Geometric Analysis of Non-convex Optimization Landscapes for Robust M-Estimation of Location

Hongyuan Yang
School of Info. Sci. and Tech.
ShanghaiTech University
Shanghai, China
yanghy@shanghaitech.edu.cn

Ziping Zhao
School of Info. Sci. and Tech.
ShanghaiTech University
Shanghai, China
zipingzhao@shanghaitech.edu.cn

Ying Sun
School of Elec. Eng. and Comp. Sci.
Pennsylvania State University
PA, USA
ysun@psu.edu

***Abstract*—In this paper, we study the classical problem of robust M-estimation of a location parameter. The problem is given by minimizing a finite sum of non-convex loss functions. We investigate the geometric structure of the empirical non-convex objective. Under mild assumptions, we find the optimization landscape can be characterized by two kinds of regions with good properties: the strong convexity in the interior of a ball centered at the minimum and the one-point strong convexity in the exterior of a ball centered at the minimum. By leveraging this result, we further establish conditions under which the estimation problem, albeit non-convex in general, has a unique global minimum close to the ground truth. Exploiting the favorable landscape properties, numerical methods like gradient descent can achieve a global convergence to the unique optimum from arbitrary initialization. Our theoretical findings are corroborated through numerical experiments.**

***Index Terms*—Robust location estimation, M-estimator, non-convex optimization, landscape analysis, worst-case noise.**

I. Introduction

The estimation of a location parameter is a fundamental problem in many data analytics related areas [1]. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. samples from distribution $F(\boldsymbol{x}-\boldsymbol{\mu}^\star)$, where $\boldsymbol{x}$ denotes the random variable and $\boldsymbol{\mu}^\star$ is the unknown location parameter. Our target is to estimate $\boldsymbol{\mu}^\star$ from $\{\boldsymbol{x}_i\}_{i=1}^n$. When F is symmetric, the location is the center of symmetry and hence is equivalent to the mean of $\boldsymbol{x}$ if it exists. The estimation of a location (or mean) parameter finds applications in various areas. For example, many dimension reduction methods in machine learning, including linear discriminant analysis [2] and principal component analysis [3], need to estimate the location parameter a prior. In federated learning [4], the server needs to obtain the mean of the gradients uploaded by the clients, and then uses it to update the model. In differential privacy [5], the system requires efficient algorithms to obtain the mean estimate in order to conduct statistical learning and analysis from the shared data.

A common and direct approach to estimate the location is using the sample mean. The sample mean corresponds to the estimator from minimizing the squared error loss and is optimal when the samples are independent and identically drawn from a Gaussian distribution. However, real-world scenarios often deviate from these ideal conditions, with samples drawn from heavy-tailed distributions or corrupted by outliers. As the sample mean is sensitive to samples significantly deviating from the majority, these outlying observations can deteriorate the estimation result. This motivates the need for robust estimation procedures [1], [6]–[9]. To obtain a robust estimator of the location parameter has raised research attention in many areas, like robotics [10], finance [11], signal processing [12], [13], machine learning [14], etc.

Among different robust estimators, the robust M-estimator, a generalization of the maximum likelihood estimator first introduced by Huber [7], is a widely applied one. An M-estimator of the location parameter is generally obtained by minimizing the sum of loss functions applied to the residual. A special choice of the loss is the quadratic function, which leads to the sample mean estimator. However, due to the fast growth of the quadratic loss, samples far away from the majority contribute to the objective value significantly. This drives the minimizer towards outliers, leading to the non-robustness of the sample mean. To fix this issue, many classical loss functions have been proposed, including the Huber function [7], Cauchy function [15], Geman-McClure function [16], Welsch function [17], "Fair" function [18], etc. These functions should be designed to satisfy certain properties [8] (increase slower than the quadratic function). With these different, typically non-convex, loss functions, extensive research has been devoted to exploring the properties of robust M-estimators [19]–[21].

From the computational perspective, though extensively studied, a standing challenge for robust M-estimation of location is the non-convexity of the optimization landscapes. Since global convergence to the optimum of non-convex optimization is in general not guaranteed, it is not always possible to obtain a good location estimator from an iterative algorithm, such as gradient descent and fixed-point iteration, which are commonly adopted in robust M-estimation problems. Despite the non-convexity landscape of the objective function, empirical observations show that these algorithms very often yield favorable outcomes in practice. Thus, investigating the underlying reasons for this phenomenon introduces an intriguing question. In recent years, proving favorable convergence behaviors of numerical algorithms for non-convex problems has gained substantial attention. One line of research develops a two-stage procedure: It first finds an initialization that is

Fig. 1: Classical "robust" functions ℓ (ϕ is a hyper-parameter).

	$\ell(r)$
Huber	$\begin{cases} \frac{r}{2} & \text{for } \sqrt{r} \le \phi \\ \phi\left(\sqrt{r} - \frac{\phi}{2}\right) & \text{for } \sqrt{r} > \phi \end{cases}$
Pseudo-Huber	$\phi\left(\sqrt{1 + r/\phi} - 1\right)$
Cauchy	$\phi \ln\left(1 + \frac{r}{\phi}\right)$
Geman-McClure	$\frac{r}{\phi + r}$
Welsch	$\frac{\phi}{2}\left[1 - \exp(-\frac{r}{\phi})\right]$
Fair	$2\phi\left(\sqrt{\frac{r}{\phi}} - \ln\left(1 + \sqrt{\frac{r}{\phi}}\right)\right)$

