arXiv:2305.06199v1 [math.ST] 10 May 2023

Computationally Efficient and Statistically Optimal Robust High-Dimensional Linear Regression

Yinan Shen、Jingyang Li、Jian-Feng Cai*和
Dong Xia†香港科技大学

(May 11, 2023)

Abstract

在重尾噪声或离群值污染下进行高维线性回归，无论在计算上还是统计上都颇具挑战。凸方法已被证明在统计上是最优的，但计算成本高昂，尤其是稳健损失函数通常是非光滑的。近年来，有人提出了通过次梯度下降实现快速计算的非凸方法，但遗憾的是，即使在次高斯噪声下，这些方法也无法得到统计一致的估计量。本文针对稀疏线性回归和低秩线性回归问题，提出了一种投影次梯度下降算法。该算法不仅计算高效、具有线性收敛性，而且在噪声为高斯或具有有限 $1+\varepsilon$ 阶矩的重尾噪声时，都能达到统计最优。我们在一个一般框架下建立了收敛理论，并研究了该框架在绝对损失、Huber 损失和分位数损失中的具体应用。与现有的非凸方法相比，我们的方法揭示了一种出人意料的两阶段收敛现象。在第一阶段，算法的表现与典型的非光滑优化相同，需要逐步减小步长。然而，第一阶段只能得到统计上次优的估计量，这一点在现有文献中已有发现。值得注意的是，在第二阶段，算法的收敛速度呈线性，仿佛是在最小化一个光滑且强凸的目标函数，因此使用恒定步长即可。第二阶段收敛的内在原因是：随机噪声在一个接近但又不太接近真实值的区域内，对非光滑稳健损失产生了平滑效应。数值模拟验证了我们的理论发现，并展示了本算法相较于以往方法的优越性。

*Jian-Feng Cai's research was partially supported by Hong Kong RGC Grant GRF 16310620 and GRF 16309219.
†Dong Xia's research was partially supported by Hong Kong RGC Grant ECS 26302019, GRF 16303320 and GRF 16300121.

1 引言

设 $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ 是一组相互独立的观测值，满足

$$Y_i = \langle \mathbf{X}_i, \mathbf{T}^* \rangle + \xi_i \tag{1}$$

其中，$\mathbf{X}_i$ 和 Y_i 分别为协变量和响应变量。这里，$\langle \cdot, \cdot \rangle$ 表示欧几里得空间中的内积。为便于分析，通常假设潜在噪声 ξ_i 是随机的。我们假设 $\mathbf{T}^*$ 位于一个高维环境空间中，但具有低维结构。我们重点研究两个高维线性回归问题：稀疏线性回归（其中 $\mathbf{T}^* \in \mathbb{R}^d$ 是稀疏的）和低秩线性回归（其中 $\mathbf{T}^* \in \mathbb{R}^{d \times d}$ 是低秩的）。对于这两个问题，如何以尽可能少的观测，在计算和统计两方面有效地恢复 $\mathbf{T}^*$，都具有重要意义。关于其动机和应用已有大量文献讨论，例如 Tibshirani (1996); Needell and Tropp (2009); Blumensath 和 Davies (2009, 2010); Loh and Wainwright (2011); Shen 和 Li (2017); Zhu 等人 (2020); Zhao 等人 (2022); Gross 等人 (2010); Xia and Koltchinskii (2016); Chen 等人 (2011); Negahban and Wainwright (2011); Chiu 等人 (2021); Siddiqi 等人 (2010)。

高维线性回归的方法大致可分为两类：凸方法和非凸方法。典型的凸方法会引入额外的惩罚项，例如矩阵核范数以促进低秩解，以及 ℓ_1 范数以促进稀疏性。凸性使得针对凸规划的成熟算法可以直接应用。因此，统计学家只需专注于研究统计性能，无须过多关注计算实现。例如，假设噪声为方差为 σ^2 的独立同分布（次）高斯噪声，并在低秩线性回归中，为估计一个 $d \times d$ 矩阵而假设满足所谓的受限等距性质（RIP）或受限强凸性（RSC），Candes 和 Plan（2011）；Rohde 和 Tsybakov（2011）；Negahban 和 Wainwright（2011）；Cai 和 Zhang（2015）；Chen 和 Wainwright（2015）；Davenport 和 Romberg（2016）；Chandrasekaran 等人（2012）证明，核范数惩罚最小二乘估计量在平方弗罗贝尼乌斯范数下的误差率为 $\tilde{O}_p(\sigma^2 rdn^{-1})$ [1]，其中 r 为秩。研究表明，这一误差率在极小极大意义下是最优的。参见例如 Xia（2014）、Ma 和 Wu（2015）及其中的参考文献。尽管凸方法在理论上表现出色，但有两个主要缺点。首先，凸方法直接在高维环境空间中运算，速度较慢，且难以扩展到超高维问题。其次，虽然已有的误差率在理论上是极小极大最优的，但其隐含常数因子似乎较大，因此在实践中，所得误差率往往不如非凸方法。非凸方法可以很好地解决这些问题

[1]这里，$\tilde{O}_p(\cdot)$ 表示忽略对数因子后的典型大 O 记号，并且以高概率成立。

。在低秩线性回归中，非凸方法可以基于矩阵分解进行运算，并采用梯度下降类算法（Chen 和 Wainwright，2015；Burer 和 Monteiro，2003；Zheng 和 Lafferty，2016；Zhao 等人，2015；Xia 和 Yuan，2021；Wei 等人，2016）；在稀疏线性回归中，典型的非凸方法包括最小角回归（Efron 等人，2004）、迭代硬阈值法（Blumensath 和 Davies，2009）以及最佳子集选择（Zhu 等人，2020）等。在相似条件下，研究表明，这些非凸算法收敛速度很快（更确切地说，是线性收敛），并且在（次）高斯噪声下也能得到极小极大最优的估计量。需要指出的是，非凸方法往往需要更多调节参数。例如，矩阵分解算法通常需要知道真实秩（Chandrasekaran 等人，2012；Chen 和 Wainwright，2015），而迭代硬阈值算法则需要掌握潜在稀疏度的信息（Blumensath 和 Davies，2009；Efron 等人，2004）。

数据技术的近期蓬勃发展给高维线性回归带来了新的挑战，其中，重尾噪声和异常值在许多应用中频繁出现，例如扩散加权成像（Chang 等人，2005）、在线广告（Sun 等人，2017）以及基因表达数据分析（Sun 等人，2020）。当噪声具有重尾特性或数据受到部分污染时，上述通过最小化平方损失进行估计的方法就会变得脆弱，甚至完全失效。为了减轻重尾噪声和数据污染的影响，一种自然的解决方案是用更稳健的损失函数替代平方损失。典型的例子包括绝对损失（Cand`es 等人，2011；Cambier 和 Absil，2016）、著名的 Huber 损失（Huber，1965）以及分位数损失（Koenker 和 Hallock，2001），这些损失函数均为凸函数，但不光滑。Elsener 和 van de Geer（2018）提出了一种基于核范数惩罚绝对损失或 Huber 损失的凸方法，用于低秩线性回归，并证明，只要噪声在原点的某个邻域内具有非零密度，该估计量的平方弗罗贝尼乌斯范数误差率即可达到 $\tilde{O}_p(rd_1n^{-1})$。有趣的是，Elsener 和 van de Geer（2018）在选择调节参数时无需估计噪声水平。Alquier 等人（2019）研究了一个更一般的框架，该框架仅要求损失函数满足利普希茨连续性和凸性，但对噪声施加了所谓的伯恩斯坦条件。另请参见 Klopp 等人（2017）。上述方法仍然依赖凸规划，并受到前文所述计算效率低下问题的困扰。事实上，当目标函数不光滑时，这一计算问题会更加严重（Boyd 等人，2004）。

基于矩阵分解的次梯度下降算法，近期由Charisopoulos 等人（2021）；Li 等人（2020）；Tong 等人（2021b）在低秩线性回归中进行了研究。这些算法在采用良好初始化并适当选择步长时会线性收敛，但在重尾噪声下，这些非凸方法的统计性能要么鲜有研究，要么通常并非最优。具体而言，Li 等人（2020）仅证明了

在无噪声且存在稀疏异常值的情况下可以精确恢复；Charisopoulos 等人（2021）和 Tong 等人（2021b）建立的平方弗罗贝尼乌斯范数误差率变为 $\tilde{O}_p(\sigma^2)$，即使噪声为方差 σ^2 的高斯噪声也是如此，因此不具备统计一致性。此外，关于重尾噪声或部分污染情况下稀疏线性回归的研究文献也非常丰富，例如 Sun 等人（2020）；Zheng 等人（2015）；Belloni 等人（2011）；Alquier 等人（2019）；Lecu´e 和 Mendelson（2017）；Lugosi 和 Mendelson（2019b）；Pan 等人（2021）；Wang 等人（2021, 2020）；Thompson（2020）。Sun 等人（2020）在噪声满足有限 $1+\varepsilon$ 阶矩条件的情况下研究了自适应 Huber 损失。Depersin（2020）研究了中位数均值法，并通过 VC 维推导出紧确的误差率，实现了首批次高斯稀疏估计量。Belloni 等人（2011）和 Wang 等人（2020）分别提出了平方根 Lasso 和绝对损失。尽管这些先前方法对重尾噪声具有统计稳健性，却没有提供可证明快速收敛的算法。在 Foucart 和 Lecu´e（2017）中，研究者提出了一种用于稀疏线性回归的迭代硬阈值算法，通过最小化绝对损失实现快速收敛，但推导出的误差率并非最优。

本文提出了一种基于投影次梯度下降、使用稳健损失函数进行高维线性回归的计算高效非凸算法。对于稀疏线性回归问题，该算法本质上是绝对损失硬阈值算法（IHT-ℓ_1，Foucart 和 Lecu´e（2017））；对于低秩线性回归问题，则本质上是黎曼次梯度下降算法（RsGrad, Vandereycken（2013）；Cambier 和 Absil（2016）；Wei 等人（2016）；Cai 等人（2021a））。在这两个问题中，次梯度下降算法都具备快速计算和线性收敛的特点。不同于 Foucart 和 Lecu´e（2017）以及 Tong 等人（2021b），我们发现次梯度下降算法呈现出一种称为两阶段收敛的奇特现象。在第一阶段，该算法表现得类似于典型的非光滑优化，需要在迭代过程中逐渐减小步长。例如，在低秩线性回归中，如果噪声服从方差为 σ^2 的高斯分布，并采用绝对损失，则第一阶段的迭代呈线性收敛，并达到一个平方弗罗贝尼乌斯范数误差为 $O_p(\sigma^2)$ 的估计量。Tong 等人（2021b）和 Charisopoulos 等人（2021）已经达到了这一速率，但他们采用的是缩放次梯度（ScaledSM）下降算法。在第二阶段，该算法（RsGrad）的表现类似于光滑优化，即采用恒定步长即可保证线性收敛。因此，在第二阶段迭代之后，RsGrad 能够输出一个在温和条件下具有统计最优性的估计量；例如，对于低秩线性回归，它可以达到 $O_p(\sigma^2 rdn^{-1})$ 的速率。稀疏线性回归中也会出现两阶段收敛现象。对于这两个问题，在重尾噪声下都可以达到极小极大最优速率。

我们的贡献是多方面的。首先，我们提出了一种用于鲁棒高维线性回归的投影次梯度下降算法，该算法适用于稀疏线性回归和低秩线性回归问题中的一大类非光滑损失函数。尽管

Cambier 和 Absil（2016）已提出黎曼次梯度下降（RsGrad）来最小化绝对损失，但据我们所知，迄今尚无关于其收敛性的理论保证。在温和条件下，我们证明该算法呈线性收敛。其次，我们证明了，对于噪声服从高斯分布或具有重尾分布的情形，我们的算法为这两个问题得到的最终估计量具有统计最优性。例如，RsGrad 在采用绝对损失、Huber 损失和分位数损失的低秩线性回归中的具体应用表明，只要噪声具有良好的 $1+\varepsilon$ 矩且其密度满足温和的正则性条件，就可以达到 rate$O_p(rdn^{-1})$。类似条件也曾出现在 Elsener 和 van de Geer（2018）的研究中。不同于 Fan 等人（2021）和 Minsker（2018），我们的速率与噪声大小成正比。第三，我们的分析揭示了两阶段收敛这一新现象，使我们能够达到统计最优的误差速率，这相较于现有工作是一个重大改进（Foucart 和 Lecu´e，2017；Li 等人，2020；Tong 等人，2021b；Charisopoulos 等人，2021）。第一阶段的收敛是非光滑优化中的典型现象，而有趣的是，第二阶段的收敛表现得类似于光滑优化。看来，随机噪声在第二阶段收敛中具有平滑化作用。

本文其余部分安排如下。在第 2节中，我们回顾三种典型的非光滑稳健损失函数。在第 3节中，我们给出一个稳健稀疏线性回归的入门示例，并探讨迭代硬阈值算法的两阶段收敛。第 4节介绍了一个通用框架，用于研究黎曼次梯度下降算法在稳健低秩线性回归中的两阶段收敛性能。非凸算法不可避免地涉及重要的调优参数，我们将在第 5节讨论这些参数。在第 6节中，我们进一步研究了所提出的方法在 Huber 污染模型下的表现。第 7节展示数值实验结果，并将其与已有方法进行比较。所有证明均放在附录中。

2 稳健损失函数

记 $\mathbb{M}$ 为该优化问题的可行集。在低秩回归（第 4节）中，它可以是秩不超过 r 的矩阵集合；在稀疏回归（第 3节）中，则可以是支撑大小不超过 $\tilde{s}$ 的向量集合。线性回归的目标是求解

$$\hat{\mathbf{T}} := \operatorname*{arg\,min}_{\mathbf{T}\in\mathbb{M}} f(\mathbf{T}) \quad \text{where } f(\mathbf{T}) := \sum_{i=1}^{n} \rho\big(\langle \mathbf{T}, \mathbf{X}_i\rangle - Y_i\big). \tag{2}$$

ˆ 此处 $\rho(\cdot): \mathbb{R} \mapsto \mathbb{R}_+$ 是经过适当选择的损失函数。 $\mathbf{T}$ ˆ 的统计性质在很大程度上取决于损失函数。例如，$\mathbf{T}$ 使用 ℓ_2-损失（平方损失）可实现，即 $\rho(x):=x^2$，已被证明是有效的（Chen 和 Wainwright，2015；Wei 等，2016；Xia 和 Yuan，2021；

Jain 等人，2014；Agarwal 等人，2010）在处理次高斯噪声时表现良好，但众所周知，它对异常值和重尾噪声较为敏感（Huber，1965）。所谓的稳健损失函数相对不易受异常值和重尾噪声影响，并能够给出更可靠的估计值$\mathbf{T}$。稳健损失函数的示例包括：

1. 绝对损失（ℓ_1-损失）：$\rho(x) := |x|$ 对任意 $x \in \mathbb{R}$；2. Huber 损失：$\rho_{H,\delta}(x) := x^2 \mathbb{I}(|x| \leq \delta) + (2\delta|x| - \delta^2)\mathbb{I}(|x| > \delta)$ 对任意 $x \in \mathbb{R}$，其中 $\delta > 0$ 是一个调节参数；3. 分位数损失：$\rho_{Q,\delta}(x) := \delta x \mathbb{I}(x \geq 0) + (\delta - 1)x \mathbb{I}(x < 0)$ 对任意 $x \in \mathbb{R}$，且有 $\delta := \mathbb{P}(\xi \leq 0)$。

绝对损失、Huber 损失和分位数损失都是凸函数。它们已见于稀疏回归文献中，包括 LASSO 型凸方法（Sun 等，2020；Fan 等，2021）以及通过迭代算法实现的非凸方法（Foucart 和 Lecu´e，2017）。这些稳健损失也已用于低秩回归研究，例如基于核范数惩罚的凸方法（Elsener和van de Geer，2018；Klopp 等，2017；Cand`es 等，2011），以及基于梯度下降型算法的非凸方法（Li 等人，2020；Tong 等人，2021a）。与平方损失相比，上述稳健损失函数并非光滑函数，即它们的导数不连续，这给计算带来了新的挑战。实际上，优化问题（2）通常会利用稳健损失函数的次梯度，记作 $\partial f(\mathbf{T})$。例如，参见 Charisopoulos 等人（2021）及其中的参考文献。最近，Thompson（2020）表明，与使用标准 Huber 损失相比，使用排序 Huber 损失进行正则化可以显著改善误差率中的常数。然而，由于在我们的框架中计算排序 Huber 损失颇具挑战，我们将其留待未来研究。

ˆ 现有统计学文献已经证明，$\mathbf{T}$ 通过使用稳健损失函数获得的结果ˆ具有针对重尾噪声的统计稳健性。然而，最优解$\mathbf{T}$ 通常通过缺乏收敛保证的启发式算法（Sun 等人，2020）或非光滑优化算法（Elsener 和 van de Geer，2018）求解，而这些算法通常可能收敛缓慢，且计算成本可能相当高，因此，前述稳健损失函数在实际高维问题中的适用性受到质疑。近年来，Charisopoulos 等人（2021）；Tong 等人（2021b）；Foucart 和 Lecu´e （2017）研究了用于最小化稳健损失函数的次梯度下降算法，并表明，只要沿迭代过程合理安排步长，这些算法就能够快速收敛。遗憾的是，他们的分析框架均未能证明，次梯度下降算法能够实现此前统计学文献所声称的稳健损失函数在统计方面的优势。统计效率与计算效率之间仍存在差距。

在后续章节3 和4中，我们将证明，采用新颖步长方案的修订的次梯度下降算法能够快速收敛，并在各种设定下达到统计最优速率。除了算法设计之外，新颖的分析框架也对发展我们的统计理论起着至关重要的作用。

3 预热：稀疏线性回归

我们先通过一个使用绝对损失的高维稀疏线性回归示例进行预热。设 $\boldsymbol{\beta}^* \in \mathbb{R}^d$ 为稀疏向量，其支撑集大小为 $|\mathrm{supp}(\boldsymbol{\beta}^*)| = s \ll d$。我们观测到一组相互独立的随机对 $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$，满足 $Y_i = \langle \mathbf{X}_i, \boldsymbol{\beta}^* \rangle + \xi_i$ 其中，预测向量 $\mathbf{X}_i \in \mathbb{R}^d$ 假设为随机采样得到。我们的目标是在噪声 $\boldsymbol{\beta}^*$ 可能具有重尾分布时估计 ξ_i。为便于阐述，我们仅研究绝对损失（Foucart and Lecu´e，2017），但也可以为其他稳健损失函数推导出类似结果（见第4节）。记

$$\hat{\boldsymbol{\beta}} := \underset{|\mathrm{supp}(\boldsymbol{\beta})| \le \tilde{s}}{\arg\min}\ f(\boldsymbol{\beta}), \quad \text{where } f(\boldsymbol{\beta}) := \|\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}\|_1 := \sum_{i=1}^n |Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i \rangle|. \tag{3}$$

这里选取的稀疏度为 $\tilde{s} \in [s, d]$，$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$，$\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$。令 $\|\cdot\|_p$表示向量的 ℓ_p-范数，其中 $1 \le p \le 2$。

3.1 算法：迭代硬阈值法

我们用于优化程序（3）的方法本质上是投影次梯度下降算法。投影梯度下降算法已被用于最小化平方损失（Agarwal 等人，2010；Chen 和 Wainwright，2015），这种算法易于实现且计算速度快。由于绝对损失不是光滑函数，我们采用次梯度。

由于将向量投影到稀疏向量集合上可以通过硬阈值实现，因此我们的算法通常称为迭代硬阈值法（Blumensath 和 Davies，2009，2010）。伪代码见算法1。算子 $\mathcal{H}_{\tilde{s}}(\mathbf{V})$ 保留$\mathbf{V}$的 $\tilde{s}$ 个绝对值最大的元素，并将其余元素置零。实践中，我们选择次梯度$\mathbf{G}_l = \sum_{i=1}^n sign(\langle \boldsymbol{\beta}_l, \mathbf{X}_i \rangle - Y_i)\mathbf{X}_i$，并设置 $sign(0) = 0$。请注意，算法1 允许任意初始化。

3.2 Theory: two-phase convergence

ˆ 算法 1的收敛性能以及 $\boldsymbol{\beta}$ 的统计性质依赖于以下两个假设。

算法 1 迭代硬阈值法（ℓ_1-损失）

输入：观测值 $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$、最大迭代次数 $l_{\max}$、步长 $\{\eta_l\}_{l=0}^{l_{\max}}$ 和稀疏参数 $\tilde{s}$。初始化：任意 $\boldsymbol{\beta}_0$，其支持大小为 $|\mathrm{supp}(\boldsymbol{\beta}_0)| \le \tilde{s}$，对于 $l = 0, \ldots\ldots, l_{\max}$ 执行 选择一个次梯度：$\mathbf{G}_l \in \partial f(\boldsymbol{\beta}_l)$ 截短：$\boldsymbol{\beta}_{l+1} = \mathcal{H}_{\tilde{s}}(\boldsymbol{\beta}_l - \eta_l \mathbf{G}_l)$ 结束循环 输出：$\boldsymbol{\beta} = \boldsymbol{\beta}_{l_{\max}}$

假设 1。协变量向量 $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 相互独立，且服从高斯分布 $\mathbf{X}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$，其中$\boldsymbol{\Sigma}_i$均为对称正定矩阵。存在绝对常数 $C_l, C_u > 0$，使得

$$C_l \le \lambda_{\min}(\boldsymbol{\Sigma}_i) \le \lambda_{\max}(\boldsymbol{\Sigma}_i) \le C_u.$$

假设2。（重尾噪声 I）噪声 $\xi_1, \cdots, \xi_n$ 为独立同分布随机变量，其密度和分布函数分别为 $h_\xi(\cdot)$和 $H_\xi(\cdot)$。存在一个 $\varepsilon > 0$，使得 $\mathbb{E}|\xi|^{1+\varepsilon} < +\infty$。噪声的中位数为零，即 $H_\xi(0) = 1/2$。记作 $\gamma = \mathbb{E}|\xi|$。存在常数 $b_0, b_1 > 0$ （可能依赖于 γ），使得 [2]

$$\begin{aligned} &h_\xi(x) \ge b_0^{-1}, \quad \textit{for all } |x| \le 8(C_u/C_l)^{1/2}\gamma; \\ &h_\xi(x) \le b_1^{-1}, \quad \forall x \in \mathbb{R}. \end{aligned}$$

根据假设2，一个简单的事实是 $b_0 \ge 8(C_u/C_l)^{1/2}\gamma$ 且 $b_0 \ge b_1$。我们要求噪声具有有限的 $1+\varepsilon$ 阶矩，这对于保证 $\lim_{x\to+\infty} x(1 - H_\xi(x)) = 0$至关重要。与现有文献相比，这一矩条件相当宽松。例如，Minsker（2018）；Depersin（2020）要求有限的二阶矩条件；Fan 等人（2021）则要求 $2+\varepsilon$ 阶矩条件。密度函数的下界也见于 Elsener 和 van de Geer（2018），且是（局部）Bernstein 条件的一个特例（Alquier等，2019；Chinot 等人，2020）。对密度函数的上界条件较为宽松，例如，Lipschitz 连续的分布函数即可保证存在这样的统一上界。

定理 1. 假设1和假设2成立。存在绝对常数 $C_1, \ldots, C_6 > 0, c_2^* \in (0, 1)$，使得如果 $n \ge C_1 C_u C_l^{-1} \tilde{s}\log(2d/\tilde{s})$，$\tilde{s} \ge C_2(C_u/C_l)^8(b_0/b_1)^8 \cdot s$ 且初始步长$\eta_0 \in (nC_u)^{-1}C_l^{1/2}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \cdot [1/8,\ 3/8]$，则以至少 $1 - \exp(-C_3\tilde{s}\log(2d/\tilde{s})) - 3\exp(-(n\log(2d/\tilde{s}))^{1/2}\log^{-1} n)$ 的概率，算法 1 的序列 $\{\boldsymbol{\beta}_l\}_{l \le l_{max}}$ 具有以下动态:

[2] The lower bound condition can be slightly relaxed to $|H_\xi(x) - H_\xi(0)| \ge |x|/b_0$ when $|x| \le 8(C_u/C_l)^{1/2}\gamma$.

(1) 在第一阶段，即当 $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma$ 时，取步长 $\eta_l = (1-c_1)^l\eta_0$，其中 $c_1 \leq C_l C_u^{-1}/64$ 为任意固定常数，则 $\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 \leq (1-c_1)^{l+1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$;

（2）在第二阶段，即当 $C_4 b_0 C_l^{-1} n^{-1/2}\big(C_u \tilde{s}\log(2d/\tilde{s})\big)^{1/2} \leq \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \leq 8C_l^{-1/2}\gamma$ 时，通过采用常数步长 $\eta_l = \eta \in C_l^{1/2} b_1^2 (n b_0 C_u)^{-1}\cdot[C_5, C_6]$，则 $\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 \leq (1-c_2^*)\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2$

因此，至多经过 $O\big(\log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma) + \log(n\gamma b_0^{-1}\log(2d/\tilde{s}))\big)$ 次迭代后，算法 *1* 会输出一个误差率为

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_4 C_u}{C_l^2}\cdot\frac{b_0^2 \tilde{s}\log(2d/\tilde{s})}{n}.$$

<escape>注意，在定理 1 中，如果某个常数依赖于 b_0 或/和 b_1，则会在其右上角标注一个星号。第一阶段收敛是非光滑优化中的典型现象，需要采用递减的步长；另见 Foucart 和 Lecu´e (2017)。令人惊讶的是，第二阶段收敛要求使用恒定步长，就如同在最小化一个光滑目标函数一样（Blumensath 和 Davies，2009；Chen 和 Wainwright，2015）。</escape>

高斯噪声 如果噪声服从高斯分布 $N(0,\sigma^2)$，则有 $b_0 \asymp b_1 \asymp \sigma$。若进一步满足 $C_u \asymp C_l \asymp 1$，我们可以选择稀疏度 $\tilde{s} \gtrsim s$，使算法 1 的输出达到如下速率 $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p\big(\sigma^2 \cdot s n^{-1}\log(d/s)\big)$，该速率是极小极大最优的（Rigollet 和 Tsybakov，2011；Ye 和 Zhang，2010；Raskutti 等人，2011；Lounici 等人，2011）。

重尾噪声 在有限二阶矩条件下，Sun 等人（2020）证明，基于 Huber 损失的估计量可以达到误差速率 $O\big(s/n\cdot\log d\big)$，且其概率超过$1-d^{-\Omega(1)}$。另请参阅 Pan 等人（2021）；Wang 等人（2021，2020）的相关研究。Depersin（2020）研究了一种中位数均值法估计量，结果表明，在有限二阶矩条件下，可以以超过 $1-\Omega(1)\cdot\exp(-s\log(ed/s))$ 的概率达到速率 $O\big(s/n\cdot\log(d/s)\big)$。我们的定理 1 仅要求有限的 $1+\varepsilon$ 矩条件，但所达到的误差速率与 Depersin（2020）的结果相当。尽管我们的估计量并非次高斯型，但大多数现有的次高斯估计量都要求有限二阶矩条件（Catoni，2012；Lugosi 和 Mendelson，2019c,a；Hopkins，2020；Depersin 和 Lecu´e，2022）。Chinot 等人（2020）放宽了噪声的矩条件，但尚不清楚如何在提供理论保证的同时高效计算其估计量。更重要的是，与上述研究不同，我们的估计量由一种可证明快速收敛的算法得到。

定理的证明概要 *1*。 注意，次梯度 $\mathbf{G}_l$ 并非所有元素都会影响 $\boldsymbol{\beta}_{l+1}$。记

$$\Omega_{l+1} := \mathrm{supp}(\boldsymbol{\beta}_{l+1}),\quad \Omega_l := \mathrm{supp}(\boldsymbol{\beta}_l),\quad \Pi_l := \mathrm{supp}\big(\mathcal{H}_{\tilde{s}}(\mathcal{P}_{\Omega_l^c}(\mathbf{G}_l))\big),\quad \Omega^* := \mathrm{supp}(\boldsymbol{\beta}^*),$$

其中，算子 $\mathcal{P}_\Omega(\mathbf{G})$ 保留 Ω 中的 $\mathbf{G}$ 条目不变，并将其他条目置零；Ω^c 是 Ω 的正交补。显然，$\Omega_{l+1} \subseteq \Omega_l \cup \Pi_l \subseteq \Omega_l \cup \Pi_l \cup \Omega^*$。我们可以将更新规则改写为

$$\boldsymbol{\beta}_{l+1} = \mathcal{H}_{\tilde{s}}\left(\boldsymbol{\beta}_l - \eta_l \mathbf{G}_l\right) = \mathcal{H}_{\tilde{s}}\left(\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l)\right).$$

因此，只需关注稀疏向量 $\mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l)$。下面的引理描述了损失函数 $f(\boldsymbol{\beta})$ 在预言机 $\boldsymbol{\beta}^*$ 附近的两阶段正则性。

引理 1。 假设条件 *1* 和 *2* 成立。存在绝对常数 C_1, C_2, C_3, $c_0 > 0$使得若 $n \geq C_1 C_u C_l^{-1} \tilde{s} \log(2d/\tilde{s})$，则以概率至少为 $1 - \exp\left(-c_0 \tilde{s} \log\left(2d/\tilde{s}\right)\right) - 3\exp\left(-(n\log(2d/\tilde{s}))^{1/2} \log^{-1} n\right)$，有

(1) 对所有 $\boldsymbol{\beta} \in \left\{\boldsymbol{\beta} \in \mathbb{R}^d : \quad \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma,\ |\mathrm{supp}(\boldsymbol{\beta})| \leq \tilde{s}\right\}$，以及所有次梯度 $\mathbf{G} \in \partial f(\boldsymbol{\beta})$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \geq \frac{n}{4} C_l^{1/2} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \quad \text{and} \quad \|\mathcal{P}_{\Omega \cup \Pi \cup \Omega^*}(\mathbf{G})\|_2 \leq n C_u^{1/2},$$

where $\Omega := supp(\boldsymbol{\beta})$ *and* $\Pi := supp\big(\mathcal{H}_{\tilde{s}}(\mathcal{P}_{\Omega^c}(\mathbf{G}))\big)$;

(2) 对于所有 $\boldsymbol{\beta} \in \left\{\boldsymbol{\beta} \in \mathbb{R}^d : \quad C_2 C_u^{1/2} C_l^{-1} \cdot b_0\big(\tilde{s}/n \cdot \log(2d/\tilde{s})\big)^{1/2} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq 8C_l^{-1/2}\gamma、\ |\mathrm{supp}(\boldsymbol{\beta})| \leq \tilde{s}\right\}$，以及所有次梯度 $\mathbf{G} \in \partial f(\boldsymbol{\beta})$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \geq \frac{nC_l}{12b_0} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \quad \text{and} \quad \|\mathcal{P}_{\Omega \cup \Pi \cup \Omega^*}(\mathbf{G})\|_2 \leq C_3 \frac{nC_u}{b_1} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,$$

其中，$\Omega := \mathrm{supp}(\boldsymbol{\beta})$，且 $\Pi := \mathrm{supp}\big(\mathcal{H}_{\tilde{s}}(\mathcal{P}_{\Omega^c}(\mathbf{G}))\big)$。

Based on Lemma 1, we show that during phase one

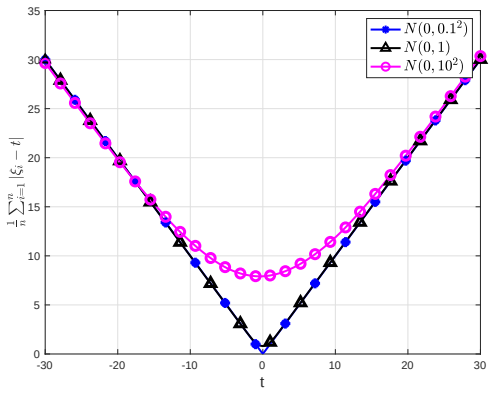
$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - \frac{n\eta_l C_l^{1/2}}{2} \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 + n^2 \eta_l^2 C_u.$$

所选步长保证 $\eta_l \asymp C_l^{1/2} C_u^{-1} \cdot n^{-1}(1 - c_1)^l \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$，由此利用以下事实推导出 $\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2$ 的收缩性 $\boldsymbol{\beta}_{l+1} = \mathcal{H}_{\tilde{s}}\left(\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l)\right)$。在第二阶段，根据引理1，我们得到

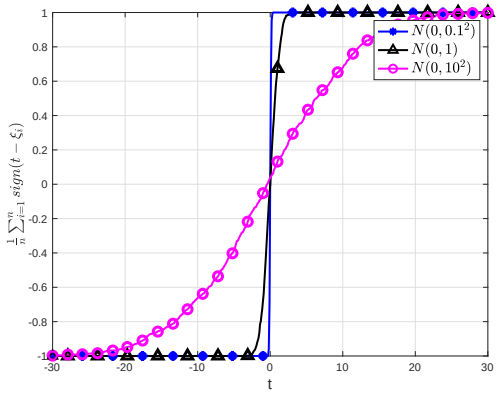
$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{\eta_l n C_l}{6b_0} + \frac{\eta_l^2 C_3^2 n^2 C_u^2}{b_1^2}\right) \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2.$$

Similarly, the selected constant stepsize guarantees the contraction of $\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2$. □

注释 1 （随机噪声的平滑效应）。根据定理 *1*，算法 *1* 的第二阶段收敛要求使用恒定步长，就像在最小化光滑函数时一样。引理 *1* 表明,



(a) Value of $\frac{1}{n}\sum_{i=1}^{n}|\xi_i - t|$ against t



(b) Value of $\frac{1}{n}\sum_{i=1}^{n}\text{sign}(t-\xi_i)$ against t

Figure 1: Sample size $n = 1000$ and $\{\xi_i\}_{i=1}^n$ follow $N(0, 0.1^2)$, $N(0, 1)$, $N(0, 10^2)$ respectively.

在发生第二阶段收敛的 $\boldsymbol{\beta}^*$ 附近的小区域内，损失函数 $f(\boldsymbol{\beta})$ 确实满足光滑优化中通常出现的正则性条件，例如平方损失（Cai 等人，2021a；Zhao 等人，2015）和逻辑损失（Lyu 等人，2021）。当个体随机噪声开始（随机地）主导 $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2$时，第二阶段便会出现。这表明，当 $\boldsymbol{\beta}^*$ 足够接近 $\boldsymbol{\beta}^*$时，随机噪声具有平滑目标函数的作用。图 1 展示了随机函数 $g(t) := n^{-1}\sum_{i=1}^{n}|\xi_i - t|$ 的形状及其次微分，其中 ξ_i's 为独立同分布的 $N(0, \tau^2)$。当 $\tau = 0$时，该函数显然并不光滑；随着 τ 增大，函数会变得更加光滑。在优化和统计学文献中，人们已观察到随机平滑现象。例如，Zhang 和 Xia（2020）证明，随机的边级观测误差可以自然地平滑网络矩统计量的 Edgeworth 展开；如果没有这种平滑效应，其经验累积分布函数可能是阶梯函数。为了以最优收敛速率优化非光滑凸函数 $h(\boldsymbol{\beta})$（例如 $\|\boldsymbol{\beta}\|_1$），Duchi 等人（2012）提出优化其随机平滑版本 $\mathbb{E}h(\boldsymbol{\beta} + Z)$，其中 Z 是一个分布已知的随机变量。

4 Low-rank Linear Regression

设 $\mathbf{M}^*$ 是一个秩为 $r \ll d_2$ 的矩阵。无需失去一般性，我们假设 $d_1 \geq d_2$。记作$\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ 满足 $Y_i = \langle \mathbf{X}_i, \mathbf{M}^* \rangle + \xi_i := \text{tr}(\mathbf{X}_i^\top \mathbf{M}^*) + \xi_i$ 的独立同分布观测集合，其中 $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ 被称为测量矩阵。我们的目标是通过求解估计 $\mathbf{M}^*$

优化程序

$$\hat{\mathbf{M}} := \underset{\mathbf{M}\in\mathbb{M}_r}{\arg\min}\ f(\mathbf{M}) \quad \text{where } f(\mathbf{M}) := \sum_{i=1}^{n} \rho\big(\langle \mathbf{M}, \mathbf{X}_i\rangle - Y_i\big), \tag{4}$$

其中 $\mathbb{M}_r := \{\mathbf{M} \in \mathbb{R}^{d_1\times d_2} : \text{秩}(\mathbf{M}) \le r\}$。这里 $\rho(\cdot)$ 是一种通用的稳健损失函数，包括绝对损失（Tong 等人，2021b；Charisopoulos 等人，2021）、Huber 损失（Elsener 和 van de Geer，2018；Sun 等人，2020）和分位数损失（Alquier 等人，2019；Chen 等人，2022）。

4.1 算法：黎曼优化

我们的算法是在黎曼流形上进行投影次梯度下降（RsGrad）。在第 l 次迭代中，给定当前的低秩估计值 $\mathbf{M}_l$，算法包含两个主要步骤。首先计算黎曼次梯度，即将普通次梯度 $\mathbf{G}_l \in \partial f(\mathbf{M}_l) \subset \mathbb{R}^{d_1\times d_2}$ 投影到 $\mathbb{M}_r$ 在点 $\mathbf{M}_l$处的切空间上，记为 $\mathbb{T}_l$。第二步是沿着负黎曼次梯度的方向更新低秩估计值，然后将其重新投影回流形 $\mathbb{M}_r$；对此，只需进行奇异值分解。这里，奇异值分解$_r(\cdot)$ 通过奇异值分解返回最佳秩 r 近似。详细内容见算法 2。

We note that the Riemannian sub-gradient is used only for the benefit of computational efficiency. The theoretical results all hold if taking the vanilla sub-gradient. The detailed computation of Riemannian sub-gradient can be found in the supplementary file. RsGrad has its origin in Cambier and Absil (2016) for minimizing the absolute loss without convergence and statistical analysis. Our framework covers general robust loss functions, and we prove its computational efficiency and statistical optimality in several important applications.

