On the Convergence of IRLS and Its Variants in Outlier-Robust Estimation: Main Paper + Supplementary Materials

Liangzu Peng
Johns Hopkins University
lpeng25@jhu.edu

Christian Kümmerle
UNC Charlotte
kuemmerle@uncc.edu

René Vidal
University of Pennsylvania
vidalr@seas.upenn.edu

Abstract

*Outlier-robust estimation involves estimating some parameters (*e.g.*, 3D rotations) from data samples in the presence of outliers, and is typically formulated as a non-convex and non-smooth problem. For this problem, the classical method called iteratively reweighted least-squares (IRLS) and its variants have shown impressive performance. This paper makes several contributions towards understanding why these algorithms work so well. First, we incorporate majorization and graduated non-convexity (GNC) into the IRLS framework and prove that the resulting IRLS variant is a convergent method for outlier-robust estimation. Moreover, in the robust regression context with a constant fraction of outliers, we prove this IRLS variant converges to the ground truth at a global linear and local quadratic rate for a random Gaussian feature matrix with high probability. Experiments corroborate our theory and show that the proposed IRLS variant converges within* 5-10 *iterations for typical problem instances of outlier-robust estimation, while state-of-the-art methods need at least* 30 *iterations. A basic implementation of our method is provided:* https://github.com/liangzu/IRLS-CVPR2023

> *... attempts to analyze this difficulty [*caused by infinite weights of IRLS for the ℓ_p-loss*] have a long history of proofs and counterexamples to incorrect claims.*
>
> Khurrum Aftab & Richard Hartley [1]

1. Introduction

1.1. The Outlier-Robust Estimation Problem

Many parameter estimation problems can be stated in the following general form. We are given some function $r : \mathcal{C} \times \mathcal{D} \to [0, \infty)$, called the *residual* function. Here $\mathcal{D}$ is the domain of data samples $d_1, \dots, d_m$, and $\mathcal{C} \subset \mathbb{R}^n$ is the *constraint set* where our (ground truth) *variable* v^* lies; $\mathcal{C}$ can be convex such as an affine subspace, or non-convex such as a special orthogonal group $\mathrm{SO}(3)$. We aim to recover v^* from data d_i's. A simple example is *linear regression*, where a sample $d_i = (\boldsymbol{a}_i, y_i)$ consists of a feature vector $\boldsymbol{a}_i \in \mathbb{R}^n$ and a scalar response $y_i \in \mathbb{R}$, and the residual function is $r(v, d_i) := |\boldsymbol{a}_i^\top v - y_i|$.

The sample d_i is called an *inlier*, if $r(v^*, d_i) \approx 0$. It is called an *outlier*, if the residual $r(v^*, d_i)$ is *large* (vaguely speaking). If all samples are inliers, one usually prefers solving the following problem as a means to estimate v^*:

$$\min_{v \in \mathcal{C}} \sum\nolimits_{i=1}^{m} r(v, d_i)^2 \tag{1}$$

Problem (1) is called *least-squares*, and is known since Legendre [47] and Gauss [30] in the linear regression context. Even before that, Boscovich [13] suggested to minimize (1) without the square. This unsquared version is called *least absolute deviation*, and is more robust to outliers than (1).

Consider the following formulation for *outlier-robust estimation* (*i.e.*, a specific type of *M-estimators* [39, 61]):

$$\min_{v \in \mathcal{C}} \sum\nolimits_{i=1}^{m} \rho\big(r(v, d_i)\big) \tag{2}$$

Here $\rho : \mathbb{R} \to \mathbb{R}$ is some outlier-robust loss (the unsquared version of (1) corresponds to $\rho(r) = |r|$ in (2)). Among many possible losses ρ [20, 24], we discuss two particular choices. The first is the ℓ_p-loss $\rho(r) = |r|^p/p$, $p \in (0, 1]$; it has been used in several research fields, *e.g.*, *geometric vision* [1, 20], *compressed sensing* [18, 23, 41], *matrix recovery* [42, 50, 51], and *subspace clustering* [27]. The other loss is due to Huber [39]: $\rho(r) = \min\{r^2, c^2\}$, with $c > 0$ a hyper-parameter; it has later been named as *Talwar* [22, 54], *Huber-type skipped mean* [32], *truncated quadratic* [6, 11], and *truncated least-squares* (TLS) [4, 78, 86]. Both losses are highly robust to outliers but make solving (2) difficult, *e.g.*, the objective of (2) becomes non-smooth or non-convex. This motivates the need to develop efficient and provably correct solvers for (2) with either of the two losses.

1.2. IRLS and Its Variants in Vision & Optimization

The General Principle of IRLS. As its name suggests, *iteratively reweighted least-squares* (IRLS) is a general algorithmic paradigm that alternates between defining a weight

for each sample and solving a weighted least squares problem. Specifically, IRLS initializes a variable $v^{(0)} \in \mathcal{C}$, and, for $t = 0, 1, \ldots$, alternates between the following two steps:

$$\text{Update weights } w_i^{(t+1)} \text{ based on } v^{(t)}, \forall i = 1, \ldots, m \quad (3)$$

$$\text{Solve: } v^{(t+1)} \leftarrow \operatorname*{argmin}_{v \in \mathcal{C}} \sum_{i=1}^{m} w_i^{(t+1)} r(v, d_i)^2 \quad (4)$$

This basic idea dates back to the seminal work of Weiszfeld [76]; see [9,34] for some historical accounts. A well-known and general rule for the weight update is (*cf*. [1, 22, 53])

$$w_i^{(t+1)} \leftarrow \rho'\left(r_i^{(t)}\right)/r_i^{(t)}, \quad r_i^{(t)} := r(v^{(t)}, d_i). \quad (5)$$

In a nutshell, the rationale behind rule (5) is to "connect" weighted least-squares (4) to outlier-robust estimation (2), allowing IRLS to optimize the latter (2). Indeed, [1] shows that IRLS with the weight update in (5) results in a non-increasing objective (2). Moreover, [1] gives conditions under which IRLS with (5) converges to a stationary point of (2). This confirms that one can apply IRLS to problem (2), as long as one can solve weighted least-squares[1] (4).

However, the conditions of the theorem of [1] are hard to verify, *e.g.*, one condition requires the minimizer of (4) to be a continuous function of weights $w_i^{(t+1)}$. Moreover, as [1] commented, directly applying (5) to non-smooth or non-convex losses (*e.g.*, ℓ_p or TLS) might create *significant theoretical and practical difficulties*, *e.g.*, (5) is undefined at non-differentiable points. This suggests that rule (5) needs to be improved if the ℓ_p or TLS loss is to be minimized.

IRLS in A Tale of Two Losses. For the non-smooth ℓ_p-loss, (5) *results in*[2] $w_i^{(t+1)} \leftarrow (r_i^{(t)})^{p-2}$, which tends to infinity as $r_i^{(t)} \to 0$. A workaround is to truncate the residual by some positive number ϵ, *i.e.*, $w_i^{(t+1)} \leftarrow \max\{r_i^{(t)}, \epsilon\}^{p-2}$ [20, 25–27, 48, 49, 71]. While [1, 66] considered this to be "*an ad-hoc procedure*", in the optimization literature, there do exist theoretical guarantees for IRLS with this revised weight update to converge, at least for some specific residual functions r, see, *e.g.*, [8, 17, 48, 49].

For the non-smooth and non-convex TLS loss $\rho(r) = \min\{r^2, c^2\}$, (5) *results in*[2] a *hard thresholding* scheme: set $w_i^{(t+1)}$ to 1 if $r_i^{(t)} \leq c$, or set it to 0 otherwise. IRLS fails with such a weight update if the outlier rate exceeds 10% for *category-level perception* as reported in [64]. This could be remedied in two ways, discussed next.

The first is to adopt a different hard thresholding method [10] from the optimization literature, which sets $w_i^{(t+1)}$ to 1 if $r_i^{(t)}$ is among the s-smallest of all residuals (s is a hyper-parameter), or set $w_i^{(t+1)}$ to 0 otherwise; this method is robust up to 50% outliers for *robust regression*, and converges globally linearly under some conditions. Note that this IRLS variant is not meant to minimize the TLS loss.

The second remedy manifests itself if one applies rule (5) to some *smoothing* approximation $\rho_\mu(r)$ of the TLS loss $\rho(r) = \min\{r^2, c^2\}$. The approximation of [12] is

$$\rho_\mu(r) = \begin{cases} r^2, & \text{if } r^2 \leq \frac{\mu c^2}{\mu+1}, \\ c^2, & \text{if } r^2 \geq \frac{\mu+1}{\mu} c^2, \\ 2c|r|\sqrt{\mu(\mu+1)} - \mu(c^2 + r^2), & \text{o/w.} \end{cases} \quad (6)$$

Since $\rho_\mu \to \rho$ as $\mu \to \infty$, a natural strategy, called *graduated non-convexity* (GNC) [12], is to alternate between optimizing ρ_μ and increasing μ at each iteration t. The method used for increasing μ is called a GNC *schedule* and the default schedule has been a *linear* one, *i.e.*, $\mu^{(t+1)} \leftarrow \gamma\mu^{(t)}$ with some hyper-parameter $\gamma > 1$ [45, 52, 64, 67, 78, 86]. For example, the `GNC-TLS` method [12, 78] incorporates this linear schedule within the IRLS framework (3)-(5) to approximate the TLS loss via ρ_μ.

However, the great engineering intuition of [12] and its follow-up works [4, 45, 69, 78, 87] on GNC comes with the lack of theoretical guarantees, thus [80, 82] refer to GNC as a "*fast heuristic*" strategy. On the other hand, in the optimization literature, similar GNC twists for the ℓ_p-loss have been empirically investigated [19, 73, 77] for compressed sensing and related problems, and empowered with global linear or local superlinear convergence rates [23, 41, 52, 58].

For outlier-robust estimation (2), either with general [4, 45, 78] or specific residual functions [27, 58], either with the ℓ_p [27, 52, 58], TLS [4, 45, 64, 78], or even other losses [66, 84], combining IRLS and GNC has pushed the empirical performance to a certain limit, which other types of methods (*e.g.*, `RANSAC` [29]) can hardly attain given the same time budget. On the other hand, theoretical guarantees for IRLS offered in the optimization literature are limited to specific problems (*e.g.*, compressed sensing [23]), and, though related, cannot be applied directly to outlier-robust estimation (2). An intriguing but under-explored theoretical question is why IRLS, GNC, and the like work so well for outlier-robust estimation (2)—can we extend, not just apply, the insights from optimization to answer this question?

1.3. Our Contribution

We present an IRLS variant called `GNC-IRLS`$_p$ (Algorithm 1) for the outlier-robust estimation problem (2) and establish general convergence properties for general constraint sets $\mathcal{C}$, providing a well-founded framework for empirically successful GNC methods. We further elucidate how appropriately chosen update rules for the smoothing parameter $\epsilon^{(t)}$ (Line 7) of `GNC-IRLS`$_p$ lead to a global and fast local convergence for outlier-robust estimation problems. More specifically, our contributions are as follows:

[1] While solving weighted least-squares (4) can be hard, many solvers for geometric vision exist, see, *e.g.*, [2, 5, 15, 37, 38, 55, 62, 64, 78, 84].

[2] Pretending that the ℓ_p or TLS losses are differentiable everywhere.

Algorithm 1: GNC−IRLS$_p$

1 Input: data $\mathcal{d}_1, \dots, \mathcal{d}_m$, $p \in (0, 1]$;
2 Let $v^{(0)} \in \mathcal{C}$ with $\|v^{(0)}\|_2 < \infty$ and $\epsilon^{(0)} \in (0, \infty)$;
3 For $t \leftarrow 0, 1, 2, \dots$:
4 Compute the residual $r_i^{(t)} \leftarrow r(v^{(t)}, \mathcal{d}_i), \forall i$;
5 $w_i^{(t+1)} \leftarrow \max\{r_i^{(t)}, \epsilon^{(t)}\}^{p-2}$; // Sec. 2
6 Solve problem (4) and get $v^{(t+1)}$; // Sec. 2
7 Calculate $\epsilon^{(t+1)}$ based on a GNC schedule ; // Sec. 3

• In Section 2, we consider outlier-robust estimation (2) for a general class of residual functions and constraints, and we prove that GNC−IRLS$_p$ converges to stationary points of (some majorizer of) the ℓ_p-loss under suitable assumptions (Theorem 1). Moreover, the assumptions are easy to verify and satisfied by many geometric vision problems (see the appendix). This challenges the viewpoint of [1, 66] that truncating the residual (Line 4, Algorithm 1) is "*ad-hoc*". Our proof is enabled by a *majorization* interpretation of GNC−IRLS$_p$, and is motivated by [23, 51, 60]. As we will discuss, our result is more general than those of [23, 51, 60].

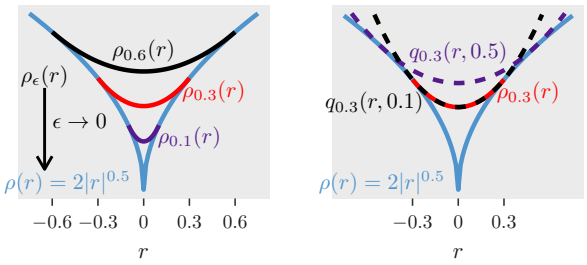
• In Section 3, we propose a *superlinear* GNC schedule for GNC−IRLS$_p$, as opposed to a *linear* one. We prove that GNC−IRLS$_p$ with such a schedule converges to the ground truth at a global linear and local superlinear rate, with high probability (Theorem 2). Moreover, GNC−IRLS$_p$ provably enjoys quadratic rates starting from the first iteration. A theoretical drawback of this powerful result is that it has a "burn-in" period and is limited to the *robust regression* setting; this is harmless though when it comes to practical use. Our proof is motivated by [52]. Their result holds only for $p = 1$, and our contribution lies not only in overcoming the non-convexity for the case of $p < 1$, but in leveraging the non-convexity to obtain a faster convergence rate.

• In Section 4 we compare the performance of GNC−TLS and GNC−IRLS$_p$ for *point cloud registration*. GNC−IRLS$_p$ terminates in 10 iterations while GNC−TLS takes 30. This is because GNC−IRLS$_p$ uses a superlinear GNC schedule, while GNC−TLS uses a linear schedule.

• In Section 5, we endow the TLS loss with a *majorization* strategy and a *superlinear* GNC schedule, leading to an IRLS method that we call MS−GNC−TLS. With majorization we prove MS−GNC−TLS converges, which challenges the viewpoint of [80, 82] that GNC is "*heuristic*". With the superlinear schedule, MS−GNC−TLS converges, say, at iteration 6, whereas GNC−TLS does so only at iteration 30.

2. GNC−IRLS$_p$: Interpretation & Convergence

In this section we show that GNC−IRLS$_p$ is a convergent method, each iteration making steady progress towards minimizing (some majorizer of) the ℓ_p-loss. We first show GNC−IRLS$_p$ involves *two-level majorization* (Section 2.1). Then we state our convergence result (Section 2.2).



(a) Smooth Majorizer $\rho_\epsilon(r)$ (b) Quadratic Majorizer $q_\epsilon(r, u)$

Figure 1. Two majorizers of $\rho(r) = 2|r|^{0.5}$, ρ_ϵ (7) and q_ϵ (8).

2.1. Interpretation of GNC−IRLS$_p$

For two functions f and g defined on $\mathbb{R}$, if $f(r) \geq g(r)$ ($\forall r \in \mathbb{R}$), we say f *majorizes* g or f is a *majorizer* of g. Behind the apparent alternating nature of GNC−IRLS$_p$, it involves *two-level majorization*, as signified by the *smooth majorizer* and *quadratic majorizer*, introduced next.

Smooth Majorizer. As the main player in the first level of majorization, we define the *smooth majorizer* $\rho_\epsilon : \mathbb{R} \to \mathbb{R}_{\geq 0}$ for each $\epsilon > 0$ [58, 68] such that

$$\rho_\epsilon(r) = \begin{cases} \frac{1}{p}|r|^p, & |r| > \epsilon, \\ \frac{1}{2}\frac{r^2}{\epsilon^{2-p}} + (\frac{1}{p} - \frac{1}{2})\epsilon^p, & |r| \leq \epsilon. \end{cases} \tag{7}$$

The smooth majorizer ρ_ϵ is a Huber-like loss [39] which coincides with the ℓ_p-loss if $|r| \geq \epsilon$ and is otherwise quadratic in r. Figure 1a shows that ρ_ϵ majorizes the ℓ_p-loss for $p = 0.5$ and different values of ϵ. More formally, we have:

Lemma 1 ($\rho_\epsilon(\cdot)$ is Smooth ℓ_p-Majorizer)**.** *For $\rho(r) = \frac{1}{p}|r|^p$ and $\rho_\epsilon(r)$ defined in (7), the following holds: (i) $\rho_\epsilon(\cdot)$ is continuously differentiable, (ii) $\rho(r) \leq \rho_\epsilon(r), \forall r \in \mathbb{R}$, (iii) $\epsilon' \leq \epsilon \Rightarrow \rho_{\epsilon'}(r) \leq \rho_\epsilon(r)$, (iv) $\rho(r) = \lim_{\epsilon \to 0} \rho_\epsilon(r)$.*

Remark 1 (*Rethink Weight Update*). The weight update of Algorithm 1 coincides with rule (5) with $\rho = \rho_{\epsilon^{(t)}}$.

Remark 2 (*GNC for the ℓ_p-Loss*). Lemma 1 prompts a GNC strategy of minimizing ρ_ϵ (7) or even the ℓ_p-loss: decrease $\epsilon^{(t)}$ at each iteration t (Line 7, Algorithm 1).

Quadratic Majorizer. The smooth majorizer (7) is non-convex, and directly minimizing it can be hard. This is why the second level of majorization comes into play; the quadratic majorizer is the following quadratic function q_ϵ:

$$q_\epsilon(r, u) = \rho_\epsilon(u) + \frac{1}{2} \cdot \frac{r^2 - u^2}{\max\{|u|, \epsilon\}^{2-p}}. \tag{8}$$

Note that $q_\epsilon(r,u)$ is a shifted version of $\rho_\epsilon(u)$ by a carefully chosen amount, which makes $q_\epsilon(\cdot,u)$ into a majorizer of $\rho_\epsilon(\cdot)$. Indeed, Figure 1b shows that $q_{0.3}(\cdot,u)$ majorizes $\rho_{0.3}(\cdot)$ for $u=0.1$ and 0.5. More formally, we have:

Lemma 2 ($q_\epsilon(\cdot,u)$ is Quadratic ℓ_p-Majorizer)**.** *With $\rho(r)=\frac{1}{p}|r|^p$, $\rho_\epsilon(r)$ and $q_\epsilon(r,u)$ defined respectively in (7) and (8), we have $\rho_\epsilon(u)=q_\epsilon(u,u)$ and $\rho_\epsilon(r)\le q_\epsilon(r,u)$, $\forall r,u\in\mathbb{R}$.*

Remark 3 (*Rethink Weighted Least-Squares*). Recall $r_i^{(t)}:=r(v^{(t)},\mathcal{A}_i)$. The WLS step (4) of Algorithm 1 minimizes the quadratic majorizer $\sum_{i=1}^m q_{\epsilon^{(t)}}\big(r(\cdot,\mathcal{A}_i),r_i^{(t)}\big)$:

$$v^{(t+1)}\in\operatorname*{argmin}_{v\in\mathcal{C}}\sum_{i=1}^m\frac{r(v,\mathcal{A}_i)^2}{\max\{|r_i^{(t)}|,\epsilon^{(t)}\}^{2-p}}$$
$$=\operatorname*{argmin}_{v\in\mathcal{C}}\sum\nolimits_{i=1}^m q_{\epsilon^{(t)}}\big(r(v,\mathcal{A}_i),r_i^{(t)}\big)$$

$\texttt{GNC-IRLS}_p$ differs from the *majorization-minimization* paradigm [28, 60, 70] in that, at different iterations, $\texttt{GNC-IRLS}_p$ minimizes different quadratic majorizers, as controlled by the smoothing parameter $\epsilon^{(t)}$; in so doing, it blends (quadratic) majorization-minimization with GNC.

2.2. Convergence of $\texttt{GNC-IRLS}_p$

To obtain convergence results, we need appropriate assumptions on the constraint set $\mathcal{C}$ and residual function r. The first assumption is standard (*cf*. [3, Section 4.2]):

Assumption 1. $\mathcal{C}$ is non-empty and closed. The residual function $r(v,\mathcal{A}):\mathcal{C}\times\mathcal{D}\to\mathbb{R}_{\ge0}$ is *weakly coercive* in v:

$$\text{Either } \mathcal{C} \text{ is bounded or } \lim_{v\in\mathcal{C},\|v\|_2\to\infty} r(v,\mathcal{A})\to\infty. \tag{9}$$

Moreover, if $\|v\|_2\neq\infty$ then $r(v,\mathcal{A})\neq\infty$.

The next assumption is about *differentiability*:

Assumption 2. The residual function $r(v,\mathcal{A})$ is continuous in v everywhere, and differentiable in v if $r(v,\mathcal{A})\neq0$. Moreover, $r(v,\mathcal{A})^2$ is continuously differentiable in v.

