arXiv:2304.11958v2 [math.ST] 6 Feb 2024

在L次指数型协变量下稀疏线性回归系数的估计

Takeyuki Sasai *

2024年2月7日

摘要

我们研究了在线性回归中估计稀疏系数的问题，其中协变量取样自 L次指数型随机向量。该向量属于一类具有比高斯随机向量更重尾部的分布。此前的研究已经建立了与高斯随机向量所得结果相似的误差界。然而，与使用高斯随机向量推导误差界的方法相比，这些方法需要更强的条件。在本研究中，我们证明了当协变量取自 L次指数型随机向量时，无需施加更强的条件，即可得到与高斯随机向量所得结果相同的误差界（相差常数因子）。有趣的是，我们采用了ℓ_1惩罚的Huber回归；众所周知，该方法针对重尾随机噪声具有稳健性，而不是针对协变量。我们认为，本研究揭示了ℓ_1惩罚的Huber回归方法的一个新方面。

1 引言

在本研究中，我们研究稀疏线性回归。我们将稀疏线性回归模型定义如下：

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \xi_i, \quad i = 1, \cdots, n, \tag{1.1}$$

其中， $\boldsymbol{\beta}^* \in \mathbb{R}^d$ 是真实的系数向量， $\{\mathbf{x}_i\}_{i=1}^n$ 是独立同分布（i.i.d.）随机向量序列， $\{\xi_i\}_{i=1}^n$ 是独立同分布随机变量序列。在整个研究中，我们假设 $\boldsymbol{\beta}^*$ 的非零元素个数为 $s(\leq d)$，且 $\{\mathbf{x}_i\}_{i=1}^n$与 $\{\xi_i\}_{i=1}^n$以及 $d/s \geq 3$相互独立。

对于向量 $\mathbf{v} \in \mathbb{R}^d$，我们将其 ℓ_2 范数定义为 $\|\mathbf{v}\|_2$。假设 $\{\mathbf{x}_i\}_{i=1}^n$ 和 $\{\xi_i\}_{i=1}^n$ 分别是从具有 $\mathbb{E}\mathbf{x}_i = 0$ 和 $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top = I$的多元高斯分布中抽取的独立同分布随机协变量向量序列，以及从具有 $\mathbb{E}\xi_i = 0$ 和 $\mathbb{E}\xi_i^2 = 1$的高斯分布中抽取的随机噪声序列。我们将 . 定义为相差一个数值因子的关系。已知，我们可以构造一个估计量 $\boldsymbol{\beta}$ ，使得

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta, \tag{1.2}$$

当 $\sqrt{\frac{s\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}$足够小时， [Bellec 等， 2018]。自套索回归的发明以来 [Tibshirani， 1996]，，人们已对根据（1.1）估计 $\boldsymbol{\beta}^*$ 开展了大量研究。在许多研究中，协变量序列 $\{\mathbf{x}_i\}_{i=1}^n$ 取样自高斯分布或 L-次高斯随机向量（见定义 2.2），噪声 $\{\xi_i\}_{i=1}^n$ 则取自高斯分布或其他重尾分布，从而得到类似于（1.2）的结果。然而，只有少数

*统计科学系，综合研究大学院大学（SOKENDAI），日本东京。电子邮件：sasai@ism.ac.jp

研究考虑了 $\{\mathbf{x}_i\}_{i=1}^n$ 取自高斯分布或 L-次高斯随机向量以外的重尾随机向量的情况 [Sivakumar 等，2015，Lecu´e 和 Mendelson，2017，Fan 等，2021，Kuchibhotla 和 Chakrabortty，2022，Genzel 和 Kipp， 2022], ，所得结果与（1.2）相比略逊一筹。在本研究中，我们假设 $\{\mathbf{x}_i\}_{i=1}^n$ 取自 L-次指数随机向量（见定义 2.2）；与 L-次高斯随机向量相比，这类向量具有更重的尾部。除数值常数和对 L的依赖性外，我们推导出的结果与（1.2）相当。

在本文余下部分中，数值常数 C 的值可以逐行变化。

2 相关工作、我们的估计方法与核心思想

2.1 一些定义

在第 2.1 节中，我们将介绍后文会用到的定义。

定义 2.1（ψ_α-范数）。对于随机变量 f，我们定义

$$\|f\|_{\psi_\alpha} := \inf\{\eta > 0 : \mathbb{E}\exp|f/\eta|^\alpha \le 2\} < \infty. \tag{2.1}$$

定义 2.2（L-次高斯和 L-次指数随机向量）。设 L 为一个数值常数。若随机向量 $\mathbf{x} \in \mathbb{R}^d$ 满足对于任意固定的 $\mathbf{v} \in \mathbb{R}^d$，则称其为 L-次高斯和 L-次指数随机向量。

$$\|\langle \mathbf{x} - \mathbb{E}\mathbf{x}, \mathbf{v}\rangle\|_{\psi_2} \le L\left(\mathbb{E}\langle \mathbf{x} - \mathbb{E}\mathbf{x}, \mathbf{v}\rangle^2\right)^{\frac{1}{2}} \ \ and \ \|\langle \mathbf{x} - \mathbb{E}\mathbf{x}, \mathbf{v}\rangle\|_{\psi_1} \le L\left(\mathbb{E}\langle \mathbf{x} - \mathbb{E}\mathbf{x}, \mathbf{v}\rangle^2\right)^{\frac{1}{2}}, \tag{2.2}$$

分别。

定义 2.3（L-次高斯和 σ-次指数随机变量）。设 σ 为数值常数。若随机变量 ξ 满足

$$\|\xi - \mathbb{E}\xi\|_{\psi_2} \le \sigma \ \ and \ \|\xi - \mathbb{E}\xi\|_{\psi_1} \le \sigma, \tag{2.3}$$

分别。

我们注意到， $\|f - \mathbb{E}f\|_{\psi_1} \le \|f - \mathbb{E}f\|_{\psi_2}$对 σ-次高斯随机变量 f成立（如 [Vershynin, 2018]中的引理2.7.7所述），并且可以看到， L-次高斯随机向量也是 L-次指数随机向量。

2.2 相关工作

对于向量 $\mathbf{v} \in \mathbb{R}^d$，其 ℓ_1 范数定义为 $\|\mathbf{v}\|_1$。在 $\{\mathbf{x}_i, y_i\}_{i=1}^n$ 的数据基础上，对式（1.1）中的 $\boldsymbol{\beta}^*$ 进行估计的问题已在 [Genzel 和 Kipp, 2022] 以及 [Sivakumar et al., 2015],中得到研究，其中 $\{\mathbf{x}_i\}_{i=1}^n$ 和 $\{\xi_i\}_{i=1}^n$ 分别取自各向同性的 L- 次指数随机向量和 σ- 次指数随机变量。在 [Sivakumar et al., 2015], 中，所用的估计方法是 ℓ_1-惩罚最小二乘法:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda_s \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.4}$$