与现有算法的比较 Tong 等人（2021b）和 Charisopoulos 等人（2021）研究了用于带绝对损失的低秩线性回归的次梯度下降算法。我们的算法在两个方面有所不同：我们采用黎曼次梯度，并使用两阶段步长方案。基于因子分解的优化方法（Zheng 和 Lafferty，2016；Zhao 等人，2015）通过 $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ 对程序 (4) 重新参数化。随后，只需依次更新 (U, V)，即可通过次梯度下降算法最小化 (2)。若 $\mathbf{M}^*$ 条件不良，这些算法的计算效率可能会大幅下降。尽管适当的逆向缩放可以从理论上解决这一问题（Tong 等人，2021a），但当 $\mathbf{M}^*$ 存在较小的非零奇异值时，可能会导致计算不稳定。最近的研究发现，直接在低秩流形 $\mathbb{M}_r$ 上优化 (2)，即采用黎曼方法，既能享有基于因子分解方法的快速计算优势，同时又能在线性收敛，且不受 $\mathbf{M}^*$ 条件数的影响。例如，Cai 等人（2021a）和 Cai 等人（2021b）证明了

最小化时的黎曼梯度下降（RGrad）算法强凸且光滑的函数及其张量相关应用。我们指出，RGrad 与投影梯度下降（PGD，Chen 和 Wainwright (2015)）相似，不同之处在于，RGrad 使用黎曼梯度，而 PGD 使用普通梯度。

Algorithm 2 Riemannian Sub-gradient Descent (RsGrad)

Input: observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, max iterations $l_{\max}$, step sizes $\{\eta_l\}_{l=0}^{l_{\max}}$.

Initialization: $\mathbf{M}_0 \in \mathbb{M}_r$

for $l = 0, \ldots, l_{\max}$ **do**

选择一个普通次梯度：$\mathbf{G}_l \in \partial f(\mathbf{M}_l)$

Compute the Riemannian sub-gradient: $\widetilde{\mathbf{G}}_l = \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)$

Retraction to $\mathbb{M}_r$: $\mathbf{M}_{l+1} = \mathrm{SVD}_r(\mathbf{M}_l - \eta_l \widetilde{\mathbf{G}}_l)$

end for

Output: $\hat{\mathbf{M}} = \mathbf{M}_{l_{\max}}$

4.2 一般收敛性能

现在，我们来介绍 RsGrad 算法 2的一般收敛性能，这主要取决于目标函数的正则性条件。从本质上说，这些条件大体沿袭了现有文献中的相关条件（Charisopoulos 等人，2021；Elsener 和 van de Geer，2018；Alquier 等人，2019；Tong 等人，2021a）。然而，正如第 1 节所述，这些先前的工作只能得到统计上次优的估计。事实证明，为了实现我们的目标，必须对这些条件进行更精细的刻画。具体而言，我们发现，上述稳健函数在 $\mathbf{M}^*$附近与远离 $\mathbf{M}^*$时呈现出截然不同的正则性条件，称为 两阶段正则性条件。这里 $\|\cdot\|_{\mathrm{F}}$ 表示矩阵的弗罗贝尼乌斯范数。

条件 1.（两阶段正则性条件）定义 $\mathbf{M}^*$ 附近的两个区域，其中 $\tau_{comp} > \tau_{stat} > 0$ ，如下：

$$\mathbb{B}_1 := \{\mathbf{M} \in \mathbb{M}_r : \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{comp}\} \quad \text{and} \quad \mathbb{B}_2 := \{\mathbf{M} \in \mathbb{M}_r : \tau_{comp} > \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{stat}\}.$$

The function $f(\cdot) : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}_+$ *satisfies*

(1) 关于 $\mathbf{M}^*$ 的秩-r 限制（τ_{comp}、τ_{stat}、μ_{comp}、μ_{stat}）两阶段尖锐性，

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \mu_{comp}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ \mu_{stat}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, & for \quad \mathbf{M} \in \mathbb{B}_2; \end{cases}$$

Loss function	Noise	τ_{comp}	τ_{stat}	μ_{comp}	L_{comp}	μ_{stat}	L_{stat}
ℓ_1-Loss	$N(0,\sigma^2)$	σ	$O\left(\sigma\left(\frac{rd_1}{n}\right)^{1/2}\right)$	$\frac{n}{12}$	$2n$	$\frac{n}{12\sigma}$	$O\left(\frac{n}{\sigma}\right)$
ℓ_1-Loss	Heavy-tailed	8γ	$O\left(b_0\left(\frac{rd_1}{n}\right)^{1/2}\right)$	$\frac{n}{4}$	$2n$	$\frac{n}{12b_0}$	$O\left(\frac{n}{b_1}\right)$
Huber $\rho_{H,\delta}$	Heavy-tailed	$8\gamma+2\delta$	$O\left(b_0\left(\frac{rd_1}{n}\right)^{1/2}\right)$	$\frac{\delta n}{2}$	$4\delta n$	$\frac{\delta n}{3b_0}$	$O\left(\frac{\delta n}{b_1}\right)$
Quantile $\rho_{Q,\delta}$	Heavy-tailed	8γ	$O\left(b_0\left(\frac{rd_1}{n}\right)^{1/2}\right)$	$\frac{n}{8}$	n	$\frac{n}{24b_0}$	$O\left(\frac{n}{b_1}\right)$

表 1：两阶段正则性条件总结。事实上， τ_{comp} 是第一阶段迭代后达到的误差率，而 τ_{stat} 是第二阶段迭代后达到的误差率，即最终的统计精度。这里 γ、 b_0 和 b_1 如假设2中所述。注意，在许多情况下，例如高斯噪声、零对称帕累托噪声和学生 t 噪声等， $b_0 \asymp \gamma$。将 b_0/γ 视为常数，当样本量较大时，有 $\tau_{\text{stat}} \ll \tau_{\text{comp}}$ 。

(2) the rank-r restricted $(\tau_{comp}, \tau_{stat}, L_{comp}, L_{stat})$ **two-phase sub-gradient bound** *with respect to* $\mathbf{M}^*$,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} L_{comp}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ L_{stat}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_2, \end{cases}$$

where $\mathbf{G} \in \partial f(\mathbf{M})$ *and the truncated Frobenius norm* $\|\mathbf{G}\|_{\mathrm{F,r}} := \|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}$.

基本上，条件 1 规定了 $f(\cdot)$ 在 $\mathbf{M}^*$ 的两个邻域中的不同表现。 τ_{stat} 反映了统计极限，而在不假设噪声分布的情况下， τ_{comp} 通常是现有文献（Tong 等人，2021a；Charisopoulos等，2021）中计算分析所达到的速率。这些先前的工作仅揭示了第一阶段正则性条件，即关于$\mathbb{B}_1$的条件。通常，我们有 $\tau_{\text{comp}} \gg \tau_{\text{stat}}$，在这种情况下，这些先前的工作只能得到统计上次优的估计。推导第二阶段正则性条件颇具挑战性，需要进行更为精确的计算。我们指出，在统计学文献中，次梯度界与 Lipschitz 连续性相关（Alquier等，2019），而尖锐性条件被称为单点间隔条件（Elsener 和 van de Geer，2018）。然而，他们的估计量建立在凸规划之上，因此分析仅在 $\mathbf{M}^*$ 的一个小邻域内进行，在该邻域中，两阶段正则性条件并不关键。我们在表 1 中总结了各种应用中的两阶段正则性条件。命题 1 刻画了算法 2 的一般动态。

命题 1. 假设条件 1 成立，初始化 $\mathbf{M}_0 \in \mathbb{M}_r$ 满足 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq c_0\sigma_r(\mathbf{M}^*) \cdot \min\{\mu^2 \text{stat} L^{-2} \text{stat}\, \mu^2 \text{comp} L^{-2} \text{comp}\}$，其中 $c_0 > 0$ 是一个较小但为绝对常数的值，且初始步长$\eta_0 \in \big[0.2\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}\mu_{comp}L^{-2}\,\text{comp} \quad 0.3\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}\mu_{comp}L^{-2}\,\text{comp}\big]$。在算法 2 的第 l 次迭代时，

(1) 当 $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{comp}$ 时，即在第一阶段，取步长 $\eta_l = \left(1 - c\mu^2{}_{comp} L^{-2}{}_{comp}\right)^l \eta_0$，其中 $c \leq 0.04$ 为任意固定常数，则我们有

$$\|\mathbf{M}_{l+1} - \mathbf{M}^*\|_{\mathrm{F}} \leq \left(1 - c\frac{\mu_{comp}^2}{L_{comp}^2}\right)^{l+1} \cdot \|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}};$$

(2) 当 $\tau_{comp} > \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{stat}$ 时，即在第二阶段，取步长 $\eta_l = \eta \in \mu_{stat} L^{-2}{}_{\text{统计}}\ [0.125,\ 0.75]$，则有

$$\|\mathbf{M}_{l+1} - \mathbf{M}^*\|_{\mathrm{F}} \leq \left(1 - \frac{\mu_{stat}^2}{32 L_{stat}^2}\right) \cdot \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}.$$

根据命题 1，RsGrad 算法 2 使用所提出的步长线性收敛，并输出一个与预言机的弗罗贝尼乌斯范数距离在 $O(\tau_{\text{stat}})$ 以内的估计值。如上所述，第一阶段收敛是非光滑优化中的典型现象，因此需要使用衰减步长。为简单起见，我们采用几何衰减步长（Tong 等人，2021a；Goffin，1977）。第二阶段收敛的表现类似于光滑优化，因此固定步长足以实现线性收敛。请注意，收敛动态与矩阵条件数无关，这是黎曼型算法的一项优势。

与先前工作的比较 RsGrad 在第一阶段的算法动态与 Charisopoulos 等人 (2021) 和 Tong 等人 (2021b) 中的类似，在第一阶段采用几何衰减步长，最终会使步长降至 $\Omega(\mu_{\text{stat}} L^{-2}{}_{\text{stat}})$ 的水平，这是第二阶段收敛所需的。然而，Charisopoulos 等人 (2021) 和 Tong 等人 (2021b) 会继续按几何方式缩小步长，因此无法得到统计最优的估计量。相比之下，我们在第二阶段设置恒定步长，最终得到的估计量在统计意义上是最优的。这一区别可在第 7 节的数值实验中观察到。

4.3 Applications

本节通过一些具体示例展示 RsGrad 的统计性能。由于篇幅限制，部分示例被放在补充文件中。我们注意到，与平方损失（Xia 等人，2021；Zhao 等人，2015；Chen 和 Wainwright，2015）相比，采用稳健损失函数的低秩回归无需满足信号与噪声比（SNR）条件（初始化阶段除外）。

4.3.1 Absolute loss with Gaussian noise

尽管我们的研究动机源于重尾噪声，首先，我们证明，在高斯噪声下，绝对损失也能够得到统计最优的估计量。目标函数为 $f(\mathbf{M}) =$

$\sum_{i=1}^{n} |Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle|$。

定理 2. 假设 $\xi_1, \ldots, \xi_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ 且 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 满足假设 1。存在绝对常数 $c_0, c_1, c_2 \in (0, 1)$ $C_1, C_2 > 0$，使得若样本量 $n \geq C_1 C_u C_l^{-1} d_1 r$、初始化 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq c_0 \sigma_r$（对于 $\mathbf{M}_0 \in \mathbb{M}_r$）、初始步长 $\eta_0 \in C_l^{1/2} (48 n C_u)^{-1} \|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \cdot [0.2, 0.3]$、第一阶段步长 $\eta_l = (1 - c_3)^l \eta_0$（其中任意固定常数 $c_3 \leq c_1 C_l C_u^{-1}$），以及第二阶段步长 $\eta_l = \eta \asymp C_l C_u^{-2} \sigma / n$，则以至少 $1 - \exp(-c_2 r d_1) - 3\exp(-\sqrt{n}/\log n)$ 的概率，在至多 $O\big(\log(\sigma_r/\sigma) + \log(n/d_1)\big)$ 次迭代后，算法 2 输出一个误差率为

$$\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}}^2 \leq C_2 \frac{C_u}{C_l^2} \cdot \frac{\sigma^2 d_1 r}{n}.$$

注意，初始步长 η_0 可以大到 $O(\sigma_r n^{-1})$，具体取决于初始化。所达到的误差率是极小极大最优的（Ma 和 Wu，2015；Xia，2014）。相比之下，Tong 等人(2021b)；Charisopoulos 等人(2021提出的算法及其分析仅达到了次优速率 $O_p(\sigma^2)$。定理2 的证明结合了命题1与 $f(\mathbf{M}$的两阶段正则性条件。第7节的数值实验展示了两阶段收敛动力学。

4.3.2 重尾噪声下的绝对损失

现在，我们展示绝对损失在处理重尾噪声时的有效性。请注意，在定理3 中，若某个常数因子依赖于 b_0 和/或 b_1，则会在该因子的右上方标注星号。

定理 3。假设 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 和 $\{\xi_i\}_{i=1}^n$ 分别满足假设1和假设2。存在常数 $c_0^*, c_1, c_2 \in (0, 1)$ $C_1, C_2 > 0$，使得若样本量 $n \geq C_1 C_u C_l^{-1} d_1 r$、初始化 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq c_0^* \sigma_r$ 对于 $\mathbf{M}_0 \in \mathbb{M}_r$、初始步长 $\eta_0 \in C_l^{1/2} (16 n C_u)^{-1} \|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \cdot [0.2, 0.3]$、第一阶段步长 $\eta_l = (1 - c_3)^l \eta_0$（取任意固定常数 $c_3 \leq c_1$）以及第二阶段步长 $\eta_l = \eta \asymp C_l C_u^{-2} b_1^2 (n b_0)^{-1}$满足相应条件，则以至少 $1 - \exp(-c_2 r d_1) - 3\exp(-\sqrt{n}/\log n)$ 的概率，在至多经过 $O\big(\log(\sigma_r/\gamma) + \log\left(\gamma n/(d_1 r b_0)\right)\big)$ 次迭代后，算法 2 输出一个误差率为下式的估计量

$$\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}}^2 \leq \frac{C_2 C_u}{C_l^2} \cdot \frac{b_0^2 d_1 r}{n}.$$

定理 3 的证明可由命题 1 直接推出，前提是我们在重尾噪声下满足 $f(\mathbf{M})$ 的两阶段正则性条件。为此，我们借助了一个强大的工具来界定经验过程（Adamczak，2008；Van de Geer 和 van de Geer，2000）。这使我们能够通过避免界定项 $\sum_{i=1}^n |\xi_i|$，推导出更强的尾概率界。实际上，由于 ξ具有有限的 $1+\varepsilon$矩，对 $|\xi_i|$求和进行界定会得到尾概率 $1 - O(n^{-\varepsilon})$。

对于较小的 ε，这样的尾概率并不理想。注意，高斯噪声也满足假设 2，在这种情况下，定理 3 中得到的速率具有相匹配的极小极大下界。该速率关于维度和样本量是极小极大最优的。注意，在许多情况下，例如高斯噪声、零对称帕累托噪声和学生 t 噪声， $b_0 \asymp \gamma = \mathbb{E}|\xi|$

假设2 中的密度要求与Elsener 和 van de Geer (2018) 中的要求类似，其估计量基于核范数惩罚凸规划，但未考虑计算效率。他们也推导出了相同的误差率 $O(b_0^2 d_1 r/n)$。而我们的方法计算速度快，且只需经过对数级别的迭代次数。若存在重尾噪声，Tong 等人 (2021b); Charisopoulos 等人 (2021) 的分析框架会得到次优误差率 $O(\gamma^2$，且该误差率不会随着样本量 n 的增大而降低。由Fan 等人 (2021); Minsker (2015) 和 Minsker (2018) 推导出的误差率与噪声水平不成正比；即使观测值没有噪声，误差率也不为零。相比之下，在没有噪声时，我们的算法可以精确恢复 $\mathbf{M}^*$ 。

4.3.3 重尾噪声下的 Huber 损失

Huber 损失在稳健统计中广泛应用（Huber，1965；Sun 等，2020；Elsener 和 van de Geer，2018），其定义为 $\rho_{H,\,\delta}(x) := x^2 \mathbb{I}(|x| \le \delta) + (2\delta|x| - \delta^2)\mathbb{I}(|x| > \delta)$，其中 $\delta > 0$ 称为稳健化参数。函数 $\rho_{H,\,\delta}(\cdot)$ 是 Lipschitz 连续的，其常数为 2δ。损失函数定义为 $f(\mathbf{M}) := \sum_{i=1}^n \rho_{H,\,\delta}(Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle)$。由于技术上的原因，我们需要对重尾噪声采用一个略有不同的假设。

假设 3。（重尾噪声 II）存在一个 $\varepsilon > 0$，使得 $\mathbb{E}|\xi|^{1+\varepsilon} < +\infty$。噪声分布函数 $H_\xi(\cdot)$ 是对称的[3]，即 $H_\xi(x) = 1 - H_\xi(-x)$。记为 $\gamma = \mathbb{E}|\xi|$。存在常数 b_0, b_1 （可能依赖于 γ 和 δ），使得

$$
\begin{aligned}
&H_\xi(x+\delta) - H_\xi(x-\delta) \ge 2\delta b_0^{-1}, \quad \textit{for all } |x| \le 8(C_u/C_l)^{1/2}\gamma + 2(C_u/C_l)^{1/2}\delta;\\
&H_\xi(x+\delta) - H_\xi(x-\delta) \le 2\delta b_1^{-1}, \quad \forall x \in \mathbb{R},
\end{aligned}
$$

where δ is the Huber loss parameter.

Theorem 4 describes the convergence and statistical performance of RsGrad Algorithm 2 for Huber loss. Similarly, the constants dependent on b_0 or/and b_1 are marked with a star.

定理 4。假设 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 和 $\{\xi_i\}_{i=1}^n$ 分别满足假设 1 和 3。存在常数 $c_0^*, c_1, c_2 \in (0,1)$ C_1，$C_2 > 0$，使得若样本量 $n \ge C_1 C_u C_l^{-1} d_1 r$、初始化 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \le c_0^* \sigma_r$ （对于 $\mathbf{M}_0 \in \mathbb{M}_r$）、初始步长 $\eta_0 \in C_l^{1/2}(32n\delta C_u)^{-1}\|\mathbf{M}_0 -$

[3]对称条件也可以替换为 $\int_{-\delta}^{\delta} H_\xi(x)\,dx = \delta$。

$\mathbf{M}^*\|_{\mathrm{F}} \cdot [0.2, 0.3]$、第一阶段步长 $\eta_l = (1-c_3)^l \eta_0$（其中任意固定常数为 $c_3 \le c_1$）以及第二阶段步长 $\eta_l = \eta \asymp C_l C_u^{-2} b_1^2 (n\delta b_0)^{-1}$，则以至少 $1 - \exp(-c_2 r d_1) - 3\exp(-\sqrt{n}/\log n)$ $O\big(\log(\sigma_r/\gamma) + \log\big(\gamma n/(b_0 d_1 r)\big)\big)$ 次迭代后，算法 2 以至少 $1 -$ 的概率输出一个误差率为

$$\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}}^2 \le \frac{C_2 C_u}{C_l^2} \cdot \frac{b_0^2 d_1 r}{n}.$$

注意，假设 3 意味着 b_0 的下界为 $\Omega(\gamma + \delta)$，因此误差率的下界为 $\Omega\big((\delta^2 + \gamma^2) r d_1/n\big)$。这表明，稳健化参数会带来一个有趣的转变：当 $\delta \lesssim \gamma$时，增大 δ 不会影响误差率；当 $\delta \gtrsim \gamma$时，增大 δ 会使误差率改善。 Sun 等人 (2020) 在未对噪声密度施加正则性条件的情况下，推导了 Huber 损失在线性回归中的统计性能，得到的误差率比 $O(n^{-1})$ 慢。

4.3.4 重尾噪声下的分位数损失

分位数损失最初由 Koenker 和 Bassett Jr (1978) 提出，并一直是稳健统计中常用的损失函数之一（Welsh，1989；Koenker 和 Hallock，2001；Wang 等人，2012；Tan 等人，2022）。用 $\rho_{Q,\delta}$ 表示分位数损失：对于任意满足 $\delta := \mathbb{P}(\xi \le 0)$ 的 $x \in \mathbb{R}$，$= \delta x \mathbb{I}(x \ge 0) + (\delta - 1) x \mathbb{I}(x < 0)$。注意，函数 $\rho_{Q,\delta}(\cdot)$ 是 Lipschitz 连续的，其常数为 $\max\{\delta, 1-\delta\}$。损失函数定义为 $f(\mathbf{M}) := \sum_{i=1}^n \rho_{Q,\delta}(Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle)$。与 Huber 损失的情形类似，出于技术原因，我们的分析框架对重尾噪声需要采用略有不同的假设。

假设 4。（重尾噪声 III）存在一个 $\varepsilon > 0$，使得 $\mathbb{E}|\xi|^{1+\varepsilon} < +\infty$。假设$\delta := H_\xi(0)$ 属于 $(0, 1)$。记为 $\gamma = \mathbb{E}|\xi|$。存在常数 b_0、$b_1 > 0$（可能依赖于γ），使得

$$h_\xi(x) \ge b_0^{-1}, \quad \textit{for all } |x| \le 8(C_u/C_l)^{1/2}\gamma;$$
$$h_\xi(x) \le b_1^{-1}, \quad \forall x \in \mathbb{R}.$$

当噪声的中位数为零，即 $\delta = 1/2$ 时，假设 4 与假设 2 完全相同。

定理 5。假设 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 和 $\{\xi_i\}_{i=1}^n$ 分别满足假设 1 和假设 4。存在常数 c_0^*、c_1、$c_2 \in (0, 1)$ C_1、$C_2 > 0$，使得当样本量 $n \ge C_1 C_u C_l^{-1} d_1 r$、初始化 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \le c_0^* \sigma_r$、初始步长 $\eta_0 \in C_l^{1/2} (8nC_u)^{-1} \|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \cdot [0.2, 0.3]$、第一阶段步长 $\eta_l = (1-c_3)^l \eta_0$（其中任意固定常数 $c_3 \le c_1$），以及第二阶段步长

$\eta_l = \eta \asymp C_l C_u^{-2} b_1^2 \left(nb_0\right)^{-1}$，则以关于 $1-\exp(-c_2 r d_1) - 3\exp(-\sqrt{n}/\log n)$的概率，在至多 $O\big(\log(\sigma_r/\gamma) + \log\left(\max\{\delta^{-1}, (1-\delta)^{-1}\}\cdot \gamma n/\left(d_1 r b_0\right)\right)\big)$ 次迭代后，算法 2输出一个误差率为

$$\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}}^2 \leq \frac{C_2 C_u \cdot \max\{\delta^2, (1-\delta)^2\}}{C_l^2} \cdot \frac{b_0^2 r d_1}{n}.$$

5 初始化与算法参数选择

初始化 稀疏线性回归算法 1 允许任意初始化。我们仅讨论低秩线性回归的初始化。算法 2 的收敛性关键取决于热启动 $\mathbf{M}_0$。可直接采用收缩低秩近似（Fan 等人，2021）进行初始化，但要求噪声满足有限 $2+\varepsilon$ 矩条件。为简单起见，我们研究标准低秩近似的性能。这里 $vec(\cdot)$ 和$mat(\cdot)$ 分别表示向量化和矩阵化算子。

定理 6.（初始化保证）假设 $\gamma_1 := \mathbb{E}|\xi|^{1+\varepsilon} < +\infty$ 对于某个 $\varepsilon \in (0,1]$ 和$\{vec(\mathbf{X}_i)\}_{i=1}^n$ 满足假设 1 ，且协方差 $\{\mathbf{\Sigma}_i\}_{i=1}^n$已知。按如下方式初始化：

$$\mathbf{M}_0 := \mathrm{SVD}_r(n^{-1}\sum_{i=1}^n mat(\mathbf{\Sigma}_i^{-1} vec(\mathbf{X}_i))Y_i).$$

对于任意小的 $c_0 > 0$，存在仅依赖于 c_0 的常数 C、c_1、c_2、$c_3 > 0$ ，使得若 $n > C\max\left\{C_u C_l^{-1}\kappa^2 d_1 r^2 \log d_1, (C_l^{-1} d_1 r)^{(1+\varepsilon)/(2\varepsilon)} \sigma_r^{-(1+\varepsilon)/\varepsilon} (\gamma_1 \log d_1)^{1/\varepsilon}\right\}$，则以 $1 - c_1 d_1^{-1} - c_2 \log^{-1} d_1 - c_3 e^{-d_1}$ 为概率，初始化满足 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq c_0 \sigma_r$。这里$\kappa := \sigma_1(\mathbf{M}^*)\sigma_r^{-1}(\mathbf{M}^*)$ 是条件数。

与 Fan 等（2021）相比，定理 6 只要求有限的 $1+\varepsilon$ 矩条件。我们注意到，如果噪声项 ξ_i 具有有限的 $2+\varepsilon$ 矩，概率项 $O(\log^{-1} d_1)$ 可以替换为 $O(n^{-\min\{1,\varepsilon/2\}})$。此外，在有限的 $2+\varepsilon$ 矩条件下，还可以通过中位数均值法（Minsker，2015；Lugosi 和 Mendelson，2019c,b）或 Catoni 估计量（Catoni，2012；Minsker，2018）获得次高斯初始估计量。由于基于收缩的方法（Fan 等，2021；Minsker，2018）已能有效提供热启动，我们不再额外费力改进定理 6。如果协方差 $\mathbf{\Sigma}_i$ 未知，我们建议采用 Fan 等（2021）或 Minsker（2018）的方法进行初始化。

初始步长的选择 尽管第 3节和 4节中的主要定理要求初始步长 $\eta_0 \asymp n^{-1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$ 和 $\eta_0 \asymp n^{-1}\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}$分别满足，但以下定理表明，即使我们只有 $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$ 和$\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}$的上界，也能得到类似的结果。

定理 1*（定理 1 的另一种表述）。假设假设1和假设2成立。存在绝对常数 $C_1, \ldots, C_6 > 0$, $c_2^* \in (0,1)$，使得如果 $n \geq C_1 C_u C_l^{-1} \tilde{s} \log(2d/\tilde{s})$，$\tilde{s} \geq C_2 (C_u/C_l)^8 (b_0/b_1)^8 \cdot s$，初始步长 $\eta_0 \in (nC_u)^{-1} C_l^{1/2} D_0 \cdot [1/8,\ 3/8]$ 对任意 $D_0 \geq \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$，第一阶段步长 $\eta_l = (1-c_1)^l \eta_0$ 对任意 $c_1 \leq C_l C_u^{-1}/64$，以及第二阶段步长 $\eta_l = \eta \in C_l^{1/2} b_1^2 (n b_0 C_u)^{-1} [C_3,\ C_4]$，那么以超过 $1 - \exp\big(-C_5 \tilde{s} \log(2d/\tilde{s})\big) - 3\big(-(n \log(2d/\tilde{s})^{1/2} \log^{-1} n\big)$ 的概率，在至多经过 $O\big(\log(D_0/\gamma) + \log(n\gamma b_0^{-1} \log(2d/\tilde{s}))\big)$ 次迭代后，算法 1 输出一个具有如下误差率的估计量。

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_6 C_u}{C_l^2} \cdot \frac{b_0^2 \tilde{s} \log(2d/\tilde{s})}{n}.$$

定理 3*（定理 3 的替代形式）。假设 $\{\xi_i\}_{i=1}^n$ 和 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 分别满足假设 1 和假设 2。存在常数 c_0^*、c_1、$c_2 \in (0,1)$ C_1、$C_2 > 0$，使得如果$n \geq C_1 C_u C_l^{-1} d_1 r$，则初始化 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq D_0$ 对任意 $D_0 \leq c_0^* \sigma_r$，初始步长$\eta_0 \in C_l^{1/2} (16 n C_u)^{-1} D_0 \cdot [0.2, 0.3]$，第一阶段步长 $\eta_l = (1-c_3)^l \eta_0$（其中任意固定常数 $c_3 \leq c_1$），以及第二阶段步长 $\eta_l = \eta \asymp C_l C_u^{-2} b_1^2 (n b_0)^{-1}$，则以至少$1 - \exp(-c_2 r d_1) - 3\exp(-n^{1/2}/\log n)$ 的概率，在至多 $O\big(\log(D_0/\gamma) + \log\big(\gamma n/(d_1 r b_0)\big)\big)$ 次迭代后，算法 2 输出一个误差率为的估计量

$$\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}}^2 \leq \frac{C_2 C_u}{C_l^2} \cdot \frac{b_0^2 d_1 r}{n}.$$

By Theorem 3*, it thus suffices to obtain an estimate of $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}$. Towards that end, one appropriate method is to take the average $n^{-1} \sum_{i=1}^n |Y_i - \langle \mathbf{M}_0, \mathbf{X}_i \rangle|$, which, under mild conditions, is of the same order of $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}}$ with high probability. The same approach is also applicable in sparse vector regression. The general quantities μ_{comp} and L_{comp} involved in Algorithm 2 are often $\asymp n$ up to a factor of loss-related parameters, which can be easily tuned in practice.

RsGrad 中的调优参数 我们的算法假设已知 $\mathbf{M}^*$ 的秩，这在非凸方法中很常见（Tong 等人，2021a；Chen 和 Wainwright，2015；Cai 等人，2021a）。可以使用 BIC 或 AIC 来选择秩（Cai 等人，2021a）。尽管我们认为，对于算法秩 r 大于 $\mathbf{M}^*$ 的真实秩的过参数化情形（Zhuo 等人，2021；Ma 和 Fattahi，2022；Luo 和 Zhang，2022），我们的算法仍然有效，但对此进行全面的理论研究超出了本文的范围。Elsener 和 van de Geer（2018）以及 Belloni 等人（2011）等凸方法无需初始化，也不需要预先知道秩。但我们主要关注的是稳健估计中的计算与统计问题。我们的算法（即第二阶段步长）还需要噪声水平 γ，而基于平方根 LASSO 或核范数惩罚绝对损失的凸方法则不需要该信息（Elsener 和 van de Geer，2018；Belloni 等人，2011；Minsker 等人，2022）。

幸运的是，我们的方法能够可靠地估计 γ。第一阶段迭代持续进行，直至$\sum_{i=1}^{n}|Y_i - \langle \mathbf{M}^{l_1}, \mathbf{X}_i\rangle| \asymp \sum_{i=1}^{n}|\xi_i| \asymp n\gamma$，其中 l_1 是第一阶段的迭代次数。随后，第二阶段步长可取为 $\eta = c \cdot f(\mathbf{M}^{l_1})/n^2$，其中 $c > 0$ 是某个调优参数。在模拟实验中，我们发现该算法对 c 的选择非常稳健。

确定阶段由于第一阶段的步长呈几何级数衰减，经过若干次迭代后，步长将进入 $O(\mu_{\text{stat}}L^{-2}{}_{\text{stat}})$所需的水平，从而满足第二阶段收敛的要求。另一方面，如果继续进行第一阶段迭代，步长会迅速减小，目标函数的值也会趋于稳定。数值实验中确实观察到了这一现象。参见第3节的图4。因此，一旦步长低于预先选定的较小阈值，就可以启动第二阶段迭代，并在此后固定步长，直至收敛。

6 Huber 污染模型

针对异常值的稳健性通常在 Huber ε污染模型（Huber 和 Ron chetti, 2009；Chen 等人，2016，2018；Diakonikolas 和 Kane，2019）中进行研究。该模型假设数据是独立同分布的，且服从共同分布 $dP = (1-\epsilon)\,dP_i + \epsilon\,dP_o$其中，$P_i$表示内点的分布，$P_o$是异常值的任意分布，$\epsilon \in [0,1)$。现在我们展示算法对异常值的稳健性。不失一般性，在本节中，我们假设响应值中有ϵ比例的 $\{Y_i\}_{i=1}^n$ 受到异常值污染。将异常值的索引记为 $\mathcal{O}$，内点的索引记为 $\mathcal{I}$，其中 $\mathcal{O}\cup\mathcal{I} = \{1,\ldots,n\}$且 $|\mathcal{O}| = \lceil \epsilon n \rceil$。基本上，$(\mathbf{X}_i,\ Y_i)$ 对于 $i \in \mathcal{I}$满足线性模型 (1)。对于异常值 $i \in \mathcal{O}$，响应值 Y_i 受到任意污染。注意，我们仍然假设所有协变量 $\mathbf{X}_i$ 或 $vec(\mathbf{X}_i)$ 满足假设 1。感兴趣的读者可参阅 Depersin (2020)、Lecu´e 和 Lerasle (2019)、Chinot 等人 (2020)、Thompson (2020) 以及 Diakonikolas 等人 (2019)，了解更一般的 Huber ϵ污染模型设定。受篇幅限制，我们仅考虑绝对损失。

6.1 Sparse linear regression

Using the same notations from Section 3, the following theorem describes the statistical guarantees of the IHT-ℓ_1 algorithm under Huber's ϵ-contamination model.

定理 7。假设假设 1 成立，且 $\{\xi_i\}_{i\in\mathcal{I}}$ 满足假设2。存在绝对常数 $C_1,\ldots,C_6 > 0, c_2^* \in (0,1)$，使得若 $n \geq C_1C_uC_l^{-1}\tilde{s}\log(2d/\tilde{s})$，$\tilde{s} \geq C_2(C_u/C_l)^8(b_0/b_1)^8\cdot s$，污染率 $\epsilon \leq C_l^{1/2}(4C_l^{1/2} + C_u^{1/2})^{-1}$，初始步长 $\eta_0 \in (nC_u)^{-1}C_l^{1/2}\cdot\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2\cdot[1/8,\ 3/8]$，第一阶段步长 $\eta_l = (1-c_1)^l\eta_0$（其中 $c_1 \leq C_lC_u^{-1}/64$ 为任意固定常数），且第二阶段步长 $\eta_l = \eta \in C_l^{1/2}b_1^2(nb_0C_u)^{-1}\cdot[C_5, C_6]$，则以关于 $1-$的概率

$2\exp(-C_3\tilde{s}\log(2d/\tilde{s}))\ -\ 3\exp(-(n\log(2d/\tilde{s}))^{1/2}\log^{-1}n)$，至多经过 $O\big(\log(\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|_2/\gamma)\ +\ \log(n\gamma b_0^{-1}\log(2d/\tilde{s}))\wedge\log(\gamma/b_0\epsilon)\big)$ 次迭代，算法 *1* 输出一个误差率为

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\le\frac{C_4C_u^{1/2}}{C_l}\cdot b_0\cdot\max\left\{\left(\frac{\tilde{s}\log(2d/\tilde{s})}{n}\right)^{1/2},\epsilon\right\}.$$

我们的定理允许 ϵ 取任意小于 $C_l^{1/2}\ (4C_l^{1/2}+C_u^{1/2})^{-1}$的值。当内点的噪声服从高斯分布时，误差率与极小极大下界相匹配（Chen 等人，2016，2018）。Chinot 等人 (2020)；Depersin (2020) 推导出的收敛率包含一项 $O(\epsilon^{1/2})$，而Thompson (2020) 给出的收敛率包含一项 $O(\epsilon\log(1/\epsilon))$。

6.2 低秩线性回归

沿用第 4 节中的记号，以下定理描述了 RsGrad 算法在 Huber ϵ-污染模型下的统计保证。

定理 8. 假设 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 和 $\{\xi_i\}_{i\in\mathcal{I}}$ 分别满足假设1和假设2。存在常数 $c_0^*,c_1,c_2^*\in(0,1)$ $C_1^*,C_2,C_3>0$，使得若样本量 $n\ge C_3C_uC_l^{-1}d_1r$、污染率 $\epsilon\le C_l^{1/2}\ (4C_l^{1/2}+C_u^{1/2})^{-1}$、初始化 $\|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}}\le c_0^*\sigma_r$、初始步长 $\eta_0\in C_l^{1/2}(16nC_u)^{-1}\|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}}\cdot[0.2,0.3]$、第一阶段步长 $\eta_l=(1-c_3)^l\eta_0$（其中 $c_3\le c_1$ 为任意固定常数），以及第二阶段步长 $\eta_l=\eta\asymp C_lC_u^{-2}b_1^2(nb_0)^{-1}$，则以至少 $1-\exp(-c_1rd_1)\ -3\exp(-\sqrt{n}/\log\ n)$ 的概率，在至多经过 $O\big(\log(\sigma_r/\gamma)\ +C_1^*\log\ \big(\gamma n/(d_1rb_0)\big)\wedge\log(\gamma/b_0\epsilon)\big)$ 次迭代后，算法 2 输出的估计量具有如下误差率

$$\|\hat{\mathbf{M}}-\mathbf{M}^*\|_{\mathrm{F}}^2\le\frac{C_2C_u}{C_l^2}\cdot b_0^2\cdot\max\left\{\frac{d_1r}{n},\epsilon^2\right\}.$$

Exact recovery under sparse corruptions has been studied by Li et al. (2020). Tong et al. (2021b) derived sub-optimal rates under sub-Gaussian noise and sparse corruptions. Our method can deal with heavy-tailed noise and is outlier-robust and minimax optimal.