Assumptions 1 and 2 are mild and easy to verify. With these assumptions, we prove the following:

Theorem 1 (Convergence of $\texttt{GNC-IRLS}_p$)**.** *Let $\{v^{(t)}\}_t$ be the iterates of Algorithm 1 with $\epsilon^{(t)}$ non-increasing and $\epsilon:=\lim_{t\to\infty}\epsilon^{(t)}>0$. Under Assumptions 1 and 2, every accumulation point of $\{v^{(t)}\}_t$ is a stationary point[3] of*

$$\min_{v\in\mathcal{C}}\sum\nolimits_{i=1}^m\rho_\epsilon\big(r(v,\mathcal{A}_i)\big). \tag{10}$$

[3]Stationary points are in the sense of [3, Section 5.3]; they satisfy a certain geometric condition that every local minimizer of (10) fulfills.

With a GNC schedule that creates a non-increasing sequence $\{\epsilon^{(t)}\}_t$ convergent to ϵ, $\texttt{GNC-IRLS}_p$ finds a stationary point of ρ_ϵ (Theorem 1), and ρ_ϵ approximates the ℓ_p-loss very well if ϵ is small (Lemma 1, Figure 1a). The convergence statement of "*accumulation points are stationary points*" in Theorem 1 is standard, and similar results can be found in optimization papers on IRLS or majorization-minimization, *e.g.*, [23, Thm 5.3 (ii)], [51, Thm 3.2], [60, Thm 1], [68, Thm 11], [53, Proposition 5], [48, Thm 1]. However, to our knowledge, Theorem 1 is the only result that holds for a *general* constraint set $\mathcal{C}$ and for minimizing *a sequence of majorizers* within the GNC framework.

Theorem 1 is proved by combing ideas of [23, 51] and [60], while generalizing their results. Unlike in Theorem 1, $\mathcal{C}$ is assumed to be convex and $\epsilon^{(t)}=\epsilon$ for all t in [60]. In [23, 51], $\mathcal{C}$ is defined by linear equality constraints and the residual function r is very specific, unlike in Theorem 1. Finally, as reviewed in Section 1.2, the result of [1] requires a condition that is hard to verify and their result does not apply to IRLS with the GNC strategy.

While stationary points are not necessarily local minimizers[3], convergence to them is perhaps the best one could guarantee in the setting where the objective (7) and constraint set $\mathcal{C}$ can both be non-convex. That said, a stronger convergence theory is possible given more assumptions on the problem and data. We will explore this in Section 3.

3. Convergence Rates for Robust Regression

While Theorem 1 is general, it does not reveal any convergence *speed*. Here, we compromise on generality and prove that $\texttt{GNC-IRLS}_p$ converges rapidly for *robust regression* [54]. Consider the following problem setup:

Problem 1 (Robust Regression)**.** For a *feature matrix* $\boldsymbol{A}=[\boldsymbol{a}_1,\dots,\boldsymbol{a}_m]^\top\in\mathbb{R}^{m\times n}$ and a *response vector* $\boldsymbol{y}=[y_1,\dots,y_m]^\top\in\mathbb{R}^m$, assume there is a ground truth vector $v^*=\boldsymbol{x}^*\in\mathbb{R}^n$ such that the residual vector $\boldsymbol{A}\boldsymbol{x}^*-\boldsymbol{y}$ has k non-zero entries; *i.e.*, there are k outliers and $m-k$ inliers among data $\{\mathcal{A}_i\}_{i=1}^m=\{(\boldsymbol{a}_i,y_i)\}_{i=1}^m$. The goal of robust regression is to recover $\boldsymbol{x}^*$ from data $\boldsymbol{A}$ and $\boldsymbol{y}$.

In Problem 1 we assume all inliers $(\boldsymbol{a}_i,y_i)$ are noiseless, *i.e.*, $r(v^*,\mathcal{A}_i)=|\boldsymbol{a}_i^\top\boldsymbol{x}^*-y_i|=0$. The extension to the noisy case is not hard (*cf*. [52, Thm 2], [41, Thm A.1]).

The GNC schedule is closely related to the convergence rates of IRLS. Informally, [52] suggests that the *linear* GNC schedule (as is commonly seen) leads to a *linear* rate. However, it is possible for IRLS to attain *superlinear* rates. In particular, defining the *superlinear GNC schedule*

$$\epsilon^{(t+1)}\leftarrow\beta(\epsilon^{(t)})^{2-p},\quad \beta>0, \tag{11}$$

we prove the following result:

Theorem 2. *Assume $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Initialize Algorithm 1 at $\boldsymbol{x}^{(0)}$ and $\epsilon^{(0)} > 0$ such that $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(0)}$. Denote by $r^*_{\text{min+}}$ the smallest non-zero number among the set of residuals $\{|\boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i|\}_{i=1}^m$. Define*

$$\alpha := \frac{\sqrt{5}\cdot 2^{2-p}}{0.99\cdot 0.516}\cdot\frac{1}{\left(r^*_{\text{min+}}\right)^{1-p}}\cdot\frac{\sqrt{k}\cdot\left(1.01\sqrt{k}+\sqrt{n}\right)}{(m-k)}. \quad (12)$$

Then the iterates $\{\boldsymbol{x}^{(t)}\}_{t\ge 0}$ produced by $\texttt{GNC-IRLS}_p$ with $p \in [0,1]$[4] and the GNC schedule (11) with $\beta \ge \alpha$ satisfy

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \begin{cases} \beta^t\cdot\epsilon^{(0)} & p = 1 \\ \beta^{\frac{(2-p)^t-1}{1-p}}\cdot\left(\epsilon^{(0)}\right)^{(2-p)^t} & p\in[0,1) \end{cases} \quad (13)$$

with probability at least $1-(P_0+P_1+P_2)$, where

$$P_0 := \exp\big(-\tilde{\Omega}(n)\big),\quad P_1 := \exp(-\tilde{\Omega}(k-n)),$$
$$P_2 := \exp\big(-\tilde{\Omega}(m-k-n\log n)\big). \quad (14)$$

We discuss several aspects of Theorem 2: **(i)** the probabilities (14), **(ii)** the condition $\beta \ge \alpha$ (12), **(iii)** its relation to prior works, and **(iv)** the interaction of the GNC schedule (11), error bound (13) and condition $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(0)}$.

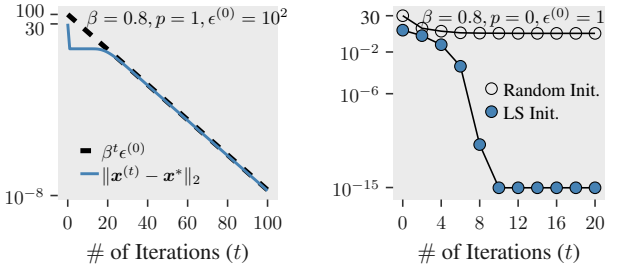
(i) In the probability terms of (14), $\tilde{\Omega}$ stands for the standard big-Ω notation, with the difference that $\tilde{\Omega}$ also suppresses logarithmic terms. We wish P_0, P_1, P_2 of (14) to be small so that (13) holds with high probability. This is true whenever n is large (P_0), $k \gg n$ (P_1), and $m-k \gg n\log n$ (P_2). It seems counterintuitive to ask for the number k of outliers to be far larger than n, but the challenging case of Problem 1 occurs exactly when k is large. If k were small, then an alternative proof would give $P_1 = \exp(-\tilde{\Omega}(n))$. Such proof is much simpler, which is why we omit it.

(ii) We wish α to be as small as possible as this would make it easier to set the factor β in the GNC schedule (11). Ignoring the values of the constants in (12), α mainly involves two terms, $r^*_{\text{min+}}$ and $O((k+\sqrt{kn})/(m-k))$. Since $r^*_{\text{min+}}$ measures the minimum residual of outliers at the ground truth $\boldsymbol{x}^*$, we expect it to be a *large* constant. Since $p \in [0,1]$, a large $(r^*_{\text{min+}})^{1-p}$ would make α small; on the other hand, for $p = 1$, α does not depend on $r^*_{\text{min+}}$ at all. Then note that we require a large k in **(i)**, but this might make the second term $O((k+\sqrt{kn})/(m-k))$ and therefore α very large. The rescue is in the denominator: α is small if the number $m-k$ of inliers (or the inlier rate) is large.

(iii) Theorem 2 is motivated by [52, Thm 1], over which we make some improvements. First, the GNC schedule of [52] sets $\epsilon^{(t+1)} \leftarrow \beta\epsilon^{(t)}$ if $\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2 \le 2\beta\epsilon^{(t)}$, or otherwise $\epsilon^{(t+1)} \leftarrow \epsilon^{(t)}$. We simplify and generalize it into (11). Also, [52] is limited to the case $p = 1$, but Theorem 2 holds for any $p \in [0,1]$; we derive some technical lemmas that overcome the challenges of the non-convex case $p < 1$.

The final point **(iv)** has more delicate interpretations and ramifications, and we discuss it in Sections 3.1-3.4 next.

[4] It is valid to run $\texttt{GNC-IRLS}_p$ with $p = 0$, as we justified in [58].



(a) Decay of Error (Bound) (13) (b) Decay of Error $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2$

Figure 2. (2a, Section 3.1): Error bound $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \beta^t\epsilon^{(0)}$ (13) with initialization $\boldsymbol{x}^{(0)} \sim \mathcal{N}(0, 100\boldsymbol{I}_n)$. (2b, Section 3.2): Errors of $\texttt{GNC-IRLS}_0$ at each iteration with least-squares versus random initialization. 100 trials, $k = 400, m = 1000, n = 10$.

3.1. Global Linear Convergence at $p = 1$?

For the error bound $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \beta^t\cdot\epsilon^{(0)}$ of (13) to make sense, one needs to set $\beta < 1$, then the condition $\alpha \le \beta$ in Theorem 2 implies $\alpha < 1$. As discussed, we have $\alpha < 1$ if the inlier rate is large. Indeed, assuming $k, m \gg n$ and bringing now the constant of (12) into the picture, we see that $\alpha < 1$ amounts to $m - k > 2.02\sqrt{5}/(0.99\times 0.516)k$. This defines an outlier rate below which Theorem 2 holds. This also implies Theorem 2 is optimal in an information-theoretical sense (*e.g.*, it only requires m to be linear in k).

For $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \beta^t\cdot\epsilon^{(0)}$ to be true, Theorem 2 requires $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(0)}$ (among other assumptions). Given any initialization $\boldsymbol{x}^{(0)}$, one can choose a large $\epsilon^{(0)}$ such that $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(0)}$, so [52] claimed this is a *global* linear convergence. But this claim is imprecise, e.g., if $\boldsymbol{x}^{(0)}$ is the least-squares initialization and $\epsilon^{(0)}$ is larger than all residuals $|\boldsymbol{a}_i^\top\boldsymbol{x}^{(0)} - y_i|$, then all weights $w_i^{(1)}$ are equal to $\epsilon^{(0)}$, and we would get $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)}$. As such, the error would not decrease until $\epsilon^{(t)}$ becomes smaller: Figure 2a shows that $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2$ "waits" for almost 20 iterations to decay together with the bound $\beta^t\epsilon^{(0)}$ ($\epsilon^{(0)} = 100$). This overlooked phenomenon caused by large $\epsilon^{(0)}$ is what we call a *burn-in* period. Interestingly, the burn-in period does not mean that our bound (13) is incorrect, but just that it might be loose for large $\epsilon^{(0)}$ in early iterations.

Figure 2a shows that $\texttt{GNC-IRLS}_1$ needs more than 100 iterations to reach machine accuracy. We improve this next, by considering $p \in [0,1)$ (Sections 3.2-3.4).

3.2. Local Quadratic Convergence at $p = 0$

Theorem 2 with $p \in [0,1)$ is better elaborated in the case $p = 0$, for which (13) gives $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le (\beta\epsilon^{(0)})^{2^t}/\beta$. This corresponds to a quadratic convergence rate. Again, the error bound $(\beta\epsilon^{(0)})^{2^t}/\beta$ only makes sense if $\beta\epsilon^{(0)} < 1$, or if we set $\epsilon^{(0)}$ small (note that this time we do not require $\beta < 1$). In turn, Theorem 2 would demand an initialization

$\boldsymbol{x}^{(0)}$ such that $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \leq \epsilon^{(0)}$. As corroborated by Figure 2b, GNC-IRLS$_0$ with random ("bad") initialization and small $\epsilon^{(0)}$ fails, but the least-squares initialization seems to be good enough, allowing GNC-IRLS$_0$ to converge at a quadratic rate, within 10 iterations, where "*the number of correct digits doubles at each iteration*" [14, Section 9.5.3]. Powerful as it might seem, quadratic (and superlinear) convergence is doomed to be local and in general cannot hold for all initializations (*cf*. Newton's method); we refer the reader to our prior work [58] for different insights into the quadratic rates of IRLS for robust regression.

We believe the *next best* convergence guarantees are these two: (i) prove that some IRLS variant has two-phase convergence, first global linear and then local quadratic, (ii) derive a suitable choice of $\epsilon^{(0)}$, β, and $\boldsymbol{x}^{(0)}$ for which quadratic convergence happens starting from the first iteration. We discuss these next in Section 3.3 and 3.4.

3.3. Graduated Rates From Linear to Quadratic?

Consider the following slight twist over Algorithm 1:

(a) With some initialization $\boldsymbol{x}^{(0)}$ and a sufficiently large $\epsilon^{(0)}$ such that $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \leq \epsilon^{(0)}$, run Algorithm 1 with $p = 1$ and GNC schedule (11), until $\beta^t \epsilon^{(0)} < 1$. Theorem 2 suggests that $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \leq \beta^t \epsilon^{(0)}$.

(b) Re-run Algorithm 1 with $\boldsymbol{x}^{(0)} := \boldsymbol{x}^{(t)}$, $\epsilon^{(0)} := \beta^t \epsilon^{(0)}$, $p = 0$, and schedule (11). Quadratic convergence (13) of Theorem 2 is now meaningful, since $\beta \epsilon^{(0)} < 1$.

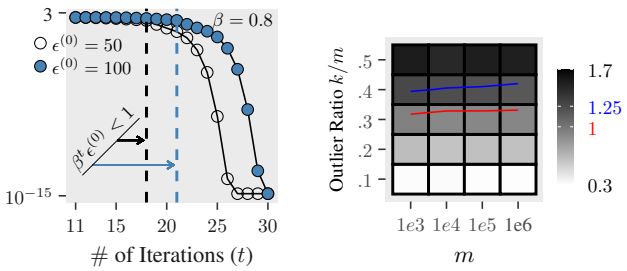
Simply put, the above twist switches from $p = 1$ to $p = 0$ if $\beta^t \epsilon^{(0)} < 1$, resulting in a graduated rate from global linear to local quadratic. Such a graduated rate guarantee seems rare; we can only find it in [21]. Figure 3a shows that when we switch to $p = 0$, the convergence ensues in the next 10 iterations. A deficiency is that this twist also comes with a burn-in period (*cf*. Section 3.1), after which it is possible that the linear convergence phase is skipped and the quadratic convergence takes place directly (Figure 3a).

3.4. Quadratic Rates From The First Iteration?

The IRLS twist of Section 3.3 can take 30 iterations to converge if $\epsilon^{(0)}$ is large (Figure 3a). But we also saw that, with the least-squares initialization and $\beta \epsilon^{(0)} = 0.8 < 1$, GNC-IRLS$_0$ converges within 10 iterations, at a quadratic rate (Figure 2b). We now argue that it is theoretically possible for GNC-IRLS$_0$ to have quadratic rates *starting from as early as the first iteration*. For this, we first prove:

Proposition 1. *Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries with $m \geq n$. Let $\boldsymbol{x}^\dagger := (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{y}$. With probability at least $1 - \exp(-\Omega(k)) - \exp(-\Omega(m))$, we have*

$$\|\boldsymbol{x}^\dagger - \boldsymbol{x}^*\|_2 \leq \frac{(1.01\sqrt{k} + \sqrt{n}) \cdot \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2}{(0.99\sqrt{m} - \sqrt{n})^2}. \quad (15)$$



(a) Decay of $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2$ (b) Error of Least-Squares

Figure 3. (3a, Section 3.3): From linear to quadratic rates, vertical lines indicating the transition takes place; $k = 400, m = 1000$, $\boldsymbol{x}^{(0)} \sim \mathcal{N}(0, \boldsymbol{I}_n)$. (3b, Section 3.4): Error $\|\boldsymbol{x}^\dagger - \boldsymbol{x}^*\|_2$ of the least-squares estimator $\boldsymbol{x}^\dagger$. We set 100 trials, $n = 10$.

We wish $\|\boldsymbol{x}^\dagger - \boldsymbol{x}^*\|_2 \leq 1$; if so we can set $\epsilon^{(0)} = 1$ and $\beta < 1$, achieving quadratic rates with initialization $\boldsymbol{x}^\dagger$ (Theorem 2). This is possible if k/m is small and $m, k \gg n$; see (15). This is also empirically confirmed in Figure 3b, where $\|\boldsymbol{x}^\dagger - \boldsymbol{x}^*\|_2 \leq 1$ for fewer than 30% outliers.

Implementation Details. The discussions so far suggest the following implementation of GNC-IRLS$_p$. Set $\epsilon^{(0)} = 1$, $p = 0$. Initialize it via least-squares. Set β smaller than 1; we always use $\beta = 0.8$. Theorem 1 suggests to let $\{\epsilon^{(t)}\}_t$ converge to some $\epsilon > 0$. In the noiseless case, we set $\epsilon = 10^{-16}$. Otherwise, if we are given an *inlier threshold* c such that $r(v^*, d_i) \leq c$ for all inliers d_i, then we set $\epsilon = c$.

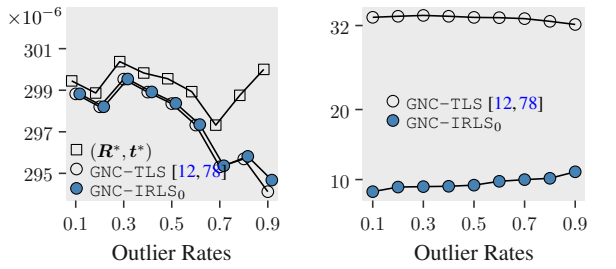
4. Experiments: Lp Versus TLS

Here we compare GNC-TLS [12, 78] and GNC-IRLS$_p$. For more extensive experiments of IRLS and its variants, see, *e.g.*, [1, 20, 25, 27, 46, 64, 65, 69, 78].

Experimental Setup. We contextualize our experiment in the application of *point cloud registration*. In this application, each sample d_i is a 3D point pair $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, the variable v consists of a 3D rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$, and the residual function is $r(v; d_i) = \|\boldsymbol{y}_i - \boldsymbol{R}\boldsymbol{x}_i - \boldsymbol{t}\|_2$. The corresponding weighted least-squares problem (4) is solved by eliminating the translation first and then applying SVD [38].

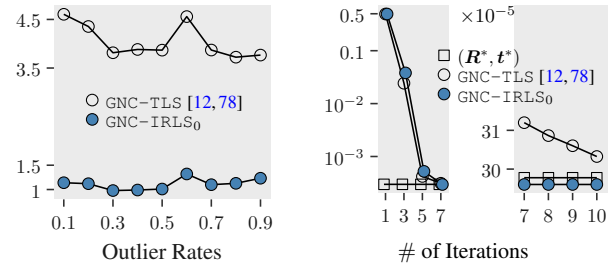
Data. We randomly sample k outlier point pairs $(\boldsymbol{y}_j, \boldsymbol{x}_j)$ so that $\boldsymbol{y}_j \sim \mathcal{N}(0, \boldsymbol{I}_3)$ and $\boldsymbol{x}_j \sim \mathcal{N}(0, \boldsymbol{I}_3)$; here $\boldsymbol{I}_3$ denotes the 3×3 identity matrix. To get $m - k$ inlier pairs $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, we randomly sample $\boldsymbol{x}_i$ from $\mathcal{N}(0, \boldsymbol{I}_3)$ and compute $\boldsymbol{y}_i = \boldsymbol{R}^* \boldsymbol{x}_i + \boldsymbol{t}^* + \boldsymbol{\epsilon}_i$. Here, $\boldsymbol{R}^*$ and $\boldsymbol{t}^*$ are randomly generated ground truth rotation and translation respectively, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 0.01^2 \boldsymbol{I}_3)$ is some Gaussian noise. We set $c^2 = 0.01^2 \times 5.54^2$, so each inlier $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ satisfies $\|\boldsymbol{y}_i - \boldsymbol{R}^* \boldsymbol{x}_i - \boldsymbol{t}^*\|_2 \leq c^2$ with probability $\geq 1 - 10^{-6}$.

Metric. Given a rotation $\boldsymbol{R}$, translation $\boldsymbol{t}$, and ground truth inlier index set $\mathcal{I}^*$, we can calculate the *average inlier resid-*



(a) Average Inlier Residual

(b) # of Iterations



(c) Running Times (msec)

(d) Average Inlier Residual

Figure 4. Comparison of GNC−IRLS$_0$ and GNC−TLS for point cloud registration. 100 trials, $m = 1000$. 4d: 900/1000 outliers.

ual $\sum_{i\in\mathcal{I}^*} \|\boldsymbol{y}_i - \boldsymbol{R}\boldsymbol{x}_i - \boldsymbol{t}\|_2/(m-k)$. This is used to measure the errors made by the algorithms to evaluate.

Results. As the outlier rate varies from 10% to 90%, GNC−IRLS$_0$ and GNC−TLS entail almost the same average inlier residual (Figure 4a). Their errors are even smaller than those at the ground truth $(\boldsymbol{R}^*, \boldsymbol{t}^*)$, which suggests that the performance of both algorithms cannot be further improved for such experiments. But note that they could fail for more than 900/1000 outliers (which was reported in prior works, so we did not provide a plot here) and that the breakdown points will change for different data distributions and different geometric problems.