其中， λ_s 是调节参数， [Genzel 和 Kipp 的估计方法 2022] 是 ℓ_1-约束最小二乘法。我们将 . 定义为差一个常数因子的关系。 [Genzel 和 Kipp、 2022] 以及 [Sivakumar 等人的方法 2015] 得出了与（1.2）类似的结果。然而，由于 $\{\mathbf{x}_i\}_{i=1}^n$的重尾性，为了获得其误差界，他们要求 n 至少大于 $s(\log(d/s))^2$,，而不是 $s\log(d/s)$。在第 2.3 节中，我们解释了 $\{\mathbf{x}_i\}_{i=1}^n$ 的重尾性如何对结果产生负面影响，并提出了一种消除附加条件的方法。

2.2.1 其他相关工作

[Fan 等人 2021] 研究了这样一种情形：协变量取自重尾随机向量 $\mathbf{x}$，对于任意 $\mathbf{v} \in \mathbb{S}^{d-1}$，$\mathbb{E}\langle \mathbf{x}, \mathbf{v}\rangle^4 \leq K^4$，其中 K 为某个常数。[Fan 等人提出的方法 2021] 要求满足附加条件，即 $(s^2 \log d)/n$ 和 $\|\boldsymbol{\beta}^*\|_1\sqrt{(\log d)/n}$ 足够小。[Lecu´e 和 Mendelson 2017] 研究了设计矩阵的受限特征值条件，其中协变量取自重尾随机向量 $\mathbf{x}$，对于任意 $\mathbf{v} \in \mathbb{S}^{d-1}$，以及满足 $4 \leq p \lesssim \log d$，$\mathbb{E}|\langle \mathbf{x}, \mathbf{v}\rangle|^p \leq L^p$的 p。接下来，[Kuchibhotla 和 Chakrabortty 2022], 研究了协变量取自边际 L-次 Weibull 随机向量的情形。该分析是在 n 大于 $s(\log(d/s))^2$的条件下进行的。边际L-次 Weibull 随机向量 $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ 是指满足以下条件的随机向量：对于 $0 < \alpha$ 和 L，每个坐标都满足 $\|x_i\|_{\psi_\alpha} \leq L$。根据 $\|\cdot\|_{\psi_\alpha}$的定义可知，σ-次指数随机变量是一个 σ-次 Weibull 随机变量，其中 $0 < \alpha \leq 1$，且后者的尾部比前者更重。因此，[Kuchibhotla 和 Chakrabortty 2022] 处理的是比本文所研究情形更为一般的情况。验证第 2.3 节中关于这些情形的观察结果的有效性将是未来的任务。

2.3 我们的估计方法与关键思想

将 Huber 损失及其导数分别定义为

$$H(t) = \begin{cases} |t| - 1/2 & (|t| > 1) \\ t^2/2 & (|t| \leq 1) \end{cases} \text{ and } h(t) = \frac{d}{dt}H(t) = \begin{cases} \operatorname{sgn}(t) & (|t| > 1) \\ t & (|t| \leq 1) \end{cases}, \tag{2.5}$$

估计量如下:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \lambda_o^2 H\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\lambda_o \sqrt{n}} \right) + \lambda_s \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.6}$$

其中，λ_o 是 Huber 损失的调节参数。当 $\{\xi_i\}_{i=1}^n$ 来自重尾分布或受到离群值影响时，ℓ_1惩罚的 Huber 损失已知具有良好的效果 [She 和 Owen，2011，Nguyen 和 Tran，2012，Dalalyan 和 Thompson，2019，Sun 等，2020]。在本研究中，我们发现，当 $\{\mathbf{x}_i\}_{i=1}^n$ 来自 L次指数型随机向量时，ℓ_1惩罚的 Huber 损失仍然有效，并且可以推导出一个误差界，该误差界除了常数因子外，与 (1.2) 等价。我们注意到，在 [Sun 等，2020], 中也对一个实质上与 (2.6) 完全相同的估计量进行了详细研究，并得出了与我们类似的结果。[Sun 等，2020]主要关注 (2.6) 对 $\{\xi\}_{i=1}^n$重尾性的稳健性，而 [Sun 等，2020] 假设$\{\mathbf{x}_i\}_{i=1}^n$ 是 L次高斯的。因此，本研究与 [Sun 等，2020] 的主要关注点不同。

下面说明平方损失为何无法有效处理 $\{\mathbf{x}_i\}_{i=1}^n$ 的重尾特征。直观而言，Huber 损失限制了 $\{\xi_i\}_{i=1}^n$ 的移动范围，从而为 $\{\mathbf{x}_i\}_{i=1}^n$ 中更大的活动空间创造了条件。

下面解释平方损失为何无法有效应对 $\{\mathbf{x}_i\}_{i=1}^n$ and$\{\xi_i\}_{i=1}^n$ 的重尾性。对于向量 $\mathbf{v} \in \mathbb{R}^d$，我们将其 ℓ_0 范数定义为 $\|\mathbf{v}\|_0$，表示 $\mathbf{v}$ 中非零元素的数量。此外，对于 $l = 0$，1，2，我们将直径为 a 的 d 维 ℓ_l 球定义为 $a\mathbb{B}_l^d = \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v}\|_l \leq a\}$。在使用平方损失推导类似于 (1.2) 的误差界时，评估以下量至关重要:

$$\sup_{\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d} \sum_{i=1}^{n} \frac{1}{n} \xi_i \mathbf{x}_i^\top \mathbf{v}. \tag{2.7}$$

当 $\mathbf{x}_i$ 是一个 L-次高斯随机向量，且 ξ_i 是一个 σ-次高斯分布时，交叉项 $\xi_i\mathbf{x}_i$ 会变成一个 $CL\sigma$-次指数随机向量（Vershynin 的引理2.7.7，[2018]）。然而，如果 $\mathbf{x}_i$ 是一个 L-次指数随机向量，即使 ξ_i 是高斯分布，交叉项 $\xi_i\mathbf{x}_i$ 通常也不再是一个 $CL\sigma$-次指数随机向量。因此，要得到

与 $\mathbf{x}_i$ 是一个 L-次高斯向量的情形类似，（2.7）中的浓度不等式也变得难以获得。

在 ℓ_1-惩罚 Huber 回归中，无需计算 (2.7)。相反，只需计算以下内容：

$$\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n\frac{1}{n}h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\mathbf{x}_i^\top\mathbf{v}. \tag{2.8}$$

当 $\mathbf{x}_i$ 是一个 L-次指数随机向量，且 ξ_i 是一个 σ-次指数随机变量时，交叉项 $h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\mathbf{x}_i$ 仍是一个 L-次指数随机向量，因为 $h(\cdot)$ 有界。基于这一观察，我们成功得到一个与 (1.2) 类似的结果，误差仅相差常数因子。