7 Numerical Experiments

本节展示数值模拟结果，以呈现两阶段收敛现象，并将达到的精度与现有算法进行比较。由于篇幅限制，部分结果放在补充材料中。

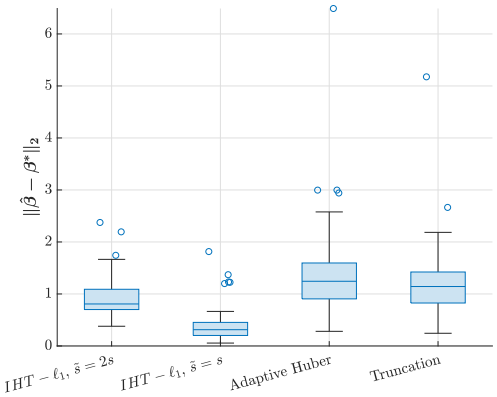
7.1 稀疏线性回归

我们的算法称为 IHT-ℓ_1。我们将其与自适应 Huber 方法（Sun 等人，(2020）以及截短方法（Fan 等人，(2021）进行比较。我们对自适应 Huber 方法和截短方法的算法参数进行了精细调优，以获得最佳性能。请注意，由于技术上的困难，我们对 IHT-ℓ_1 的理论分析只能处理 $\tilde{s} > s$ 这种情况。模拟结果表明，当设置为ℓ_1 时，ITH- $\tilde{s} = s$仍表现良好。ITH-ℓ_1 的两阶段收敛动态与低秩线性回归的情况类似，因此不再赘述。

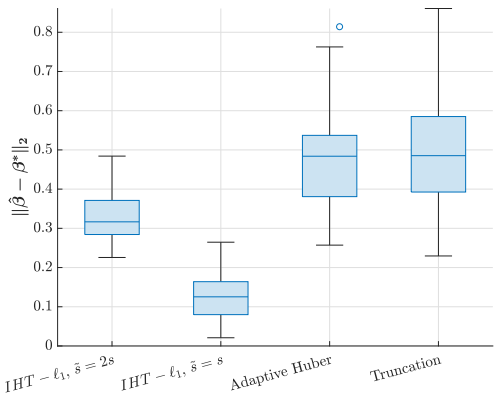
精度 我们设定 $\boldsymbol{\beta}^* = (16, 4, 1, 0, ..., 0)^\top$ 其中 $d = 50$ 且 $s = 3$。协变量向量$\{\mathbf{X}_i\}_{i=1}^n$的各项独立同分布，且服从 $N(0, 1)$ 分布。样本量从 $n \in \{50, 300\}$中选取。噪声可以服从高斯分布，也可以服从自由度为 $\nu = 2$的 Student t 分布。图 2 展示了基于 50 次模拟得到的 $\hat{}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ 箱线图。

真实数据 我们将 IHT-ℓ_1 应用于 NCI-60 癌细胞系数据集。请注意，由于该数据集最近进行了更新，我们的实验结果可能与 Sun 等人（2020）的结果略有不同。该数据集包含两部分：蛋白质：抗体阵列 DBT，以及 RNA：Affy HG-U133(A-B) - GCRMA；两者均可在网站 https://discover.nci.nih.gov/cellminer/loadDownload.do 上获取。第一部分“蛋白质：抗体阵列 DBT”包含 60 个细胞系中 162 种抗体的水平（Nishizuka 等人，2003）。我们聚焦于抗体“KRT19”，它是乳腺癌患者淋巴结、外周血和骨髓中播散肿瘤细胞最常用的标志物（Nakata 等人，2004）。第二部分“RNA：Affy HG-U133(A-B) - GCRMA”反映了 60 个细胞系中的基因转录本（Shankavaram 等人，2007）。我们的目标是筛选出参与生成蛋白质“KRT19”的显著基因。通过合并重叠基因和多个检测的基因，并剔除一个含缺失数据的细胞系，我们得到一个设计矩阵 X，其 $n = 59$ 和$d = 14$，342。有关该数据集的更多详情，请参见 Sun 等人（2020）和 Shankavaram 等人（2007）。

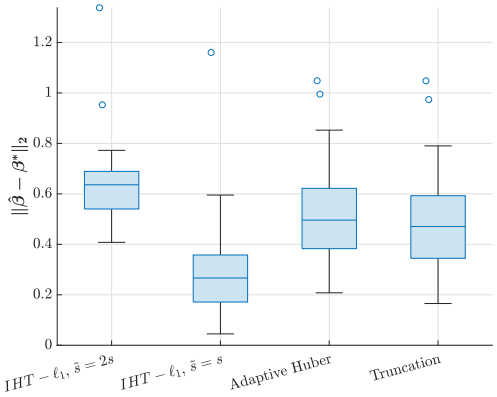
我们测试了预测性能并报告筛选出的基因。我们将数据划分为训练集和测试集。设计矩阵经对数2 转换后再进行中心化。自适应 Huber、截短方法和 IHT-ℓ_1 的算法参数通过 10 折交叉验证进行调优。测试数据上的平均绝对误差、最终模型大小以及筛选出的基因列于表 2。请注意，筛选出的基因是通过将算法应用于整个数据集获得的。这里我们仅报告 IHT-ℓ_1 在以下两种情形下的性能 $\tilde{s} = 7$ 和$\tilde{s} = 12$。表 2 显示，基因“KRT19”具有显著性，并被所有算法选中。三个算法均选中了基因“NRN1”和“EPS8L2”，其对乳腺癌的影响已由 Kim 等人 (2011); Colas 等人 (2011) 证实。基因“MALL”也



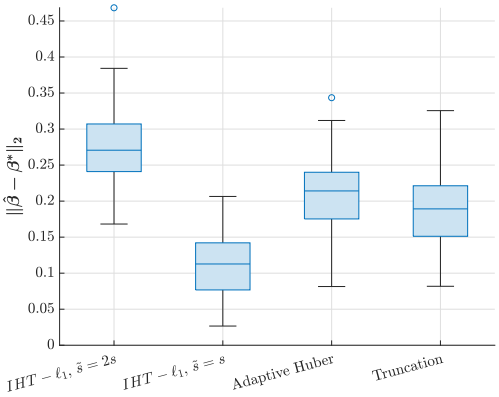
(a) 学生 t 噪声，自由度 $\nu = 2$, $n = 50$



(b) 学生 t 噪声，自由度 $\nu = 2$, $n = 300$



(c) 高斯噪声， $n = 50$



(d) 高斯噪声， $n = 300$

图 2：IHT-ℓ_1、自适应 Huber（Sun 等人，2020）和截短方法（Fan 等人，2021）在稀疏线性回归中的准确率比较。

也被所有算法选中，目前已知它与其他一些癌症有关。IHT-ℓ_1 选中的基因“TGFBI”“DSP”“DKK1”“S100A6”“EPCAM”“ZIC1”“MYO5C”“LPXN”也与乳腺癌密切相关（Li 等人，2012；Yang 等人，2012；Niu 等人，2019；Zhang 等人，2017；Osta 等人，2004；Han 等人，2018；Ye 等人，2022；Abe 等人，2016）。此外，我们注意到，自适应 Huber 和截短方法选中的基因“ANXA3”“ADRB2”“CA2”“MAMBI”“CEMIP”已被现有文献证实具有显著性（Zhou 等人，2017；Xie 等人，2019；Annan 等人，2019；Yu 和 Chai，2017；Xue 等人，2022）。

7.2 低秩线性回归

现在，我们展示 RsGrad 的两阶段收敛，并将其达到的精度与凸 Huber、凸ℓ_1 (Elsener 和 van deGeer，2018）、凸ℓ_2 (Candes 和 Plan，2011）进行比较，

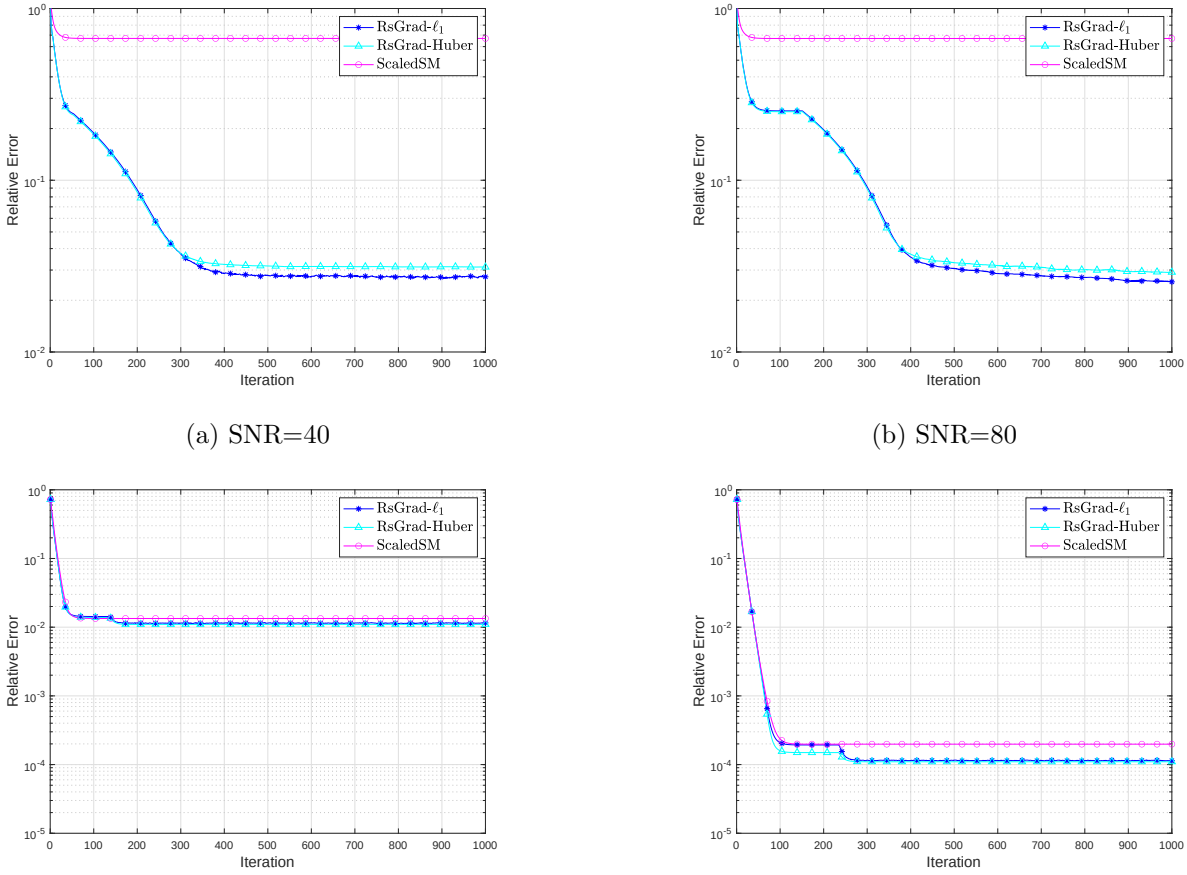
Method	MAE	Size	Selected Genes
Adaptive Huber	2.13	11	MALL, TM4SF4, ANXA3, ADRB2, NRN1, AUTS2, CA2, BAMBI, EPS8L2, CEMIP, KRT19
Truncation	2.01	10	MALL, ANXA3, ADRB2, NRN1, CA2, BAMBI, COCH, CEMIP, KRT19, EMP3
IHT-ℓ_1 ($\tilde{s} = 7$)	1.91	7	MALL, TGFBI, NRN1, DSP, EPS8L2, KRT19, LPXN
IHT-ℓ_1 ($\tilde{s} = 12$)	2.12	12	S100A6, EPCAM, MALL, ZIC1, TGFBI, NRN1, DSP, DKK1, EPS8L2, MYO5C, KRT19, LPXN

Table 2: Prediction performance on NCI-60 data. The mean absolute error of test data, final model size and selected genes are reported for three methods: IHT-ℓ_1, the adaptive Huber (Sun et al., 2020) and truncation method (Fan et al., 2021).

以及（非凸）ScaledSM 方法（Tong 等人，2021b）。精度比较详见补充材料。对于 ScaledSM 和 RsGrad 的第一阶段迭代，步长设置为 $\eta_0 = c_1\|\mathbf{M}_0\|n^{-1}$ 和 $\eta_l = q \cdot \eta_{l-1}$，其中 $c_1 > 0$、$0 < q < 1$ 是待调整参数，$\|\cdot\|$ 表示矩阵算子范数。按照 Tong 等人（2021b）的建议，我们选择 $q = 0.91$。对于 RsGrad 的第二阶段迭代，我们将常数步长设为 $\eta_l = c_2\mathbb{E}|\xi|n^{-1}$，其中 $c_2 > 0$ 是一个待调整参数。为简单起见，这里假设 $\mathbb{E}|\xi|$ 已知。实际应用中，$\mathbb{E}|\xi|$ 可以通过 $n^{-1}\sum_{i=1}^{n}|Y_i - \langle \mathbf{M}^{l_1}, \mathbf{X}_i\rangle|$ 估计，其中 $\mathbf{M}^{l_1}$ 是第一阶段迭代后的输出。遵循 Tong 等人（2021b），将信噪比 SNR 定义为：$= 20\log_{10}\left(\|\mathbf{M}^*\|_{\mathrm{F}}/\mathbb{E}|\xi|\right)$。相对误差计算为 $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}^*\|_{\mathrm{F}}^{-1}$。

Two-phase convergence We compare the convergence dynamics between RsGrad and ScaledSM. We experiment RsGrad with the absolute loss (*RsGrad-ℓ_1*) and Huber loss (*RsGrad-Huber*), but ScaledSM was proposed only for the absolute loss. The dimensions and ranks are fixed at $d_1 = d_2 = 80$ and rank $r = 5$, respectively. The phase one iterations are terminated when the stepsize becomes rather small, e.g., smaller than 10^{-10}, and then the phase two iterations are initiated.

第一组模拟旨在展示高斯噪声下的收敛性能。信噪比设为 $\{40, 80\}$，样本量则在以下数值中变化 $\{2.5, 5\} \times rd_1$。收敛动力学如图3所示，清楚地显示了 RsGrad 的两阶段收敛。我们注意到，当样本量较小时，ScaledSM 的表现较差（$n = 2.5rd_1$）。这可能是因为



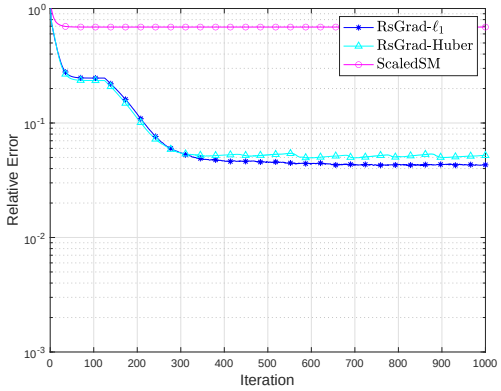
(a) SNR=40 (b) SNR=80

(c) SNR=40 (d) SNR=80

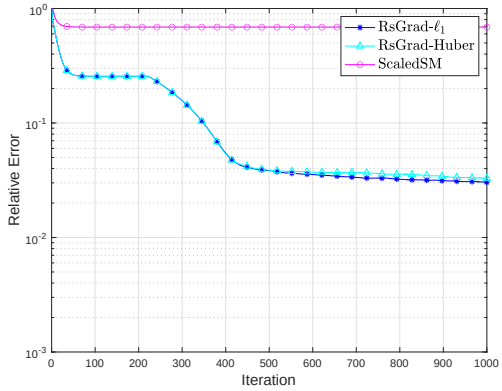
Figure 3: Convergence dynamics of RsGrad and ScaledSM (Tong et al., 2021b) under Gaussian noise. Top two figures: $n = 2.5rd_1$; bottom two figures: $n = 5rd_1$.

小样本量下逆向缩放的不稳定性。图3 中的底部两幅图显示，经过第一阶段迭代后，RsGrad 的表现与 ScaledSM 相近。然而，经过第二阶段收敛后，RsGrad（使用 Huber 损失或绝对损失）最终得到的估计值比 ScaledSM 更准确。

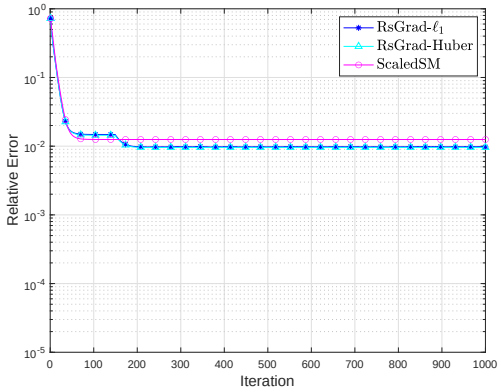
第二组模拟实验旨在测试重尾噪声下的收敛性能。为简化起见，噪声从具有 d t-自由度的 Student t分布. $\nu = 2$中独立采样，该分布具有有限的一阶（绝对）矩，但二阶矩无穷大。其他参数的选择与高斯情形相同。图 4 展示了收敛性能，表明 RsGrad（无论采用绝对损失还是 Huber 损失）和 ScaledSM 均能抵抗重尾噪声。同样，我们可以观察到 RsGrad 的两阶段收敛，其第二阶段的迭代结果比 ScaledSM 得到的估计值更加准确，尤其是在信噪比较大时。



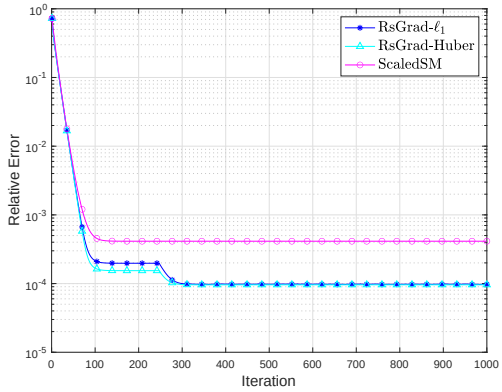
(a) SNR=40



(b) SNR=80



(c) SNR=40



(d) SNR=80

Figure 4: Convergence dynamics of RsGrad and ScaledSM (Tong et al., 2021b) under Student's t-distribution with d.f. $\nu = 2$. Top two figures: $n = 2.5rd_1$; bottom two figures: $n = 5rd_1$.

References

Akihiro Abe、Yukiya Yamamoto、Sachiko Iba、Tadaharu Kanie、Akinao Okamoto、Masutaka Tokuda、Yoko Inaguma、Masamitsu Yanada、Satoko Morishima、Shuichi Mizuta 等。伴有 NUP98-HOXA9 的复发性急性髓系白血病患者中由 t (11; 12)(q12. 1; p13) 产生的 Etv6-LPXN 融合转录本。《基因、染色体与癌症》，55(3):242–250，2016 年。

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23, 2010.

Pierre Alquier、Vincent Cottet 和 Guillaume Lecu´e。正则化程序在 Lipschitz损失函数下的估计界与锐利预言机不等式。统计年刊，47(4): 2117–2144，2019。

Dorcas A Annan、Nako Maishi、Tomoyoshi Soga、Randa Dawood、Cong Li、Hiroshi Kikuchi、Takayuki Hojo、Masahiro Morimoto、Tetsuya Kitamura、Mohammad To wfik Alam 等。碳酸酐酶 2（CAII）在肿瘤微环境的乳酸性酸中毒条件下支持肿瘤血管内皮细胞存活。细胞通讯与信号传导，17(1):1–15，2019。

Pierre C Bellec、Guillaume Lecu´e 和 Alexandre B Tsybakov。Slope方法与 Lasso相遇：改进的预言机界与最优性。统计年刊，46(6B):3603–3642，2018。

Alexandre Belloni、Victor Chernozhukov 和 Lie Wang。平方根Lasso：通过锥规划恢复稀疏信号的枢轴方法。Biometrika，98(4):791–806，2011。

Thomas Blumensath 和 Mike E Davies。用于压缩感知的迭代硬阈值法。《应用与计算调和分析》，27(3):265–274，2009。

Thomas Blumensath 和 Mike E Davies。归一化迭代硬阈值：稳定性与性能保证。IEEE信号处理专题期刊，4(2):298–309，2010。

Stephen Boyd、Stephen P Boyd 和 Lieven Vandenberghe。凸优化。剑桥大学出版社，2004。

Samuel Burer 和 Renato DC Monteiro。一种通过低秩分解求解半定规划的非线性规划算法。《数学规划》，95(2):329–357，2003。

Jian-Feng Cai、Jingyang Li 和 Dong Xia。通过快速黎曼优化进行广义低秩加稀疏张量估计。arXiv预印本 arXiv:2103.08895，2021a。

Jian-Feng Cai, Jingyang Li, and Dong Xia. Provable tensor-train format tensor completion by riemannian optimization. *arXiv preprint arXiv:2108.12163*, 2021b.

T Tony Cai and Anru Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

Léopold Cambier and P-A Absil. Robust low-rank matrix completion by riemannian optimization. *SIAM Journal on Scientific Computing*, 38(5):S440–S460, 2016.

Emmanuel J Candes 和 Yaniv Plan。利用最少数量的含噪随机测量进行低秩矩阵恢复的紧预言机不等式。《IEEE 信息论汇刊》，57(4):2342–2359，2011。

Emmanuel J Cand`es、Xiaodong Li、Yi Ma 和 John Wright。稳健主成分分析？《美国计算机协会会刊》（JACM），58(3):1–37，2011。

Olivier Catoni。对经验均值和经验方差提出挑战：一项偏差研究。载于《IHP 概率与统计年刊》，第 48 卷，第 1148–1185 页，2012。

Venkat Chandrasekaran、Benjamin Recht、Pablo A Parrilo 和 Alan S Willsky。线性逆问题的凸几何。《计算数学基础》，12(6):805–849，2012。

Lin-Ching Chang、Derek K Jones 和 Carlo Pierpaoli。Restore：通过剔除离群值稳健估计张量。《医学磁共振：国际医学磁共振学会官方期刊》，53(5):1088–1095，2005。

Vasileios Charisopoulos、Yudong Chen、Damek Davis、Mateo Díaz、Lijun Ding 和 Dmitriy Drusvyatskiy。通过复合优化进行低秩矩阵恢复：良好的条件数与快速收敛。《计算数学基础》，21(6):1505–1593，2021。

Jianhui Chen、Jiayu Zhou 和 Jieping Ye。融合低秩与组稀疏结构的鲁棒多任务学习。载于第 17 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 42–50 页，2011 年。

Mengjie Chen、Chao Gao 和 Zhao Ren。Huber 污染模型 ϵ的一般决策理论。Electronic Journal of Statistics，10(2):3752–3774，2016 年。

Mengjie Chen、Chao Gao 和 Zhao Ren。Huber 污染模型下的稳健协方差矩阵与散布矩阵估计。《统计年刊》，46(5):1932–1960，2018 年。

Xi Chen, Weidong Liu, and Xiaojun Mao. Robust reduced rank regression in a distributed setting. *Science China Mathematics*, 65(8):1707–1730, 2022.

Yudong Chen 和 Martin J. Wainwright。《通过投影梯度下降进行快速低秩估计：一般统计与算法保证》，2015年。

Geoffrey Chinot、Guillaume Lecu´e 和 Matthieu Lerasle。《基于 Lipschitz 凸损失函数的稳健统计学习》。《概率论及相关领域》，176(3-4):897–940，2020年。Justin Chiu、Yuntian Deng 和 Alexander Rush。《结构化模型快速推断的低秩约束》。《神经信息处理系统进展》，34，2021年。Eva Colas、Cristina Perez、Silvia Cabrera、Nuria Pedrola、Marta Monge、Josep Castellvi、Fernando Eyzaguirre、Jesus Gregorio、Anna Ruiz、Marta Llaurado 等。《子宫抽吸物中检测到的子宫内膜癌分子标志物》。《国际癌症杂志》，129(10):2435–2444，2011年。Mark A Davenport 和 Justin Romberg。《从不完整观测中恢复低秩矩阵概述》。《IEEE信号处理专题期刊》，10(4):608–622，2016年。Jules Depersin。《基于 VC 维数的稳健次高斯估计》。《arXiv》预印本 arXiv:2004.11734，2020年。

Jules Depersin and Guillaume Lecué. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.

Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Bradley Efron、Trevor Hastie、Iain Johnstone 和 Robert Tibshirani。最小角回归。《统计年刊》，32(2):407–499，2004。Andreas Elsener 和 Sara van de Geer。稳健低秩矩阵估计。《统计年刊》，46(6B):3481–3509，2018。Jianqing Fan、Weichen Wang 和 Ziwei Zhu。重尾数据的收缩原理：高维稳健低秩矩阵恢复。《统计年刊》，49(3):1239，2021。Simon Foucart 和 Guillaume Lecu´e。一种从次指数测量中进行稀疏恢复的 IHT 算法。《IEEE 信号处理快报》，24(9):1280–1283，2017。

Jean-Louis Goffin。关于次梯度优化方法的收敛速度。《数学规划》，13(1):329–347，1977。Michael Grant 和 Stephen Boyd。Cvx：用于规范凸规划的 MATLAB 软件，2.1 版，2014。David Gross、Yi-Kai Liu、Steven T Flammia、Stephen Becker 和 Jens Eisert。通过压缩感知进行量子态层析。《物理评论快报》，105(15):150401，2010。Wei Han、Fang Cao、Xiao-Jiao Gao、Hua-Bing Wang、Fang Chen、Si-Jian Cai、Cong Zhang、Yong-Wei Hu、Jun Ma、Xing Gu 等。ZIC1 通过靶向 survivin 在乳腺癌中发挥肿瘤抑制作用。《国际肿瘤学杂志》，53(3):937–948，2018。Samuel B Hopkins。多项式时间内以次高斯速率进行均值估计。2020。Peter J Huber。概率比检验的稳健版本。《统计数学年刊》，页码 1753–1758，1965。Peter J Huber 和 Elvezio M Ronchetti。稳健统计学。新泽西州霍博肯：Wiley。doi，10(1002): 9780470434697，2009。Prateek Jain、Ambuj Tewari 和 Purushottam Kar。关于用于高维 M 估计的迭代硬阈值法。《神经信息处理系统进展》，27，2014。Ju Hee Kim、Han-Sung Kang、Tae Woo Kim 和 Sun Jung Kim。差异甲基化杂交分析揭示了 STAT1 介导的通路参与乳腺癌。《国际肿瘤学杂志》，39(4):955–963，2011。Olga Klopp、Karim Lounici 和 Alexandre B Tsybakov。稳健矩阵补全。《概率论及相关领域》，169(1):523–564，2017。Roger Koenker 和 Gilbert Bassett Jr.。回归分位数。《计量经济学》，页码 33–50，1978。Roger Koenker 和 Kevin F Hallock。分位数回归。《经济展望杂志》，15(4): 143–156，2001。Vladimir Koltchinskii、Karim Lounici 和 Alexandre B Tsybakov。核范数惩罚与含噪低秩矩阵补全的最优速率。《统计年刊》，39(5):2302–2329，2011。

Guillaume Lecu´e 和 Matthieu Lerasle。《从妈妈的原则中学习：Le Cam 的方法》。随机过程及其应用，129(11):4385–4410，2019。Guillaume Lecu´e 和 Shahar Mendelson。《弱矩假设下的稀疏恢复》。欧洲数学学会杂志，19(3):881–904，2017。Bingyan Li、Gengyun Wen、Yongliang Zhao、Jian Tong 和 Tom K Hei。《TGFBI 在间皮瘤和乳腺癌中的作用：与肿瘤抑制的关联》。BMC 癌症，12(1):1–12，2012。Xiao Li、Zhihui Zhu、Anthony Man-Cho So 和 Rene Vidal。《非凸稳健低秩矩阵恢复》。SIAM 优化杂志，30(1):660–686，2020。Po-Ling Loh 和 Martin J Wainwright。《含噪声和缺失数据的高维回归：非凸方法的可证明保证》。神经信息处理系统进展，24，2011。Karim Lounici、Massimiliano Pontil、Sara Van De Geer 和 Alexandre B Tsybakov。《群稀疏下的预言机不等式与最优推断》，2011。M Ludoux 和 M Talagrand。《巴拿赫空间中的概率：等周不等式与随机过程》，1991。G´abor Lugosi 和 Shahar Mendelson。《重尾分布下的均值估计与回归：综述》。计算数学基础，19(5):1145–1190，2019a。G´abor Lugosi 和 Shahar Mendelson。《正则化、稀疏恢复与中位数均值竞赛》。Bernoulli，25(3):2075–2106，2019b。Gabor Lugosi 和 Shahar Mendelson。《通过中位数均值竞赛实现风险最小化》。欧洲数学学会杂志，22(3):925–965，2019c。Yuetian Luo 和 Anru R Zhang。《张量对张量回归：黎曼优化、过参数化、统计—计算差距及其相互作用》。arXiv 预印本 arXiv:2206.08756，2022。Zhongyuan Lyu、Dong Xia 和 Yuan Zhang。《高阶网络的潜在空间模型与广义张量分解》。arXiv 预印本 arXiv:2106.16042，2021。Jianhao Ma 和 Salar Fattahi。《稳健矩阵恢复的次梯度方法的全局收敛：小初始化、含噪测量与过参数化》。arXiv 预印本 arXiv:2202.08788，2022。

Zongming Ma 和 Yihong Wu。《酉不变范数下的体积比、稀疏性与极小极大性》。IEEE 信息论汇刊，61(12):6939–6956，2015。Stanislav Minsker。《巴拿赫空间中的几何中位数与稳健估计》。Bernoulli，21(4): 2308–2335，2015。Stanislav Minsker。《重尾条目随机矩阵均值的次高斯估计量》。统计年刊，46(6A):2871–2903，2018。Stanislav Minsker、Mohamed Ndaoud 和 Lang Wang。《通过平方根 Slope 方法实现稳健且无需调参的稀疏线性回归》。arXiv 预印本 arXiv:2210.16808，2022。Bamdev Mishra、Gilles Meyer、Silvere Bonnabel 和 Rodolphe Sepulchre。《固定秩矩阵分解与黎曼低秩优化》。计算统计学，29(3):591–621，2014。

B Nakata、T Takashima、Y Ogawa、T Ishikawa 和 K Hirakawa。血清 cyfra 21-1（细胞角蛋白 19 片段）是检测乳腺癌复发和评估治疗效果的有用肿瘤标志物。《英国癌症杂志》，91(5):873–878，2004。

Deanna Needell 和 Joel A Tropp。Cosamp：从不完整且不准确的样本中迭代恢复信号。《应用与计算调和分析》，26(3):301–321，2009。

Sahand Negahban 和 Martin J Wainwright。含噪声且具有高维尺度的（近似）低秩矩阵估计。《统计年刊》，第 1069–1097 页，2011。

Satoshi Nishizuka、Lu Charboneau、Lynn Young、Sylvia Major、William C Reinhold、Mark Waltham、Hosein Kouros-Mehr、Kimberly JBussey、Jae K Lee、Virginia Espina 等。使用新型高密度反相裂解液微阵列对 NCI-60 癌细胞系进行蛋白质组分析。《美国国家科学院院刊》，100(24):14229–14234，2003。

Jie Niu、Xiao-Meng Li、Xiao Wang、Chao Liang、Yi-Dan Zhang、Hai-Ying Li、Fan-Ye Liu、Hua Sun、Song-Qiang Xie 和 Dong Fang。Dkk1 通过抑制 β-catenin/mmp7 信号通路，抑制乳腺癌细胞的迁移和侵袭。《Cancer Cell International》，19(1):1–13，2019。

Walid A Osta、Yian Chen、Kaidi Mikhitarian、Michael Mitas、Mohamed Salem、Yusuf A Hannun、David J Cole 和 William E Gillanders。Epcam 在乳腺癌中过度表达，是乳腺癌基因治疗的潜在靶点。《Cancer Research》，64(16):5818–5824，2004。

Xiaoou Pan、Qiang Sun 和 Wen-Xin Zhou。迭代重加权 ℓ_1 惩罚稳健回归。Electronic Journal of Statistics，15(1):3287–3348，2021。

Garvesh Raskutti、Martin J Wainwright 和 Bin Yu。针对 ℓ_q 球上高维线性回归的极小极大估计速率。IEEE 信息论汇刊，57(10): 6976–6994，2011。Philippe Rigollet 和 Alexandre Tsybakov。指数筛选与稀疏估计的最优速率。2011。Angelika Rohde 和 Alexandre B Tsybakov。高维低秩矩阵的估计。统计年刊，39(2):887–930，2011。

Uma T Shankavaram、William C Reinhold、Satoshi Nishizuka、Sylvia Major、Daisaku Morita、Krishna K Chary、Mark A Reimers、Uwe Scherf、Ari Kahn、Douglas Dolginow 等。NCI-60 癌细胞系面板的转录本与蛋白质表达谱：一项整合组学微阵列研究。《分子癌症治疗》，6(3):820–832，2007。

Jie Shen 和 Ping Li。硬阈值的紧致界。《机器学习研究杂志》，18(1):7650–7691，2017。

Sajid Siddiqi、Byron Boots 和 Geoffrey Gordon。降秩隐马尔可夫模型。载于第十三届国际人工智能与统计学会议论文集，第 741–748 页。JMLR Workshop and Conference Proceedings, 2010。

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

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

Kean Ming Tan、Lan Wang 和 Wen-Xin Zhou。高维分位数回归：卷积平滑与凹正则化。《皇家统计学会会刊 B 辑：统计方法》，84(1)：205–233，2022。Philip Thompson。抗离群点稀疏/低秩最小二乘回归与稳健矩阵补全。arXiv 预印本 arXiv:2012.06750，2020。Robert Tibshirani。通过 Lasso 进行回归收缩与变量选择。《皇家统计学会会刊：B 辑（方法论）》，58(1)：267–288，1996。Tian Tong、Cong Ma 和 Yuejie Chi。通过缩放梯度下降加速病态低秩矩阵估计，2021a。

Tian Tong、Cong Ma 和 Yuejie Chi。采用缩放次梯度方法的低秩矩阵恢复：无需条件数的快速且稳健收敛。IEEE Transactions on Signal Processing，69:2396–2409，2021b。Sara A Van de Geer 和 Sara van de Geer。M-估计中的经验过程，第 6 卷。剑桥大学出版社，2000。Aad W Van Der Vaart、Aad van der Vaart、Adrianus Willem van der Vaart 和 Jon Wellner。弱收敛与经验过程：及其在统计学中的应用。Springer Science & Business Media，1996。Bart Vandereycken。通过黎曼优化进行低秩矩阵补全。SIAM Journal on Optimization，23(2):1214–1236，2013。Roman Vershynin。高维概率：数据科学中的应用导论，第 47 卷。剑桥大学出版社，2018。Lan Wang、Yichao Wu 和 Runze Li。用于分析超高维异质性的分位数回归。Journal of the American Statistical Association，107(497):214–222，2012。Lan Wang、Bo Peng、Jelena Bradic、Runze Li 和 Yunan Wu。高维回归的一种免调参稳健且高效的方法。Journal of the American Statistical Association，115 (532):1700–1714，2020。Lili Wang、Chao Zheng、Wen Zhou 和 Wen-Xin Zhou。免调参 Huber 回归的新原则。Statistica Sinica，31(4):2153–2177，2021。Ke Wei、Jian-Feng Cai、Tony F Chan 和 Shingyu Leung。低秩矩阵恢复的黎曼优化保证。SIAM Journal on Matrix Analysis and Applications，37(3):1198–1222，2016。AH Welsh。关于 M-过程与 M-估计。统计年刊，页码 337–361，1989。Dong Xia。低秩矩阵估计的最优 Schatten-q 和 Ky-Fan-k 范数速率。arXiv 预印本 arXiv:1403.6499，2014。Dong Xia。奇异子空间的正态近似与置信区域。Electronic Journal of Statistics，15(2):3798–3851，2021。Dong Xia 和 Vladimir Koltchinskii。低秩密度矩阵的估计：Schatten 范数及其他距离下的界。Electronic Journal of Statistics，10(2):2717–2745，2016。

Dong Xia 和 Ming Yuan。《含噪矩阵补全中线性形式的统计推断》。《英国皇家统计学会杂志：B辑（统计方法学）》，83(1):58–77，2021。

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49(1):76–99, 2021.

Wan-Ying Xie、Ruo-Hui He、Jun Zhang、Yi-Jing He、Zan Wan、Cheng-Fang Zhou、Yong-Jun Tang、Zhi Li、Howard L Mcleod 和 Jie Liu。 β受体阻滞剂通过调节 ERK/COX-2 信号通路抑制乳腺癌细胞的存活能力，其药物反应受 ADRB2 单核苷酸多态性的影响。《肿瘤学报告》，41(1):341–350，2019年。

Jinqi Xue、Xudong Zhu、Xinbo Qiao、Yulun Wang、Jiawen Bu、Xiaoying Zhang、Qingtian Ma、Lu Liang、Lisha Sun 和 Caigang Liu。CEMIP 作为乳腺癌患者的潜在生物标志物和治疗靶点。《国际医学科学杂志》，19(3):434，2022。

Linlin Yang、Yuan Chen、Tiantian Cui、Thomas Kn¨osel、Qing Zhang、Kai Frederik Albring、Otmar Huber 和 Iver Petersen。Desmoplakin 通过抑制人类肺癌中的 wnt/β-catenin 信号通路发挥肿瘤抑制作用。《致癌作用》，33(10):1863–1870，2012 年。

Fei Ye 和 Cun-Hui Zhang。《lr 球中 lq 损失下 Lasso 和 Dantzig 选择器的速率极小极大性》。《机器学习研究杂志》，11:3519–3540，2010 年。

Xinxing Ye、Jie Zhou、Dandan Tong、Dandan Wang、Hui Wang、Jixue Guo 和 Xinmei Kang。E2f1 影响乳腺癌对新辅助治疗的治疗反应。《疾病标志物》，2022，2022。

Wenzhen Yu 和 Hong Chai。通过在体外和体内激活 TGF-β/SMAD 通路，抑制 BAMBI 可降低结肠癌细胞的活力和运动能力。撤稿。DOI：10.3892/ol.2021.12505。《Oncology Letters》，14(4):4793–4799，2017。

Shizhen Zhang, Zhen Wang, Weiwei Liu, Rui Lei, Jinlan Shan, Ling Li, and Xiaochen Wang. Distinct prognostic values of s100 mrna expression in breast cancer. *Scientific reports*, 7(1):1–11, 2017.