What can actually be improved is the convergence rate: GNC−IRLS$_0$ terminates in 10 iterations, while GNC−TLS takes 32 (Figure 4b), indicating that GNC−IRLS$_0$ is 3 times faster (Figure 4c). For fair[5] comparison, both methods are implemented to terminate under the same condition, that is whenever the difference of the minimum values of (4) between two consecutive iterations is smaller than 10^{-10} (other thresholds, *e.g.*, $10^{-6}, 10^{-16}$, lead to similar results).

The errors of GNC−TLS decrease as fast as GNC−IRLS$_0$ (Figure 4d, left). At first glance, this seems counterintuitive because GNC−TLS comes with a linear GNC schedule (*cf*. Section 1.2) and is thus expected to converge linearly (*cf*. [52]), as opposed to the quadratic rate of GNC−IRLS$_0$ (*cf*. Theorem 2). With hindsight, this might be a natural consequence of the weighting strategy of GNC−TLS (*cf*. [78, Eq. (14)], (5), (6)): Weight 0 is set if the residual is *particularly large*, and this could completely rule out some *obvious* outliers at early iterations (and similarly for *particularly small* residuals), resulting in a fast decrease of errors. But this weighting scheme brings diminishing gains in later iterations, where the errors of GNC−TLS decrease only linearly (Figure 4d, right). The final observation is that GNC−IRLS$_0$ reaches an error smaller than that of $(\boldsymbol{R}^*, \boldsymbol{t}^*)$ at iteration 7, but it requires a few more iterations to terminate (similarly for GNC−TLS). This implies the termination criterion is sub-optimal (it is hard to design a provably better one).

Finally, Figures 4b and 4d show that GNC−TLS has an error of $\leq 10^{-3}$ already at iteration 10, but, unnecessarily, it terminates at iteration 32. We improve this in Section 5, without even changing the termination criterion.

[5] It is slightly unfair to GNC−IRLS$_0$ as its weights are typically larger.

5. MS−GNC−TLS: Improving GNC−TLS

> *... it indicates that GNC can fail, and that there is therefore no point in looking for a general proof of correctness.*
>
> Andrew Blake & Andrew Zisserman [12]

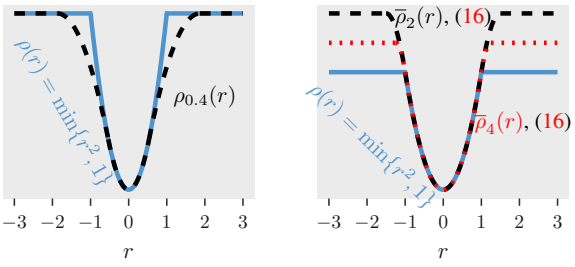
In this section, we improve GNC−TLS [12, 78] from two aspects, as respectively motivated by two ideas that we have developed for the ℓ_p-loss, namely *majorization* (Section 2) and *superlinear GNC schedule* (Section 3). Majorization guarantees a monotonic decrease of the objective and the eventual convergence (*cf*. Theorem 1), and the superlinear GNC schedule speeds up convergence (*cf*. Theorem 2).

Majorization. To motivate the need for majorizing the TLS loss $\rho(r) = \min\{r^2, c^2\}$, recall GNC−TLS uses ρ_μ (6) to approximate $\rho(\cdot)$. The issue is that $\rho_\mu(\cdot)$ relaxes $\rho(\cdot)$ and approximates it *from below*, and hence $\rho_\mu(r) \leq \rho(r), \forall \mu > 0$ (Figure 5a). This makes a convergence analysis difficult.

We propose the following smooth function

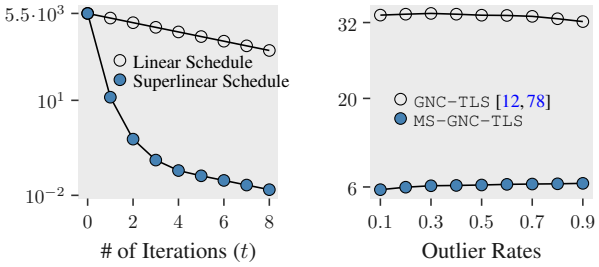
$$\overline{\rho}_\mu(r) = \begin{cases} r^2, & \text{if } |r| \leq c, \\ \frac{\mu+1}{\mu}c^2, & \text{if } |r| \geq \frac{\mu+1}{\mu}c, \\ -\mu r^2 + 2(1+\mu)c|r| - (1+\mu)c^2, & \text{o/w}, \end{cases} \tag{16}$$

to majorize the TLS loss $\rho(r)$; see Figure 5b. Since both $\rho_\mu(r)$ (6) and $\overline{\rho}_\mu(r)$ approach $\rho(r)$ as $\mu \to \infty$, one might expect comparable performance. However, a crucial difference is that, with the majorizer $\overline{\rho}_\mu(r)$, convergence guarantees easily ensue. Indeed, $\overline{\rho}_\mu(r)$ is akin to the smooth majorizer (7) of the ℓ_p-loss, and one could construct a quadratic majorizer for $\overline{\rho}_\mu(r)$, which enables an IRLS + GNC scheme (*cf*. Remarks 1-3, Section 1.2). In particular, this IRLS vari-



(a) Approximation $\rho_\mu(r)$ of [12] (b) Smooth Majorizer (16)

Figure 5. The TLS loss $\rho(r)$ and its surrogates.



(a) $c/\mu^{(t)}$ (b) # of Iterations (t)

Figure 6. 6a: Length $c/\mu^{(t)}$ of the interval that corresponds to non-binary weights (17) with $c = 0.0554$, $\mu^{(0)} = 10^{-5}$, $\gamma = 1.4$. 6b: Number of iterations at which the algorithms terminate.

ant involves (i) weight update using (5) with $\rho = \overline{\rho}_{\mu^{(t)}}$, *i.e.*,

$$w_i^{(t+1)} = \begin{cases} 1, & \text{if } r_i^{(t)} \le c, \\ 0, & \text{if } r_i^{(t)} \ge \frac{\mu^{(t)}+1}{\mu^{(t)}} c, \\ \frac{c(1+\mu^{(t)})}{r_i^{(t)}} - \mu^{(t)}, & \text{o/w}, \end{cases} \tag{17}$$

and (ii) updating $\mu^{(t+1)}$ based on some GNC schedule.

We prove the following result to accompany Theorem 1.

Theorem 3 (Convergence of majorized GNC-TLS)**.** *Let $\{v^{(t)}\}_t$ be the iterates of IRLS with weight update (17) and a GNC schedule $\{\mu^{(t)}\}_t$. Assume $\{v^{(t)}\}_t$ is bounded, i.e., $\|v^{(t)}\|_2 < \infty$ $(\forall t)$. Suppose $\{\mu^{(t)}\}_t$ is non-decreasing and converges to $\mu < \infty$. Under Assumptions 1 and 2, every accumulation point of $\{v^{(t)}\}_t$ is a stationary point of*

$$\min_{v \in \mathcal{C}} \sum_{i=1}^{m} \overline{\rho}_\mu\big(r(v, \acute{a}_i)\big). \tag{18}$$

Superlinear Schedule. Motivated by (11) (with $p = 0$) and discussions in Sections 3.2-3.3, we propose the update rule

$$\mu^{(t+1)} \leftarrow \begin{cases} \gamma\sqrt{\mu^{(t)}} & \mu^{(t)} \le 1 \\ \gamma\mu^{(t)} & \mu^{(t)} > 1 \end{cases}, \quad \gamma > 1, \tag{19}$$

as our GNC schedule. Denote by MS-GNC-TLS the resulting IRLS method that optimizes (16) with schedule (19).

The intuition behind the superlinear schedule (19) is as follows. With (19), the interval $(c, c + c/\mu^{(t)})$ of (17) that produces non-binary weights shrinks faster than the linear schedule $\mu^{(t+1)} \leftarrow \gamma\mu^{(t)}$ (Figure 6a), thus the superlinear schedule makes it happen earlier that all weights become binary, which is a good indicator for convergence. Note though that this argument does not prove (MS-)GNC-TLS converges, as it does not exclude the case that (MS-)GNC-TLS could produce different binary weights at consecutive iterations (*cf*. [10] and [4, Thm 15]).

Under the setting of Figure 4c, MS-GNC-TLS takes 6 iterations to converge (Figure 6b). It is even faster than GNC-IRLS$_0$ as it benefits from combining soft and hard thresholding (17). In this experiment, MS-GNC-TLS and GNC-TLS result in basically the same error upon convergence; it is just that GNC-TLS does not monotonically decrease the objective, and that its linear GNC schedule is more conservative than the proposed superlinear one.

Implementation Details. With the superlinear schedule (19), $\mu^{(t)}$ increases very fast, so one could set $\mu^{(0)} \leftarrow 10^{-15}$ such that MS-GNC-TLS can still terminate within 10 iterations. However, schedule (19) is *double-edged*: If $\mu^{(t)}$ increases so fast that all residuals are larger than $\frac{\mu^{(t)}+1}{\mu^{(t)}}c$, then all weights would be zero as per (17) and MS-GNC-TLS might fail. Fortunately, this situation can be prevented if we *slow down*: replace $\mu^{(t+1)} \leftarrow \gamma\sqrt{\mu^{(t)}}$ of (19) with $\mu^{(t+1)} \leftarrow \gamma(\mu^{(t)})^{1/(2-p)}$ for a larger $p \in (0, 1]$.

6. Conclusion, Limitations, and Future Work

Conclusion. While IRLS and GNC have often been viewed as different techniques [44,45,64,78,83], we reconcile them with an emphasis on a theoretical understanding of convergence properties and their relation with GNC schedules. Two messages are (i) that a majorization strategy should be constructed for guaranteeing *convergence* (Theorems 1 and 3), and (ii) that a superlinear GNC schedule should be considered for guaranteeing *convergence rates* (Theorem 2).

Limitations & Future Work. IRLS and its variants would break down if the number $m - k$ of inliers is close to the number n of variables, say if $m - k < 3n$. For geometric vision problems, n is small (*e.g.*, $n = 6$ for point cloud registration), so IRLS might fail if, for example, $m - k < 18$. In fact, for small m, other methods (*e.g.*, RANSAC [7, 29], outlier removal [56], or semidefinite relaxations [33, 57, 81]) are efficient, accurate, and are thus recommended.

A limitation of the TLS loss $\rho(r) = \min\{r^2, c^2\}$ is the need to choose a threshold parameter c. Ideally, it should be chosen as small as possible but larger than every inlier residual; see [59] for a related discussion. Prior works [4, 69]

tried to dispense with c^2, but it was at the expense of introducing other parameters. This issue might be solved by changing c in a GNC style at each iteration, which implies future work of designing a GNC schedule for c and studying its interplay with another GNC parameter $\mu^{(t)}$. On the theory side, we note that extending the analysis of Theorem 2 beyond ℓ_p-losses remains to be studied in future work.

Acknowledgements. This work was supported by grants NSF 1704458, NSF 1934979, ONR MURI 503405-78051, and the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative.

References

[1] Khurrum Aftab and Richard Hartley. Convergence of iteratively re-weighted least squares to robust M-estimators. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 1, 2, 3, 4, 6, 22

[2] Chris Aholt, Sameer Agarwal, and Rekha Thomas. A QCQP approach to triangulation. In *European Conference on Computer Vision*, 2012. 2

[3] Niclas Andréasson, Anton Evgrafov, and Michael Patriksson. *An Introduction to Continuous Optimization: Foundations and Fundamental Algorithms*. Courier Dover Publications, 2020. 4, 16, 23

[4] Pasquale Antonante, Vasileios Tzoumas, Heng Yang, and Luca Carlone. Outlier-robust estimation: Hardness, minimally tuned algorithms, and applications. *IEEE Transactions on Robotics*, 2021. 1, 2, 8

[5] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):698–700, 1987. 2

[6] Erik Ask, Olof Enqvist, and Fredrik Kahl. Optimal geometric fitting under the truncated L_2-norm. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[7] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[8] Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015. 2

[9] Amir Beck and Shoham Sabach. Weiszfeld's method: Old and new results. *Journal of Optimization Theory and Applications*, 164(1):1–40, 2015. 2

[10] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in Neural Information Processing Systems*, 2015. 2, 8

[11] Michael J Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996. 1

[12] Andrew Blake and Andrew Zisserman. *Visual Reconstruction*. MIT Press, 1987. 2, 6, 7, 8

[13] Roger Joseph Boscovich. De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impessa. *Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii*, 4:353–396, 1757. 1

[14] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 6

[15] Jesus Briales and Javier Gonzalez-Jimenez. Convex global 3D registration with lagrangian duality. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[16] Lucas Brynte, Viktor Larsson, José Pedro Iglesias, Carl Olsson, and Fredrik Kahl. On the tightness of semidefinite relaxations for rotation estimation. *Journal of Mathematical Imaging and Vision*, 64(1):57–67, 2022. 14

[17] Tony F Chan and Pep Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM Journal on Numerical Analysis*, 36(2):354–367, 1999. 2

[18] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007. 1

[19] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008. 2

[20] Avishek Chatterjee and Venu Madhav Govindu. Robust relative rotation averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):958–972, 2017. 1, 2, 6, 22

[21] Bintong Chen and Naihua Xiu. A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen–Mangasarian smoothing functions. *SIAM Journal on Optimization*, 9(3):605–623, 1999. 6

[22] David Coleman, Paul Holland, Neil Kaden, Virginia Klema, and Stephen C Peters. A system of subroutines for iteratively reweighted least squares computations. *ACM Transactions on Mathematical Software*, 6(3):327–336, 1980. 1, 2

[23] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010. 1, 2, 3, 4

[24] DQF De Menezes, Diego Martinez Prata, Argimiro R Secchi, and José Carlos Pinto. A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, 147:107254, 2021. 1

[25] Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P Robinson, René Vidal, Laurent Kneip, and Manolis C Tsakiris. Robust homography estimation via dual principal component pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6

[26] Tianyu Ding, Zhihui Zhu, Tianjiao Ding, Yunchen Yang, René Vidal, Manolis C. Tsakiris, and Daniel Robinson. Noisy dual principal component pursuit. In *International Conference on Machine Learning*, 2019. 2, 14

[27] Wenhua Dong, Xiao-jun Wu, and Josef Kittler. Sparse subspace clustering via smoothed ℓ_p minimization. *Pattern Recognition Letters*, 125:206–211, 2019. 1, 2, 6

[28] Taosha Fan and Todd Murphey. Majorization minimization methods for distributed pose graph optimization with convergence guarantees. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. 4

[29] Martin A Fischler and Robert C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 8

[30] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser, 1809. 1

[31] Stuart Geman and Donald E. McClure. Bayesian image analysis: An application to single photon emission tomography. In *Proceedings of the American Statistical Association*, 1985. 22

[32] Frank R Hampel. The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27(2):95–107, 1985. 1

[33] Linus Härenstam-Nielsen, Niclas Zeller, and Daniel Cremers. Semidefinite relaxations for robust multiview triangulation. Technical report, arXiv:2301.11431 [cs.CV], 2023. 8

[34] Richard Hartley. Tutorial notes on IRLS. `http://users.cecs.anu.edu.au/~hartley/Papers/PDF/Hartley:IRLS14.pdf`, 2014. Accessed: September 2022. 2

[35] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International journal of computer vision*, 103(3):267–305, 2013. 14

[36] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 14, 22

[37] Joel A Hesch and Stergios I Roumeliotis. A direct least-squares (DLS) method for PnP. In *International Conference on Computer Vision*, 2011. 2, 13

[38] Berthold KP Horn, Hugh M Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7):1127–1135, 1988. 2, 6, 13

[39] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. 1, 3

[40] Laurent Kneip, Hongdong Li, and Yongduek Seo. UPnP: An optimal $O(n)$ solution to the absolute pose problem with universal applicability. In *European Conference on Computer Vision*, 2014. 13

[41] Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for basis pursuit with global linear convergence rate. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 4

[42] Christian Kümmerle and Claudio M Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *International Conference on Machine Learning*, 2021. 1

[43] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. 33

[44] Huu Le and Christopher Zach. A graduated filter method for large scale robust estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[45] Huu Le and Christopher Zach. Robust fitting with truncated least squares: A bilevel optimization approach. In *International Conference on 3D Vision*, 2021. 2, 8, 22

[46] Seong Hun Lee and Javier Civera. HARA: A hierarchical approach for robust rotation averaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[47] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Courcier, 1806. 1

[48] Gilad Lerman and Tyler Maunu. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018. 2, 4

[49] Gilad Lerman, Michael B. McCoy, Joel A. Tropp, and Teng Zhang. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics*, 15(2):363–410, 2015. 2

[50] Goran Marjanovic and Victor Solo. On ℓ_q optimization and matrix completion. *IEEE Transactions on Signal Processing*, 60(11):5714–5724, 2012. 1

[51] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012. 1, 3, 4

[52] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *International Conference on Artificial Intelligence and Statistics*, 2019. 2, 3, 4, 5, 7

[53] Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, 2015. 2, 4

[54] Dianne P O'Leary. Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications*, 11(3):466–480, 1990. 1, 4

[55] Frank C Park and Bryan J Martin. Robot sensor calibration: Solving $AX = XB$ on the Euclidean group. *IEEE Transactions on Robotics and Automation*, 10(5):717–721, 1994. 2

[56] Álvaro Parra Bustos and Tat-Jun Chin. Guaranteed outlier removal for point cloud registration with correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2868–2882, 2018. 8

[57] Liangzu Peng, Mahyar Fazlyab, and René Vidal. Semidefinite relaxations of truncated least-squares in robust rotation search: Tight or not. In *European Conference on Computer Vision*, 2022. 8

[58] Liangzu Peng, Christian Kümmerle, and René Vidal. Global linear and local superlinear convergence of IRLS for nonsmooth robust regression. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 5, 6

[59] Liangzu Peng, Manolis C. Tsakiris, and René Vidal. ARCS: Accurate rotation and correspondences search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 8

[60] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013. 3, 4

[61] William J.J. Rey. *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer Science & Business Media, 1983. 1

[62] David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. SE-Sync: A certifiably correct algorithm for synchronization over the special Euclidean group. *The International Journal of Robotics Research*, 38(2-3):95–125, 2019. 2, 14

[63] Gerald Schweighofer and Axel Pinz. Globally optimal $O(n)$ solution to the PnP problem for general camera models. In *British Machine Vision Conference*, 2008. 13

[64] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal and robust category-level perception: Object pose and shape estimation from 2D and 3D semantic keypoints. Technical report, arXiv:2206.12498 [cs.CV], 2022. 2, 6, 8, 13

[65] Yunpeng Shi and Gilad Lerman. Message passing least squares framework and its application to rotation synchronization. In *International Conference on Machine Learning*, 2020. 6

[66] Chitturi Sidhartha and Venu Madhav Govindu. It is all in the weights: Robust rotation averaging revisited. In *International Conference on 3D Vision*, 2021. 2, 3

[67] Torbjorn Smith and Olav Egeland. Dynamical pose estimation with graduated non-convexity for outlier robustness. *Modeling, Identification and Control*, 43(2):79–89, 2022. 2

[68] Junxiao Song, Prabhu Babu, and Daniel P Palomar. Sparse generalized eigenvalue problem via smooth optimization. *IEEE Transactions on Signal Processing*, 63(7):1627–1642, 2015. 3, 4

[69] Lei Sun. IMOT: General-purpose, fast and robust estimation for spatial perception problems with outliers. Technical report, arXiv:2204.01324v1 [cs.CV], 2022. 2, 6, 8

[70] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016. 4

[71] Manolis C. Tsakiris and René Vidal. Dual principal component pursuit. *Journal of Machine Learning Research*, 19(18):1–50, 2018. 2

[72] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. 33

[73] Sergey Voronin and Ingrid Daubechies. An iteratively reweighted least squares algorithm for sparse regularization. Technical report, arXiv:1511.08970v3 [math.NA], 2015. 2

[74] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. 33

[75] Jianqiao Wangni, Dahua Lin, Ji Liu, Kostas Daniilidis, and Jianbo Shi. Towards statistically provable geometric 3D human pose recovery. *SIAM Journal on Imaging Sciences*, 14(1):246–270, 2021. 13

[76] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937. 2

[77] David Wipf and Srikantan Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, 2010. 2

[78] Heng Yang, Pasquale Antonante, Vasileios Tzoumas, and Luca Carlone. Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection. *IEEE Robotics and Automation Letters*, 5(2):1127–1134, 2020. 1, 2, 6, 7, 8

[79] Heng Yang and Luca Carlone. In perfect shape: Certifiably optimal 3D shape reconstruction from 2D landmarks.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 13

[80] Heng Yang and Luca Carlone. One ring to rule them all: Certifiably robust geometric perception with outliers. In *Advances in Neural Information Processing Systems*, 2020. 2, 3

[81] Heng Yang and Luca Carlone. Certifiable outlier-robust geometric perception: Exact semidefinite relaxations and scalable global optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 8

[82] Heng Yang, Jingnan Shi, and Luca Carlone. TEASER: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021. 2, 3

[83] Christopher Zach and Guillaume Bourmaud. Descending, lifting or smoothing: Secrets of robust cost optimization. In *European Conference on Computer Vision*, 2018. 8

[84] Ji Zhao. An efficient solution to non-minimal case essential matrix estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[85] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Astrom, and Masatoshi Okutomi. Revisiting the PnP problem: A fast, general and optimal solution. In *IEEE International Conference on Computer Vision*, 2013. 13

[86] Pengwei Zhou, Xuexun Guo, Xiaofei Pei, and Ci Chen. T-TOAM: Truncated least squares Lidar-only odometry and mapping in real time. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 1, 2

[87] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, 2016. 2

[88] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3D shape reconstruction from 2D landmarks: A convex formulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 13

[89] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics*, 33(4):1–12, 2014. 22

A. Structure of The Appendix

We structure the appendix as follows:

- In Section B, we present some geometric vision problems whose residuals and constraints satisfy Assumptions 1 and 2.
- In Section C, we prove Lemma 1, Lemma 2, and Theorem 1 of the main paper.
- In Section D, we discuss the design of smooth and quadratic majorizers for the TLS loss, and prove Theorem 3.
- In Section E, we extend our convergence theory for the ℓ_p and TLS losses to a general class of loss functions. The result is Theorem 4, which can be applied to general classes of loss functions, residual functions, and constraints. We also list some losses in Example 1. As a side product, this section sheds light on how the weight update rule (5) shows up.
- In Section F, we prove Theorem 2 of the main paper. For this, we invoke Theorem 5 of Section G multiple times.
- In Section H, we prove Proposition 1 of the main paper.
- In Section I, we reference several results from high-dimensional statistics.

