Matrix Sensing with Kernel Optimal Loss: Robustness and Optimization Landscape

Xinyuan Song[1] Ziye Ma[1*]
[1]Department of Computer Science, City University of Hong Kong
ziyema@cityu.edu.hk

In this paper, we study how the choice of loss functions of non-convex optimization problems affects their robustness and optimization landscape, through the study of noisy matrix sensing. In traditional regression tasks, mean squared error (MSE) loss is a common choice, but it can be unreliable for non-Gaussian or heavy-tailed noise. To address this issue, we adopt a robust loss based on nonparametric regression, which uses a kernel-based estimate of the residual density and maximizes the estimated log-likelihood. This robust formulation coincides with the MSE loss under Gaussian errors but remains stable under more general settings. We further examine how this robust loss reshapes the optimization landscape by analyzing the upper-bound of restricted isometry property (RIP) constants for spurious local minima to disappear. Through theoretical and empirical analysis, we show that this new loss excels in handling large noise and remains robust across diverse noise distributions. This work provides initial insights into improving the robustness of machine learning models through simple loss modification, guided by an intuitive and broadly applicable analytical framework.

1. Introduction

Noisy low-rank matrix optimization arises in numerous settings, including matrix sensing [1], matrix completion [2], and robust PCA [3]. In practical applications such as recommender systems [4], motion detection [5, 6], phase synchronization and retrieval [7–9], and power system estimation [10], one often encounters an unknown positive semidefinite matrix $M \in \mathbb{R}^{n\times n}$ of rank at most r. Measurements are collected through a known linear operator $\mathcal{A}\colon \mathbb{R}^{n\times n} \to \mathbb{R}^m$, and are corrupted by an additive noise vector $w \in \mathbb{R}^m$ whose distribution may be unknown. A standard formulation is

$$\min_{M\in\mathbb{R}^{n\times n}} f(M, w) \quad \text{subject to} \quad \operatorname{rank}(M) \le r, \quad M \succeq 0, \tag{1.1}$$

where f denotes a loss function evaluated at (M, w). In recovery problems, the target quantity is M, while the noise can vary widely in scale or distribution.

A widely used choice for f is the mean squared error (MSE),

$$f(M) = \tfrac{1}{2}\|\mathcal{A}(M) - \tilde{b}\|_F^2, \tag{1.2}$$

where $\tilde{b} = \mathcal{A}(M^*) + w$ and $\|\cdot\|_F$ is the Frobenius norm. Although the MSE objective performs well when the noise is Gaussian, its sensitivity to heavy-tailed corruption, outliers, or heterogeneous errors has been well documented [11–14].

Because M is constrained to be low-rank and positive semidefinite, many algorithms adopt the Burer–Monteiro (BM) factorization [15], which writes $M = XX^\top$ for $X \in \mathbb{R}^{n\times r}$. Substituting this representation into (1.1) yields an unconstrained non-convex problem:

$$\min_{X\in\mathbb{R}^{n\times r}} f\big(XX^\top, w\big). \tag{1.3}$$

Theoretical studies [16–19] have characterized the geometry of (1.3), including properties of its local minima and its global recovery guarantees.

*Corresponding author

Third Conference on Parsimony and Learning (CPAL 2026).

To improve robustness in the presence of unknown or irregular noise, we draw inspiration from nonparametric regression. In that setting [20–30], one estimates an unknown function g from samples $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ without assuming a parametric form for the noise. A kernel density estimator $\hat{f}(\cdot)$ is used to construct the log-likelihood objective

$$\hat{g} = \arg\max_{g:\, \mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}\big(Y_i - g(\mathbf{X}_i)\big). \tag{1.4}$$

The loss proposed in [31] is Lipschitz continuous, avoids committing to a specific noise model, and reduces to the classical MSE estimator under Gaussian noise. Crucially, it continues to behave reliably when the corruption is heavy-tailed or heterogeneous.

In this work, we adapt the robust loss (1.4) to the BM formulation (1.3) by applying the kernel estimator to the residuals

$$\mathcal{A}(XX^\top) - \tilde{b}, \tag{1.5}$$

which play the same role as the terms $Y_i - g(\mathbf{X}_i)$ in nonparametric regression. In this correspondence, the unknown matrix $M = XX^\top$ replaces the unknown function g, and remains the central object of estimation throughout. We analyze the resulting optimization landscape, with particular attention to how the restricted isometry property (RIP) constants influence global recovery under noise. Both favorable and unfavorable RIP regimes are considered, and convergence properties for local search methods are derived. Numerical experiments under various corruption models demonstrate that this robust loss yields more stable estimation and improved convergence.

Our main contributions are as follows. (1) We provide theoretical recovery guarantees for M^* and establish convergence behavior of the kernel-based loss (1.4) in the BM setting. The results indicate enhanced robustness under heavy-tailed and heterogeneous noise. (2) We show that when the noise is Gaussian, the robust loss coincides with the MSE objective, thereby retaining its statistical efficiency. (3) Through empirical studies, we demonstrate accurate recovery and stable convergence across diverse corruption scenarios.

2. Preliminaries

2.1. The Kernel Loss

A second choice for the data fidelity term is a kernel-based loss derived from a log-likelihood formulation. Given observations $(X_i, Y_i)_{i=1}^n$, define:

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \widehat{R}_n(g) = \arg\min_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} \left(-\log \frac{1}{n} \sum_{j=1}^{n} K_h\big(Y_j - g(X_j), Y_i - g(X_i)\big) \right). \tag{2.1}$$

This loss function is based on the log-likelihood formulation of a kernel density estimator using a kernel function K_h, applied to the residuals. Unlike the mean squared error (MSE) loss, which lacks adaptability to varying noise distributions, the proposed loss function provides robustness by accommodating different noise settings. This characteristic ensures that the estimator remains statistically reliable in the large-sample regime. The proposed $\hat{\mathcal{R}}_n(g)$ involves a tuning parameter h. For implementation we use the exponential kernel

$$K_h(u, v) = \exp\big(-(u - v)^2 / h^2\big). \tag{2.2}$$

Then we use $\mathcal{A}(\cdot)$ to replace $g(\cdot)$ and BM factorization form to get the explicit form $\hat{g}$:

$$\hat{g} = n^{-1} \sum_{i=1}^{n} \left(-\log \frac{1}{n} \sum_{j=1}^{n} \exp\left(-\frac{((Y_j - (\mathcal{A}(XX^\top)_j)) - (Y_i - \mathcal{A}(XX^\top)_i))^2}{h^2} \right) \right). \tag{2.3}$$

Using this loss, problem (1.1) becomes

$$\min_{M \in \mathbb{R}^{n \times n}} \hat{g}(M, w) \quad \text{s.t.} \quad \operatorname{rank}(M) \le r, M \succeq 0 \tag{2.4}$$

or, in factorized form,

$$\min_{X \in \mathbb{R}^{n \times r}} \hat{g}(XX^\top, w), \tag{2.5}$$

where $\hat{g}$ is defined in (2.1).

2.2. RIP Conditions

This line of work usually assumes the Restricted Isometry Property (RIP) for the problem, which is defined below:

$$(1-\delta_p)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_F^2 \leq (1+\delta_p)\|X\|_F^2, \tag{2.6}$$

here the $\|\cdot\|_F$ denotes the Frobenius norm, $\mathcal{A}$ is the linear operator, and δ_p is the RIP-constant, which is usually simplified as δ in the following content.

2.3. Lemmas on Noise

Assumption 2.1. The noise w has a finite influence on the gradient and Hessian of the objective function in the sense that there exist two constants $\zeta_1 \geq 0$ and $\zeta_2 \geq 0$ such that

$$\left|\langle \nabla_M f(M,w) - \nabla_M f(M,0), K\rangle\right| \leq \zeta_1 \|w\|_2 \|K\|_F, \tag{2.7}$$

$$\left|\left[\nabla_M^2 f(M,w) - \nabla_M^2 f(M,0)\right](K,L)\right| \leq \zeta_2 \|w\|_2 \|K\|_F \|L\|_F, \tag{2.8}$$

for all matrices $M, K, L \in \mathbb{R}^{n\times n}$ with $\text{rank}(M), \text{rank}(K), \text{rank}(L) \leq 2r$.
Assumption 2.1 is the bounded noise assumption, for instance: $\|\langle \nabla_X \hat{g}(X+w) - \nabla_X \hat{g}(X), K\rangle\|$ or possibly $\|\langle \nabla_X^2 \hat{g}\big(X+w\big) - \nabla_X^2 \hat{g}\big(X\big), (K,L)\rangle\|$ is bounded with $\|X\|, \|K\|, \|L\|,$ and$\|w\|$. In addition to Assumption 2.1, there exist other forms of noise assumptions used in the subsequent proof. The following two lemmas illustrate how alternative noise assumptions can be incorporated under fixed constants ρ, λ_1, and λ_2.

Lemma 2.2. *There exists a constant ρ such that the gradient of the function (2.3) $\hat{g}(\cdot, w)$ with respect to the first argument M is $\rho-$restricted Lipschitz continuous, meaning that:*

$$\|\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M',w)\|_F \leq \rho \|M - M'\|_F \tag{2.9}$$

for all matrices $M, M' \in \mathbb{R}^{n\times n}$ with $\text{rank}(M) \leq r$ and $\text{rank}(M') \leq r$.
The proof is provided in Section C. Lemma 2.2 establishes that the gradient of the function $\hat{g}(M,w)$ with respect to the variable M satisfies a restricted Lipschitz continuity property over the set of rank-r matrices. Specifically, there exists a constant $\rho > 0$ such that, for any pair of matrices $M, M' \in \mathbb{R}^{n\times n}$ with $\text{rank}(M) \leq r$ and $\text{rank}(M') \leq r$, the Frobenius norm of the gradient difference is bounded above by ρ times the Frobenius norm of the matrix difference. This condition ensures that, within the low-rank manifold, the gradient of $\hat{g}$ does not change too rapidly, which is critical for the analysis of optimization algorithms constrained to low-rank structures.

Lemma 2.3. *Gradient Lipschitz Continuity with Respect to Noise:*

$$\|\nabla_M \hat{g}(M,w_1) - \nabla_M \hat{g}(M,w_2)\|_F \leq \lambda_1 \|w_1 - w_2\|_2. \tag{2.10}$$

Hessian Lipschitz Continuity with Respect to Noise:

$$\|\nabla_M^2 \hat{g}(M,w_1) - \nabla_M^2 \hat{g}(M,w_2)\|_F \leq \lambda_2 \|w_1 - w_2\|_2. \tag{2.11}$$

The proof is provided in Section D. Based on the two lemmas we know that Assumption 2.1 must hold in the kernel loss.

Lemma 2.3 formalizes the regularity of the function $\hat{g}(M,w)$ with respect to the noise variable w. The first inequality establishes that the gradient of $\hat{g}$ with respect to M is Lipschitz continuous in w, with Lipschitz constant λ_1. This implies that small changes in the noise vector induce proportionally small changes in the gradient. The second inequality states that the Hessian of $\hat{g}$ with respect to M is also Lipschitz continuous in w, with Lipschitz constant λ_2. Together, these conditions ensure that both first- and second-order derivatives of $\hat{g}$ vary in a controlled manner as a function of the noise, which is essential for stability and convergence guarantees in noise-sensitive optimization problems.

Due to space limitations, additional preliminaries and the background are provided in Appendix B.

3. Comparison and Relationship of MSE Loss and the kernel loss

Assumption 3.1. Assume the noise w follows a centered symmetric distribution.

Based on Assumption 3.1 and heavy tailed assumptions and analysis in Appendix B, we have:

Theorem 3.2 (Noise sensitivity of $f(M,w)$ and $\hat{g}(M,w)$)**.** *Let $w \in \mathbb{R}^n$ be the noise vector and assume $\|w\|_2 \le \epsilon$ with probability at least $\mathbb{P}(\|w\| \le \epsilon)$ for some positive constant ϵ. We measure noise sensitivity by the Euclidean norm of the gradient with respect to w, that is, $\|\nabla_w L(M,w)\|_2$. Then the sensitivities of the MSE loss $f(M,w)$ and the robust kernel loss $\hat{g}(M,w)$ are bounded as follows:*

- **(A)** **(*Case: MSE loss* $f(M,w)$)** *For the MSE loss, the gradient norm satisfies*
$$\left\|\nabla_w f(M,w)\right\|_2 = \mathcal{O}\left(\frac{\epsilon}{n}\right). \tag{3.1}$$
- **(B)** **(*Case: robust kernel loss* $\hat{g}(M,w)$ *in Equation* (2.3))** *For the robust kernel loss, there exists a constant $C > 0$ (independent of n and ϵ) such that*
$$\left\|\nabla_w \hat{g}(M,w)\right\|_2 = \mathcal{O}\left(\frac{\epsilon\, e^{-\epsilon^2/h^2}}{nh^2}\right), \tag{3.2}$$
for all noise vectors w with $\|w\|_2 \le \epsilon$.

The detailed background together with the proof is provided in Section F.

Theorem 3.2 compares the sensitivity of two loss functions-namely, the standard mean squared error (MSE) loss $f(M,w)$ and a kernel-based robust loss $\hat{g}(M,w)$—with respect to the noise variable w. In case (A), for the MSE loss, the gradient with respect to w grows linearly with ϵ, leading to a derivative of order $\mathcal{O}(\epsilon/n)$, where n is the number of samples. This indicates that large noise directly amplifies the gradient, potentially causing instability in optimization.

In contrast, case (B) demonstrates that the derivative of the robust loss $\hat{g}(M,w)$ with respect to w is exponentially suppressed for large $\|w\|$, scaling as $\mathcal{O}(\epsilon e^{-\epsilon^2/h^2}/(nh^2))$. The exponential decay introduces a natural attenuation of the influence of large noise components, making the loss function more robust to outliers. This distinction underscores the regularizing effect of the proposed loss and its improved stability in high-noise regimes.

Due to space limitations, additional preliminaries and background are provided in Appendix G.

4. Main Result

We now officially characterize the recovery guarantees of ground truth M^* under our new loss setting:

Theorem 4.1 (Behavior of $\hat{g}(M,w)$)**.** *Let $\hat{g}(\cdot,w)$ be our twice-differentiable loss function (Equation (2.3)), M^* is the ground truth, and suppose $\nabla^2_M \hat{g}(M^*,w)$ is positive definite at the ground truth point M^*. Suppose δ satisfies (5.1) in Section 5 for the guarantee of no local minima, with probability at least $\mathbb{P}(\|w\| \le \epsilon)$, if M is a local minimizer of $\hat{g}$, then*
$$\|M - M^*\|_F \le \mathcal{O}\left(\max\left\{1, \frac{\epsilon e^{-\epsilon^2}}{h^2}\right\}\right). \tag{4.1}$$

Meanwhile, for equation:

$$\|\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0)\| \le \lambda\epsilon \quad \textit{with:} \quad \lambda = \mathcal{O}\Big(\frac{8(1+\delta)}{h^4}\epsilon e^{-\epsilon^2} + C\Big). \tag{4.2}$$

The proof is provided in Section I. Theorem 4.1 provides a probabilistic characterization of the behavior of the robust loss function $\hat{g}(M,w)$ in the vicinity of local minimizers. When the minimizer M lies close to M^*, the estimation error $\|M - M^*\|_F$ is shown to be bounded above by $\mathcal{O}\left(\max\left\{1, \frac{\epsilon e^{-\epsilon^2}}{h^2}\right\}\right)$ with high probability, indicating that the error decays exponentially in the noise magnitude ϵ. Moreover, the gradient variation with respect to noise is controlled via a Lipschitz constant λ, which itself satisfies an exponential decay bound given by $\lambda = \mathcal{O}\left(\frac{8(1+\delta)}{h^4}\epsilon e^{-\epsilon^2} + C\right)$. These results together highlight the smoothing effect of the exponential kernel in $\hat{g}$, which yields improved stability and robustness to noise in regions close to the ground truth.

4.1. Turning Point of the Upper Bound

Now we want to calculate the turning point of the Theorem 4.1 to illustrate the specific optimization landscape of $\|M - M^*\|_F$.

Lemma 4.2 (Comparison of Terms in Theorem 4.1)**.** *Let $T_1 = \mathcal{O}(1)$ be the constant term in Equation (4.1) and $T_2 = \mathcal{O}\left(\frac{\|w\|e^{-w^2}}{h^2}\right)$ be a noise-dependent term in Equation (4.1). Then with probability at least $\mathbb{P}(\|w\| \le \epsilon)$: for very small ϵ, specifically when $\epsilon \ll h^2$, we have $T_2 = \mathcal{O}\left(\frac{\epsilon e^{-\epsilon^2}}{h^2}\right) = \mathcal{O}\left(\frac{\epsilon}{h^2}\right) \ll \mathcal{O}(1) = T_1$, Conversely, for small ϵ, this roughly corresponds to $\epsilon \sim h^2$. In regimes where h^2 is significantly smaller than the maximum value of $\epsilon e^{-\epsilon^2}$ (which occurs at $\epsilon = \frac{1}{\sqrt{2}}$ with value $\frac{1}{\sqrt{2}}e^{-1/2}$), $T_2 \gg T_1$. When ϵ and h satisfy $\epsilon e^{-\epsilon^2} \sim h^2$, then $T_2 \sim T_1$.*

The proof is provided in Section I.10. Lemma 4.2 analyzes the interplay between two key terms in the upper bound of the estimation error: a constant term $T_1 = \mathcal{O}(1)$ and a noise-dependent term $T_2 = \mathcal{O}\left(\frac{\epsilon e^{-\epsilon^2}}{h^2}\right)$. As illustrated in Figure 1, for very small noise magnitude $\epsilon \ll h^2$, the exponential term $e^{-\epsilon^2} \sim 1$, so $T_2 \sim \frac{\epsilon}{h^2} \ll 1$, making the constant term T_1 dominant. In this case, the estimation error remains effectively bounded and insensitive to small perturbations in ϵ. Conversely, as ϵ increases and approaches the regime where $\epsilon e^{-\epsilon^2} \sim h^2$, the two terms become comparable, marking a critical threshold where noise starts to meaningfully affect the bound. Moreover, when h^2 becomes much smaller than the peak value of the function $\epsilon e^{-\epsilon^2}$ (attained at $\epsilon = 1/\sqrt{2}$), the term T_2 may become larger than T_1, and the error becomes dominated by the noise effect.

