}.$$

根据（9）在定理 9的证明中，存在一个草图 S_{ReLU}，其具有 $m_1 = O\left(k\log(d/\varepsilon\delta)/\varepsilon^2\right)$ 行，以及一个函数 g_{ReLU} 使得其概率至少为 $1-\delta$,

$$\left|g_{\mathrm{ReLU}}(S_{\mathrm{ReLU}}A, x) - \|Ax\|_{\mathrm{ReLU}}\right| \leq \varepsilon\|Ax\|_1. \tag{13}$$

我们给出一个包含 m_2 行的草图，用于近似 $R(Ax) = \sum_{i=1}^{n} (f(Ax)_i - \mathrm{ReLU}(Ax)_i)$。考虑对 A 的行进行均匀有放回采样。令 $S_u \in \mathbb{R}^{m_2\times n}$ 为对应于均匀采样 m_2 行的草图矩阵。我们将证明:

$$R(S_uAx) = \frac{n}{m_2}\sum_{i=1}^{m_2} (f(S_uAx)_i - \mathrm{ReLU}(S_uAx)_i)$$

可以吗近似 $R(Ax)$，使其误差在 $O(\varepsilon(n + \|Ax\|_1))$ 以内，对于某个合适的 m_2 值，即 .,

$$|R(S_uAx) - R(Ax)| \leq O(\varepsilon(n + \|Ax\|_1)). \tag{14}$$

固定任意 k-稀疏支撑，索引为 $I \subseteq [d]$，且满足 $|I| = k$。我们将证明 (14) 对所有 x 在此固定支撑中成立，然后对所有 $\binom{d}{k}$个支撑取并集界。令 A_I 为矩阵，其 k 列正是 A_{*i} 在 A 中满足 $i \in I$ 的那些列。我们不妨假设 A_I 的列是标准正交的，因为向量集合 $\{A_I x \mid x$ 中支撑为 $I\}$ 者在将其标准正交化后保持不变。在此假设下，$\|Ax\|_2 = \|A_I x\|_2 = \|x\|_2$ 对所有 x 且支撑为 I 成立。我们考虑两种情形。

大范数 在这种情况下，我们考虑满足 $\|Ax\|_1 \geq na_1/\varepsilon$ 的 x。由于 f 是一个（L，a_1，a_2）类铰链损失函数，因此对所有 $1 \leq i \leq n$，$|f(Ax)_i - \mathrm{ReLU}(Ax)_i| \leq a_1$。因此，对所有 $1 \leq i \leq m_2$，同样有$|f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i| \leq a_1$。我们有

$$\begin{aligned}\left|R(S_u Ax) - R(Ax)\right| &= \left|\frac{n}{m_2}\sum_{i=1}^{m_2}(f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i) - \sum_{i=1}^{n}(f(Ax)_i - \mathrm{ReLU}(Ax)_i)\right| \\ &\leq \frac{n}{m_2} m_2 a_1 + n a_1 = 2n a_1 \leq 2\varepsilon\|Ax\|_1. \end{aligned} \tag{15}$$

小范数 现在我们考虑满足 $\|Ax\|_1 < na_1/\varepsilon$ 的 x。这意味着 $\|x\|_2 = \|Ax\|_2 \leq \|Ax\|_1 < na_1/\varepsilon$。我们在 ℓ_2 范数下构造一个 $\frac{\varepsilon}{L+1}$-网 $\mathcal{N}_I^k$，使其覆盖 ℓ_2 球 $\mathcal{B}_I$ =中所有 x，且 $\{x$ 的支撑集为 $I \mid \|x\|_2 \leq na_1/\varepsilon\}$。根据标准的体积论证，$\left|\mathcal{N}_I^k\right| < (3(L+1)na_1/\varepsilon^2)^k$。接下来，我们考虑一个支撑集为 I 的固定向量 x，然后对网 $\mathcal{N}_I^k$ 中的所有向量应用并集界。

设 $X_1, X_2, \ldots, X_{m_2}$为随机变量，使得 $X_i = f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i$。由于 f是一个(L, a_1, a_2)类铰链损失函数，$|X_i| \leq a_1$ 对所有 i成立。由Hoeffding不等式，

$$\mathbb{P}\left(\left|\frac{n\sum_{i=1}^{m_2} X_i}{m_2} - R(Ax)\right| \geq \varepsilon n\right) = \mathbb{P}\left(\left|\frac{\sum_{i=1}^{m_2} X_i}{m_2} - \frac{R(Ax)}{n}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{m_2\varepsilon^2}{4a_1^2}\right). \tag{16}$$