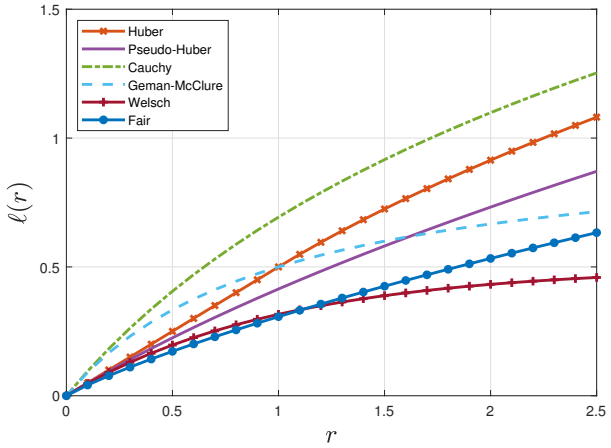


Fig. 2: "Robust" functions ℓ ($\phi = 1$).

close to the ground truth and then refines the estimate to make it closer to ground truth by performing a few local search iterations. Related applications include matrix completion [22], tensor decomposition [23], and phase retrieval [24]. This analysis indicates careful initialization is required for obtaining a good solution. However, very often descent algorithms can converge effectively even without a smart initialization. This motivates another line of research, which focuses on analyzing the properties of the local minima of the non-convex problems. [25]–[27] show that every stationary point of the objective is within a statistically good region around the ground truth. [28] proves that, under mild assumptions, the M-estimation for linear regression has a unique local minimum. These analyses rely on a predetermined region, which is related to the ground truth. The size of this region also determines the initialization and the step size of the numerical algorithm studied therein. Nevertheless, in practice it is hard to decide the radius of such a region since we cannot obtain any information about the ground truth.

In this paper, we investigate the geometric property of the non-convex landscape of the problem of the robust M-estimation of location. Different from the idea in [25]–[28], our analysis does not reply on a predetermined region for geometric analysis, and hence is a universal characterization for the optimization landscape. Besides, instead of the stochastic measurement noise assumption [25]–[28], which aims to analyze the long-term average performance, our analysis is based on the worst-case noise assumption [29], which entails analyzing the instantaneous estimation performance [30] and helps us establish a connection between the geometric property and the noise level. We prove that, under mild restrictions, the estimation problem exhibits two regions characterized by favorable properties: the strong convexity property in the interior of a ball around the minimum and the one-point strong convexity property in the exterior of a ball around the minimum. By leveraging this result, we show that the landscape of robust M-estimation of location, albeit non-convex, has only one unique local minimum when noise is sufficiently small. Exploiting these favorable landscape properties, numerical methods like gradient descent can achieve a global convergence to the unique minimum from arbitrary initialization. Numerical experiments on synthetic data are carried out, which corroborate our theoretical findings.

Detailed proofs for the results in this paper are available at: www.ncvxopt.com/pubs/YangZhaoSun-RobustLocation.pdf.

II. Problem Formulation

Suppose $\{\boldsymbol{x}_i \in \mathbb{R}^p\}_{i=1}^n$ are generated from the model:

$$\boldsymbol{x}_i = \boldsymbol{\mu}^\star + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{\mu}^\star$ is the unknown location parameter, and $\{\boldsymbol{\varepsilon}_i\}_{i=1}^n$ are the measurement noise. Denote the distance between $\boldsymbol{x}_i$ and $\boldsymbol{\mu}$ as $d_i(\boldsymbol{\mu}) = \|\boldsymbol{x}_i - \boldsymbol{\mu}\|$ where $\|\cdot\|$ is the Euclidean norm. The robust M-estimation of location is obtained through minimizing the sum of losses for n samples. Specifically, the estimation problem is written as

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^p} \quad \mathcal{L}(\boldsymbol{\mu}) = \frac{1}{n} \sum_{i=1}^n \ell\left(d_i^2(\boldsymbol{\mu})\right). \tag{2}$$

A class of prototypical loss functions ℓ are listed in Table 1 and Figure 2. Also, we have the following assumption on ℓ.

Assumption 1. *The loss function ℓ defined on $[0, +\infty)$ is increasing and twice differentiable. Besides, it has bounded first and second derivatives; i.e., there exist $L, U > 0$ such that $L \le \min\{\ell', |\ell''|\} \le \max\{\ell', |\ell''|\} \le U$.*

Assumption 1 is satisfied by all robust loss functions in Table 1 except Huber function due to the fact that it is not twice differentiable at one point. The pseudo-Huber function [18], with slight modification to the Huber function, can easily mitigate this issue.

Remark 2. It should be noted that although the concavity of ℓ is usually regarded as a natural suggestion by the robustness [15], the analysis in this paper is independent of this property. That is to say, results derived in this paper are applicable to a broader class of (non-robust) M-estimators of location.

III. Geometric Analysis

We first introduce the following assumption.

Assumption 3 (Worst-case Noise)**.** *For noises $\{\boldsymbol{\varepsilon}_i\}_{i=1}^n$, there exists $\kappa \geq 0$ such that $\|\boldsymbol{\varepsilon}_i\| \leq \kappa$ for $i = 1, \ldots, n$.*

Denote the stationary points of (2) as $\hat{\boldsymbol{\mu}}$, i.e. $\nabla \mathcal{L}(\hat{\boldsymbol{\mu}}) = \mathbf{0}$. Given $\boldsymbol{a} \in \mathbb{R}^p$ and $r \geq 0$, we define $\mathcal{B}(\boldsymbol{a}, r) = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x} - \boldsymbol{a}\| \leq r\}$. Then, we have the following result.

Proposition 4 (Error Bound of Robust M-Estimator)**.** *Suppose Assumption 3 holds. It follows that $\hat{\boldsymbol{\mu}} \in \mathcal{B}(\boldsymbol{\mu}^\star, \kappa)$.*

Proposition 4 states that all stationary points of (2) are within the ball $\mathcal{B}(\boldsymbol{\mu}^\star, \kappa)$, which implies that local search algorithms, like gradient descent, are guaranteed to converge to this region. Since a stationary point can be a local minimum, a local maximum, or a saddle point, with Proposition 4, it is hard to tell which kind of solution the algorithm will eventually converge to. In practice, we are more concerned about local minima, on which the following theorem summarizes an interesting result.