3 我们的主要结果与关键命题

在第 3 节中，我们给出主要结果（定理 3.1）以及命题 3.1 和 3.2，这些命题用于证明定理 3.1。定理 3.1 以及命题 3.1 和 3.2 的证明见附录。在陈述主要结果之前，我们先引入以下假设：

假设3.1. 假设

(i) $\{\mathbf{x}_i\}_{i=1}^n$ 是从具有 L-次指数分布的随机向量中抽取的独立同分布随机向量序列，且满足 $\mathbb{E}\mathbf{x}_i=0, \mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top=I$,

(ii) $\{\xi_i\}_{i=1}^n$ 是满足 $\mathbb{E}\xi_i^2\le\sigma^2$的独立同分布随机变量序列。

我们注意到，假设3.1中的条件(ii)弱于 σ-次指数性质。定义

$$r_{d,s}=\sqrt{\frac{s\log(d/s)}{n}},\quad r_\delta=\sqrt{\frac{\log(1/\delta)}{n}}. \tag{3.1}$$

接下来，我们陈述主要定理：

定理 3.1. 假设假设3.1成立。令 c_1 和 c_2 为数值常数，其定义见命题3.1和命题3.2。对于调节参数 λ_o 和 λ_s，我们假设

$$\lambda_o\sqrt{n}\ge 576\sigma L^2,\quad \lambda_s\sqrt{s}=c_s\lambda_o\sqrt{n}L(r_{d,s}+r_\delta), \tag{3.2}$$

其中，c_s 表示一个数值常数，使得 $c_s\ge 5c_{1\circ}$ 假设 $r_1=3\sqrt{s}r_2$ 且 $r_2=5L\lambda_o\sqrt{n}(c_1+c_2+c_s)(r_{d,s}+r_\delta)$，并且 n 足够大，以至于

$$1\ge\max\left\{r_{d,s},r_\delta,r_2,320L^4(c_1+c_2+c_s)(r_{d,s}+r_\delta)\right\}. \tag{3.3}$$

那么，以至少 $1-2\delta$ 的概率，(2.6) 的最优解满足

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\le r_1,\quad \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\le r_2. \tag{3.4}$$

假设定理 3.1 中的条件以及 $\lambda_o\sqrt{n}=576\sigma L^2$ 成立，则有

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\le 2880L^3\sigma(c_1+c_2+c_s)(r_{d,s}+r_\delta), \tag{3.5}$$

这与 (1.2) 一致，但数值因子和对 L 的依赖性除外。从定理 3.1 的条件可以看出，定理 3.1 不需要诸如 n 至少大于 $s(\log(d/s))^2$, 这样的附加条件；这与 [Sivakumar 等人， 2015] 以及 [Genzel 和 Kipp， 2022] 的结果不同。需要指出的是，我们并未优化定理 3.1 中的数值因子，因此仍有很大的改进空间。

以下两个命题在定理 3.1 的证明中发挥着重要作用：

命题 3.1. 假设定理 3.1 中的条件成立。令 c_1 为一个（充分大的）数值常数。那么，对于任意 $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$，以至少 $1-\delta$ 的概率，有

$$\left|\sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right) \mathbf{x}_i^\top \mathbf{v}\right| \leq c_1 \lambda_o \sqrt{n} L \left(\sqrt{\frac{\log(d/s)}{n}} r_1 + \sqrt{\frac{\log(1/\delta)}{n}} r_2\right). \tag{3.6}$$

正如我们在第 2.2 节中提到的，在与命题 3.1 相对应的部分中，[Sivakumar 等人、 2015] 以及 [Genzel 和 Kipp 的结果要求 n 足够大，从而使 n至少大于 $s(\log(d/s))^2$。然而，命题 3.1 不要求 n 对 log(·) 具有二次依赖关系。

命题 3.2。假设定理 3.1 中的条件成立。令 c_2 为一个（足够大的）数值常数。则对于任意 $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$，以至少 $1-\delta$ 的概率，有

$$\sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} \left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} - \mathbf{x}_i^\top \mathbf{v}\right) + h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right) \mathbf{x}_i^\top \mathbf{v} \geq \frac{\|\mathbf{v}\|_2^2}{2} - c_2 L \lambda_o \sqrt{n} \left(\sqrt{\frac{\log(d/s)}{n}} r_1 + \sqrt{\frac{\log(1/\delta)}{n}} r_2\right). \tag{3.7}$$

致谢

我们感谢统计数学研究所的 Hironori Fujisawa，感谢他参与有益的讨论。

参考文献

[Bellec 等人、 2018] Bellec, Pierre C 和 Lecu´e, Guillaume 以及 Tsybakov, Alexandre B（2018）。斜率与套索回归的结合：改进的预言界与最优性。《统计年鉴》，46(68)：3603–3642。

[Dalalyan 和 Thompson， 2019] Dalalyan, Arnak 和 Thompson, Philip。（2019）。使用 ℓ_1-惩罚 Huber M 估计量对稀疏线性模型进行抗离群值估计。神经信息处理系统进展，13188–13198。

[Fan 等， 2021] Fan, Jianqing 和 Wang, Weichen 和 Zhu, Ziwei（2021）。重尾数据的收缩原理：高维稳健低秩矩阵恢复。《统计年鉴》，49(3)：1239–1266。

[Genzel 和 Kipp， 2022] Genzel, Martin 和 Kipp, Christian（2022）。次指数数据下广义套索回归的一般误差界。《抽样理论、信号处理与数据分析》，20(2)：15。

[Kuchibhotla 和 Chakrabortty， 2022] Kuchibhotla, Arun Kumar 和 Chakrabortty, Abhishek（2022）。超越高维统计学中的次高斯性：在协方差估计和线性回归中的应用。《信息与推断：IMA 期刊》，11(4)：1389–1456。

[Lecu´e 和 Mendelson， 2017] Lecu´e，Guillaume 和 Mendelson，Shahar（2017）。弱矩假设下的稀疏恢复。《欧洲数学学会杂志》，19(3)：881– 904。

[Nguyen 和 Tran， 2012] Nguyen, Nam H 和 Tran, Trac D.（2012）。含缺失和严重损坏观测的稳健套索回归。IEEE Transactions on Information Theory，59(4)：2036– 2058。

[Sivakumar 等， 2015] Sivakumar, Vidyashankar 和 Banerjee, Arindam 和 Ravikumar, Pradeep K（2015）。超越次高斯测量：基于次指数设计的高维结构化估计。神经信息处理系统进展，46(68)：3603–3642。

[She 和 Owen， 2011] She、Yiyuan 和 Owen，Art B.（2011）。使用非凸惩罚回归进行异常值检测。《美国统计学会会刊》，106(494)：626–639。