注意（16）对于固定向量 x成立。令

$$m_2 = O\left(\frac{a_1^2}{\varepsilon^2}\left(k\log\frac{d(L+1)na_1}{\varepsilon k} + \log\frac{1}{\delta}\right)\right)$$

并对至多 $(3(L+1)na_1/\varepsilon^2)^k$ 个 $\mathcal{N}_I^k$ 中的点取并集界，我们得到 (16) 对 $\mathcal{N}_I^k$ 中所有点都成立的概率为 $1-\delta'$，其中

$$\delta' = \left(\frac{3(L+1)na_1}{\varepsilon^2}\right)^k \exp\left(-\frac{m_2\varepsilon^2}{4a_1^2}\right) \leq \delta\left(\frac{k}{ed}\right)^k.$$

接下来我们将证明，$\left|R(S_u Ax') - R(Ax')\right| \leq \varepsilon n$ 对所有 x' 属于 $\mathcal{N}_I^k$都意味着

$$\left|R(S_u Ax) - R(Ax)\right| \leq 3\varepsilon n \tag{17}$$

for 所有 $x \in B_I$。令 $x' \in \mathcal{N}_I^k$使得 $\|x - x'\|_2 \leq \frac{\varepsilon}{L+1}$。根据三角不等式 ，

$$\left|R(S_u Ax) - R(Ax)\right| \leq \left|R(Ax) - R(Ax')\right| + \left|R(S_u Ax) - R(S_u Ax')\right| + \left|R(S_u Ax') - R(Ax')\right|.$$

由于x' 属于 $\mathcal{N}$，最后一项至多为 εn。我们将利用 f是 L-利普希茨且 $\|x - x'\|_2 \leq \frac{\varepsilon}{L+1}$这一事实来界定另外两项。再次应用三角不等式，我们有

$$\left|R(Ax) - R(Ax')\right| \leq \left|\sum_{i=1}^{n}\left(\mathrm{ReLU}(Ax)_i - \mathrm{ReLU}(Ax')_i\right)\right| + \left|\sum_{i=1}^{n}\left(f(Ax)_i - f(Ax')_i\right)\right|$$

$$
\begin{aligned}
&\le \|A(x-x')\|_1 + L\|A(x-x')\|_1 \\
&\le (1+L)\sqrt{n}\|x-x'\|_2 \\
&\le \varepsilon\sqrt{n}.
\end{aligned}
$$

同样,

$$
\begin{aligned}
&\left|R(S_uAx) - R(S_uAx')\right| \\
&\le \left|\frac{n}{m_2}\sum_{i=1}^{m_2}\Big(\mathrm{ReLU}(S_uAx)_i - \mathrm{ReLU}(S_uAx')_i\Big)\right| + \left|\frac{n}{m_2}\sum_{i=1}^{m_2}\Big(f(S_uAx)_i - f(S_uAx')_i\Big)\right| \\
&\le \frac{n}{m_2}\|S_uA(x-x')\|_1 + \frac{n}{m_2}L\|S_uA(x-x')\|_1 \\
&\le (L+1)\frac{n}{\sqrt{m_2}}\|S_uA(x-x')\|_2 \\
&\le (L+1)\frac{n}{\sqrt{m_2}}\|A(x-x')\|_2 \\
&\le (L+1)\frac{n}{\sqrt{m_2}}\|x-x'\|_2 \\
&\le \varepsilon\frac{n}{\sqrt{m_2}}.
\end{aligned}
$$

这就完成了对式（17）的证明，也完成了小范数情形的论证。结合式（15）和（17），对于固定的支撑集 I，以至少 $1-\delta\left(\frac{k}{ed}\right)^k$ 的概率，

$$
\left|R(S_uAx) - R(Ax)\right| \le 3\varepsilon n + 2\varepsilon\|Ax\|_1.
$$

对 $\binom{d}{k} < (ed/k)^k$ k-稀疏支撑集应用并集界，可得成功概率至少为 $1-\delta$。令 $S = \begin{bmatrix} S_{\mathrm{ReLU}} \\ S_u \end{bmatrix}$。定义

$$
g_f(SA, x) \overset{\mathrm{def}}{=} R(S_uAx) + g_{\mathrm{ReLU}}(S_{\mathrm{ReLU}}A, x).
$$