4.2. MSE Loss Result Comparison

Based on the above Lemma we have:

Theorem 4.3 (Behavior of MSE loss function $f(M, w)$)**.** *Let $f(\cdot, w)$ vanilla MSE loss function. M^* is the ground truth. with probability at least $\mathbb{P}(\|w\| \le \epsilon)$:*

If M is a local minimizer of $\hat{g}$, then

$$\|M - M^*\|_F \le \mathcal{O}(\epsilon). \tag{4.3}$$

Meanwhile, for equation:

$$\|\nabla_M f(M, w) - \nabla_M f(M, 0)\| \le \lambda\epsilon \quad \textit{with:} \quad \lambda = \mathcal{O}\left(2\sqrt{1 + \delta_p}\right). \tag{4.4}$$

The proofs are provided in Section I.6.1. Theorem 4.3 describes the behavior of the standard MSE loss $f(M, w)$ near the ground truth M^*, which is a tighter bound than in [32]. Unlike the robust loss, the MSE does not benefit from exponential decay in noise and is more sensitive to outliers. Also, in Theorem 4.1, the new loss $\hat{g}(M, w)$ is smoother (it has a smaller Lipschitz constant λ), and consequently M transitions toward M^* when ϵ decreases. In contrast, in Theorem 4.3, as ϵ grows, $\|M - M^*\|_F$ can grow linearly in ϵ. Here the new loss exhibits a rougher landscape (with a larger Lipschitz constant ρ), and hence when ϵ decreases, M may actually move away from M^*. The loss function still becomes smoother as ϵ changes, but the local minimum remains distant from the ground truth.

5. The Condition of δ

Lemma 5.1 (δ condition with explicit choice of bandwidth Parameter h)**.** *Assume Assumption 2.1 and 2.3 hold and use the definitions in Lemma J.1 in Appendix J. Let the bandwidth be chosen as $h = \frac{\sqrt{2}B}{\sqrt{G_{\min}}}$. Then, the quantity δ must satisfy*

$$\delta \le \sqrt{\frac{B^2}{4(G_{\min} + 2)}\left(2 - \frac{G}{\sigma_r} - \frac{L_2}{B}\right)} - 1. \tag{5.1}$$

In particular, under the conservative (worst-case) assumption, this reduces to

$$\delta \leq \frac{B}{\sqrt{2(G_{\min}+2)}} - 1 \sim \frac{1}{3}. \tag{5.2}$$

The proof is provided in Section J. Lemma 5.1 provides an explicit upper bound on the parameter δ in terms of the kernel bandwidth h, by setting $h = \frac{\sqrt{2}B}{\sqrt{G_{\min}}}$, and shows that δ depends on a combination of spectral and residual quantities; in particular, this value can be calculated explicitly as around $1/3$, which aligns well with Ma's δ bound [32]. The bound simplifies to $\delta \leq \frac{B}{\sqrt{2(G_{\min}+2)}} - 1$, illustrating how the choice of h directly regulates estimation stability through the tradeoff between noise, kernel concentration, and curvature. As shown in Figure 2, the value of δ decreases monotonically as w increases when h is held constant. For fixed w, increasing h suggests a non-monotonic dependence on the smoothing parameter.

The δ bound shown in Equation (5.1) corresponds directly to the theoretical setting (as in Theorem 4.1) in which the convergence guarantees for ground truth recovery. Specifically, these results are valid only when δ remains below the threshold, which aligns with the classical requirement $\delta < 1/2$, as stated in [32]. When δ exceeds this threshold, matrix recovery is no longer guaranteed to succeed in recovering the ground truth. Therefore, in the following sections, we shift our focus to analyzing the behavior of the recovery landscape in the regime where δ is above $1/2$. We only consider the optimization landscape in a region around the ground truth and show that local minimizers are all very close to M^*.

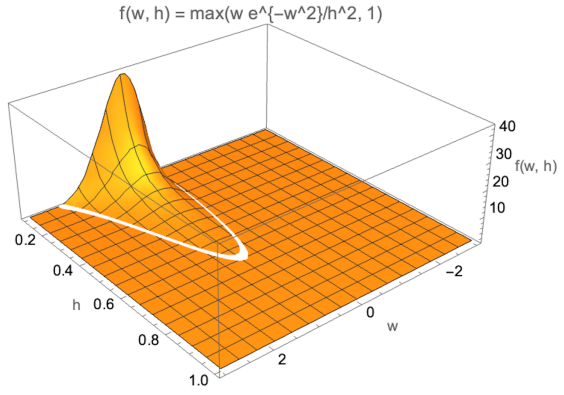

Figure 1: The kernel loss bound result: the yellow mesh depicts the three-dimensional surface of the upper bound on $\|M - M^*\|_F$, expressed as a function of the noise magnitude $\|w\|$ and the kernel bandwidth h.

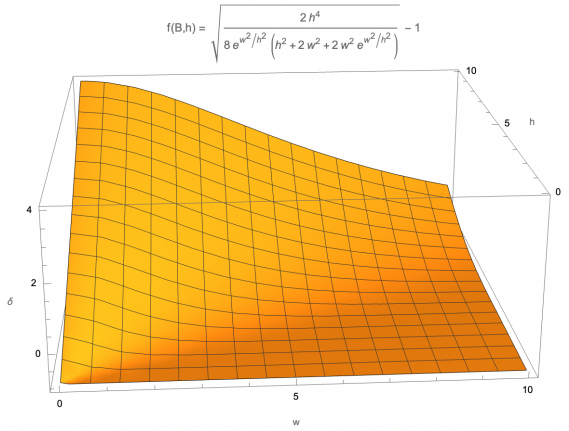

Figure 2: Our δ bound result is illustrated in the figure, where the yellow mesh represents the boundary surface of δ as a function of the noise magnitude $\|w\|$ and the kernel bandwidth h.

6. Upper Bound When $\delta > 1/2$

6.1. The Kernel Loss Function

Theorem 6.1 (Upper Bound for $\delta > 1/2$)**.** *Let $\hat{M} \in \mathbb{R}^{n\times n}$ be an estimator of the true matrix M^*, and with the new loss function (Equation (2.3)), assume that: the derivative terms $\nabla_M u_{ij}(M)$, $\nabla^2_M u_{ij}(M)$ are bounded in norm by L_1, L_2, respectively, and the quantity $G_i(M)$ is uniformly lower bounded: $G_i(M) \geq \Gamma_{\min} > 0$. Also assume the minimum pairwise squared residual satisfies $u^2_{\min} := \min_{i,j,t} u_{ij}(M_t)^2$. Then with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$ the estimation error $\|\hat{M} - M^*\|_F$ satisfies the bound:*

$$\|\hat{M} - M^*\|_F \leq \frac{-\left[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\right] + \sqrt{\left[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\right]^2 + 8B\epsilon^2\zeta_1\zeta_2(1-\delta)}}{4\zeta_2}, \tag{6.1}$$

where

$$B = \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|}\left[\frac{L_1^2\delta^2}{h^4\Gamma^2_{\min}}\exp\left(-\frac{2u^2_{\min}}{h^2}\right) + \frac{L_2\delta}{h^2\Gamma_{\min}}\exp\left(-\frac{u^2_{\min}}{h^2}\right)\right]. \tag{6.2}$$

and $Q := \hat{X}U^\top + U\hat{X}^\top$. Here, ζ_1, ζ_2 are structural noise constants as in Assumptions 2.3 and 2.1, and $\lambda_{r^*}(\hat{X}\hat{X}^\top)$ denotes the r^*-th eigenvalue of the data covariance matrix.

The proofs and detailed math are provided in Section K.

Corollary 6.2. *The $\|\hat{M} - M^*\|_F$ in Theorem 6.1 can be roughly written as:*

$$\|\hat{M} - M^*\|_F \leq \mathcal{O}\Big(-\left[1 - B\epsilon^2\right] + \sqrt{\left[1 - B\epsilon^2\right]^2 + B\epsilon^2}\Big). \tag{6.3}$$

It is a simplified result and the proof is omitted here due to page limit.

6.2. MSE Loss Function

Lemma 6.3 (Estimation Bound under MSE Loss)**.** *Assume that the objective function is given by the mean squared error (MSE) loss and that it satisfies Assumptions 2.3 and 2.1. Further suppose that the noise-free mapping $f(M, 0)$ satisfies the δ-restricted isometry property (RIP) for some constant $\delta \in (0, 1)$. Let $\tau \in (0, 1 - \delta^2)$ be arbitrary. We have, with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, the refined upper bound*

$$\|M - M^*\|_F \leq \mathcal{O}\left(\frac{\epsilon(1+\epsilon)}{\sqrt{1-\epsilon}}\right). \tag{6.4}$$

The proofs and detailed math are provided in Section P, which alignes well with Ma's result [32]. The upper bounds for the proposed loss in Theorem 6.1 and MSE loss in Theorem 6.3 illustrate a direct generalization of the results in [33], Theorem 6.1 shows that even when $\delta > 1/2$, a recovery guarantee still holds under more strict conditions on the model and the noise structure. Hence, unlike certain MSE-based approaches that may fail to provide meaningful error bounds once δ exceeds the $1/2$ threshold, the new kernel-based method retains theoretical validity in this regime. For scenarios in which δ is only marginally above $1/2$, MSE might still be competitive if its assumptions are not severely violated, but it generally does not offer the same level of robustness provided by the new loss.

7. Lower Bound for the $\|M - M^*\|_F$

Now we want to provide the lower bound for the result of the kernel loss and MSE loss for comparison. The lower bound refers to a threshold below which the Frobenius norm of the error, $\|M - M^*\|_F$, cannot decrease unless the estimated matrix M coincides with the ground truth M^*. Thus, if $M \neq M^*$, then $\|M - M^*\|_F$ must lie above a positive value

7.1. MSE Loss Function

Theorem 7.1 (Lower Bound under MSE Loss)**.** *Let $M^* \in \mathbb{R}^{n\times n}$ be the ground truth matrix. $L > 0$, $\delta \in (0, 1)$, and w is the noise term. Provided that $L > 2(1 + \delta)$, then with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$ either $M = M^*$ (i.e., perfect recovery), or the Frobenius norm of the error is lower bounded by*

$$\|M - M^*\|_F \geq \frac{4\sqrt{1+\delta}\epsilon}{L - 2(1+\delta)}. \tag{7.1}$$

The proofs and mathematics are provided in M.4.

7.2. The Kernel Loss Function

Theorem 7.2 (Lower Bound for the kernel loss)**.** *Let M^* be the ground truth matrix. Then, with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$ either $M = M^*$ (i.e., exact recovery), or the Frobenius norm of the error satisfies the lower bound:*

$$\|M - M^*\|_F \geq \frac{2L(1+\delta)e^{-\epsilon^2}}{(1-\delta)h^2}, \tag{7.2}$$

here $L > 0$, and $0 < \delta < 1$.

The proofs and mathematics are provided in M.5. Comparing Theorems 7.1 and 7.2, we observe that the kernel loss yields a tighter and more informative lower bound for the recovery error. Specifically, the bound in Theorem 7.2 scales favorably with respect to both the noise level ϵ and the kernel bandwidth h, and decays exponentially in ϵ^2, which reflects stronger robustness to small noise. In contrast, the lower bound under the MSE loss in Theorem 7.1 grows linearly with ϵ and inversely with the gap $L - 2(1+\delta)$, which can be loose in high-noise or near-threshold regimes. Therefore, the proposed kernel-based formulation not only enhances recovery in practice but also enjoys a more favorable theoretical guarantee in terms of error lower bounds.

However, a closer look at Theorem 7.2 also shows that when ϵ is very small, the exponential term $e^{-\epsilon^2}$ does not significantly decrease, and the resulting lower bound can remain comparatively large. In such low-noise settings, this may hinder the solver's ability to achieve a close approximation to M^*, making the MSE approach more favorable in that specific scenario. This explains why we combine the two losses.

8. Combined Loss

Table 1: Comparison of theoretical properties the between the kernel loss and MSE loss, with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$

Property	**The Kernel Loss**	**MSE Loss**
Loss Behavior (Section 3)	$\mathcal{O}\left(\frac{\epsilon e^{-\frac{\epsilon^2}{h^2}}}{nh^2}\right)$	$\mathcal{O}(\frac{\epsilon}{n})$
Optimization Landscape (Section 4)	$\mathcal{O}\left(\max\left\{1, \frac{\epsilon e^{-\epsilon^2}}{h^2}\right\}\right)$	$\mathcal{O}(\epsilon)$
Continuity Result (Section 4)	$\mathcal{O}\left(\frac{8(1+\delta)}{h^4}\epsilon e^{-\epsilon^2} + C\right)$	$2\sqrt{1+\delta_p}$
$\delta > \frac{1}{2}$ (Section 5)	$\mathcal{O}\left(-\left[1 - B\epsilon^2\right] + \sqrt{\left[1 - B\epsilon^2\right]^2 + B\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\epsilon(1+\epsilon)}{\sqrt{1-\epsilon}}\right)$
Lower Bound (Section 7)	$\frac{2L(1+\delta)e^{-\epsilon^2}}{(1-\delta)h^2}$	$\frac{4\sqrt{1+\delta}\epsilon}{L-2(1+\delta)}$
Convergence Result (Section L)	$\eta \leq \frac{h^2}{\left(12\rho^{1/2}C_0\right)}$	$\eta \leq \frac{1}{\left(12\rho^{1/2}C_0\right)}$

The Convergence analysis is provided in Section L, and the analysis of the behavior of the two loss functions under non-centered noise is provided in Section G. Table 1 summarizes the theoretical properties of the kernel loss and the standard MSE loss. Compared to MSE, the kernel loss exhibits smoother behavior, provides benign optimization landscapes, and maintains more robust theoretical guarantees even when the RIP constant exceeds $\frac{1}{2}$. It also offers tighter lower bounds and provable convergence under heavy-tailed noise, highlighting its robustness and theoretical advantages.

As discussed in Table 1, a comprehensive comparison between the mean squared error (MSE) and our proposed loss reveals that each has distinct advantages and limitations. To leverage the strengths of both, we construct a combined loss function that incorporates MSE and the proposed kernel-based loss, weighted by a learnable trade-off parameter λ. The resulting objective function is given in Equation (8.1). A concise theoretical analysis of the combined formulation is provided in Appendix S, where we highlight its optimization properties. In the following empirical study, we evaluate the performance of three loss functions: the MSE loss, the kernel loss, and the combined loss.

$$\begin{aligned}\mathbf{L}_{\text{combined}}(X) &= \lambda \cdot \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathcal{A}(XX^\top)_i\right)^2 \\ &+ (1-\lambda)\cdot n^{-1}\sum_{i=1}^{n}\left(-\log\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{((Y_j - (\mathcal{A}(XX^\top)_j)) - (Y_i - \mathcal{A}(XX^\top)_i))^2}{h^2}\right)\right).\end{aligned} \tag{8.1}$$

9. Empirical Study Of the Three Different Loss

In this section, we provide an illustrative example of the theoretical results. We examine the proximity of an arbitrary local minimizer $\hat{X}$ of the three losses to the ground truth in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, and study the effect of the step size on convergence. The setting follows the 1-bit Matrix Completion problem, a low-rank optimization task commonly used in recommendation systems [34, 35], with the same experimental configuration as [32]. Additional implementation details are given in Appendix T.

Figure 3 compares the bounds in Theorem 4.1 and Theorem 4.3 with $n = 40$ and $r = 5$. The y-axis reports the distance from an arbitrary local minimum $\hat{M}$ to M^*, measured in units of $\lambda_r(M^*)$, while the x-axis represents the probability lower bound, corresponding to the quantile of the noise norm $\|w\|$. The numerical results lie in the regime $\delta < 1/3$. The real error is the Frobenius norm of the recovered matrix under additive noise, and the numerical error is the corresponding computed upper bound. The kernel loss maintains stable error as the noise increases, with mild compression in some cases, consistent with Theorem 4.1. In contrast, the MSE loss exhibits an almost linear increase in error as the noise strengthens, matching Theorem 4.3. This confirms that the kernel loss can limit the influence of large noise and preserves robustness.

For low noise, the MSE loss achieves smaller error than the kernel loss, revealing a trade-off between precision and robustness. The composite loss combines the strengths of both objectives. As shown in the last columns of Figure 3, it reduces the noise sensitivity of MSE while improving the low-noise performance of the kernel loss, yielding more reliable recovery across a broader range of noise levels. Readers may refer to Appendix T for more visualizations of how the values of ζ_1, ζ_2, and Lipschitz

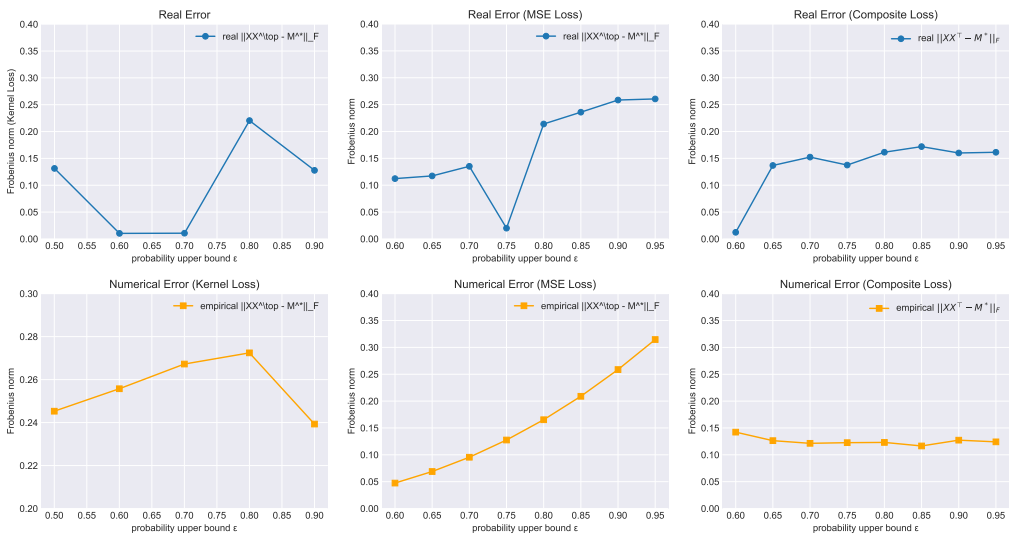

Figure 3: Comparison of real (top row) and numerical (bottom row) errors under different loss functions: kernel loss (left), MSE loss (middle), and composite loss (right), for $\delta < 1/3$.

constant L affect Theorem 4.1, especially under different uniform noise levels. More empirical results for $\delta > 1/3$ are also provided in Appendix T.