Yuan Zhang and Dong Xia. Edgeworth expansions for network moments. *arXiv preprint arXiv:2004.06615*, 2020.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 2022.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 28, 2015.

Qi Zheng, Limin Peng, and Xuming He. Globally adaptive quantile regression with ultra-high dimensional data. *Annals of statistics*, 43(5):2225, 2015.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements, 2016.

Tao Zhou、Yong Li、Li Yang、Liang Liu、Yingchao Ju 和 Chunxiao Li。通过 RNA 干扰沉默 ANXA3 表达可抑制乳腺癌细胞的增殖和侵袭。《肿瘤学报告》，37(1):388–398，2017 年。

Junxian Zhu、Canhong Wen、Jin Zhu、Heping Zhang 和 Xueqin Wang. 最佳子集选择问题的多项式时间算法。《美国国家科学院院刊》，117(52)：33117–33123，2020年。

Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

《计算高效且统计最优的鲁棒高维线性回归》附录

Yinan Shen、Jingyang Li、Jian-Feng Cai 和 Dong Xia香
港科技大学

第 A 节介绍了因篇幅限制而未在正文中展示的内容，包括黎曼优化的细节和模拟结果。在 B 节中，我们给出稀疏向量恢复的证明。C 节证明了命题 1，即两阶段正则性条件下低秩回归的收敛动力学。D 节包含四个小节 D.1-D.4，分别在绝对损失、Huber 损失和分位数损失等不同情形下验证正则性条件。初始化分析见 E 节。接着，F 节列出了技术引理。最后，低秩回归中两个重要引理（引理 14 矩阵扰动理论）和命题3 （经验过程）的冗长证明分别见 G 节和 H 节。在最后一节 I 中，我们分析离群点鲁棒性。

以下关键要素为低秩线性回归的证明提供指导。

1. 关于 $\mathbb{M}_r$ 上的切空间。在算法中，它按照 $\mathrm{T}_{l+1} = \text{奇异值分解}_r(\mathrm{T}_l - \eta_l \mathcal{P}_{\mathbb{T}_l}(\mathrm{G}_l))$ 的方式进行更新。随后，在收敛性分析中，涉及 $\|\mathcal{P}_{\mathbb{T}_l}(\mathrm{G}_l)\|_{\mathrm{F}}^2$ 和 $\left|\langle \mathcal{P}_{\mathbb{T}_l}^{\perp}\mathrm{G}_l, \mathrm{T}_l - \mathrm{T}^*\rangle\right|$ 的上界。对于第一项，注意到

$$\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) = \mathbf{U}_l\mathbf{U}_l^\top\mathbf{G}_l + \mathbf{G}_l\mathbf{V}_l\mathbf{V}_l^\top - \mathbf{U}_l\mathbf{U}_l^\top\mathbf{G}_l\mathbf{V}_l\mathbf{V}_l^\top = \mathbf{U}_l\mathbf{U}_l^\top\mathbf{G}_l + (\mathbf{I} - \mathbf{U}_l\mathbf{U}_l^\top)\mathbf{G}_l\mathbf{V}_l\mathbf{V}_l^\top$$

可分解为两个秩至多为 r 的矩阵之和。接着，我们使用“部分 Frobenius 范数”（参见 Tong 等人 (2021a) 或引理 12）来界定 $\|\mathcal{P}_{\mathbb{T}_l}\mathbf{G}_l\|_{\mathrm{F}}$；更多细节请参见引理 12 和引理 13 。对于第二项，我们使用

$$\left|\langle \mathcal{P}_{\mathbb{T}_l}^{\perp}\mathbf{G}_l, \mathbf{T}_l - \mathbf{T}^*\rangle\right| = \left|\langle \mathbf{G}_l, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{T}_l - \mathbf{T}^*)\rangle\right|,$$

并使用 $\mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathrm{T}_l - \mathrm{T}^*) = -(\mathrm{I} - \mathrm{U}_l\mathrm{U}_l^\top)\mathrm{T}^*(\mathrm{V}\mathrm{V}^\top - \mathrm{V}_l\mathrm{V}_l^\top) = (\mathrm{I} - \mathrm{U}_l\mathrm{U}_l^\top)(\mathrm{T}_l - \mathrm{T}^*)(\mathrm{V}\mathrm{V}^\top - \mathrm{V}_l\mathrm{V}_l^\top)$。因此，根据引理 12 中部分 Frobenius 范数的定义，有

$$\left|\langle \mathcal{P}_{\mathbb{T}_l}^{\perp}\mathbf{G}_l, \mathbf{T}_l - \mathbf{T}^*\rangle\right| \le \|(\mathbf{I} - \mathbf{U}_l\mathbf{U}_l^\top)(\mathbf{T}_l - \mathbf{T}^*)(\mathbf{V}\mathbf{V}^\top - \mathbf{V}_l\mathbf{V}_l^\top)\|_{\mathrm{F}}\|\mathbf{G}_l\|_{\mathrm{F,r}}.$$

2. 扰动引理。基于 Xia（2021）中的定理 1，我们构建了一个用于矩阵扰动的强大分析方案，其中有 $\|\mathrm{T}_{l+1} - \mathrm{T}^*\|_{\mathrm{F}} \le \|\mathrm{T}_l - \mathrm{T}^* - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathrm{G}_l)\|_{\mathrm{F}} + O(\|\mathrm{T}_l - \mathrm{T}^* - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathrm{G}_l)\|_{\mathrm{F}}^2)$。为清晰起见，我们将 Xia（2021）中的定理 1 表述为定理 11。引理18讨论了对称矩阵扰动，并以此为基础得到了一般矩阵扰动引理 14。

3. 经验过程。 我们使用经验过程理论（Ludoux 和 Talagrand，1991；Van Der Vaart 等人，1996；Adamczak，2008）来证明正则性。其优点之一是能够得到强指数概率界。

4. 初始化保证。 请注意，扰动界和收敛动力学方程（9）要求 $\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \lesssim \sigma_r$。因此，我们从热启动开始。这里给出一种简洁明了的谱初始化方案。第 E 节给出了证明。

5. 命题 1。 命题 1 分析了两阶段正则性假设下的一般收敛性，并描述了步长选择。引理4、6、8 和 10 证明了若干设定下的两阶段正则性。具体而言，当损失函数远离预言矩阵 $\mathbf{M}^*$时，其表现出典型的非光滑特性；而当它足够接近 $\mathbf{M}^*$时，损失函数的表现则类似于光滑的 ℓ_2 损失。

A 补充结果

本节汇总了一些因篇幅限制而未在正文中呈现的结果。

A.1 黎曼次梯度算法的计算优势

黎曼次梯度的计算速度很快。设 $\mathbf{M}_l = \mathbf{U}_l \mathbf{\Sigma}_l \mathbf{V}_l^\top$为 $\mathbf{M}_l$的经济奇异值分解。众所周知（Absil 等，2009；Vandereycken，2013），切空间 $\mathbb{T}_l$ 可表示为 $\mathbb{T}_l := \{\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2} : \mathbf{Z} = \mathbf{U}_l \mathbf{R}^\top + \mathbf{L}\mathbf{V}_l^\top, \mathbf{R} \in \mathbb{R}^{d_2 \times r}, \mathbf{L} \in \mathbb{R}^{d_1 \times r}\}$。于是，$\mathbf{G}_l$ 在 $\mathbb{T}_l$上的投影为

$$\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) = \mathbf{U}_l \mathbf{U}_l^\top \mathbf{G}_l + \mathbf{G}_l \mathbf{V}_l \mathbf{V}_l^\top - \mathbf{U}_l \mathbf{U}_l^\top \mathbf{G}_l \mathbf{V}_l \mathbf{V}_l^\top,$$

which is of rank at most $2r$. The first step in Algorithm 2 could be written as

$$\begin{aligned} &\mathbf{M}_l - \eta_l \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) \\ =\ &\mathbf{U}_l \left(\mathbf{\Sigma}_l - \eta_l \mathbf{U}_l^\top \mathbf{G}_l \mathbf{V}_l \right) \mathbf{V}_l^\top - \eta_l \mathbf{U}_l \mathbf{U}_l^\top \mathbf{G}_l \left(\mathbf{I}_{d_2} - \mathbf{V}_l \mathbf{V}_l^\top \right) - \eta_l \left(\mathbf{I}_{d_1} - \mathbf{U}_l \mathbf{U}_l^\top \right) \mathbf{G}_l \mathbf{V}_l \mathbf{V}_l^\top. \end{aligned}$$

Let $\left(\mathbf{I}_{d_2} - \mathbf{V}_l \mathbf{V}_l^\top \right) \mathbf{G}_l^\top \mathbf{U}_l = \mathbf{Q}_1 \mathbf{R}_1$ and $\left(\mathbf{I}_{d_1} - \mathbf{U}_l \mathbf{U}_l^\top \right) \mathbf{G}_l \mathbf{V}_l = \mathbf{Q}_2 \mathbf{R}_2$ be the QR decomposition. Then the update is

$$\mathbf{M}_l - \eta_l \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) = [\mathbf{U}_l \ \mathbf{Q}_2] \begin{bmatrix} \mathbf{\Sigma}_l - \eta_l \mathbf{U}_l^\top \mathbf{G}_l \mathbf{V}_l & -\eta_l \mathbf{R}_1^\top \\ -\eta_l \mathbf{R}_2 & \mathbf{0} \end{bmatrix} [\mathbf{V}_l \ \mathbf{Q}_1]^\top,$$

其中 $[\mathbf{U}_l \ \mathbf{Q}_2]$, $[\mathbf{V}_l \mathbf{Q}_1]$ 为正交矩阵。因此，回缩的最后一步只需要对中间项（一个 $2r \times 2r$ 矩阵）进行奇异值分解，其计算量为 $O(r^3)$ 次浮点运算。更多细节请参见 Vandereycken (2013) 和 Mishra 等人 (2014)。

A.2 低秩线性回归的更多模拟结果

精度 现在，我们比较 RsGrad、ScaledSM 输出的最终估计量与凸优化方法所得估计量的统计精度，其中凸优化方法包括核范数惩罚（Elsener 和 van de Geer，2018）绝对损失（凸 ℓ_1）、Huber 损失（凸 Huber ）以及 ℓ_2- 损失（平方损失），这些方法基于开创性工作 Candes 和 Plan (2011)。我们借助 Matlab 库 cvx（Grant 和 Boyd，2014）中的函数来实现凸 ℓ_1、凸 Huber 和 ℓ_2-损失。对于每种设置，每种方法均重复运行 10 次，并展示误差率的箱线图。

参数设置与上文类似，分别对高斯噪声和 Student's t- 分布进行实验。高斯噪声下统计精度的比较结果见图 5。上方两个图表明，当样本量较小时，所有凸方法和 ScaledSM 的表现都较差（$n = 2.5rd_1$）。另一方面，样本量较大时，ScaledSM 的表现会好得多，但仍不如 RsGrad，这与我们的理论预测一致。噪声服从自由度为 t 的 Student's $\nu = 2$-分布时，也观察到了类似结果。见图 6。

从图 5 和图 6可以看出，非凸模型的精度更高，这可能是因为与凸方法相比，它们能够更精细地调节参数。我们比较的凸方法在理论上也已被证明具有统计最优性，而且计算开销更大。例如，在 $n = 5rd_1$ 的情形下，凸 ℓ_1、凸 Huber、凸 ℓ_2 分别需要超过 1.5 小时、1.5 小时和 3 小时才能完成十次重复。相比之下，RsGrad 和 ScaledSM 只需约十分钟。对于 $n = 10d_1r$这样更大的样本量，在我们的计算平台上，Matlab 软件包完成凸 ℓ_1 和凸 Huber 的模拟（重复 10 次）分别需要 3 小时和 6 小时，而 RsGrad 和 ScaledSM 各自只需约 20 分钟。

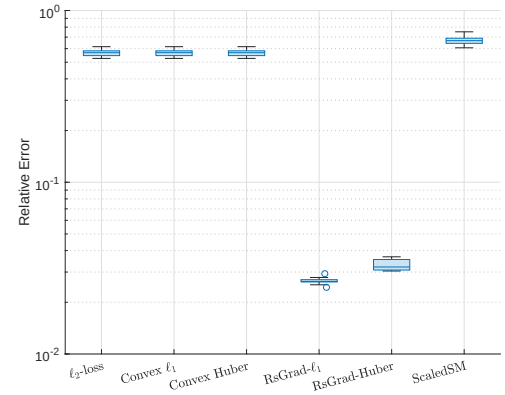
B Proofs of Sparse Linear Regression

In general, there are three key technical points, Proposition 2, Theorem 9 and Lemma 2; see below.

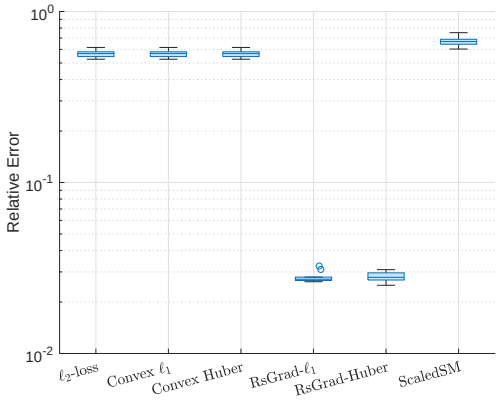
Proposition 2. *Suppose function $f(\cdot)$ is given in Equation (3). Then there exist absolute constants $C, C_1 > 0$ such that*

$$|f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta}) - \mathbb{E}\left[f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta})\right]| \leq C\sqrt{nsC_u \log(2d/s)}\|\Delta\boldsymbol{\beta}\|_2$$

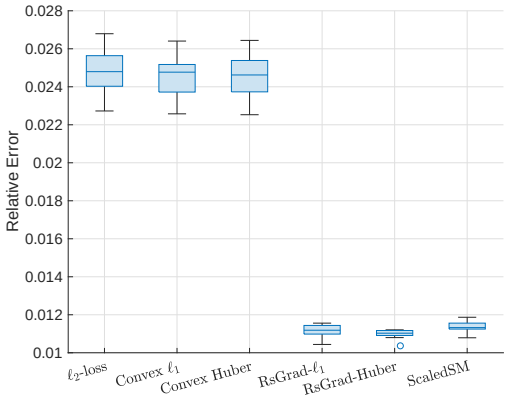
以概率大于 $1-\exp(_ C_1^2 s\log(2d/s)\,3)-3\exp\left(-\sqrt{n}\log(2d/s\ n\right))\log$的概率，对所有 $\boldsymbol{\beta} \in \mathbb{R}^d$and $\Delta\boldsymbol{\beta} \in \{$b $\in \mathbb{R}^d$: $|\overline{\text{supp(b)}|} \leq s\}$，均成立。



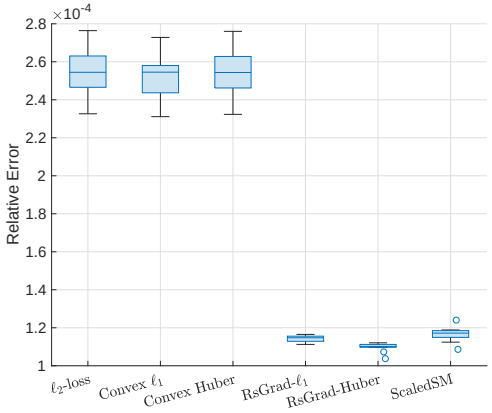
(a) SNR=40



(b) SNR=80



(c) SNR=40



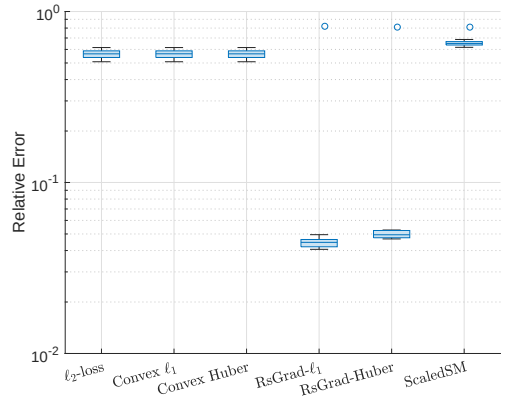
(d) SNR=80

Figure 5: Accuracy comparisons of RsGrad, ScaledSM, robust convex models (Elsener and van de Geer, 2018) and convex ℓ_2 loss (Candes and Plan, 2011) under Gaussian noise. Top two figures: $n = 2.5rd_1$; bottom two figures: $n = 5rd_1$.

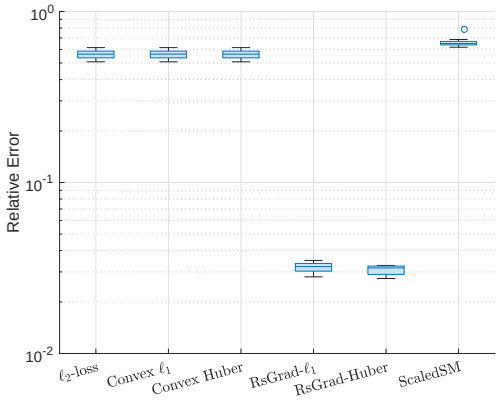
命题2 给出了 $f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta})$ 的偏差界。其证明基于经验过程理论，并将在本节后面给出。此外，我们还需要扰动界。粗略的扰动界 $\|\mathcal{H}_K(\mathbf{b}) - \mathbf{x}\|_2 \le 2\|\mathbf{b} - \mathbf{x}$通常用于 ℓ_2 损失（Blumensath 和 Davies，2009，2010），并且足以推导出 ℓ_2 损失下的统计最优性。然而，系数 2 对于 ℓ_1 损失的第二阶段收敛而言过大。值得庆幸的是，Shen 和 Li (2017) 的研究给出了硬阈值扰动的紧界。

定理9 (硬阈值的紧界Shen 和 Li (2017))。 设 $\mathbf{b} \in \mathbb{R}^d$ 为任意向量，且 $\mathbf{x} \in \mathbb{R}^d$ 为任意 K-稀疏信号。对于任意 $k \ge K$，有如下界：

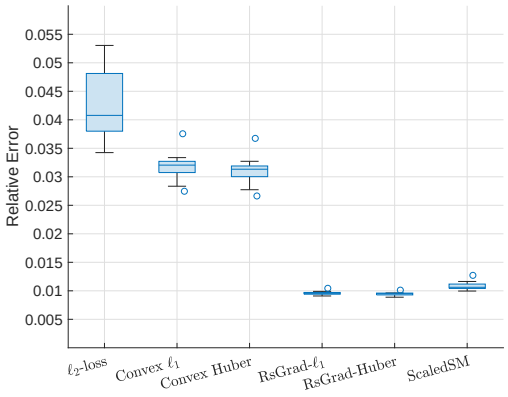
$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 \le \sqrt{\nu}\|\boldsymbol{b} - \boldsymbol{x}\|_2, \quad \nu = 1 + \frac{\rho + \sqrt{(4+\rho)\rho}}{2}, \quad \rho = \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}.$$



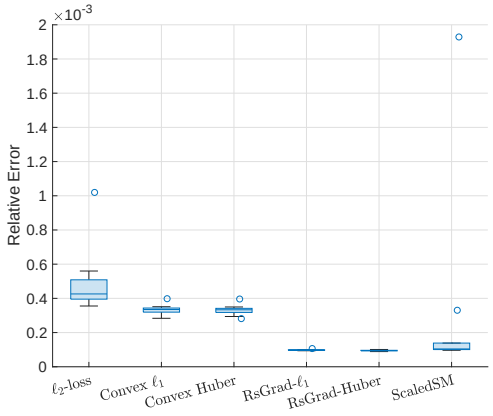
(a) SNR=40



(b) SNR=80



(c) SNR=40



(d) SNR=80

Figure 6: Accuracy comparisons of RsGrad, ScaledSM, robust convex models (Elsener and van de Geer, 2018) and convex ℓ_2 loss (Candes and Plan, 2011) under Students' t-distribution noise with d.f. $\nu = 2$. Top two figures: $n = 2.5rd_1$; bottom two figures: $n = 5rd_1$.

请注意，在某种程度上，较大的硬阈值参数 k 意味着较小的上界系数 ν。此外，定理 9 适用于任意 $\mathbf{b}$ ，且对 $\mathbf{b}$不作任何假设。这意味着，在稀疏向量恢复中，我们不需要良好的初始化，但需要实际的阈值参数 $\tilde{s} \geq \min\{cs, d\}$ 才能获得收敛保证。我们猜想，当 $\|\mathbf{b} - \mathbf{x}\|_2$ 足够小时，存在更紧的界，据此我们可以选择 $\tilde{s} = s$。这留待未来研究。

Lemma 2. *Suppose function $f(\cdot)$ is given in Equation (3). Suppose Assumptions 1 and 2 hold.*

Then for any vector $\boldsymbol{\beta} \in \left\{\mathbf{b} \in \mathbb{R}^d : \|\mathbf{b} - \boldsymbol{\beta}^*\|_2 \leq 8\sqrt{C_u C_l^{-1}}\gamma\right\}$, *we have*

$$\mathbb{E}f(\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}^*) \geq \frac{n}{6b_0}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

对于任意的 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$，我们有

$$\mathbb{E}f(\boldsymbol{\beta}_1) - \mathbb{E}f(\boldsymbol{\beta}_2) \leq nb_1^{-1}C_u \cdot \left[\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + 2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\|_2\right].$$

引理 2 给出了以下各项期望的上界和下界： $f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2)$ 以及 $f(\boldsymbol{\beta}), -f(\boldsymbol{\beta}^*$。随后，我们即可证明引理 1中的正则性性质。

引理的证明*1.* 我们假设该事件

$$\boldsymbol{\mathcal{E}} := \left\{\sup_{\boldsymbol{\beta}\in\mathbb{R}^d,\, |\mathrm{supp}(\Delta\boldsymbol{\beta})|\leq 3\tilde{s}} |f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta}) - \mathbb{E}\left[f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta})\right]| \cdot \|\Delta\boldsymbol{\beta}\|_2^{-1} \leq C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\right\}$$

成立。命题2证明了 $\mathbb{P}(\mathbf{E}) \geq 1 - \exp(-^{C^2\tilde{s}}\log(2d/\tilde{s})\,3) - 3\exp\left(-\frac{\sqrt{n}\log(2d/\tilde{s}))\log}{n}\right)$。基于事件 $\mathbf{E}$, 我们将得到正则性性质。

首先，分析第一阶段的性质，即当 $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq 8\sqrt{C_l^{-1}}\gamma$ 时的情况。对于 $f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)$，在事件 $\mathbf{E}$ 下，它具有

$$\begin{aligned}
f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) &\geq \mathbb{E}f(\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}^*) - C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
&= \mathbb{E}\left[\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \boldsymbol{\xi}\|_1 - \|\boldsymbol{\xi}\|_1\right] - C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
&\geq \mathbb{E}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_1 - 2\mathbb{E}\|\boldsymbol{\xi}\|_1 - C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.
\end{aligned}$$

注意到 $\left|\mathbf{X}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right|$ 服从折叠高斯分布，因此我们有 $\mathbb{E}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_1 \geq n\sqrt{\frac{2}{\pi}}\sqrt{C_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$。因此，结合 $\mathbb{E}|\xi| = \gamma$、$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq 8\sqrt{C_l^{-1}}\gamma$ 和 $n \geq C_1C_uC_l^{-1}\tilde{s}\log(2d/\tilde{s})$，它具有

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \geq \frac{n}{4}\sqrt{C_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

Then consider $\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2$. By definition of sub-gradient, it has

$$\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2^2 = \langle\mathbf{G}, \mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\rangle \leq f(\boldsymbol{\beta} + \mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - f(\boldsymbol{\beta}). \tag{5}$$

On the other hand, under the event $\boldsymbol{\mathcal{E}}$, it has

$$\begin{aligned}
&f(\boldsymbol{\beta} + \mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - f(\boldsymbol{\beta}) \\
&\leq \mathbb{E}f(\boldsymbol{\beta} + \mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - \mathbb{E}f(\boldsymbol{\beta}) + C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \\
&\leq \mathbb{E}\|\mathbf{X}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_1 + C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2.
\end{aligned}$$

Then together with folded Gaussian distributions, it has $\mathbb{E}\|\mathbf{X}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_1 \le n\sqrt{\frac{2}{\pi}}\sqrt{C_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2$. More details of calculations can be found in Corollary 2. Furthermore, with $n \ge C_1\tilde{s}\log(2d/\tilde{s})$, it has

$$f(\boldsymbol{\beta}+\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - f(\boldsymbol{\beta}) \le n\sqrt{C_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2.$$

那么，由上式和公式（5）可得

$$\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \le n\sqrt{C_u}.$$

其次，分析第二阶段的性质，当 $C_2 b_0\sqrt{\frac{C_u}{C_l^2}\cdot\frac{\tilde{s}\log(2d/\tilde{s})}{n}} \le \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le 8\sqrt{C_l^{-1}}\gamma)$时。根据引理2 和事件$\boldsymbol{\mathcal{E}}$，可得

$$\begin{aligned} f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) &\ge \mathbb{E}f(\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}^*) - C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \\ &\ge \frac{n}{6b_0}C_l\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 - C\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \\ &\ge \frac{n}{12b_0}C_l\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2, \end{aligned}$$

其中最后一行由 $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \ge C_2\sqrt{\frac{\tilde{s}\log(2d/\tilde{s}}{n}\cdot\frac{C_u}{C_l^2}b_0)}$得出。另一方面，有

$$\begin{aligned} &f(\boldsymbol{\beta}+\frac{b_1}{2nC_u}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - f(\boldsymbol{\beta}) \\ &\le \mathbb{E}f(\boldsymbol{\beta}+\frac{b_1}{2nC_u}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - \mathbb{E}f(\boldsymbol{\beta}) + C_1\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\cdot\frac{b_1}{2nC_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \\ &\le \frac{nC_u}{b_1}\left[\left(\frac{b_1}{2nC_u}\right)^2\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2^2 + 2\frac{b_1}{2nC_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\right] \\ &\qquad + C_1\sqrt{n\tilde{s}C_u\log(2d/\tilde{s})}\cdot\frac{b_1}{2nC_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \\ &\le \frac{b_1}{4nC_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2^2 + (2+C_1)\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2, \end{aligned}$$

其中第二个等式由引理2 得出，最后一个等式则来自第二阶段区域约束 $\sqrt{\tilde{s}\log(}$ $\frac{2d/\tilde{s}}{C_u n}$ $b_1 \le C_2 b_0\sqrt{\frac{C_u}{C_l^2}\cdot\tilde{s})\log(2d/\tilde{s})_n}$ $\le \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2$。同时，根据次梯度的定义，我们得到 $f($ $\boldsymbol{\beta}+\frac{b_1}{2nC_u}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G}))$ $-f(\boldsymbol{\beta})$ 的一个下界，

$$f(\boldsymbol{\beta}+\frac{b_1}{2nC_u}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})) - f(\boldsymbol{\beta}) \ge \langle\mathbf{G}, \frac{b_1}{2nC_u}\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\rangle = \frac{b_1}{2nC_u}\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2^2.$$

Combine the above two equations and then the quadratic inequality leads to

$$\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \le C_3C_u\frac{n}{b_1}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2.$$

□

接下来，我们就可以证明 IHT-ℓ_1的收敛动力学，定理 1。

定理的证明 *1.* 在第一阶段，为方便起见，记 $D_l := (1-c_1)^l\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$. 用归纳法证明，对 $l=0$, 结论显然成立。假设 $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \le D_l$，接下来证明$\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 \le D_{l+1}$。我们有

$$\begin{aligned}\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 &= \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - 2\eta_l\langle\boldsymbol{\beta}_l - \boldsymbol{\beta}^*, \mathbf{G}_l\rangle + \eta_l^2\|\mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)\|_2^2\\ &\le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - 2\eta_l\left(f(\boldsymbol{\beta}_l) - f(\boldsymbol{\beta}^*)\right) + \eta_l^2\|\mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)\|_2^2\\ &\le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - \eta_l\frac{n}{2}\sqrt{C_l}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 + n^2\eta_l^2 C_u,\end{aligned}$$

其中，第一行是由 $\mathrm{supp}(\boldsymbol{\beta}_l - \boldsymbol{\beta}^*) \subseteq \Omega_l \cup \Pi_l \cup \Omega^*$ 得到的，$\langle\boldsymbol{\beta}_l - \boldsymbol{\beta}^*, \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)\rangle = \langle\boldsymbol{\beta}_l - \boldsymbol{\beta}^*, \mathbf{G}_l\rangle$；第二行使用次梯度定义，最后一行使用引理 1。由归纳法可得 $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \le D_l$，而步长的选取说明 $\eta_l \in n^{-1}\sqrt{\frac{C_l}{C_u^2}}\cdot[\frac{1}{8}D_l, \frac{3}{8}D_l]$。根据二次函数性质命题 1，我们有

$$\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - \eta_l\frac{n}{2}\sqrt{C_l}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \le D_l^2 - \eta_l\frac{n}{2}\sqrt{C_l}D_l.$$

Then insert stepsize selection and with quadratic function properties, it becomes

$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 \le \left(1 - \frac{3}{64}\frac{C_l}{C_u}\right)\cdot D_l^2.$$

It implies

$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2 \le \left(1 - \frac{3}{128}\frac{C_l}{C_u}\right) D_l.$$

Then with hard thresholding perturbation Theorem 9 and $\tilde{s}$ selection, we obtain

$$\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 = \|\mathcal{H}_{\tilde{s}}(\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)) - \boldsymbol{\beta}^*\|_2 \le \left(1 - \frac{1}{64}\frac{C_l}{C_u}\right) D_l \le (1-c_1)D_l = D_{l+1},$$

其中常数 $c_1 \le \frac{1}{64}\frac{C_u}{C_l}$。

Similarly, in phase two, based on Lemma 1, we have

$$\begin{aligned}\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 &= \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - 2\eta_l\langle\boldsymbol{\beta}_l - \boldsymbol{\beta}^*, \mathbf{G}_l\rangle + \eta_l^2\|\mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)\|_2^2\\ &\le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - 2\eta_l\left(f(\boldsymbol{\beta}_l) - f(\boldsymbol{\beta}^*)\right) + \eta_l^2\|\mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l)\|_2^2\\ &\le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 - \eta_l\frac{n}{6b_0}C_l\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 + C_3^2 n^2 b_1^{-2}\eta_l^2 C_u^2\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2.\end{aligned}$$

With stepsize $\eta_l \in \frac{C_l}{C_u^2}\cdot\frac{b_1^2}{b_0}\cdot[C_4, C_5]$, we have

$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l\cup\Pi_l\cup\Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2^2 \le \left(1 - \tilde{c}_2\frac{C_l^2}{C_u^2}\cdot\frac{b_1^2}{b_0^2}\right)\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2.$$

It leads to

$$\|\boldsymbol{\beta}_l - \eta_l \mathcal{P}_{\Omega_l \cup \Pi_l \cup \Omega^*}(\mathbf{G}_l) - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \tilde{c}_2 \frac{C_l^2}{2C_u^2} \cdot \frac{b_1^2}{b_0^2}\right) \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2.$$

那么，当 $\tilde{s} \geq$最小值 $\left\{C_6 s \cdot \frac{C_u^8}{C_l^8} \cdot \frac{b_0^8}{b_1^8},\ d\right\}$扰动界定理 9 可推出

$$\begin{aligned}\|\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}^*\|_2 &\leq \left(1 + \sqrt{\frac{s}{\tilde{s}} + \sqrt{\frac{s}{\tilde{s}}}}\right) \cdot \left(1 - \tilde{c}_2 \frac{C_l^2}{2C_u^2} \cdot \frac{b_1^2}{b_0^2}\right) \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \\ &\leq \left(1 - c_2 \frac{C_l^2}{C_u^2} \cdot \frac{b_1^2}{b_0^2}\right) \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2.\end{aligned}$$

□

命题 2的证明，向量恢复经验过程。 记为 $\mathbb{V}_s := \{\mathbf{b} \in \mathbb{R}^d : \ |\mathrm{supp}(\mathbf{b})| \leq s\}$。然后考虑

$$Z := \sup_{\boldsymbol{\beta} \in \mathbb{R}^d,\, \Delta\boldsymbol{\beta} \in \mathbb{V}_s} |f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta}) - \mathbb{E}\left[f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta})\right]| \,/ \|\Delta\boldsymbol{\beta}\|_2.$$

设 $f(\boldsymbol{\beta}) := \sum_{i=1}^n |\mathbf{X}_i^\top \boldsymbol{\beta} - Y_i|$，我们有

$$\begin{aligned}\mathbb{E}Z &\leq 2\mathbb{E} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d,\, \Delta\boldsymbol{\beta} \in \mathbb{V}_s} \left|\sum_{i=1}^n \tilde{\epsilon}_i \cdot \left(|\mathbf{X}_i^\top(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - Y_i| - |\mathbf{X}_i^\top \boldsymbol{\beta} - Y_i|\right)\right| / \|\Delta\boldsymbol{\beta}\|_2 \\ &\leq 4\mathbb{E} \sup_{\Delta\boldsymbol{\beta} \in \mathbb{V}_s} \left|\sum_{i=1}^n \tilde{\epsilon}_i \cdot \mathbf{X}_i^\top \Delta\boldsymbol{\beta}\right| / \|\Delta\boldsymbol{\beta}\|_2 \\ &= 4\mathbb{E} \sup_{\Delta\boldsymbol{\beta} \in \mathbb{V}_s} \left|\left(\sum_{i=1}^n \tilde{\epsilon}_i \mathbf{X}_i\right)^\top \Delta\boldsymbol{\beta}\right| / \|\Delta\boldsymbol{\beta}\|_2 \\ &\leq 4\mathbb{E} \sup_{|\Omega| \leq s} \left\|\mathcal{P}_\Omega\left(\sum_{i=1}^n \tilde{\epsilon}_i \mathbf{X}_i\right)\right\|_2.\end{aligned}$$

其中 $\{\tilde{\epsilon}_i\}_{i=1}^n$是与 $\mathbf{X}_1$、.....、$\mathbf{X}_n$ 独立的拉德马赫序列，且算子 $\mathcal{P}_\Omega(\cdot)$ 仅保留 Ω 中的条目。第一个不等式来自定理 12 ，第二个不等式来自定理13。

注意到 $\sum_{i=1}^n \tilde{\epsilon}_i \mathbf{X}_i \sim N(\mathbf{0},\ \sum_{i=1}^n \boldsymbol{\Sigma}_i)$ 且 $\|\sum_{i=1}^n \boldsymbol{\Sigma}_i\| \leq nC_u$，由推论 1，我们有 $\mathbb{E}$上确界$_{|\Omega| \leq s} \|\mathcal{P}_\Omega\left(\sum_{i=1}^n \tilde{\epsilon}_i \mathbf{X}_i\right)\|_2 \leq \left(\mathbb{E}\text{上确界}_{|\Omega| \leq s} \|\mathcal{P}_\Omega\left(\sum_{i=1}^n \tilde{\epsilon}_i \mathbf{X}_i\right)\|_2^2\right)^{1/2} \leq C\sqrt{nsC_u \log(2d/s)}$。因此，$\mathbb{E}Z$ 可用以下内容给出上界

$$\mathbb{E}Z \leq C\sqrt{nsC_u \log(2d/s)}.$$

With similar analyses in matrix case Section H, we have

$$\left\|\max_{i=1,\ldots,n} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d,\, \Delta\boldsymbol{\beta} \in \mathbb{V}_s} |\mathbf{X}_i^\top(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - Y_i| - |\mathbf{X}_i^\top \boldsymbol{\beta} - Y_i| - \mathbb{E}\left[|\mathbf{X}_i^\top(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - Y_i| - |\mathbf{X}_i^\top \boldsymbol{\beta} - Y_i|\right]\right\|_{\Psi_1}$$

$$\leq C_1 \sqrt{C_u s} \log n,$$

and

$$\mathbb{E}\left(|\mathbf{X}_i^\top(\boldsymbol{\beta}+\Delta\boldsymbol{\beta})-Y_i|-|\mathbf{X}_i^\top\boldsymbol{\beta}-Y_i|\right)^2/\|\Delta\boldsymbol{\beta}\|_2^2\le C_u.$$

Thus with Theorem 14, we have

$$\mathbb{P}\left(Z\ge C\sqrt{nsC_u\log(2d/s)}\right)\le\exp\left(-\frac{C^2s\log(2d/s)}{3}\right)+3\exp\left(-\frac{\sqrt{n\log(2d/s)}}{\log n}\right),$$

which completes the proof. □

Proof of Lemma 2. Notice that $\mathbb{E}f(\boldsymbol{\beta})-\mathbb{E}f(\boldsymbol{\beta}^*)$ has the following expansion

$$\begin{aligned}\mathbb{E}f(\boldsymbol{\beta})-\mathbb{E}f(\boldsymbol{\beta}^*)&=\sum_{i=1}^n\mathbb{E}\left|\xi_i-\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle\right|-\mathbb{E}\left|\xi_i\right|\\&=\sum_{i=1}^n\mathbb{E}\left[\mathbb{E}\left[\left|\xi-\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle\right|-\left|\xi\right|\middle|\mathbf{X}_i\right]\right].\end{aligned}$$