[Sun 等， 2020] Sun、Qiang、Zhou、Wen-Xin 和 Fan、Jianqing。（2020）。自适应 Huber 回归。《美国统计学会会刊》，115(529)：254–265。

[Tibshirani， 1996] Tibshirani，Robert。（1996）。通过套索回归进行回归收缩与变量选择。英国皇家统计学会会刊：B辑，58(1)：267–288。

[Vershynin， 2018] Vershynin，Roman。（2018）。高维概率：数据科学中的介绍与应用。剑桥大学出版社。

A 证明主要命题

本节证明命题 3.1 和 3.2。

A.1 准备工作

在 A.1 节中，我们介绍引理 A.1 和引理 A.2，它们用于命题 3.1 和命题 3.2 的证明。

引理 A.1. 对于一个 L-次指数随机向量 $\mathbf{x} \in \mathbb{R}^d$ 满足 $\mathbb{E}\mathbf{x} = 0$，对于任意固定向量 $\mathbf{v} \in \mathbb{R}^d$ 以及 $p \geq 4$，有

$$\left(\mathbb{E}|\langle \mathbf{x}^\top \mathbf{v}\rangle|^p\right)^{\frac{1}{p}} \leq 2Lp\left(\mathbb{E}|\langle \mathbf{x}^\top \mathbf{v}\rangle|^2\right)^{\frac{1}{2}}. \tag{A.1}$$

证明。我们定义 $c_L = L\left(\mathbb{E}|\langle \mathbf{x}^\top \mathbf{v}\rangle|^2\right)^{\frac{1}{2}}$。本证明基于 Vershynin[的命题 2.5.2 的证明 2018]。根据 L-次指数随机向量的定义，有 $\mathbb{E}\exp\left(|\mathbf{x}^\top \mathbf{v}|/c_L\right) \leq 2$。由此以及马尔可夫不等式，对于所有 $t \geq 0$，有

$$\mathbb{P}\left(|\mathbf{x}^\top \mathbf{v}|/c_L \geq t\right) = \mathbb{P}\left(e^{|\mathbf{x}^\top \mathbf{v}|/c_L} \geq e^t\right) \leq e^{-t}\mathbb{E}e^{|\mathbf{x}^\top \mathbf{v}|/c_L} \leq 2e^{-t}. \tag{A.2}$$

于是，有

$$\mathbb{E}|\mathbf{x}^\top \mathbf{v}|^p = \int_0^\infty \mathbb{P}\left(|\mathbf{x}^\top \mathbf{v}|^p \geq u\right) du \overset{(a)}{\leq} c_L^p \int_0^\infty 2e^{-t} p t^{p-1} dt \overset{(b)}{\leq} 2p(c_L p)^p, \tag{A.3}$$

其中，(a) 根据 (A.2) 得出，(b) 根据 $\Gamma(x) \leq x^x$ 得出，其中 $\Gamma(x)$ 是伽马函数，且使用了斯特林公式。由 $p^{1/p} \leq e^{1/e}$ 可知，证明完成。 □

在介绍引理 A.2 之前，我们先介绍 γ_α-泛函。

定义 A.1（γ_α-泛函 [Dirksen， 2018]）。设 (T, d) 为一个半度量空间；也就是说，对于 $x, y, z \in T$，有 $d(x, z) \leq d(x, y) + d(y, z)$，且 $d(x, y) = d(y, x)$。子集序列 $\mathcal{T} = \{T_n\}_{n\geq 0}$ 称为可容许的，若对于所有 $n \geq 1$，都有 $|T_0| = 1$ 且 $|T_n| \leq 2^{2^n}$。对于任意 $\alpha \in (0, \infty)$，(T, d) 的 γ_α-泛函定义为

$$\gamma_\alpha(T, d) = \inf_{\mathcal{T}} \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{\alpha}} \inf_{s\in T_n} d(t, s), \tag{A.4}$$

其中，下确界取遍所有可容许序列。

引理 A.2. 假设定理 3.1 中关于 r_1 和 r_2 的假设成立，则有

$$\gamma_1\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right) \le Cr_1\sqrt{s}\log(d/s), \quad \gamma_2\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right) \le Cr_1\sqrt{\log(d/s)}. \tag{A.5}$$

证明。根据链式理论中的标准熵界，对于 $\alpha = 1,2$，我们有

$$\gamma_\alpha\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right) \le C\int_0^{r_2}\left(\log N\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \epsilon\right)\right)^{\frac{1}{\alpha}} d\epsilon, \tag{A.6}$$

其中 $N\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \epsilon\right)$ 是 $(r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d$ 的覆盖数，即 $(r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d$ 的 ϵ-网的最小基数。

对于 $\gamma_1(\cdot)$，根据 $r_1/r_2 = 3\sqrt{s}$, $d/s \ge 3$ 以及 Vershynin 的《Plan》中的引理 3.3， 2020], 我们有

$$\gamma_1\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right) \lesssim \int_0^{r_2}\log N\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \epsilon\right)d\epsilon \lesssim \int_0^{r_2} 9s\log\frac{r_2 d}{s\epsilon}d\epsilon \lesssim r_1\sqrt{s}\log(d/s). \tag{A.7}$$

对于 $\gamma_2(\cdot)$，我们可以重复几乎与 $\gamma_1(\cdot)$ 情形相同的论证，因此省略。

□

A.2 命题 3.1 的证明

对于集合 A，我们将 conv(A) 定义为凸包。首先，根据 [Plan 和 Vershynin 的引理 3.1， 2020], $r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \subset 2\text{conv}\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d\right)$ 以及 [Oymak 的引理 D.8， 2018], 我们有

$$\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^n h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}\rangle\right| \le 2\sup_{\mathbf{v}\in\frac{r_1^2}{r_2^2}\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^n h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}\rangle\right|. \tag{A.8}$$

我们验证推论 5.2 的假设，该推论出自 [Dirksen, 2018]。首先，根据引理 A.1，我们有 $\left(\mathbb{E}\left|\langle\mathbf{x}^\top\mathbf{v}\rangle\right|^p\right)^{\frac{1}{p}} \le 2Lp\left(\mathbb{E}\langle\mathbf{x}^\top\mathbf{v}\rangle^2\right)^{\frac{1}{2}} = 2Lpr_2$；根据斯特林公式 $p^p \le p!e^p$，我们有 $\mathbb{E}|\langle\mathbf{v}^\top\mathbf{x}_i\rangle|^p \le (2pLr_2)^p \le (2eLr_2)^p p!$。然后根据 $-1 \le h(\cdot) \le 1$，我们有

$$\mathbb{E}\left|h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}\rangle\right|^p \le \mathbb{E}\left|\langle\mathbf{x}_i,\mathbf{v}\rangle\right|^p \le \frac{p!}{2}\left(4eL\sigma r_2\right)^2\left(2eLr_2\right)^{p-2}. \tag{A.9}$$