Theorem 5 (Local Strong Convexity and Smoothness [31])**.** *Suppose Assumptions 1 and 3 hold. Given $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$ with*

$$R < \sqrt{\frac{L}{2U}} - 2\kappa, \tag{3}$$

we have

$$\alpha \boldsymbol{I} \preceq \nabla^2 \mathcal{L}(\boldsymbol{\mu}) \preceq \beta \boldsymbol{I}$$

for all $\boldsymbol{\mu}$ in $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$; i.e., $\mathcal{L}$ is α-strongly convex and β-smooth in $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$, where non-negative constants α, β depend on L, U, κ, R.

In Theorem 5, since $\sqrt{\frac{L}{2U}} > 0$ always holds, (3) can always be satisfied when κ and R are sufficiently small. Theorem 5 indicates that while the loss $\ell\left(d_i^2(\boldsymbol{\mu})\right)$ related to the i-th sample is non-convex, the sum of n non-convex losses introduces a favorable strong convexity property. More specifically, under condition (3), $\mathcal{L}$ is strongly convex and smooth around $\hat{\boldsymbol{\mu}}$ and hence all stationary points are local minimum. Moreover, an important insight drawn from Theorem 5 is the dependency of the radius R on the noise level κ. Based on condition 3, the smaller the noise level is, the larger the region of strong convexity will be.

While Theorem 5 ensures algorithms like gradient descent can always convergence to a local minimum, it is hard to say whether other local minima can be better. In view of this, it is more desirable to investigate the existence of unique local minimum. To attain this, we have the following landscape characterization called one-point strong convexity.

Theorem 6 (One-point Strong Convexity [32])**.** *Suppose Assumptions 1 and 3 hold. Given $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)$ with*

$$r \geq 4\kappa,$$

we have

$$\langle \nabla \mathcal{L}(\boldsymbol{\mu}), \boldsymbol{\mu} - \hat{\boldsymbol{\mu}} \rangle \geq \gamma \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 \tag{4}$$

for all $\boldsymbol{\mu}$ in $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$ (the complement of set $\mathcal{B}(\boldsymbol{a}, r)$); i.e., $\mathcal{L}$ is one-point strongly convex in $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$, where non-negative constant γ depend on L, κ.

Note that the region of one-point strong convexity always exists since κ is finite. One-point strong convexity of $\mathcal{L}$ in $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$ guarantees that $\mathcal{L}$ cannot attain any stationary point in this region. This is because if there exists some $\boldsymbol{\mu} \neq \hat{\boldsymbol{\mu}}$ satisfies $\nabla \mathcal{L}(\boldsymbol{\mu}) = \mathbf{0}$ in $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$, then the left hand side of (4) will be zero while the right hand side of (4) is strictly positive, leading to a contradiction. Similar to Theorem 5, the size of $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$ also depends on the noise level κ. The larger the noise is, the smaller the region of the one-point strong convexity is. If $\kappa = 0$, we can always choose $r = 0$, implying the global one-point strong convexity of $\mathcal{L}$.

So far we have shown that under some conditions the landscape of $\mathcal{L}$ can have two benign geometry properties: strong convexity and smoothness on $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$ given by Theorem 5 and one-point strong convexity on $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$ given by Theorem 6. Combining these two property, we can establish the following result so that the two regions overlap, which leads to the uniqueness of the local minimum.

Corollary 7 (Unique local minimum condition)**.** *Suppose Assumptions 1 and 3 hold. There exists a sufficiently small κ satisfying*

$$\kappa < \frac{1}{6}\sqrt{\frac{L}{2U}}$$

so that $\mathcal{L}$ has a unique global minimum.

Since $\sqrt{\frac{L}{2U}}$ is a positive constant independent of κ, there always exists a sufficiently small κ such that Corollary 7 holds. From Theorems 5 and 6, we find as κ decreases, both the region of strong convexity and the region of one-point strong convexity will be enlarged. Corollary 7 is a direct consequence of this observation; when κ is sufficiently small, there is an overlap between $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$ and $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$, namely $R > r$. By strong convexity there can be only one local minimum in $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$ and by one-point strong convexity there is no local minimum in $\mathcal{B}(\hat{\boldsymbol{\mu}}, r)^c$. Then only unique local minimum exists. We conclude when the noise power is sufficiently small, the problem of robust M-estimation of location, albeit non-convex in general, has a unique local minimum close to the ground truth.