10. Conclusion

We compared the MSE loss and a kernel-based loss for low-rank matrix recovery in the presence of heavy-tailed noise. The exponential decay term in the proposed kernel-based loss reduces the impact of large outliers, whereas the MSE loss exhibits a linearly growing gradient. Our theoretical findings show that the kernel-based loss has more favorable upper and lower bounds in noisy regimes, and a combined formulation further balances these properties. Empirical results support these conclusions, indicating that the kernel-based and combined losses perform reliably under varying noise levels and values of δ.

11. Acknowledgment

The work was done while Xinyuan Song was at City University of Hong Kong. This work was supported by the Natural Science Foundation of China (62506314), the Research Grants Council of Hong Kong (21208525), and City University of Hong Kong (9382001).

References

[1] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[2] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[3] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[4] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems, 2009.

[5] Salar Fattahi and Somayeh Sojoudi. Efficient learning of bounded-rank positive semidefinite matrices from linear measurements. In *International Conference on Artificial Intelligence and Statistics*, pages 3015–3025. PMLR, 2020.

[6] Ross Anderson and Somayeh Sojoudi. Learning low-rank models for forecasting traffic flow. *IEEE Transactions on Intelligent Transportation Systems*, 20(11):4082–4093, 2019.

[7] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

[8] Nicolas Boumal. Nonconvex phase synchronization. In *Proceedings of the 2016 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 530–539. PMLR, 2016.

[9] Yoav Shechtman, Yonina C Eldar, Yair Cohen, Henry N Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

[10] Baosen Zhang, Albert Lam, David S Lee, and David Tse. Distributed optimal power flow using admm with a trust region method. *IEEE Transactions on Smart Grid*, 8(6):2784–2795, 2017.

[11] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

[12] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[13] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

[14] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[15] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[16] Xuyang Li, Cong Li, Yuejie Zhang, and Yuejie Chi. Towards fast and global convergence of gradient descent for nonconvex matrix factorization. *Information and Inference: A Journal of the IMA*, 9(4):915–935, 2020.

[17] Davis Park, Anastasios Kyrillidis, and Constantine Caramanis. Finding global minimizers of nonconvex low rank optimization problems. *arXiv preprint arXiv:1802.08463*, 2018.

[18] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[19] Ziye Ma, Yingjie Bi, Javad Lavaei, and Somayeh Sojoudi. Sharp restricted isometry property bounds for low-rank matrix recovery problems with corrupted measurements, 2023. URL https://arxiv.org/abs/2105.08232.

[20] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[21] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[22] Philip E Cheng. Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis*, 15(1):63–72, 1984.

[23] Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3): 1371–1385, 1994.

[24] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

[25] Geoffrey S Watson. Smooth regression analysis. *Sankya, Series A*, 26:359–372, 1964.

[26] Peter Hall and Li-Shan Huang. Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29(3):624–647, 2001.

[27] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018.

[28] Larry Schumaker. *Spline functions: basic theory*. Cambridge university press, 2007.

[29] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[30] Shaogao Lv, Huazhen Lin, Heng Lian, and Jian Huang. Oracle inequalities for sparse additive quantile regression in reproducing kernel hilbert space. *The Annals of Statistics*, 46(2):781–813, 2018.

[31] Xuancheng Wang, Ling Zhou, and Huazhen Lin. Deep regression learning with optimal loss function, 2023. URL https://arxiv.org/abs/2309.12872.

[32] Ziye Ma and Somayeh Sojoudi. Noisy low-rank matrix optimization: Geometry of local minima and convergence rate, 2023. URL https://arxiv.org/abs/2203.03899.

[33] Chong Bi and Javad Lavaei. Delocalization and stability in matrix sensing, 2020.

[34] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[35] Nadi Ghadermarzy, Ali Zandieh, Ioannis Mitliagkas, and Rebecca M Willett. Learning low-rank matrices from 1-bit observations under a general sampling model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[36] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20(1):37–39, 1953. doi: 10.1215/S0012-7094-53-02004-1.

[37] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912. doi: 10.1007/BF01456804.

[38] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Short proofs for the sub-gaussian concentration of norm and projection for isotropic log-concave distributions, 2019.

Appendix

A. Convergence Behavior Under Uniform Noise

Figures 4 and 5 illustrate the loss curves obtained under uniform noise sampled from the interval $[0, 1]$. The left plot corresponds to a setting where the final error converges to 0.5032, with a theoretically computed upper bound of 0.803, a Lipschitz constant of 2.2825922, and a Hessian constant of 3.3899682. In contrast, the right plot achieves a lower final error of 0.40832, with a corresponding upper bound of 0.9104874, a Lipschitz constant of 2.2984617, and a Hessian constant of 3.390177.

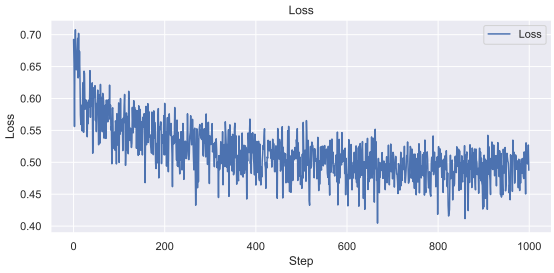

Figure 4: Loss under uniform noise setting 1.

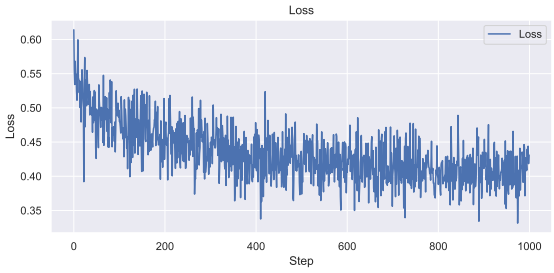

Figure 5: Loss under uniform noise setting 2.

The observed decrease in the loss values over time indicates a favorable convergence behavior. In both scenarios, the empirical loss steadily approaches a stable minimum, validating the optimization stability of the proposed approach. The final errors lie significantly below the respective theoretical upper bounds, suggesting that the training dynamics are well-controlled and that the theoretical estimates are conservative. These results confirm the practical viability of the method in the presence of stochastic perturbations.

B. Additional preliminaries

B.1. Difference and Geometry of Different Noise Bound

B.1.1. $\left|\langle \nabla_M \hat{g}(M, w) - \nabla_M \hat{g}(M, 0), K\rangle\right| \leq \zeta_1 \|\mathbf{w}\|_2 \|K\|_F$

Here, for a fixed model parameter M, the change in the gradient $\nabla_M \hat{g}(M, w)$ when moving from $w = 0$ to a nonzero noise vector $\mathbf{w}$ is measured by its projection onto some direction K, expressed via the inner product $\langle \cdot, K\rangle$. This inequality indicates that when the noise changes from 0 to $\mathbf{w}$, the change in the gradient (with respect to M) in any direction K is bounded by $\zeta_1 \|\mathbf{w}\|_2 \|K\|_F$.

B.1.2. $\|\nabla_M \hat{g}(M, w) - \nabla_M \hat{g}(M', w)\|_F \leq \rho \|M - M'\|_F$

Here, for a fixed noise vector w, we compare how the gradient $\nabla_M \hat{g}(M, w)$ changes when the model parameter moves from M to M'. This describes a ρ-Lipschitz continuity of the gradient with respect to M. This inequality shows that if the model parameter M changes slightly (while w is fixed), then the change in $\nabla_M \hat{g}(M, w)$ is controlled by $\|M - M'\|_F$ with the proportionality constant ρ.

B.1.3. $\|\nabla_M \hat{g}(M, w_1) - \nabla_M \hat{g}(M, w_2)\|_F \leq \lambda_1 \|w_1 - w_2\|_2$

Here, for a fixed model parameter M, we compare how the gradient $\nabla_M \hat{g}(M, w)$ changes when the noise vector goes from $\mathbf{w}_1$ to $\mathbf{w}_2$. This shows that when the noise vector $\mathbf{w}$ changes, the gradient $\nabla_M \hat{g}(M, w)$ with respect to M changes by at most $\lambda_1 \|\mathbf{w}_1 - \mathbf{w}_2\|_2$.

B.2. Definitions and Notations

In this paper:

- I_n refers to the identity matrix of size $n \times n$.
- $M \succeq 0$ means that M is a symmetric and positive semidefinite matrix.
- $\sigma_i(M)$ denotes the i-th largest singular value of a matrix M, and $\lambda_i(M)$ denotes the i-th largest eigenvalue of M.
- $\|v\|$ denotes the Euclidean norm of a vector v, while $\|M\|_F$ and $\|M\|_2$ denote the Frobenius norm and the operator norm of a matrix M, respectively.
- The inner product $\langle A, B\rangle$ is defined as $\text{tr}(A^\top B)$ for two matrices A and B of identical dimensions.
- For a matrix M, $\text{vec}(M)$ is the usual vectorization operation by stacking the columns of M into a vector.

The Hessian of the function $\hat{g}(\cdot,\cdot)$ with respect to the first argument M, denoted as $\nabla^2_M \hat{g}(\cdot,\cdot)$, can be regarded as a quadratic form whose action on any two matrices $K, L \in \mathbb{R}^{n\times n}$ is given by

$$[\nabla^2_M \hat{g}(M,w)](K,L) = \sum_{i,j,k,l=1}^{n} \frac{\partial^2 \hat{g}}{\partial M_{ij}\partial M_{kl}}(M,w) K_{ij} L_{kl}. \tag{B.1}$$

In this paper, $\nabla^2_M \hat{g}(M)$ and $\nabla^2 \hat{g}(M,w)$ are used interchangeably since w is an unknown fixed parameter, and it is impossible to take a derivative with respect to w.

B.3. Distance to Low-Rank Matrices

Define $M^* \in \arg\min_M f(M,0)$. We also characterize the distance of an arbitrary factorized point $X \in \mathbb{R}^{n\times r}$ to a rank-r positive semidefinite matrix M with the function $\text{dist}(X,M)$, defined as:

$$\text{dist}(X,M) = \min_{Z\in\mathcal{Z}} \|X - Z\|_F, \tag{B.2}$$

where

$$\mathcal{Z} = \{Z \in \mathbb{R}^{n\times r} \mid M = ZZ^\top\}. \tag{B.3}$$

Given a matrix $\hat{X} \in \mathbb{R}^{n\times r}$, define $\hat{X} \in \mathbb{R}^{n^2 \times nr}$ to be the matrix satisfying

$$\hat{X}\text{vec}(U) = \text{vec}(\hat{X}U^\top + U\hat{X}^\top), \quad \forall U \in \mathbb{R}^{n\times r}. \tag{B.4}$$

B.4. Projection onto a Low-Rank Manifold

Define $\mathcal{P}_r(M)$ of an arbitrary matrix M to be the projection of M onto a low-rank manifold of rank at most r:

$$\mathcal{P}_r(M) = \arg\min_{M_r\in\mathcal{M}} \|M_r - M\|_F, \tag{B.5}$$

where

$$\mathcal{M} := \{M \in \mathbb{S}^{n\times n} \mid \text{rank}(M) \le r, M \succeq 0\}. \tag{B.6}$$

For problem (1.2), $\mathcal{A} \in \mathbb{R}^{m\times n^2}$ is defined such that $\mathcal{A}\text{vec}(M) = \mathcal{A}(M)$. Finally, define:

$$h(X,w) := f(XX^\top, w). \tag{B.7}$$

The old loss function, it appears this is a quadratic loss: $h(X) = \sum_i \big(\mathcal{A}(\mathbf{X}\mathbf{X}^\top)_i - b_i\big)$.

Theorem B.1. *The objective function $\hat{g}(\cdot,w)$ of (2.4) has a first-order critical point M^w for every w such that it is symmetric, positive semidefinite, and $\text{rank}(M^w) \le r$.*

The proof is provided in Section E.

B.5. Heavy Tailed Analysis

We start with for fixed x, the probability $P(r_i > x) \geq P(p_i > x)$,

$$\frac{\partial L_i}{\partial p_i} = -\frac{2}{h^2}\frac{1}{nZ_i}\sum_{j=1}^{n}(p_j - p_i)\exp\Big(-\frac{(p_j - p_i)^2}{h^2}\Big). \tag{B.8}$$

$$\frac{\partial L_i}{\partial r_i} = -\frac{2}{h^2}\frac{1}{nZ_i}\sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{B.9}$$

Now we want to prove $\frac{\partial L_i}{\partial r_i} \geq \frac{\partial L_i}{\partial p_i}$. By Heavy tailed definition and we assume that r_i is more heavy tail than p_i, for every real number x:

$$P(r_i > x) \geq P(p_i > x). \tag{B.10}$$

here r_i and p_i are noise. A standard result is that if ϕ is any (measurable) function which is non-decreasing then under mild integrability conditions one has

$$\mathbb{E}[\phi(r_i)] \geq \mathbb{E}[\phi(p_i)]. \tag{B.11}$$

For a fixed index i one may imagine that the set $\{p_j\}_{j=1}^n$ (or $\{r_j\}_{j=1}^n$) provides independent draws from the underlying distribution. Then for any non-decreasing function ϕ we expect, probabilistically, that

$$E\big[\phi(r_j - r_i)\big] \leq E\big[\phi(p_j - p_i)\big]. \tag{B.12}$$

Thus we can prove that for larger r, which is also w, the effective loss (B.8) is larger.

C. Proof of Lemma 2.2

Proof. In (2.3), we let:

$$f_{ij}(M) = \exp\left(-\frac{((Y_j - (\mathcal{A}(M)_j)) - (Y_i - \mathcal{A}(M)_i))^2}{h^2}\right), \tag{C.1}$$

Then:

$$\hat{g}(M) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\right). \tag{C.2}$$

So we get:

$$\nabla_M \hat{g}(M) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)}\cdot\frac{1}{n}\sum_{j=1}^{n}\nabla_M f_{ij}(M), \tag{C.3}$$

$$f_{ij}(M) = \exp\left(-\frac{((Y_j - (\mathcal{A}(M)_j)) - (Y_i - \mathcal{A}(M)_i))^2}{h^2}\right). \tag{C.4}$$

We have $u_{ij}(M) = (Y_j - (\mathcal{A}(M)_j)) - (Y_i - \mathcal{A}(M)_i)$,$f_{ij}(M) = \exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)$. The gradient of $f_{ij}(M)$ with respect to M is:

$$\nabla_M f_{ij}(M) = f_{ij}(M)\cdot\left(-\frac{2u_{ij}(M)}{h^2}\right)\cdot\nabla_M u_{ij}(M). \tag{C.5}$$

In this process: $u_{ij}(M) = (Y_j - (\mathcal{A}(M)_j)) - (Y_i - \mathcal{A}(M)_i)$.,$\nabla_M u_{ij}(M) = -\nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i$. Using the expressions derived above, we have:

$$\nabla_M \hat{g}(M) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)}\cdot\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\cdot\left(-\frac{2u_{ij}(M)}{h^2}\right)\cdot\nabla_M u_{ij}(M). \tag{C.6}$$

Similarly for M':

$$\nabla_M \hat{g}(M') = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M')} \cdot \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M') \cdot \left(-\frac{2u_{ij}(M')}{h^2}\right) \cdot \nabla_M u_{ij}(M'). \tag{C.7}$$

And:

$$\begin{aligned}\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M') = -\frac{1}{n}\sum_{i=1}^{n} &\left(\frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)} \cdot \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M) \cdot \left(-\frac{2u_{ij}(M)}{h^2}\right) \cdot \nabla_M u_{ij}(M) \right) \\ &- \left(\frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M')} \cdot \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M') \cdot \left(-\frac{2u_{ij}(M')}{h^2}\right) \cdot \nabla_M u_{ij}(M') \right).\end{aligned} \tag{C.8}$$

Let:

$$A_i(M) = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)}, \tag{C.9}$$

$$B_{ij}(M) = f_{ij}(M) \cdot \left(-\frac{2u_{ij}(M)}{h^2}\right) \cdot \nabla_M u_{ij}(M). \tag{C.10}$$

Then:

$$\nabla_M \hat{g}(M) = -\frac{1}{n}\sum_{i=1}^{n} A_i(M) \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M). \tag{C.11}$$

Similarly for M':

$$\nabla_M \hat{g}(M') = -\frac{1}{n}\sum_{i=1}^{n} A_i(M') \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M'). \tag{C.12}$$

The difference is:

$$\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M') = -\frac{1}{n}\sum_{i=1}^{n} \left(A_i(M) \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) - A_i(M') \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M') \right). \tag{C.13}$$

Bounding the Difference. We can split Equation (C.13) into two parts:

$$\begin{aligned}&A_i(M) \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) - A_i(M') \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M') \\ &\quad = (A_i(M) - A_i(M')) \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) + A_i(M') \cdot \left(\frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) - \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M') \right).\end{aligned} \tag{C.14}$$

Term 1: $(A_i(M) - A_i(M')) \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M)$. Since $A_i(M) = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)}$, we have:

$$\|A_i(M) - A_i(M')\|_F \leq \left\| \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)} - \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M')} \right\|_F. \tag{C.15}$$

Using the mean value theorem for the function $f(x) = \frac{1}{x}$, we get:

$$\left\| \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)} - \frac{1}{\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M')} \right\|_F \leq \frac{C}{\left(\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\right)^2} \cdot \left\| \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M) - \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M') \right\|_F, \tag{C.16}$$

where C is a constant. Since $f_{ij}(M)$ is Lipschitz continuous with respect to M, we have:

$$|f_{ij}(M) - f_{ij}(M')| \leq L_f \|M - M'\|_F, \tag{C.17}$$

where L_f is the Lipschitz constant. Thus:

$$\left\| \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M) - \frac{1}{n}\sum_{j=1}^{n} f_{ij}(M') \right\|_F \leq L_f \|M - M'\|_F. \tag{C.18}$$

Combining Equation (C.18), we get:

$$\|A_i(M) - A_i(M')\|_F \leq \frac{CL_f \|M - M'\|_F}{\left(\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\right)^2}. \tag{C.19}$$

Term 2: $A_i(M') \cdot \left(\frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) - \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M')\right)$. Since $B_{ij}(M)$ involves $f_{ij}(M)$, $u_{ij}(M)$, and $\nabla_M u_{ij}(M)$, we need to bound each term:

$$\|B_{ij}(M) - B_{ij}(M')\|_F \leq L_B \|M - M'\|_F, \tag{C.20}$$

where L_B is a Lipschitz constant that depends on the Lipschitz constants of $f_{ij}(M)$, $u_{ij}(M)$, and $\nabla_M u_{ij}(M)$. Thus:

$$\left\| \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) - \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M') \right\|_F \leq L_B \|M - M'\|_F. \tag{C.21}$$

Combining the bounds for both terms in Equation (C.21) , we get:

$$\begin{aligned} &\|\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M')\|_F \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left(\frac{CL_f \|M - M'\|_F}{\left(\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\right)^2} \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) + A_i(M') \cdot L_B \|M - M'\|_F \right). \end{aligned} \tag{C.22}$$

Factoring out $\|M - M'\|_F$, we get:

$$\begin{aligned} &\|\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M')\|_F \\ &\leq \left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{CL_f}{\left(\frac{1}{n}\sum_{j=1}^{n} f_{ij}(M)\right)^2} \cdot \frac{1}{n}\sum_{j=1}^{n} B_{ij}(M) + A_i(M') \cdot L_B \right) \right) \|M - M'\|_F. \end{aligned} \tag{C.23}$$

Thus, we have:

$$\|\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M')\|_F \leq \rho \|M - M'\|_F, \tag{C.24}$$

where ρ is a constant that depends on the Lipschitz constants L_f and L_B, and the terms involving $A_i(M)$ and $B_{ij}(M)$.