B. Geometric Vision Problems Satisfy Assumptions 1 and 2

Here, we list several geometric vision problems that satisfy Assumptions 1 and 2, as one can verify very easily. Therefore, Theorems 1 and 3 of the main paper (and the more general Theorem 4 in Section E) can be applied to these problems, yielding convergence guarantees for the proposed IRLS variants ($\texttt{GNC-IRLS}_p$ and $\texttt{MS-GNC-TLS}$).

Point Cloud Registration. In the point cloud registration problem [38] that we discussed in the main paper, the variable v lies in the special Euclidean group $\mathrm{SE}(3) = \{(\boldsymbol{R}, \boldsymbol{t}) : \boldsymbol{R} \in \mathrm{SO}(3), \boldsymbol{t} \in \mathbb{R}^3\}$, each sample d_i is a point pair $(\boldsymbol{y}_i, \boldsymbol{x}_i) \in \mathbb{R}^3 \times \mathbb{R}^3$, and the residual function $r(v, d_i)$ is

$$r(v, d_i) = \|\boldsymbol{y}_i - \boldsymbol{R}\boldsymbol{x}_i - \boldsymbol{t}\|_2.$$

One can easily check that Assumptions 1 and 2 are satisfied.

In the more general and more challenging problem of *category-level registration*, point $\boldsymbol{x}_i$ is not given directly. We are instead given $\boldsymbol{b}_{i1}, \ldots, \boldsymbol{b}_{is} \in \mathbb{R}^3$ such that $\boldsymbol{x}_i$ can be represented as a convex combination of $\boldsymbol{b}_{ij}$'s, i.e., $\boldsymbol{x}_i = \sum_{j=1}^s c_j \boldsymbol{b}_{ij}$ for some unknown non-negative coefficients c_j's with $\sum_{j=1}^s c_j = 1$. This setting is under the assumption that the 3D point cloud object $\{\boldsymbol{x}_i\}_{i=1}^m$ (*e.g.*, a car) can be represented as a convex combination of s objects $\{\{\boldsymbol{b}_{ij}\}_{i=1}^m\}_{j=1}^s$ of the same category (*e.g.*, s different cars). As such, our variable consists of a 3D rotation $\boldsymbol{R}$, translation $\boldsymbol{t}$, and coefficients c_j's. Therefore, the constraint set $\mathcal{C}$ becomes $\mathrm{SE}(3) \times \{(c_1, \ldots, c_s) \in \mathbb{R}^s : c_j \geq 0, \forall j, c_1 + \cdots + c_s = 1\}$, and the residual function $r(v, d_i)$ is

$$r(v, d_i) = \big\|\boldsymbol{y}_i - \boldsymbol{R}\big(\sum\nolimits_{j=1}^s c_j \boldsymbol{b}_{ij}\big) - \boldsymbol{t}\big\|_2.$$

Both for this problem and a similar one where $\boldsymbol{R}(\sum_{j=1}^s c_j \boldsymbol{b}_{ij}) + \boldsymbol{t}$ of the residual function $r(v, d_i)$ is replaced by its 2D projection (*cf*. [64, 75, 79, 88]), Assumptions 1 and 2 are satisfied.

Absolute Pose Estimation. The *absolute pose estimation* problem has a few different formulations [37, 40, 63, 85]. In one of the formulations, the variable v lies in $\mathrm{SE}(3)$ and consists of a 3D rotation $\boldsymbol{R} \in \mathrm{SO}(3)$ and translation $\boldsymbol{t} \in \mathbb{R}^3$, each data sample $d_i = (\boldsymbol{u}_i, \boldsymbol{x}_i) \in \mathbb{S}^2 \times \mathbb{R}^3$ consists of a unit vector $\boldsymbol{u}_i$ and a 3D point $\boldsymbol{x}_i$, and the residual function is

$$r(v, d_i) = \|[\boldsymbol{u}_i]_\times (\boldsymbol{R}\boldsymbol{x}_i + \boldsymbol{t})\|_2,$$

where $[\cdot]_\times$ denotes the cross product matrix of some 3D vector. One verifies that Assumptions 1 and 2 are satisfied.

Essential Matrix Estimation. In the problem of *essential matrix estimation*, our variable v lies in $\overline{\mathrm{SE}}(3) := \{(\boldsymbol{R}, \boldsymbol{t}/\|\boldsymbol{t}\|_2) : (\boldsymbol{R}, \boldsymbol{t}) \in \mathrm{SE}(3)\}$. Therefore, the translation variable is assumed to be a unit vector. The data sample d_i is a 2D point pair $(\mathbf{y}_i, \mathbf{x}_i) \in \mathbb{R}^2 \times \mathbb{R}^2$ representing pixel locations in two different images respectively. The residual function in this case is

$$r(v, d_i) = \left|[\mathbf{y}_i^\top \; 1]([\boldsymbol{t}]_\times \boldsymbol{R}) \begin{bmatrix}\mathbf{x}_i \\ 1\end{bmatrix}\right|.$$

The matrix $[\boldsymbol{t}]_\times \boldsymbol{R}$ is called an essential matrix–thus the name of the problem. One easily verifies that Assumptions 1 and 2 are satisfied. And one further verifies the assumptions are true for related geometric vision problems (*e.g.*, *fundamental matrix estimation*, *homography estimation*, *trifocal tensor estimation*). See, *e.g.*, [36] for a summary of these problems.

Hand-Eye Calibration. In this *hand-eye calibration* problem (*cf*. [16, Example 3]), one aims to find the relative transformation $v = \boldsymbol{R} \in \mathrm{SO}(3)$ between a robot hand and a camera. Each data sample $\mathcal{d}_i$ consists of a pair of rotations $(\boldsymbol{U}_i, \boldsymbol{V}_i) \in \mathrm{SO}(3) \times \mathrm{SO}(3)$, each of which represents the location of the robot hand or the camera respectively in some global coordinate system. The residual function for this problem is

$$r(v, \mathcal{d}_i) = \|\boldsymbol{U}_i\boldsymbol{R} - \boldsymbol{R}\boldsymbol{V}_i\|_{\mathrm{F}}.$$

One can then verify that Assumptions 1 and 2 are satisfied.

Rotation Averaging. In this problem [35], our variable v consists of s 3D rotations $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_s$. And we are given $s(s+1)/2$ samples $\{\tilde{\boldsymbol{R}}_{ij}\}_{1\le i<j\le s} \subset \mathrm{SO}(3)$, and each sample $\mathcal{d}_{ij} = \tilde{\boldsymbol{R}}_{ij}$ is assumed to satisfy $\tilde{\boldsymbol{R}}_{ij} \approx \boldsymbol{R}_i^\top \boldsymbol{R}_j$ if it is an inlier. Therefore, a residual function in this problem is

$$r(v, \mathcal{d}_{ij}) = \|\boldsymbol{R}_i\tilde{\boldsymbol{R}}_{ij} - \boldsymbol{R}_j\|_{\mathrm{F}}.$$

One verifies Assumptions 1 and 2 are satisfied. One further verifies that the assumptions also hold for a more general problem called $\mathrm{SE}(3)$ synchronization [62, Eq. (12)].

Affine Hyperplane Recovery. In this problem [26, Section 5], each data sample $\mathcal{d}$ is an n-dimensional point $\boldsymbol{x} \in \mathbb{R}^n$, the variable $v = (\boldsymbol{u}, t) \in \mathbb{S}^{n-1} \times \mathbb{R}$ consists of some vector $\boldsymbol{u} \in \mathbb{S}^{n-1}$ and a scale translation $t \in \mathbb{R}$. The variable $(\boldsymbol{u}, t)$ defines an affine hyperplane in $\mathbb{R}^n$, and in the 3D case, it could represent some road in a real point cloud. The residual function is

$$r(v; \mathcal{d}) = |\boldsymbol{x}_i^\top \boldsymbol{u} - t|.$$

It is easy to verify that Assumptions 1 and 2 are satisfied.

C. Proof of Theorem 1

Proof of Lemma 1. To show statement (i) of Lemma 1, recall the definition of (7) and we note that the $\rho_\epsilon(\cdot)$ is differentiable for all $|\rho| \neq \epsilon$ and its derivative satisfies

$$\rho_\epsilon'(r) = \begin{cases} \frac{|r|^p}{r}, & \text{if } |r| > \epsilon, \\ \frac{r}{\epsilon^{2-p}}, & \text{if } |r| \le \epsilon. \end{cases}$$

To check continuous differentiability of $\rho_\epsilon(\cdot)$ at $r = \epsilon$, we compute

$$\lim_{h>0, h\to 0} \frac{\rho_\epsilon(\epsilon+h) - \rho_\epsilon(\epsilon)}{h} = \lim_{h>0, h\to 0} \frac{\frac{1}{p}(\epsilon+h)^p - \frac{1}{p}\epsilon^p}{h} = \epsilon^{p-1}$$

and

$$\lim_{h<0, h\to 0} \frac{\rho_\epsilon(\epsilon+h) - \rho_\epsilon(\epsilon)}{h} = \lim_{h>0, h\to 0} \frac{\frac{1}{2}\frac{(\epsilon+h)^2}{\epsilon^{2-p}} + (\frac{1}{p} - \frac{1}{2})\epsilon^p - \frac{1}{p}\epsilon^p}{h} = \epsilon^{p-2} \lim_{h>0, h\to 0} \frac{\frac{1}{2}(\epsilon+h)^2 - \frac{1}{2}\epsilon^2}{h} = \epsilon^{p-1},$$

which shows that

$$\rho_\epsilon'(\epsilon) = \epsilon^{p-1} = \lim_{r\to\epsilon} \rho_\epsilon'(r), \tag{20}$$

and analogously, it can be shown that ρ_ϵ' exists and is continuous at $r = -\epsilon$, which shows (i).

For (ii), due to $0 < p \le 1$, we know that $z \to z^{p/2}$ is concave, and therefore

$$\rho(r) = \frac{1}{p}|r|^p \le \frac{1}{p}(r^2)^{p/2} \le \frac{1}{p}\left[(\epsilon^2)^{\frac{p}{2}} + \frac{p}{2}(\epsilon^2)^{\frac{p}{2}-1}(r^2 - \epsilon^2)\right] = \frac{1}{2}\frac{r^2}{\epsilon^{2-p}} + \left(\frac{1}{p} - \frac{1}{2}\right)\epsilon^p = \rho_\epsilon(r)$$

for $|r| \le \epsilon$, and $\rho(r) = \rho_\epsilon(r)$ for $|r| > \epsilon$, which shows (ii).

For (iii), we can define the function $g_r(\epsilon) := \rho_\epsilon(r)$, which is differentiable for $\epsilon \geq |r|$: Computing its derivative, we obtain

$$g_r'(\epsilon) = \left(\frac{1}{p} - \frac{1}{2}\right) p\epsilon^{p-1} + (p-2)\frac{1}{2}\frac{r^2}{\epsilon^{3-p}} = \epsilon^{p-1}\left[\left(1 - \frac{r^2}{\epsilon^2}\right) + \frac{p}{2}\frac{r^2}{\epsilon^2}\right],$$

which is ≥ 0 for all $\epsilon \geq |r|$, so we conclude that $g_r(\cdot)$ is monotonously increasing in this interval and $\rho_{\epsilon'}(r) \leq \rho_\epsilon(r)$ if $\epsilon' \leq \epsilon$ for all $\epsilon \geq |r|$, and $\rho_{\epsilon'}(r) = \rho_\epsilon(r)$ if $|r| \leq \epsilon' \leq \epsilon$. If $\epsilon' < |r| \leq \epsilon$, $\rho_{\epsilon'}(r) = \rho(r) \leq \rho_\epsilon(r)$, which finishes the proof of (iii).

(iv) finally follows as for each $r \neq 0$, $\rho_\epsilon(r)$ coincides with $\rho(r)$ for ϵ small enough. □

Proof of Lemma 2. Lemma 2 follows from the fact that $\rho_\epsilon(\sqrt{r})$ is concave in r, and therefore, it is majorized by its tangent line at any given point $u \in \mathbb{R}$. More formally, it follows from (34) below with $\theta = u$ and the fact that $\frac{\rho_\epsilon'(u)}{|u|} = \frac{1}{\max(|u|,\epsilon)^{2-p}}$ for the smoothed ℓ_p-loss $\rho_\epsilon(\cdot)$.

□

Proof of Theorem 1. Recall the definition $r_i^{(t)} := r(v^{(t)}, a_i)$, $\forall t$. We have

$$\begin{aligned}
\sum_{i=1}^m \rho_{\epsilon^{(t+1)}}\left(r_i^{(t+1)}\right) &\overset{\text{(i)}}{\leq} \sum_{i=1}^m \rho_{\epsilon^{(t)}}\left(r_i^{(t+1)}\right) \\
&\overset{\text{(ii)}}{\leq} \sum_{i=1}^m q_{\epsilon^{(t)}}\left(r_i^{(t+1)}, r_i^{(t)}\right) \\
&\overset{\text{(iii)}}{\leq} \sum_{i=1}^m q_{\epsilon^{(t)}}\left(r_i^{(t)}, r_i^{(t)}\right) \\
&= \sum_{i=1}^m \rho_{\epsilon^{(t)}}\left(r_i^{(t)}\right).
\end{aligned} \tag{21}$$

The three inequalities in (21) are due to (i) Lemma 1, (ii) Lemma 2, and (iii) Remark 3, respectively. Hence $\{\sum_{i=1}^m \rho_{\epsilon^{(t)}}(r_i^{(t)})\}_t$ is a non-increasing sequence of non-negative numbers, thus convergent. We now show that $\{v^{(t)}\}_t$ is bounded. This is true if $\mathcal{C}$ is bounded, so we assume that $\mathcal{C}$ is unbounded. Suppose for the sake of contradiction that $\|v^{(t)}\|_2 \to \infty$ as $t \to \infty$. By Assumption 1, for every $i = 1, \ldots, m$ and as $t \to \infty$, we have $r_i^{(t)} = r(v^{(t)}, a_i) \to \infty$ and thus $\rho_{\epsilon^{(t)}}(r_i^{(t)}) \to \infty$. Therefore

$$\sum_{i=1}^m \rho_{\epsilon^{(0)}}(r_i^{(0)}) \geq \lim_{t\to\infty} \sum_{i=1}^m \rho_{\epsilon^{(t)}}(r_i^{(t)}) = \infty. \tag{22}$$

But Algorithm 1 ensures $\|v^{(0)}\|_2 < \infty$, thus $r_i^{(0)}$ is finite (by Assumption 1), and so is $\rho_{\epsilon^{(0)}}(r_i^{(0)})$, contradicting (22). We have thus proved that $\{v^{(t)}\}_t$ is a bounded sequence. It now follows from the famous *Bolzano–Weierstrass* theorem that this bounded sequence $\{v^{(t)}\}_t$ has a convergent subsequence. Let $\{v^{(t_j)}\}_j$ be such subsequence that converges to some accumulation point, say $\hat{v}$. We will next show that $\hat{v}$ is a stationary point of (10). In fact, we have

$$\begin{aligned}
\sum_{i=1}^m q_{\epsilon^{(t_{j+1})}}\left(r_i^{(t_{j+1})}, r_i^{(t_{j+1})}\right) &= \sum_{i=1}^m \rho_{\epsilon^{(t_{j+1})}}\left(r_i^{(t_{j+1})}\right) \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^m \rho_{\epsilon^{(t_j+1)}}\left(r_i^{(t_j+1)}\right) \\
&\overset{\text{(ii)}}{\leq} \sum_{i=1}^m \rho_{\epsilon^{(t_j)}}\left(r_i^{(t_j+1)}\right) \\
&\overset{\text{(iii)}}{\leq} \sum_{i=1}^m q_{\epsilon^{(t_j)}}\left(r_i^{(t_j+1)}, r_i^{(t_j)}\right) \\
&\overset{\text{(iv)}}{\leq} \sum_{i=1}^m q_{\epsilon^{(t_j)}}\left(r(v, a_i), r_i^{(t_j)}\right), \quad \forall v \in \mathcal{C}.
\end{aligned} \tag{23}$$

The above inequalities are due respectively to (i) $t_{j+1} \geq t_j + 1$ and (21), (ii) Lemma 1, (iii) Lemma 2, (iv) Remark 3. Note that $\rho_{\epsilon^{(t)}}$, $q_{\epsilon^{(t)}}$ are continuous for any t, and $r(v, d_i)$ is continuous in v by Assumption 2; with $t_j \to \infty$ in (23) we get

$$\sum_{i=1}^{m} q_\epsilon\Big(r(\hat{v}, d_i), r(\hat{v}, d_i)\Big) \leq \sum_{i=1}^{m} q_\epsilon\Big(r(v, d_i), r(\hat{v}, d_i)\Big), \quad \forall v \in \mathcal{C}.$$

With $\hat{w}_i := \max\{r(\hat{v}, d_i), \epsilon\}^{2-p}$ and Remark 3, the above implies

$$\hat{v} \in \operatorname*{argmin}_{v \in \mathcal{C}} \sum_{i=1}^{m} q_\epsilon\Big(r(v, d_i), r(\hat{v}, d_i)\Big) = \operatorname*{argmin}_{v \in \mathcal{C}} \sum_{i=1}^{m} \hat{w}_i \cdot r(v, d_i)^2.$$

Thus, $\hat{v}$ must satisfy the geometric optimality condition [3, Section 5.3]

$$\Big(\sum_{i=1}^{m} \hat{w}_i \cdot \nabla\big(r(\hat{v}, d_i)^2\big)\Big)^\top b \geq 0, \quad \forall b \in T_{\hat{v}}\mathcal{C}, \tag{24}$$

where $T_{\hat{v}}\mathcal{C}$ denotes the tangent cone of $\mathcal{C}$ at $\hat{v}$.

We now verify $\nabla\rho_\epsilon\big(r(\hat{v}, d_i)\big) = \frac{1}{2}\hat{w}_i \cdot \nabla\big(r(\hat{v}, d_i)^2\big)$. Indeed, if $r(\hat{v}, d_i) = 0$, then by the definition of ρ_ϵ (7) we have

$$\nabla\rho_\epsilon\big(r(\hat{v}, d_i)\big) = \nabla\Big(\frac{1}{2}\frac{r(\hat{v}, d_i)^2}{\epsilon^{2-p}} + (\frac{1}{p} - \frac{1}{2})\epsilon^p\Big) = \frac{1}{2}\nabla\Big(\frac{r(\hat{v}, d_i)^2}{\epsilon^{2-p}}\Big) = \frac{1}{2}\hat{w}_i \cdot \nabla\big(r(\hat{v}, d_i)^2\big).$$

On the other hand, $r(\hat{v}, d_i) \neq 0$, then $r(v, d_i)$ is differentiable at $v = \hat{v}$ by Assumption 2. Therefore, applying the chain rule and the equality $\rho'_\epsilon\big(r(\hat{v}, d_i)\big) = \hat{w}_i \cdot r(\hat{v}, d_i)$, we get

$$\nabla\rho_\epsilon\big(r(\hat{v}, d_i)\big) = \rho'_\epsilon\big(r(\hat{v}, d_i)\big) \cdot \nabla r(\hat{v}, d_i) = \hat{w}_i \cdot r(\hat{v}, d_i) \cdot \nabla r(\hat{v}, d_i) = \frac{1}{2}\hat{w}_i \cdot \nabla\big(r(\hat{v}, d_i)^2\big).$$

Substitute this into (24), and we obtain

$$\Big(\sum_{i=1}^{m} \nabla\rho_\epsilon\big(r(\hat{v}, d_i)\big)\Big)^\top b \geq 0, \quad \forall b \in T_{\hat{v}}\mathcal{C},$$

which means $\hat{v}$ is a stationary point of (10). The proof is complete. □

D. Majorizing Truncated Least-Squares and Proof of Theorem 3

In this section, we consider the TLS loss with some hyper-parameter c^2 ($c > 0$),

$$\rho(r) = \min\{r^2, c^2\}. \tag{25}$$

As mentioned in the main paper, here we present a smooth majorizer and a quadratic majorizer for the TLS loss (25), in Section D.1 and Section D.2 respectively. After that, we provide a proof of Theorem 3. The development of this section follows much from Section 2, but this time we discuss how the majorizers are designed in more detail.