由式（13），以至少 $1-2\delta$ 的概率，

$$
\begin{aligned}
\left|g_f(SA,x) - \|Ax\|_f\right| &\le \left|R(S_uAx) - R(Ax)\right| + \left|g_{\mathrm{ReLU}}(S_{\mathrm{ReLU}}A, x) - \|Ax\|_{\mathrm{ReLU}}\right| \\
&\le 3\varepsilon\left(n + \|Ax\|_1\right) \qquad (18)
\end{aligned}
$$

根据引理 5，$\|Ax\|_f \ge (n+\|Ax\|_1)/(C\mu)$，其中 $C = 16\max(1, L, a_1, 1/a_2)^4$。将其与式（18）结合可得

$$
\left|g_f(SA,x) - \|Ax\|_f\right| \le 3C\mu\varepsilon\|Ax\|_f.
$$

最后，将 ε 按 $O(1/(C\mu))$ 缩放后，S 中的行数为

$$
\begin{aligned}
m_1 + m_2 &= O\left(\frac{C^2\mu^2 k(1+a_1^2)}{\varepsilon^2}\cdot\log\left(\frac{Cna_1(L+1)\mu d}{\varepsilon\delta}\right)\right) \\
&= O\left(\frac{c^{10}\mu^2 k}{\varepsilon^2}\cdot\log\left(\frac{cn\mu d}{\varepsilon\delta}\right)\right)
\end{aligned}
$$

其中 $c = \max(1,\ L,\ a_1,\ 1/a_2)$。 □

为完整起见，我们给出以下简单结果，它揭示了在草图空间中使用 k-稀疏仿射嵌入进行最小化（如上述所有上界所示）与原始问题之间的联系。

推论 *11.* 设 $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$。设 $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ 是任意损失函数。设 S 是 $[A, b]$ 的一个无偏线性草图，并设 $E(SA, Sb, x)$ 是满足 $\forall x \in \Psi_k: (1-\varepsilon)\|Ax-b\| \leq E(SA, Sb, x) \leq (1+\varepsilon)\|Ax-b\|$. 的估计量。那么 $\tilde{x} \in \mathit{argmin}_{x\in\Psi_k} E(SA, Sb, x)$ 满足$\|A\tilde{x}-b\| \leq (1+O(\varepsilon))\ \mathit{min}_{x\in\Psi_k}\|Ax-b\|$.

证明。 (推论 11的证明）令 $x^* \in \operatorname{argmin}_{x\in\Psi_k}\|Ax-b\|$。则

$$\begin{aligned}\|A\tilde{x}-b\| &\leq E(SA, Sb, \tilde{x})/(1-\varepsilon) \leq E(SA, Sb, x^*)/(1-\varepsilon)\\ &\leq \|Ax^*-b\|(1-\varepsilon)/(1-\varepsilon) \leq (1+4\varepsilon)\|Ax^*-b\|.\end{aligned}$$

□

C LASSO回归的无偏草图

LASSO回归是 k-稀疏 ℓ_2 回归的一种凸松弛形式，作为诱导稀疏性和进行特征选择的启发式方法广受欢迎 [Tib96]。这里我们给出一个草图的上界，该上界取决于 ℓ_1 正则化参数 λ以及 log(d)。

C.1 上界

定理12。考虑 $A \in \mathbb{R}^{n\times d}$、 $b \in \mathbb{R}^n$ 和 $\lambda \in (0、1)$。假设 $\|A\|_2 \leq 1$ 和 $\|b\|_2 \leq 1$，若 $S \in \mathbb{R}^{m\times n}$ 是一个随机高斯矩阵（即每个元素均独立同分布地从 $N(0、 1/m)$中采样），则对任意 $\varepsilon\ \delta \in 0,1 \quad m = O \quad 1-\delta \quad \tilde{x} = \quad x\in\mathbb{R}^d \|SAx-, ()$ 和 $(\frac{\log d/\delta}{\lambda^2\cdot\varepsilon^2})$，以至少 $Sb\|_2^2+\lambda\|x\|_1 \quad \|A\tilde{x}-b\|_2^2+\lambda\|\tilde{x}\|_1 \leq \frac{}{1+\varepsilon}\cdot \quad x\in\mathbb{R}^d \|Ax-b\|_2^2+\lambda\|x\|_1.$ 的概率，若 argmin则 () min