□

D. Proof of Lemma 2.3

D.1. Jacobian Case

Recall that the function 2.3 regard to the gradient $\nabla_M \hat{g}(M, w)$ is:

$$\nabla_M \hat{g}(M, w) = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} f_{ij}(M, w)} \sum_{j=1}^{n} \frac{\partial f_{ij}(M, w)}{\partial M}, \tag{D.1}$$

where $f_{ij}(M, w) = \exp\left(-\frac{((Y_j + w_j - (AM)_j) - (Y_i + w_i - (AM)_i))^2}{h^2}\right)$

$$\begin{aligned} &\nabla_M \hat{g}(M, w_1) - \nabla_M \hat{g}(M, w_2) = \\ &-\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{\sum_{j=1}^{n} f_{ij}(M, w_1)} \sum_{j=1}^{n} \frac{\partial f_{ij}(M, w_1)}{\partial M} - \frac{1}{\sum_{j=1}^{n} f_{ij}(M, w_2)} \sum_{j=1}^{n} \frac{\partial f_{ij}(M, w_2)}{\partial M} \right). \end{aligned} \tag{D.2}$$

To bound the difference, we use the fact that the exponential function is Lipschitz continuous based on Lemma 2.2. Specifically, for any x and y,

$$\|e^x - e^y\| \leq e^{\max(x,y)}\|x - y\|. \tag{D.3}$$

Applying Equation (D.3) to our case, we get:

$$\begin{aligned}&\|f_{ij}(M, w_1) - f_{ij}(M, w_2)\|_F \leq \\ &f_{ij}(M, w_1) \cdot \Big\| - \frac{2}{h^2}\left((Y_j + w_{1j} - (AM)_j) - (Y_i + w_{1i} - (AM)_i)\right) - \\ &\left((Y_j + w_{2j} - (AM)_j) - (Y_i + w_{2i} - (AM)_i)\right)\Big\|_F.\end{aligned} \tag{D.4}$$

Simplifying further, we get:

$$\|f_{ij}(M, w_1) - f_{ij}(M, w_2)\|_F \leq f_{ij}(M, w_1) \cdot \frac{2}{h^2}\left\|(w_{1j} - w_{2j}) - (w_{1i} - w_{2i})\right\|_F. \tag{D.5}$$

Combining these terms in Equation (D.5), we get:

$$\|\nabla_M \hat{g}(M, w_1) - \nabla_M \hat{g}(M, w_2)\|_F \leq \frac{2}{h^2}\left(\sum_{i=1}^n \sum_{j=1}^n |(w_{1j} - w_{2j}) - (w_{1i} - w_{2i})|\,\|A_j - A_i\|_F\right). \tag{D.6}$$

Since $\|A_j - A_i\|_F$ is bounded by a constant C, we have:

$$\|\nabla_M \hat{g}(M, w_1) - \nabla_M \hat{g}(M, w_2)\|_F \leq \lambda\|w_1 - w_2\|_2, \tag{D.7}$$

where λ is a constant that depends on C, h, and the properties of the function $\hat{g}(M, w)$. This verifies the given inequality. Let the second w_2 to be zero and we get the original assumption 2.1.

D.2. Hessian Case

For notational convenience we use Equation (C.1) and Equation (C.2). and if we define $S_i(M, w) = \frac{1}{n}\sum_{j=1}^n f_{ij}(M, w)$,, one may show by the chain and quotient rules that:

$$\nabla_M \hat{g}(M, w) = -\frac{1}{n}\sum_{i=1}^n \frac{1}{S_i(M, w)} \sum_{j=1}^n \nabla_M f_{ij}(M, w), \tag{D.8}$$

and a second differentiation gives

$$\begin{aligned}\nabla_M^2 \hat{g}(M, w) =& \frac{1}{n}\sum_{i=1}^n \Big\{ \frac{1}{S_i(M, w)^2}\left[\sum_{j=1}^n \nabla_M f_{ij}(M, w)\right]\left[\sum_{j=1}^n \nabla_M f_{ij}(M, w)\right]^\top \\ &- \frac{1}{S_i(M, w)}\sum_{j=1}^n \nabla_M^2 f_{ij}(M, w)\Big\}.\end{aligned} \tag{D.9}$$

Notice that in each term the only dependence on w comes through the combination $(Y_j + w_j) - (Y_i + w_i)$, so that for any two vectors w_1 and w_2 we have

$$\Delta_{ij}(M, w_1) - \Delta_{ij}(M, w_2) = (w_{1j} - w_{2j}) - (w_{1i} - w_{2i}), \tag{D.10}$$

where we set

$$\Delta_{ij}(M, w) = (Y_j + w_j - (\mathcal{A}(M))_j) - (Y_i + w_i - (\mathcal{A}(M))_i). \tag{D.11}$$

Because the exponential function is smooth and each derivative (that is, $\nabla_M f_{ij}$ and $\nabla_M^2 f_{ij}$) involves factors such as $\exp\left(-\frac{\Delta_{ij}(M,w)^2}{h^2}\right)$ and $\frac{\Delta_{ij}(M,w)}{h^2}$ and $\frac{1}{h^2}$, a Taylor expansion shows that these derivatives (and hence also their sums and the quotient factors $1/S_i(M, w)$) are Lipschitz continuous with respect to w. More precisely, if we define

$$H(w) = \nabla_M^2 \hat{g}(M, w), \tag{D.12}$$

then using the triangle inequality we can write

$$\|H(w_1) - H(w_2)\|_F \le \frac{1}{n}\sum_{i=1}^{n}\Bigg\| \underbrace{\frac{1}{S_i(M,w_1)^2}\Big(\sum_j \nabla_M f_{ij}(M,w_1)\Big)\Big(\sum_j \nabla_M f_{ij}(M,w_1)\Big)^\top}_{\text{(I)}} \\ \underbrace{- \frac{1}{S_i(M,w_2)^2}\Big(\sum_j \nabla_M f_{ij}(M,w_2)\Big)\Big(\sum_j \nabla_M f_{ij}(M,w_2)\Big)^\top}_{\text{(I)}} \\ \underbrace{- \left[\frac{1}{S_i(M,w_1)}\sum_j \nabla_M^2 f_{ij}(M,w_1) - \frac{1}{S_i(M,w_2)}\sum_j \nabla_M^2 f_{ij}(M,w_2)\right]}_{\text{(II)}}\Bigg\|_F. \tag{D.13}$$

For notational clarity we denote

$$\begin{aligned} \text{(I)}(w) &= \frac{1}{S_i(M,w)^2}\Big(\sum_j \nabla_M f_{ij}(M,w)\Big)\Big(\sum_j \nabla_M f_{ij}(M,w)\Big)^\top, \\ \text{(II)}(w) &= \frac{1}{S_i(M,w)}\sum_j \nabla_M^2 f_{ij}(M,w). \end{aligned} \tag{D.14}$$

Then, by linearly combining these ingredients and applying the triangle inequality we have

$$\|H(w_1) - H(w_2)\|_F \le \frac{1}{n}\sum_{i=1}^{n}\Big\|\underbrace{\text{(I)}(w_1) - \text{(I)}(w_2)}_{\text{(I)}} - \underbrace{\big[\text{(II)}(w_1) - \text{(II)}(w_2)\big]}_{\text{(II)}}\Big\|_F. \tag{D.15}$$

Then we have the following two terms: 1.For (I): Both the scaling $1/S_i(M,w)^2$ and the gradient sums $\sum_j \nabla_M f_{ij}(M,w)$ depend on w only through the combination

$$\Delta_{ij}(M,w) = (Y_j + w_j - (\mathcal{A}(M))_j) - (Y_i + w_i - (\mathcal{A}(M))_i), \tag{D.16}$$

which is affine in w. Using the chain rule, one can show that there is a constant $L_{I,i}$ (depending on h, bounds on f_{ij} and on $S_i(M,w)$, etc.) such that

$$\Big\|\text{(I)}(w_1) - \text{(I)}(w_2)\Big\|_F \le L_{I,i}\|w_1 - w_2\|_2. \tag{D.17}$$

2.For (II): A similar analysis shows that there exists a constant $L_{II,i}$ such that

$$\Big\|\text{(II)}(w_1) - \text{(II)}(w_2)\Big\|_F \le L_{II,i}\|w_1 - w_2\|_2. \tag{D.18}$$

Thus, for each index i we obtain

$$\Big\|\text{(I)}(w_1) - \text{(I)}(w_2) - \big(\text{(II)}(w_1) - \text{(II)}(w_2)\big)\Big\|_F \le \Big(L_{I,i} + L_{II,i}\Big)\|w_1 - w_2\|_2. \tag{D.19}$$

Averaging over i we conclude

$$\|H(w_1) - H(w_2)\|_F \le \frac{1}{n}\sum_{i=1}^{n}\Big(L_{I,i} + L_{II,i}\Big)\|w_1 - w_2\|_2. \tag{D.20}$$

Defining

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}\Big(L_{I,i} + L_{II,i}\Big), \tag{D.21}$$

we then have

$$\|H(w_1) - H(w_2)\|_F \le \lambda\|w_1 - w_2\|_2. \tag{D.22}$$

Let the second w_2 to be zero and we get the original Assumption 2.1.

E. Proof of Theorem B.1

Proof. Given the loss function (2.3), and using the notations (C.1) with the (C.2), we rewrite the derivative form by:

$$\nabla_M f_{ij}(M) = \exp\left(\frac{u_{ij}(M)^2}{h^2}\right) \cdot \left(-\frac{2u_{ij}(M)}{h^2}\right) \cdot (-\nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i)\,. \tag{E.1}$$

Thus:

$$\begin{aligned}\nabla_M \hat{g}(M) =& \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)} \cdot \frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right) \\ & \cdot \left(-\frac{2u_{ij}(M)}{h^2}\right) \cdot (-\nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i)\,.\end{aligned} \tag{E.2}$$

Simplifying, we get:

$$\begin{aligned}\nabla_M \hat{g}(M) =& \frac{2}{nh^2}\sum_{i=1}^{n} \frac{1}{\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)} \cdot \sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right) \\ & \cdot u_{ij}(M) \cdot (\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j)\,.\end{aligned} \tag{E.3}$$

To find the critical points, we set $\nabla_M \hat{g}(M) = 0$:

$$\frac{2}{nh^2}\sum_{i=1}^{n} \frac{1}{\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)} \cdot \sum_{j=1}^{n}\exp\left(-\frac{u_{ij}(M)^2}{h^2}\right) \cdot u_{ij}(M) \cdot (\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j) = 0. \tag{E.4}$$

This equation must hold for all i. Solving this equation will give us the critical points M.

□

F. Proof of Theorem 3.2

Proof. We want to first prove that (2.3) can be better than the standard squared loss $\ell(M, w) = \|Y - \mathcal{A}(M)\|^2$,

$$\hat{g}(M, w) = \frac{1}{n}\sum_{i=1}^{n} -\log\left[\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\Delta_{ij}^2}{h^2}\right)\right], \tag{F.1}$$

with $\Delta_{ij} = (Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i)$. We use:

$$r_i \coloneqq Y_i + w_i - \mathcal{A}(M)_i, \quad i = 1, \ldots, n. \tag{F.2}$$

Then Equation (F.1) can be written as

$$\hat{g}(M, w) = \frac{1}{n}\sum_{i=1}^{n}\left[-\log\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{(r_j - r_i)^2}{h^2}\right)\right]. \tag{F.3}$$

We focus on the loss term for a fixed index i. For a given i define

$$L_i \coloneqq -\log\left(\frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big)\right). \tag{F.4}$$

We now compute the derivative of L_i with respect to r_i. Define

$$Z_i \coloneqq \frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{F.5}$$

Here $L_i = -\log Z_i$. Then, do differentiate L_i with respect to r_i. Using the chain rule we have

$$\frac{\partial L_i}{\partial r_i} = -\frac{1}{Z_i}\frac{\partial Z_i}{\partial r_i}. \tag{F.6}$$

Let us now differentiate Z_i:

$$\frac{\partial Z_i}{\partial r_i} = \frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial r_i}\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{F.7}$$

For any fixed j, note that:

$$\frac{\partial}{\partial r_i}\left(-\frac{(r_j - r_i)^2}{h^2}\right) = -\frac{\partial}{\partial r_i}\frac{(r_j - r_i)^2}{h^2} = -\frac{-2(r_j - r_i)}{h^2} = \frac{2(r_j - r_i)}{h^2}. \tag{F.8}$$

Putting Equations (F.6) (F.7) and (F.8) together we obtain:

$$\frac{\partial Z_i}{\partial r_i} = \frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big)\frac{2(r_j - r_i)}{h^2}. \tag{F.9}$$

Finally, the overall gradient of the loss with respect to the residuals is given by averaging over i:

$$\frac{\partial L_i}{\partial r_i} = -\frac{2}{h^2}\cdot\frac{1}{nZ_i}\sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{F.10}$$

This derivative shows how the loss changes with respect to the prediction error and illustrates its robust nature; large errors have a reduced effect because of the exponential weighting. The effective loss for $\hat{g}(X, w)$ is (F.3), and according to calculation in Section C, we have:

$$\frac{\partial \hat{g}}{\partial r_i} = -\frac{2}{n^2h^2}\left[\frac{1}{Z_i}\sum_{j=1}^{n}(r_j - r_i)e^{-\frac{(r_j - r_i)^2}{h^2}} + \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{Z_j}(r_i - r_j)e^{-\frac{(r_i - r_j)^2}{h^2}}\right], \tag{F.11}$$

So the two parts are likely the same. We only need to consider one part. The normalizing constants $Z_i = \sum_{j=1}^{n} e^{-\frac{(w_j - w_i)^2}{h^2}}$, are roughly of order n (i.e. $Z_i = O(n)$). The same holds for Z_j. Then for the first term,

$$\frac{1}{Z_i}\sum_{j=1}^{n}(w_j - w_i)e^{-\frac{(w_j - w_i)^2}{h^2}}, \tag{F.12}$$

we note that if the typical magnitude of w is, say, $\|w\|$, then the difference $w_j - w_i$ is at most $O(\|w\|)$. With approximately n terms in the sum and a normalization factor $1/Z_i = O(1/n)$, we obtain an overall size:

$$\frac{1}{Z_i}\sum_{j=1}^{n}(w_j - w_i)e^{-\frac{(w_j - w_i)^2}{h^2}} = \mathcal{O}(\|w\|e^{-\frac{\|w\|^2}{h^2}}). \tag{F.13}$$

A similar argument applies for the summation. In summary, we can write

$$\frac{\partial \hat{g}}{\partial w} = \mathcal{O}\Big(\frac{\|w\|e^{-\frac{\|w\|^2}{h^2}}}{n^2h^2}\Big). \tag{F.14}$$

A similar argument applies for the summation of vectors. In summary, we can write:

$$\left\|\frac{\partial \hat{g}}{\partial w}\right\| = \mathcal{O}\Big(\frac{\|w\|e^{-\frac{\|w\|^2}{h^2}}}{n^2h^2}\Big). \tag{F.15}$$

For the MSE loss defined by:

$$\ell_{\text{MSE}}(M, w) = \frac{1}{n}\sum_{i=1}^{n}\left(r_i^2\right) \quad \text{with} \quad r_i = Y_i + w_i - \mathcal{A}(M)_i, \tag{F.16}$$

the (partial) derivative with respect to each r_i is

$$\frac{\partial \ell_{\text{MSE}}}{\partial r_i} = \frac{2}{n} r_i. \tag{F.17}$$

Thus the gradient norm is:

$$\left\| \frac{\partial \ell_{\text{MSE}}}{\partial w} \right\| = \mathcal{O}\left(\frac{\|w\|}{n} \right). \tag{F.18}$$

Notice that this derivative is linear in w and hence its magnitude grows without bound as $\|w\| \to \infty$.

□

G. Relationship between different loss

To analyze the behavior and robustness of alternative loss functions under noisy observations, we present the following theorem, which compares covariance-based and kernel-based loss formulations to the standard mean squared error (MSE) criterion.

Theorem G.1. *Let $w = (w_1, w_2, \ldots, w_n)$ be a noise vector, and suppose w follows a centered distribution. That is, for residuals*

$$r_i = Y_i + w_i - \mathcal{A}(M)_i, \quad \textit{we have} \quad \frac{1}{n} \sum_{j=1}^{n} r_j = 0. \tag{G.1}$$

Define the covariance loss *by*

$$L_{\text{cov}}(r_i) = \frac{1}{n} \sum_{j=1}^{n} \frac{(r_i - r_j)^2}{h^2}. \tag{G.2}$$

Under the centered noise assumption, L_{cov} reduces to the mean squared error (MSE) loss. Next, define the exponential kernel loss *by*

$$L_{\exp}(r_i) = \frac{1}{n} \sum_{j=1}^{n} \exp\Big(-\big(\tfrac{r_i - r_j}{h}\big)^2 \Big). \tag{G.3}$$

This exponential kernel loss aligns with the following form of an optimal loss:

$$\hat{g}(M, w) = \frac{1}{n} \sum_{i=1}^{n} \left(-\log\Big[\frac{1}{n} \sum_{j=1}^{n} \exp\Big(-\frac{\big((Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i)\big)^2}{h^2} \Big) \Big] \right). \tag{G.4}$$

As the noise amplitude $\|w\|$ increases, the exponential kernel loss is suppressed, and the effective residual tends to be small within clusters. In particular, the weighted average of the residual r_i is

$$\overline{r}_i = \frac{\sum_{j=1}^{n} r_j \exp\Big(-\dfrac{(r_j - r_i)^2}{h^2} \Big)}{\sum_{j=1}^{n} \exp\Big(-\dfrac{(r_j - r_i)^2}{h^2} \Big)}, \tag{G.5}$$

and

$$\frac{\partial L_i}{\partial r_i} = -\frac{2}{h^2} \frac{1}{n Z_i} \sum_{j=1}^{n} (r_j - r_i) \exp\Big(-\frac{(r_j - r_i)^2}{h^2} \Big), \tag{G.6}$$

which indicates that $L_{\exp}$ is not sensitive to large residuals.