To bound the conditional expectation $\mathbb{E}\left[\left|\xi-\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle\right|-\left|\xi\right|\middle|\mathbf{X}_i\right]$, we shall first analyze $\mathbb{E}|\xi-t_0|$, with $t_0\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}|\xi-t_0|&=\int_{s\ge t_0}(s-t_0)\,dH_\xi(s)+\int_{s<t_0}(t_0-s)\,dH_\xi(s)\\&=2\int_{s\ge t_0}(s-t_0)\,dH_\xi(s)+\int_{-\infty}^{+\infty}(t_0-s)\,dH_\xi(s)\\&=2\int_{s\ge t_0}(1-H_\xi(s))\,ds+t_0-\int_{-\infty}^{\infty}s\,dH_\xi(s),\end{aligned}\tag{6}$$

where the last line is due to

$$\begin{aligned}\int_{s\ge t_0}(s-t_0)\,dH_\xi(s)&=-\int_{s\ge t_0}(s-t_0)\,d(1-H_\xi(s))\\&=-(s-t_0)(1-H_\xi(s))\Big|_{t_0}^{+\infty}+\int_{s\ge t_0}(1-H_\xi(s))\;ds\\&=\lim_{s\to+\infty}(s-t_0)(1-H_\xi(s))+\int_{s\ge t_0}(1-H_\xi(s))\;ds,\end{aligned}$$

whose first term vanishes

$$0\le\lim_{s\to+\infty}(s-t_0)(1-H_\xi(s))\le\lim_{s\to+\infty}(s-t_0)\cdot\mathbb{P}(\xi>s)\le\lim_{s\to+\infty}(s-t_0)\cdot\frac{\mathbb{E}|\xi|^{1+\varepsilon}}{s^{1+\varepsilon}}=0.$$

令 $t_0=\langle\mathbf{X}_i,\ \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle$ 则方程（6）变为

$$\mathbb{E}\left[\left|\xi-\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle\right|\middle|\mathbf{X}_i\right]=2\int_{s\ge\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}(1-H_\xi(s))\,ds+\langle\mathbf{X}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle-\int_{-\infty}^{+\infty}sdH_\xi(s).$$

With $t_0 = 0$, it becomes

$$\mathbb{E}|\xi| = 2\int_{s\geq 0}(1 - H_\xi(s))\,ds - \int_{-\infty}^{+\infty} s dH_\xi(s).$$

Thus, by taking difference of the above two equations, we have the conditional expectation $\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| - |\xi\right.$

$$\begin{aligned}
\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| - |\xi| \,\middle|\, \mathbf{X}_i\right] &= -2\int_0^{\langle \mathbf{X}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}(1 - H_\xi(s))\,ds + \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \\
&= 2\int_0^{\langle \mathbf{X}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}(H_\xi(s) - H_\xi(0))\,ds \\
&= 2\int_0^{\langle \mathbf{X}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}\int_0^s h_\xi(w)\,dw\,ds,
\end{aligned}$$

其中使用了 $H_\xi(0) = 1/2$。注意，$\langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \sim N(0, (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}_i(\boldsymbol{\beta} - \boldsymbol{\beta}))$。记 $z_i := \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle$, $\sigma_i^2 := (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}_i(\boldsymbol{\beta} - \boldsymbol{\beta})$，并令 $f_{z_i}(\cdot)$ 表示 z_i 的密度。对上述等式两边分别关于 $\mathbf{X}_i$ 取期望，便得到，

$$\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| - |\xi|\right] = 2\int_{-\infty}^{+\infty}\int_0^t\int_0^s h_\xi(w) f_{z_i}(t)\,dw\,ds\,dt.$$

Note that with Assumption 2, we have

$$\begin{aligned}
\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| - |\xi|\right] &\geq 2\int_{-\sigma_i}^{+\sigma_i}\int_0^t\int_0^s h_\xi(w) f_{z_i}(t)\,dw\,ds\,dt \\
&\geq \frac{2}{b_0}\int_{-\sigma_i}^{+\sigma_i}\int_0^t s f_{z_i}(t)\,ds\,dt \\
&= \frac{1}{b_0}\int_{-\sigma_i}^{+\sigma_i} t^2 f_{z_i}(t)\,dt \\
&= \frac{\sigma_i^2}{b_0}\int_{-1}^{1} t^2 \cdot \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\,dt \geq \frac{\sigma_i^2}{6b_0},
\end{aligned}$$

where the last line uses $\int_{-1}^1 t^2 \cdot \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\,dt \geq \frac{1}{6}$. Besides, Assumption 1 implies $\sigma_i^2 \geq C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$. Thus we have

$$\begin{aligned}
\mathbb{E}f(\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}^*)| &= \sum_{i=1}^n \mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| - |\xi|\right] \\
&\geq \frac{n}{6b_0} C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.
\end{aligned}$$

另一方面，对于任意 $\boldsymbol{\beta}_1$，$\boldsymbol{\beta}_2 \in \mathbb{R}^d$根据公式（6），可得

$$\begin{aligned}
&\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| - |\xi - \langle \mathbf{X}_i, \boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\rangle| \,\middle|\, \mathbf{X}_i\right] \\
&= 2\int_{\langle \mathbf{X}_i, \boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\rangle}^{\langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle} \int_0^s h_\xi(w)\, dw\, ds \\
&\le \frac{2}{b_1}\left|\int_{\langle \mathbf{X}_i, \boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\rangle}^{\langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle} \int_0^s dw\, ds\right| \\
&= b_1^{-1} \cdot \left|\langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle^2 + 2\langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle\langle \mathbf{X}_i, \boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\rangle\right|,
\end{aligned}$$

其中第三行由噪声密度条件（假设2）得出。然后对上式两边分别关于$\mathbf{X}_i$取期望，得到

$$\mathbb{E}\left[|\xi - \langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| - |\xi - \langle \mathbf{X}_i, \boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\rangle|\right] \le b_1^{-1} C_u \cdot \left[\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + 2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\|_2\right],$$

这里用到了霍尔德不等式。对 $i = 1, \ldots, n$ 求和可得

$$\mathbb{E}f(\boldsymbol{\beta}_1) - \mathbb{E}f(\boldsymbol{\beta}_2) \le n b_1^{-1} C_u \cdot \left[\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + 2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\|_2\right],$$

由此完成证明。 □

引理 3 (Bellec等人(2018)的命题E.1)。 设 $g_1, \ldots, g_p$ 为方差至多为 σ^2的零均值高斯随机变量。记 $\left(g_1^\sharp, \ldots, g_p^\sharp\right)$ 为$(|g_1|, \ldots, |g_p|)$的非递增重排。则

$$\mathbb{P}\left(\frac{1}{s\sigma^2}\sum_{j=1}^s \left(g_j^\sharp\right)^2 > t\log\left(\frac{2p}{s}\right)\right) \le \left(\frac{2p}{s}\right)^{1-\frac{3t}{8}}$$

对所有 $t > 0$和 $s \in \{1, \ldots, p\}$，均成立。

推论1 (引理的推论3)。 在引理3的相同条件下，有

$$\mathbb{E}\sum_{i=1}^s \left(g_i^\sharp\right)^2 \le Cs\log(p/s)\sigma^2,$$

where C is some absolute constant.

Proof. We have,

$$
\begin{aligned}
&\mathbb{E}\sum_{i=1}^{s}\left(g_i^\sharp\right)^2 \\
&=\mathbb{E}\left[\sum_{i=1}^{s}\left(g_i^\sharp\right)^2\cdot 1_{\sum_{i=1}^{s}\left(g_i^\sharp\right)^2\le 8s\log(2p/s)\sigma^2}\right]+\sum_{j=8}^{+\infty}\mathbb{E}\left[\sum_{i=1}^{s}\left(g_i^\sharp\right)^2\cdot 1_{js\log(2p/s)\sigma^2\le\sum_{i=1}^{s}\left(g_i^\sharp\right)^2\le(j+1)s\log(2p/s)\sigma^2}\right] \\
&\le 8s\log(2p/s)\sigma^2+\sum_{j=8}^{+\infty}(j+1)s\log(2p/s)\sigma^2\cdot\mathbb{P}\left(\sum_{i=1}^{s}\left(g_i^\sharp\right)^2>js\log(2p/s)\sigma^2\right) \\
&\le 8s\log(2p/s)\sigma^2+s\log(2p/s)\sigma^2\cdot\sum_{j=8}^{+\infty}(j+1)\left(\left(\frac{s}{2p}\right)^{\frac{1}{4}}\right)^j,
\end{aligned}
$$

其中最后一行来自引理 3 且 $1-\frac{3}{8}t\le-\frac{1}{4}t$ 当 $t\ge 8$。此外，注意 $\left(\frac{s}{2p}\right)^{\frac{1}{4}}\le 0.5^{\frac{1}{4}}\le 0.85$，然后通过数学分析，我们有

$$
\sum_{j=8}^{+\infty}(j+1)\left(\left(\frac{s}{2p}\right)^{\frac{1}{4}}\right)^j\le\sum_{j=1}^{+\infty}(j+1)\cdot(0.85)^j\le\frac{1}{(1-0.85)^2}\le 49.
$$

因此，存在某个绝对常数 $C>0$，使得

$$
\mathbb{E}\sum_{i=1}^{s}\left(g_i^\sharp\right)^2\le Cs\log(2p/s)\sigma^2.
$$

□

C 命题证明 1

第一阶段

为方便起见，记 $d_l:=\left(1-c\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2}\right)^l\cdot\|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}}$。用归纳法证明。当 $l=0$,$\|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}}\le d_0$ 显然成立。假设我们已经有 $\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}\le d_l$，并且要证明 $\|\mathbf{M}^{l+1}-\mathbf{M}^*\|_{\mathrm{F}}\le d_{l+1}$。

注意到 $\mathbf{M}_{l+1}=$ 奇异值分解$_r(\mathbf{M}_l-\eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l))$，我们将使用引理 14 来界定 $\mathbf{M}^{l+1}$ 与真实矩阵 $\mathbf{M}^*$之间的距离。首先考虑 $\|\mathbf{M}_l-\eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)-\mathbf{M}^*\|_{\mathrm{F}}^2$,

$$
\begin{aligned}
&\|\mathbf{M}_l-\eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)-\mathbf{M}^*\|_{\mathrm{F}}^2 \\
&\qquad=\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2-2\eta_l\langle\mathbf{M}_l-\mathbf{M}^*,\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\rangle+\eta_l^2\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}^2.
\end{aligned}
\tag{7}
$$

By definition of sub-gradient, the second term can be bounded in the way of

$$
\begin{aligned}
\langle\mathbf{M}_l-\mathbf{M}^*,\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\rangle&=\langle\mathbf{M}_l-\mathbf{M}^*,\mathbf{G}_l\rangle-\langle\mathbf{M}_l-\mathbf{M}^*,\mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle \\
&\ge f(\mathbf{M}_l)-f(\mathbf{M}^*)-\langle\mathbf{M}_l-\mathbf{M}^*,\mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle.
\end{aligned}
\tag{8}
$$

Notice that $\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle$ has the following expansion, see Lemma 4.1 of Wei et al. (2016) for more detailed calculations,

$$\begin{aligned}\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle &= \langle \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{M}_l - \mathbf{M}^*), \mathbf{G}_l\rangle \\ &= -\langle \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{M}^*), \mathbf{G}_l\rangle \\ &= \langle (\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top)(\mathbf{M}_l - \mathbf{M}^*)(\mathbf{I} - \mathbf{V}_l\mathbf{V}_l^\top), \mathbf{G}_l\rangle.\end{aligned}$$

此外，由于 $\mathrm{rank}(\mathbf{M}^*) \le r$，项 $(\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top)(\mathbf{M}_l - \mathbf{M}^*)(\mathbf{I} - \mathbf{V}_l\mathbf{V}_l^\top) = -(\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top)\mathbf{M}^*(\mathbf{I} - \mathbf{V}_l\mathbf{V}_l^\top)$ 的秩至多为 r。因此，根据引理 12 中部分 Frobenius 范数的性质，我们有

$$\begin{aligned}\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle &= \langle (\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top)(\mathbf{M}_l - \mathbf{M}^*)(\mathbf{I} - \mathbf{V}_l\mathbf{V}_l^\top), \mathbf{G}_l\rangle \\ &\le \|(\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top)(\mathbf{M}_l - \mathbf{M}^*)(\mathbf{I} - \mathbf{V}_l\mathbf{V}_l^\top)\|_{\mathrm{F}}\|\mathbf{G}_l\|_{\mathrm{F,r}}.\end{aligned}$$

此外，引理 14 给出了扰动界 $\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}_l\mathbf{U}_l^\top\| \le \frac{8\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}}{\sigma_r}$，因此上式变为

$$\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle \le 8\frac{1}{\sigma_r}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2\|\mathbf{G}_l\|_{\mathrm{F,r}}.$$

此外，条件 1 假设 $\|\mathbf{G}_l\|_{\mathrm{F,r}} \le L_{\mathsf{comp}}$，于是上式变为

$$\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathbf{G}_l)\rangle \le 8\frac{L_{\mathsf{comp}}}{\sigma_r}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

因此，方程（8）的下界为

$$\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\rangle \ge f(\mathbf{M}_l) - f(\mathbf{M}^*) - 8\frac{L_{\mathsf{comp}}}{\sigma_r}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

接下来分析方程（7）的最后一项，$\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}^2$。尽管

$$\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) = \mathbf{U}_l\mathbf{U}_l^\top\mathbf{G}_l + \mathbf{G}_l\mathbf{V}_l\mathbf{V}_l^\top - \mathbf{U}_l\mathbf{U}_l^\top\mathbf{V}_l\mathbf{V}_l^\top$$

并非奇异值分解$_r(\mathbf{G}_l)$，因为 $\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)$ 可以写成两个秩至多为 r 的矩阵之和，$\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}$ 可以由 $\|$奇异值分解$_r(\mathbf{G}_l)\|_{\mathrm{F}}$ 在相差一个常数因子的意义下给出上界。引理 13 表明

$$\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}} \le \sqrt{2}\|\mathbf{G}_l\|_{\mathrm{F,r}}.$$

Then with phase one regularity conditions in Condition 1,

$$\|\mathbf{G}_l\|_{\mathrm{F,r}} \le L_{\mathsf{comp}}, \quad f(\mathbf{M}_l) - f(\mathbf{M}^*) \ge \mu_{\mathsf{comp}}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}},$$

Equation (7) could be bounded in the way of

$$\begin{aligned}&\|\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2 \\ &\quad\le \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - 2\eta_l\mu_{\mathsf{comp}}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} + 16\eta_l\frac{L_{\mathsf{comp}}}{\sigma_r}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 + 2\eta_l^2L_{\mathsf{comp}}^2 \qquad (9) \\ &\quad\le \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - \eta_l\mu_{\mathsf{comp}}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} + 2\eta_l^2L_{\mathsf{comp}}^2,\end{aligned}$$

where the last line uses the initialization condition and the induction that $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} \le d_l \le \|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \le \frac{1}{16}\frac{\mu_{\mathsf{comp}}}{L_{\mathsf{comp}}}\sigma_r$.

断言 1. 对于二次函数 $q(x) = x^2 - ax : \mathbb{R} \to \mathbb{R}$，其参数为 $a > 0$，且对于任意$x_2 \ge x_1 \ge 0$，若 $q(x_2) > 0$，则有

$$q(x_1) \le q(x_2).$$

断言 1 的证明。一方面，若 $q(x_1) \le 0$，则 $q(x_1) \le 0 < q(x_2)$。另一方面，若 $q(x_1) \ge 0$，则可推出 $x_1 > a$。注意，当限制在 $(a/2, +\infty)$ 上时，$q(\cdot)$ 是增函数。因此，关系 $x_1 \le x_2$ 意味着 $q(x_1) \le q(x_2)$。 □

将 $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - \eta_l\mu_{\mathsf{comp}}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}$ 视为关于 $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}$ 的二次函数。根据归纳法，已有 $\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} \le d_l$，而通过步长选择，我们知道 η_l 满足

$$\frac{1}{5}d_l\mu_{\mathsf{comp}}L_{\mathsf{comp}}^{-2} \le \eta_l = (1 - c\mu_{\mathsf{comp}}^2/L_{\mathsf{comp}}^2)^l \cdot \eta_0 \le \frac{3}{10}d_l\mu_{\mathsf{comp}}L_{\mathsf{comp}}^{-2}.$$

Thus, the quadratic function is positive at point d_l, namely $d_l^2 - \eta_l\mu_{\mathsf{comp}}d_l > 0$. Then Claim 1 implies

$$\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - \eta_l\mu_{\mathsf{comp}}\|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} \le d_l^2 - \eta_l\mu_{\mathsf{comp}}d_l.$$

因此，方程（9）可以用以下方式进行界定

$$\begin{aligned}\|\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2 &\le d_l^2 - \eta_l\mu_{\mathsf{comp}}d_l + 2\eta_l^2L_{\mathsf{comp}}^2 \\ &\le (1 - \frac{3}{25}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})d_l^2,\end{aligned} \tag{10}$$

where the last line is from $\frac{1}{5}d_l\mu_{\mathsf{comp}}L_{\mathsf{comp}}^{-2} \le \eta_l \le \frac{3}{10}d_l\mu_{\mathsf{comp}}L_{\mathsf{comp}}^{-2}$. Take square root on each side of Equation (10),

$$\|\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}} \le (1 - \frac{3}{50}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})d_l < d_0 \le \sigma_r/8. \tag{11}$$

Then apply matrix perturbation bound Lemma 14,

$$\begin{aligned}\|\mathbf{M}_{l+1} - \mathbf{M}^*\|_{\mathrm{F}} &= \|\mathrm{SVD}_r(\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)) - \mathbf{M}^*\|_{\mathrm{F}} \\ &\le \|\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}} + \frac{40}{\sigma_r}\|\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2,\end{aligned} \tag{12}$$

Insert Equation (11) into Equation (12) :

$$\begin{aligned}\|\mathbf{M}_{l+1} - \mathbf{M}^*\|_{\mathrm{F}} &\le (1 - \frac{3}{50}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})d_l + \frac{40}{\sigma_r}(1 - \frac{3}{50}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})^2d_l^2 \\ &\le (1 - \frac{1}{25}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})d_l,\end{aligned}$$

其中使用了 $\frac{40}{\sigma_r}d_l \le \frac{40}{\sigma_r}d_0 \le \frac{1}{50}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2}$。也就是说。

$$\|\mathbf{M}_{l+1}-\mathbf{M}^*\|_{\mathrm{F}} \le (1-\frac{1}{25}\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2})d_l \le \left(1-c\frac{\mu_{\mathsf{comp}}^2}{L_{\mathsf{comp}}^2}\right)d_l = d_{l+1},$$

其中常数为 $c \le \frac{1}{25}$。因此，第一阶段收敛性的分析到此完成。

第二阶段

考虑第二阶段，此时 $\tau_{\mathsf{stat}} \le \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}} < \tau_{\mathsf{comp}}$。其推导与第一阶段的情形略有相似，

$$\begin{aligned}
&\|\mathbf{M}_l - \eta_l \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2 \\
&= \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - 2\eta_l\langle \mathbf{M}_l - \mathbf{M}^*, \mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\rangle + \eta_l^2\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}^2 \\
&\le \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - 2\eta_l(f(\mathbf{M}_l)-f(\mathbf{M}^*)) + 2\eta_l\frac{L_{\mathsf{stat}}}{\sigma_r}\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^3 + 2\eta_l^2\|\mathbf{G}_l\|_{\mathrm{F,r}}^2 \\
&\le \|\mathbf{M}_l - \mathbf{M}^*\|_{\mathrm{F}}^2 - 2\eta_l\mu_{\mathsf{stat}}\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2 + 2\eta_l\frac{L_{\mathsf{stat}}}{\sigma_r}\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^3 + 2\eta_l^2L_{\mathsf{stat}}^2\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2 \\
&\le (1-\frac{7}{4}\eta_l\mu_{\mathsf{stat}})\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2 + 2\eta_l^2L_{\mathsf{stat}}^2\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2,
\end{aligned}$$

其中第四行由条件 1 可得，即 $f(\mathbf{M}_l) - f(\mathbf{M}^*) \ge \mu_{\mathsf{stat}}\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2$, $\|\mathbf{G}_l\|_{\mathrm{F,\ r}} \le L_{\mathsf{stat}}\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}$，最后一行利用了 $\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}} \le \|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}} \le \frac{1}{8}\frac{\mu_{\mathsf{stat}}}{L_{\mathsf{stat}}}\sigma_r$。取步长 $\frac{1}{8}\cdot\frac{\mu_{\mathsf{stat}}}{L_{\mathsf{stat}}^2} \le \eta_l \le \frac{3}{4}\cdot\frac{\mu_{\mathsf{stat}}}{L_{\mathsf{stat}}^2}$，则有

$$\|\mathbf{M}_l - \eta\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2 \le (1-\frac{1}{8}\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2})\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2.$$

对上述等式两边开平方，得到

$$\|\mathbf{M}_l - \eta\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}} \le (1-\frac{1}{16}\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2})\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}} < \|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}} \le \sigma_r/8.$$

然后使用引理 14 处理矩阵扰动，

$$\begin{aligned}
\|\mathbf{M}_{l+1}-\mathbf{M}^*\|_{\mathrm{F}} &= \|\mathrm{SVD}_r(\mathbf{M}_l - \eta_l\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)) - \mathbf{M}^*\|_{\mathrm{F}} \\
&\le \|\mathbf{M}_l - \eta\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}} + \frac{40}{\sigma_r}\|\mathbf{M}_l - \eta\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l) - \mathbf{M}^*\|_{\mathrm{F}}^2 \\
&\le (1-\frac{1}{16}\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2})\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}} + \frac{40}{\sigma_r}(1-\frac{1}{8}\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2})^2\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}}^2 \\
&\le (1-\frac{1}{32}\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2})\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}},
\end{aligned}$$

which uses the initialization condition $40\|\mathbf{M}_l-\mathbf{M}^*\|_{\mathrm{F}} \le 40\|\mathbf{M}_0-\mathbf{M}^*\|_{\mathrm{F}} \le \frac{1}{32}\cdot\frac{\mu_{\mathsf{stat}}^2}{L_{\mathsf{stat}}^2}\sigma_r$. It finishes the proof.

D 低秩线性回归的证明

D.1 定理 2 的证明

The following lemma establishes the two-phase regularity conditions of the absolute loss under Gaussian noise.

引理 4. 假设 $\xi_1, \ldots, \xi_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ 且 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 满足假设 1。存在绝对常数 $C_1, C_2, C_3, c_1 > 0$，使得若 $n \geq C_1 r d_1 C_u C_l^{-1}$ 且 $\mathbb{B}_1, \mathbb{B}_2$ 定义如下

$$\mathbb{B}_1 := \left\{ \mathbf{M} \in \mathbb{M}_r : \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \sqrt{C_l^{-1}}\sigma \right\},$$
$$\mathbb{B}_2 := \left\{ \mathbf{M} \in \mathbb{M}_r : C_2\sigma \left(\frac{C_u}{C_l^2} \cdot \frac{r d_1}{n} \right)^{1/2} \leq \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \leq \sqrt{C_l^{-1}}\sigma \right\},$$

则以至少 $1 - \exp(-c_1 r d_1) - 3\exp(-\sqrt{n}/\log n)$ 的概率，绝对损失 $f(\mathbf{M}) = \sum_{i=1}^n |Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle|$ 满足条件 1:

(1) 相对于 $\mathbf{M}^*$ 的秩-r 受限两阶段锐度,

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \frac{n}{12}\sqrt{C_l}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ \frac{n}{12\sigma} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, & for \quad \mathbf{M} \in \mathbb{B}_2; \end{cases}$$

(2) 相对于 M 的秩-r 受限两阶段次梯度界 *,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} 2n\sqrt{C_u}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ C_3 n \sigma^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_2, \end{cases}$$

where $\mathbf{G} \in \partial f(\mathbf{M})$ *is any sub-gradient.*

By putting together Lemma 4 and Proposition 1, we immediately obtain the convergence and statistical performance of RsGrad Algorithm 2. The proof is a combination of Proposition 1 and Lemma 4, and is hence omitted.

引理的证明 4。 我们假设事件 $\boldsymbol{\mathcal{E}} = \{\sup_{\Delta\mathbf{M} \in \mathbb{M}_{2r}} |f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M}) - \mathbf{M} \in \mathbb{R}^{d_1 \times d_2} \mathbb{E}(f(\mathbf{M} + \Delta\mathbf{M}) - f(\mathbf{M}))| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1} \leq C_2\sqrt{n d_1 r C_u}\}$ 成立。具体而言，定理 3 证明

$$\mathbb{P}(\boldsymbol{\mathcal{E}}) \geq 1 - \exp(-c d_1 r/4) - 3\exp(-\sqrt{n}/\log n).$$

首先考虑第一阶段中次梯度 $\|\mathbf{G}\|_{\mathrm{F},\,\mathrm{r}}$ 部分 Frobenius 范数的上界，其中 $\mathbf{G} \in \partial f(\mathbf{M})$，且 $\mathbf{M}$ 的秩至多为 r。根据事件 E，有

$$\begin{aligned}
&f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\\
\le\ &\mathbb{E}f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-\mathbb{E}f(\mathbf{M})+C_2\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}\\
=\ &\mathbb{E}\sum_{i=1}^{n}|Y_i-\langle\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}),\mathbf{X}_i\rangle|-\mathbb{E}\sum_{i=1}^{n}|Y_i-\langle\mathbf{M},\mathbf{X}_i\rangle|+C_2\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}.
\end{aligned}$$

Then by triangular inequality, we have

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\le\sum_{i=1}^{n}\mathbb{E}|\langle\mathrm{SVD}_r(\mathbf{G}),\mathbf{X}_i\rangle|+C_2\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}.$$

Moreover, Corollary 2 proves $\sum_{i=1}^{n}\mathbb{E}|\langle\mathrm{SVD}_r(\mathbf{G}),\mathbf{X}_i\rangle|\le n\sqrt{C_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}$. Thus, together with sample size condition $n\ge c\sqrt{C_uC_l^{-1}}d_1r\ge cd_1r$, we obtain

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\le 2n\sqrt{C_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}.$$

Then, recall definition of partial Frobenius norm $\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}=\|\mathbf{G}\|_{\mathrm{F,r}}$; see Lemma 12, and it leads to

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\le 2n\sqrt{C_u}\|\mathbf{G}\|_{\mathrm{F,r}}. \tag{13}$$

On the other hand, by definition of sub-gradient, we have

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\ge\langle\mathrm{SVD}_r(\mathbf{G}),\mathbf{G}\rangle=\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}^2=\|\mathbf{G}\|_{\mathrm{F,r}}^2,$$

and together with Equation (13), we have

$$\|\mathbf{G}\|_{\mathrm{F,r}}\le 2n\sqrt{C_u},$$

which implies $L_{\mathsf{comp}}=2n\sqrt{C_u}$.

Next consider the lower bound for $f(\mathbf{M})-f(\mathbf{M}^*)$. We discuss the two phases respectively, namely, phase one when $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\ge\sqrt{C_l^{-1}}\sigma$ and phase two when $C\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1r}{n}}\sigma\le\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}<\sqrt{C_l^{-1}}\sigma$.

Case 1: When $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\ge\sqrt{C_l^{-1}}\sigma=\tau_{\mathsf{comp}}$, under event $\boldsymbol{\mathcal{E}}$, we have

$$\begin{aligned}
f(\mathbf{M})-f(\mathbf{M}^*)&\ge\mathbb{E}\left[f(\mathbf{M})-f(\mathbf{M}^*)\right]-C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\\
&\ge n\sqrt{2/\pi}\frac{C_l}{\sqrt{C_l\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2+\sigma^2}+\sigma}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2-C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\\
&\ge n\sqrt{2/\pi}\frac{\sqrt{C_l}}{\sqrt{2}+1}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}-C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\\
&\ge\frac{1}{12}n\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}},
\end{aligned}$$

where the second line uses $\mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*)$, see Lemma 16, the penultimate line is from $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \sqrt{C_l^{-1}}\sigma$ and the last line uses $n \geq Cd_1 r C_u C_l^{-1}$ for some absolute constant $C > 0$. It shows with $\tau_{\mathsf{stat}} = \sqrt{C_l^{-1}}\sigma$, $\mu_{\mathsf{comp}} = \frac{1}{12}n\sqrt{C_l}$.
情况 2：当 $\tau_{\mathsf{stat}} = C\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1 r}{n}}\sigma \leq \|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} < \sqrt{C_l^{-1}}\sigma = \tau_{\mathsf{comp}}$ 时，在事件 E 下，我们有

$$
\begin{aligned}
f(\mathbf{M}) - f(\mathbf{M}^*) &\geq \mathbb{E}\left[f(\mathbf{M}) - f(\mathbf{M}^*)\right] - C_2\sqrt{nd_1 r C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\
&\geq n\sqrt{2/\pi}\frac{C_l}{\sqrt{C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2 + \sigma^2} + \sigma}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2 - C_2\sqrt{nd_1 r C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\
&\geq \frac{1}{6}\frac{n}{\sigma}C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2 - C_2\sqrt{nd_1 r C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\
&\geq \frac{1}{12}\frac{n}{\sigma}C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2,
\end{aligned}
$$

其中，倒数第二行利用了 $\sqrt{C_l}\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} < \sigma$，最后一行则由 $\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq C\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1 r}{n}}\sigma$ 得出。这表明 $\mu_{\mathsf{stat}} = \frac{1}{12}\frac{n}{\sigma}C_l$ 在 $\tau_{\mathsf{stat}} = C\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1 r}{n}}\sigma$ 条件下成立。

最后，下面的引理 5 证明了在事件 $\boldsymbol{\mathcal{E}}$ 下，$L_{\mathsf{stat}} = Cn\sigma^{-1}C_u$ 因此，引理4 的证明也就完成了。□

引理 5（子梯度上界）。设 $\mathbf{M} \in \mathbb{M}_r$ 满足 $\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{stat} := \sqrt{\frac{d_1 r}{n}\cdot\frac{C_u}{C_l^2}}\sigma$。令 $\mathbf{G} \in \partial f(\mathbf{M})$ 为任意次梯度。在事件 E 下，我们有 $\|\mathbf{G}\|_{\mathrm{F},\ \mathrm{r}} \leq Cn\sigma^{-1}C_u\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}}$，其中 $C > 0$ 为某个绝对常数。

Proof of Lemma 5. Take $\mathbf{M}_1 = \mathbf{M} + \frac{\sigma}{2nC_u}\mathrm{SVD}_r(\mathbf{G})$, where $\mathrm{SVD}_r(\mathbf{G})$ is the best rank r approximation of $\mathbf{G}$. So $\mathrm{rank}(\mathbf{M} - \mathbf{M}_1) \leq r$. Then by Lemma 16, we have

$$
\begin{aligned}
\mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) \leq& n\sqrt{2/\pi}\sqrt{\sigma^2 + C_u\|\mathbf{M}_1 - \mathbf{M}^*\|_{\mathrm{F}}^2} - n\sqrt{2/\pi}\sqrt{\sigma^2 + C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2} \\
=& \sqrt{2/\pi}n\frac{C_u\|\mathbf{M}_1 - \mathbf{M}^*\|_{\mathrm{F}}^2 - C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2}{\sqrt{\sigma^2 + C_u\|\mathbf{M}_1 - \mathbf{M}^*\|_{\mathrm{F}}^2} + \sqrt{\sigma^2 + C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2}} \\
=& \sqrt{2/\pi}n\frac{C_l\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 2C_l\langle\mathbf{M} - \mathbf{M}^*, \mathbf{M}_1 - \mathbf{M}\rangle + (C_u - C_l)\|\mathbf{M}_1 - \mathbf{M}^*\|_{\mathrm{F}}^2}{\sqrt{\sigma^2 + C_u\|\mathbf{M}_1 - \mathbf{M}^*\|_{\mathrm{F}}^2} + \sqrt{\sigma^2 + C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2}} \\
\leq& \frac{n}{\sigma}C_u\left(\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 2\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} + \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2\right),
\end{aligned}
$$

Besides, event $\boldsymbol{\mathcal{E}}$ shows

$$
\begin{aligned}
f(\mathbf{M}_1) - f(\mathbf{M}) \leq& \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1\sqrt{nd_1 r C_u}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\
\leq& \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1\frac{n}{\sigma}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\
\leq& \frac{n}{\sigma}C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + (2 + C_1)\frac{n}{\sigma}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} + \frac{n}{\sigma}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2,
\end{aligned}
$$

where the second inequality is from the condition $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{\mathsf{stat}} \geq \sqrt{\frac{d_1 r}{nC_u}}\sigma$. Then insert $\mathbf{M}_1 = \mathbf{M} + \frac{\sigma}{2nC_u}\mathrm{SVD}_r(\mathbf{G})$ into the equation and then we obtain

$$f(\mathbf{M}+\frac{\sigma}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \leq \frac{\sigma}{4nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2 + (1+C_1/2)\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{G}\|_{\mathrm{F,r}} + \frac{n}{\sigma}C_u\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2. \tag{14}$$

另一方面，根据次梯度的定义，我们有

$$f(\mathbf{M}+\frac{\sigma}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \geq \langle \mathbf{G}, \frac{\sigma}{2nC_u}\mathrm{SVD}_r(\mathbf{G})\rangle = \frac{\sigma}{2nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2. \tag{15}$$

结合 E 式（14）与方程（15），通过求解二次不等式可得 t

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq Cn\sigma^{-1}C_u\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}.$$

□

D.2 定理 3 的证明

下面的引理证明了重尾噪声下绝对损失的两阶段正则性条件。

引理 6. 假设 $\{\xi_i\}_{i=1}^n$ 和 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 分别满足假设 2 和 1。存在绝对常数 C_1, C_2, C_3, $c_1>0$，使得若 $n \geq C_1 r d_1 C_u C_l^{-1}$ 和 $\mathbb{B}_1$, $\mathbb{B}_2$ 由下式给出

$$\mathbb{B}_1 := \left\{\mathbf{M}\in\mathbb{M}_r : \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq 8\sqrt{C_l^{-1}}\gamma\right\},$$
$$\mathbb{B}_2 := \left\{\mathbf{M}\in\mathbb{M}_r : C_2 b_0\left(\frac{C_u}{C_l^2}\cdot\frac{rd_1}{n}\right)^{1/2} \leq \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} < 8\sqrt{C_l^{-1}}\gamma\right\},$$

then with probability at least $1-\exp(-c_1 r d_1) - 3\exp(-\sqrt{n}/\log n)$, *the absolute loss* $f(\mathbf{M}) = \sum_{i=1}^n |Y_i - \langle \mathbf{M}, \mathbf{X}_i\rangle|$ *satisfies Condition 1:*

(1) the rank-r restricted ***two-phase sharpness*** *with respect to* $\mathbf{M}^*$,

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \frac{n}{4}\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}, & \text{for} \quad \mathbf{M}\in\mathbb{B}_1; \\ \frac{n}{12b_0}C_l\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2, & \text{for} \quad \mathbf{M}\in\mathbb{B}_2; \end{cases}$$

(2) the rank-r restricted ***two-phase sub-gradient bound*** *with respect to* $\mathbf{M}^*$,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} 2n\sqrt{C_u}, & \text{for} \quad \mathbf{M}\in\mathbb{B}_1; \\ C_3 n b_1^{-1} C_u\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}, & \text{for} \quad \mathbf{M}\in\mathbb{B}_2, \end{cases}$$

where $\mathbf{G}\in\partial f(\mathbf{M})$ *is any sub-gradient.*

The proof of Theorem 3 is a combination of Proposition 1 and Lemma 6. Thus we only prove Lemma 6 here.