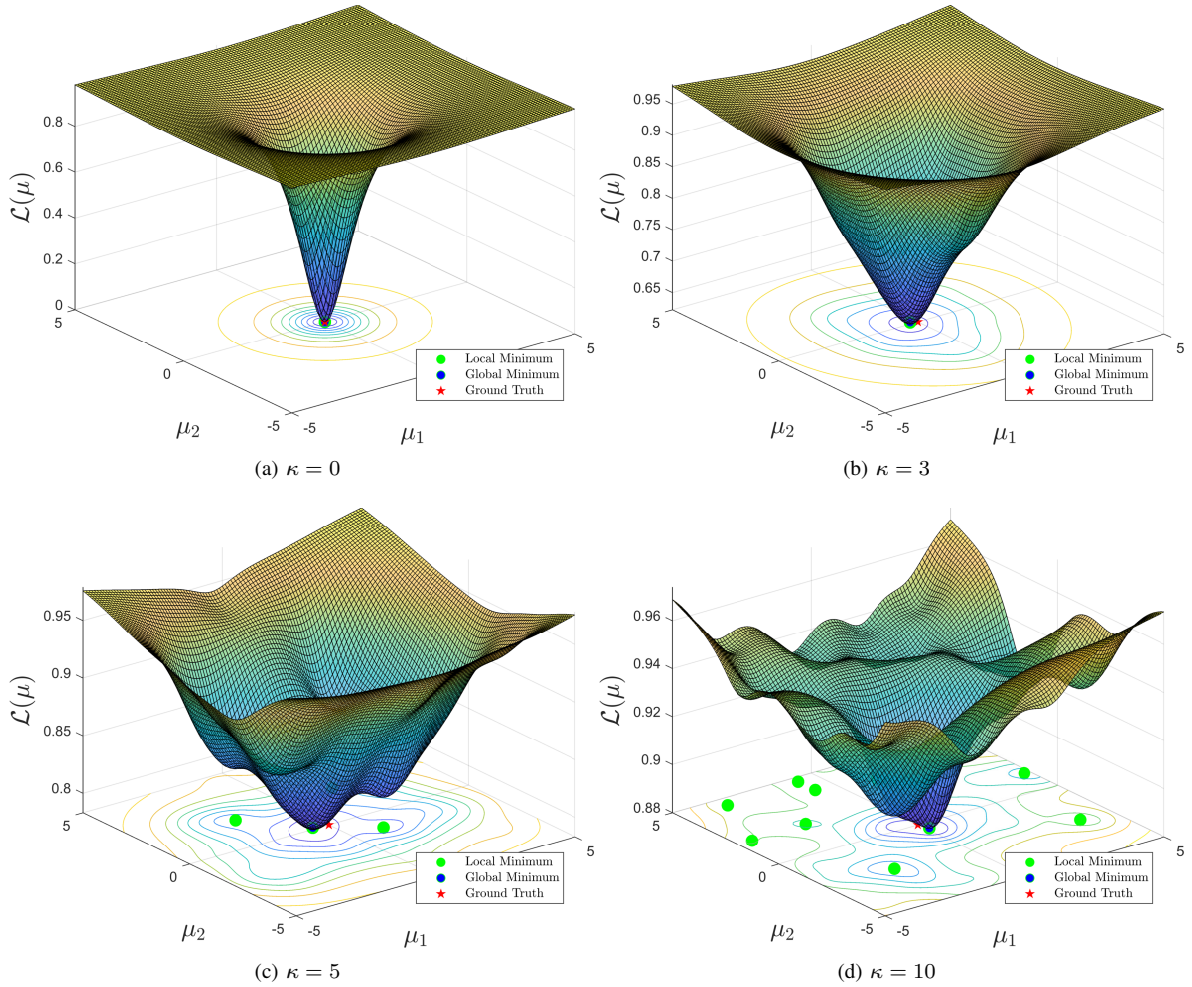
IV. Numerical Experiments

A. Landscape of $\mathcal{L}$

Firstly, we numerically demonstrate the landscape of $\mathcal{L}$ under various noise levels to validate our theoretical findings about the regions with benign properties. The "robust" function is chosen as Geman-McClure function:

$$\ell(r) = \frac{r}{\phi + r}, \tag{5}$$

where we choose $\phi = 1$. The samples $\{\boldsymbol{x}_i \in \mathbb{R}^p\}_{i=1}^n$ are generated through (1), where $n = 50$ and $p = 2$. The ground truth is $\boldsymbol{\mu}^\star = [0, 0]^\top$. The noise $\{\boldsymbol{\varepsilon}_i\}_{i=1}^n$ is generated in the



(a) $\kappa = 0$ (b) $\kappa = 3$

(c) $\kappa = 5$ (d) $\kappa = 10$

Fig. 3: Landscape of $\mathcal{L}$ with different noise level

following way: we first generate sample $\tilde{\boldsymbol{\varepsilon}}_i$ where each element follows a uniform distribution defined on $[-\kappa, \kappa]$, and then we set $\boldsymbol{\varepsilon}_i = \kappa \frac{\tilde{\boldsymbol{\varepsilon}}_i}{\|\tilde{\boldsymbol{\varepsilon}}_i\|}$.

We consider κ at four different levels for the measurement noise: (i) $\kappa = 0$; (ii) $\kappa = 3$; (iii) $\kappa = 5$; and (iv) $\kappa = 10$. The corresponding results are illustrated in Fig. 3a, Fig. 3b, Fig. 3c, and Fig. 3d, respectively. Fig. 3a represents the noiseless scenario.

In Fig. 3a, when $\kappa = 0$, it exhibits the global one-point strong convexity by Theorem 6 so that there is a unique local minimum. And this minimum is aligned with the ground truth, which indicates the exact recovery by Proposition 4.

Fig. 3b represents the small noise scenario. It shows when κ is sufficiently small, $\mathcal{L}$ has a unique local minimum, which is no longer equal to the ground truth but resides in the region $\mathcal{B}(\boldsymbol{\mu}^\star, \kappa)$. This observation verifies the theoretical findings in Corollary 7 and Proposition 4.

In Fig. 3c, as κ increases, the landscape of $\mathcal{L}$ exhibits numerous local minima. When κ can satisfy condition 3, then around all local minima, there is a strong convexity region.