请注意，我们对 $\|A\|_2$、 $\|b\|_2$ 的约束是必要的，因为经典 LASSO 回归不具有尺度不变性。若将 A 放大任意大的因子 α，放大 A α 倍后，我们可以保持相同的误差 $\|Ax-b\|_2^2$，同时将 x 缩小 $1/\alpha$倍。因此， $\lambda\cdot\|x\|_1$ 项变得可以忽略，问题也就简化为普通最小二乘回归。类似地，若将 b 放大 α倍，则可达到的最小 $\|Ax-b\|_2^2$ 会缩放 α^2倍，而达到该最小值的 x 仅缩放 α倍，其 $\lambda\|x\|_1$ 如此变化。因此，当 α 任意增大时，问题同样会变成普通最小二乘回归。已知，对于普通最小二乘回归，草图维度 $\Theta(d/\varepsilon)$ 是必要且充分的 [Sar06; CW09]。因此，若要超越这一结果，就需要对 $\|A\|_2$、 $\|b\|_2$ 加以限制，以确保正则化 $\lambda\|x\|_1$ 具有不可忽略的影响。

还要注意，我们无法期望实现 $o(d/\varepsilon^2)$ 的界限，以对所有 $x \in \mathbb{R}^d$ 保持 LASSO 成本，因为对于 x 取足够大的 $\|x\|_2^2$，以及对于 $b = 0$，该问题等价于保持 $\|Ax\|_2^2$，这要求 $\Theta(d/\varepsilon^2)$ 的草图大小 [NN14]。

证明定理 *12.* 设 $OPT = \min_{x\in\mathbb{R}^d}\|Ax-b\|_2^2+\lambda\|x\|_1$。注意到由于 $\|b\|_2 \leq 1$，$OPT \leq \|b\|_2^2+\lambda\|0\|_1 \leq 1$。由标准的Johnson-Lindenstrauss引理，我们以高概率得到，对于 $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\|Ax-b\|_2^2+\lambda\|x\|_1$，
$\|SAx^*-Sb\|_2^2+\lambda\|x^*\|_1 \leq (1+\varepsilon)\cdot OPT$。

因此，我们必须有 $\|\tilde{x}\|_1 \leq \frac{(1+\varepsilon)\cdot OPT}{\lambda} \leq \frac{2\cdot OPT}{\lambda}$，否则就会有 $\|SA\tilde{x}-Sb\|_2^2+\lambda\|\tilde{x}\|_1 \geq \|SAx^*-Sb\|_2^2+\lambda\|x^*\|_1$，这与 $\tilde{x}$ 是草图问题的最小值点这一事实矛盾。出于同样的原因，我们必须有 $\|SA\tilde{x}-Sb\|_2 \leq \|Sb\|_2 \leq 2$.

设 $\mathcal{T} = \{y = Ax-b: \|x\|_1 \leq \frac{2\cdot OPT}{\lambda}\}$。设 $\mathcal{T}' = \{y = Ax-b: \|x\|_1 = 1\}$。注意到根据我们关于 $\|A\|_2 \leq 1$的假设， A 的每一列 a_i 都具有 $\|a_i\|_2 \leq 1$。因此， $\mathcal{T}'$ 是凸包

的 $2d+1$ 单位球中的点： a_1， $-a_1$， a_2， $-a_2$， $\ldots a_d$， $-a_d$， b。根据 [NN19], 对于 $m = O\left(\frac{\log(d/\delta)}{\lambda^2 \cdot \varepsilon^2}\right)$，以至少 $1-\delta$的概率，对所有 $y' \in \mathcal{T}'$，

$$|\|Sy\|_2 - \|y\|_2| \leq \varepsilon\lambda.$$

注意，任何 $y \in \mathcal{T}$ 都可以写成 $\alpha \cdot y'$，对于 $y' \in \mathcal{T}'$，对于某个 $\alpha \leq \frac{2 \cdot OPT}{\lambda}$。因此，我们有：以至少 $1-\delta$ 的概率，对于所有 $y \in \mathcal{T}$，

$$|\|Sy\|_2 - \|y\|_2| \leq 2\varepsilon \cdot OPT.$$

特别地　特别地，对于 $\tilde{x} = \text{argmin}_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$，我们有 $A\tilde{x} - b \in \mathcal{T}$因此这给出 S

$$\begin{aligned} \left|\|SAx - Sb\|_2^2 - \|Ax - b\|_2^2\right| &\leq |\|SAx - Sb\|_2 - \|Ax - b\|_2| \cdot |\|SAx - Sb\|_2 + \|Ax - b\|_2| \\ &\leq 12\varepsilon \cdot OPT, \end{aligned} \tag{19}$$