引理的证明 6。 假设事件 $\boldsymbol{\mathcal{E}} = \{$上确界 $_{\Delta\mathbf{M}\in\mathbb{M}_{2r}} |f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M}) - \mathbb{E}(f(\mathbf{M}+\Delta\mathbf{M}) - \langle_{\mathbf{M}\in\mathbb{R}^{d_1\times d_2}} f(\mathbf{M}), \Delta\mathbf{M}\rangle)| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1} \le C_2\sqrt{nd_1rC_u}\}$ 成立。具体而言，定理 3 证明了

$$\mathbb{P}(\boldsymbol{\mathcal{E}}) \ge 1 - \exp(-cd_1r/4) - 3\exp(-\sqrt{n}/\log n).$$

该 $L_{\mathsf{comp}} = 2n\sqrt{C_u}$ 证明与引理证明中的证明相同 4。接下来考虑 $f(\mathbf{M}) - f(\mathbf{M}^*)$ 的下界。

情况 1：当 $\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} \ge 8\sqrt{C_l^{-1}}\gamma =: \tau_{\mathsf{comp}}$，由事件 $\mathbf{E}$ 可得

$$\begin{aligned}
f(\mathbf{M}) - f(\mathbf{M}^*) &\ge \mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*) - C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&= \sum_{i=1}^{n}\mathbb{E}\left(|\langle\mathbf{M}-\mathbf{M}^*, \mathbf{X}_i\rangle - \xi_i| - |\xi_i|\right) - C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&\ge \sum_{i=1}^{n}\mathbb{E}\,|\langle\mathbf{M}-\mathbf{M}^*, \mathbf{X}_i\rangle| - 2\sum_{i=1}^{n}\mathbb{E}|\xi_i| - C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&\ge n\sqrt{\frac{2C_l}{\pi}}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} - 2n\gamma - C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}},
\end{aligned}$$

第三行来自三角不等式，最后一行来自推论 2。接着结合 $\|\mathbb{M} - \mathbf{M}^*\|_{\mathrm{F}} \ge 8\sqrt{C_l^{-1}}\gamma$ 和 $n \ge cd_1rC_uC_l^{-1}$，并且 $f(\mathbf{M}) - f(\mathbf{M}^*)$ 还可以进一步界定为

$$f(\mathbf{M}) - f(\mathbf{M}^*) \ge \frac{n}{4}\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}},$$

这验证了 $\mu_{\mathsf{comp}} = \frac{n}{4}\sqrt{C_l}$ 在 $\tau_{\mathsf{comp}} = 8\sqrt{C_l^{-1}}\gamma$ 条件下成立。

Case 2: Then consider the second phase when $C_2\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1r}{n}}b_0 = \tau_{\mathsf{stat}} \le \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} < \tau_{\mathsf{comp}} = 8\sqrt{C_l^{-1}}\gamma$. To bound expectation of $f(\mathbf{M}) - f(\mathbf{M}^*)$, we first consider the conditional expectation $\mathbb{E}\left[f(\mathbf{M}) - f(\mathbf{M}^*)\middle|\mathbf{X}_1, \ldots, \mathbf{X}_n\right]$. Notice for all $t_0 \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}|\xi - t_0| &= \int_{s\ge t_0}(s-t_0)\,dH_\xi(s) + \int_{s<t_0}(t_0-s)\,dH_\xi(s) \\
&= 2\int_{s\ge t_0}(s-t_0)\,dH_\xi(s) + \int_{-\infty}^{+\infty}(t_0-s)\,dH_\xi(s) \\
&= 2\int_{s\ge t_0}(1-H_\xi(s))\,ds + t_0 - \int_{-\infty}^{\infty} s\,dH_\xi(s),
\end{aligned} \tag{16}$$

where the last line is due to

$$\begin{aligned}\int_{s\geq t_0}(s-t_0)\,dH_\xi(s) &= -\int_{s\geq t_0}(s-t_0)\,d(1-H_\xi(s))\\ &= -(s-t_0)(1-H_\xi(s))\Big|_{t_0}^{+\infty}+\int_{s\geq t_0}(1-H_\xi(s))\;ds\\ &= \lim_{s\to+\infty}(s-t_0)(1-H_\xi(s))+\int_{s\geq t_0}(1-H_\xi(s))\;ds,\end{aligned}$$

其第一项为零

$$0\leq\lim_{s\to+\infty}(s-t_0)(1-H_\xi(s))\leq\lim_{s\to+\infty}(s-t_0)\cdot\mathbb{P}(\xi>s)\leq\lim_{s\to+\infty}(s-t_0)\cdot\frac{\mathbb{E}|\xi|^{1+\varepsilon}}{s^{1+\varepsilon}}=0.$$

令 $t_0=\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle$，则式（16）变为

$$\mathbb{E}\left[|\xi-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle|\,\middle|\,\mathbf{X}_i\right]=2\int_{s\geq\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle}(1-H_\xi(s))\,ds+\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle-\int_{-\infty}^{+\infty}sdH_\xi(s).$$

与此同时，当 $t_0=0$ 时，式（16）变为

$$\mathbb{E}\,|\xi|=2\int_{s\geq0}(1-H_\xi(s))\,ds-\int_{-\infty}^{+\infty}s\,dH_\xi(s).$$

将上述两个等式相减，得到

$$\mathbb{E}\left[|\xi-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle|\,\middle|\,\mathbf{X}_i\right]-\mathbb{E}\,|\xi|=2\int_0^{\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle}H_\xi(s)ds-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle.\tag{17}$$

注意到 $\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle\sim N(0,\mathrm{vec}(\mathbf{M}-\mathbf{M}^*)^\top\mathbf{\Sigma}_i\mathrm{vec}(\mathbf{M}-\mathbf{M}^*))$，并记 $z_i:=\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle,\sigma_i^2:=\mathrm{vec}(\mathbf{M}-\mathbf{M}^*)^\top\mathbf{\Sigma}_i\mathrm{vec}(\mathbf{M}-\mathbf{M}^*)$。注意，假设 1 中关于感知矩阵的条件意味着

$$C_l\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2\leq\sigma_i^2\leq C_u\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2$$

记 the density of z_i as $f_{z_i}(\cdot)$. Take expectation of z_i on both sides of Equation (17)

$$\begin{aligned}\mathbb{E}\left[|\xi_i-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle|-|\xi_i|\right] &= 2\int_{-\infty}^{+\infty}\int_0^t(H_\xi(s)-0.5)\,f_{z_i}(t)\,ds\,dt\\ &= 2\int_{-\infty}^{+\infty}\int_0^t\int_0^s h_\xi(w)f_{z_i}(t)\,dw\,ds\,dt.\end{aligned}$$

Notice that by conditions on sensing matrices Assumption 1, we have $\sigma_i\leq\sqrt{C_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\leq 8\sqrt{C_uC_l^{-1}}\gamma$. Hence, when $w\in(-\sigma_i,\sigma_i)$, by Assumption 2, we have $h_\xi(w)\geq b_0^{-1}$. Thus the above equation could be further bounded with

$$\begin{aligned}\mathbb{E}\left[|\xi_i-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle|-|\xi_i|\right] &\geq 2\int_{-\sigma_i}^{\sigma_i}\int_0^t f_{z_i}(t)b_0^{-1}s\,ds\,dt\\ &= b_0^{-1}\int_{-\sigma_i}^{\sigma_i}t^2f_{z_i}(t)\,dt=b_0^{-1}\sigma_i^2\int_{-1}^{1}t^2\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt\\ &\geq\frac{1}{6b_0}\sigma_i^2,\end{aligned}$$

where the last line is from $\int_{-1}^{1} t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \geq 1/6$. Therefore, sum over $i = 1\ldots, n$ and then it has

$$\mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*) = \sum_{i=1}^{n} \mathbb{E}\left[|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| - |\xi_i|\right] \geq \frac{1}{6b_0} \sum_{i=1}^{n} \sigma_i^2 \geq \frac{n}{6b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

Invoke event $\boldsymbol{\mathcal{E}}$,

$$\begin{aligned} f(\mathbf{M}) - f(\mathbf{M}^*) &\geq \mathbb{E}[f(\mathbf{M}) - f(\mathbf{M}^*)] - C\sqrt{nd_1 r}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\ &\geq \frac{n}{6b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2 - C\sqrt{nd_1 r C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\ &\geq \frac{n}{12b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, \end{aligned}$$

where the last inequality uses $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq C_2 \sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_0$. This proves $\mu_{\mathsf{stat}} = \frac{n}{12b_0} C_l$.

Finally, from the following lemma, we see that $L_{\mathsf{stat}} \leq C_3 n b_1^{-1} C_u$ when $\boldsymbol{\mathcal{E}}$ holds. And this finishes the proof. □

引理 7（子梯度上界）。设 $\mathbf{M} \in \mathbb{M}_r$ 满足 $\|\mathbf{M}–\mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{stat} := C_1 \sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_1$。令 $\mathbf{G} \in \partial f(\mathbf{M})$ 为任意一个次梯度。在事件 $\mathbf{E}$ 下，我们有

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq C_2 n b_1^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}},$$

where $C_1, C_2 > 0$ are absolute constants.

引理的证明 7。 取 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{2nC_u}$奇异值分解$_r(\mathbf{G})$，其中 奇异值分解$_r(\mathbf{G})$ 是 r 阶最佳近似 $\mathbf{G}$。结合 $f(\mathbf{M}_1)$ 的下界和上界 $- f(\mathbf{M})$，即可完成证明。

First consider $\mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M})$. Use the same notation as the above proof and similarly,

$$\begin{aligned} &\mathbb{E}\left[|\xi_i - \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}^* \rangle| - |\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \,\middle|\, \mathbf{X}_i\right] \\ =&2 \int_{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}^{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle} \int_0^{\xi} h_\xi(x)\, dx\, d\xi \\ \leq&2b_1^{-1} \left| \int_{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}^{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle} \xi\, d\xi \right| \\ =&b_1^{-1} \left| (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle)^2 - (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle)^2 \right|, \end{aligned}$$

where the inequality follows from the upper bound for $h_\xi(x)$. Then take expectation with respect to $\mathbf{X}_i$ on each side and sum up over i :

$$\begin{aligned} \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) =& \sum_{i=1}^{n} \mathbb{E}\left[|\xi_i - \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}^* \rangle| - |\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \,\middle|\, \mathbf{X}_i\right] \\ \leq& b_1^{-1} \sum_{i=1}^{n} \mathbb{E}\left| (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle)^2 - (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle)^2 \right| \\ \leq& n b_1^{-1} C_u \left[\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 2\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \right], \end{aligned}$$

where the last inequality uses Holder's inequality. Then, bound $f(\mathbf{M}_1) - f(\mathbf{M})$ under event $\boldsymbol{\mathcal{E}}$, we have

$$\begin{aligned} f(\mathbf{M}_1) - f(\mathbf{M}) \leq & \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1\sqrt{nd_1rC_u}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\ \leq & \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1nb_1^{-1}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\ \leq & nb_1^{-1}C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + (2 + C_1)nb_1^{-1}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}, \end{aligned}$$

其中，第二个不等式来自条件 $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{\mathsf{stat}} = C_2\sqrt{\frac{d_1r}{n} \cdot \frac{C_u}{C_l^2}}b_0 \geq C_2\sqrt{\frac{d_1r}{n}}\sqrt{C_u^{-1}}b_1$。代入 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{2nC_u}\mathsf{SVD}_r(\mathbf{G})$ 后可得

$$\begin{aligned} & f(\mathbf{M} + \frac{b_1}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \\ & \qquad \leq \frac{b_1}{4nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2 + (1 + C_1/2)\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{G}\|_{\mathrm{F,r}}. \end{aligned} \tag{18}$$

另一方面，根据次梯度的定义，$f(\mathbf{M}_1) - f(\mathbf{M})$ 的下界为

$$f(\mathbf{M} + \frac{b_1}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \geq \langle\frac{b_1}{2nC_u}\mathrm{SVD}_r(\mathbf{G}), \mathbf{G}\rangle = \frac{b_1}{2nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2. \tag{19}$$

结合公式 (18) 与公式 (19)，然后求解二次不等式，可得

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq C_2nb_1^{-1}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}.$$

□

D.3 定理 4 的证明

下面的引理证明了重尾噪声下 Huber 损失的两阶段正则性条件。

引理 8。假设 $\{\xi_i\}_{i=1}^n$ 且 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 分别满足假设 *3* 和 *1*。存在绝对常数 $C_1, C_2, C_3, c_1 > 0$ 使得若 $n \geq C_1rd_1C_uC_l^{-1}$ 并定义 $\mathbb{B}_1, \mathbb{B}_2$ 为

$$\begin{aligned} \mathbb{B}_1 &:= \left\{\mathbf{M} \in \mathbb{M}_r : \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq 8\sqrt{C_l^{-1}}\gamma + 2\sqrt{C_l^{-1}}\delta\right\}, \\ \mathbb{B}_2 &:= \left\{\mathbf{M} \in \mathbb{M}_r : C_2b_0\left(\frac{C_u}{C_l^2} \cdot \frac{rd_1}{n}\right)^{1/2} \leq \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} < 8\sqrt{C_l^{-1}}\gamma + 2\sqrt{C_l^{-1}}\delta\right\}, \end{aligned}$$

then with probability at least $1-\exp(-c_1rd_1)-3\exp(-\sqrt{n}/\log n)$, *the Huber loss* $f(\mathbf{M}) = \sum_{i=1}^n \rho_{H,\delta}(Y_i - \langle\mathbf{M}, \mathbf{X}_i\rangle)$ *satisfies Condition 1:*

(1) the rank-r restricted ***two-phase sharpness*** *with respect to* $\mathbf{M}^*$,

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \frac{\delta n}{2}\sqrt{C_l}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & \textit{for} \quad \mathbf{M} \in \mathbb{B}_1; \\ \frac{\delta n}{3b_0}C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, & \textit{for} \quad \mathbf{M} \in \mathbb{B}_2; \end{cases}$$

(2) the rank-r restricted ***two-phase sub-gradient bound*** *with respect to* $\mathbf{M}^*$,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} 4\delta n\sqrt{C_u}, & \text{for} \quad \mathbf{M} \in \mathbb{B}_1; \\ C_3\delta n b_1^{-1} C_u \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}, & \text{for} \quad \mathbf{M} \in \mathbb{B}_2, \end{cases}$$

其中 $\mathsf{G} \in \partial f(\mathsf{M})$ 是任意一个次梯度。

定理 4 的证明结合了命题 1 和引理 8。因此，我们这里只证明引理8 。

引理 8 的证明。我们接下来假设事件 $\mathsf{E} = \{\sup_{\Delta\mathsf{M}\in\mathbb{M}_{2r}} |f(\mathsf{M}+\Delta\mathsf{M})-f(\mathsf{M})-\mathsf{M}_{\in\mathbb{R}^{d_1\times d_2}}\mathbb{E}(f(\mathsf{M}+\Delta\mathsf{M})-f(\mathsf{M}))| \cdot \|\Delta\mathsf{M}\|_{\mathsf{F}}^{-1} \leq C_2\delta\sqrt{nd_1rC_u}\}$ 成立。具体而言，定理 3 证明了

$$\mathbb{P}(\boldsymbol{\mathcal{E}}) \geq 1-\exp(-cd_1r/4)-3\exp(-\sqrt{n}/\log n).$$

首先考虑次梯度 $\mathsf{G} \in f(\mathsf{M})$ 的上界，其中矩阵 M 的秩至多为 r。根据事件 E 和 2δ 的$\rho_{H,\delta}(\cdot)$ 的 Lipschitz 连续性，我们有

$$\begin{aligned} & f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M}) \\ =& \sum_{i=1}^n \rho_{H,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}+\mathrm{SVD}_r(\mathbf{G})\rangle)-\rho_{H,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}\rangle) \\ \leq& \mathbb{E}\sum_{i=1}^n \rho_{H,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}+\mathrm{SVD}_r(\mathbf{G})\rangle)-\rho_{H,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}\rangle)+C_2\delta\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}} \\ \leq& \mathbb{E}\sum_{i=1}^n 2\delta|\langle\mathbf{X}_i,\mathrm{SVD}_r(\mathbf{G})\rangle|+C_2\delta\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}}. \end{aligned}$$

注意，推论 2 给出了 $|\langle\mathsf{X}_i,\mathsf{SVD}_r(\mathsf{G})\rangle| \leq n\sqrt{2/\pi}\sqrt{C_u}\|\mathsf{SVD}_r(\mathsf{G})\|_{\mathrm{F}} = n\sqrt{2/\pi}\sqrt{C_u}\|\mathsf{G}\|_{\mathrm{F,r}}$和 $n \geq cd_1r$ 的界。因此，上式可以用以下内容给出上界

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M}) \leq 4\delta n\sqrt{C_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}} = 4\delta n\sqrt{C_u}\|\mathbf{G}\|_{\mathrm{F,r}}$$

On the other hand, by definition of sub-gradient, it has

$$f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M}) \geq \langle\mathrm{SVD}_r(\mathbf{G}),\mathbf{G}\rangle = \|\mathbf{G}\|_{\mathrm{F,r}}^2.$$

Thus the above two equations lead to

$$\|\mathbf{G}\|_{\mathrm{F,r}}^2 \leq f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M}) \leq 4\delta n\sqrt{C_u}\|\mathbf{G}\|_{\mathrm{F,r}},$$

这意味着 $\|\mathsf{G}\|_{\mathrm{F,\ r}} \leq 4\delta n\sqrt{C_u}$，即 $L_{\mathsf{comp}} = 4\delta n\sqrt{C_u}$。

Then consider the lower bound of $f(\mathbf{M})-f(\mathbf{M}^*)$.

Case 1: When $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq 8\gamma\sqrt{C_l^{-1}}+2\delta\sqrt{C_l^{-1}}$, use the inequality $2\delta|x|-\delta^2 \leq \rho_{H,\delta}(x) \leq 2\delta|x|$. Under the event $\boldsymbol{\mathcal{E}}$, we have

$$\begin{aligned}
f(\mathbf{M})-f(\mathbf{M}^*) &\geq \mathbb{E}f(\mathbf{M})-\mathbb{E}f(\mathbf{M}^*)-C_2\delta\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&= \mathbb{E}\sum_{i=1}^{n}\rho_{H,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle)-\rho_{H,\delta}(\xi_i)-C_2\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&\geq 2\delta\mathbb{E}\sum_{i=1}^{n}\left(|\xi_i-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle|-|\xi_i|\right)-n\delta^2-C_2\delta\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&\geq 2\delta\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\mathbb{E}\left|\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle\right|-4\delta\sum_{i=1}^{n}\mathbb{E}|\xi_i|-n\delta^2-C_2\delta\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
&\geq 2n\delta\sqrt{\frac{2C_l}{\pi}}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}-4\delta n\gamma-n\delta^2-C_2\delta\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}},
\end{aligned}$$

其中最后一行来自推论 2，且 $\mathbb{E}|\xi_i|=\gamma_\circ$ 于是，当 $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq 8\gamma\sqrt{C_l^{-1}}+2\delta\sqrt{C_l^{-1}}$ 且样本量为 $n \geq cd_1rC_uC_l^{-1}$ 时，有

$$f(\mathbf{M})-f(\mathbf{M}^*) \geq \frac{1}{2}n\delta\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}.$$

这表明，当 $\tau_{\mathsf{comp}}=8\gamma\sqrt{C_l^{-1}}+2\delta\sqrt{C_l^{-1}}$ 时，我们有 $\mu_{\mathsf{comp}}=\delta n\sqrt{C_l}/2_\circ$

情况 2：当 $C_2b_0\sqrt{\frac{C_u}{C_l^2}\cdot\frac{d_1r}{n}} \leq \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \leq 8\gamma\sqrt{C_l^{-1}}+2\delta\sqrt{C_l^{-1}}$ 时，首先考虑在给定 $\mathbf{X}_i$ 条件下的期望。对于任意 $z\in\mathbb{R}$，有

$$\begin{aligned}
&\mathbb{E}\left[\rho_{H,\delta}(\xi_i-z)\right] \\
&= \int_{\delta+z}^{+\infty}2\delta(x-z)-\delta^2\,dH_\xi(x)+\int_{-\infty}^{z-\delta}2\delta(z-x)-\delta^2dH_\xi(x)+\int_{z-\delta}^{z+\delta}(z-x)^2\,dH_\xi(x).
\end{aligned} \tag{20}$$

对变量 z求导，

$$\begin{aligned}
&\frac{d}{d\,z}\mathbb{E}\left[\rho_{H,\delta}(\xi_i-z)\right] \\
&= -\delta^2h_\xi(\delta+z)-2\delta\int_{\delta+z}^{+\infty}dH_\xi(x)+\delta^2h_\xi(z-\delta)+2\delta\int_{-\infty}^{z-\delta}dH_\xi(x) \\
&\quad +\delta^2h_\xi(z+\delta)-\delta^2h_\xi(z-\delta)-2\int_{z-\delta}^{z+\delta}(x-z)\,dH_\xi(x) \\
&= -2\delta+2\delta H_\xi(z+\delta)+2\delta H_\xi(z-\delta)-2(x-z)H_\xi(x)\Big|_{z-\delta}^{z+\delta}+2\int_{z-\delta}^{z+\delta}H_{\xi_i}(x)\,dx \\
&= 2\int_{z-\delta}^{z+\delta}H_\xi(x)\,dx-2\delta,
\end{aligned}$$

and take second order derivative,

$$\frac{d^2}{dz^2}\mathbb{E}\left[\rho_{H,\delta}(\xi_i-z)\right]=2(H_\xi(z+\delta)-H_\xi(z-\delta)).$$

Assumption 3 guarantees $\frac{d}{dz}\mathbb{E}\left[\rho_{H,\delta}(\xi_i - z)\right]\Big|_{z=0} = 0$. Therefore the Taylor expansion at 0 is

$$\mathbb{E}\left[\rho_{H,\delta}(\xi_i - z)\right] = \mathbb{E}\left[\rho_{H,\delta}(\xi_i)\right] + 2\int_0^z (H_\xi(y+\delta) - H_\xi(y-\delta))y\,dy.$$

插入 $z = \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle \sim N(0, \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M} - \mathbf{M}^*))$ 代入上式，

$$\mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle)\Big|\mathbf{X}_i\right] - \mathbb{E}\left[\rho_{H,\delta}(\xi_i)\Big|\mathbf{X}_i\right] = 2\int_0^z (H_\xi(y+\delta) - H_\xi(y-\delta))y\,dy.$$

记 $\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle$ 的密度为 $f_{z_i}(\cdot)$，方差记为 $\sigma_i^2 := \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)$。对 $\mathbf{X}_i$ 取期望，得到

$$\begin{aligned}\mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle)\right] - \mathbb{E}\left[\rho_{H,\delta}(\xi_i)\right] &= 2\int_{-\infty}^{\infty}\int_0^t (H_\xi(y+\delta) - H_\xi(y-\delta))y f_{z_i}(t)\,dy\,dt \\ &\geq 2\int_{-\sigma_i}^{\sigma_i}\int_0^t f_{z_i}(t)(H_\xi(y+\delta) - H_\xi(y-\delta))y\,dy\,dt.\end{aligned}$$

注意到，感知矩阵的条件（假设 1 和第二阶段区域）意味着 $\sigma_i \leq \sqrt{C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \leq 8\gamma\sqrt{C_u C_l^{-1}} + 2\delta\sqrt{C_u C_l^{-1}}$。因此，对于 $y \in (-\sigma_i, \sigma_i)$，根据假设 3，有 $H_\xi(y+\delta) - H_\xi(y-\delta) \geq 2\delta b_0^{-1}$。因此，上式的下界为

$$\begin{aligned}\mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle)\right] - \mathbb{E}\left[\rho_{H,\delta}(\xi_i)\right] &\geq 4\delta b_0^{-1}\int_{-\sigma_i}^{\sigma_i} t^2 f_{z_i}(t)\,dt \\ &= 4\delta b_0^{-1}\sigma_i^2\int_{-1}^{1} t^2 \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\,dt \\ &\geq \frac{2}{3}\delta b_0^{-1}\sigma_i^2,\end{aligned}$$

其中最后一行用到了 $\int_{-1}^1 t^2 \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\,dt \geq \frac{1}{6}$。感知矩阵的条件（假设 1）还意味着 $\sigma_i^2 \geq C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2$。因此，$\mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*)$ 有如下下界

$$\mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*) = \sum_{i=1}^n \mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle)\right] - \mathbb{E}\left[\rho_{H,\delta}(\xi_i)\right] \geq \frac{2}{3}n\delta b_0^{-1} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

Based on event $\boldsymbol{\mathcal{E}}$, we obtain

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \mathbb{E}[f(\mathbf{M}) - f(\mathbf{M}^*)] - C_2\delta\sqrt{n d_1 r C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}.$$

因此，当 $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq b_0 C_2\sqrt{\frac{d_1 r}{n}\cdot\frac{C_u}{C_l^2}}$ 时，我们有

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \frac{1}{3}\delta n b_0^{-1} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

它证明在 $\tau_{\mathsf{stat}} = b_0 C_2\sqrt{\frac{d_1 r}{n}\cdot\frac{C_u}{C_l^2}}$ 条件下，它具有 $\mu_{\mathsf{stat}} = \frac{1}{3}\delta n b_0^{-1} C_l$。下面的引理表明，当 $\mathbf{E}$ 成立时，它具有 $L_{\mathsf{stat}} = C_2\delta C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}$，至此证明完成。 □

引理 9（次梯度上界）。设 $\mathbf{M} \in \mathbb{M}_r$ 满足 $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{stat} := c\sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_0$。令 $\mathbf{G} \in \partial f(\mathbf{M})$ 为次梯度。在事件 $\mathbf{E}$ 下，我们有 $\|\mathbf{G}\|_{\mathrm{F},\,\mathrm{r}} \leq C_2\delta\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}$，其中某个绝对常数为 $C_2 > 0$。

证明。取 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{4n\delta C_l}$奇异值分解$_r(\mathbf{G})$。首先考虑 $\mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M})$。从方程（20）继续，

$$
\begin{aligned}
&\mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}^*\rangle)\middle|\mathbf{X}_i\right] - \mathbb{E}\left[\rho_{H,\delta}(\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle)\middle|\mathbf{X}_i\right] \\
&= 2\int_{\langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle}^{\langle \mathbf{X}_i, \mathbf{M}_1-\mathbf{M}^*\rangle} [H_\xi(y+\delta) - H_\xi(y-\delta)]y\,dy \\
&\leq 4\delta b_1^{-1}\left|\int_{\langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle}^{\langle \mathbf{X}_i, \mathbf{M}_1-\mathbf{M}^*\rangle} y\,dy\right| \\
&= 2\delta b_1^{-1}\left|\langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}\rangle^2 + 2\langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}\rangle\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^*\rangle\right|.
\end{aligned}
$$

然后对两边关于 $\mathbf{X}_i$ 取期望，并对 i求和：

$$
\mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) \leq 2\delta n b_1^{-1} C_u[\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 2\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}],
$$

这使用了关于感知算子的假设。然后利用事件 $\mathbf{E}$，$f(\mathbf{M}_1) - f(\mathbf{M})$ 可由以下式子界定：

$$
\begin{aligned}
&f(\mathbf{M}_1) - f(\mathbf{M}) \\
&\leq \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1\delta\sqrt{nd_1rC_u}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\
&\leq 2\delta n b_1^{-1} C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 4\delta n b_1^{-1} C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} + C_1\delta\sqrt{nd_1rC_u}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\
&\leq 2\delta n b_1^{-1} C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + (C_1+4)\delta n b_1^{-1} C_u\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}},
\end{aligned}
$$

其中，最后一个不等式使用了 $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{\mathrm{stat}} = c\sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_0 \geq c\sqrt{\frac{d_1 r}{n} \cdot \frac{1}{C_u}} b_1$。代入 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{4n\delta C_u}$奇异值分解$_r(\mathbf{G})$，得到

$$
f(\mathbf{M} + \frac{b_1}{4n\delta C_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \leq \frac{b_1}{8n\delta C_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2 + \frac{C_1+4}{4}\|\mathbf{G}\|_{\mathrm{F,r}}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}. \tag{21}
$$

另一方面，根据次梯度的定义，我们有

$$
f(\mathbf{M} + \frac{b_1}{4n\delta C_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \geq \langle \mathbf{G}, \frac{b_1}{4n\delta C_u}\mathrm{SVD}_r(\mathbf{G})\rangle = \frac{b_1}{4n\delta C_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2. \tag{22}
$$

Combine Equation (21) with Equation (22) and then the quadratic inequality of $\|\mathbf{G}\|_{\mathrm{F,r}}$ gives

$$
\|\mathbf{G}\|_{\mathrm{F,r}} \leq C_2\delta n b_1^{-1} C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}.
$$

□

D.4 定理 5的证明

The two-phase regularity of the quantile loss is provided by the following lemma.

引理 10. 假设 *4* 成立。存在绝对常数 C_1, C_2, C_3, $c_1 > 0$ 使得若 $n \geq C_1 r d_1 \max\{\delta^2, (1-\delta)^2\} C_u C_l^{-1}$ *abd* $\mathbb{B}_1$, $\mathbb{B}_2$ 由下式给出

$$\mathbb{B}_1 := \left\{ \mathbf{M} \in \mathbb{M}_r : \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq 8\sqrt{C_l^{-1}}\gamma \right\},$$
$$\mathbb{B}_2 := \left\{ \mathbf{M} \in \mathbb{M}_r : C_2 \max\{\delta, 1-\delta\} b_0 \Big(\frac{C_u}{C_l^2} \cdot \frac{r d_1}{n}\Big)^{1/2} \leq \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} < 8\sqrt{C_l^{-1}}\gamma \right\},$$

则以关于 $1-\exp(-c_1 r d_1) - 3\exp(-\sqrt{n}/\log\, n)$的概率，分位数损失 $f(\mathbf{M}) = \sum_{i=1}^n \rho_{Q,\,\delta}(Y_i - \langle \mathbf{M},\ \mathbf{X}_i \rangle)$ 满足条件 *1*:

(1) 关于 $\mathbf{M}^*$ 的秩-r 限制两阶段尖锐性,

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \frac{n}{8}\sqrt{C_l}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ \frac{n}{24 b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, & for \quad \mathbf{M} \in \mathbb{B}_2; \end{cases}$$

(2) 关于 $\mathbf{M}^*$ 的秩-r 限制两阶段次梯度界,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} n\sqrt{C_u}, & for \quad \mathbf{M} \in \mathbb{B}_1; \\ C_3 n b_1^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & for \quad \mathbf{M} \in \mathbb{B}_2, \end{cases}$$

其中 $\mathbf{G} \in \partial f(\mathbf{M})$ 是任意次梯度。

结合命题 1 与引理 10，我们得到 RsGrad 算法 2 在分位数损失下的收敛动力学和统计精度，即定理 5。

引理 *10* 的证明。注意，分位数损失函数 $\rho_{Q,\delta}(\cdot)$ 是利普希茨连续的，其最大值为 $\{\delta, 1-\delta\}$，并满足 $f(x_1 + x_2) \leq f(x_1) + f(x_2)$。

假设事件 $\mathrm{E} = \{\sup_{\Delta\mathbf{M} \in \mathbb{M}_{2r}} |f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M}) - \mathbb{E}(f(\mathbf{M}+\Delta\mathbf{M}) - f(\mathbf{M}))| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1} \leq_{\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}} C_2 \max\{\delta, 1-\delta\}\sqrt{n d_1 r C_u}\}$ 成立。具体而言，定理 3 证明了

$$\mathbb{P}(\boldsymbol{\mathcal{E}}) \geq 1 - \exp(-c d_1 r/4) - 3\exp(-\sqrt{n}/\log n).$$

First consider upper bound of sub-gradient $\mathbf{G} \in f(\mathbf{M})$, where matrix $\mathbf{M}$ has rank at most r. By

event $\boldsymbol{\mathcal{E}}$, we have

$$
\begin{aligned}
& f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M}) \\
\le & \mathbb{E}f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-\mathbb{E}f(\mathbf{M})+C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathrm{SVD}_r(\mathbf{G})\|_{\mathrm{F}} \\
= & \sum_{i=1}^{n}\mathbb{E}\left[\rho_{Q,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}+\mathrm{SVD}_r(\mathbf{G})\rangle)-\rho_{Q,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}\rangle)\right]+C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{G}\|_{\mathrm{F,r}} \\
\le & \sum_{i=1}^{n}\mathbb{E}\rho_{Q,\delta}(-\langle\mathbf{X}_i,\mathrm{SVD}_r(\mathbf{G})\rangle)+C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{G}\|_{\mathrm{F,r}} \\
\le & \sqrt{\frac{1}{2\pi}}n\sqrt{C_u}\|\mathbf{G}\|_{\mathrm{F,r}}+C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{G}\|_{\mathrm{F,r}} \\
\le & n\sqrt{C_u}\|\mathbf{G}\|_{\mathrm{F,r}},
\end{aligned}
$$

where the last line uses $n\ge cd_1r\cdot\max\{\delta^2,(1-\delta)^2\}$. On the other hand, by definition of sub-gradient, we have

$$
f(\mathbf{M}+\mathrm{SVD}_r(\mathbf{G}))-f(\mathbf{M})\ge\langle\mathbf{G},\mathrm{SVD}_r(\mathbf{G})\rangle=\|\mathbf{G}\|_{\mathrm{F,r}}^2.
$$

Then combine the above two equations and it leads to

$$
\|\mathbf{G}\|_{\mathrm{F,r}}\le n\sqrt{C_u}.
$$

Then consider the lower bound of $f(\mathbf{M})-f(\mathbf{M}^*)$.

情况 1：当 $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\ge 8\sqrt{C_l^{-1}}\gamma$ 时。注意，分位数损失函数满足三角不等式，$\rho_{Q,\delta}(x_1+x_2)\le\rho_{Q,\delta}(x_1)+\rho_{Q,\delta}(x_2)$，由此我们有

$$
\begin{aligned}
& f(\mathbf{M})-f(\mathbf{M}^*) \\
\ge & \mathbb{E}f(\mathbf{M})-\mathbb{E}f(\mathbf{M}^*)-C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
= & \sum_{i=1}^{n}\mathbb{E}\left[\rho_{Q,\delta}(\xi_i-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle)-\rho_{Q,\delta}(\xi_i)\right]-C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
\ge & \sum_{i=1}^{n}\mathbb{E}\rho_{Q,\delta}(-\langle\mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle)-\sum_{i=1}^{n}\mathbb{E}\left[\rho_{Q,\delta}(\xi_i)+\rho_{Q,\delta}(-\xi_i)\right]-C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \\
\ge & \sqrt{\frac{1}{2\pi}}n\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}-n\gamma-C_2\max\{\delta,1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}.
\end{aligned}
$$

因此，在条件 $\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}\ge 8\sqrt{C_l^{-1}}\gamma$ 且 $n\ge Cd_1rC_uC_l^{-1}\cdot\max\{\delta^2,(1-\delta)^2\}$ 下，有

$$
f(\mathbf{M})-f(\mathbf{M}^*)\ge\frac{n}{8}\sqrt{C_l}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}},
$$

这验证了 $\tau_{\mathsf{stat}}=8\sqrt{C_l^{-1}}\gamma$、$\mu_{\mathsf{comp}}=\frac{1}{8}n\sqrt{C_l}$。

Case 2: Then consider the second phase where $c\sqrt{\frac{d_1 r}{n}}b_0 \le \|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \le 8\sqrt{C_l^{-1}}\gamma$. Notice for all fixed $t_0 \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\rho_{Q,\delta}(\xi - t_0) &= \delta\int_{s\ge t_0}(s-t_0)\,dH_\xi(s) + (1-\delta)\int_{s<t_0}(t_0-s)\,dH_\xi(s)\\ &= \int_{s\ge t_0}(s-t_0)\,dH_\xi(s) + (1-\delta)\int_{-\infty}^{+\infty}(t_0-s)\,dH_\xi(s)\\ &= \int_{s\ge t_0}(1-H_\xi(s))\,ds + (1-\delta)t_0 - (1-\delta)\int_{-\infty}^{\infty} s\,dH_\xi(s),\end{aligned}$$

其中最后一行与绝对损失情形相同。令 $t_0 = \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle$，则可得

$$\begin{aligned}&\mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle)\middle|\mathbf{X}_i\right]\\ =&\int_{s\ge\langle \mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle}(1-H_\xi(s))\,ds + (1-\delta)\langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle - (1-\delta)\int_{-\infty}^{+\infty} s\,dH_\xi(s).\end{aligned}$$

取 $t_0 = 0$，则可得

$$\mathbb{E}\left[\rho_{Q,\delta}(\xi)\middle|\mathbf{X}_i\right] = \int_{s\ge 0}(1-H_\xi(s))\,ds - (1-\delta)\int_{-\infty}^{+\infty} s\,dH_\xi(s).$$

上述两个公式推出

$$\mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle)\middle|\mathbf{X}_i\right] - \mathbb{E}\left[\rho_{Q,\delta}(\xi)\middle|\mathbf{X}_i\right] = \int_0^{\langle \mathbf{X}_i,\mathbf{M}-\mathbf{M}^*\rangle} H_\xi(s)\,ds - \delta\langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle. \tag{23}$$

注意，$\langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle \sim N(0, \mathrm{vec}(\mathbf{M}-\mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M}-\mathbf{M}^*))$。记 $z_i := \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle$，且$\sigma_i^2 :=$ $\mathrm{vec}(\mathbf{M}-\mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M}-\mathbf{M}^*)$。令 $f_{z_i}(\cdot)$ 为 z_i 的密度。对公式（23）两边关于 $\mathbf{X}_i$ 取期望

$$\begin{aligned}\mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle) - \rho_{Q,\delta}(\xi)\right] &= \int_{-\infty}^{+\infty}\int_0^t (H_\xi(s)-\delta)\, f_{z_i}(t)\,ds\,dt\\ &= \int_{-\infty}^{+\infty}\int_0^t\int_0^s h_\xi(w) f_{z_i}(t)\,dw\,ds\,dt\end{aligned}$$

Notice that sensing matrices conditions imply $\sigma_i \le \sqrt{C_u}\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}} \le 8\sqrt{C_u C_l^{-1}}\gamma$. Then when $w \in (-\sigma_i, \sigma_i)$, under Assumption 4, it has $h_\xi(w) \ge b_0^{-1}$. Thus it has the following upper bound,

$$\begin{aligned}\mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}-\mathbf{M}^*\rangle) - \rho_{Q,\delta}(\xi)\right] &\ge b_0^{-1}\int_{-\sigma_i}^{\sigma_i}\int_0^t f_{z_i}(t) s\,ds\,dt\\ &= \frac{1}{2}b_0^{-1}\int_{-\sigma_i}^{\sigma_i} t^2 f_{z_i}(t)\,dt = \frac{1}{2}b_0^{-1}\sigma_i^2\int_{-1}^{1} t^2\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt\\ &\ge \frac{1}{12b_0}\sigma_i^2,\end{aligned}$$

and the last line is from $\int_{-1}^{1} t^2 \cdot \frac{1}{\sqrt{2}} e^{-t^2/2} dt \geq 1/6$. Therefore,

$$\mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*) = \sum_{i=1}^{n} \mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle) - \rho_{Q,\delta}(\xi)\right] \geq \frac{1}{12b_0} \sum_{i=1}^{n} \sigma_i^2 \geq \frac{n}{12b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2.$$

Invoke event $\boldsymbol{\mathcal{E}}$,

$$\begin{aligned} f(\mathbf{M}) - f(\mathbf{M}^*) &\geq \mathbb{E}f(\mathbf{M}) - \mathbb{E}f(\mathbf{M}^*) - C_2 \max\{\delta, 1-\delta\} \sqrt{n d_1 r C_u} \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\ &\geq \frac{n}{12b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2 - C_2 \max\{\delta, 1-\delta\} \sqrt{n d_1 r C_u} \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \\ &\geq \frac{n}{24b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, \end{aligned}$$

其中最后一个不等式利用了 $\|\mathbf{X} - \mathbf{X}^*\|_{\mathrm{F}} \geq c\max\{\delta, 1-\delta\} \sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_0$。这就证明了第二阶段中的下界，并表明 $\mu_{\mathsf{stat}} = \frac{n}{24b_0} C_l$。