Finally, if the loss function is not centered, then both the kernel loss and the covariance loss may converge to a suboptimal solution that does not match the ground truth. In that setting, even the optimal loss described above can perform worse than the MSE loss.

The proof is provided in Section G.1. Theorem G.1 establishes a theoretical relationship between the traditional mean squared error (MSE) loss and a kernel-based robust alternative by analyzing the behavior of residual-dependent loss functions under a centered noise assumption. Specifically, it shows that the covariance loss, defined by normalized pairwise squared differences of residuals, reduces to the MSE when the noise vector w is centered, i.e., when the empirical mean of the residuals is zero.

The theorem then introduces the exponential kernel loss, which applies a Gaussian-type weighting to the residual differences. This loss aligns with the structure of the proposed robust objective $\hat{g}(M, w)$, where the inner exponential acts to suppress the contribution of large residuals. As the noise amplitude increases, the exponential weights diminish for outlier points, making the loss function focus on local, more coherent clusters of residuals. The weighted mean $\overline{r}_i$ and the derivative $\frac{\partial L_i}{\partial r_i}$ further demonstrate that this kernel loss downweights large deviations and is inherently more robust to noise than the MSE.

However, the theorem also cautions that this robustness depends critically on the centering of the noise. When the noise is not centered, both the covariance and kernel-based losses may fail to identify the correct minimizer, potentially performing worse than the MSE. This highlights an important limitation of these methods: their effectiveness relies on structural assumptions about the data distribution, particularly the unbiasedness of the residuals.

G.1. Proof of Theorem G.1

G.2. For Covariance Loss

We want to analyze the loss function

$$L(r_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{(r_i - r_j)^2}{h^2}, \tag{G.7}$$

then the derivative of Equation (G.7) is

$$\frac{dL}{dr_i} = \frac{2}{nh^2}\sum_{j=1}^{n}(r_i - r_j). \tag{G.8}$$

Assume that the values r_j are drawn from a centered distribution, so $\frac{1}{n}\sum_{j=1}^{n} r_j = 0$. Then the gradient (G.7) becomes

$$\frac{\partial L}{\partial r_i} = \frac{2r_i}{h^2}. \tag{G.9}$$

Thus, for a centered distribution the gradient of our normalized loss is $2r_i$, implying that every deviation from zero is penalized linearly. So such loss function is the same with MSE loss.

G.3. Old Kernel Loss

We start with the old kernel loss function only regard to i:

$$L(r_i) = \frac{1}{n}\sum_{j=1}^{n}\exp\left(-\left(\frac{r_i - r_j}{h}\right)^2\right). \tag{G.10}$$

Simplify to obtain the final answer:

$$\frac{\partial L}{\partial r_i} = -\frac{2}{nh^2}\sum_{j=1}^{n}(r_i - r_j)\exp\left(-\left(\frac{r_i - r_j}{h}\right)^2\right). \tag{G.11}$$

G.4. Distribution approximation of w

G.4.1. When the Loss Becomes Good

We start with:

$$\frac{\partial L_i}{\partial r_i} = -\frac{2}{h^2}\frac{1}{nZ_i}\sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big), \tag{G.12}$$

and we want to choose the value (or distribution) of r_i such that the magnitude of this derivative is small. A natural way to do this is to set the derivative to zero. Ignoring the constant factor $-\frac{2}{h^2}\frac{1}{nZ_i}$, we see that

$$\sum_{j=1}^{n}(r_j - r_i)w(r_j, r_i) = 0, \quad \text{with} \quad w(r_j, r_i) = \exp\left(-\frac{(r_j - r_i)^2}{h^2}\right). \tag{G.13}$$

Thus, if we define the weighted average:

$$\overline{r}_i = \frac{\sum_{j=1}^{n} r_j \exp\left(-\frac{(r_j - r_i)^2}{h^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{(r_j - r_i)^2}{h^2}\right)}, \tag{G.14}$$

the vanishing of the derivative occurs when

$$r_i = \overline{r}_i. \tag{G.15}$$

This is exactly the mean of the r_j's weighted by a Gaussian kernel centered at r_i. In words, the derivative $\frac{\partial L_i}{\partial r_i}$ is small (or zero) when r_i is at the center of a symmetric, tight cluster of points $\{r_j\}$.

G.4.2. When the Loss Becomes Bad

We start with:

$$\frac{\partial L_i}{\partial r_i} = -\frac{2}{h^2}\frac{1}{nZ_i}\sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{G.16}$$

Notice that aside from the overall constant, the value of the derivative is determined by the weighted sum:

$$S(r_i) = \sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big). \tag{G.17}$$

A large (in magnitude) derivative will occur when the sum $S(r_i)$ is as far from zero as possible. To make $S(r_i)$ large, we need the differences $r_j - r_i$ to have essentially the same sign so that they add up rather than cancel. This happens when r_i is located at one end of the data.

The derivative is bigger (in magnitude) when the points r_j are not symmetrically distributed about r_i but are instead all on one side of r_i; that is, when r_i is at one extreme (e.g. at the left or right boundary) of the r_j distribution.

G.5. Example, when r_i follows Gaussian

Assume that the neighborhood values r_k (with k indexing the neighbors, here j) are i.i.d. samples from a Gaussian distribution

$$p(r) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right). \tag{G.18}$$

For large n the sums:

$$\begin{aligned} Z_i &= \frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big), \\ S(r_i) &= \frac{1}{n}\sum_{j=1}^{n}(r_j - r_i)\exp\Big(-\frac{(r_j - r_i)^2}{h^2}\Big) \end{aligned} \tag{G.19}$$

can be well approximated by integrals over the density $\mathrm{p(r)}$. That is, we have:

$$\begin{aligned} Z_i &\sim \int \exp\Big(-\frac{(r-r_i)^2}{h^2}\Big)p(r)dr, \\ S(r_i) &\sim \int (r-r_i)\exp\Big(-\frac{(r-r_i)^2}{h^2}\Big)p(r)dr. \end{aligned} \tag{G.20}$$

Because the exponential kernel is symmetric, if you choose $r_i = \mu$, then in the integrals the term $(r-\mu)$ is integrated against a function that is symmetric about $r=\mu$. In that case, positive and negative contributions cancel and one obtains $S(\mu)=0$, so that:

$$\left.\frac{\partial L_i}{\partial r_i}\right|_{r_i=\mu} = 0. \tag{G.21}$$

Now, if r_i differs from μ, then the weighting is no longer symmetric. For example, if $r_i < \mu$, then most of the weight in the Gaussian density $p(r)$ lies to the right of r_i (i.e. for $r > \mu$) and thus most terms in $S(r_i)$ are positive; similarly if $r_i > \mu$ the sum is negative. In both cases, the difference $|r_i - \mu|$ produces a nonzero—and generally larger—value of $|\frac{\partial L_i}{\partial r_i}|$, since the cancellation in the weighted sum is diminished.

H. Rough Bound and For the Kernel Loss

H.1. Rough Bound For $\|\Delta\|_F = \|M - M^*\|_F$

Lemma H.1 (Behavior of $\hat{g}(M,w)$ Around Local Minima and Ground Truth)**.** *Let $\hat{g}(\cdot,w)$ be a twice-differentiable loss function, and suppose $\nabla^2_M \hat{g}(M^*,w)$ is positive definite at the ground truth minimizer M^*. Define $\Delta \coloneqq M - M^*$, the loss function $\hat{g}(M)$ is from (2.3), if we left the $\lambda_{\min}\big(\nabla^2_M \hat{g}(M^*,w)\big)$ uncomputed, then we have:*

(*A*) (Case: M is a local minimizer near M^*) *If M is a local minimizer of $\hat{g}$ close to M^*, then*

$$\|\Delta\|_F = \|M - M^*\|_F \le \sqrt{\frac{2}{\lambda_{\min}\big(\nabla^2_M \hat{g}(M^*,w)\big)}\Big[\hat{g}(M,w) - \hat{g}(M^*,w)\Big]}. \tag{H.1}$$

Moreover, as the noise level $\|w\|$ increases, $\|\Delta\|_F$ does not *increase linearly but instead exhibits an* exponential decay *trend in $\|w\|$. In this regime, the* new loss $\hat{g}(M,w)$ *is* smoother (*it has a smaller Lipschitz constant ρ*)*, and consequently M transitions* toward M^* *when $\|w\|$ decreases.*

(*B*) (Case: M is a local minimizer *not* at M^*) *Alternatively, consider a local minimizer M that is* not *the ground truth M^*. Assume $\|\Delta\|_F$ cannot be made arbitrarily small. Then one may write*

$$\|\Delta\|_F = \|M - M^*\|_F \ge \sqrt{\frac{2}{L}\Big[\hat{g}(M,w) - \hat{g}(M^*,w)\Big]}, \tag{H.2}$$

for some constant $L > 0$. In this scenario, as $\|w\|$ grows, $\|\Delta\|_F$ can grow faster than linearly in $\|w\|$. Here the kernel loss exhibits a rougher landscape (with a larger Lipschitz constant ρ), and hence when $\|w\|$ decreases, M may actually move away from M^*.The loss function still becomes smoother as $\|w\|$ changes, but the local minimum remains distant from the ground truth.

The proof are provided in Section I. In short, **(A)** describes a desirable regime in which $\hat{g}$ is smooth and strongly convex near M^*, causing Δ to shrink with decreasing noise; **(B)** describes a regime where a spurious local minimum persists, with Δ potentially growing under increased noise and not converging to M^*.

Lemma H.1 characterizes the behavior of the proposed loss function $\hat{g}(M,w)$ in the neighborhood of the ground truth minimizer M^*, under the assumption that the Hessian $\nabla^2_M \hat{g}(M^*,w)$ is positive definite. The lemma considers two scenarios depending on the location of a local minimizer M relative to M^*.

In case **(A)**, when M is a local minimizer close to M^*, the difference $\|\Delta\|_F = \|M - M^*\|_F$ can be bounded above by a term proportional to the square root of the suboptimality gap, scaled by the inverse of the smallest eigenvalue of the Hessian at M^*. Importantly, due to the structure of $\hat{g}(M, w)$, this bound implies that as the noise norm $\|w\|$ increases, the effect on $\|\Delta\|_F$ is not linear but rather decays exponentially. In this regime, the loss landscape becomes smoother (i.e., has smaller gradient Lipschitz constant ρ), and M tends to move closer to the ground truth as the noise level decreases.

In contrast, case **(B)** considers the situation where M is a spurious local minimizer not coinciding with M^*, and the deviation $\|\Delta\|_F$ is non-negligible. In this case, a lower bound on $\|\Delta\|_F$ is given in terms of the suboptimality gap and a generic smoothness constant L. Here, $\|\Delta\|_F$ may grow faster than linearly with $\|w\|$, and the loss landscape becomes rougher (with larger ρ). Even though the overall smoothness of $\hat{g}$ improves as $\|w\|$ changes, the minimizer M does not necessarily converge to the ground truth, highlighting the sensitivity of the optimization process to local geometry.

This lemma therefore distinguishes between a favorable regime (near-global minimum) where robustness and convergence are preserved, and an unfavorable regime (spurious local minimum) where robustness may deteriorate despite the loss surface becoming smoother in a global sense.

H.2. Lower Bound For $\lambda_{\min}$

Lemma H.2 (Lower Bound on the Minimum Eigenvalue of the Hessian)**.** *Let $\hat{g}(M, w)$ be a smooth function defined via a kernel-smoothed loss involving the operator $\mathcal{A}$, and suppose that: the operator $\mathcal{A}$ satisfies the Restricted Isometry Property (RIP) with constant δ_p in Equation (2.6), The residual terms $z_{ij}(M^*)$ are uniformly bounded: $|z_{ij}(M^*)| \leq B$. The kernel average $G_i(M^*) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(-\frac{z_{ij}(M^*)^2}{h^2}\right)$ is bounded below: $G_{\min} = \min_i G_i(M^*) > 0$. The smoothness constants satisfy:*

$$L_1 = 2(1 + \delta_p), \quad L_2 = 0. \tag{H.3}$$

The Hessian $\nabla^2_M \hat{g}(M^*, w)$ is symmetric positive definite, with $v^\top \nabla^2_M \hat{g}(M^*, w)v \geq c\|v\|^2$ for some constant $c > 0$ and all v. Then the minimum eigenvalue of the Hessian satisfies the lower bound:

$$\lambda_{\min}\big(\nabla^2_M \hat{g}(M^*, w)\big) \geq c, \tag{H.4}$$

where an explicit expression for c is given by

$$c = \frac{2}{G_{\min} h^2}\left(L_1^2\left(1 + \frac{2B^2}{h^2}\right) + BL_2\right) + \frac{4}{G_{\min}^2 h^4} B^2 L_1^2. \tag{H.5}$$

The proof is provided in Section I.4. Lemma H.2 provides a lower bound on the smallest eigenvalue of the Hessian of the smoothed loss function $\hat{g}(M, w)$, evaluated at the ground truth matrix M^*. This result is important for establishing local strong convexity, which in turn guarantees stability and convergence of optimization algorithms near M^*.

The lemma assumes that the loss function is constructed using a kernel-based smoothing over residuals defined by a linear operator $\mathcal{A}$ that satisfies the Restricted Isometry Property (RIP). Under this condition, as well as uniform boundedness of the residuals $z_{ij}(M^*)$ and the positivity of the kernel averages $G_i(M^*)$, an explicit lower bound c for the minimal eigenvalue is derived. The expression for c depends on the kernel bandwidth parameter h, the residual bound B, the RIP constant δ_p, and the kernel average lower bound $G_{\min}$.

Crucially, this bound ensures that the Hessian $\nabla^2_M \hat{g}(M^*, w)$ is well-conditioned near M^*, with eigenvalues bounded away from zero. This guarantees that $\hat{g}(M, w)$ is locally strongly convex around the ground truth, which is essential for ensuring that gradient-based methods converge efficiently to M^* and that local perturbations in M lead to bounded variations in the objective.

I. Proof of the main Theorem 4.1

I.1. Error Upper Bound

Lemma I.1 (Upper Bound on the recovery Error)**.** *Let $\hat{g}(M, w)$ be a smooth loss function with minimizer M^* at fixed noise w, and suppose the Hessian $\nabla^2_M \hat{g}(M^*, w)$ is symmetric positive definite, with smallest eigenvalue $\lambda_{\min}\left(\nabla^2_M \hat{g}(M^*, w)\right) > 0$. $\delta > 0$ is a constant and $R(w)$ is a auxiliary residual term. The residual term $R(w)$ is bounded by*

$$R(w) = O\left(\frac{\|w\| e^{-w^2}}{h^2}\right). \tag{I.1}$$

Then the estimation error satisfies the upper bound

$$\|M - M^*\|_F \leq \max\left\{\sqrt{\frac{2}{\lambda_{\min}\left(\nabla^2_M \hat{g}(M^*, w)\right)}}, \frac{2(1+\delta)R(w)}{1 - \delta - \lambda_{\min}\left(\nabla^2_M \hat{g}(M^*, w)\right)}\right\}. \tag{I.2}$$

The proof is provided in Section I.9 together with the tight estimation of $\lambda_{\min}(\nabla^2_M \hat{g}(M^*, w)$. Lemma I.1 provides an upper bound on the estimation error $\|M - M^*\|_F$ between a candidate solution M and the true minimizer M^* of the kernel-smoothed loss function $\hat{g}(M, w)$, in the presence of fixed noise w.

The resulting bound has a two-part structure. The first term corresponds to the curvature-controlled region, where small suboptimality in $\hat{g}(M, w)$ implies proximity to M^* via standard strong convexity arguments. The second term dominates when the residual term is significant, capturing the influence of noise through the interaction between $R(w)$, the loss curvature, and the parameter δ. Notably, due to the exponential decay of $R(w)$, the estimation error does not grow linearly in $\|w\|$, indicating robustness of the estimator even under moderate levels of noise.

I.2. Gradient Continuity Result

Lemma I.2 (Continuity of Gradient with Respect to Noise)**.** *Let $\hat{g}(M, w)$ be the smoothed loss function depending on noise variable w, and suppose that the data residuals $z_j - z_i$ satisfy Assumptions 2.3 and 2.1, which is: $|z_j - z_i| \leq C\|w\|$ for some constant $C > 0$ and all relevant i, j. Assume also that the operator $\mathcal{A}$ satisfies the RIP condition with constant δ. Then the gradient of $\hat{g}(M, w)$ with respect to M is Lipschitz continuous in w, with Lipschitz constant λ satisfying*

$$\|\nabla_M \hat{g}(M, w) - \nabla_M \hat{g}(M, 0)\| \leq \lambda \|w\|, \tag{I.3}$$

where the Lipschitz constant is bounded above by

$$\lambda = \sup_{\xi \in [0,w]} \left\| \frac{d}{dw} \nabla_M \hat{g}(M, \xi) \right\| \leq \frac{8(1+\delta)}{h^4} \|w\| e^{-w^2}. \tag{I.4}$$

The proof is provided in Section I.11.

Lemma I.2 establishes a form of Hölder-type continuity for the gradient of the smoothed loss function $\hat{g}(M, w)$ with respect to the noise variable w. Under the assumption that the data residuals $z_j - z_i$ grow at most linearly in $\|w\|$, and that the measurement operator $\mathcal{A}$ satisfies the Restricted Isometry Property (RIP), the gradient $\nabla_M \hat{g}(M, w)$ is shown to be Lipschitz continuous with respect to w, with an explicitly computable upper bound on the Lipschitz constant λ.

Importantly, this upper bound decays exponentially in $\|w\|$, which reflects the robustness of the kernel-smoothed loss to large noise perturbations. Unlike traditional losses where gradient sensitivity may grow linearly or quadratically with noise, the bound here reveals that the influence of noise on the gradient diminishes rapidly as $\|w\|$ increases—due to the presence of exponential terms in the kernel weights. This behavior is characteristic of smoothing via Gaussian kernels and is key to the stability of the optimization process in high-noise regimes.