其中第二个不等式利用了 $\|SA\tilde{x} - Sb\|_2 \leq 2 \cdot OPT \leq 2$ 因此 $\|A\tilde{x} - b\|_2 \leq (2 + 2\varepsilon) \cdot OPT \leq 4$. 最后，利用（19）我们有：

$$\begin{aligned} \|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 &\leq \|SA\tilde{x} - Sb\|_2^2 + \lambda\|\tilde{x}\|_1 + 5\varepsilon \cdot OPT \\ &\leq \|SA\tilde{x}^* - Sb\|_2^2 + \lambda\|x^*\|_1 + 12\varepsilon \cdot OPT \\ &\leq (1 + 13\varepsilon) \cdot [\|SA\tilde{x}^* - Sb\|_2^2 + \lambda\|x^*\|_1]. \end{aligned}$$

将 ε 调整一个常数因子后，定理得证。 □

C.2 下界

定理13。令 $A \in \mathbb{R}^{n \times d}$、$b \in \mathbb{R}^n$ 具有有界的 $\|A\|_2 \leq 1$ 和 $\|b\|_2 \leq 1$，并令 $\lambda \in (0,\ 1)$。假-设 $S \in \mathbb{R}^{m \times n}$ 是用于 *LASSO* 回归的无关线性草图，其估计器为 $E_\lambda(SA、\ Sb、\ x)$，使得以常数概率 $\tilde{x} \in \text{argmin}_{x \in \Psi_k} E_\lambda(SA、\ Sb、\ x)$ 满足 $\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1+\varepsilon) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$. 则 $m = \Omega(\frac{\log(\lambda d)}{\lambda^2})$。

定理13的证明。我们将以类似于定理1中下界的方式证明这一点。我们将取$[b\ A] = X \sim \frac{1}{\sqrt{n}} G(I + zz^T)^{1/2}$，其中 z如推论15中那样，且带有 $\varepsilon = 1/2$. 不失一般性，我们可以假设 S 具有标准正交行，从而 $SG(I + zz^T)^{1/2}$ 与行数更少的$G(I + zz^T)^{1/2}$ 具有相同的分布。我们还令 $\lambda = 1/(2\sqrt{k})$。归一化因子 $\frac{1}{\sqrt{n}}$ 确保范数条件 $\|A\|_2$、$\|b\|_2 \leq 1$ 以高概率成立。于是我们以高概率有

$$\|Ax - b\|_2^2 + \lambda\|x\|_1 \approx 1 + \|x\|_2^2 + (1 - \varepsilon v^T x)^2 + \lambda\|x\|_1 =: L(x).$$

上述近似误差（以及概率）可以通过令 n 足够大而任意变小。

我们注意到 $L(x)$ 是一个 1-强凸函数，因此

$$L(\hat{x}) \geq L(x^*) + \|\hat{x} - x^*\|_2^2.$$

对于任意 $\hat{x}$，其中 x^* 是 $L(x)$ 的最小化器. 通过直接计算，我们还可得到 $x^* = v/5$ 和 $L(x^*) = 1.95$. 假设对于充分小的 $c_1.$，有 $L(\hat{x}) \leq (1 + c_1)L(x^*)$。那么我们有

$$(1 + c_1)L(x^*) \geq L(x^*) + \|\hat{x} - x^*\|_2^2.$$

这意味着 $\|\hat{x} - x^*\|_2^2 \leq c_1 L(x^*) \leq 1.95 \cdot c_1$. 因此，选择 c_1 为一个足够小的常数后，我们可以恢复 $\text{supp}(19/20)$ 的一部分x^*，即 $=\text{supp}(v)$。推论 15 进而给出所需的下界 $\Omega(k\log(d/k))$，即草图的大小至少为该值，其中 k 设为 $1/(4\lambda^2)$，这与我们对 λ的选择相对应。

□

D 区分稀疏恢复和稀疏回归

现在，我们来说明 k-稀疏恢复问题与 k-稀疏回归问题之间的差异。结合我们针对 k-稀疏回归得到的下界，这表明 k-稀疏恢复问题严格来说更容易。下面给出的上界与 k-稀疏输出的 [PW11] 的下界相差一个 $\log(k/\varepsilon)$ 因子，改进了朴素的上界 $O(k(\log d)/\varepsilon^2)$。