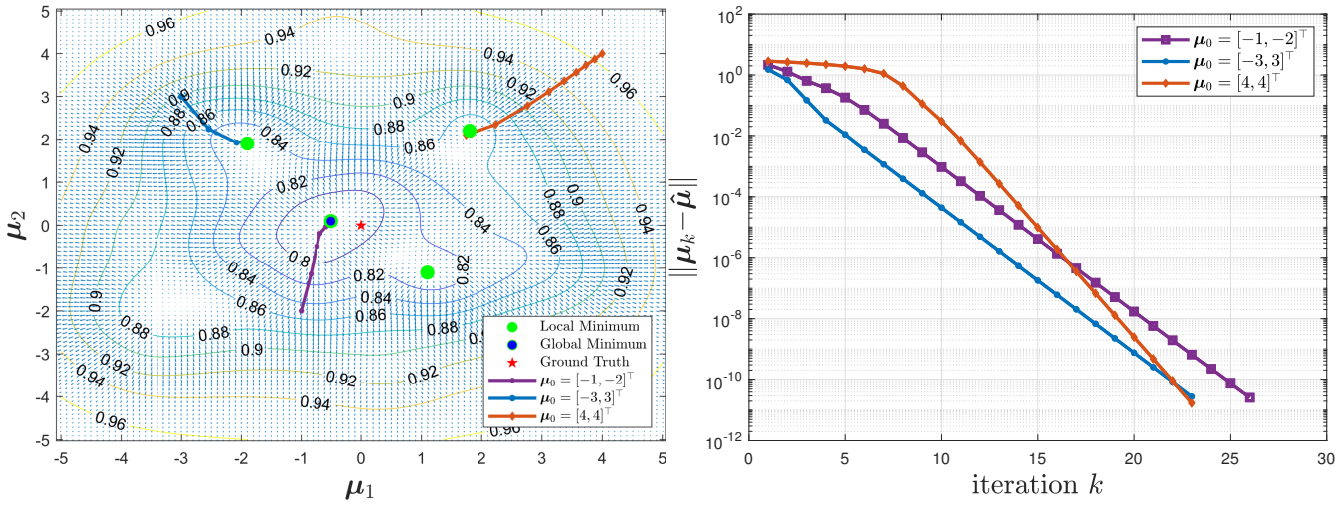
In Fig. 3d, κ is not satisfied condition 3, which leads to the existence of the local maxima and saddle points.

B. Convergence behavior of gradient descent

Next, we show the numerical results about the convergence of gradient descent. We run gradient descent to minimize $\mathcal{L}$, using Geman-McClure function (5) with $\phi = 1$ as the "robust" function. The function $\mathcal{L}$ is the same in Section IV-A with noise level $\kappa = 3$ and $\kappa = 5$. The stopping criterion of gradient descent is set to $\|\nabla \mathcal{L}(\boldsymbol{\mu}_k)\| \leq 10^{-12}$, where k is the current iteration number. The results are illustrated in Fig. 4 and Fig. 5.

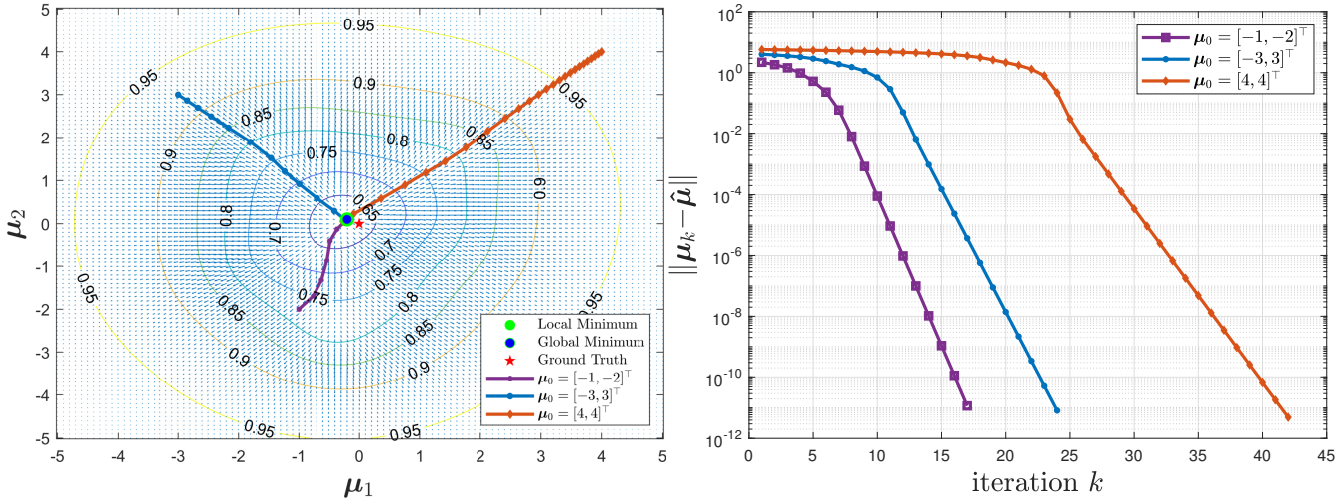
Fig. 4a is aligned with the landscape of Fig. 3c. Since there are many different local minima in this landscape, when we set different initialization, gradient descent can converge to different local minima. Around each local minimum, there exists a strong convexity region, which leads to a linear convergence in Fig. 4b.

Fig. 5a is aligned with the landscape of Fig. 3b, where noise level κ satisfies the condition in Corollary 7, leading to the unique local minimum. When we set the initialization $\boldsymbol{\mu}_0 = [-1, -2]^\top$, which represents a favorable initialization



(a) (b)

Fig. 4: Convergence curve and contour of gradient descent from different initializations μ_0 under noise level $\kappa = 5$.



(a) (b)

Fig. 5: Convergence curve and contour of gradient descent from different initializations μ_0 under noise level $\kappa = 3$.

close to the ground truth, it is obvious that gradient descent obtains a linear convergence rate to $\hat{\mu}$. It suggests the existence of a strongly convex and smooth region in the interior of a ball around $\hat{\mu}$. When we progressively increase the distance between the initialization μ_0 and the ground truth $\mu^\star$, i.e., $\mu_0 = [-3,3]^\top$ and $\mu_0 = [4,4]^\top$, we observe that the number of iterations required for convergence in gradient descent gradually increases. Notably, the convergence curve is piece-wise linear in 5b. This phenomenon corresponds to the existence of two regions, namely the region of one-point strong convexity and the region of strong convexity and smoothness. All three initialization can converge to the nearly same solution $\hat{\mu}$, which corroborates the existence of the unique local minimum.