I.3. Upper Bound For $\|M - M^*\|_F$

We want to use the Taylor expansion of such loss. Under the assumption that $\hat{g}$ is twice differentiable in a neighborhood of M^*, we have

$$\hat{g}(M, w) = \hat{g}(M^*, w) + \langle \nabla_M \hat{g}(M^*, w), \Delta \rangle + \frac{1}{2} \langle \Delta, \nabla_M^2 \hat{g}(M^*, w)[\Delta] \rangle + o(\|\Delta\|_F^2). \tag{I.5}$$

Since M^* minimizes $\hat{g}$ (at least locally), $\nabla_M \hat{g}(M^*, w) = 0$, so that to second order we have

$$\hat{g}(M, w) - \hat{g}(M^*, w) = \frac{1}{2} \langle \Delta, \nabla_M^2 \hat{g}(M^*, w)[\Delta] \rangle + o(\|\Delta\|_F^2). \tag{I.6}$$

This local strong convexity (or quadratic error bound) is beneficial for optimization—gradient-based methods will enjoy a linear (or even faster) convergence rate provided their step sizes are chosen appropriately.

$$\hat{g}(M, w) - \hat{g}(M^*, w) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{\displaystyle\sum_{j=1}^{n} \exp\left(- \frac{((Y_j + w_j - \mathcal{A}(M^*)_j) - (Y_i + w_i - \mathcal{A}(M^*)_i))^2}{h^2} \right)}{\displaystyle\sum_{j=1}^{n} \exp\left(- \frac{((Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i))^2}{h^2} \right)} \right]. \tag{I.7}$$

$$\begin{aligned} G_i(M) &= \frac{1}{n} \sum_{j=1}^{n} \exp\left(- \frac{z_{ij}(M)^2}{h^2} \right), \\ \text{with} \quad z_{ij}(M) &= \Big((Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i) \Big). \end{aligned} \tag{I.8}$$

Then we have

$$\hat{g}(M, w) = -\frac{1}{n} \sum_{i=1}^{n} \log \Big[G_i(M) \Big]. \tag{I.9}$$

A standard application of the chain rule shows that, for a given i,

$$\nabla_M \log G_i(M) = \frac{1}{G_i(M)} \nabla_M G_i(M) \tag{I.10}$$

and hence the Hessian (i.e. second derivative with respect to M) satisfies

$$\nabla_M^2 \log G_i(M) = \frac{1}{G_i(M)} \nabla_M^2 G_i(M) - \frac{1}{[G_i(M)]^2} \nabla_M G_i(M) \nabla_M G_i(M)^\top. \tag{I.11}$$

Thus,

$$\begin{aligned} \nabla_M^2 \hat{g}(M, w) &= -\frac{1}{n} \sum_{i=1}^{n} \nabla_M^2 \log G_i(M) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{G_i(M)} \nabla_M^2 G_i(M) - \frac{1}{[G_i(M)]^2} \nabla_M G_i(M) \nabla_M G_i(M)^\top \right\}. \end{aligned} \tag{I.12}$$

For completeness we now indicate the first– and second–order derivatives of $G_i(M)$. Note that only the dependence of $G_i(M)$ on M appears through the residuals

$$z_{ij}(M) = \Big((Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i) \Big), \tag{I.13}$$

so that using the chain rule we obtains

$$\nabla_M G_i(M) = \frac{1}{n} \sum_{j=1}^{n} \exp\Big(- \frac{z_{ij}(M)^2}{h^2} \Big) \left(- \frac{2 z_{ij}(M)}{h^2} \right) \Big[- \nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i \Big], \tag{I.14}$$

or, equivalently,

$$\nabla_M G_i(M) = \frac{2}{nh^2}\sum_{j=1}^{n}\exp\Big(-\frac{z_{ij}(M)^2}{h^2}\Big) z_{ij}(M)\Big[\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j\Big]. \tag{I.15}$$

Likewise, one may differentiate once more to obtain

$$\begin{aligned}\nabla^2_M G_i(M) = \frac{2}{nh^2}\sum_{j=1}^{n}\exp\Big(-\frac{z_{ij}(M)^2}{h^2}\Big)\Bigg\{&\Big[\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j\Big] \\ \times\Big[\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j\Big]^\top &\times\left(1-\frac{2z_{ij}(M)^2}{h^2}\right) + z_{ij}(M)\Big[\nabla^2_M \mathcal{A}(M)_i - \nabla^2_M \mathcal{A}(M)_j\Big]\Bigg\}.\end{aligned} \tag{I.16}$$

We start from:

$$\hat{g}(M,w) - \hat{g}(M^*,w) = \frac{1}{2}\langle \Delta, \nabla^2_M \hat{g}(M^*,w)[\Delta]\rangle + o(\|\Delta\|_F^2), \tag{I.17}$$

where $\Delta = M - M^*$. Assume that the Hessian at M^* is (strictly) positive definite from B.1 and let

$$\lambda_{\min} = \lambda_{\min}\big(\nabla^2_M \hat{g}(M^*,w)\big) \tag{I.18}$$

denote its smallest eigenvalue. Then, using the spectral properties of the Hessian,

$$\langle \Delta, \nabla^2_M \hat{g}(M^*,w)[\Delta]\rangle \geq \lambda_{\min}\|\Delta\|_F^2. \tag{I.19}$$

Thus, for $\|\Delta\|$ sufficiently small (so that the $o(\|\Delta\|_F^2)$ term is negligible) we have

$$\hat{g}(M,w) - \hat{g}(M^*,w) \geq \frac{1}{2}\lambda_{\min}\|\Delta\|_F^2. \tag{I.20}$$

$$\|\Delta\|_F^2 \leq \frac{2}{\lambda_{\min}}\Big(\hat{g}(M,w) - \hat{g}(M^*,w)\Big). \tag{I.21}$$

Taking square roots of Equation (I.21) yields the desired bound for Δ in Frobenius norm,

$$\|\Delta\|_F \leq \sqrt{\frac{2}{\lambda_{\min}\big(\nabla^2_M \hat{g}(M^*,w)\big)}\Big(\hat{g}(M,w) - \hat{g}(M^*,w)\Big)}. \tag{I.22}$$

This provides the upper bound for the deviation $\Delta = M - M^*$ in terms of the function gap and the smallest eigenvalue of the Hessian at M^*. This result is Section H.1

I.4. Lower Bound For $\lambda_{\min}$

Then we want to compute the upper bound for $\lambda_{\min}$. Firstly we want to calculate the bound for $\|\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j\|$, and $\|\nabla^2_M \mathcal{A}(M)_i - \nabla^2_M \mathcal{A}(M)_j\| \leq L_2$ We begin by noting that RIP condition

$$(1-\delta_p)\|X\|_F^2 \leq \|AX\|_F^2 \leq (1+\delta_p)\|X\|_F^2 \tag{I.23}$$

We set $f(M) = \|\mathcal{A}M\|_F^2$. Because $\mathcal{A}$ is linear, this function is quadratic. Let us compute its gradient and Hessian.

$$\nabla f(M_i) - \nabla f(M_j) = 2\mathcal{A}^\top \mathcal{A}(M_i - M_j). \tag{I.24}$$

Taking norms and using the fact that all singular values of $\mathcal{A}^\top\mathcal{A}$ lie between $1-\delta_p$ and $1+\delta_p$ (from the RIP condition) we deduce

$$\|\nabla f(M_i) - \nabla f(M_j)\| \leq 2\|\mathcal{A}^\top\mathcal{A}\|\|M_i - M_j\| \leq 2(1+\delta_p)\|M_i - M_j\|. \tag{I.25}$$

Thus

$$L_1 = 2(1+\delta_p). \tag{I.26}$$

Differentiating the gradient (I.25) we see that the Hessian is

$$\nabla^2 f(M) = 2\mathcal{A}^\top\mathcal{A}, \tag{I.27}$$

which is independent of M. (In other words, the Hessian is constant throughout the domain.) Therefore, for any two points M_i and M_j we have:

$$\nabla^2 f(M_i) - \nabla^2 f(M_j) = 0. \tag{I.28}$$

Thus the Lipschitz constant for the Hessian (i.e. the constant L_2 satisfying

$$\|\nabla^2_M \mathcal{A}(M)_i - \nabla^2_M \mathcal{A}(M)_j\| \leq L_2 \|M_i - M_j\| \tag{I.29}$$

can be taken as $L_2 = 0 \leq C$. This result is a direct consequence of the fact that for a quadratic function defined by a linear operator the Hessian is constant.

$$\|\nabla_M \mathcal{A}(M)_i - \nabla_M \mathcal{A}(M)_j\| \leq L_1; \tag{I.30}$$

$$\|\nabla^2_M \mathcal{A}(M)_i - \nabla^2_M \mathcal{A}(M)_j\| \leq L_2; \tag{I.31}$$

Here Suppose also that the residuals are bounded in absolute value,

$$|z_{ij}(M^*)| \leq B, \quad \text{for all } i, j. \tag{I.32}$$

Also, note that for each i the quantity

$$G_i(M^*) = \frac{1}{n} \sum_{j=1}^{n} \exp\Big(-\frac{z_{ij}(M^*)^2}{h^2}\Big) \tag{I.33}$$

is a positive average. Define the minimum of (I.33) over i by

$$G_{\min} = \min_i G_i(M^*). \tag{I.34}$$

Since the exponential is always positive this is bounded away from 0 under Assumption 2.1.

Bound the norm of the first derivative $\nabla_M G_i(M^*)$. We have

$$\nabla_M G_i(M^*) = \frac{2}{nh^2} \sum_{j=1}^{n} \exp\Big(-\frac{z_{ij}(M^*)^2}{h^2}\Big) z_{ij}(M^*) \Big[\nabla_M \mathcal{A}(M^*)_i - \nabla_M \mathcal{A}(M^*)_j\Big]. \tag{I.35}$$

Using the bounds, we get

$$\|\nabla_M G_i(M^*)\| \leq \frac{2}{h^2} B L_1, \tag{I.36}$$

since the exponential is at most 1. Now we want to bound the norm of the second derivative $\nabla^2_M G_i(M^*)$. Its schematic form is

$$\nabla^2_M G_i(M^*) = \frac{2}{nh^2} \sum_{j=1}^{n} e^{-\frac{z_{ij}(M^*)^2}{h^2}} \Bigg\{ \Big[\nabla_M \mathcal{A}(M^*)_i - \nabla_M \mathcal{A}(M^*)_j\Big] \times \Big[\nabla_M \mathcal{A}(M^*)_i - \nabla_M \mathcal{A}(M^*)_j\Big]^\top \times \Big(1 - \frac{2 z_{ij}(M^*)^2}{h^2}\Big) + z_{ij}(M^*) \Big[\nabla^2_M \mathcal{A}(M^*)_i - \nabla^2_M \mathcal{A}(M^*)_j\Big] \Bigg\}. \tag{I.37}$$

Taking norms and using triangle and sub-multiplicative properties we obtain an upper bound of the form

$$\|\nabla^2_M G_i(M^*)\| \leq \frac{2}{h^2} \left(L_1^2 \Big(1 + \frac{2B^2}{h^2}\Big) + B L_2 \right). \tag{I.38}$$

Now, look at each term in the Hessian from Equation (I.37). The first term contributes

$$\left\| \frac{1}{G_i(M^*)} \nabla^2_M G_i(M^*) \right\| \leq \frac{2}{G_{\min} h^2} \left(L_1^2 \Big(1 + \frac{2B^2}{h^2}\Big) + B L_2 \right), \tag{I.39}$$

and the second term gives

$$\left\| \frac{1}{\big[G_i(M^*)\big]^2} \nabla_M G_i(M^*) \nabla_M G_i(M^*)^\top \right\| \leq \frac{1}{G^2_{\min}} \|\nabla_M G_i(M^*)\|^2 \leq \frac{4}{G^2_{\min} h^4} B^2 L_1^2. \tag{I.40}$$

Because the Hessian $\nabla_M^2 \hat{g}(M^*, w)$ is an average (with a minus sign) of these terms, a bound for its operator norm is

$$\|\nabla_M^2 \hat{g}(M^*, w)\| \leq \frac{1}{n} \sum_{i=1}^{n} \left[\frac{2}{G_{\min} h^2} \left(L_1^2 \Big(1 + \frac{2B^2}{h^2}\Big) + BL_2 \right) + \frac{4}{G_{\min}^2 h^4} B^2 L_1^2 \right]. \tag{I.41}$$

Since the bound is uniform in i, we obtain

$$\|\nabla_M^2 \hat{g}(M^*, w)\| \leq \frac{2}{G_{\min} h^2} \left(L_1^2 \Big(1 + \frac{2B^2}{h^2}\Big) + BL_2 \right) + \frac{4}{G_{\min}^2 h^4} B^2 L_1^2. \tag{I.42}$$

From Assumption B.1, one can reasonably assume that—in view of the problem's structure—there exists a constant $c > 0$ such that

$$v^\top \nabla_M^2 \hat{g}(M^*, w) v \geq c\|v\|^2, \quad \text{for all } v. \tag{I.43}$$

Then one may choose

$$\lambda_{\min}\big(\nabla_M^2 \hat{g}(M^*, w)\big) \geq c. \tag{I.44}$$

In summary, one obtains an explicit lower bound (depending on h, B, G_{min}, L_1, L_2, and the problem dimension) of the form

$$\lambda_{\min}\big(\nabla_M^2 \hat{g}(M^*, w)\big) \geq c \quad \text{with} \quad c = \frac{2}{G_{\min} h^2} \left(L_1^2 \left(1 + \frac{2B^2}{h^2}\right) + BL_2 \right) + \frac{4}{G_{\min}^2 h^4} B^2 L_1^2, \tag{I.45}$$

I.5. Upper Bound For $\|M - M^*\|_F$

Now we want to calculate the lower bound for δ. We start from the second-order Taylor expansion around the optimum M^*. Defining

$$\Delta = M - M^*, \tag{I.46}$$

we have

$$\hat{g}(M, w) - \hat{g}(M^*, w) = \frac{1}{2} \langle \Delta, \nabla_M^2 \hat{g}(M^*, w)[\Delta] \rangle + o(\|\Delta\|_F^2). \tag{I.47}$$

Assume that in a neighborhood of M^* the function $\hat{g}$ is twice differentiable and its Hessian satisfies the following two assumptions:

Assumption I.3 (Strong Convexity)**.** There exists $\lambda_{\min} > 0$ such that for all Δ,

$$\langle \Delta, \nabla_M^2 \hat{g}(M^*, w)[\Delta] \rangle \geq \lambda_{\min} \|\Delta\|_F^2. \tag{I.48}$$

Assumption I.4 (Smoothness)**.** There exists a constant $L > 0$ such that for all Δ,

$$\langle \Delta, \nabla_M^2 \hat{g}(M^*, w)[\Delta] \rangle \leq L \|\Delta\|_F^2. \tag{I.49}$$

Then the Taylor expansion yields the following two-sided bounds for small Δ: Using strong convexity we obtain

$$\hat{g}(M, w) - \hat{g}(M^*, w) \geq \frac{1}{2} \lambda_{\min} \|\Delta\|_F^2 + o(\|\Delta\|_F^2). \tag{I.50}$$

Neglecting the higher-order term for sufficiently small Δ and rearranging gives an upper bound on $\|\Delta\|_F$

$$\|\Delta\|_F \leq \sqrt{\frac{2}{\lambda_{\min}} \Big[\hat{g}(M, w) - \hat{g}(M^*, w) \Big]}. \tag{I.51}$$

Using the smoothness (an upper bound on the Hessian) we similarly obtain

$$\hat{g}(M, w) - \hat{g}(M^*, w) \leq \frac{1}{2} L \|\Delta\|_F^2 + o(\|\Delta\|_F^2), \tag{I.52}$$

so that for sufficiently small Δ

$$\|\Delta\|_F \geq \sqrt{\frac{2}{L} \Big[\hat{g}(M, w) - \hat{g}(M^*, w) \Big]}. \tag{I.53}$$

Thus, the lower bound for $\|\Delta\|_F$ is given by

$$\|\Delta\|_F \geq \sqrt{\frac{2}{L} \left(\hat{g}(M, w) - \hat{g}(M^*, w) \right)}. \tag{I.54}$$

Here L is an upper bound on the eigenvalues of $\nabla_M^2 \hat{g}(M^*, w)$ (or equivalently a smoothness constant for $\hat{g}$). In many practical settings one might be able to compute or bound L using known Lipschitz constants of the components involved in $\hat{g}$.