在 k-稀疏恢复问题中，人们旨在对向量 $x \in \mathbb{R}^d$ 进行草图处理，从而输出一个 k-稀疏的 $\hat{x} \in \mathbb{R}^d$，使得

$$\|x - \hat{x}\|_2^2 \leq (1+\varepsilon)\|x - x_k\|_2^2, \tag{20}$$

其中，x_k 由 x 中幅值最大的前 k 个条目组成，平局时任意打破平局。我们需要以下关于 CountSketch 的定理。

定理 16。 *([CCF04])* 存在一种关于草图矩阵的分布 $S \in \mathbb{R}^{bt \times d}$，称为 ***Count-Sketch*** 矩阵，其参数由桶的数量 b 和表的数量 t 决定。对于向量 x，存在一个过程，给定 $S \cdot x$ 和一个坐标 $i \in [d]$，返回一个估计值 $\hat{x}_i$，满足

$$|\hat{x}_i - x_i| \leq C \cdot \|x - x_k\|_2 / \sqrt{b},$$

失败概率至多为 $2^{-C't}$，其中 C、$C' > 0$ 为绝对常数。

考虑以下过程：

1. 运行 CountSketch，使用 $b = O(k/\varepsilon)$ 个桶和 $t = O(\log d)$ 个表，并令 A 为满足对应估计值 $\hat{}\, i \in \{1, 2, \ldots, d\}$ 位于绝对值最大的 x_i' 之列的索引集合 $O(k/\varepsilon)$。这里我们用 $\hat{x}_i'$ 表示定理16返回的估计值，以便将其与下一步返回的估计值区分开来。

2. 同时，运行具有 $b = O(k/\varepsilon^2)$ 个桶和 $t = O(\log(k/\varepsilon))$ 个表的CountSketch，并为每个 $i \in A$ 计算估计值 $\hat{x}_i$。令 B 为限制在 $i \in A$ 内的幅度估计值 $\hat{x}_i$ 中最大的 k 个估计值所构成的集合。

3. 返回由 B 支持的向量 $\hat{x}$，并为每个 $i \in B$ 提供相应的估计值 $\hat{x}_i$。

定理 14. 对于输入 $x \in \mathbb{R}^d$，上述稀疏恢复方案使用 $O(k\log(k/\varepsilon)/\varepsilon^2 + k\log(d)/\varepsilon)$ 次测量，并以至少 $1 - 1/\operatorname{poly}(d) - 1/\operatorname{poly}(k/\varepsilon)$ 的概率，返回一个 k-稀疏向量 $\hat{x} \in \Psi_k$，满足 $\|x - \hat{x}\|_2^2 \leq (1+\varepsilon)\min_{x_k \in \Psi_k} \|x - x_k\|_2^2$.

证明。 由于草图 S 是线性的，我们可以不失一般性地假设 $\|x - x_k\|_2^2 = k$。根据定理16，以概率 $1 - 1/\operatorname{poly}(d$，步骤 1 中的 CountSketch 返回满足条件的估计值 $\hat{x}_i'$，使得

$$|\hat{x}_i' - x_i| \leq \sqrt{\varepsilon}$$

对所有 $i = 1$、2、…、d 同时成立。因此，对于每个满足 $|x_i| \geq 3\sqrt{\varepsilon}$ 的 i，都有 $i \in A$。为说明这一点，请注意，若 $|x_i| \geq 3\sqrt{\varepsilon}$，则 $|\hat{x}_i'| \geq 2\sqrt{\varepsilon}$。另一方面，若对某个 $j \in [d]$ 有 $|\hat{x}_j'| \geq 2\sqrt{\varepsilon}$，则 $|x_j| \geq \sqrt{\varepsilon}$。满足 $|x_j| \geq \sqrt{\varepsilon}$ 的 j 的数量至多为 $k + k/\varepsilon$，因为 $\|x - x_k\|_2^2 = k$。此外，满足 $|\hat{x}_i| \geq 2\sqrt{\varepsilon}$ 的 i 的数量至多为 $k + k/(4\varepsilon)$。因此，若对于某个足够大的常数，$|A| = O(k/\varepsilon)$，则当 $|x_i| \geq 3\sqrt{\varepsilon}$ 时，有 $i \in A$。