V. Conclusion

In this paper, we have demonstrated a benign geometric structure of the problem of robust M-estimation of a location parameter. Instead of using a predetermined region, we established an explicit characterization of the overall non-convex landscape of the optimization objective. Under mild assumptions, the problem of robust M-estimation of a location parameter, albeit non-convex, has a unique local minimum close to the ground truth. We also establish the correlation between the benign geometric structure and the noise level. With this results, gradient descent can converge to the unique minimum from arbitrary initialization. Experimental results validate the theoretical findings and corroborate the robustness of the robust M-estimator of location.

REFERENCES

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972.

[2] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[3] I. T. Jolliffe. *Principal Component Analysis for Special Types of Data*. Springer, 2002.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[5] X. Liu, W. Kong, S. Kakade, and S. Oh. Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems*, volume 34, pages 3887–3901, 2021.

[6] J. W. Tukey. A survey of sampling from contaminated distributions. *In: Olkin I (ed) Contributions to probability and statistics*, pages 448–485, 1960.

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

[8] R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

[9] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Hoboken, NJ: Wiley, 2009.

[10] M. Bosse, G. Agamennoni, and I. Gilitschenski. Robust estimation and applications in robotics. *Foundations and Trends® in Robotics*, 4(4):225–269, 2016.

[11] Y. Feng and D. P. Palomar. A signal processing perspective on financial engineering. *Foundations and Trends® in Signal Processing*, 9(1/2):1–231, 2016.

[12] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80, 2012.

[13] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma. *Robust Statistics for Signal Processing*. Cambridge University Press, 2018.

[14] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] R. A. Maronna, D. R. Martin, V. J. Yohai, and M. Salibian-Barrera. *Robust Statistics: Theory and Methods: Theory and Methods*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley, 2006.

[16] S. Geman and D. E. McClure. Bayesian image analysis: An application to single photon emission tomography. *American Statistical Society*, pages 12–18, 1985.

[17] J. E. Dennis Jr. and R. E. Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics - Simulation and Computation*, 7(4):345–359, 1978.

[18] W. J. Rey. Introduction to robust and quasi-robust statistical methods. 1983.

[19] P. J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

[20] D. L. Donoho and P. J. Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, pages 157–184, 1983.

[21] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley, 1986.

[22] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[23] A. Anandkumar, R. Ge, and M. Janzamin. Learning overcomplete latent variable models through tensor methods. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 36–112. PMLR, 2015.

[24] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[25] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[26] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[27] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation under nonconvexity. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2472–2481. PMLR, 2016.

[28] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[29] Angelia Nedić and Dimitri P Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical programming*, 125(1):75–99, 2010.

[30] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu. Robust federated learning with noisy communication. *IEEE Transactions on Communications*, 68(6):3452–3464, 2020.

[31] J.-P. Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1):187–205, 1982.

[32] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Supplementary Materials for "Geometric Analysis of Non-Convex Optimization Landscape for Robust M-Estimation of Location"

Hongyuan Yang, Ziping Zhao, and Ying Sun

APPENDIX A
PROOF OF PROPOSITION 3

Considering the first-order optimality condition of problem (2), we have

$$\begin{aligned}\nabla\mathcal{L}\left(\hat{\boldsymbol{\mu}}\right) &= \frac{1}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)\nabla d_i^2\left(\boldsymbol{\mu}\right)\\ &= \frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)\left(\hat{\boldsymbol{\mu}}-\boldsymbol{x}_i\right)=\mathbf{0},\end{aligned} \tag{6}$$

where $\hat{\boldsymbol{\mu}}$ denotes any stationary point. Then we have

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n}\frac{\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}{\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}\boldsymbol{x}_i. \tag{7}$$

Note that $\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)\neq 0$, since ℓ' is strictly positive by Assumption 1. The right hand side (RHS) of (7) can be interpreted as a weighted average of the sample $\{\boldsymbol{x}_i\}_{i=1}^n$. (For robust functions, ℓ' is a non-increasing function for large values, so outlying samples will receive smaller weights, which brings the robustness.)

Substitute $\boldsymbol{x}_i=\boldsymbol{\mu}^\star+\boldsymbol{\varepsilon}_i$ into (7), and we obtain

$$\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star = \sum_{i=1}^{n}\frac{\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}{\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}\boldsymbol{\varepsilon}_i. \tag{8}$$

Taking Euclidean norm to both sides of (8) leads to

$$\begin{aligned}\left\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\right\| &= \left\|\sum_{i=1}^{n}\frac{\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}{\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}\boldsymbol{\varepsilon}_i\right\|\\ &\leq \frac{\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}{\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}\right)\right)}\left\|\boldsymbol{\varepsilon}_i\right\|\leq\kappa,\end{aligned}$$

where the second inequality follows from Assumption 3.

APPENDIX B
PROOF OF THEOREM 5

Our goal is to prove $\mathcal{L}$ is strongly convex and smooth in the interior of $\mathcal{B}\left(\hat{\boldsymbol{\mu}},R\right)$. We first prove the strong convexity of $\mathcal{L}$. Since ℓ is twice differentiable according to Assumption 1, we can compute the Hessian of $\mathcal{L}$ as follows:

$$\begin{aligned}\nabla^2\mathcal{L}\left(\boldsymbol{\mu}\right) = \frac{4}{n}\sum_{i=1}^{n}\ell''\left(d_i^2\left(\boldsymbol{\mu}\right)\right)\left(\boldsymbol{\mu}-\boldsymbol{x}_i\right)\left(\boldsymbol{\mu}-\boldsymbol{x}_i\right)^\top\\ +\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\boldsymbol{\mu}\right)\right)\boldsymbol{I}.\end{aligned}$$

Then, it is equivalent to prove $\nabla^2\mathcal{L}\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\succ\mathbf{0}$ for all $\boldsymbol{r}$ within some ball.