最后，根据下面的引理可知，在事件 E 下，$L_{\mathsf{stat}} \leq C_2 n b_1^{-1}$。至此，引理的证明完成。

□

引理 11（子梯度上界）。设 $\mathbf{M} \in \mathbb{M}_r$ 满足 $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq \tau_{\mathsf{stat}} := c\max\{\delta, 1-\delta\} \sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_1$。令 $\mathbf{G} \in \partial f(\mathbf{M})$ 为次梯度。在事件 E 下，我们有 $\|\mathbf{G}\|_{\mathrm{F},\, \mathrm{r}} \leq C n b_1^{-1} \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}$，其中 $C > 0$ 为某个绝对常数。

证明。取 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{2n}\mathbf{G}$ 的奇异值分解 $r(\mathbf{G})$，其中 $\mathsf{SVD}_r(\mathbf{G})$ 是 $\mathbf{G}$ 的最佳秩 r 近似，因此 $\mathsf{rank}(\mathbf{M}_1) \leq 2r$。首先考虑 $\mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M})$。使用上述证明中的相同记号，类似地，

$$\begin{aligned} &\mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}^* \rangle) - \rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle) \middle| \mathbf{X}_i\right] \\ =& \int_{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}^{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle} \int_0^{\xi} h_\xi(x)\, dx\, d\xi \\ \leq& b_1^{-1} \left| \int_{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}^{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle} \xi\, d\xi \right| \\ =& \frac{1}{2b_1} \left| (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle)^2 - (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle)^2 \right|, \end{aligned}$$

where the inequality follows from the upper bound for $h_\xi(x) \leq b_1^{-1}$. Then take expectation over $\mathbf{X}_i$ on each side and sum up over $i = 1, \ldots, n$,

$$\begin{aligned} \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) =& \sum_{i=1}^{n} \mathbb{E}\left[\rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M}^* \rangle) - \rho_{Q,\delta}(\xi - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle)\right] \\ \leq& \frac{1}{2b_1} \sum_{i=1}^{n} \mathbb{E}\left| (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle + \langle \mathbf{X}_i, \mathbf{M}_1 - \mathbf{M} \rangle)^2 - (\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle)^2 \right| \\ \leq& \frac{n}{2b_1} C_u \left| \|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + 2\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \right|. \end{aligned}$$

Bound $f(\mathbf{M}_1) - f(\mathbf{M})$ under event $\boldsymbol{\mathcal{E}}$:

$$\begin{aligned} f(\mathbf{M}_1) - f(\mathbf{M}) \leq & \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1 \max\{\delta, 1-\delta\}\sqrt{nd_1rC_u}\|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\ \leq & \mathbb{E}f(\mathbf{M}_1) - \mathbb{E}f(\mathbf{M}) + C_1 n b_1^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}} \\ \leq & \frac{n}{2b_1} C_u \|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}^2 + (0.5 + C_1) n b_1^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \|\mathbf{M}_1 - \mathbf{M}\|_{\mathrm{F}}, \end{aligned}$$

其中，第二个不等式源于条件 $\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq c\max\{\delta, 1-\delta\}\sqrt{\frac{d_1 r}{n} \cdot \frac{C_u}{C_l^2}} b_1 \geq c \max\{\delta, 1-\delta\}\sqrt{\frac{d_1 r}{n} \cdot \frac{1}{C_l}} b_1$。由于 $\mathbf{M}_1 = \mathbf{M} + \frac{b_1}{2nC_u}$奇异值分解$_r(\mathbf{G})$，因此可得

$$\begin{aligned} & f(\mathbf{M} + \frac{b_1}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \\ & \qquad \leq \frac{b_1}{8nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2 + (1/2 + C_1)\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}\|\mathbf{G}\|_{\mathrm{F,r}}. \end{aligned} \tag{24}$$

另一方面，根据次梯度的定义，$f(\mathbf{M}_1) - f(\mathbf{M})$ 有下界 nd

$$f(\mathbf{M} + \frac{b_1}{2nC_u}\mathrm{SVD}_r(\mathbf{G})) - f(\mathbf{M}) \geq \frac{b_1^2}{2nC_u}\|\mathbf{G}\|_{\mathrm{F,r}}^2. \tag{25}$$

合并公式（24）、公式（25），然后求解二次不等式，可得

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq Cnb_1^{-1}C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}.$$

□

E 初始化的证明

定理 *6* 的证明。首先，考虑

$$\begin{aligned} & \|\frac{1}{n}\sum_{i=1}^n Y_i \mathrm{mat}(\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\| \\ & \quad = \|\frac{1}{n}\sum_{i=1}^n (\xi_i + \langle \mathbf{X}_i, \mathbf{M}^*\rangle)\,\mathrm{mat}(\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\| \\ & \quad \leq \underbrace{\|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{M}^*\rangle \mathrm{mat}(\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\|}_{B_1} + \underbrace{\|\frac{1}{n}\sum_{i=1}^n \xi_i \mathrm{mat}(\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)\|}_{B_2}. \end{aligned}$$

Analysis of B_1 First show $\langle \mathbf{X}_i, \mathbf{M}^*\rangle \mathrm{mat}(\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))$ is an unbiased estimator of $\mathbf{M}^*$. Or equivalently, $\langle \mathbf{X}_i, \mathbf{M}^*\rangle \boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)$ is an unbiased estimator of $\mathrm{vec}(\mathbf{M}^*)$,

$$\begin{aligned} \mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^*\rangle \boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i) &= \mathbb{E}\mathrm{vec}(\mathbf{X}_i)^\top \mathrm{vec}(\mathbf{M}^*)\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i) \\ &= \mathbb{E}\boldsymbol{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^\top \mathrm{vec}(\mathbf{M}^*) \\ &= \mathrm{vec}(\mathbf{M}^*), \end{aligned}$$

where the second line switches vector and scalar. Then we shall use Bernstein Theorem 10 to bound B_1. Note that

$$\langle \mathbf{X}_i, \mathbf{M}^* \rangle \sim N(0, \mathrm{vec}(\mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M}^*)), \quad \mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i) \sim N(0, \mathbf{\Sigma}_i^{-1}), \tag{26}$$

and it implies

$$\begin{aligned}
\left\| \|\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\| \right\|_{\Psi_1} &\le \|\langle \mathbf{X}_i, \mathbf{M}^* \rangle\|_{\Psi_2} \left\| \|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\| \right\|_{\Psi_2} \\
&\le c\sqrt{\mathrm{vec}(\mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M}^*)} \cdot \|\mathbf{\Sigma}_i^{-\frac{1}{2}}\| \\
&\le c\sqrt{C_u C_l^{-1}} \|\mathbf{M}^*\|_{\mathrm{F}}.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
&\|\mathbb{E}\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)^\top\| \\
\le\ &\mathbb{E}\|\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)^\top\| \\
\le\ &\|\mathbf{M}^*\|^2 + 2\|\mathbf{M}^*\|\mathbb{E}\|\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\| + \mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^2 \|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^2 \\
\le\ &2\|\mathbf{M}^*\|^2 + 2\mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^2 \|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^2.
\end{aligned}$$

注意， $\|\mathbf{M}^*\| = \sigma_1$ ，并且根据 $ab \le \frac{a^2}{2c} + \frac{cb^2}{2}$ ，我们有

$$\begin{aligned}
&\mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^2 \|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^2 \\
\le\ &\frac{d_1 C_l^{-1}}{2C_u\|\mathbf{M}^*\|_{\mathrm{F}}^2}\mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^4 + \frac{C_u\|\mathbf{M}^*\|_{\mathrm{F}}^2}{2d_1 C_l^{-1}}\mathbb{E}\|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^4
\end{aligned}$$

Equation (26) suggests $\mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^4 \le 3C_u^2\|\mathbf{M}^*\|_{\mathrm{F}}^4$ and $\mathbb{E}\|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^4 \le cC_l^{-2}d_1^2$. Therefore we have

$$\mathbb{E}\langle \mathbf{X}_i, \mathbf{M}^* \rangle^2 \|\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\|^2 \le \tilde{c}d_1 C_u C_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2.$$

Thus, we have

$$\begin{aligned}
&\|\mathbb{E}\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)^\top\| \\
\le\ &2\|\mathbf{M}^*\|_{\mathrm{F}}^2 + 2\tilde{c}d_1 C_u C_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2,
\end{aligned}$$

and it verifies

$$\begin{aligned}
&\Lambda_{\max}\left(\sum_{i=1}^n \left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)\left(\langle \mathbf{X}_i, \mathbf{M}^* \rangle \mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i)) - \mathbf{M}^*\right)^\top\right)/n \\
\le\ &2\|\mathbf{M}^*\|_{\mathrm{F}}^2 + 2\tilde{c}d_1 C_u C_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2 \le cd_1 C_u C_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2
\end{aligned}$$

Similarly, we have

$$\Lambda_{\max}\left(\sum_{i=1}^{n}\left(\langle\mathbf{X}_i,\mathbf{M}^*\rangle\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))-\mathbf{M}^*\right)^\top\left(\langle\mathbf{X}_i,\mathbf{M}^*\rangle\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))-\mathbf{M}^*\right)\right)/n$$
$$\leq 2\|\mathbf{M}^*\|_{\mathrm{F}}^2+2\tilde{c}d_1C_uC_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2\leq cd_1C_uC_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2$$

Thus, Bernstein Theorem 10 implies the following equation

$$B_1\leq c_1\sqrt{d_1C_uC_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2}\sqrt{\frac{\log d_1}{n}}+c_1\frac{\log d_1}{n}\log\left(\frac{c\sqrt{C_uC_l^{-1}}\|\mathbf{M}^*\|_{\mathrm{F}}}{\sqrt{cd_1C_uC_l^{-1}\|\mathbf{M}^*\|_{\mathrm{F}}^2}}\right)$$
$$\leq c_0\sqrt{C_uC_l^{-1}}\sqrt{\frac{d_1r\log d_1}{n}}\sigma_1$$

以$1-c_1d_1^{-1}$为概率成立。

Analysis of B_2 注意到，给定$\{\xi_i\}_{i=1}^n$，以超过 $1-c_2\exp(-d_1)$ 的概率，我们有 e,

$$B_2\leq c_3\sqrt{C_l^{-1}}\frac{\sqrt{d_1}}{n}\left(\sum_{i=1}^{n}\xi_i^2\right)^{1/2}.$$

Then by Lemma 17，我们以概率大于 $1-c_3(\log d_1)^{-1}$，得到以下结论 s

$$\sum_{i=1}^{n}\xi_i^2\leq(2n\gamma_1\log d_1)^{\frac{2}{1+\varepsilon}}.$$

Thus, we have the upper bound for B_2,

$$B_2\leq c_4\sqrt{C_l^{-1}}\frac{\sqrt{d_1}}{n^{\frac{\varepsilon}{1+\varepsilon}}}(\log d_1)^{\frac{1}{1+\varepsilon}}\gamma_1^{\frac{1}{1+\varepsilon}}.$$

Combine B_1 and B_2 and then we obtain

$$\|\frac{1}{n}\sum_{i=1}^{n}Y_i\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))-\mathbf{M}^*\|\leq c_0\sqrt{C_uC_l^{-1}}\sqrt{\frac{d_1r\log d_1}{n}}\sigma_1+c_4\sqrt{C_l^{-1}}\frac{\sqrt{d_1}}{n^{\frac{\varepsilon}{1+\varepsilon}}}(\log d_1)^{\frac{1}{1+\varepsilon}}\gamma_1^{\frac{1}{1+\varepsilon}}$$

holds with probability exceeding $1-c_1d_1^{-1}-c_2\exp(-d_1)-c_3(\log d_1)^{-1}$. Then with perturbation theories Lemma 14 and sample size condition

$$n\geq C\max\left\{C_uC_l^{-1}\kappa^2d_1r^2\log d_1,\left(C_l^{-1}d_1r\right)^{\frac{1+\varepsilon}{2\varepsilon}}\cdot\gamma_1^{\frac{1}{\varepsilon}}\sigma_r^{-\frac{1+\varepsilon}{\varepsilon}}(\log d_1)^{\frac{1}{\varepsilon}}\right\},$$

we have

$$\|\mathbf{M}_0-\mathbf{M}^*\|=\|\mathrm{SVD}_r\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\mathrm{mat}(\mathbf{\Sigma}_i^{-1}\mathrm{vec}(\mathbf{X}_i))\right)-\mathbf{M}^*\|$$
$$\leq\tilde{c}_0\sqrt{C_uC_l^{-1}}\sqrt{\frac{d_1r\log d_1}{n}}\sigma_1+\tilde{c}_4\sqrt{C_l^{-1}}\frac{\sqrt{d_1}}{n^{\frac{\varepsilon}{1+\varepsilon}}}(\log d_1)^{\frac{1}{1+\varepsilon}}\gamma_1^{\frac{1}{1+\varepsilon}},$$

which implies

$$\|\mathbf{M}_0 - \mathbf{M}^*\|_{\mathrm{F}} \leq \sqrt{2r}\|\mathbf{M}_0 - \mathbf{M}^*\| \leq c\sigma_r,$$

and it completes the proof. □

F Technical Lemma

The following norm relations are used as facts,

$$\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|, \quad |\langle \mathbf{C}, \mathbf{D}\rangle| \leq \|\mathbf{C}\|_{\mathrm{F}}\|\mathbf{D}\|_{\mathrm{F}},$$

其中，A、B、C、D 均为维度适当的矩阵。

Lemma 12 (Partial Frobenius Norm, Lemma 15 of Tong et al. (2021a))**.** *Define partial Frobenius norm of matrix* $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ *to be*

$$\|\mathbf{M}\|_{\mathrm{F,r}} := \sqrt{\sum_{i=1}^{r} \sigma_i^2(\mathbf{M})} = \|\mathrm{SVD}_r(\mathbf{M})\|_{\mathrm{F}}.$$

For any $\mathbf{M}, \hat{\mathbf{M}} \in \mathbb{R}^{d_1 \times d_2}$ *with* $\mathrm{rank}(\hat{\mathbf{M}}) \leq r$*, one has*

$$|\langle \mathbf{M}, \hat{\mathbf{M}}\rangle| \leq \|\mathbf{M}\|_{\mathrm{F,r}}\|\hat{\mathbf{M}}\|_{\mathrm{F}}.$$

对于任意的 $\mathrm{M} \in \mathbb{R}^{d_1 \times d_2}$、$\mathrm{V} \in \mathbb{R}^{d_2 \times r}$，均有

$$\|\mathbf{MV}\|_{\mathrm{F}} \leq \|\mathbf{M}\|_{\mathrm{F,r}}\|\mathbf{V}\|.$$

Note that in proofs of main theorems, we actually need to bound $\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}$ rather than the term $\|\mathbf{G}_l\|_{\mathrm{F,r}}$. While, $\|\mathbf{G}_l\|_{\mathrm{F,r}}$ provides a satisfactory upper bound for $\|\mathcal{P}_{\mathbb{T}_l}(\mathbf{G}_l)\|_{\mathrm{F}}$, see the followed lemma.

Lemma 13. *(Frobenius norm of projected sub-gradient) Let* $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ *be a matrix with rank at most* r*. Suppose* $\mathbf{G} \in \partial f(\mathbf{M})$ *is the sub-gradient at* $\mathbf{M}$ *and* $\mathbb{T}$ *is the tangent space for* $\mathbb{M}_r$ *at* $\mathbf{M}$*. Then we have*

$$\|\mathcal{P}_{\mathbb{T}}(\mathbf{G})\|_{\mathrm{F}} \leq \sqrt{2}\|\mathbf{G}\|_{\mathrm{F,r}}.$$

Proof. Let $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be SVD of $\mathbf{M}$ with orthogonal matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$. Notice that

$$\|\mathcal{P}_{\mathbb{T}}(\mathbf{G})\|_{\mathrm{F}}^2 = \|\mathbf{U}\mathbf{U}^T\mathbf{G}\|_{\mathrm{F}}^2 + \|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{G}\mathbf{V}\mathbf{V}^T\|_{\mathrm{F}}^2,$$

and $\operatorname{rank}(\mathbf{U}\mathbf{U}^\top) = \operatorname{rank}(\mathbf{V}\mathbf{V}^\top) = r$. Then by properties of partial Frobenius norm Lemma 12, one has

$$\|\mathbf{U}\mathbf{U}^T\mathbf{G}\|_{\mathrm{F}} = \|\mathbf{U}^T\mathbf{G}\|_{\mathrm{F}} \le \|\mathbf{G}\|_{\mathrm{F,r}}\|\mathbf{U}\| = \|\mathbf{G}\|_{\mathrm{F,r}},$$
$$\|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{G}\mathbf{V}\mathbf{V}^T\|_{\mathrm{F}} \le \|\mathbf{G}\|_{\mathrm{F,r}},$$

这意味着 $\|\mathcal{P}_{\mathbb{T}}(\mathbf{G})\|_{\mathrm{F}}^2 \le 2\|\mathbf{G}\|_{\mathrm{F,r}}^2$。 □

以下 引理对于证明黎曼优化的收敛性至关重要 .

引理 14.（矩阵扰动）设矩阵 $\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$ 的秩为 r，且其奇异值分解为 $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$，其中 $\mathbf{\Sigma} = \operatorname{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ 且 $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$。那么，对于任意满足 $\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} < \sigma_r/8$ 的 $\hat{\mathbf{M}} \in \mathbb{R}^{d\times d}$，若 $\hat{\mathbf{U}} \in \mathbb{R}^{d_1\times r}$ 和 $\hat{\mathbf{V}} \in \mathbb{R}^{d_2\times r}$ 分别是 r 个最大奇异值对应的左、右奇异向量，则有

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \le \frac{8}{\sigma_r}\|\hat{\mathbf{M}} - \mathbf{M}\|, \quad \|\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\| \le \frac{8}{\sigma_r}\|\hat{\mathbf{M}} - \mathbf{M}\|,$$
$$\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\| \le \|\hat{\mathbf{M}} - \mathbf{M}\| + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|^2}{\sigma_r},$$
$$\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}} \le \|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}}{\sigma_r}.$$

证明。参见第 G节。 □

引理 15. 设随机矩阵 $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ 服从均值为零的多元高斯分布 $N(\mathbf{0}, \mathbf{\Sigma})$，其中 $\mathbf{\Sigma} \in \mathbb{R}^{d_1d_2\times d_1d_2}$ 满足

$$\lambda_{\max}(\mathbf{\Sigma}) \le C_u.$$

Then its operator norm could be upper bounded by

$$\mathbb{P}\left(\|\mathbf{X}\| \ge t + c\sqrt{C_u(d_1 + d_2)}\right) \le \exp(-t^2),$$

where $c > 0$ is some constant. It implies $\|\|\mathbf{X}\|\|_{\Psi_2} \le c_1\sqrt{d_1C_u}$ and $\|\|\mathbf{X}\|\|_{\Psi_1} \le c_2\sqrt{d_1C_u}$, with some constants $c_1, c_2 > 0$.

Proof. The proof follows ε-net arguments. Corollary 4.2.13 of work Vershynin (2018) shows that take $\varepsilon = \frac{1}{4}$ and then there exists an ε-net $\mathcal{N}$ of sphere $\mathbb{S}^{d_1-1}$ and an ε-net $\mathcal{M}$ of sphere $\mathbb{S}^{d_2-1}$ with cardinalities

$$|\mathcal{N}| \le 9^{d_1}, \quad |\mathcal{M}| \le 9^{d_2}.$$

For any fixed unit vector $\mathbf{u} \in \mathcal{N}$, $\mathbf{v} \in \mathcal{M}$, consider $\mathbf{u}^\top \mathbf{X} \mathbf{v}$,

$$\mathbf{u}^\top \mathbf{X} \mathbf{v} = \operatorname{vec}(\mathbf{X})^\top (\mathbf{u} \otimes \mathbf{v}) \sim N(0, (\mathbf{u} \otimes \mathbf{v})^\top \mathbf{\Sigma} (\mathbf{u} \otimes \mathbf{v})),$$

其中 $\otimes$ 是克罗内克积。此外，它具有 $\|\mathbf{u} \otimes \mathbf{v}\|_2 = 1$ 和 $(\mathbf{u} \otimes \mathbf{v})^\top \mathbf{\Sigma} (\mathbf{u} \otimes \mathbf{v}) \leq C_u$。因此，对于任意 $t > 1$,

$$\mathbb{P}\left(\mathbf{u}^\top \mathbf{X} \mathbf{v} \geq t\right) \leq \exp\left(-\frac{t^2}{2C_u}\right).$$

对 $\in \mathcal{N}$ 和 $\in \mathcal{M}$ 取并集，然后得到的是

$$\mathbb{P}\left(\max_{\mathbf{u} \in \mathcal{N}, \mathbf{v} \in \mathcal{M}} \mathbf{u}^\top \mathbf{X} \mathbf{v} \geq t\right) \leq 9^{d_1 + d_2} \exp\left(-\frac{t^2}{2C_u}\right).$$

另一方面，根据 Vershynin （2018）的定理 4.4.5，我们有 $\|\mathbf{X}\| \leq 2 \max_{\mathbf{u} \in \mathcal{N},\ \mathbf{v} \in \mathcal{M}} \mathbf{u}^\top \mathbf{X} \mathbf{v}$，由此我们得到,

$$\mathbb{P}\left(\|\mathbf{X}\| \geq 2t\right) \leq \mathbb{P}\left(\max_{\mathbf{u} \in \mathcal{N}, \mathbf{v} \in \mathcal{M}} \mathbf{u}^\top \mathbf{X} \mathbf{v} \geq t\right) \leq 9^{d_1 + d_2} \exp\left(-\frac{t^2}{2C_u}\right).$$

It implies $\mathbb{P}\left(\|\mathbf{X}\| \geq t + c\sqrt{C_u(d_1 + d_2)}\right) \leq \exp(-t^2)$ and the proof completes. □

引理 16. （高斯噪声下绝对损失的期望）假设噪声项服从高斯分布 $\xi_1, \cdots, \xi_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ 且高斯感知算子的向量化 $\operatorname{vec}(\mathbf{X}_i)$ 满足假设 *1*。取绝对值损失函数

$$f(\mathbf{M}) = \sum_{i=1}^{n} |Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle|,$$

and then for any matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, *we have*

$$\mathbb{E}[f(\mathbf{M}^*)] = n\sqrt{\frac{2}{\pi}}\sigma,$$

$$n\sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2} \leq \mathbb{E}[f(\mathbf{M})] \leq n\sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2}.$$

Proof. Note that $f(\mathbf{M}) = \sum_{i=1}^n |Y_i - \langle \mathbf{X}_i, \mathbf{M} \rangle| = \sum_{i=1}^n |\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle|$, $f(\mathbf{M}^*) = \sum_{i=1}^n |Y_i - \langle \mathbf{X}_i, \mathbf{M}^* \rangle| = \sum_{i=1}^n |\xi_i|$.

First calculate the conditional expectation $\mathbb{E}[|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \big| \mathbf{X}_i]$. Notice that conditioned on $\mathbf{X}_i$, $\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle$ is normally distributed with mean $-\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle$ and variance σ^2. Then $|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle|$ has a folded normal distribution.

$$\begin{aligned}
&\mathbb{E}\left[|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \big| \mathbf{X}_i\right] \\
&= \sigma\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle^2}{2\sigma^2}\right) - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle \left[1 - 2\phi\left(\frac{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}{\sigma}\right)\right],
\end{aligned}$$

where $\phi(\cdot)$ is the normal cumulative distribution function. For each $i = 1, 2, \cdots, n$, one has

$$\begin{aligned}\mathbb{E}|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| &= \mathbb{E}\left[\mathbb{E}[|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \big| \mathbf{X}_i]\right] \\ &= \sigma\sqrt{\frac{2}{\pi}}\mathbb{E}\left[\exp\left(-\frac{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle^2}{2\sigma^2}\right)\right] \\ &\quad - \mathbb{E}\left[\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle \cdot \left[1 - 2\phi(\frac{\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle}{\sigma})\right]\right].\end{aligned}$$

Notice that $\langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle \sim N(0, \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M} - \mathbf{M}^*))$ and then after integration calculation it is

$$\mathbb{E}|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| = \sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)}.$$

Specifically, when $\mathbf{M} = \mathbf{M}^*$, it leads to $\mathbb{E}|\xi| = \sqrt{\frac{2}{\pi}}\sigma$. Thus, sum over $i = 1, \ldots, n$, we have

$$\begin{aligned}\mathbb{E}f(\mathbf{M}^*) &= n\mathbb{E}|\xi| = n\sqrt{\frac{2}{\pi}}\sigma, \\ \mathbb{E}f(\mathbf{M}) &= \sum_{i=1}^n \mathbb{E}|\xi_i - \langle \mathbf{X}_i, \mathbf{M} - \mathbf{M}^* \rangle| \\ &= \sum_{i=1}^n \sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)^\top \mathbf{\Sigma}_i \mathrm{vec}(\mathbf{M} - \mathbf{M}^*)}.\end{aligned}$$

Then by conditions of sensing operators Assumption 1, we obtain

$$n\sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + C_l\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2} \le \mathbb{E}f(\mathbf{M}) \le n\sqrt{\frac{2}{\pi}}\sqrt{\sigma^2 + C_u\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2}.$$

□

Corollary 2. *Under same conditions of Lemma 16, we have*

$$n\sqrt{\frac{2}{\pi}}\sqrt{C_l}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \le \mathbb{E}\sum_{i=1}^n |\langle \mathbf{M} - \mathbf{M}^*, \mathbf{X}_i \rangle| \le n\sqrt{\frac{2}{\pi}}\sqrt{C_u}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}.$$

Proof. Take noiseless case in Lemma 16 and it yields $\sum_{i=1}^n |\langle \mathbf{M} - \mathbf{M}^*, \mathbf{X}_i \rangle|$. □

命题3。（矩阵恢复中的经验过程）假设 $f(\mathbf{M})$ 定义于公式（4）˜中，感知矩阵满足假设 *1*，且 $\rho(\cdot)$ 是利普希茨连续的，满足 L。那么存在常数 C、$C_1 > 0$，使得，以概率至少为 $1 - \exp\left(-\frac{C^2 d_1 r}{3}\right) - 3\exp\left(-\frac{\sqrt{n}}{\log n}\right)$，

$$\left|f(\mathbf{M} + \Delta\mathbf{M}) - f(\mathbf{M}) - (\mathbb{E}f(\mathbf{M} + \Delta\mathbf{M}) - \mathbb{E}f(\mathbf{M}))\right| \le C\tilde{L}\sqrt{nd_1 r C_u}\|\Delta\mathbf{M}\|_{\mathrm{F}} \tag{27}$$

对所有 $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ 和 $\Delta\mathbf{M} \in \mathbb{M}_r$ 均成立。

Proof. See Section H. □

引理 17。（平方重尾随机变量之和的尾部不等式）假设 ξ、ξ_1、$\cdots$、ξ_n 是独立同分布的随机变量，且存在 $0<\varepsilon\leq 1$ 使其 $1+\varepsilon$ 矩有限，即 $\mathbb{E}|\xi|^{1+\varepsilon}<+\infty$。记 $\|\xi\|_{1+\varepsilon}=\left(\mathbb{E}|\xi|^{1+\varepsilon}\right)^{1/(1+\varepsilon)}$。那么对于任意 $s>0$，平方和的尾部界可由以下上界给出。

$$\mathbb{P}\left(\sum_{i=1}^{n}\xi_i^2>s\right)\leq 2n\|\xi\|_{1+\varepsilon}^{1+\varepsilon}s^{-\frac{1+\varepsilon}{2}}.$$

Proof. For any $s>0$, it has $\mathbb{P}\left(\xi^2\geq s\right)\leq s^{-\frac{1+\varepsilon}{2}}\mathbb{E}|\xi|^{1+\varepsilon}$ which uses Markov's inequality. Introduce truncated random variable $\varphi_i:=\xi_i^2 1_{\{\xi_i^2<s\}}$. Note that $\sum_{i=1}^n\xi_i^2$ could be written into sum of two parts

$$\sum_{i=1}^{n}\xi_i^2=\sum_{i=1}^{n}\varphi_i+\sum_{i=1}^{n}(\xi_i^2-\varphi_i).$$

Hence, we analyze these two parts respectively. First consider probability of $\sum_{i=1}^n\xi_i^2\neq\sum_{i=1}^n\varphi_i$:

$$\mathbb{P}\left(\sum_{i=1}^{n}\xi_i^2\neq\sum_{i=1}^{n}\varphi_i\right)\leq n\mathbb{P}\left(\xi^2>s\right)\leq n\frac{\mathbb{E}|\xi|^{1+\varepsilon}}{s^{\frac{1+\varepsilon}{2}}}=n\|\xi\|_{1+\varepsilon}^{1+\varepsilon}s^{-\frac{1+\varepsilon}{2}}.$$

Then bound tail for sum of φ_i:

$$\mathbb{P}\left(\sum_{i=1}^{n}\varphi_i>s\right)\leq ns^{-1}\mathbb{E}\varphi=ns^{-1}\mathbb{E}\varphi^{\frac{1+\varepsilon}{2}}\varphi^{\frac{1-\varepsilon}{2}}\leq n\|\xi\|_{1+\varepsilon}^{1+\varepsilon}s^{-\frac{1+\varepsilon}{2}}.$$

Finally, combine the above two equations, it has

$$\mathbb{P}\left(\sum_{i=1}^{n}\xi_i^2>s\right)\leq\mathbb{P}\left(\sum_{i=1}^{n}\varphi_i>s\right)+\mathbb{P}\left(\sum_{i=1}^{n}\xi_i^2\neq\sum_{i=1}^{n}\varphi_i\right)\leq 2n\|\xi\|_{1+\varepsilon}^{1+\varepsilon}s^{-\frac{1+\varepsilon}{2}},$$

which completes the proof. □

Theorem 10. *(Bernstein's Inequality, Proposition 2 in Koltchinskii et al. (2011)) Let* $\mathbf{A},\mathbf{A}_1,\cdots,\mathbf{A}_n$ *be independent* $p\times q$ *matrices that satisfy for some* $\alpha\geq 1$ *(and all* i*)*

$$\mathbb{E}\mathbf{A}_i=0,\qquad\|\Lambda_{\max}(\mathbf{A}_i)\|_{\Psi_\alpha}=:K<+\infty.$$

Define

$$S^2:=\max\left\{\Lambda_{\max}\left(\sum_{i=1}^{n}\mathbb{E}\mathbf{A}_i\mathbf{A}_i^\top\right)/n,\Lambda_{\max}\left(\sum_{i=1}^{n}\mathbb{E}\mathbf{A}_i^\top\mathbf{A}_i\right)/n\right\}.$$

那么，对于某个常数 $C>0$，且对于所有 $t>0$,

$$\mathbb{P}\left(\Lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{A}_i\right)/n\geq CS\sqrt{\frac{t+\log(p+q)}{n}}+CK\log^{1/\alpha}\left(\frac{K}{S}\right)\frac{t+\log(p+q)}{n}\right)\leq\exp(-t).$$

G 矩阵扰动引理证明14

To prove Lemma 14, first consider a simpler case, when the matrix and the perturbation are symmetric. The core of the proof is Theorem 1 of work Xia (2021) and it is presented here.

设 $\mathbf{M}$ 和 $\mathbf{Z}$ 为对称 $d \times d$ 矩阵。矩阵 $\mathbf{M}$ 的秩为 $r = \mathsf{rank}(\mathbf{M}) \leq d$，其特征分解为 $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$，其中 $\mathbf{U} \in \mathbb{R}^{d\times r}$ 包含标准正交列，且 $\mathbf{\Lambda} = \mathsf{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r\times r}$ 为对角矩阵。

设 $\mathbf{U}_\perp \in \mathbb{R}^{d\times(d-r)}$ 为 $\mathbf{U}$ 的正交补，使得 $[\mathbf{U}, \mathbf{U}_\perp] \in \mathbb{R}^{d\times d}$ 为标准正交。假设 $\hat{}\mathbf{U} \in \mathbb{R}^{d\times r}$ 包含 $\mathbf{M}+\mathbf{Z}$ 的特征向量，对应于绝对值最大的 r 特征值。

定义以下投影算子

$$\mathfrak{P}^0 := \mathbf{U}_\perp \mathbf{U}_\perp^\top, \quad \mathfrak{P}^{-k} := \mathbf{U}\mathbf{\Lambda}^{-k}\mathbf{U}^\top, \quad \text{for all } k \geq 1, \tag{28}$$

并将 k 阶扰动项记为

$$S_{\mathbf{M},k}(\mathbf{Z}) := \sum_{\mathbf{s}: s_1+\cdots+s_{k+1}=k} (-1)^{1+\tau(\mathbf{s})} \cdot \mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\mathbf{Z}\cdots\mathbf{Z}\mathfrak{P}^{-s_{k+1}}, \tag{29}$$

其中 $\mathbf{s} = (s_1, \ldots, s_{k+1})$ 包含非负整数指标，且 $\tau(\mathbf{s}) := \sum_{j=1}^{k+1} 1_{s_j>0}$。

Theorem 11 (Theorem 1 of Xia (2021)). *If* $\|\mathbf{Z}\| < \frac{\min_{1\leq i\leq r}|\lambda_i|}{2}$, *then it has*

$$\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top = \sum_{k\geq 1} S_{\mathbf{M},k}(\mathbf{Z}).$$

Lemma 18. *(Symmetric Matrix Perturbation) Suppose symmetric matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *has rank* r. *Let* $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ *be its singular value decomposition with* $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$, $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0$. *Then for any matrix* $\hat{\mathbf{M}} \in \mathbb{R}^{d\times d}$ *satisfying* $\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} < |\lambda_r|/8$ *with* $\hat{\mathbf{U}} \in \mathbb{R}^{d\times r}$ *eigenvectors of* r *largest absolute eigenvalues of* $\hat{\mathbf{M}}$, *then we have*

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \leq \frac{8\|\hat{\mathbf{M}} - \mathbf{M}\|}{|\lambda_r|},$$

$$\|\,\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\| \leq \|\hat{\mathbf{M}} - \mathbf{M}\| + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|^2}{|\lambda_r|},$$

$$\|\,\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}} \leq \|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}}{|\lambda_r|}.$$

证明。 记 $\mathbf{Z} = \hat{\mathbf{M}} - \mathbf{M}$。定义 $\mathbf{U}_\perp \in \mathbb{R}^{d\times(d-r)}$ 使得 $[\mathbf{U}, \mathbf{U}_\perp] \in \mathbb{R}^{d\times d}$ 为正交归一的。对于非负整数 $\mathfrak{P}^{-k}$，投影算子 k 以及 k 阶扰动项分别由公式（28）和公式（29）给出。

তাহলে $\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})$-এর নর্ম নিম্নলিখিত উপায়ে সীমাবদ্ধ করা যতে পারে।

$$\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\| \le \binom{2k}{k}\frac{\|\mathbf{Z}\|^k}{|\lambda_r|^k} \le \left(\frac{4\|\mathbf{Z}\|}{|\lambda_r|}\right)^k, \quad \|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\|_{\mathrm{F}} \le \binom{2k}{k}\frac{\|\mathbf{Z}\|_{\mathrm{F}}^k}{|\lambda_r|^k} \le \left(\frac{4\|\mathbf{Z}\|_{\mathrm{F}}}{|\lambda_r|}\right)^k.$$

因此，将以上方程对 k 求和，并根据定理 11，可得

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \le \sum_{k\ge1}\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\| \le \frac{8\|\mathbf{Z}\|}{|\lambda_r|}, \quad \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}} \le \sum_{k\ge1}\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\|_{\mathrm{F}} \le \frac{8\|\mathbf{Z}\|_{\mathrm{F}}}{|\lambda_r|}. \tag{30}$$

Then consider $\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}$:

$$\begin{aligned}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M} &= \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k}(-1)^{1+\tau(\mathbf{s})}\mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\cdots\mathfrak{P}^{-s_k}\mathbf{Z}\mathfrak{P}^{-s_{k+1}}\mathbf{M}\\ &= \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k}(-1)^{1+\tau(\mathbf{s})}\mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\cdots\mathfrak{P}^{-s_k}\mathbf{Z}\mathfrak{P}^{-s_{k+1}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top.\end{aligned}$$

Note that when $s_{k+1}=0$, $\mathfrak{P}^{-s_{k+1}} = \mathbf{U}_\perp\mathbf{U}_\perp^\top$, it has $\mathfrak{P}^{-s_{k+1}}\mathbf{M} = 0$ and thus we have

$$\begin{aligned}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M} &= \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k,\, s_{k+1}>0}(-1)^{1+\tau(\mathbf{s})}\mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\cdots\mathfrak{P}^{-s_k}\mathbf{Z}\mathfrak{P}^{-s_{k+1}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\\ &= \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k,\, s_{k+1}=1}(-1)^{1+\tau(\mathbf{s})}\mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\cdots\mathfrak{P}^{-s_k}\mathbf{Z}\mathbf{U}\mathbf{U}^\top\\ &\quad+\sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k,\, s_{k+1}\ge2}(-1)^{1+\tau(\mathbf{s})}\mathfrak{P}^{-s_1}\mathbf{Z}\mathfrak{P}^{-s_2}\cdots\mathfrak{P}^{-s_k}\mathbf{Z}\mathfrak{P}^{1-s_{k+1}}\end{aligned}$$

The operator norm of $\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\|$ could be bounded with

$$\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\| \le \binom{2k-1}{k}\frac{\|\mathbf{Z}\|^k}{|\lambda_r|^{k-1}} \le \|\mathbf{Z}\|\left(\frac{4\|\mathbf{Z}\|}{|\lambda_r|}\right)^{k-1},$$

which implies

$$\|\sum_{k\ge2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\| \le \frac{8\|\mathbf{Z}\|^2}{|\lambda_r|}.$$

Specifically, when $k=1$, it is $\mathcal{S}_{\mathbf{M},1}(\mathbf{Z})\mathbf{M} = \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top$. Similarly, $\|\sum_{k\ge2}\mathbf{M}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\| \le \frac{8\|\mathbf{Z}\|^2}{|\lambda_r|}$ holds. In parallel, Frobenius norm of $\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}$ has

$$\|\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\|_{\mathrm{F}} \le \binom{2k-1}{k}\frac{\|\mathbf{Z}\|^{k-1}}{|\lambda_r|^{k-1}}\|\mathbf{Z}\|_{\mathrm{F}} \le \|\mathbf{Z}\|_{\mathrm{F}}\left(\frac{4\|\mathbf{Z}\|}{|\lambda_r|}\right)^{k-1},$$

$$\|\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\|_{\mathrm{F}} \leq \frac{8\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|}, \quad \|\sum_{k\geq 2}\mathbf{M}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\|_{\mathrm{F}} \leq \frac{8\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|}. \tag{31}$$

Then expand $\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}$,

$$\begin{aligned}
&\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M} \\
&= \hat{\mathbf{U}}\hat{\mathbf{U}}^\top(\mathbf{M}+\mathbf{Z})\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{M} \\
&= \hat{\mathbf{U}}\hat{\mathbf{U}}^\top(\mathbf{M}+\mathbf{Z})\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{M}\mathbf{U}\mathbf{U}^\top \\
&= (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{M}\mathbf{U}\mathbf{U}^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{M}(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top) + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top \\
&= \sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\mathbf{U}\mathbf{U}^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{M}\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z}) + \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top \\
&\quad + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top,
\end{aligned}$$

where the last equation follows from Theorem 11 and $\mathcal{S}_{\mathbf{M},1}(\mathbf{Z})\mathbf{M} = \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top$, $\mathbf{M}\mathcal{S}_{\mathbf{M},1}(\mathbf{Z}) = \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top$. Then consider the latter three terms of the above equations,

$$\begin{aligned}
&\mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top + \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top \\
&\quad + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top) \\
&= \mathbf{Z} - \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top \\
&\quad + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top),
\end{aligned}$$

其中，最后一个等式是由于 $\mathbf{U}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top = \mathbf{I}$ 和 $\mathbf{Z} = (\mathbf{U}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top)\mathbf{Z}(\mathbf{U}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top)$ 得到的。因此，我们得到了 $\mathrm{SVD}_r(\mathbf{M})$ 的展开式，

$$\begin{aligned}
&\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M} \\
&= \sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\mathbf{U}\mathbf{U}^\top + \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{M}\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z}) + \mathbf{Z} - \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top \\
&\quad + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top).
\end{aligned}$$

Take Frobenius norm on each side of the above equation and it leads to

$$\begin{aligned}
&\|\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}} \\
&\leq \underbrace{\|\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}}_{A_1} + \underbrace{\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{M}\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\|_{\mathrm{F}}}_{A_2} + \underbrace{\|\mathbf{Z} - \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top\|_{\mathrm{F}}}_{A_3} \\
&\quad + \underbrace{\|(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top\|_{\mathrm{F}}}_{A_4} + \underbrace{\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}}_{A_5}.
\end{aligned} \tag{32}$$

Note that by Equation (31), we have upper bound of A_1 and A_2,

$$A_1 \vee A_2 \leq \|\sum_{k\geq 2} \mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\mathbf{M}\|_{\mathrm{F}} \vee \|\mathbf{M}\sum_{k\geq 2} \mathcal{S}_{\mathbf{M},k}(\mathbf{Z})\|_{\mathrm{F}} \leq \frac{8\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|}.$$

此外，它还具有 $\mathbf{Z} - \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top = \mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top + \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top + \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top$ 以及 $\|\mathbf{Z}\|_{\mathrm{F}}^2 = \|(\mathbf{U}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top)\mathbf{Z}(\mathbf{U}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top)\|_{\mathrm{F}}^2 = \|\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 + \|\mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 + \|\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top\|_{\mathrm{F}}^2 + \|\mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top\|_{\mathrm{F}}^2$，并由此得到

$$A_3^2 = \|\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 + \|\mathbf{U}_\perp\mathbf{U}_\perp^\top\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 + \|\mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{U}_\perp\mathbf{U}_\perp^\top\|_{\mathrm{F}}^2 \leq \|\mathbf{Z}\|_{\mathrm{F}}^2,$$

which shows $A_3 \leq \|\mathbf{Z}\|_{\mathrm{F}}$. With Equation (30), A_4 could be upper bounded by

$$A_4 \leq \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}\|\mathbf{Z}\| \leq \frac{8\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|}.$$

Note that with triangular inequality and in a similar fashion, it has

$$A_5 \leq \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\mathbf{Z}(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\|_{\mathrm{F}} + \|(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{Z}\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}} \leq \frac{16\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|}.$$

Thus, Equation (32) could be further upper bounded by

$$\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}} \leq \|\mathbf{Z}\|_{\mathrm{F}} + 40\frac{\|\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{Z}\|}{|\lambda_r|} = \|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}\|\hat{\mathbf{M}} - \mathbf{M}\|}{|\lambda_r|}.$$

Likewise, one has the bound of the operator norm

$$\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}^*\| \leq \|\hat{\mathbf{M}} - \mathbf{M}\| + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|^2}{|\lambda_r|}.$$

The proof completes. □

Then we are ready to prove Lemma 14, the assymetric case.