其次，对于任意 $\mathbf{v}$, $\mathbf{v}' \in \frac{r_1^2}{r_2^2}\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d$ 以及任意 $i \in 1、\cdots、n$，由 $-1 \le h(\cdot) \le 1$，我们有 $\left|h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i、\mathbf{v}-\mathbf{v}'\rangle\right| \le |\langle\mathbf{x}_i、\mathbf{v}-\mathbf{v}'\rangle|$，并且根据 L-次指数随机向量的定义，

$$\left\|h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}-\mathbf{v}'\rangle\right\|_{\psi_1} \le \left\|\langle\mathbf{x}_i,\mathbf{v}-\mathbf{v}'\rangle\right\|_{\psi_1} \le L\|\mathbf{v}-\mathbf{v}'\|_2. \tag{A.10}$$

由 (A.8)、(A.9)、(A.10)、 r_δ、 $r_2 \le 1$ 以及 [Dirksen 的推论 5.2，以至少 $1-\delta$ 的概率，我们有

$$\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^n h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}\rangle\right| \le CL\left(\frac{\gamma_1\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right)}{n} + \frac{\gamma_2\left((r_1/r_2)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d, \|\cdot\|_2\right)}{\sqrt{n}} + r_\delta r_2\right). \tag{A.11}$$

由引理 A.2 和 $r_{d,\ s} \le 1$，以至少 $1-\delta$ 的概率，有

$$\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^n h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\langle\mathbf{x}_i,\mathbf{v}\rangle\right| \le CL\left(\frac{r_{d,s}}{\sqrt{s}}r_1 + r_\delta r_2\right). \tag{A.12}$$

A.3 命题3.2的证明

在继续证明命题 3.2 之前，我们指出，类似的证明可见于[Sun 等人、 2020] 以及 [Chen 和 Zhou， 2020]。由 Huber 损失的凸性，对于任意a、 $b \in \mathbb{R}$，我们有 $H(a) - H(b) \geq h(b)(a-b)$，以及 $H(b) - H(a) \geq h(a)(b-a)$，并且我们有

$$0 \leq (h(a) - h(b))(a-b). \tag{A.13}$$

F对于事件 $\mathfrak{E}$，令 $\mathrm{I}_{\mathfrak{E}}$ 表示事件 $\mathfrak{E}$的示性函数。由 (A.13)，我们有 e

$$\lambda_o^2 \sum_{i=1}^{n} \left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} - \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right) + h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} \right) \right) \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \geq \lambda_o^2 \sum_{i=1}^{n} \left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} - \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right) + h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} \right) \right) \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \mathrm{I}_{E_i}, \tag{A.14}$$

其中 E_i 表示一个事件 $E_i := \left(\left| \frac{\xi_i}{\lambda_o\sqrt{n}} \right| \leq 1/2 \right) \cap \left(\left| \frac{\mathbf{x}_i^\top}{\lambda_o\sqrt{n}} \right| \leq 1/2 \right)^{\mathbf{v}}$。定义函数 w s

$$\varphi(x) = \begin{cases} x^2 & \text{if } |x| \leq 1/4 \\ (x-1/4)^2 & \text{if } 1/4 \leq x \leq 1/2 \\ (x+1/4)^2 & \text{if } -1/2 \leq x \leq -1/4 \\ 0 & \text{if } |x| > 1/2 \end{cases} \quad \text{and} \quad \psi(x) = \mathrm{I}_{|x| \leq 1/2}. \tag{A.15}$$

令 $f_i(\mathbf{v}) = \varphi\left(\frac{\mathbf{x}_i^\top}{\lambda_o\sqrt{n}} \right) \psi\left(\frac{\xi_i}{\lambda_o\sqrt{n}} \right)^{\mathbf{v}}$，并且我们有

$$\sum_{i=1}^{n} \left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} - \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right) + h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} \right) \right) \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \mathrm{I}_{E_i} \overset{(a)}{\geq} \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right) \psi\left(\frac{\xi_i}{\lambda_o\sqrt{n}} \right) = \sum_{i=1}^{n} f_i(\mathbf{v}), \tag{A.16}$$

其中 (a) 由 $\varphi(v) \leq v^2$ for $|v| \leq 1/2$ 得出。注意到

$$f_i(\mathbf{v}) \leq \varphi\left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right) \leq \min\left\{ \left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right)^2, \frac{1}{4} \right\}. \tag{A.17}$$

为了从下方界定 $\sum_{i=1}^{n} f_i(\mathbf{v})$，我们有

$$\inf_{\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d} \sum_{i=1}^{n} f_i(\mathbf{v}) \geq \inf_{\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d} \mathbb{E} \sum_{i=1}^{n} f(\mathbf{v}) - \sup_{\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d} \left| \sum_{i=1}^{n} f_i(\mathbf{v}) - \mathbb{E} \sum_{i=1}^{n} f_i(\mathbf{v}) \right|. \tag{A.18}$$

定义一个由 $r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$ 索引的随机过程 Δ 的上确界为 $\sup_{\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d} |\sum_{i=1}^{n} f_i(\mathbf{v}) - \mathbb{E}\sum_{i=1}^{n} f_i(\mathbf{v})|$。由 (A.15) 和 (A.16)，我们有

$$\mathbb{E} \sum_{i=1}^{n} f_i(\mathbf{v}) \geq \frac{\|\mathbf{v}\|_2^2}{\lambda_o^2} - \sum_{i=1}^{n} \mathbb{E} \left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right|^2 \mathrm{I}_{\left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right| \geq 1/2} - \sum_{i=1}^{n} \mathbb{E} \left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right|^2 \mathrm{I}_{\left| \frac{\xi_i}{\lambda_o\sqrt{n}} \right| \geq 1/2}. \tag{A.19}$$

我们注意到，根据引理 A.1 以及关于 λ_o 和 $320L^4(c_1+c_2+c_2)(r_{d,\ s}+r_\delta) \leq 1$, $r_2 = 5L^2\lambda_o\sqrt{n}(c_1+c_2+c_2)(r_{d,\ s}+r_\delta)$ 的假设，

$$\mathbb{E}(\mathbf{x}_i^\top \mathbf{v})^4 \leq 16 \times 81 \times L^4 \times \|\mathbf{v}\|_2^4, \quad \frac{64 \times 16}{\lambda_o^4 n} L^4 \|\mathbf{v}\|_2^4 \leq \frac{\|\mathbf{v}\|_2^2}{4\lambda_o^2}. \tag{A.20}$$