We have

$$\begin{aligned}&\lambda_{\min}\left(\nabla^2\mathcal{L}\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\\ =&\lambda_{\min}\left(\frac{4}{n}\sum_{i=1}^{n}\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)^\top\right.\\ &\left.+\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\boldsymbol{I}\right).\end{aligned}$$

Using Weyl's inequality, we get

$$\begin{aligned}&\lambda_{\min}\left(\nabla^2\mathcal{L}\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\\ \geq&\lambda_{\min}\left(\frac{4}{n}\sum_{i=1}^{n}\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)^\top\right)\\ &+\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right),\end{aligned} \tag{9}$$

Consider the first term in (9). We have

$$\begin{aligned}&\lambda_{\min}\left(\frac{4}{n}\sum_{i=1}^{n}\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)^\top\right)\\ \geq&\frac{4}{n}\sum_{i=1}^{n}\lambda_{\min}\left(\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)^\top\right),\end{aligned}$$

where the inequality follows from the Weyl's inequality. We further have

$$\begin{aligned}&\frac{4}{n}\sum_{i=1}^{n}\lambda_{\min}\left(\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right)^\top\right)\\ \geq&-\frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\right|\left\|\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\right\|^2\\ \geq&-\frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\right|\left(\|\boldsymbol{r}\|+\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\|+\|\boldsymbol{\varepsilon}_i\|\right)^2\\ \geq&-\frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\right|\left(\|\boldsymbol{r}\|+2\kappa\right)^2,\end{aligned} \tag{10}$$

where the first inequality is due to $\lambda_{\min}\left(\omega\boldsymbol{a}\boldsymbol{a}^\top\right)=-\left|\omega\right|\left\|\boldsymbol{a}\right\|^2$ for any scalar ω, and the last inequality follows from $\|\boldsymbol{\varepsilon}_i\|\leq\kappa$ by Assumption 2 and $\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\|\leq\kappa$ by Proposition 4. Upon substituting (10) into (9), we get

$$\begin{aligned}&\lambda_{\min}\left(\nabla^2\mathcal{L}\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\\ \geq&-\frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\right|\left(\|\boldsymbol{r}\|+2\kappa\right)^2\\ &+\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2\left(\hat{\boldsymbol{\mu}}+\boldsymbol{r}\right)\right)\\ \geq&-4U\left(r+2\kappa\right)^2+2L,\end{aligned} \tag{11}$$

where we define $\|\boldsymbol{r}\|=r$ and the second inequality follows from Assumption 1. The RHS of (11) is positive whenever

$$r<\sqrt{\frac{L}{2U}}-2\kappa. \tag{12}$$

Denote the supremum of the solution of r to (12) as R. (Note that there always exists a solution r satisfying (12). Since $\sqrt{\frac{L}{2U}} > 0$ always holds in (12). Thus, (12) can always be satisfied when κ and r are sufficiently small.) Then we have that in the interior of the ball $\mathcal{B}(\hat{\boldsymbol{\mu}}, R)$, $\mathcal{L}$ α-strongly convex with $\alpha = -4U(r+2\kappa)^2 + 2L > 0$.

Next, we prove the β-smoothness of $\mathcal{L}$. It is equivalent to show $\nabla^2\mathcal{L}(\hat{\boldsymbol{\mu}}+\boldsymbol{r}) \preceq \beta\boldsymbol{I}$, i.e., $\lambda_{\min}\left(\beta\boldsymbol{I} - \nabla^2\mathcal{L}(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right) \geq 0$ for a constant $\beta > 0$. We have

$$\begin{aligned}
&\lambda_{\min}\left(\beta\boldsymbol{I} - \nabla^2\mathcal{L}(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\\
=&\lambda_{\min}\left(\left(\beta - \frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\right)\boldsymbol{I}\right.\\
&\left.-\frac{4}{n}\sum_{i=1}^{n}\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)^\top\right)\\
\geq&-\frac{4}{n}\sum_{i=1}^{n}\lambda_{\min}\left(\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)^\top\right)\\
&+\beta - \frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right), \quad (13)
\end{aligned}$$

where the inequality follows from Weyl's inequality. For the first term in (13), we have

$$\begin{aligned}
&-\frac{4}{n}\sum_{i=1}^{n}\lambda_{\min}\left(\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)(\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i)^\top\right)\\
\geq&-\frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\right|\|\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\|^2, \quad (14)
\end{aligned}$$

where the first inequality is due to $\lambda_{\min}\left(\omega\boldsymbol{a}\boldsymbol{a}^\top\right) = -|\omega|\|\boldsymbol{a}\|^2$ for any scalar ω. Upon substituting (14) into (13), we get

$$\begin{aligned}
&\lambda_{\min}\left(\beta\boldsymbol{I} - \nabla^2\mathcal{L}(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\\
\geq&\beta - \frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\right|\|\hat{\boldsymbol{\mu}}+\boldsymbol{r}-\boldsymbol{x}_i\|^2\\
&-\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\\
\geq&\beta - \frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\right|(\|\boldsymbol{r}\|+\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\|+\|\boldsymbol{\varepsilon}_i\|)^2.\\
&-\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\\
\geq&\beta - \frac{4}{n}\sum_{i=1}^{n}\left|\ell''\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\right|(\|\boldsymbol{r}\|+2\kappa)^2\\
&-\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right),
\end{aligned}$$

where the second inequality follows from triangular inequality and the third inequality follows from Proposition 4 and Assumption 3. By using Assumption 1, we then have

$$\begin{aligned}
&\lambda_{\min}\left(\beta\boldsymbol{I} - \nabla^2\mathcal{L}(\hat{\boldsymbol{\mu}}+\boldsymbol{r})\right)\\
\geq&\beta - 4U(r+2\kappa)^2 - 2\overline{G}(\kappa, r)\\
\geq&\beta - 2L - 2U \quad (15)
\end{aligned}$$

The RHS of (15) is positive whenever

$$\beta \geq 2(L+U).$$

And it is easy to see $\alpha \leq \beta$.