Proof of Lemma 14. To apply Lemma 18, first construct symmetrization of $\mathbf{M}$ and $\hat{\mathbf{M}}$,

$$\mathbf{Y} := \begin{pmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}, \ \hat{\mathbf{Y}} := \begin{pmatrix} 0 & \hat{\mathbf{M}} \\ \hat{\mathbf{M}}^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}.$$

Denote the perturbation matrix as $\mathbf{Z} := \hat{\mathbf{M}} - \mathbf{M}$ and similarly define

$$\hat{\mathbf{Z}} := \begin{pmatrix} 0 & \mathbf{Z} \\ \mathbf{Z}^\top & 0 \end{pmatrix} = \begin{pmatrix} 0 & \hat{\mathbf{M}} - \mathbf{M} \\ \hat{\mathbf{M}}^\top - \mathbf{M}^\top & 0 \end{pmatrix}.$$

Note that $\hat{}\,\mathsf{rank(Y)} = 2\mathsf{rank(M)} = 2r$。首先分析矩阵 Y 和 Y 的奇异值分解。注意，t

$$\frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{U} & \mathbf{U} \\ \mathbf{V} & -\mathbf{V} \end{pmatrix} \in \mathbb{R}^{2d\times 2r}$$

包含标准正交列。此外，它具有

$$\frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{U} & \mathbf{U}\\ \mathbf{V} & -\mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & 0\\ 0 & -\mathbf{\Sigma}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{U} & \mathbf{U}\\ \mathbf{V} & -\mathbf{V}\end{pmatrix}^\top = \begin{pmatrix}0 & \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\\ \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top & 0\end{pmatrix} = \begin{pmatrix}0 & \mathbf{M}\\ \mathbf{M}^\top & 0\end{pmatrix} = \mathbf{Y},$$

由此我们得到 Y 的奇异值分解。不失一般性，设 $d_1 \geq d_2$ and $\hat{\mathbf{M}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$ 是 $\hat{\mathbf{M}}$ 的完整奇异值分解，其中 $\tilde{\mathbf{U}} \in \mathbb{R}^{d_1\times d_2}$, $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{d_2}) \in \mathbb{R}^{d_2\times d_2}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{d_2\times d_2}$, 且 $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_{d_2} \geq 0$。因此，$\hat{\mathbf{M}}$ 的前导 r 个分量的奇异值分解为 $_r(\hat{\mathbf{M}}) = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$，其中 $\hat{\mathbf{U}}$ 和 $\hat{\mathbf{V}}$ 分别是完整矩阵 U 和 V 的前 r 列，且 $\mathbf{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_r)$。

注意到，由于 Y 是 M 的对称化，其完整奇异值分解为

$$\hat{\mathbf{Y}} = \frac{1}{\sqrt{2}}\begin{pmatrix}\tilde{\mathbf{U}} & \tilde{\mathbf{U}}\\ \tilde{\mathbf{V}} & -\tilde{\mathbf{V}}\end{pmatrix}\begin{pmatrix}\tilde{\mathbf{\Sigma}} & 0\\ 0 & -\tilde{\mathbf{\Sigma}}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\tilde{\mathbf{U}} & \tilde{\mathbf{U}}\\ \tilde{\mathbf{V}} & -\tilde{\mathbf{V}}\end{pmatrix}^\top, \tag{33}$$

其特征值为 $\tilde{\sigma}_1$, $-\tilde{\sigma}_1$, ..., $\tilde{\sigma}_{d_2}$, $-\tilde{\sigma}_{d_2}$。因此，最大的 $2r$ 个特征值的绝对值 $\mathbf{Y}$ 分别为 $\tilde{\sigma}_1$, $-\tilde{\sigma}_1$, ..., $\tilde{\sigma}_r$, $-\tilde{\sigma}_r$。因此，由于 $\mathbf{U}$、$\mathbf{V}$ 分别是 r 列 $\mathbf{U}$、$\mathbf{V}$，公式（33）可得

$$\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix}^\top \hat{\mathbf{Y}}\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{\Sigma}} & 0\\ 0 & -\hat{\mathbf{\Sigma}}\end{pmatrix},$$

这意味着 $\mathrm{SVD}_{2r}(\mathbf{Y})$ 为

$$\mathrm{SVD}_{2r}(\hat{\mathbf{Y}}) = \frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{\Sigma}} & 0\\ 0 & -\hat{\mathbf{\Sigma}}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix}^\top = \begin{pmatrix}0 & \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top\\ \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}\hat{\mathbf{U}}^\top & 0\end{pmatrix}$$
$$= \begin{pmatrix}0 & \mathrm{SVD}_r(\hat{\mathbf{M}})\\ \mathrm{SVD}_r(\hat{\mathbf{M}})^\top & 0\end{pmatrix},$$

并且 $\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}}\ \hat{\mathbf{U}}\\ \hat{\mathbf{V}}\ -\hat{\mathbf{V}}\end{pmatrix}$ 包含 $\hat{\mathbf{Y}}$ 的前导 $2r$ 特征向量。因此，我们有

$$\mathrm{SVD}_{2r}(\hat{\mathbf{Y}}) - \mathbf{Y} = \begin{pmatrix}0 & \mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\\ \mathrm{SVD}_r(\hat{\mathbf{M}})^\top - \mathbf{M}^\top & 0\end{pmatrix}.$$

由此可得

$$\|\mathrm{SVD}_{2r}(\hat{\mathbf{Y}}) - \mathbf{Y}\|_{\mathrm{F}} = \sqrt{2}\|\mathrm{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}}. \tag{34}$$

Then apply Lemma 18 to the rank $2r$ symmetric matrix $\mathbf{Y}$ and the symmetric perturbed matrix $\hat{\mathbf{Y}}$,

$$\|\mathrm{SVD}_{2r}(\hat{\mathbf{Y}}) - \mathbf{Y}^*\|_{\mathrm{F}} \leq \|\hat{\mathbf{Y}} - \mathbf{Y}\|_{\mathrm{F}} + 40\frac{\|\hat{\mathbf{Y}} - \mathbf{Y}\|_{\mathrm{F}}\|\hat{\mathbf{Y}} - \mathbf{Y}\|}{\sigma_r}, \tag{35}$$

which uses $|\lambda_{2r}(\mathbf{Y})| = \sigma_r$. Note that definition of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ implies $\|\hat{\mathbf{Y}} - \mathbf{Y}\|_{\mathrm{F}} = \sqrt{2}\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}$ and $\|\hat{\mathbf{Y}} - \mathbf{Y}\| = \|\hat{\mathbf{M}} - \mathbf{M}\|$, respectively. Then, Equation (34), Equation (35) infer,

$$\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\|_{\mathrm{F}} \leq \|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}\|\hat{\mathbf{M}} - \mathbf{M}\|}{\sigma_r}.$$

注意，算子范数情形的证明 $\|\operatorname{SVD}_r(\hat{\mathbf{M}}) - \mathbf{M}\| \leq \|\hat{\mathbf{M}} - \mathbf{M}\| + 40\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|^2}{\sigma_r}$ 与 Frobenius 范数情形类似，因此略去。

将引理 18 中的扰动特征向量结果应用于对称秩 $2r$ 矩阵 $\hat{\mathbf{Y}}$ 和对称扰动矩阵 $\mathbf{Y}$,

$$\left\|\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{\mathbf{U}} & \hat{\mathbf{U}}\\ \hat{\mathbf{V}} & -\hat{\mathbf{V}}\end{pmatrix}^{\top} - \frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{U} & \mathbf{U}\\ \mathbf{V} & -\mathbf{V}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{U} & \mathbf{U}\\ \mathbf{V} & -\mathbf{V}\end{pmatrix}^{\top}\right\| \leq \frac{8\|\hat{\mathbf{Y}} - \mathbf{Y}\|}{\sigma_r},$$

可将其简化为

$$\left\|\begin{pmatrix}\hat{\mathbf{U}}\hat{\mathbf{U}}^{\top} & 0\\ 0 & \hat{\mathbf{V}}\hat{\mathbf{V}}^{\top}\end{pmatrix} - \begin{pmatrix}\mathbf{U}\mathbf{U}^{\top} & 0\\ 0 & \mathbf{V}\mathbf{V}^{\top}\end{pmatrix}\right\| \leq \frac{8\|\hat{\mathbf{Y}} - \mathbf{Y}\|}{\sigma_r}.$$

接着，根据 $\|\mathbf{Y} - \mathbf{Y}\| = \|\mathbf{M} - \mathbf{M}\|$ 及分块矩阵的性质，可得

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^{\top} - \mathbf{U}\mathbf{U}^{\top}\| \leq \frac{8}{\sigma_r}\|\hat{\mathbf{M}} - \mathbf{M}\|, \quad \|\hat{\mathbf{V}}\hat{\mathbf{V}}^{\top} - \mathbf{U}\mathbf{U}^{\top}\| \leq \frac{8}{\sigma_r}\|\hat{\mathbf{M}} - \mathbf{M}\|.$$

□

H Proo矩阵恢复经验过程命题 3

证明。 证明由定理 14 得出。回顾 $\mathbb{M}_r := \{\mathbf{M} \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(\mathbf{M}) \leq r\}$ 是秩不超过 r 的矩阵集合。考虑

$$Z := \sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f(\mathbf{M} + \Delta\mathbf{M}) - f(\mathbf{M}) - \mathbb{E}(f(\mathbf{M} + \Delta\mathbf{M}) - f(\mathbf{M}))| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}.$$

注意到 $f(\mathbf{M}) = \sum_{i=1}^{n} \rho(\langle \mathbf{X}_i, \mathbf{M}\rangle - Y_i)$，于是我们有

$$\begin{aligned}
\mathbb{E}Z &\leq 2\mathbb{E}\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} \left|\sum_{i=1}^{n}\widetilde{\epsilon}_i \cdot \big(\rho(\langle \mathbf{X}_i, \mathbf{M} + \Delta\mathbf{M}\rangle - Y_i) - \rho(\langle \mathbf{X}_i, \mathbf{M}\rangle - Y_i)\big)\right| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}\\
&\leq 4\tilde{L} \cdot \mathbb{E}\sup_{\mathbf{M}\in\mathbb{M}_r}\left|\sum_{i=1}^{n}\widetilde{\epsilon}_i \cdot \langle \mathbf{X}_i, \Delta\mathbf{M}\rangle\right| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}\\
&\leq 4\tilde{L} \cdot \mathbb{E}\|\sum_{i=1}^{n}\widetilde{\epsilon}_i\mathbf{X}_i\|_{\mathrm{F,r}} \leq 4\tilde{L} \cdot \sqrt{r}\mathbb{E}\|\sum_{i=1}^{n}\widetilde{\epsilon}_i\mathbf{X}_i\|,
\end{aligned}$$

其中 $\{\widetilde{\epsilon}_i\}_{i=1}^n$ 是与 $\mathbf{X}_1,\ldots,\mathbf{X}_n$ 独立的 Rademacher 序列。第一个不等式来自定理 12，第二个不等式来自定理 13。注意，$\mathrm{vec}\,(\sum_{i=1}^n \widetilde{\epsilon}_i\mathbf{X}_i) \sim N(\mathbf{0}, \sum_{i=1}^n \mathbf{\Sigma}_i)$，且 $\|\sum_{i=1}^n \mathbf{\Sigma}_i\| \le nC_u$。引理 15 意味着 $\mathbb{E}\|\sum_{i=1}^n \widetilde{\epsilon}_i\mathbf{X}_i\| \le C_1\sqrt{nd_1C_u}$ 对于某个绝对常数$C_1 > 0$成立。因此我们有

$$\mathbb{E}Z \le C_1\tilde{L}\sqrt{nd_1rC_u}.$$

为简便起见，记 $f_i(\mathbf{M}\ \Delta\mathbf{M}) := |\rho(Y_i - \langle \mathbf{X}_i, \mathbf{M}\ + \Delta\mathbf{M}\rangle)\ - \rho(Y_i - \langle \mathbf{X}_i, \mathbf{M}\rangle)| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}$。于是有

$$\begin{aligned}
&\|\max_{i=1,\ldots,n}\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f_i(\mathbf{M},\Delta\mathbf{M}) - \mathbb{E}f_i(\mathbf{M},\Delta\mathbf{M})|\|_{\Psi_1}\\
&\le 2\log n \max_{i=1,\ldots,n}\|\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f_i(\mathbf{M}+\Delta\mathbf{M})|\|_{\Psi_1} + 2\log n \max_{i=1,\ldots,n}\mathbb{E}\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f_i(\mathbf{M}+\Delta\mathbf{M})|,
\end{aligned}$$

这里用到了次指数随机变量最大值的不等式。此外，对于每个 $i = 1, \ldots, n$，有

$$\begin{aligned}
&\|\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f_i(\mathbf{M}+\Delta\mathbf{M})|\|_{\Psi_1}\\
&= \|\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |\rho(Y_i - \langle\mathbf{X}_i, \mathbf{M}+\Delta\mathbf{M}\rangle) - \rho(Y_i - \langle\mathbf{X}_i,\mathbf{M}\rangle)| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}\|_{\Psi_1}\\
&\le \tilde{L}\|\sup_{\Delta\mathbf{M}\in\mathbb{M}_r} |\langle\mathbf{X}_i,\Delta\mathbf{M}\rangle| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}\|_{\Psi_1}\\
&\le \tilde{L}\|\|\mathbf{X}_i\|_{\mathrm{F,r}}\|_{\Psi_1} \le \tilde{L}\sqrt{r}\|\|\mathbf{X}_i\|\|_{\Psi_1} \le C_2\tilde{L}\sqrt{d_1rC_u},
\end{aligned}$$

对于某个绝对常数 $C_2 > 0$ 最后一行由部分 Frobenius 范数的定义（引理 12 ）和随机多元高斯矩阵范数（引理 15）得到。类似地，我们有

$$\begin{aligned}
\mathbb{E}\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |f_i(\mathbf{M})| &= \mathbb{E}\sup_{\substack{\Delta\mathbf{M}\in\mathbb{M}_r\\ \mathbf{M}\in\mathbb{R}^{d_1\times d_2}}} |\rho(Y_i - \langle\mathbf{X}_i,\mathbf{M}+\Delta\mathbf{M}\rangle) - \rho(Y_i - \langle\mathbf{X}_i,\mathbf{M}\rangle)| \cdot \|\Delta\mathbf{M}\|_{\mathrm{F}}^{-1}\\
&\le \tilde{L}\cdot\mathbb{E}\|\mathbf{X}_i\|_{\mathrm{F,r}} \le C_3\tilde{L}\sqrt{d_1rC_u}.
\end{aligned}$$

结合上述三个等式，可得

$$\|\max_{i=1,\ldots,n}\sup_{\mathbf{M}\in\mathbb{M}_r} |f_i(\mathbf{M}) - \mathbb{E}f_i(\mathbf{M})|\|_{\Psi_1} \le C_4\tilde{L}\sqrt{d_1rC_u}\log n.$$

Also, note that $\langle\mathbf{X}_i,\Delta\mathbf{M}\rangle \sim N(0, \mathrm{vec}(\Delta\mathbf{M})^\top\mathbf{\Sigma}_i\mathrm{vec}(\Delta\mathbf{M}))$ and $\mathrm{vec}(\mathbf{M}-\mathbf{M}^*)^\top\mathbf{\Sigma}_i\mathrm{vec}(\mathbf{M}-\mathbf{M}^*) \le C_u\|\mathbf{M}-\mathbf{M}^*\|_{\mathrm{F}}^2$. Then we have

$$\begin{aligned}
&\mathbb{E}\big[\big(\rho(Y_i - \langle\mathbf{X}_i,\mathbf{M}+\Delta\mathbf{M}\rangle) - \rho(Y_i - \langle\mathbf{X}_i,\mathbf{M}\rangle)\big)^2\big]/\|\Delta\mathbf{M}\|_{\mathrm{F}}^2\\
&\le \tilde{L}^2\mathbb{E}\langle\mathbf{X}_i,\Delta\mathbf{M}\rangle^2/\|\Delta\mathbf{M}\|_{\mathrm{F}}^2\\
&\le C_u\tilde{L}^2.
\end{aligned}$$

Invoke Theorem 14 and Remark 2, and take $\alpha = 1$, $\eta = 1$, $\delta = 0.5$ and there exists some constant $C > 0$, such that

$$\mathbb{P}(Z \geq C_2\tilde{L}\sqrt{C_u n d_1 r} + t) \leq \exp\left(-\frac{t^2}{3nC_u\tilde{L}^2}\right) + 3\exp\left(-\frac{t}{C_5\tilde{L}\sqrt{C_u d_1 r}\log n}\right)$$

对任意 $t > 0$均成立。取 $t = C\sqrt{n d_1 r C_u}\tilde{L}$ ，则有

$$\mathbb{P}(Z \geq C\tilde{L}\sqrt{n d_1 r C_u}) \leq \exp\left(-\frac{C^2 d_1 r}{3}\right) + 3\exp\left(-\frac{\sqrt{n}}{\log n}\right),$$

由此完成证明。 □

定理 12（期望的对称化，（Van Der Vaart 等人，1996））。设 $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ 为 χ 中相互独立的矩阵，且 $\mathcal{F}$ 是定义在 χ 上的实值函数类。设 $\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_n$ 是与 $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ 独立的 Rademacher 序列，则

$$\mathbb{E}\big[\sup_{f\in\mathcal{F}}\big|\sum_{i=1}^{n}(f(\mathbf{X}_i) - \mathbb{E}f(\mathbf{X}_i))\big|\big] \leq 2\mathbb{E}\big[\sup_{f\in\mathcal{F}}\big|\sum_{i=1}^{n}\tilde{\varepsilon}_i f(\mathbf{X}_i)\big|\big] \tag{36}$$

定理 13 （压缩定理，（Ludoux 和 Talagrand，1991））。设非随机元素 x_1，……，x_n 属于 χ。设 $\mathcal{F}$ 是定义在 χ 上的实值函数类。设利普希茨连续函数 ρ_i： $\mathbb{R} \to \mathbb{R}$ ，其利普希茨常数为 L，即

$$|\rho_i(\mu) - \rho_i(\tilde{\mu})| \leq L|\mu - \tilde{\mu}|, \ for\ all\ \mu, \tilde{\mu} \in \mathbb{R}$$

令 $\tilde{\varepsilon}_1$……，$\tilde{\varepsilon}_n$ 是一个 Rademacher 序列。于是，对于任意函数 f^*: $\chi \to \mathbb{R}$，我们有 e

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\tilde{\varepsilon}_i\left\{\rho_i\left(f\left(x_i\right)\right) - \rho_i\left(f^*\left(x_i\right)\right)\right\}\right|\right] \leq 2\mathbb{E}\left[L\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\tilde{\varepsilon}_i\left(f\left(x_i\right) - f^*\left(x_i\right)\right)\right|\right] \tag{37}$$

定理 14（经验过程上确界的尾不等式（Adamczak，2008））。设 $\mathbf{X}_1$，……，$\mathbf{X}_n$是取值于可测空间 $(\mathcal{S}, \mathcal{B})$ 中的独立随机变量，并设 $\mathcal{F}$ 是可测函数 f: $\mathcal{S} \to \mathbb{R}$ 的可数类。假设对每个 $f \in \mathcal{F}$ 和每个 i， $\mathbb{E}f\left(\mathbf{X}_i\right) = 0$，且对于某个 $\alpha \in (0, 1]$，以及所有 i， $\|\sup_{\mathcal{F}}|f(\mathbf{X}_i)|\|_{\Psi_\alpha} < \infty$. 令

$$Z = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f\left(\mathbf{X}_i\right)\right|$$

Define moreover

$$\sigma^2 = \sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathbb{E}f\left(\mathbf{X}_i\right)^2$$

Then, for all $0 < \eta < 1$ *and* $\delta > 0$, *there exists a constant* $C = C(\alpha, \eta, \delta)$, *such that for all* $t \geq 0$

$$\begin{aligned}
&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t) \\
&\quad \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\left\|\max_i \sup_{f\in\mathcal{F}} |f(\mathbf{X}_i)|\right\|_{\Psi_\alpha}}\right)^\alpha\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(Z \leq (1-\eta)\mathbb{E}Z - t) \\
&\quad \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\left\|\max_i \sup_{f\in\mathcal{F}} |f(\mathbf{X}_i)|\right\|_{\Psi_\alpha}}\right)^\alpha\right)
\end{aligned}$$

注释 2。注意，这里我们要求 $\mathbb{E}f(\mathbf{X}_i) = 0$ 且当 $\mathbb{E}f(\mathbf{X}_i) \neq 0$ 时，Z 应为

$$Z = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n f(\mathbf{X}_i) - \mathbb{E}f(\mathbf{X}_i) \right|,$$

而 Z 的尾部不等式应为

$$\begin{aligned}
&\mathbb{P}(Z \geq (1+\eta)\mathbb{E}Z + t) \\
&\quad \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\left\|\max_i \sup_{f\in\mathcal{F}} |f(\mathbf{X}_i) - \mathbb{E}f(\mathbf{X}_i)|\right\|_{\Psi_\alpha}}\right)^\alpha\right)
\end{aligned}$$

I 异常值情形的证明

定理 7 的证明

我们只需证明以下两阶段正则性性质，由此即可得到与第 B 节类似的收敛动力学。

引理 19。假设内点满足假设 1 和假设 2。存在绝对常数$C_1, C_2, C_3, c_0 > 0$，使得若 $n \geq C_1 C_u C_l^{-1} \tilde{s}\log(2d/\tilde{s})$ 且 $\epsilon \leq \frac{\sqrt{C_l}}{4\sqrt{C_l}+\sqrt{C_u}}$，则以概率至少为 1- $1 - 2\exp\left(-c_0\tilde{s}\log\left(2d/\tilde{s}\right)\right) - 3\exp\left(-((1-\epsilon)n\log(2d/\tilde{s}))^{1/2}\log^{-1}(1-\epsilon)n\right)$，我们有

(1) for all $\boldsymbol{\beta} \in \left\{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma,\ |supp(\boldsymbol{\beta})| \leq \tilde{s}\right\}$ *and for all sub-gradient* $\mathbf{G} \in \partial f(\boldsymbol{\beta})$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \geq \frac{n}{5}C_l^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \quad \text{and} \quad \|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \leq nC_u^{1/2},$$

其中 $\Omega := \text{supp}(\boldsymbol{\beta})$ 且 $\Pi := \text{supp}\left(\mathcal{H}_{\tilde{s}}(\mathcal{P}_{\Omega^c}(\mathbf{G}))\right)$;

(2) for all

$$\boldsymbol{\beta} \in \left\{\boldsymbol{\beta} \in \mathbb{R}^d:\ C_2 C_u^{1/2} C_l^{-1} \cdot b_0 \max\left\{\left(\tilde{s}/n \cdot \log(2d/\tilde{s})\right)^{1/2}, \epsilon\right\} \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le 8 C_l^{-1/2}\gamma, |supp(\boldsymbol{\beta})| \le \tilde{s}\right\}$$

且对于所有次梯度 $\mathsf{G} \in \partial f(\boldsymbol{\beta})$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \ge \frac{nC_l}{12b_0}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \quad \text{and} \quad \|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \le C_3\frac{nC_u}{b_1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,$$

其中 $\Omega := \mathrm{supp}(\boldsymbol{\beta})$, 且 $\Pi := \mathrm{supp}\left(\mathcal{H}_{\tilde{s}}(\mathcal{P}_{\Omega^c}(\mathbf{G}))\right)$。

引理 *19* 的证明。与引理 1 的区别在于出现了异常值，因此我们将以内点和异常值分别处理。记

$$f_i(\boldsymbol{\beta}) := \sum_{i\in\mathcal{I}} |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}\rangle|, \quad f_o(\boldsymbol{\beta}) := \sum_{i\in\mathcal{O}} |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}\rangle|.$$

注意，我们有 $f(\boldsymbol{\beta}) = f_i(\boldsymbol{\beta}) + f_o(\boldsymbol{\beta})$。我们假设事件

$$\boldsymbol{\mathcal{E}}_i := \left\{\sup_{\boldsymbol{\beta}\in\mathbb{R}^d,\ |\mathrm{supp}(\Delta\boldsymbol{\beta})|\le 3\tilde{s}} |f_i(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}) - \mathbb{E}\left[f_i(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) - f_i(\boldsymbol{\beta})\right]| \cdot \|\Delta\boldsymbol{\beta}\|_2^{-1} \le C\sqrt{(1-\varepsilon)n\tilde{s}C_u\log(2d/\tilde{s})}\right\}$$

成立。命题 2 证明了 $\mathbb{P}(\mathsf{E}_i) \ge 1 - \exp(-C^2\tilde{s}\log(2d/\tilde{s})3) - 3\exp\left(-\sqrt{(1-\epsilon)n\log(2d/\tilde{s})}\right)\log(1-\epsilon)$。此外，我们假设事件

$$\boldsymbol{\mathcal{E}}_o := \left\{\sup_{|\mathrm{supp}(\Delta\boldsymbol{\beta})|\le 3\tilde{s}} \sum_{i\in\mathcal{O}} \left||\langle \mathbf{X}_i, \Delta\boldsymbol{\beta}\rangle| - \mathbb{E}|\langle \mathbf{X}_i, \Delta\boldsymbol{\beta}\rangle|\right| \cdot \|\Delta\boldsymbol{\beta}\|_2^{-1} \le C\sqrt{\epsilon n\tilde{s}C_u\log(2d/\tilde{s})}\right\}$$

成立，并且 ε-网理论或高斯过程理论表明 $\mathbb{P}(\mathsf{E}_o) \ge 1 - \exp(-C^2\tilde{s}\log(2d/\tilde{s})3)$。

Phase One First, show μ-sharpness. Notice that

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) = f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) + f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*).$$

Same as Section B, under event $\boldsymbol{\mathcal{E}}_i$, we have

$$f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) \ge (1-\epsilon)n\sqrt{C_l\pi/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - C\sqrt{(1-\epsilon)n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - 2(1-\epsilon)n\gamma.$$

At the same time, under event $\boldsymbol{\mathcal{E}}_o$, we have

$$\begin{aligned}
\left|f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*)\right| &= \left|\sum_{i\in\mathcal{O}} |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}\rangle| - \sum_{i\in\mathcal{O}} |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}^*\rangle|\right| \\
&\le \sum_{i\in\mathcal{O}} |\langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| \\
&\le \mathbb{E}\sum_{i\in\mathcal{O}} |\langle \mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| + C\sqrt{\epsilon n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
&\le \epsilon n\sqrt{\frac{2}{\pi}C_u}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C\sqrt{\epsilon n\tilde{s}C_u\log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,
\end{aligned}$$

where the last line follows from Lemma 16. Thus, we have lower bound of $f(\mathbf{M}) - f(\mathbf{M}^*)$,

$$\begin{aligned}
& f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) \\
\geq\ & f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) - |f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*)| \\
\geq\ & (1-\epsilon)n\sqrt{\frac{2}{\pi}C_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - C\sqrt{(1-\epsilon)n\tilde{s}C_u \log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
& - \epsilon n\sqrt{\frac{2}{\pi}C_u}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - C\sqrt{\epsilon n\tilde{s}C_u \log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - 2(1-\epsilon)n\gamma \\
\geq\ & \frac{n}{5}\sqrt{C_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,
\end{aligned}$$

其中，最后一行是由 $\epsilon \leq \sqrt{C_l}/(\sqrt{C_l}+4\sqrt{C_u}) \leq 0.2$, $1-\epsilon \geq 0.8$, $n \geq C_1 C_u C_l^{-1}\tilde{s}\log(2d/\tilde{s})$ 和第一阶段区域约束 $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma$ 得出的。我们可以用类似的方法证明 $\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2$ 的上界，

$$\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \leq n\sqrt{C_u}.$$

第二阶段 首先考虑 $f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)$ 的下界。按照第 B 节，我们得到当$C_2\sqrt{\frac{C_u}{C_l^2}}b_0 \max\{\sqrt{\frac{\tilde{s}}{\log(2d/\tilde{s})\, n}}, \epsilon\} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq 8\sqrt{C_l^{-1}}\gamma$,

$$f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) \geq (1-\epsilon)\frac{n}{12b_0}C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

Besides, same as phase one analyses, under event $\boldsymbol{\mathcal{E}}_o$, it has

$$\left|f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*)\right| \leq \epsilon n\sqrt{\frac{2}{\pi}C_u}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C\sqrt{\epsilon n\tilde{s}C_u \log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

Thus, the above two equations lead to

$$\begin{aligned}
f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*) &= f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) + f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*) \\
&\geq f_i(\boldsymbol{\beta}) - f_i(\boldsymbol{\beta}^*) - |f_o(\boldsymbol{\beta}) - f_o(\boldsymbol{\beta}^*)| \\
&\geq (1-\epsilon)\frac{n}{12b_0}C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - \epsilon n\sqrt{\frac{2}{\pi}C_u}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - C\sqrt{\epsilon n\tilde{s}C_u \log(2d/\tilde{s})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
&\geq \frac{n}{C_3 b_0}C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,
\end{aligned}$$

where the last line uses $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq C_2\sqrt{\frac{C_u}{C_l^2}}b_0 \max\{\sqrt{\frac{\tilde{s}\log(2d/\tilde{s})}{n}}, \epsilon\}$. Similar to Section B, we have

$$\|\mathcal{P}_{\Omega\cup\Pi\cup\Omega^*}(\mathbf{G})\|_2 \leq C_4 C_u\frac{n}{b_1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

□

Proof of Theorem 8

We only need to prove the following regularity properties.

引理 20。 假设 $\{\xi_i\}_{i\in\mathcal{I}}$ 和 $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ 分别满足假设 2和 1。存在绝对常数 $C_1, C_2, C_3, c_1 > 0$，使得若 $n \geq C_1 r d_1 C_u C_l^{-1}$，$\epsilon \leq \sqrt{C_l}/(4\sqrt{C_l} + \sqrt{C_u})$和 $\mathbb{B}_1, \mathbb{B}_2$ 由下式给出

$$\mathbb{B}_1 := \left\{ \mathbf{M} \in \mathbb{M}_r : \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} \geq 8\sqrt{C_l^{-1}}\gamma \right\},$$
$$\mathbb{B}_2 := \left\{ \mathbf{M} \in \mathbb{M}_r : C_2 b_0 \cdot \sqrt{\frac{C_u}{C_l^2}} \cdot \max\left\{ \sqrt{\frac{d_1 r}{n}}, \epsilon \right\} \leq \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}} < 8\sqrt{C_l^{-1}}\gamma \right\},$$

那么，以至少 $1 - 2\exp(-c_1 r d_1) - 3\exp(-\sqrt{n}/\log n)$ 的概率，绝对损失 $f(\mathbf{M}) = \sum_{i=1}^n |Y_i - \langle \mathbf{M}, \mathbf{X}_i \rangle|$ 满足条件 1:

(1) the rank-r restricted ***two-phase sharpness*** *with respect to* $\mathbf{M}^*$,

$$f(\mathbf{M}) - f(\mathbf{M}^*) \geq \begin{cases} \frac{n}{4}\sqrt{C_l}\|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & \text{for} \quad \mathbf{M} \in \mathbb{B}_1; \\ \frac{n}{12 b_0} C_l \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2, & \text{for} \quad \mathbf{M} \in \mathbb{B}_2; \end{cases}$$

(2) the rank-r restricted ***two-phase sub-gradient bound*** *with respect to* $\mathbf{M}^*$,

$$\|\mathbf{G}\|_{\mathrm{F,r}} \leq \begin{cases} 2n\sqrt{C_u}, & \text{for} \quad \mathbf{M} \in \mathbb{B}_1; \\ C_3 n b_1^{-1} C_u \|\mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}, & \text{for} \quad \mathbf{M} \in \mathbb{B}_2, \end{cases}$$

where $\mathbf{G} \in \partial f(\mathbf{M})$ *is any sub-gradient.*

Proof of Lemma 20 would be a combination of absolute loss low-rank study in Lemma 6 and outlier analyses in Lemma 19. We also study loss function of inliers and outliers seperately,

$$f_i(\mathbf{M}) := \sum_{i\in\mathcal{I}} |Y_i - \langle \mathbf{X}_i, \mathbf{M} \rangle|, \quad f_o(\mathbf{M}) := \sum_{i\in\mathcal{O}} |Y_i - \langle \mathbf{X}_i, \mathbf{M} \rangle|.$$

Hence, we skip the proof.