我们评估 对每一项取 (A.19) 的右端。首先，对于任意 $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$，我们有

$$\sum_{i=1}^{n} \mathbb{E} \left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right|^2 \mathrm{I}_{\left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right| \geq 1/2} \overset{(a)}{\leq} \sum_{i=1}^{n} \sqrt{\mathbb{E} \left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right|^4} \sqrt{\mathbb{E}\mathrm{I}_{\left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right| \geq 1/2}} \overset{(b)}{\leq} 4 \sum_{i=1}^{n} \mathbb{E} \left| \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}} \right|^4 \overset{(c)}{\leq} \frac{\|\mathbf{v}\|_2^2}{4\lambda_o^2}, \tag{A.21}$$

其中，(a) 根据 Hölder 不等式得出，(b) 根据示性函数与期望之间的关系以及马尔可夫不等式得出，(c) 根据 (A.20) 得出。其次，对于任意 $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$，我们有

$$\sum_{i=1}^n \mathbb{E}\left|\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|^2 \mathrm{I}_{\left|\frac{\xi_i}{\lambda_o\sqrt{n}}\right|\geq 1/2} \overset{(a)}{\leq} \sum_{i=1}^n \sqrt{\mathbb{E}\left|\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|^4}\sqrt{\mathbb{E}\mathrm{I}_{\left|\frac{\xi_i}{\lambda_o\sqrt{n}}\right|\geq 1/2}} \overset{(b)}{\leq} \sum_{i=1}^n \sqrt{\mathbb{E}\left|\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|^4}\sqrt{\frac{4}{\lambda_o^2 n}\mathbb{E}\xi_i^2} \overset{(c)}{\leq} \frac{\|\mathbf{v}\|_2^2}{4\lambda_o^2}, \tag{A.22}$$

其中，(a) 由 H¨older 不等式得出，(b) 由示性函数与期望之间的关系以及 Markov 不等式得出，(c) 由 (A.20) 和关于 ξ_i 的假设得出。因此，由 (A.21) 和 (A.22) 可得

$$\frac{\|\mathbf{v}\|_2^2}{2} - \lambda_o^2\Delta \leq \inf_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d} \lambda_o^2 \sum_{i=1}^n \left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}} - \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right) + h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}. \tag{A.23}$$

接下来，我们估计随机项 Δ。注意，由 (A.17) 和 (A.20)，我们有 $\mathbb{E}(f_i(\mathbf{v}) - \mathbb{E}f_i(\mathbf{v}))^2 \leq \mathbb{E}f_i^2(\mathbf{v}) \leq \frac{L^2}{\lambda_o^2 n}\|\mathbf{v}\|_2^2$。由此，根据 (A.17) 和 [Massart 的定理 3，以至少 $1-\delta$ 的概率，我们有 2000],

$$\Delta \leq 2\mathbb{E}\Delta + \sqrt{\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n \mathbb{E}(f_i(\mathbf{v}) - \mathbb{E}f_i(\mathbf{v}))^2}\sqrt{8\log(1/\delta)} + 18\log(1/\delta) \leq 2\mathbb{E}\Delta + \frac{L\sqrt{8\log(1/\delta)}}{\lambda_o}\|\mathbf{v}\|_2 + 18\log(1/\delta). \tag{A.24}$$

F由对称化不等式（ [Boucheron等人， 2013]的定理 11.4），我们有 e

$$\mathbb{E}\Delta \leq 2\,\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\varphi\left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right)\psi\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right|, \tag{A.25}$$

其中，$\{a_i\}_{i=1}^n$ 是一列与 $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$ 相互独立的独立同分布 Rademacher随机变量。我们将条件期望定义为 $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$，即 $\mathbb{E}[\,\cdot \mid \{\mathbf{x}_i, \xi_i\}_{i=1}^n]$。由 [Boucheron等人， 2013], 的定理 11.6，可得

$$\mathbb{E}\left[\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\varphi\left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right)\psi\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right|\middle|\{\mathbf{x}_i, \xi_i\}_{i=1}^n\right] \leq \mathbb{E}\left[\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|\middle|\{\mathbf{x}_i, \xi_i\}_{i=1}^n\right], \tag{A.26}$$

根据期望的基本性质，并采用与命题 3.1 的证明类似的论证，我们得到

$$\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\varphi\left(\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right)\psi\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right| \leq \mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right| \leq \mathbb{E}\sup_{\mathbf{v}\in \frac{r_1^2}{r_2^2}\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^n a_i\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|. \tag{A.27}$$

我们将应用 [Dirksen 2018], 的推论 5.2，并验证 [Dirksen 2018] 的推论 5.2 的假设。由引理 A.1 和斯特林公式 $p^p \leq p!e^p$，我们得到

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left|a_i\frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o\sqrt{n}}\right|^p \leq \frac{\mathbb{E}\left|\mathbf{x}_i^\top \mathbf{v}\right|^p}{(\lambda_o\sqrt{n})^p} \leq \frac{2^p p^p L^p(\mathbb{E}\langle \mathbf{x}_i, \mathbf{v}\rangle^2)^{\frac{p}{2}}}{(\lambda_o\sqrt{n})^p} \leq 2^p p^p\left(\frac{Lr_2}{\lambda_o\sqrt{n}}\right)^p \leq \frac{p!}{2}\left(\frac{4eLr_2}{\lambda_o\sqrt{n}}\right)^p. \tag{A.28}$$

对于任意 $\mathbf{v}$，$\mathbf{v}' \in \frac{r_1^2}{r_2^2}\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d$ 且对于任意 $i = 1$、$\cdots$、n，我们有 $\left|a_i\mathbf{x}_i^\top \mathbf{v} - a_i\mathbf{x}_i^\top \mathbf{v}'\right| \leq |\langle \mathbf{x}_i, \mathbf{v} - \mathbf{v}'\rangle|$。然后，根据 L-次指数随机向量的定义，我们有

$$\left\|a_i\mathbf{x}_i^\top \mathbf{v} - a_i\mathbf{x}_i^\top \mathbf{v}'\right\|_{\psi_1} \leq L\left\|\mathbf{v} - \mathbf{v}'\right\|_2. \tag{A.29}$$

由 (A.28) 和 (A.29)，采用与命题 3.1 的证明类似的论证，我们得到

$$\frac{1}{n}\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i=1}^n a_i\frac{\mathbf{x}_i^\top\mathbf{v}}{\lambda_o\sqrt{n}}\right|\le C\frac{L}{\lambda_o\sqrt{n}}\left(\frac{r_{d,s}}{\sqrt{s}}r_1+r_\delta r_2\right). \tag{A.30}$$

由 (A.23)、(A.25)、(A.30) 和 $\lambda_o^2\log(1/\delta)\le(\lambda_o\sqrt{n})^2r_\delta^2\le\lambda_o\sqrt{n}r_2r_\delta$，对于任意 $\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d$，我们有

$$\lambda_o^2\sum_{i=1}^n\left(-h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}-\frac{\mathbf{x}_i^\top\mathbf{v}}{\lambda_o\sqrt{n}}\right)+h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\frac{\mathbf{x}_i^\top\mathbf{v}}{\lambda_o\sqrt{n}}\ge\frac{\|\mathbf{v}\|_2^2}{2}-CL\lambda_o\sqrt{n}\left(\frac{r_{d,s}}{\sqrt{s}}r_1+r_\delta r_2\right). \tag{A.31}$$