I.6. MSE case

I.6.1. Optimization Landscape $\|M - M^*\|_F$ case

Given that for any matrix M the linear operator $\mathcal{A}$ satisfies the Restricted Isometry Property (RIP) (2.6). The function is

$$f(M) = \|Y + w - \mathcal{A}(M)\|_2^2. \tag{I.55}$$

$r(M) = Y + w - \mathcal{A}(M)$, the function can be written as $f(M) = \langle r(M), r(M)\rangle$. Taking the gradient with respect to M yields $\nabla f(M) = -2\mathcal{A}^*\big(r(M)\big)$, and then the Hessian (the second derivative with respect to M) is $\nabla^2 f(M) = 2\mathcal{A}^*\mathcal{A}$. Notice that this Hessian is independent of the value of M. In particular, at $M = M^*$, we have $\nabla^2 f(M^*, w) = 2\mathcal{A}^*\mathcal{A}$. For any matrix R,

$$\langle R, \mathcal{A}^*\mathcal{A}(R)\rangle = \langle \mathcal{A}(R), \mathcal{A}(R)\rangle = \|\mathcal{A}(R)\|_2^2. \tag{I.56}$$

Thus, for the Hessian we have

$$\langle R, \nabla^2 f(M^*, w)[R]\rangle = 2\|\mathcal{A}(R)\|_2^2, \tag{I.57}$$

and consequently,

$$2(1 - \delta_p)\|R\|_F^2 \le \langle R, \nabla^2 f(M^*, w)[R]\rangle \le 2(1 + \delta_p)\|R\|_F^2. \tag{I.58}$$

The smallest eigenvalue is thus at least

$$\lambda_{\min}\big(\nabla_M^2 f(M^*, w)\big) \ge 2(1 - \delta_p). \tag{I.59}$$

Then we have:

$$f(M, 0) - f(M^*, 0) \ge \delta\|M - M^*\|_F^2 + \frac{1 - 3\delta}{2}\|M - M^*\|_F^2. \tag{I.60}$$

Now, since the full difference is a sum of the noise–free part plus a noise correction we may write

$$f(M, w) - f(M^*, w) = \Big[f(M, 0) - f(M^*, 0)\Big] + E(w), \tag{I.61}$$

Now we want to calculate $E(w)$

$$f(M, w) = \|Y + w - \mathcal{A}(M)\|_2^2, \tag{I.62}$$

and in the noise–free case

$$f(M, 0) = \|Y - \mathcal{A}(M)\|_2^2. \tag{I.63}$$

We set:

$$u_1 = Y - \mathcal{A}(M) \quad \text{and} \quad u_2 = Y - \mathcal{A}(M^*). \tag{I.64}$$

We also set:

$$\begin{aligned} f(M, w) &= \|Y - \mathcal{A}(M)\|_2^2 + 2\langle Y - \mathcal{A}(M), w\rangle + \|w\|_2^2, \\ f(M^*, w) &= \|Y - \mathcal{A}(M^*)\|_2^2 + 2\langle Y - \mathcal{A}(M^*), w\rangle + \|w\|_2^2. \end{aligned} \tag{I.65}$$

Consider the difference $f(M, w) - f(M^*, w)$:

$$\begin{aligned} f(M, w) - f(M^*, w) &= \Big[\|Y - \mathcal{A}(M)\|_2^2 - \|Y - \mathcal{A}(M^*)\|_2^2\Big] \\ &\quad + 2\Big[\langle Y - \mathcal{A}(M), w\rangle - \langle Y - \mathcal{A}(M^*), w\rangle\Big] + \Big[\|w\|_2^2 - \|w\|_2^2\Big] \\ &= \Big[f(M, 0) - f(M^*, 0)\Big] + 2\Big[\langle Y - \mathcal{A}(M), w\rangle - \langle Y - \mathcal{A}(M^*), w\rangle\Big]. \end{aligned} \tag{I.66}$$

$$E(w) = 2\Big[\langle Y - \mathcal{A}(M), w\rangle - \langle Y - \mathcal{A}(M^*), w\rangle\Big]. \tag{I.67}$$

Notice that:

$$\langle Y - \mathcal{A}(M), w\rangle - \langle Y - \mathcal{A}(M^*), w\rangle = \langle \mathcal{A}(M^*) - \mathcal{A}(M), w\rangle. \tag{I.68}$$

Thus, we obtain

$$E(w) = 2\langle \mathcal{A}(M^*) - \mathcal{A}(M), w\rangle. \tag{I.69}$$

For low rank recovery shown in [32], we still have:

$$\big\|f(M,w)-f(M^*,w)\big\| \geq \frac{1-\delta}{2}\|M-M^*\|_F^2 - \big\|E(w)\big\|. \tag{I.70}$$

Using the Cauchy–Schwarz inequality we have

$$|E(w)| = 2\Big|\langle \mathcal{A}(M^*)-\mathcal{A}(M), w\rangle\Big| \leq 2\|\mathcal{A}(M^*)-\mathcal{A}(M)\|_2\|w\|. \tag{I.71}$$

Again applying the (2.6), we obtain

$$\|\mathcal{A}(M^*)-\mathcal{A}(M)\|_2 \leq \sqrt{1+\delta}\|M-M^*\|_F. \tag{I.72}$$

Thus,

$$|E(w)| \leq 2\sqrt{1+\delta}\|w\|\|M-M^*\|_F. \tag{I.73}$$

Taking absolute values and using the triangle inequality on (I.66), we conclude that

$$\begin{aligned}\big\|f(M,w)-f(M^*,w)\big\| &\leq \Big\|f(M,0)-f(M^*,0)\Big\| + \|E(w)\| \\ &\leq (1+\delta)\|M-M^*\|_F^2 + 2\sqrt{1+\delta}\|w\|\|M-M^*\|_F.\end{aligned} \tag{I.74}$$

With the previous result (I.51):

$$\|M-M^*\|_F^2 \leq \frac{2}{\lambda_{\min}}\Big[f(M,w)-f(M^*,w)\Big]. \tag{I.75}$$

$$\lambda_{\min}\big(\nabla_M^2 f(M^*,w)\big) \geq 2(1-\delta_p), \tag{I.76}$$

Put Equation (I.75) and (I.76) into (I.74), we have derived an upper bound on the function-difference:

$$f(M,w)-f(M^*,w) \leq (1+\delta)\|M-M^*\|_F^2 + 2\sqrt{1+\delta}\|w\|\|M-M^*\|_F. \tag{I.77}$$

$$\|M-M^*\|_F^2 \leq \frac{2}{\lambda_{\min}}\Big[(1+\delta)\|M-M^*\|_F^2 + 2\sqrt{1+\delta}\|w\|\|M-M^*\|_F\Big]. \tag{I.78}$$

Since $\|M-M^*\|_F \geq 0$ and $\delta_p > 0$, reversing the inequality leads to:

$$(2(1+\delta)-\lambda_{\min})\|M-M^*\|_F^2 - 4\sqrt{1+\delta}\|w\|\|M-M^*\|_F \leq 0. \tag{I.79}$$

$$\|M-M^*\|_F \leq \frac{\sqrt{1+\delta_p}\|w\|}{\delta_p}. \tag{I.80}$$

Thus it is:

$$\|M-M^*\|_F = O(\|w\|). \tag{I.81}$$

I.6.2. Hölder Continuous Case

Since the derivative is independent of w, we get

$$\lambda = \sup_{\xi\in[0,w]}\left\|\frac{d}{dw}\nabla_M f(M,\xi)\right\|. \tag{I.82}$$

Often the Restricted Isometry Property (RIP) guarantees that the operator $\mathcal{A}$ satisfies $\|\mathcal{A}(X)\|_2 \leq \sqrt{1+\delta}\|X\|_F$, which implies that the operator norm of $\mathcal{A}$ is bounded by $\sqrt{1+\delta}$. Since the operator norm of $\mathcal{A}^*$ equals that of $\mathcal{A}$, we have $\|\mathcal{A}^*\| \leq \sqrt{1+\delta}$. Thus, we obtain the upper bound

$$\lambda = O(1), \quad \text{more precisely} \quad \lambda \leq 2\sqrt{1+\delta}. \tag{I.83}$$

I.7. Intermediate Result and Proof of Lemma I.5

First, we want to prove the intermediate result:

Lemma I.5. *With the assumptions in (2.3) and (2.3) loss function, suppose we have RIP 2.6 conditions and M^* is the ground truth matrix, we have:*

$$\hat{g}(M,0) - \hat{g}(M^*,0) \geq \delta\|M - M^*\|_F^2 + \frac{1-3\delta}{2}\|M - M^*\|_F^2. \tag{I.84}$$

Proof. Define $\hat{M} \coloneqq \hat{X}\hat{X}^\top$ and

$$\bar{M} \coloneqq \hat{M} - \frac{1}{1+\delta+\zeta_2 q}\nabla_M f(\hat{M}, w). \tag{I.85}$$

Additionally, define $\phi(\cdot)$ as

$$\phi(M) \coloneqq \langle \nabla_M f(\hat{M}, w), M - \hat{M}\rangle + \frac{1+\delta+\zeta_2 q}{2}\|M - \hat{M}\|_F^2. \tag{I.86}$$

Now,

$$\begin{aligned}\frac{1+\delta+\zeta_2 q}{2}\|M - \bar{M}\|_F^2 &= \frac{1+\delta+\zeta_2 q}{2}\|M - \hat{M} + \frac{1}{1+\delta+\zeta_2 q}\nabla_M f(\hat{M}, w)\|_F^2 \\ &= \frac{1+\delta+\zeta_2 q}{2}\|M - \hat{M}\|_F^2 + \langle \nabla_M f(\hat{M}, w), M - \hat{M}\rangle + \\ &\frac{1}{(1+\delta+\zeta_2 q)^2}\|\nabla_M f(\hat{M}, w)\|_F^2 \\ &= \phi(M) + \text{constant with respect to } M.\end{aligned} \tag{I.87}$$

Next, we apply the Taylor expansion to $f(M, w)$ at $\hat{M}$ and combine it with the RIP property to obtain

$$f(M^*, w) \geq f(\hat{M}, w) + \langle \nabla_M f(\hat{M}, w), M^* - \hat{M}\rangle + \frac{1-\delta-\zeta_2 q}{2}\|M^* - \hat{M}\|_F^2. \tag{I.88}$$

Additionally, by expanding at M^*, we can also write:

$$\begin{aligned}f(\hat{M}, w) - f(M^*, w) &\geq \langle \nabla_M f(M^*, w), \hat{M} - M^*\rangle + \frac{1-\delta-\zeta_2 q}{2}\|\hat{M} - M^*\|_F^2 \\ &\geq \frac{1-\delta-\zeta_2 q}{2}\|\hat{M} - M^*\|_F^2 - \zeta_1 q\|\hat{M} - M^*\|_F\end{aligned} \tag{I.89}$$

□

I.8. Proof of Main Theorem (Theorem 4.1)

We start with the definition:

$$\hat{g}(M, w) = -\frac{1}{n}\sum_{i=1}^{n}\log\Big[F_i(M, w)\Big], \tag{I.90}$$

with

$$F_i(M, w) = \frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\Delta_{ij}(w)^2}{h^2}\right) \tag{I.91}$$

and

$$\Delta_{ij}(w) = \Big(Y_j + w_j - \mathcal{A}(M)_j\Big) - \Big(Y_i + w_i - \mathcal{A}(M)_i\Big). \tag{I.92}$$

$$\nabla_M \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right] = \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right]\cdot\left(-\frac{2\Delta_{ij}(w)}{h^2}\right)\nabla_M\Delta_{ij}(w). \tag{I.93}$$

Note that:

$$\Delta_{ij}(w) = \big(Y_j + w_j\big) - \big(Y_i + w_i\big) - \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big], \tag{I.94}$$

so that:

$$\nabla_M \Delta_{ij}(w) = -\nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]. \tag{I.95}$$

Thus, we obtain

$$\nabla_M F_i(M,w) = -\frac{1}{n}\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right] \frac{2\Delta_{ij}(w)}{h^2} \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]. \tag{I.96}$$

with its gradient:

$$\nabla_M \hat{g}(M,w) = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{F_i(M,w)} \nabla_M F_i(M,w). \tag{I.97}$$

Substituting $\nabla_M F_i(M,w)$ into (I.96) gives

$$\nabla_M \hat{g}(M,w) = \frac{2}{h^2 n}\sum_{i=1}^{n} \frac{\frac{1}{n}\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right] \Delta_{ij}(w) \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]}{F_i(M,w)}. \tag{I.98}$$

Therefore, the difference between the gradients is

$$\begin{aligned}\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0) = &\frac{2}{h^2 n}\sum_{i=1}^{n}\Bigg\{ \frac{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right] \Delta_{ij}(w) \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]}{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right]} \\ &- \frac{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(0)^2}{h^2}\right] \Delta_{ij}(0) \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]}{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(0)^2}{h^2}\right]} \Bigg\}.\end{aligned} \tag{I.99}$$

We assume $w \sim 0$ and we can simplify this by:

$$\begin{aligned}\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0) = &\frac{2}{h^2 n}\sum_{i=1}^{n}\Bigg\{ \frac{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right] \Delta_{ij}(w) \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]}{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(0)^2}{h^2}\right]} \\ &- \frac{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(0)^2}{h^2}\right] \Delta_{ij}(0) \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]}{\sum_{j=1}^{n} \exp\left[-\frac{\Delta_{ij}(0)^2}{h^2}\right]} \Bigg\}.\end{aligned} \tag{I.100}$$

Because the noise only enters via $\Delta_{ij}(w) = \Delta_{ij}(0) + (w_j - w_i)$, we first write the term in the numerator of the Equation (I.100) as

$$\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big]\Delta_{ij}(w) = \exp\left[-\frac{\Delta_{ij}(w)^2}{h^2}\right]\Big[\Delta_{ij}(0) + (w_j - w_i)\Big]. \tag{I.101}$$

Thus, the whole difference inside the braces can be written as

$$\begin{aligned}&\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big]\Big[\Delta_{ij}(0) + (w_j - w_i)\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Delta_{ij}(0) \\ &\qquad = \Big\{\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j - w_i) + \Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0)\Big\}.\end{aligned} \tag{I.102}$$

We now substitute this back into (I.100). Keeping the $\exp[-\Delta_{ij}(w)^2/h^2]$ factor intact, our first–order (in w) approximation for small w is:

$$\begin{aligned}\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0) \sim &\frac{2}{h^2 n}\sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]} \times \sum_{j=1}^{n}\Bigg\{ \exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j - w_i) \\ &+ \Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0)\Bigg\} \times \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big].\end{aligned} \tag{I.103}$$

So we have:

$$\hat{g}(M,w) - \hat{g}(M^*,w) = \int_0^1 \Big\langle \nabla_M \hat{g}\Big(M^* + t(M-M^*), w\Big), M - M^* \Big\rangle \, dt. \tag{I.104}$$

Next, we decompose the gradient into its noise–free part plus its noise–dependent correction:

$$\nabla_M \hat{g}(M,w) = \nabla_M \hat{g}(M,0) + \Big[\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0)\Big]. \tag{I.105}$$

Thus, we write:

$$\begin{aligned} \hat{g}(M,w) - \hat{g}(M^*,w) = &\int_0^1 \Big\langle \nabla_M \hat{g}\Big(M^* + t(M-M^*), 0\Big), M - M^* \Big\rangle \, dt \\ &+ \int_0^1 \Big\langle \Delta\nabla_M \hat{g}\Big(M^* + t(M-M^*), w\Big), M - M^* \Big\rangle \, dt, \end{aligned} \tag{I.106}$$

where

$$\Delta\nabla_M \hat{g}(M,w) = \nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0). \tag{I.107}$$

The first term of Eqation (I.106) is the usual noise–free difference $\hat{g}(M,0) - \hat{g}(M^*,0)$. Now, using our previous approximation (I.103) we have:

$$\begin{aligned} \Delta\nabla_M \hat{g}(M,w) \sim \frac{2}{h^2 n} \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]} \times \sum_{j=1}^n \Bigg\{ &\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j - w_i) \\ &+ \Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0) \Bigg\} \times \nabla_M \Big[\mathcal{A}(M)_j - \mathcal{A}(M)_i\Big]. \end{aligned} \tag{I.108}$$

Here we want to recall our notations:

$$\Delta_{ij}(w) = \Delta_{ij}(0) + (w_j - w_i) \quad \text{with} \quad \Delta_{ij}(0) = \Big(Y_j - \mathcal{A}(M)_j\Big) - \Big(Y_i - \mathcal{A}(M)_i\Big). \tag{I.109}$$

$$\begin{aligned} \hat{g}(M,w) - \hat{g}(M^*,w) \sim &\Big[\hat{g}(M,0) - \hat{g}(M^*,0)\Big] + \frac{2}{h^2 n} \int_0^1 \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]} \\ &\times \sum_{j=1}^n \Bigg\{ \exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j - w_i) + \Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0) \Bigg\} \\ &\times \Big\langle \nabla_M \Big[\mathcal{A}\big(M^* + t(M-M^*)\big)_j - \mathcal{A}\big(M^* + t(M-M^*)\big)_i\Big], M - M^* \Big\rangle \, dt. \end{aligned} \tag{I.110}$$

We are given a lower bound on the noise-free part from Equation (I.70):

$$\hat{g}(M,0) - \hat{g}(M^*,0) \geq \delta \|M - M^*\|_F^2 + \frac{1-3\delta}{2} \|M - M^*\|_F^2. \tag{I.111}$$

so that:

$$\hat{g}(M,0) - \hat{g}(M^*,0) \geq \frac{1-\delta}{2} \|M - M^*\|_F^2. \tag{I.112}$$

Now, since the full difference is a sum of the noise–free part plus a noise correction we may write:

$$\hat{g}(M,w) - \hat{g}(M^*,w) = \Big[\hat{g}(M,0) - \hat{g}(M^*,0)\Big] + E(w), \tag{I.113}$$

with

$$\begin{aligned} E(w) = &\frac{2}{h^2 n} \int_0^1 \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]} \times \sum_{j=1}^n \Bigg\{ \exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j - w_i) \\ &+ \Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big] - \exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0) \Bigg\} \\ &\times \Big\langle \nabla_M \Big[\mathcal{A}\Big(M^* + t(M-M^*)\Big)_j - \mathcal{A}\Big(M^* + t(M-M^*)\Big)_i\Big], M - M^* \Big\rangle \, dt. \end{aligned} \tag{I.114}$$

Taking absolute values on Equation (I.113), we obtain by the triangle inequality

$$\big\|\hat{g}(M,w)-\hat{g}(M^*,w)\big\| \geq \frac{1-\delta}{2}\|M-M^*\|_F^2 - \big\|E(w)\big\|. \tag{I.115}$$

At this point, we may bound the correction term $\|E(w)\|$ by assuming that the perturbative factors and the derivative of $\mathcal{A}(\cdot)$ are bounded. Assume that there exists a constant $L>0$ such that:

$$\left\|\nabla_M\Big[\mathcal{A}\Big(M^*+t(M-M^*)\Big)_j - \mathcal{A}\Big(M^*+t(M-M^*)\Big)_i\Big]\right\| \leq L = 1+\delta, \tag{I.116}$$

and denote by:

$$\begin{aligned} R(w) := \frac{2}{h^2 n}\sup_{t\in[0,1]}\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{n}\exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]}\sum_{j=1}^{n}\Big|\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big](w_j-w_i) \\ +\Big[\exp\Big[-\frac{\Delta_{ij}(w)^2}{h^2}\Big]-\exp\Big[-\frac{\Delta_{ij}(0)^2}{h^2}\Big]\Big]\Delta_{ij}(0)\Big|. \end{aligned} \tag{I.117}$$

We need to mention that actually $R(w)<0$. Then we may bound

$$\|E(w)\| \leq LR(w)\|M-M^*\|_F. \tag{I.118}$$

In summary, we have the lower bound

$$\big\|\hat{g}(M,w)-\hat{g}(M^*,w)\big\| \geq \frac{1-\delta}{2}\|M-M^*\|_F^2 - (1+\delta)R(w)\|M-M^*\|_F. \tag{I.119}$$