D.1. A Smooth Majorizer for The TLS Loss

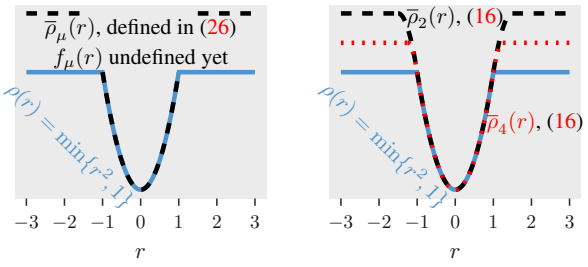
In designing a majorizer $\overline{\rho}_\mu$ for the TLS loss ρ (25), we expect the following properties that it should yield:

(S1) $\overline{\rho}_\mu$ is continuously differentiable.

(S2) $\overline{\rho}(r) \leq \overline{\rho}_\mu(r)$ for every $r \in \mathbb{R}$.

(S3) $\mu' \leq \mu \Rightarrow \overline{\rho}_\mu(r) \leq \overline{\rho}_{\mu'}(r)$.

(S4) $\overline{\rho}_\mu \to \rho$ as $\mu \to \infty$.



(a) Smooth Majorizer (26) (b) Proposition 2

Figure 7. The TLS loss $\rho(r)$ (25) and its smooth majorizer $\overline{\rho}_\mu(r)$ (*cf*. (26), (28), and (16)).

The above four properties correspond to those of Lemma 1. (S1) ensures the weighting strategy (5) makes sense. (S2) ensures majorization. (S3) is in different direction than that of Lemma 1; this is because we are increasing μ to approach ρ, whereas ϵ is decreased in Lemma 1 to achieve similar effects. (S4) ensures the approximation quality, and as a side effect, it guarantees binary weights as $\mu \to \infty$. Such majorizers do exist, as we will show next.

We choose majorizers based on the following idea. First we make sure $\overline{\rho}_\mu(r) = r^2$ whenever $r^2 \in [0, c^2]$. Then we set $\overline{\rho}_\mu(r) = \frac{(\mu+1)}{\mu}c^2$ if $r^2 \geq \nu_\mu^2 c^2$, where ν_μ is some number to be determined and we require $\nu_\mu > 1$ and $\nu_\mu \to 1$ as $\mu \to \infty$. Specifically, we parameterize $\overline{\rho}_\mu(r)$ as

$$\overline{\rho}_\mu(r) = \begin{cases} r^2, & \text{if } r^2 \in \big[0, c^2\big], \\ f_\mu(|r|), & \text{if } r^2 \in \big(c^2, \nu_\mu^2 c^2\big), \\ \frac{\mu+1}{\mu}c^2, & \text{if } r^2 \in \big[\nu_\mu^2 c^2, \infty\big), \end{cases} \tag{26}$$

where the function $f : \mathbb{R} \to \mathbb{R}$ is to be determined. This is illustrated in Figure 7a (dashed, black). This construction ensures (S4). It remains to find f_μ and ν_μ to fulfill properties (S1)-(S3). In view of the requirement of continuous differentiability (S1), f_μ itself must be continuously differentiable, which means that it should satisfy the following constraints with ν_μ:

$$\begin{aligned} f_\mu(c) &= c^2 \\ f'_\mu(c) &= 2c \\ f'_\mu(\nu_\mu c) &= 0 \end{aligned} \tag{27}$$

Interestingly, if f_μ belongs to some specific class of functions (*e.g*., quadratic functions), then it is determined as the *unique* solution of (27). And if so, this construction (26) automatically satisfies the remaining properties, namely (S2) and (S3), which makes ρ_μ (26) into a smooth majorizer, as desired. More formally, we have the following proposition:

Proposition 2. *Let ρ be the truncated least-squares loss (25). Let ρ_μ be defined in (26), where $f_\mu : \mathbb{R} \to \mathbb{R}$ is a quadratic function and $\nu_\mu > 1$ with $\nu_\mu \to 1$ as $\mu \to \infty$. If $\overline{\rho}_\mu$ satisfies properties (S1)-(S4), then it is uniquely determined by the following expressions:*

$$\overline{\rho}_\mu(r) = \begin{cases} r^2, & \text{if } r^2 \in \big[0, c^2\big], \\ -\mu r^2 + 2(1+\mu)c|r| - (1+\mu)c^2, & \text{if } r^2 \in \big(c^2, \frac{(\mu+1)^2}{\mu^2}c^2\big), \\ \frac{\mu+1}{\mu}c^2, & \text{if } r^2 \in \big[\frac{(\mu+1)^2}{\mu^2}c^2, \infty\big). \end{cases} \tag{28}$$

Pictorial Proof of Proposition 2. We need to "attach" a quadratic function f_μ into Figure 7a so that it "connects" the black dashed curves of Figure 7a and makes the overall curve continuously differentiable. We attach such $f_\mu(|r|) = -\mu r^2 + 2(1+\mu)c|r| - (1+\mu)c^2$ (28) in Figure 7b with different values of μ, where, visually, all properties (S1)-(S4) are fulfilled. □

Proof of Proposition 2. Note that $\overline{\rho}_\mu$ is symmetric with $\overline{\rho}_\mu(r) = \overline{\rho}_\mu(-r)$, so we assume $r \geq 0$ without loss of generality. Since f_μ is a quadratic function, we write $f_\mu(r) = ar^2 + br + d$, where a, b and d are some real-valued coefficients to be

determined. These coefficients depend on μ, but we drop the indices for simplicity. Note that $\nu_\mu c \neq 0$ and f is quadratic, for $f'_\mu(\nu_\mu c) = 0$ of (27) to hold, we must have $\nu_\mu c = -\frac{b}{2a}$ and $f(-\frac{b}{2a}) = \frac{\mu+1}{\mu}c^2$. Hence, (27) gives the following equations:

$$\begin{aligned} ac^2 + bc + d &= c^2 \\ 2ac + b &= 2c \\ -\frac{b}{2a} &= \nu_\mu c \\ -\frac{b^2}{4a} + d &= \frac{\mu+1}{\mu}c^2 \end{aligned}$$

Note that μ and c are given, the above are four equations with four unknowns, which admit a unique solution:

$$\begin{aligned} a &= -\mu \\ b &= 2c(1+\mu) \\ d &= -(1+\mu)c^2 \\ \nu_\mu &= \frac{1+\mu}{\mu} \end{aligned}$$

This gives a unique expression of $\overline{\rho}_\mu$ (28). By construction, $\overline{\rho}_\mu$ is continuously differentiable (S1) and approximates the TLS loss ρ as $\mu \to \infty$ (S4). It is a simple routine to verify that $\overline{\rho}_\mu$ also satisfies (S2) and (S3). The proof is complete. □

Remark 4 (Other classes of functions). The majorizer of the TLS loss exists also for other classes of functions. For example, if $f_\mu : \mathbb{R} \to \mathbb{R}$ is a trigonometric function parameterized as $f_\mu(r) = a\sin(b(r-c)) + c^2$, then one can solve (27) for a and b. Pictorally, this amounts to attaching a trigonometric function to Figure 7a that ensures countinuous differentiablity. However, we did not find more benefits of doing so than what (28) gives, so we omitted this construction.

Remark 5 (Other quadratic functions f_μ). While Proposition 2 shows f_μ is a unique quadratic function, it is based on the fact that ρ_μ is parameterized as (26). Different parameterizations lead to different quadratic functions, and finding an optimal one that accelerates convergence is left to future work.

D.2. A Quadratic Majorizer for The TLS Loss

We will construct a quadratic majorizer $q_\mu(r, \cdot) : \mathbb{R} \to \mathbb{R}$ of $\overline{\rho}_\mu(r)$ (28). We expect $q_\mu(r, \cdot)$ to satisfy three properties:

(Q1) $q_\mu(r, \theta)$ is a quadratic *symmetric* function in r; by *symmetric* we mean that $q_\mu(r, \theta) = q_\mu(-r, \theta)$ for any $r, \theta \in \mathbb{R}$.

(Q2) $\overline{\rho}_\mu(r) \leq q_\mu(r, \theta)$ for every $r, \theta \in \mathbb{R}$.

(Q3) $\overline{\rho}_\mu(\theta) = q_\mu(\theta, \theta)$ for every $\theta \in \mathbb{R}$.

(Q1) and (Q2) are to make sure that $q_\mu(r, \theta)$ is indeed a quadratic majorizer of $\overline{\rho}_\mu(r)$; requiring $q_\mu(r, \theta)$ to be symmetric in r is because $\overline{\rho}_\mu(r)$ is symmetric in r. (Q3) makes sure that $q_\mu(r, \theta)$ is tight and, as such, minimizing $q_\mu(r, \theta)$ is expected to lead to minimization of $\overline{\rho}_\mu(\theta)$.

Constructing such a quadratic function to majorize $\overline{\rho}_\mu(\theta)$ at a point θ is not hard. In particular, note that $\overline{\rho}_\mu(r)$ is locally quadratic, or otherwise concave (or in particular linear) in r. Its concave part (Figure 8, red) can be majorized at any given point θ by any tangent line passing through θ (Figure 8, blue), and furthermore by a quadratic convex function (Figure 8, black); the latter is what we need. Next, we make this argument more formal.

To start with, let us assume $\theta > 0$. If $\theta \in [0, c]$ we could simply set $q_\mu(r, \theta) = r^2$, which fulfills all desired properties (Q1)-(Q3). So let us next focus on the case $\theta > c$. To satisfy (Q1), write $q(r, \theta) = ar^2 + d$ for some unknown coefficients a and d, which are to be determined and implicitly depend on θ. (Q3) gives us

$$a\theta^2 + d = \overline{\rho}_\mu(\theta),$$

and we would expect that (Q2) gives us one more equation so that we can solve them for a and d. Since $\overline{\rho}_\mu(\theta)$ is concave in (c, ∞), it is majorized by its tangent line passing through θ with derivative $\overline{\rho}'_\mu(\theta)$, and thus by a quadratic (convex) function passing through θ with derivative $\overline{\rho}'_\mu(\theta)$. Therefore, for $q(r, \theta)$ to majorize $\overline{\rho}_\mu(r)$ when $\theta > c$, it suffices to have the equality

$$\frac{\partial q(r, \theta)}{\partial r} = \overline{\rho}'_\mu(\theta) \Leftrightarrow 2\theta a = \overline{\rho}'_\mu(\theta). \tag{29}$$

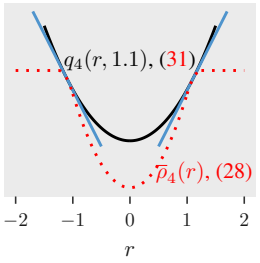


Figure 8. The quadratic majorizer (31) for the TLS loss.

Now, solve the above two equations for a and d and generalize to the case $\theta^2 \geq c^2$. Then we get the following expression of the quadratic majorizer $q(\cdot, \theta)$ for $\theta^2 \geq c^2$:

$$q_\mu(r,\theta) = \overline{\rho}_\mu(\theta) + \frac{\overline{\rho}'_\mu(\theta)}{2|\theta|}(r^2-\theta^2) = \begin{cases} \overline{\rho}_\mu(\theta) + (\frac{c(1+\mu)}{|\theta|} - \mu)\cdot(r^2-\theta^2), & \text{if } \theta^2 \in (c^2, \frac{(\mu+1)^2}{\mu^2}c^2), \\ \overline{\rho}_\mu(\theta), & \text{if } \theta^2 \geq \frac{(\mu+1)^2}{\mu^2}c^2. \end{cases} \tag{30}$$

We now summarize our discovery about $q_\mu(r,\theta)$:

Proposition 3. *Recall the definition of the smooth majorizer $\overline{\rho}_\mu(\cdot)$ (28) for the TLS loss. The quadratic function*

$$q_\mu(r,\theta) = \begin{cases} r^2, & \text{if } \theta^2 \in [0, c^2], \\ \overline{\rho}_\mu(\theta) + (\frac{c(1+\mu)}{|\theta|} - \mu)\cdot(r^2-\theta^2), & \text{if } \theta^2 \in (c^2, \frac{(\mu+1)^2}{\mu^2}c^2), \\ \overline{\rho}_\mu(\theta), & \text{if } \theta^2 \geq \frac{(\mu+1)^2}{\mu^2}c^2, \end{cases} \tag{31}$$

is a quadratic majorizer of $\overline{\rho}_\mu(r)$ that fulfills (Q1), (Q2), and (Q3).

Proof. Note that, by design, $q_\mu(r,\theta)$ (31) fulfills (Q1) and (Q3). However, it should be noted that, by design, the majorizer (30) only makes sure $q_\mu(r,\theta) \geq \overline{\rho}_\mu(r)$ for any r and θ that fulfills $r^2 > c^2$ and $\theta^2 > c^2$ (and similarly for the case $\theta^2 \leq c^2$); therefore, slightly more work is needed to verify (Q2).

If $\theta^2 \leq c^2$, we have

$$q_\mu(r,\theta) = r^2, \geq \begin{cases} r^2 & \text{if } r^2 \in \big[0, c^2\big], \\ -\mu r^2 + 2(1+\mu)c|r| - (1+\mu)c^2, & \text{if } r^2 \in \big(c^2, \frac{(\mu+1)^2}{\mu^2}c^2\big), \\ \frac{\mu+1}{\mu}c^2, & \text{if } r^2 \in \big[\frac{(\mu+1)^2}{\mu^2}c^2, \infty\big), \end{cases} = \overline{\rho}_\mu(r).$$

On the other hand, consider the case $\theta^2 \geq c^2$. We need to prove $q_\mu(r,\theta) \geq \overline{\rho}_\mu(r)$ whenever $r^2 \leq c^2$ and $\theta^2 > c^2$, in which case $\overline{\rho}_\mu(r)$ is quadratic and the inequality to prove is $q_\mu(r,\theta) \geq r^2$. This is true if $\theta^2 \geq (\mu+1)^2c^2/\mu^2$, so we consider the case $\theta^2 \in (c^2, \frac{(\mu+1)^2}{\mu^2}c^2)$. In this case, we have

$$\begin{aligned} q_\mu(r,\theta) &= \overline{\rho}_\mu(\theta) + \Big(\frac{c(1+\mu)}{|\theta|} - \mu\Big)\cdot(r^2-\theta^2) \\ &= -\mu\theta^2 + 2(1+\mu)c|\theta| - (1+\mu)c^2 + \Big(\frac{c(1+\mu)}{|\theta|} - \mu\Big)\cdot(r^2-\theta^2) \\ &= (1+\mu)c|\theta| - (1+\mu)c^2 + \Big(\frac{c(1+\mu)}{|\theta|} - \mu\Big)\cdot r^2 \\ &= (1+\mu)\cdot\Big(c|\theta| - c^2 + \frac{c}{|\theta|}r^2 - r^2\Big) + r^2 \\ &= (1+\mu)\cdot(|\theta| - c)\cdot\Big(c - \frac{r^2}{|\theta|}\Big) + \overline{\rho}_\mu(r) \\ &\geq \overline{\rho}_\mu(r), \end{aligned}$$

by which the proof is complete. □

Remark 6. Similar statements in Remarks 1 and 3 can also be made for MS−GNC−TLS and its smooth and quadratic majorizers. In particular, first note that the coefficient of the quadratic term of $q_\mu(r,\theta)$ is exactly $\overline{\rho}_\mu(\theta)/(2|\theta|)$, which coincides with the weight update rule (17) of MS−GNC−TLS if θ is the residual $r_i^{(t)}$ at the t-th iteration. Therefore, the weighted least-squares problem that MS−GNC−TLS solves at each iteration t is minimizing the quadratic majorizer

$$\begin{aligned} v^{(t+1)} &\in \operatorname*{argmin}_{v\in\mathcal{C}} \sum\nolimits_{i=1}^{m} w_i^{(t+1)} r(v,\mathcal{d}_i)^2 \\ &= \operatorname*{argmin}_{v\in\mathcal{C}} \sum\nolimits_{i=1}^{m} \frac{\overline{\rho}_\mu(r_i^{(t)})}{2r_i^{(t)}} r(v,\mathcal{d}_i)^2 \\ &= \operatorname*{argmin}_{v\in\mathcal{C}} \sum\nolimits_{i=1}^{m} q_{\mu^{(t)}}(r(v,\mathcal{d}_i), r_i^{(t)}). \end{aligned}$$

D.3. Proof of Theorem 3

The proof follows very similarly from that of Theorem 1, therefore we provide a high-level proof here and refer to the proof of Theorem 1 for omitted steps. Recall $r_i^{(t)} := r(v^{(t)},\mathcal{d}_i), \forall t$. Similarly to (21), based on majorization properties of smooth and quadratic majorizers, we can show that the objective, despite keeping changing, has non-increasing values:

$$\sum_{i=1}^{m} \overline{\rho}_{\mu^{(t+1)}}\left(r_i^{(t+1)}\right) \le \sum_{i=1}^{m} \overline{\rho}_{\mu^{(t)}}\left(r_i^{(t)}\right)$$

Since $\{v^{(t)}\}_t$ is a bounded sequence, it has a convergent subsequence, say, $v_j^{(t_j)}$, which converges to $\hat{v}$. It remains to show that $\hat{v}$ is a stationary point of (18). Towards this end, one can follow the reasoning of (23) and show that

$$\sum_{i=1}^{m} q_{\mu^{(t_{j+1})}}\left(r_i^{(t_{j+1})}, r_i^{(t_{j+1})}\right) \le \sum_{i=1}^{m} q_{\mu^{(t_j)}}\left(r(v,\mathcal{d}_i), r_i^{(t_j)}\right), \forall v\in\mathcal{C}.$$

As $t_j\to\infty$, the above equality becomes

$$\begin{aligned} &\sum_{i=1}^{m} q_\mu\big(r(\hat{v},\mathcal{d}_i), r(\hat{v},\mathcal{d}_i)\big) \le \sum_{i=1}^{m} q_\mu\big(r(v,\mathcal{d}_i), r(\hat{v},\mathcal{d}_i)\big), \forall v\in\mathcal{C} \\ \Leftrightarrow\ &\hat{v} \in \operatorname*{argmin}_{v\in\mathcal{C}} \sum_{i=1}^{m} q_\mu\big(r(v,\mathcal{d}_i), r(\hat{v},\mathcal{d}_i)\big) = \operatorname*{argmin}_{v\in\mathcal{C}} \sum_{i=1}^{m} \hat{w}_i \cdot r(v,\mathcal{d}_i)^2, \end{aligned}$$

where $\hat{w}_i$ is defined to be the weight (17) as t_j goes to infinity, that is

$$\hat{w}_i = \begin{cases} 1, & \text{if } \ r(\hat{v},\mathcal{d}_i) \le c, \\ 0, & \text{if } \ r(\hat{v},\mathcal{d}_i) \ge \frac{\mu+1}{\mu}c, \\ \frac{c(1+\mu)}{r(\hat{v},\mathcal{d}_i)} - \mu, & \text{if } \ r(\hat{v},\mathcal{d}_i) \in (c, \frac{\mu+1}{\mu}c). \end{cases}$$

Therefore, $\hat{v}$ must satisfy the geometric condition (*cf*. (24))

$$\left(\sum_{i=1}^{m} \hat{w}_i \cdot \nabla\big(r(\hat{v},\mathcal{d}_i)^2\big)\right)^\top \mathcal{b} \ge 0, \quad \forall \mathcal{b}\in T_{\hat{v}}\mathcal{C},$$

which is equivalent to (see the proof of Theorem 1 for a similar derivation)

$$\left(\sum_{i=1}^{m} \nabla\overline{\rho}_\mu\big(r(\hat{v},\mathcal{d}_i)\big)\right)^\top \mathcal{b} \ge 0, \quad \forall \mathcal{b}\in T_{\hat{v}}\mathcal{C},$$

meaning that $\hat{v}$ is a stationary point of (18).

E. Convergence Theory of IRLS +GNC for A General Class of Losses

In this section we develop a general convergence theory for IRLS and GNC applied to the outlier-robust estimation problem (2). We repeat the optimization problem here for convenience:

$$\min_{v\in\mathcal{C}} \sum_{i=1}^{m} \rho_\zeta\big(r(v, d_i)\big) \tag{32}$$

This time we added a subscript $\zeta \in \mathbb{R}\cup\{\infty\}$ to the loss ρ_ζ for notational convenience. Since we are now working with a general loss function ρ_ζ, we need some assumptions on it for deriving meaningful convergence results. Our previous experience suggests that ρ_ζ should admit some smooth majorizer and quadratic majorizer. We start with assuming the first:

Assumption 3 (Existence of A Smooth Majorizer)**.** There exists a function $\rho_\epsilon : \mathbb{R}\to\mathbb{R}$ parametrized by $\epsilon > 0$, which satisfies the following properties: (i) ρ_ϵ is continuously differentiable, (ii) $\rho_\epsilon \geq \rho_\zeta$ for every $\epsilon > 0$, (iii) $\rho_\epsilon(r) \geq \rho_{\epsilon'}(r)$ as long as $\epsilon \geq \epsilon'$, (iv) $\rho_\epsilon(r) \to \rho_\zeta(r)$ whenever $\epsilon \to \zeta$, (v) ρ_ϵ is symmetric, *i.e.*, $\rho_\epsilon(r) = \rho_\epsilon(-r)$.