B 主定理的证明

本节中，我们证明定理 3.1。在开始证明定理 3.1 之前，我们指出， [Alquier 等人 2019]对命题 9.1 的证明中也可以看到与定理 3.1 非常相似的思路。证明分为两个主要步骤。假设定理 3.1 中的假设成立，则 (3.6) 和 (3.7) 以至少 $1-2\delta$ 的概率成立。在 B 节余下部分，我们假设 (3.6) 和 (3.7) 成立。

B.1 定理 3.1 的证明：步骤 1

^^ 我们将 $\boldsymbol{\delta}$ 定义为 $\boldsymbol{\beta}-\boldsymbol{\beta}^*$，将 $\boldsymbol{\delta}_\eta$ 定义为 $\eta(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$。若 $\|\boldsymbol{\delta}\|_1>r_1$ 和 $\|\boldsymbol{\delta}\|_2>r_2$成立，则可导出矛盾。假设 $\|\boldsymbol{\delta}\|_1>r_1$ 和 $\|\boldsymbol{\delta}\|_2>r_2$成立。于是，我们可以找到某些 η_1、$\eta_2\in(0,1)$，使得$\|\boldsymbol{\delta}_{\eta_1}\|_1=r_1$和 $\|\boldsymbol{\delta}_{\eta_2}\|_2=r_2$成立。定义 $\eta_3=\min\{\eta_1,\eta_2\}$。我们在 B.1.1 节讨论 $\eta_3=\eta_2$的情形，在 B.1.2 节讨论 $\eta_3=\eta_1$ 的情形。

B.1.1 定理 3.1 的证明：步骤 1(a)

假设 $\eta_3=\eta_2$，并注意到 $\|\boldsymbol{\delta}_{\eta_3}\|_2=r_2$ 和 $\|\boldsymbol{\delta}_{\eta_3}\|_1\le r_1$ 成立。在 B.1.1 节中，令 $\eta=\eta_3(=\eta_1)$。令

$$Q'(\eta)=\frac{\lambda_o}{\sqrt{n}}\sum_{i=1}^n\left(-h\left(\frac{y_i-\mathbf{x}_i^\top(\boldsymbol{\beta}^*+\boldsymbol{\delta}_\eta)}{\lambda_o\sqrt{n}}\right)+h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\mathbf{x}_i^\top\boldsymbol{\delta}. \tag{B.1}$$

根据 [Fan 等的引理 F.2. 证明，我们有 $\eta Q'(\eta)\le\eta Q'(1)$，这意味着

$$\sum_{i=1}^n\frac{\lambda_o}{\sqrt{n}}\left(-h\left(\frac{y_i-\mathbf{x}_i^\top(\boldsymbol{\beta}^*+\boldsymbol{\delta}_\eta)}{\lambda_o\sqrt{n}}\right)+h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\mathbf{X}_i^\top\boldsymbol{\delta}_\eta\le\sum_{i=1}^n\frac{\lambda_o}{\sqrt{n}}\eta\left(-h\left(\frac{\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}}{\lambda_o\sqrt{n}}\right)+h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\mathbf{X}_i^\top\boldsymbol{\delta}. \tag{B.2}$$

对于向量 $\mathbf{v}\in\mathbb{R}$，令 $\partial\mathbf{v}$ 为 $\|\mathbf{v}\|_1$ 的次微分，并注意到 $\|\hat{}\boldsymbol{\beta}\|_1-\|\boldsymbol{\beta}^*\|_1\le\langle\partial\hat{}\boldsymbol{\beta},\boldsymbol{\delta}\rangle,\hat{}$，这正是次微分的定义。因此，将 $\eta\lambda_s(\|\boldsymbol{\beta}\|_1-\|\boldsymbol{\beta}^*\|_1)$ 加到 (B.2) 的两边，我们有

$$\sum_{i=1}^n\frac{\lambda_o}{\sqrt{n}}\left(-h\left(\frac{y_i-\mathbf{x}_i^\top(\boldsymbol{\beta}^*+\boldsymbol{\delta}_\eta)}{\lambda_o\sqrt{n}}\right)+h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\right)\mathbf{X}_i^\top\boldsymbol{\delta}_\eta\le\sum_{i=1}^n\frac{\lambda_o}{\sqrt{n}}h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right)\mathbf{X}_i^\top\boldsymbol{\delta}_\eta+\eta\lambda_s(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1), \tag{B.3}$$

^^ 这里我们利用了 $\boldsymbol{\beta}$ 的最优性。注意到 $\eta\lambda_s(\|\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}\|_1)\le\lambda_s\|\boldsymbol{\delta}_\eta\|_1$，并由 (3.6)、(3.7) 和 $r_1=3\sqrt{s}r_2$，得到 $\|\boldsymbol{\delta}_\eta\|_2^2/4\le(c_1+c_2+c_s)\lambda_o\sqrt{n}L(r_{d,\ s}+r_\delta)\ \|\boldsymbol{\delta}_2\|_2$。由此及$\|\boldsymbol{\delta}_\eta\|_2\ge0$，得到 $\|\boldsymbol{\delta}_\eta\|_2\le4(c_1+c_2+c_s)\lambda_o\sqrt{n}L(r_{d,\ s}+r_\delta)$。这与$\|\boldsymbol{\delta}_\eta\|_2=r_2$矛盾。

B.1.2 定理 3.1 的证明：步骤 1(b)

假设 $\eta_3 = \eta_1$，则我们看到 $\|\boldsymbol{\delta}_{\eta_3}\|_2 \le r_2$ 和 $\|\boldsymbol{\delta}_{\eta_3}\|_1 = r_1$ 成立。在 B.1.2 节中，令$\eta = \eta_3(= \eta_1)$。首先假设 $\|\boldsymbol{\delta}_\eta\|_2 \ge \|\boldsymbol{\delta}_\eta\|_1/\sqrt{s}$。然后我们立即观察到 $\|\boldsymbol{\delta}_\eta\|_1 \le\sqrt{s}\|\boldsymbol{\delta}_\eta\|_2$ 成立，并由 $\|\boldsymbol{\delta}_\eta\|_1 = r_1 = 3\sqrt{s}r_2$，这与之矛盾。