Appendix C
Proof of Theorem 7

The gradient of $\mathcal{L}$ is given by

$$\nabla\mathcal{L}(\boldsymbol{\mu}) = \frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\boldsymbol{\mu}-\boldsymbol{x}_i).$$

Then we have

$$\begin{aligned}
&\nabla\mathcal{L}(\boldsymbol{\mu})^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
=&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\boldsymbol{\mu}-\boldsymbol{x}_i)^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
=&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star-\boldsymbol{\varepsilon}_i)^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
=&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|^2\\
&+\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star-\boldsymbol{\varepsilon}_i)^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}). \quad (16)
\end{aligned}$$

For the second term in (16), by using the Cauchy-Schwarz inequality and the triangle inequality, we have

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star-\boldsymbol{\varepsilon}_i)^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
\geq&-\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star-\boldsymbol{\varepsilon}_i\|\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|\\
\geq&-\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\|+\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|. \quad (17)
\end{aligned}$$

Upon substituting (17) into (16), we have

$$\begin{aligned}
&\nabla\mathcal{L}(\boldsymbol{\mu})^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
\geq&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|-\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}^\star\|-\|\boldsymbol{\varepsilon}_i\|)\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|\\
\geq&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)(\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|-2\kappa)\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|,
\end{aligned}$$

where the last inequality follows from Proposition 4 and Assumption 3.

When $\kappa = 0$, we have

$$\begin{aligned}
\nabla\mathcal{L}(\boldsymbol{\mu})^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}) \geq&\frac{2}{n}\sum_{i=1}^{n}\ell'\left(d_i^2(\boldsymbol{\mu})\right)\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|^2\\
=&\gamma_1\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|^2,
\end{aligned}$$

where $\gamma_1 = \frac{2}{n}\sum_{i=1}^{n} \ell'\left(d_i^2\left(\boldsymbol{\mu}\right)\right) > 0$ is a positive constant.

When $\kappa > 0$, if $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| \geq 4\kappa$, we have $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| - 2\kappa \geq 2\kappa$. Then there exists a positive constant $C_1 = \frac{2\kappa}{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|}$ such that $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| - 2\kappa \geq C_1 \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|$. For the lower bound of C_1, we have

$$C_1 = \frac{2\kappa}{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|} \geq \frac{2\kappa}{\|\boldsymbol{\mu} - \boldsymbol{\mu}^\star\| + \|\boldsymbol{\mu}^\star - \hat{\boldsymbol{\mu}}\|}. \tag{18}$$

Consider term $\|\boldsymbol{\mu} - \boldsymbol{\mu}^\star\|$, we have

$$\begin{aligned} \|\boldsymbol{\mu} - \boldsymbol{\mu}^\star\| &\leq \|\boldsymbol{\mu} - \boldsymbol{x}_i\| + \|\boldsymbol{x}_i - \boldsymbol{\mu}^\star\| \\ &= \|\boldsymbol{\mu} - \boldsymbol{x}_i\| + \|\boldsymbol{\varepsilon}_i\| \\ &\leq \max_{i=1,\ldots,n} \|\boldsymbol{\mu} - \boldsymbol{x}_i\| + \kappa, \end{aligned} \tag{19}$$

where the first inequality follows from the triangle inequality and the second inequality follows from $\|\boldsymbol{\mu}^\star - \hat{\boldsymbol{\mu}}\| \leq \kappa$ by Proposition 4. Upon substituting (19) into (18), we have

$$C_1 \geq \frac{2\kappa}{\max_{i=1,\ldots,n} \|\boldsymbol{\mu} - \boldsymbol{x}_i\| + 2\kappa}.$$

Then we have

$$\begin{aligned} \nabla\mathcal{L}\left(\boldsymbol{\mu}\right)^\top \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\right) &\geq \frac{2C_1}{n}\sum_{i=1}^{n} \ell'\left(d_i^2\left(\boldsymbol{\mu}\right)\right) \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 \\ &\geq \gamma_2 \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2, \end{aligned}$$

where $\gamma_2 = \frac{4\kappa}{\max_{i=1\ldots,n}\|\boldsymbol{\mu} - \boldsymbol{x}_i\| + 2\kappa} L > 0$ when $\kappa > 0$. Then we have a positive constant

$$\gamma = \min\left\{\gamma_1, \gamma_2\right\}$$

such that

$$\nabla\mathcal{L}\left(\boldsymbol{\mu}\right)^\top \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\right) \geq \gamma \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2.$$

Since we have $\gamma > 0$, we have $\mathcal{L}$ is one-point strongly convex in $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, r\right)$, where $r \geq 4\kappa$.

APPENDIX D
PROOF OF COROLLARY 8

From Theorem 5, $\mathcal{L}$ is locally strong convex and smooth in the interior of the ball $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, R\right)$, so there is only one local minimum in $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, R\right)$. From Theorem 6, $\mathcal{L}$ is one-point strong convex in the interior of the ball $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, r\right)^c$, so there is no local minimum in $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, r\right)^c$. If $R > r$, there is an overlap between $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, R\right)$ and $\mathcal{B}\left(\hat{\boldsymbol{\mu}}, r\right)^c$. Then we have $\mathcal{L}$ has a unique local minimum, which is the global minimum.

Furthermore, we demonstrate that $R > r$ can not have no solution. If $R > r$, it indicates that there exists an R satisfying $R < \sqrt{\frac{L}{2U}} - 2\kappa$ and $r \geq 4\kappa$. Combining these two condition, we need $\kappa < \frac{1}{6}\sqrt{\frac{L}{2U}}$, which can be satisfied by a small enough κ since $\sqrt{\frac{L}{2U}} > 0$.