A simple case is when ρ_ζ is already continuously differentiable and symmetric, which makes itself a smooth majorizer. We knew from Lemma 1 that the ℓ_p-loss admits a smooth majorizer, and in this case we have $\zeta = 0$. We knew from Section D.1 that the TLS loss admits a smooth majorizer (if one of the inequalities in the property (iii) is reversed), and in this case we have $\zeta = \infty$. Indeed, smooth majorizers of ρ_ζ are not hard to find. A simple way to do so is to plot ρ_ζ and then specify a "higher" continuously differentiable curve (recall Figures 1a and 5b, Section D.1).

From a loss function ρ_ζ and its smooth majorizer ρ_ϵ, a quadratic majorizer $q_\epsilon(r,\theta)$ ensues if $\rho_\epsilon(\sqrt{r})$ is concave and differentiable in r over $[0,\infty)$. In particular, the concave function $\rho_\epsilon(\sqrt{r})$ ($r \geq 0$) is majorized by its tangent line at a given point $\theta \geq 0$, meaning that

$$\rho_\epsilon(\sqrt{r}) \leq \big(\rho_\epsilon(\sqrt{\theta})\big)' \cdot (r-\theta) + \rho_\epsilon(\sqrt{\theta}) = \begin{cases} \frac{\rho_\epsilon'(\sqrt{\theta})}{2\sqrt{\theta}} \cdot (r-\theta) + \rho_\epsilon(\sqrt{\theta}), & \forall r \geq 0, \forall \theta > 0, \\ \big(\rho_\epsilon(\sqrt{\theta})\big)' \cdot r + \rho_\epsilon(0), & \forall r \geq 0, \theta = 0. \end{cases} \tag{33}$$

Make a substitution of variables (say $\sqrt{r} \leftarrow |r|$ and $\sqrt{\theta} \leftarrow |\theta|$) in (33), and we arrive at

$$\rho_\epsilon(r) \leq \frac{\rho_\epsilon'(\theta)}{2\theta} \cdot (r^2 - \theta^2) + \rho_\epsilon(\theta), \quad \forall r \geq 0, \forall \theta \neq 0,$$

and $\rho_\epsilon(r) \leq \big(\rho_\epsilon(\sqrt{\theta})\big)' \cdot r^2 + \rho_\epsilon(0)$ if $\theta = 0$. In other words, we now have the quadratic majorizer

$$q_\epsilon(r,\theta) = \begin{cases} \frac{\rho_\epsilon'(\theta)}{2|\theta|} \cdot (r^2 - \theta^2) + \rho_\epsilon(\theta), & \text{if} \quad \theta \neq 0, \\ \big(\rho_\epsilon(\sqrt{\theta})\big)' \cdot r^2 + \rho_\epsilon(0), & \text{if} \quad \theta = 0. \end{cases} \tag{34}$$

One verifies that the quadratic majorizers for the ℓ_p and TLS losses are special cases of (34). Note that in the above development, we have distinguished the case $\theta \neq 0$ and $\theta = 0$. This is because $\sqrt{\theta}$ is not differentiable at $\theta = 0$, and the chain rule cannot be applied directly. That said, for the case $\theta = 0$, the derivative $\big(\rho_\epsilon(\sqrt{\theta})\big)'$ can be easily evaluated once the exact form of ρ_ϵ is given. Finally, since $\rho_\epsilon(\sqrt{\theta})$ is differentiable in θ and $\rho_\epsilon(\cdot)$ is continuously differentiable, we can write

$$\begin{aligned} \big(\rho_\epsilon(\sqrt{\theta})\big)'\big|_{\theta=0} &= \lim_{\theta=0} \big(\rho_\epsilon(\sqrt{\theta})\big)' \\ &= \lim_{\theta=0} \frac{\rho_\epsilon'(\sqrt{\theta})}{2\sqrt{\theta}} \\ &= \lim_{\theta=0} \frac{\rho_\epsilon'(\theta)}{2\theta}. \end{aligned} \tag{35}$$

This equality will be useful in the sequel.

We have therefore justified the following assumption:

Assumption 4 (Existence of A Quadratic Majorizer)**.** For a loss function ρ_ζ, assume its smooth majorizer ρ_ϵ composed with the square root function is concave and differentiable on $[0,\infty)$, that is $\rho_\epsilon(\sqrt{r})$ is concave and differentiable in r over $[0,\infty)$.

Example 1. Here are several examples of continuously differentiable losses $\rho(\cdot)$ for which $\rho_\epsilon(\sqrt{r})$ is differentiable and concave in r. See [36, Appendix 6.8], [1], and [20, Table 1] for more examples.

- The smooth majorizers for the ℓ_p-loss and TLS loss, as we have discussed.
- The Geman-McClure loss [31], which is defined as

$$\rho_\epsilon(r) = \frac{\epsilon c^2 r^2}{\epsilon c^2 + r^2}.$$

- A surrogate function of the TLS loss used in [45, 89]:

$$\rho_\epsilon(r) = \begin{cases} r^2(2 - \frac{r^2}{c^2}), & \text{if} \quad r^2 \leq c^2, \\ c^2, & \text{if} \quad r^2 > c^2. \end{cases}$$

  Note that this loss function does not depend on ϵ. It is a smooth majorizer of itself.

- The Blake-Zisserman loss (as [36, Appendix 6.8] called):

$$\rho_\epsilon(r) = -\log(\exp(-r^2) + c)$$

With the smooth and quadratic majorizers of ρ_ζ, we can now use the IRLS framework with the weight update rule

$$w_i^{(t+1)} \leftarrow \begin{cases} \frac{\rho'_{\epsilon^{(t)}}\left(r_i^{(t)}\right)}{2r_i^{(t)}}, & \text{if} \quad r_i^{(t)} \neq 0, \\ \left(\rho_{\epsilon^{(t)}}\left(\sqrt{r_i^{(t)}}\right)\right)', & \text{if} \quad r_i^{(t)} = 0, \end{cases} \tag{36}$$

and some GNC schedule $\{\epsilon^{(t)}\}_t$ to optimize ρ_ζ. Compared to rule (5), our rule (36) explicitly takes the case $r_i^{(t)} = 0$ into account and makes it mathematically correct (computationally, this makes no difference if all the differentiability assumptions are satisfied). We are now ready to state the main result of this section:

Theorem 4. *Let $\{v^{(t)}\}_t$ be the iterates generated by IRLS with weight update rule (36) and some GNC schedule $\{\epsilon^{(t)}\}_t$. Assume $\|v^{(t)}\|_2 < \infty$ for every t. Suppose that $\{\epsilon^{(t)}\}_t$ is non-increasing and converges to ζ with $\zeta > 0$. Under Assumptions 1-4, every accumulation point of $\{v^{(t)}\}_t$ is a stationary point of the outlier-robust estimation problem (32).*

Proof of Theorem 4. The proof follows closely from that of Theorem 1. Recall the definition $r_i^{(t)} = r(v^{(t)}, d_i)$. Note that, by the weighting strategy (36), $v^{(t+1)}$ is a global minimizer of the quadratic majorizer (34) (also recall Remark 3):

$$\begin{aligned} v^{(t+1)} &\in \operatorname*{argmin}_{v \in \mathcal{C}} \sum\nolimits_{i=1}^m w_i^{(t+1)} r(v, d_i)^2 \\ &= \operatorname*{argmin}_{v \in \mathcal{C}} \sum\nolimits_{i=1}^m q_{\epsilon^{(t)}}\left(r(v, d_i), r_i^{(t)}\right) \end{aligned}$$

Since $\{v^{(t)}\}_t$ is bounded, it has a subsequence $\{v^{(t_j)}\}_j$ convergent to some point say $\hat{v} \in \mathcal{C}$. For this subsequence, it holds that (by identical arguments as in (21) and (23))

$$\sum_{i=1}^m q_{\epsilon^{(t_{j+1})}}\left(r_i^{(t_{j+1})}, r_i^{(t_{j+1})}\right) \leq \sum_{i=1}^m q_{\epsilon^{(t_j)}}\left(r(v, d_i), r_i^{(t_j)}\right), \quad \forall v \in \mathcal{C}.$$

Since $\rho'_{\epsilon^{(t_j)}}$ is continuously differentiable, we see that $q_{\epsilon^{(t_j)}}$ (34) is continuous in both of its parameters. Therefore, as $t_j \to \infty$, the above inequality becomes

$$\begin{aligned} &\sum_{i=1}^m q_\zeta\left(r(\hat{v}, d_i), r(\hat{v}, d_i)\right) \leq \sum_{i=1}^m q_\zeta\left(r(v, d_i), r(\hat{v}, d_i)\right), \quad \forall v \in \mathcal{C} \\ \Rightarrow\ &\hat{v} \in \operatorname*{argmin}_{v \in \mathcal{C}} \sum_{i=1}^m q_\zeta\left(r(v, d_i), r(\hat{v}, d_i)\right) = \hat{v} \in \operatorname*{argmin}_{v \in \mathcal{C}} \sum_{i=1}^m \hat{w}_i \cdot r(v, d_i)^2 \end{aligned}$$

where $\hat{w}_i$ is the weight of the i-th sample as $t_j \to \infty$, that is

$$\hat{w}_i = \frac{\rho_\zeta'\big(r(\hat{v}, \acute{a}_i)\big)}{2r(\hat{v}, \acute{a}_i)} \qquad \text{if } \; r(\hat{v}, \acute{a}_i) \neq 0$$
$$\hat{w}_i = \Big(\rho_\zeta\big(\sqrt{r(\hat{v}, \acute{a}_i)}\big)\Big)' \quad \text{if } \; r(\hat{v}, \acute{a}_i) = 0$$

Global optimality of $\hat{v}$ for the weighted least-squares problem in the limit implies the optimality condition [3, Section 5.3]

$$\Big(\sum_{i=1}^m \hat{w}_i \cdot \nabla\big(r(\hat{v}, \acute{a}_i)^2\big)\Big)^\top \acute{b} \geq 0, \quad \forall \acute{b} \in T_{\hat{v}}\mathcal{C}.$$

As before, we finish the proof by showing that $\hat{w}_i \cdot \nabla\big(r(\hat{v}, \acute{a}_i)^2\big)$ is equal to $\nabla\rho_\zeta\big(r(\hat{v}, \acute{a}_i)\big)$. Indeed, if $r(\hat{v}, \acute{a}_i) \neq 0$ then

$$\begin{aligned}\hat{w}_i \cdot \nabla\big(r(\hat{v}, \acute{a}_i)^2\big) &= \hat{w}_i \cdot 2r(\hat{v}, \acute{a}_i) \cdot \nabla\big(r(\hat{v}, \acute{a}_i)\big) \\ &= \rho_\zeta'\big(r(\hat{v}, \acute{a}_i)\big) \cdot \nabla\big(r(\hat{v}, \acute{a}_i)\big) \\ &= \nabla\rho_\zeta\big(r(\hat{v}, \acute{a}_i)\big).\end{aligned}$$

If $r(\hat{v}, \acute{a}_i) = 0$, using (35) and basic limit operations we obtain

$$\begin{aligned}\hat{w}_i \cdot \nabla_{r(\hat{v},\acute{a}_i)=0}\big(r(\hat{v}, \acute{a}_i)^2\big) &= \Big(\rho_\zeta\big(\sqrt{r(\hat{v}, \acute{a}_i)}\big)\Big)' \cdot 2r(\hat{v}, \acute{a}_i) \cdot \nabla_{r(\hat{v},\acute{a}_i)=0}\big(r(\hat{v}, \acute{a}_i)\big) \\ &= \nabla_{r(\hat{v},\acute{a}_i)=0}\big(r(\hat{v}, \acute{a}_i)\big) \cdot \lim_{r(\hat{v},\acute{a}_i)\to 0} 2r(\hat{v}, \acute{a}_i) \cdot \lim_{r(\hat{v},\acute{a}_i)\to 0} \frac{\rho_\zeta'(r(\hat{v}, \acute{a}_i))}{2r(\hat{v}, \acute{a}_i)} \\ &= \nabla_{r(\hat{v},\acute{a}_i)=0}\big(r(\hat{v}, \acute{a}_i)\big) \cdot \lim_{r(\hat{v},\acute{a}_i)\to 0} \rho_\zeta'(r(\hat{v}, \acute{a}_i)) \\ &= \nabla_{r(\hat{v},\acute{a}_i)=0}\big(r(\hat{v}, \acute{a}_i)\big) \cdot \rho_\zeta'(r(\hat{v}, \acute{a}_i))\big|_{r(\hat{v},\acute{a}_i)=0} \\ &= \nabla_{r(\hat{v},\acute{a}_i)=0}\Big(\rho_\zeta\big(r(\hat{v}, \acute{a}_i)\big)\Big)\end{aligned}$$

The proof is now complete. □

F. Proof of Theorem 2

Here we collect notations that are used throughout our proofs. Let $p \in [0, 1]$. Let the weight $w_i^{(t+1)} \in \mathbb{R}$, vector $\boldsymbol{x}^{(t)} \in \mathbb{R}^n$, and smoothing parameter $\epsilon^{(t)} \in \mathbb{R}$ be defined as per Algorithm 1. Let $\boldsymbol{W}^{(t+1)} \in \mathbb{R}^{m\times m}$ be the diagonal matrix with $w_i^{(t+1)}$ its diagonal, *i.e.*, $\boldsymbol{W}^{(t+1)} = \operatorname{diag}\big(w_1^{(t+1)}, \ldots, w_m^{(t+1)}\big)$. The ground truth signal is denoted by $\boldsymbol{x}^*$ and $\boldsymbol{r}^* = [r_1^*, \ldots, r_m^*]^\top \in \mathbb{R}^m$ with $r_i^* := \boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i$ is the associated residual vector, which is assumed to be k-sparse. Denote by S^* the support of $\boldsymbol{r}^*$, that is $S^* := \{i : r_i^* \neq 0\} \subset \{1, \ldots, m\}$.

By our assumption, we have $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2 \leq \epsilon^{(0)}$. Therefore, under other assumptions of Theorem 2, we can invoke Theorem 5 to obtain

$$\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2 \leq \beta(\epsilon^{(0)})^{2-p}$$

with probability at least $1 - P_0' - P_1' - P_2'$, where each P_i' is defined in (38). In fact, with a union bound, we can invoke Theorem 5 for t times and get

$$\|\boldsymbol{x}^{(t_0+1)} - \boldsymbol{x}^*\|_2 \leq \beta(\epsilon^{(t_0)})^{2-p}, \quad \forall t_0 = 0, 1, \ldots, t-1, \tag{37}$$

with probability at least $1 - tP_0' - tP_1' - tP_2'$. Via simple algebraic manipulation, we see that (37) implies the desired bound (13) of Theorem 2. It is easy to show that the probability term $1 - tP_0' - tP_1' - tP_2'$ matches the probability term $1 - P_0 - P_1 - P_2$ of Theorem 2. For example, we have

$$tP_0' = \exp(-\Omega(n - \log m - \log t)) = \exp(-\tilde{\Omega}(n)) = P_0.$$

Note that the logarithmic term $\log t$ is suppressed by $\tilde{\Omega}$ as t is a constant. Similarly we have $tP_1' = P_1$ and $tP_2' = P_2$. The proof is complete, provided we can prove Theorem 5. This is the most technical part and is done in the next section.

G. Theorem 5 and Its Proof

Theorem 5. *Suppose $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Denote by $r^*_{\text{min+}}$ the smallest non-zero residual among $\{|\boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i|\}_{i=1}^m$. Define*

$$\alpha := \frac{\sqrt{5}\cdot 2^{2-p}}{0.99\cdot 0.516}\cdot\frac{1}{\big(r^*_{\text{min+}}\big)^{1-p}}\cdot\frac{\sqrt{k}\cdot\big(1.01\sqrt{k}+\sqrt{n}\big)}{(m-k)}.$$

If $\alpha \le \beta$ (β is defined in (11)), then we have

$$\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\|_2 \le \beta(\epsilon^{(t)})^{2-p}$$

with probability at least $1 - P_0' - P_1' - P_2'$, where

$$\begin{aligned} P_0' &:= \exp(-\Omega(n-\log m)) \\ P_1' &:= \exp(-\Omega(k-n)) \\ P_2' &:= \exp\big(-\Omega(m-k-n\log n)\big). \end{aligned} \tag{38}$$

Proof of Theorem 5. To begin with, consider the event $\mathscr{E}$

$$\mathscr{E} := \{\sqrt{0.5n} \le M \le \sqrt{5n}\}, \qquad M := \max_{i=1,\dots,m}\|\boldsymbol{a}_i\|_2 \tag{39}$$

By Lemma 10 and a union bound we have with probability at least $1 - m\cdot\exp(-n) - \exp(-n/16) = 1 - P_0'$ that

$$\mathbb{P}(\mathscr{E}) \ge 1 - m\cdot\exp(-n) - \exp(-n/16) = 1 - \exp(-\Omega(n-\log m)). \tag{40}$$

As indicated by (40), we will heavily use the big-Ω notation to suppress lower-order and constant terms.

Our proof will condition on the event $\mathscr{E}$. In other words, We assume $\mathscr{E}$ happens, and makes derivations. After all of the derivations say in Proposition 4 and Proposition 5, we will get some high probability bounds. We can apply a union bound and take the probability bound $1 - \exp(-\Omega(n-\log m))$ for $\mathscr{E}$ into account.

Note that $\boldsymbol{r}^* = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$, so we have

$$\begin{aligned} \boldsymbol{x}^{(t+1)} &= \big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\big)^{-1}\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{y} \\ &= \big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\big)^{-1}\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{r}^*) \\ &= \boldsymbol{x}^* - \big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\big)^{-1}\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*. \end{aligned}$$

This gives us an upper bound on the distance between $\boldsymbol{x}^{(t+1)}$ and $\boldsymbol{x}^*$:

$$\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\|_2 \le \frac{\|\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2}{\lambda_{\min}\big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\big)} \tag{41}$$

We can finish the proof by invoking Proposition 4 and Proposition 5, which provides high probability bounds for $\|\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2$ and $\lambda_{\min}\big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\big)$ respectively. Specifically, conditioned on $\mathscr{E}$ (39), Proposition 4 gives

$$\|\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2 \le \sqrt{5}\cdot 2^{2-p}\cdot\left(\frac{1}{r^*_{\text{min+}}}\right)^{1-p}\cdot\sqrt{k}\cdot(1.01\sqrt{k}+\sqrt{n})$$

with probability at least $1 - \exp(-\Omega(k-n)) = 1 - P_1'$. On the other hand, by assumption we have $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}$. Therefore, conditioned on $\mathscr{E}$ (39), we can invoke Proposition 5 and obtain

$$\lambda_{\min}\Big(\boldsymbol{A}^\top\boldsymbol{W}^{(t+1)}\boldsymbol{A}\Big) \ge 0.99\cdot 0.516\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2}$$

with probability at least $1 - \exp\big(-\Omega(m-k-n\log n)\big) = 1 - P_2'$. Combining the above with a union bound, we have

$$\begin{aligned} \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\|_2 &\le \frac{\sqrt{5}\cdot 2^{2-p}}{0.99\cdot 0.516}\cdot\frac{1}{\big(r^*_{\text{min+}}\big)^{1-p}}\cdot\frac{\sqrt{k}\cdot\big(1.01\sqrt{k}+\sqrt{n}\big)}{(m-k)}\cdot\big(\epsilon^{(t)}\big)^{2-p} \\ &= \alpha\cdot\big(\epsilon^{(t)}\big)^{2-p} \\ &\le \beta\cdot\big(\epsilon^{(t)}\big)^{2-p}. \end{aligned}$$

Here we used the definition of α and the condition $\alpha \le \beta$. The proof is complete. □

G.1. Proposition 4 and Its Proof

Proposition 4. *Assume* $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}$ *and* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *has i.i.d.* $\mathcal{N}(0,1)$ *entries. Assume event* $\mathscr{E}$ (39) *happens. Then*

$$\|\boldsymbol{A}^\top \boldsymbol{W}^{(t+1)} \boldsymbol{r}^*\|_2 \le \sqrt{5}\cdot 2^{2-p}\cdot \left(\frac{1}{r^*_{\text{min+}}}\right)^{1-p} \cdot \sqrt{k}\cdot (1.01\sqrt{k}+\sqrt{n})$$

with probability at least $1-\exp(-\Omega(k-n))$.

Proof. Define $\boldsymbol{z}^{(t)} := \boldsymbol{x}^{(t)} - \boldsymbol{x}^*$ and $\overline{\boldsymbol{z}}^{(t)} := \boldsymbol{z}^{(t)}/\|\boldsymbol{z}^{(t)}\|_2$. Define $r_i^* := \boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i$. We first bound $\|\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2^2$. Recall that S^* is the support of $\boldsymbol{r}^*$ and $0 < r^*_{\text{min+}} \le |r_i^*|$ for any $i \in S^*$. Then we have

$$\|\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2^2 = \sum_{i\in S^*} (w_i^{(t+1)} r_i^*)^2 \le \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \sum_{i\in S^*} \left(w_i^{(t+1)}\right)^2 \cdot (r_i^*)^2 \cdot (r_i^*)^{2-2p}.$$

We need to bound $\left(w_i^{(t+1)}\right)^2 \cdot (r_i^*)^{4-2p}$ for every $i \in S^*$. To do so, we consider three cases.