在步骤 2 中，通过对所有 $i \in A$ 应用并集界，我们得到：以 $1 - 1/\operatorname{poly}(k/\varepsilon)$ 的概率，对于每个 $i \in A$，$|\hat{x}_i - x_i| \leq \varepsilon$ 都同时成立。令 H 为幅值最大的前 k

x 的坐标集合，我们有 $\|x_{H\setminus A} - \hat{x}_{H\setminus A}\|_2^2 \le k \cdot 9 \cdot \varepsilon$，因为任意 $i \in H \setminus A$ 必然满足 $|x_i| < 3\sqrt{\varepsilon}$ 和 $\hat{x}_i = 0$。

我们还有 $\|x_{H\cap B} - \hat{x}_{H\cap B}\|_2^2 \le k \cdot \varepsilon^2$，因为对于每个 $i \in H \cap B$，都有 $|\hat{x}_i - x_i| \le \varepsilon$。

最后，考虑这些 $i \in H\cap(A\setminus B)$。对于每个这样的 i，必然存在一个 $j = j(i) \in A\setminus H$，使得 $|\hat{x}_{j(i)}| \ge |\hat{x}_i|$；并且由于 $|\hat{x}_{j(i)} - x_{j(i)}| \le \varepsilon$ 和 $|\hat{x}_i - x_i| \le \varepsilon$，这意味着 $|x_{j(i)}| \ge |x_i| - 2\varepsilon$。另请注意，根据 H 的定义，$|x_{j(i)}| \le |x_i|$。因此，草图解 $\hat{x}$ 的代价至多是

$$(x_{j(i)} - \hat{x}_{j(i)})^2 + x_i^2 \le \varepsilon^2 + (x_{j(i)} + 2\varepsilon)^2 = O(\varepsilon^2 + x_{j(i)}^2 + |x_{j(i)}|\varepsilon),$$

在该坐标上，而最优的非草图化解 x_k 需付出 $x_{j(i)}^2$。由于 $|A \setminus B| \le k$，草图化解相对于最优解所支付的额外总成本至多为

$$O(k\varepsilon^2 + \varepsilon \sum_{i\in H\cap(A\setminus B)} |x_{j(i)}|). \tag{21}$$

最后，请注意$\sum_{i\in H\cap(A\setminus B)} |x_{j(i)}|$ 至多为 ℓ_1个按绝对值计最大的、不在 k 坐标的 H范数。由于这些 ℓ_2 坐标的范数至多为$\sqrt{k}$，因此这些 ℓ_1坐标的范数至多为 k。结合（21），草图化解所付出的总额外误差为 $O(\varepsilon k)$。

由此可得

$$\|x - \hat{x}\|_2^2 \le \|x_{H\setminus A} - \hat{x}_{H\setminus A}\|_2^2 + \|x_{H\cap B} - \hat{x}_{H\cap B}\|_2^2 + \|x - x_k\|_2^2 + O(\varepsilon k) \le (1 + O(\varepsilon))\|x - x_k\|_2^2$$

并且（20）通过将其按常数因子重新缩放即可得到 ε。测量总数和总体失败概率由定理16得出。

□

E 信息论基础

我们需要信息论中的以下概念。

定义 3（熵和互信息）。随机变量 X 在某个支撑集 S 上的熵是

$$H(X) = \sum_{i\in S} p_i \log_2 \frac{1}{p_i}.$$

给定两个随机变量 X 和 Y，条件熵为

$$H(X|Y) = \sum_y H(X|Y = y) \cdot \mathbb{P}[Y = y]$$

且它们的联合熵为

$$H(X,Y) = \sum_{x,y} \mathbb{P}[X = x \bigwedge Y = y] \log_2 \frac{1}{\mathbb{P}[X = x \wedge Y = y]}.$$

两个随机变量的互信息是

$$I(X;Y) = H(X) - H(X|Y)$$

最后，我们需要 Fano 不等式。

引理6（Fano 不等式）。令 X 是根据分布 μ_X 从定义域 $\mathcal{X}$ 中选取的随机变量，令 Y 是根据分布 μ_Y 从定义域 $\mathcal{Y}$ 中选取的随机变量。那么对于任意重构函数 $g: Y \to X$，其误差为 ε_g，都有

$$H(X|Y) \le H(\varepsilon_g) + \varepsilon_g \log(|\mathcal{X}| - 1).$$