接下来，假设$\|\boldsymbol{\delta}_\eta\|_2 \le \|\boldsymbol{\delta}_\eta\|_1/\sqrt{s}$。由 (A.13) 可知，(B.2) 左侧为正，因此有

$$0 \le \frac{1}{n}\sum_{i=1}^{n} \lambda_o\sqrt{n}h\left(\frac{y_i - \mathbf{x}_i^\top\boldsymbol{\beta}^*}{\lambda_o\sqrt{n}}\right)\mathbf{X}_i^\top\boldsymbol{\delta}_\eta + \eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1). \tag{B.4}$$

由 (B.4)、(3.6)、$r_1 = 3\sqrt{s}r_2$、$\max\{r_2, r_\delta, r_{d,s}\} \le 1$ 以及假设 $\|\boldsymbol{\delta}_\eta\|_2 \le \|\boldsymbol{\delta}_\eta\|_1/\sqrt{s}$，可得

$$0 \le c_1\lambda_o\sqrt{n}L\left(r_{d,s}r_1/\sqrt{s} + r_\delta r_2\right) + \eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \le C_s\|\boldsymbol{\delta}_\eta\|_1 + \eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1), \tag{B.5}$$

其中 $C_s = c_1\lambda_o\sqrt{n}L(r_{d,\ s} + r_\delta)/\sqrt{s}$。定义 $\mathcal{J}_\mathbf{a}$ 为 a 的非零元素的指标集，并令$\boldsymbol{\delta}_{\eta,\ \mathcal{J}_\mathbf{a}}$为满足 $(\boldsymbol{\delta}_{\eta,\ \mathcal{J}_\mathbf{a}})_i = (\boldsymbol{\delta}_\eta)_i$（对于 $i \in \mathcal{J}_\mathbf{a}$）以及 $(\boldsymbol{\delta}_{\eta,\ \mathcal{J}_\mathbf{a}})_i = 0$（对于 $i \notin \mathcal{J}_\mathbf{a}$）的向量。此外，我们注意到

$$\begin{aligned}
&0 \le C_s\|\boldsymbol{\delta}_\eta\|_1 + \eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \\
&\le C_s(\|\boldsymbol{\delta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \|\boldsymbol{\delta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) + \eta\lambda_s(\|\boldsymbol{\beta}^*_{\mathcal{J}_{\boldsymbol{\beta}^*}} - \hat{\boldsymbol{\beta}}_{\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) = (\lambda_s + C_s)\,\|\boldsymbol{\delta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + (-\lambda_s + C_s)\,\|\boldsymbol{\delta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1.
\end{aligned} \tag{B.6}$$

随后，由 $c_s \ge 5c_1$ 可得

$$\|\boldsymbol{\delta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1 \le \frac{\lambda_s + C_s}{\lambda_s - C_s}\|\boldsymbol{\delta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \le (3/2)\|\boldsymbol{\delta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1, \tag{B.7}$$

并且有$\|\boldsymbol{\delta}_\eta\|_1 = \|\boldsymbol{\delta}_{\eta,\ \mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \|\boldsymbol{\delta}_{\eta,\ \mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1 \le (3/2+1)\|\boldsymbol{\delta}_{\eta,\ \mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \le (5/2)\sqrt{s}\|\boldsymbol{\delta}_\eta\|_2$。于是产生矛盾，因为 $\|\boldsymbol{\delta}_\eta\|_1 = r_1 = 3\sqrt{s}r_2$。

B.2 定理 3.1：步骤 2 的证明

根据 B.1 节中的论证，$\|\boldsymbol{\delta}\|_2 \le r_2$ 或 $\|\boldsymbol{\delta}\|_1 \le r_1$ 成立。在 B.2.1 节中，假设 $\|\boldsymbol{\delta}\|_2 > r_2$ 和 $\|\boldsymbol{\delta}\|_1 \le r_1$ 成立，然后导出矛盾。在 B.2.2 节中，假设 $\|\boldsymbol{\delta}\|_2 \le r_2$ 和 $\|\boldsymbol{\delta}\|_1 > r_1$ 成立，然后导出矛盾。最后，我们有 ˆ ˆ$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le r_2$ 和 $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le r_1$，证明完成。

B.2.1 定理 3.1 的证明：步骤 2(a)

假设 $\|\boldsymbol{\delta}\|_2 > r_2$ 和 $\|\boldsymbol{\delta}\|_1 \le r_1$ 成立，则我们观察到 $\eta_4 \in (0, 1)$，使得 $\|\boldsymbol{\delta}_{\eta_4}\|_2 = r_2$成立。注意到 $\|\boldsymbol{\delta}_{\eta_4}\|_1 \le r_1$ 也成立。然后，沿用 B.1.1 节中的相同论证，可得矛盾。

B.2.2 定理 3.1 的证明：步骤 2(b)

假设 $\|\boldsymbol{\delta}\|_2 \le r_2$ 和 $\|\boldsymbol{\delta}\|_1 > r_1$ 成立，则我们观察到 $\eta_5 \in (0, 1)$，使得 $\|\boldsymbol{\delta}_{\eta_5}\|_1 = r_1$成立。注意到 $\|\boldsymbol{\delta}_{\eta_5}\|_2 \le r_2$ 也成立。然后，沿用 B.1.2 节中的相同论证，可得矛盾。

参考文献

[Alquier 等人， 2019] Alquier, Pierre and Cottet, Vincent and Lecu´e, Guillaume。（2019）。带 Lipschitz 损失函数的正则化方法的估计界与尖锐 oracle 不等式。《统计年鉴》，47(4):2117–2144。

[Boucheron等人， 2013] Boucheron, St´ephane and Lugosi, G´abor and Massart, Pascal。（2013）。集中不等式：独立性的非渐近理论。牛津大学出版社。[Chen and Zhou， 2020] Chen, Xi and Zhou, Wen-Xin。（2020）。通过乘子自助法进行稳健推断。《统计年鉴》，48(3):1665–1691。[Dirksen， 2018] Dirksen, Sjoerd。（2015）。通过泛型链获得尾界。《电子概率杂志》，20:1–29。[Fan 等， 2018] Fan, Jianqing and Liu, Han and Sun, Qiang and Zhang, Tong。（2018）。用于稀疏学习的 I-LAMM：同时控制算法复杂度与统计误差。《统计年鉴》，46(2):814。[Massart， 2000] Massart, Pascal。（2000）。关于经验过程 Talagrand 集中不等式中的常数。《概率年鉴》，28(2):863-884。[Oymak， 2018] Oymak, Samet。（2018）。通过正则化学习紧凑型神经网络。国际机器学习大会，3966–3975。[Plan and Vershynin， 2020] Plan, Yaniv and Vershynin, Roman。（2013）。通过线性规划进行单比特压缩感知。《纯粹与应用数学通讯》，66(8):1275–1297。[Sun 等， 2020] Sun, Qiang and Zhou, Wen-Xin and Fan, Jianqing。（2020）。自适应 Huber 回归。《美国统计学会会刊》，115(529):254–265。