Case 1. If $|r_i^*| > 2\cdot |\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)}|$ then we have

$$\begin{aligned} |\boldsymbol{a}_i^\top \boldsymbol{x}^{(t)} - y_i| &= |\boldsymbol{a}_i^\top \boldsymbol{x}^{(t)} - \boldsymbol{a}_i^\top \boldsymbol{x}^* + r_i^*| \\ &= |\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)} + r_i^*| \\ &\ge |r_i^*| - |\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)}| \\ &\ge |r_i^*|/2, \end{aligned}$$

which by definition of the weights $w_i^{(t+1)}$ of Algorithm 1 implies

$$\left(w_i^{(t+1)}\right)^2 \cdot (r_i^*)^{4-2p} \le \left(\frac{|r_i^*|}{|\boldsymbol{a}_i^\top \boldsymbol{x}^{(t)} - y_i|}\right)^{4-2p} \le 2^{4-2p} = 4^{2-p}.$$

Case 2. If $|\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)}| \ne 0$ and $|r_i^*| \le 2\cdot |\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)}|$, then, with $w_i^{(t+1)} \le \left(\epsilon^{(t)}\right)^{p-2}$ (Algorithm 1), the assumption $\|\boldsymbol{z}^{(t)}\|_2 \le \epsilon^{(t)}$, and the definition $\overline{\boldsymbol{z}}^{(t)} := \boldsymbol{z}^{(t)}/\|\boldsymbol{z}^{(t)}\|_2$, we obtain

$$\begin{aligned} \left(w_i^{(t+1)}\right)^2 \cdot (r_i^*)^{4-2p} &\le 2^{4-2p}\cdot \frac{(\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)})^{4-2p}}{\left(\epsilon^{(t)}\right)^{4-2p}} \\ &\le 4^{2-p}\cdot \left(\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}^{(t)}\right)^{4-2p}. \end{aligned}$$

Case 3. If $|\boldsymbol{a}_i^\top \boldsymbol{z}^{(t)}| = 0$, then Case 1 implies that $\left(w_i^{(t+1)}\right)^2 \cdot (r_i^*)^{4-2p} \le 1$.

Combining the above, we obtain that

$$\begin{aligned} \|\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2^2 &\le \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \sum_{i\in S^*} \max\left\{1, 4^{2-p}, 4^{2-p}\cdot \left(\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}^{(t)}\right)^{4-2p}\right\} \\ &= \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \sum_{i\in S^*} \max\left\{4^{2-p}, 4^{2-p}\cdot \left(\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}^{(t)}\right)^{4-2p}\right\} \\ &\le 4^{2-p}\cdot \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \cdot \left(k + \sum_{i\in S^*} \left(\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}^{(t)}\right)^{4-2p}\right) \\ &\overset{\text{(i)}}{\le} 4^{2-p}\cdot \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \cdot (k+4k) \\ &= 5\cdot 4^{2-p}\cdot \left(\frac{1}{r^*_{\text{min+}}}\right)^{2-2p} \cdot k \end{aligned}$$

where step (i) follows from Lemma 5 which assumes event $\mathscr{E}$ happens and it holds with probability at least $1-\exp(-\Omega(k-n))$. Therefore, we have

$$\|\boldsymbol{A}^\top \boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2 \le \|\boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2 \cdot \sqrt{\lambda_{\max}\Big(\sum_{i\in S^*}\boldsymbol{a}_i\boldsymbol{a}_i^\top\Big)},$$

we can invoke Lemma 11 with $\delta_1 = 0.01$ to upper bound the maximum eigenvalue of $\sum_{i\in S^*}\boldsymbol{a}_i\boldsymbol{a}_i^\top$. Therefore, invoking a union bound, it holds with probability at least $1-\exp(-\Omega(k-n))-\exp(-\Omega(k)) = 1-\exp(-\Omega(k-n))$ that

$$\|\boldsymbol{A}^\top \boldsymbol{W}^{(t+1)}\boldsymbol{r}^*\|_2 \le \sqrt{5}\cdot 2^{2-p}\cdot\left(\frac{1}{r^*_{\text{min+}}}\right)^{1-p}\cdot\sqrt{k}\cdot(1.01\sqrt{k}+\sqrt{n}).$$

The proof is now complete. □

G.1.1 Auxiliary Lemmas for Proposition 4

Lemma 3 (Locally Lipschitz Continuity)**.** *Let $p\in[0,1]$ and $c>0$. The function $\phi:[-d,d]\to\mathbb{R}$ with $\phi(a)=a^{4-2p}$ is convex and differentiable, and thus for any $a,b\in[-d,d]$ we have*

$$\phi(a)-\phi(b)\le\phi'(a)\cdot(a-b)\le(4-2p)\cdot d^{3-2p}\cdot|a-b|.$$

In other words, ϕ is Lipschitz over $[-d,d]$ with constant $(4-2p)\cdot d^{3-2p}$.

Proof. This follows directly from the definition of convexity for differentiable functions, and the fact that $\phi'(a)=(4-2p)\cdot a^{3-2p}$ is bounded above by $(4-2p)\cdot c^{3-2p}$. □

Lemma 4 (Moments of Gaussian Scalar)**.** *For $p\in[0,1]$ and a standard Gaussian variable $u\sim\mathcal{N}(0,1)$, we have*

$$\mathbb{E}[u^q]\le 4.$$

Proof. we have $\mathbb{E}[u^2]=1$ and $\mathbb{E}[u^4]=(4-1)!!=3$.[6] Therefore, for $p\in[0,1]$ it holds that

$$\begin{aligned}
\mathbb{E}[u^{4-2p}] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u^{4-2p}\exp(-u^2/2)\,du\\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\Big(\int_0^1 u^{4-2p}\exp(-u^2/2)\,du+\int_1^\infty u^{4-2p}\exp(-u^2/2)\,du\Big)\\
&\le \frac{\sqrt{2}}{\sqrt{\pi}}\Big(\int_0^1 u^{2}\exp(-u^2/2)\,du+\int_1^\infty u^{4}\exp(-u^2/2)\,du\Big)\\
&\le \frac{\sqrt{2}}{\sqrt{\pi}}\Big(\int_0^\infty u^{2}\exp(-u^2/2)\,du+\int_0^\infty u^{4}\exp(-u^2/2)\,du\Big)\\
&= \mathbb{E}[u^2]+\mathbb{E}[u^4]=4.
\end{aligned}$$

The proof is complete. Note that, mathematically, this bound 4 might be sub-optimal, as one would expect that $\mathbb{E}[u^{4-2p}]$ takes values between $\mathbb{E}[u^2]=1$ and $\mathbb{E}[u^4]=3$. However, it is harmless for our purpose as it is just a constant. □

Lemma 5. *Suppose $\boldsymbol{a}_i\in\mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries, $\forall i=1,\dots,k$. Assume event $\mathscr{E}$ (39) happens and $p\in[0,1]$. We have*

$$\max_{\boldsymbol{z}\in\mathbb{S}^{n-1}}\sum_{i=1}^{k}\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^{4-2p}\le 4k \tag{42}$$

with probability at least $1-\exp(-\Omega(k-n))$.

[6] More generally, it holds that $\mathbb{E}[u^{2q}]=(2q-1)!!$ for every positive integer q. Here $(\cdot)!!$ denotes the *double factorial* of some number.

Proof. Note that event $\mathscr{E}$ (39) happens by assumption. In the proof we implicitly condition on $\mathscr{E}$ (39). First, by (39), we know that $(\boldsymbol{a}_1^\top \boldsymbol{z})^{4-2p}, \ldots, (\boldsymbol{a}_k^\top \boldsymbol{z})^{4-2p}$ are independent random variables, and they are bounded with $0 \le |\boldsymbol{a}_i^\top \boldsymbol{z}|^{4-2p} \le \max_{i=1,\ldots,m} \|\boldsymbol{a}_i\|_2^{4-2p} \le M^{4-2p}$.

Pointwise Bound. We first consider a fixed $\boldsymbol{z} \in \mathbb{S}^{n-1}$, and derive a high probability upper bound for $\sum_{i=1}^k \big(\boldsymbol{a}_i^\top \boldsymbol{z}\big)^{4-2p}$. Since $\boldsymbol{a}_i^\top \boldsymbol{z} \sim \mathcal{N}(0,1)$ for every $i = 1, \ldots, k$, it follows from Lemma 4 that

$$\sum_{i=1}^k \mathbb{E}\big[(\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p}\big] \le 4k. \tag{43}$$

By the standard Hoeffding bound on independent zero mean bounded random variables (i), for any $\delta > 0$, we have

$$\mathbb{P}\bigg(\sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p} \ge (4+\delta)k\bigg) \overset{(43)}{\le} \mathbb{P}\bigg(\sum_{i=1}^k \Big((\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p} - \mathbb{E}\big[(\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p}\big]\Big) \ge \delta k\bigg) \tag{44}$$

$$\overset{\text{(i)}}{\le} \exp\Big(-\frac{2k\delta^2}{M^{8-4p}}\Big) =: P_\delta \tag{45}$$

Union Bound. Similarly to the proof of Lemma 8 we consider a $(1/4)$-net $\mathcal{N}_{0.25}$ of the sphere $\mathbb{S}^{n-1}$; $\mathcal{N}_{0.25}$ has at most 9^n points. A direct application of the union bound yields

$$\sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p} \ge (4+\delta)k, \quad \forall \boldsymbol{z} \in \mathcal{N}_{0.25},$$

with probability at most $9^n \cdot P_\delta$.

Approximation Bound. With (43), we now bound $\sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z})^{4-2p}$ for every $\boldsymbol{z} \in \mathbb{S}^{n-1}$. Indeed, for any $\boldsymbol{z}_1 \in \mathbb{S}^n$, there is some $\boldsymbol{z}_2 \in \mathcal{N}_{0.25}$ such that $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 \le 1/4$. Lemma 3 implies that

$$\bigg|\sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z}_1)^{4-2p} - \sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z}_2)^{4-2p}\bigg| \le \sum_{i=1}^k (4-2p) \cdot M^{3-2p} \cdot |\boldsymbol{a}_i^\top \boldsymbol{z}_1 - \boldsymbol{a}_i^\top \boldsymbol{z}_2| \le (1-p/2) \cdot M^{4-2p} \cdot k. \tag{46}$$

Now, take (45) with $\delta = (1-p/2) \cdot M^{4-2p}$, and we have $P_\delta = \exp\big(-k(2-p)^2/2\big)$. Combining (45) with (46) gives

$$\mathbb{P}\bigg(\sum_{i=1}^k (\boldsymbol{a}_i^\top \boldsymbol{z}_1)^{4-2p} \ge 4k\bigg) \le 9^n \exp\big(-k(2-p)^2/2\big), \quad \forall \boldsymbol{z}_1 \in \mathbb{S}^{n-1}.$$

In other words, (42) holds with probability at least $1 - \exp(-\Omega(k-n))$. The proof is complete. □

G.2. Proposition 5 and Its Proof

Proposition 5. *Let $\mathbb{B}^{(t+1)}(\boldsymbol{x}^*; \epsilon^{(t)})$ be the ℓ_2-ball centered at the ground truth $\boldsymbol{x}^*$ with radius $\epsilon^{(t)}$, that is*

$$\mathbb{B}^{(t+1)}(\boldsymbol{x}^*; \epsilon^{(t)}) := \Big\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}\Big\}.$$

Associate each vector $\boldsymbol{x}$ in the ball with an $m \times m$ diagonal matrix of weights, $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_m)$, where $w_i = \max\big\{|\boldsymbol{a}_i^\top \boldsymbol{x} - y_i|, \epsilon^{(t)}\big\}^{p-2}$. This gives a product set

$$\mathcal{Q}_p^{(t+1)} := \bigg\{\big(\boldsymbol{W}, \boldsymbol{x}\big) = \big(\mathrm{diag}(w_1, \ldots, w_m), \boldsymbol{x}\big) \in \mathbb{R}^{m\times m} \times \mathbb{B}^{(t+1)}(\boldsymbol{x}^*; \epsilon^{(t)}) : w_i = \max\big\{|\boldsymbol{a}_i^\top \boldsymbol{x} - y_i|, \epsilon^{(t)}\big\}^{p-2}\bigg\}. \tag{47}$$

Assume event $\mathscr{E}$ (39) happens. Then we have

$$\lambda_{\min}\Big(\boldsymbol{A}^\top \boldsymbol{W} \boldsymbol{A}\Big) \ge 0.99 \cdot 0.516 \cdot (m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2}, \quad \forall (\boldsymbol{W}, \boldsymbol{x}) \in \mathcal{Q}_p^{(t+1)},$$

with probability at least $1 - \exp\big(-\Omega(m - k - n\log n)\big)$.

Thus, the same conclusion holds for the iterate $\big(\boldsymbol{W}^{(t+1)}, \boldsymbol{x}^{(t)}\big)$ of our Algorithm 1 with as least the same probability, as long as $\big(\boldsymbol{W}^{(t+1)}, \boldsymbol{x}^{(t)}\big) \in \mathcal{Q}_p^{(t+1)}$ or equivalently $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}$.

Proof. Note that we assume event $\mathscr{E}$ (39) takes place. The derivation here implicitly conditions on this event.

Since $\boldsymbol{A}^\top \boldsymbol{W} \boldsymbol{A} = \sum_{i=1}^m w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \succeq \sum_{i \in (S^*)^c} w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$, we will bound the minimum eigenvalue of the latter. Recall the definition of M in (39), and consider a τ-net $\mathcal{N}_\tau$ over the ball $\mathbb{B}^{(t+1)}(\boldsymbol{x}^*; \epsilon^{(t)})$, with

$$\tau = \frac{0.516 \epsilon^{(t)}}{400 \cdot (2-p) \cdot M^3}.$$

Since event $\mathscr{E}$ happens (39), we have $M \in [\sqrt{0.5n}, 5n]$, therefore

$$\frac{\tau M}{\epsilon^{(t)}} = \frac{0.516}{400 \cdot (2-p) \cdot M^2} \leq \frac{0.516}{400 \cdot (2-p) \cdot n/2} \leq 1, \tag{48}$$

$$\frac{\tau}{\epsilon^{(t)}} = \frac{0.516}{400 \cdot (2-p) \cdot M^3} \leq 1. \tag{49}$$

Note that Lemma 13 and (49) imply that this τ-net is of size at most

$$T_\tau := \Big(\frac{3\epsilon^{(t)}}{\tau}\Big)^n = \Big(\frac{1200 \cdot (2-p) \cdot M^3}{0.516}\Big)^n.$$

Note that every $\boldsymbol{x}_\tau \in \mathcal{N}_\tau$ induces a weight matrix $\boldsymbol{W}_\tau = \mathrm{diag}(w_{\tau 1}, \ldots, w_{\tau m})$ such that $(\boldsymbol{W}_\tau, \boldsymbol{x}_\tau) \in \mathcal{Q}_p^{(t+1)}$. We can now invoke Lemma 8 with a union bound, to obtain

$$\lambda_{\min}\bigg(\sum_{i \in (S^*)^c} w_{\tau i} \boldsymbol{a}_i \boldsymbol{a}_i^\top\bigg) \geq 0.995 \cdot 0.516 \cdot (m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2}, \quad \forall \boldsymbol{x}_\tau \in \mathcal{N}_\tau,$$

which holds with probability at least $1 - P_\tau$, where $P_\tau := T_\tau \cdot \exp(-\Omega(m-k-n))$ with $\Omega(\cdot)$ hiding constants from Lemma 8. Since for any $(\boldsymbol{W}, \boldsymbol{x}) \in \mathcal{Q}_p^{(t+1)}$, there exists some $\boldsymbol{x}_\tau \in \mathcal{N}_\tau$ with $\|\boldsymbol{x}_\tau - \boldsymbol{x}\|_2 \leq \tau$, combine this with (48) and we can now invoke Lemma 9, which implies

$$\begin{aligned}\bigg|\lambda_{\min}\bigg(\sum_{i \in (S^*)^c} w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top\bigg) - \lambda_{\min}\bigg(\sum_{i \in (S^*)^c} w_{\tau i} \boldsymbol{a}_i \boldsymbol{a}_i^\top\bigg)\bigg| &\leq 2 \cdot (2-p) \cdot (m-k) \cdot \frac{\tau M^3}{\big(\epsilon^{(t)}\big)^{3-p}} \\ &= 0.005 \cdot 0.516 \cdot (m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2}.\end{aligned}$$

Combining the above with a union bound, we now have

$$\lambda_{\min}\bigg(\sum_{i \in (S^*)^c} w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top\bigg) \geq 0.99 \cdot 0.516 \cdot (m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2}, \quad \forall (\boldsymbol{W}, \boldsymbol{x}) \in \mathcal{Q}_p^{(t+1)}$$

with probability at least $1 - P_\tau$. To finish the proof, it remains to simplify the expression of $P_\tau = T_\tau \cdot \exp(-\Omega(m-k-n))$. Note that $M \leq \sqrt{5n}$, which implies $T_\tau = O(M^{3n}) = O(n^{3n/2})$ and $\log T_\tau = O(n \log n)$. Therefore

$$\begin{aligned}P_\tau &= \exp(-\Omega(m-k-n) + \log T_\tau) \\ &= \exp\big(-\Omega(m-k-n-n\log n)\big) \\ &= \exp\big(-\Omega(m-k-n\log n)\big).\end{aligned}$$

The proof of Proposition 5 is now complete. □

G.2.1 Auxiliary Lemmas for Proposition 5

Lemma 6 (Binomial Approximation). *For any $a \in [0,1]$ and $p \in [0,1]$, we have:*

$$(1+a)^{2-p} - 1 \leq 2 \cdot (2-p) \cdot a$$

Proof. Let p be fixed. Define a function $f : [0,1] \to \mathbb{R}$ such that

$$f(a) = (1+a)^{2-p} - 1 - 2 \cdot (2-p) \cdot a.$$

We see that $f(0) = 0$, and

$$\begin{aligned} f'(a) &= (2-p) \cdot (1+a)^{1-p} - 2 \cdot (2-p) \\ &= (2-p) \cdot \big((1+a)^{1-p} - 2\big). \end{aligned}$$

Since $1 \le 1+a \le 2$ for $0 \le a \le 1$ and $1-p \in [0,1]$, we know that $(1+a)^{1-p} \le 2$ and thus $f'(a) \le 0$. Hence, f is a decreasing function in $[0,1]$ with $f(a) \le f(0) = 0$ for any $a \in [0,1]$. □

Lemma 7 (Pointwise Expectation)**.** *Suppose for every $i = 1, \dots, m$, $\boldsymbol{a}_i \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries. Recall the definition of the product set $\mathcal{Q}_p^{(t+1)}$ (47). Let $\boldsymbol{v} \in \mathbb{S}^{n-1}$ and $\big(\boldsymbol{W}, \boldsymbol{x}\big) = \big(\operatorname{diag}(w_1, \dots, w_m), \boldsymbol{x}\big) \in \mathcal{Q}_p^{(t+1)}$ be fixed. Then it holds that*

$$(m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2} \ge \mathbb{E}\left[\sum_{i \in (S^*)^c} w_i \big(\boldsymbol{a}_i^\top \boldsymbol{v}\big)^2\right] \ge 0.516 \cdot (m-k) \cdot \big(\epsilon^{(t)}\big)^{p-2}. \tag{50}$$

Proof. The first inequality of (50) is obvious, and we now prove the second. For any $i \in (S^*)^c$ we have $y_i = \boldsymbol{a}_i^\top \boldsymbol{x}^*$. Note that $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \le \epsilon^{(t)}$. Define $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^*$ and $\overline{\boldsymbol{z}} = \boldsymbol{z}/\|\boldsymbol{z}\|_2$. Then

$$\begin{aligned} w_i^{(t+1)} &= \max\big\{\big|\boldsymbol{a}_i^\top \boldsymbol{x} - y_i\big|, \epsilon^{(t)}\big\}^{p-2} \\ &= \big(\epsilon^{(t)}\big)^{p-2} \cdot \max\left\{\frac{\big|\boldsymbol{a}_i^\top \boldsymbol{z}\big|}{\epsilon^{(t)}}, 1\right\}^{p-2} \\ &= \big(\epsilon^{(t)}\big)^{p-2} \cdot \max\left\{\frac{\big|\boldsymbol{a}_i^\top \boldsymbol{z}\big|}{\|\boldsymbol{z}\|_2} \cdot \frac{\|\boldsymbol{z}\|_2}{\epsilon^{(t)}}, 1\right\}^{p-2} \\ &\ge \big(\epsilon^{(t)}\big)^{p-2} \cdot \max\Big\{\big|\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}\big|, 1\Big\}^{p-2}. \end{aligned}$$

This leads to

$$\begin{aligned} \mathbb{E}\left[\sum_{i \in (S^*)^c} w_i \big(\boldsymbol{a}_i^\top \boldsymbol{v}\big)^2\right] &\ge \big(\epsilon^{(t)}\big)^{p-2} \cdot \mathbb{E}\left[\sum_{i \in (S^*)^c} \max\Big\{\big|\boldsymbol{a}_i^\top \overline{\boldsymbol{z}}\big|, 1\Big\}^{p-2} \cdot \big(\boldsymbol{a}_i^\top \boldsymbol{v}\big)^2\right] \\ &= \big(\epsilon^{(t)}\big)^{p-2} \cdot (m-k) \cdot \mathbb{E}_{\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{I}_n)}\left[\frac{\big(\boldsymbol{a}^\top \boldsymbol{v}\big)^2}{\max\big\{\big|\boldsymbol{a}^\top \overline{\boldsymbol{z}}\big|, 1\big\}^{2-p}}\right] \\ &\ge \big(\epsilon^{(t)}\big)^{p-2} \cdot (m-k) \cdot c_p, \end{aligned}$$

where c_p is defined as

$$c_p := \inf_{\boldsymbol{u} \in \mathbb{S}^{n-1}, \boldsymbol{q} \in \mathbb{S}^{n-1}} \mathbb{E}_{\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{I}_n)}\left[\frac{\big(\boldsymbol{a}^\top \boldsymbol{q}\big)^2}{\max\big\{\big|\boldsymbol{a}^\top \boldsymbol{u}\big|, 1\big\}^{2-p}}\right].$$

It then remains to show that $c_p \ge 0.516$. By rotation invariance, we can assume without loss of generality that $\boldsymbol{u} =$

$[u_1, 0, \ldots, 0]^\top$ and $\boldsymbol{q} = [q_1, q_2, 0, \ldots, 0]^\top$. Then we have

$$\begin{aligned}
c_p &= \inf_{\substack{u_1: u_1^2 = 1 \\ q_1, q_2: q_1^1 + q_2^2 = 1}} \mathbb{E}_{a_1 \sim \mathcal{N}(0,1), a_2 \sim \mathcal{N}(0,1)} \left[\frac{a_1^2 q_1^2 + a_2^2 q_2^2 + 2 a_1 a_2 q_1 q_2}{\max\left\{|a_1 u_1|, 1\right\}^{2-p}} \right] \\
&= \inf_{q_1, q_2: q_1^1 + q_2^2 = 1} \mathbb{E}_{a_1 \sim \mathcal{N}(0,1)} \left[\frac{a_1^2 q_1^2 + q_2^2}{\max\left\{|a_1|, 1\right\}^{2-p}} \right] \\
&= \frac{2}{\sqrt{2\pi}} \cdot \inf_{q_1, q_2: q_1^1 + q_2^2 = 1} \int_0^1 (t^2 q_1^2 + q_2^2) e^{-t^2/2} \, dt + \int_1^\infty \frac{t^2 q_1^2 + q_2^2}{t^{2-p}} \cdot e^{-t^2/2} \, dt \\
&\geq \inf_{q_1, q_2: q_1^1 + q_2^2 = 1} 0.516 q_1^2 + 0.849 q_2^2 \\
&\geq 0.516.
\end{aligned}$$

In the above, the (lower bound on the) integral was calculated by the MATLAB function `integral` with $p = 0$. This finishes the proof. □

Lemma 8 (Pointwise Concentration)**.** *Suppose $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Let S^* be the support of $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$. Recall the definition of the product set $\mathcal{Q}_p^{(t+1)}$ (47). With $\left(\boldsymbol{W}, \boldsymbol{x}\right) = \left(\operatorname{diag}(w_1, \ldots, w_m), \boldsymbol{x}\right) \in \mathcal{Q}_p^{(t+1)}$ fixed, we have*

$$\lambda_{\min}\left(\sum_{i \in (S^*)^c} w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \right) \geq 0.995 \cdot 0.516 \cdot (m-k) \cdot \left(\epsilon^{(t)}\right)^{p-2}$$

with probability at least $1 - \exp(-\Omega(m-k-n))$.

Proof. Define $\boldsymbol{B}_i := w_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$ and $\boldsymbol{B} := \sum_{i \in (S^*)^c} \boldsymbol{B}_i$. We now bound the minimum eigenvalue of $\boldsymbol{B}$. We first derive a high probability lower bound for $\sum_{i \in (S^*)^c} g_i^2$, where $g_i := \sqrt{w_i} \boldsymbol{a}_i^\top \boldsymbol{v}$ with a fixed spherical vector $\boldsymbol{v} \in \mathbb{S}^{n-1}$. Then we apply a union bound.

Bounding $\sum_{i \in (S^*)^c} g_i^2$. Since $w_i \leq (\epsilon^{(t)})^{p-2}$, the sub-Gaussian norm $\|g_i\|_{\psi_2}$ can be bounded above:

$$\begin{aligned}
\|g_i\|_{\psi_2} &\leq \|(\epsilon^{(t)})^{p/2-1} \cdot \boldsymbol{a}_i^\top \boldsymbol{v}\|_{\psi_2} \\
&= (\epsilon^{(t)})^{p/2-1} \cdot \|\boldsymbol{a}_i^\top \boldsymbol{v}\|_{\psi_2} \\
&= (\epsilon^{(t)})^{p/2-1} \cdot \|a\|_{\psi_2},
\end{aligned}$$

where $a \sim \mathcal{N}(0,1)$, whose sub-Gaussian norm $\|a\|_{\psi_2}$ is bounded above by some constant C_1, so we have $\|g_i\|_{\psi_2} \leq (\epsilon^{(t)})^{p/2-1} \cdot C_1$. Thus, g_i is a sub-Gaussian random variable with parameter $(\epsilon^{(t)})^{p/2-1} \cdot C_1$, which implies that $g_i^2 = w_i \cdot (\boldsymbol{a}_i^\top \boldsymbol{v})^2$ is sub-exponential with parameter $(\epsilon^{(t)})^{p-2} \cdot C_1^2$. Furthermore, using the centering technique we know that $g_i^2 - \mathbb{E}[g_i^2]$ is also sub-exponential, with parameter $2(\epsilon^{(t)})^{p-2} \cdot C_1^2$. Invoking Bernstein's inequality (Lemma 12), we have for any $\delta > 0$ that

$$\left| \sum_{i \in (S^*)^c} \left(g_i^2 - \mathbb{E}[g_i^2] \right) \right| \geq \delta \cdot (m-k) \cdot \left(\epsilon^{(t)}\right)^{p-2}$$

with probability at most

$$\begin{aligned}
P_\delta &:= 2 \cdot \exp\left[- C_2 \cdot \min\left\{ \frac{\delta^2 \cdot (m-k)^2 \cdot \left(\epsilon^{(t)}\right)^{2p-4}}{4(\epsilon^{(t)})^{2p-4} \cdot C_1^4 \cdot (m-k)}, \frac{\delta \cdot (m-k) \cdot \left(\epsilon^{(t)}\right)^{p-2}}{2 \cdot (\epsilon^{(t)})^{p-2} \cdot C_1^2} \right\} \right] \\
&= 2 \cdot \exp\left[- C_2 \cdot \min\left\{ \frac{\delta^2}{4 \cdot C_1^4}, \frac{\delta}{2 \cdot C_1^2} \right\} \cdot (m-k) \right],
\end{aligned}$$

where C_2 is a universal constant.

Union Bound. Consider a $(1/4)$-net $\mathcal{N}_{0.25}$ of the sphere $\mathbb{S}^{n-1}$, which has at most 9^n points (Lemma 13). Lemma 14 implies

$$\begin{aligned}\big\|\boldsymbol{B}-\mathbb{E}[\boldsymbol{B}]\big\|_2 &\le 2\cdot\max_{\boldsymbol{v}\in\mathcal{N}_{0.25}}\big|\boldsymbol{v}^\top(\boldsymbol{B}-\mathbb{E}[\boldsymbol{B}])\boldsymbol{v}\big|\\ &= 2\cdot\max_{\boldsymbol{v}\in\mathcal{N}_{0.25}}\sum_{i\in(S^*)^c}\Big|g_i^2-\mathbb{E}[g_i^2]\Big|.\end{aligned}$$

Thus, combine the above with a union bound and to obtain

$$\big\|\boldsymbol{B}-\mathbb{E}[\boldsymbol{B}]\big\|_2 \le 2\cdot\delta\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2}$$

with probability at least $1-9^n\cdot P_\delta$. Thus, with at least this probability we have

$$\boldsymbol{v}^\top\boldsymbol{B}\boldsymbol{v}\ge\boldsymbol{v}^\top\mathbb{E}[\boldsymbol{B}]\boldsymbol{v}-2\cdot\delta\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2},\forall\boldsymbol{v}\in\mathbb{S}^{n-1}.$$

In particular, for an eigenvector $\boldsymbol{v}_0$ of $\boldsymbol{B}$ that corresponds to its minimum eigenvalue, we have

$$\begin{aligned}\lambda_{\min}(\boldsymbol{B}) &\ge \boldsymbol{v}_0^\top\mathbb{E}[\boldsymbol{B}]\boldsymbol{v}_0-2\cdot\delta\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2}\\ &\ge \lambda_{\min}(\mathbb{E}[\boldsymbol{B}])-2\cdot\delta\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2}\\ &\ge 0.516\cdot(1-2\cdot\delta)\cdot(m-k)\cdot\big(\epsilon^{(t)}\big)^{p-2}\end{aligned}$$

with probability at least $1-9^n\cdot P_\delta$. In the last step, we used the lower bound of $\lambda_{\min}(\mathbb{E}[\boldsymbol{B}])$ derived in Lemma 7. Now, with $\delta=0.0025$, and noticing that $1-9^n\cdot P_\delta$ is of order $\exp(-\Omega(m-k-n))$, we finished the proof. □

Lemma 9 (Lipschitz Continuity of Minimum Eigenvalues)**.** *Recall $M:=\max_{i=1,\dots,m}\|\boldsymbol{a}_i\|_2$ (39) and the definition of the product set $\mathcal{Q}_p^{(t+1)}$ (47). Let $(\boldsymbol{W}_1,\boldsymbol{x}_1)$ and $(\boldsymbol{W}_2,\boldsymbol{x}_2)$ be two elements of $\mathcal{Q}_p^{(t+1)}$ with $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2\le\tau$. Write $\boldsymbol{W}_1=\mathrm{diag}(w_{1i},\dots,w_{1m})$ and $\boldsymbol{W}_2=\mathrm{diag}(w_{2i},\dots,w_{2m})$. If $\tau M\le\epsilon^{(t)}$, then*

$$\left|\lambda_{\min}\Big(\sum_{i\in(S^*)^c}w_{1i}\boldsymbol{a}_i\boldsymbol{a}_i^\top\Big)-\lambda_{\min}\Big(\sum_{i\in(S^*)^c}w_{2i}\boldsymbol{a}_i\boldsymbol{a}_i^\top\Big)\right|\le 2\cdot(2-p)\cdot(m-k)\cdot\frac{\tau M^3}{\big(\epsilon^{(t)}\big)^{3-p}}.$$

Proof. The proof is divided into four cases, depending on whether the i-th weights, w_{1i} and w_{2i}, are truncated by $\epsilon^{(t)}$ or not.
Case 1. If both $|\boldsymbol{a}_i^\top\boldsymbol{x}_1-y_i|\le\epsilon^{(t)}$ and $|\boldsymbol{a}_i^\top\boldsymbol{x}_2-y_i|\le\epsilon^{(t)}$, then we have $w_{1i}=w_{2i}$.
Case 2. Assume $|\boldsymbol{a}_i^\top\boldsymbol{x}_1-y_i|\le\epsilon^{(t)}$ and $|\boldsymbol{a}_i^\top\boldsymbol{x}_2-y_i|>\epsilon^{(t)}$. First note that

$$|\boldsymbol{a}_i^\top\boldsymbol{x}_2-y_i|=|\boldsymbol{a}_i^\top\boldsymbol{x}_2-\boldsymbol{a}_i^\top\boldsymbol{x}_1+\boldsymbol{a}_i^\top\boldsymbol{x}_1-y_i|\le\|\boldsymbol{a}_i\|_2\cdot\tau+|\boldsymbol{a}_i^\top\boldsymbol{x}_1-y_i|. \tag{51}$$

Then we have

$$\begin{aligned}|w_{i1}-w_{i2}| &= \Big|\big(\epsilon^{(t)}\big)^{p-2}-|\boldsymbol{a}_i^\top\boldsymbol{x}_2-y_i|^{p-2}\Big|\\ &= \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}}-\frac{1}{|\boldsymbol{a}_i^\top\boldsymbol{x}_2-y_i|^{2-p}}\\ &\le \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}}-\frac{1}{\big(\|\boldsymbol{a}_i\|_2\cdot\tau+|\boldsymbol{a}_i^\top\boldsymbol{x}_1-y_i|\big)^{2-p}}\\ &\le \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}}-\frac{1}{\big(\|\boldsymbol{a}_i\|_2\cdot\tau+\epsilon^{(t)}\big)^{2-p}}\\ &= \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}}\left(1-\frac{1}{\big(\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}}+1\big)^{2-p}}\right)\\ &\le \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}}\left(1-\frac{1}{1+2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}}}\right).\end{aligned}$$

In the last step we used the assumption $\|\boldsymbol{a}_i\|_2 \cdot \tau \le \epsilon^{(t)}$ and Lemma 6. Then we get

$$\begin{aligned}
\big|w_{i1} - w_{i2}\big| &\le \frac{1}{\big(\epsilon^{(t)}\big)^{2-p}} \cdot \frac{2\cdot(2-p)\cdot \frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}}}{1+2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}}} \\
&= 2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\big(\epsilon^{(t)}\big)^{3-p}} \cdot \frac{1}{1+2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}}} \\
&\le 2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\big(\epsilon^{(t)}\big)^{3-p}}.
\end{aligned}$$

Case 3. If $|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i| > \epsilon^{(t)}$ and $|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i| \le \epsilon^{(t)}$, then we can bound $|w_{i1} - w_{i2}|$ similarly as in Case 2 (by symmetry), and obtain the same upper bound.

Case 4. Suppose $|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i| > \epsilon^{(t)}$ and $|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i| > \epsilon^{(t)}$, then, assuming $|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p}$ is larger than or equal to $|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i|^{2-p}$ without loss of generality, we have

$$\begin{aligned}
\big|w_{i1} - w_{i2}\big| &= \left|\frac{1}{|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p}} - \frac{1}{|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i|^{2-p}}\right| \\
&= \frac{|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p} - |\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i|^{2-p}}{|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p}\cdot|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i|^{2-p}} \\
&= \frac{\frac{\big|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i\big|^{2-p}}{\big|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i\big|^{2-p}} - 1}{|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p}}.
\end{aligned}$$

Similarly in (51), we have $|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i| \le |\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i| + \|\boldsymbol{a}_i\|_2 \cdot \tau$, and consequently

$$\begin{aligned}
\big|w_{i1} - w_{i2}\big| &\le \frac{\left|\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\big|\boldsymbol{a}_i^\top \boldsymbol{x}_2 - y_i\big|} + 1\right|^{2-p} - 1}{|\boldsymbol{a}_i^\top \boldsymbol{x}_1 - y_i|^{2-p}} \\
&\le \frac{\left|\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\epsilon^{(t)}} + 1\right|^{2-p} - 1}{\big(\epsilon^{(t)}\big)^{2-p}} \\
&\le 2\cdot(2-p)\cdot\frac{\|\boldsymbol{a}_i\|_2\cdot\tau}{\big(\epsilon^{(t)}\big)^{3-p}}
\end{aligned}$$

where we used the assumption $\|\boldsymbol{a}_i\|_2 \cdot \tau \le \epsilon^{(t)}$ and Lemma 6. To summarize, we have shown

$$\begin{aligned}
\sum_{i\in(S^*)^c} \big|w_{i1} - w_{i2}\big| &\le 2\cdot(2-p)\cdot(m-k)\cdot\frac{\tau}{\big(\epsilon^{(t)}\big)^{3-p}}\cdot\max_{i\in(S^*)^c}\|\boldsymbol{a}_i\|_2 \\
&\le 2\cdot(2-p)\cdot(m-k)\cdot\frac{\tau M}{\big(\epsilon^{(t)}\big)^{3-p}}.
\end{aligned}$$

Next, suppose without loss of generality that the minimum eigenvalue of $\sum_{i\in(S^*)^c} w_{1i}\boldsymbol{a}_i\boldsymbol{a}_i^\top$ is larger than or equal to that of $\sum_{i\in(S^*)^c} w_{2i}\boldsymbol{a}_i\boldsymbol{a}_i^\top$, and denote by $\boldsymbol{v}_0 \in \mathbb{S}^{n-1}$ an eigenvector of $\sum_{i\in(S^*)^c} w_{2i}\boldsymbol{a}_i\boldsymbol{a}_i^\top$ corresponding to its minimum eigenvalue. Then we have

$$\begin{aligned}
\lambda_{\min}\bigg(\sum_{i\in(S^*)^c} w_{1i}\boldsymbol{a}_i\boldsymbol{a}_i^\top\bigg) - \lambda_{\min}\bigg(\sum_{i\in(S^*)^c} w_{2i}\boldsymbol{a}_i\boldsymbol{a}_i^\top\bigg) &\le \sum_{i\in(S^*)^c} w_{1i}(\boldsymbol{a}_i^\top\boldsymbol{v}_0)^2 - \sum_{i\in(S^*)^c} w_{2i}(\boldsymbol{a}_i^\top\boldsymbol{v}_0)^2 \\
&\le \Big(\max_{i\in(S^*)^c}(\boldsymbol{a}_i^\top\boldsymbol{v}_0)^2\Big)\sum_{i\in(S^*)^c}\big|w_{i1} - w_{i2}\big| \\
&\le 2\cdot(2-p)\cdot(m-k)\cdot\frac{\tau M^3}{\big(\epsilon^{(t)}\big)^{3-p}}
\end{aligned}$$

and we finished the proof. □

H. Proof of Proposition 1

Similarly to (41), for the least-squares estimator $\boldsymbol{x}^\dagger$ we have

$$\|\boldsymbol{x}^\dagger - \boldsymbol{x}^*\|_2 \le \frac{\|\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y})\|_2}{\lambda_{\min}\big(\boldsymbol{A}^\top\boldsymbol{A}\big)} \le \frac{\|\boldsymbol{A}\|_2 \cdot \|(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y})\|_2}{\lambda_{\min}\big(\boldsymbol{A}^\top\boldsymbol{A}\big)}.$$

Invoke Lemma 11 with $\delta_1 = \delta_2 = 0.01$ to bound $\|\boldsymbol{A}\|_2$ and $\lambda_{\min}\big(\boldsymbol{A}^\top\boldsymbol{A}\big)$, then use a union bound, and we get that the desired bound (15) holds with probability at least $1 - \exp(-\Omega(k)) - \exp(-\Omega(m))$. The proof is complete.

I. Auxiliary and Known Results in High Dimensional Statistics

Lemma 10 (Lemma 1 of [43]). *Assume $\boldsymbol{a} \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries, then for any $\delta_1 > 0$ and $\delta_2 > 0$, it holds that*

$$\mathbb{P}\Big(\|\boldsymbol{a}\|_2^2 > n + 2\sqrt{n\delta_1} + 2\delta_1\Big) \le \exp(-\delta_1)$$
$$\mathbb{P}\Big(\|\boldsymbol{a}\|_2^2 < n - 2\sqrt{n\delta_2}\Big) \le \exp(-\delta_2).$$

In particular, set $\delta_1 = n$ and $\delta_2 = n/16$, and we get

$$\mathbb{P}\Big(\|\boldsymbol{a}\|_2^2 > 5n\Big) \le \exp(-n),$$
$$\mathbb{P}\Big(\|\boldsymbol{a}\|_2^2 < n/2\Big) \le \exp(-n/16).$$

Finally, if $\boldsymbol{a}_i \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries for every $i = 1, \dots, s$, then

$$\mathbb{P}\Big(\max_{i=1,\dots,s} \|\boldsymbol{a}_i\|_2^2 > 5n\Big) \le s \cdot \exp(-n).$$

Lemma 11 (Theorem 6.1 and Example 6.2 of [74]). *Suppose $\boldsymbol{a}_i \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries for each $i = 1, \dots, m$. Then, for any $\delta_1 > 0$, it holds with probability at most $\exp(-k\delta_1^2/2)$ that*

$$\lambda_{\max}\bigg(\sum_{i=1}^k \boldsymbol{a}_i\boldsymbol{a}_i^\top\bigg) \ge \big((1+\delta_1)\cdot\sqrt{k} + \sqrt{n}\big)^2.$$

Moreover, if $m \ge n$ then, for any $\delta_2 \in (0,1)$, it holds with probability at most $\exp(-m\delta_2^2/2)$ that

$$\lambda_{\min}\bigg(\sum_{i=1}^m \boldsymbol{a}_i\boldsymbol{a}_i^\top\bigg) \le \big((1-\delta_2)\cdot\sqrt{m} - \sqrt{n}\big)^2.$$

Lemma 12 (Bernstein's inequality, *cf*. Theorem 2.8.1 of [72]). *Let $X_1, \dots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $\delta \ge 0$, we have*

$$\mathbb{P}\bigg[\Big|\sum_{i=1}^N X_i\Big| \ge \delta\bigg] \le 2\exp\bigg[-C\min\Big(\frac{\delta^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{\delta}{\max_i \|X_i\|_{\psi_1}}\Big)\bigg].$$

Here C is some absolute constant and $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm.

Lemma 13 (Corollary 4.2.13 of [72]). *For any $\varepsilon > 0$, an ε-net of the unit Euclidean ball of $\mathbb{R}^n$ has at most $(2/\varepsilon + 1)^n$ points. This is also true for the unit Euclidean sphere $\mathbb{S}^{n-1}$.*

Lemma 14 (Exercise 4.4.3 of [72]). *For a symmetric matrix $\boldsymbol{B}$ and an ε-net $\mathcal{N}$ with $\varepsilon \in [0, 1/2)$, we have*

$$\|\boldsymbol{B}\|_2 \le \frac{1}{1-2\varepsilon} \cdot \max_{\boldsymbol{v}\in\mathcal{N}} |\boldsymbol{v}^\top\boldsymbol{B}\boldsymbol{v}|.$$