I.9. Finally We Calculate the Upper Bound For $\|M-M^*\|_F$

we recall that:

$$\lambda_{\min}\big(\nabla_M^2\hat{g}(M^*,w)\big)\geq c \quad \text{with} \quad c=\frac{2}{G_{\min}h^2}\left(L_1^2\left(1+\frac{2B^2}{h^2}\right)+BL_2\right)+\frac{4}{G_{\min}^2h^4}B^2L_1^2, \tag{I.120}$$

$$\|\Delta\|_F = \|M-M^*\|_F \leq \sqrt{\frac{2}{\lambda_{\min}\left(\nabla_M^2\hat{g}(M^*,w)\right)}\Big[\hat{g}(M,w)-\hat{g}(M^*,w)\Big]}, \tag{I.121}$$

where $\lambda_{\min}\big(\nabla_M^2\hat{g}(M^*,w)\big)$ is the smallest eigen-value of the Hessian at M^*:

$$\|M-M^*\|_F \leq \sqrt{\frac{2}{\lambda_{\min}\Big(\nabla_M^2\hat{g}(M^*,w)\Big)}\Big[\frac{1-\delta}{2}\|M-M^*\|_F^2-(1+\delta)R(w)\|M-M^*\|_F\Big]}. \tag{I.122}$$

To eliminate the square root, square both sides (noting that all terms are nonnegative):

$$\|M-M^*\|_F^2 \leq \frac{2}{\lambda_{\min}\Big(\nabla_M^2\hat{g}(M^*,w)\Big)}\Big[\frac{1-\delta}{2}\|M-M^*\|_F^2-(1+\delta)R(w)\|M-M^*\|_F\Big]. \tag{I.123}$$

Multiply both sides by $\lambda_{\min}\Big(\nabla_M^2\hat{g}(M^*,w)\Big)$:

$$\lambda_{\min}\Big(\nabla_M^2\hat{g}(M^*,w)\Big)\|M-M^*\|_F^2 \leq (1-\delta)\|M-M^*\|_F^2 - 2(1+\delta)R(w)\|M-M^*\|_F. \tag{I.124}$$

Rearrange by bringing all terms to one side:

$$\|M-M^*\|_F \leq \frac{2(1+\delta)R(w)}{\Big(-\lambda_{\min}\Big(\nabla_M^2\hat{g}(M^*,w)\Big)+1-\delta\Big)}. \tag{I.125}$$

Recall that:

$$R(w) := \frac{2}{h^2 n} \sup_{t\in[0,1]} \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} e^{-\Delta_{ij}(0)^2/h^2}} \sum_{j=1}^{n} \Big\| e^{-\Delta_{ij}(w)^2/h^2}(w_j - w_i) + \Big(e^{-\Delta_{ij}(w)^2/h^2} - e^{-\Delta_{ij}(0)^2/h^2}\Big)\Delta_{ij}(0) \Big\|. \tag{I.126}$$

and:

$$e^{-\Delta_{ij}(w)^2/h^2}(w_j - w_i), \tag{I.127}$$

$$e^{-\Delta_{ij}(w)^2/h^2}(w_j - w_i) = \mathcal{O}\big(e^{-w^2}|w_j - w_i|\big). \tag{I.128}$$

We assume smooth dependence on w based on Assusmption 2.3, then a first-order Taylor expansion (and using Lipschitz properties) gives:

$$e^{-\Delta_{ij}(w)^2/h^2} - e^{-\Delta_{ij}(0)^2/h^2} = \mathcal{O}\Big(e^{-w^2}\frac{\|w_j - w_i\|}{h^2}\Big). \tag{I.129}$$

Multiplying by the bounded $\|\Delta_{ij}(0)\|$ we obtain an additional term whose order is also

$$\mathcal{O}\Big(e^{-w^2}\frac{\|w_j - w_i\|}{h^2}\Big). \tag{I.130}$$

After summing over j (and i) and dividing by n, these estimates lead to

$$R(w) = \mathcal{O}\Big(\frac{1}{h^2}e^{-w^2}\max_{i,j}\|w_j - w_i\|\Big). \tag{I.131}$$

When the differences $\|w_j - w_i\|$ are controlled by a norm $\|w\|$, we may write:

$$R(w) = \mathcal{O}\Big(\frac{\|w\|e^{-w^2}}{h^2}\Big). \tag{I.132}$$

Putting (I.121) and (I.125) together, the error-bound is:

$$\|M - M^*\|_F \le \max\left\{\sqrt{\frac{2}{\lambda_{\min}\left(\nabla^2_M \hat{g}(M^*, w)\right)}}, \frac{2(1+\delta)R(w)}{\Big(-\lambda_{\min}\Big(\nabla^2_M \hat{g}(M^*, w)\Big) + 1 - \delta\Big)}\right\}, \tag{I.133}$$

We then have a approximation term (with respect to w): $\sqrt{\frac{2}{\lambda_{\min}}} = \mathcal{O}(1)$ and a term that depends on w: $\frac{2(1+\delta)R(w)}{-\lambda_{\min}+1-\delta} = \mathcal{O}(R(w)) = \mathcal{O}\left(\frac{\|w\|e^{-w^2}}{h^2}\right)$. Therefore, the final result is:

$$\|M - M^*\|_F = \mathcal{O}\left(\max\left\{1, \frac{\|w\|e^{-w^2}}{h^2}\right\}\right). \tag{I.134}$$

with probability at least $\mathbb{P}(\|w\|_2 \le \epsilon)$

$$\|M - M^*\|_F = \mathcal{O}\left(\max\left\{1, \frac{\epsilon e^{-\epsilon^2}}{h^2}\right\}\right). \tag{I.135}$$

I.10. The Turning Point For the Upper bound

Write the inequality when term 1 of (I.133) is larger than term 2 of (I.133):

$$\frac{2}{\lambda} > \frac{2(1+\delta)R(w)}{1-\delta-\lambda}. \tag{I.136}$$

We now wish to determine when these two terms are of comparable size. That is, when $\mathcal{O}(1) \sim \mathcal{O}\left(\frac{\|w\|e^{-w^2}}{h^2}\right)$. That is $\|w\|e^{-w^2} \sim h^2$. When $\|w\|$ is very small the exponential may be approximated by $e^{-w^2} \sim 1$, so that

$$\|w\|e^{-w^2} \sim \|w\|. \tag{I.137}$$

Then the condition becomes $\|w\| \sim h^2$. When $\|w\|$ is larger the product $\|w\|e^{-w^2}$ achieves a maximum at $\|w\| = \frac{1}{\sqrt{2}}$, since

$$\frac{d}{dw}\big(\|w\|e^{-w^2}\big) = e^{-w^2}(1-2w^2) = 0 \quad \Longrightarrow \quad w^2 = \frac{1}{2}. \tag{I.138}$$

The maximum value is $\frac{1}{\sqrt{2}}e^{-1/2}$. Hence if $h^2 > \frac{1}{\sqrt{2}}e^{-1/2}$, then even the maximum of the w–dependent term does not exceed a constant and the constant term dominates. Thus, summarizing the comparison: For very small $\|w\|$, specifically when $\|w\| \ll h^2$, with probability at least $\mathbb{P}(\|w\|_2 \le \epsilon)$, we have

$$T_2 = \mathcal{O}\Big(\frac{\epsilon e^{-\epsilon^2}}{h^2}\Big) = \mathcal{O}\Big(\frac{\epsilon}{h^2}\Big) \ll \mathcal{O}(1), \tag{I.139}$$

so the error is dominated by the constant term T_1. Conversely, when ϵ and h are such that $\epsilon e^{-\epsilon^2} \sim h^2$, (which for small ϵ roughly means $\epsilon \sim h^2$) the two terms become comparable. In other regimes (for instance if h^2 is very small compared to the maximum value $\frac{1}{\sqrt{2}}e^{-1/2}$) the ϵ–dependent term may become larger than the constant term. Any answer that shows the turning-point is determined by $\epsilon e^{-\epsilon^2} \sim h^2$.

I.11. Continuity Study of New Loss

We now provide the full proof of Theorem 4.2, establishing the stated continuity property under the conditions specified earlier.

Proof. Based on Assumption 2.1,

$$\|\nabla_M \hat{g}(M,w) - \nabla_M \hat{g}(M,0)\| \le \lambda\|w\| \tag{I.140}$$

So we have:

$$\lambda = \sup_{\xi\in[0,w]} \left\| \frac{d}{dw}\nabla_M \hat{g}(M,\xi) \right\|. \tag{I.141}$$

Differentiating $\nabla_M \hat{g}(M,w)$ with respect to w, we encounters two types of contributions: $\exp\Big(-\frac{(z_j-z_i)^2}{h^2}\Big)$ introduces the chain-rule factor:

$$\frac{d}{dw}\exp\Big(-\frac{(z_j-z_i)^2}{h^2}\Big) = \exp\Big(-\frac{(z_j-z_i)^2}{h^2}\Big)\left(-\frac{2(z_j-z_i)}{h^2}\right)\frac{d}{dw}(z_j-z_i). \tag{I.142}$$

Since $\frac{d}{dw}(z_j - z_i) = I$ (for the appropriate components) the derivative includes a factor proportional to

$$\frac{2}{h^2}(z_j-z_i)\exp\Big(-\frac{(z_j-z_i)^2}{h^2}\Big). \tag{I.143}$$

Under a worst-case scenario the size of $z_j - z_i$ may be bounded by a quantity that depends linearly on $\|w\|$. In many applications one may write for some constant C (which may absorb additional contributions from the data Y and $\mathcal{A}(M)$) derivative includes a factor proportional to

$$\|z_j - z_i\| \le C\|w\|. \tag{I.144}$$

Then one obtains an extra factor of order

$$\frac{2C\|w\|}{h^2}\exp\Big(-\frac{C^2\|w\|^2}{h^2}\Big). \tag{I.145}$$

Meanwhile

$$\|\nabla_M \mathcal{A}(M)\| \le 1+\delta. \tag{I.146}$$

Combining the above equation and Equation (I.141), one obtains that the norm of the mixed derivative satisfies

$$\left\| \frac{d}{dw}\nabla_M \hat{g}(M,w) \right\| \le \frac{8(1+\delta)}{h^4}\|w\|e^{-w^2}, \tag{I.147}$$

Then, with probability at least $\mathbb{P}(\|w\|_2 \le \epsilon)$

$$\left\| \frac{d}{d\epsilon}\nabla_M \hat{g}(M,\epsilon) \right\| \le \frac{8(1+\delta)}{h^4}\epsilon e^{-\epsilon^2}, \tag{I.148}$$

□

J. Proof of δ Condition

J.1. Proof of Lemma 5.1

From [32], we already know that:

$$\hat{g}(M,w) - \hat{g}(M^*,w) - (\delta+\zeta_2 q)\|\hat{M}-M^*\|_F^2 \geq \frac{1-3\delta-3\zeta_2 q}{2}\|\hat{M}-M^*\|_F^2 - \zeta_1 q\|\hat{M}-M^*\|_F. \quad \text{(J.1)}$$

with:

$$G > \sigma_r(1+\delta+\zeta_2 q), \quad \text{(J.2)}$$

and:

$$\|\Delta\|_F = \|M-M^*\|_F \leq \sqrt{\frac{2}{\lambda_{\min}\big(\nabla_M^2 \hat{g}(M^*,w)\big)}\|\hat{g}(M,w)-\hat{g}(M^*,w)\|}, \quad \text{(J.3)}$$

with:

$$L_1 = 2(1+\delta_p). \quad \text{(J.4)}$$

Now we rewrite Equation (I.45),

$$\lambda_{\min}\big(\nabla_M^2 \hat{g}(M^*,w)\big) \geq c \quad \text{with} \quad c = \frac{2}{G_{\min}h^2}\left(L_1^2\left(1+\frac{2B^2}{h^2}\right)+BL_2\right) + \frac{4}{G_{\min}^2 h^4}B^2L_1^2, \quad \text{(J.5)}$$

Combining Equation (J.1) and (J.3), we observe that:

$$(\delta+\zeta_2 q) + \frac{1-3\delta-3\zeta_2 q}{2} = \frac{2(\delta+\zeta_2 q)+\big(1-3\delta-3\zeta_2 q\big)}{2} = \frac{1-\delta-\zeta_2 q}{2}. \quad \text{(J.6)}$$

Thus:

$$\hat{g}(M,w) - \hat{g}(M^*,w) \geq \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_1 q\|M-M^*\|_F. \quad \text{(J.7)}$$

Squaring both sides of Equation (J.3), Now use the lower bound for $\hat{g}(M,w)-\hat{g}(M^*,w)$. Inserting it in the inequality above yields

$$\|M-M^*\|_F^2 \leq \frac{2}{\lambda_{\min}\Big(\nabla_M^2 \hat{g}(M^*,w)\Big)}\left[\frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_1 q\|M-M^*\|_F\right]. \quad \text{(J.8)}$$

Multiply both sides by $\lambda_{\min}(\nabla_M^2 \hat{g}(M^*,w))$, Bring the quadratic term to the left-side to find

$$\|M-M^*\|_F \leq \frac{2\zeta_1 q}{\lambda_{\min}\Big(\nabla_M^2 \hat{g}(M^*,w)\Big) - (1-\delta-\zeta_2 q)}. \quad \text{(J.9)}$$

Combining the two results Equation (J.7) and (J.9), in full detail,

$$\|M-M^*\|_F \leq \frac{2\zeta_1 q}{\frac{2}{G_{\min}h^2}\Big(L_1^2\Big(1+\frac{2B^2}{h^2}\Big)+BL_2\Big) + \frac{4}{G_{\min}^2h^4}B^2L_1^2 - \big(1-\delta-\zeta_2 q\big)}. \quad \text{(J.10)}$$

so we have:

$$\|M-M^*\|_F \leq \frac{2\zeta_1 q}{\underbrace{\frac{2}{G_{\min}h^2}\Big[4(1+\delta)^2\Big(1+\frac{2B^2}{h^2}\Big)+BL_2\Big] + \frac{16B^2(1+\delta)^2}{G_{\min}^2h^4}}_{A} - \Big(1-\delta-\zeta_2 q\Big)}, \quad \text{(J.11)}$$

that is,

$$\|M-M^*\|_F \leq \frac{2\zeta_1 q}{A - \Big(1-\delta-\zeta_2 q\Big)}. \quad \text{(J.12)}$$

Solving for q we have:

$$\sigma_r\Big(1+\delta+\zeta_2 q\Big) < G \quad \Longrightarrow \quad \zeta_2 q < \frac{G}{\sigma_r} - (1+\delta). \quad \text{(J.13)}$$

Thus we may choose:

$$q = \frac{1}{\zeta_2}\Big[\frac{G}{\sigma_r} - (1+\delta)\Big], \tag{J.14}$$

which is the largest allowable choice given the condition (or an upper bound on q). Substitute this choice (J.14) for q into the original bound Equation (J.11). Thus the overall bound becomes:

$$\|M - M^*\|_F \leq \frac{2\zeta_1}{\zeta_2}\frac{\frac{G}{\sigma_r} - (1+\delta)}{A + \frac{G}{\sigma_r} - 2}, \tag{J.15}$$

with:

$$A = \frac{2}{G_{\min}h^2}\left[4(1+\delta)^2\Big(1 + \frac{2B^2}{h^2}\Big) + BL_2\right] + \frac{16B^2(1+\delta)^2}{G_{\min}^2 h^4}. \tag{J.16}$$

Since:

$$\frac{2\zeta_1}{\zeta_2}\frac{\frac{G}{\sigma_r} - (1+\delta)}{A + \frac{G}{\sigma_r} - 2} \geq 0, \tag{J.17}$$

so:

$$0 < \delta < \sqrt{\frac{2 - \frac{G}{\sigma_r} - \frac{2BL_2}{G_{\min}h^2}}{\frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2} + \frac{16B^2}{G_{\min}^2h^4}}} - 1. \tag{J.18}$$

We set:

$$\frac{2B^2}{h^2} = G_{\min} \quad \Longrightarrow \quad h^2 = \frac{2B^2}{G_{\min}}. \tag{J.19}$$

The denominator of (J.18) is:

$$D(h) = \frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2} + \frac{16B^2}{G_{\min}^2h^4}. \tag{J.20}$$

With Equation (J.19) we have

$$\frac{2B^2}{h^2} = G_{\min} \quad \text{and} \quad h^4 = \left(\frac{2B^2}{G_{\min}}\right)^2 = \frac{4B^4}{G_{\min}^2}. \tag{J.21}$$

The numerator is:

$$N(h) = 2 - \frac{G}{\sigma_r} - \frac{2BL_2}{G_{\min}h^2}. \tag{J.22}$$

Again, using Equation (J.19) we have

$$\frac{2BL_2}{G_{\min}h^2} = \frac{2BL_2}{G_{\min}\left(\frac{2B^2}{G_{\min}}\right)} = \frac{2BL_2}{2B^2} = \frac{L_2}{B}. \tag{J.23}$$

Thus,

$$N(h) = 2 - \frac{G}{\sigma_r} - \frac{L_2}{B}. \tag{J.24}$$

Finally:

$$\begin{aligned}\delta &= \sqrt{\frac{N(h)}{D(h)}} - 1 \\ &= \sqrt{\frac{2 - \frac{G}{\sigma_r} - \frac{L_2}{B}}{\frac{4(G_{\min}+2)}{B^2}}} - 1 \\ &= \sqrt{\frac{B^2}{4\left(G_{\min}+2\right)}\Big(2 - \frac{G}{\sigma_r} - \frac{L_2}{B}\Big)} - 1.\end{aligned} \tag{J.25}$$

Thus, by choosing:

$$h = \frac{\sqrt{2}B}{\sqrt{G_{\min}}} \tag{J.26}$$

we specifically have:

$$\delta = \sqrt{\frac{B^2}{4(G_{\min}+2)}\left(2 - \frac{G}{\sigma_r} - \frac{L_2}{B}\right)} - 1. \tag{J.27}$$

In many applications the nonnegative parameters G/σ_r and L_2/B are present in a subtractive term. In the worst-case (largest) scenario the subtraction is minimized, that is, one may assume:

$$\frac{G}{\sigma_r} = 0 \quad \text{and} \quad \frac{L_2}{B} = 0. \tag{J.28}$$

Then we have

$$2 - \frac{G}{\sigma_r} - \frac{L_2}{B} \leq 2. \tag{J.29}$$

Substituting into the expression gives:

$$\delta \leq \sqrt{\frac{B^2}{4(G_{\min}+2)} \cdot 2} - 1 = \sqrt{\frac{2B^2}{4(G_{\min}+2)}} - 1 = \sqrt{\frac{B^2}{2(G_{\min}+2)}} - 1, \tag{J.30}$$

which is:

$$\delta \leq \frac{B}{\sqrt{2(G_{\min}+2)}} - 1. \tag{J.31}$$

J.2. δ Result With Regard to $\|w\|$

Lemma J.1 (Upper Bound on δ under Bounded Noise and Kernel Smoothing)**.** *Let $\hat{g}(M, w)$ be a smoothed loss function involving kernel weights depending on the residuals, and suppose: the residuals satisfy $|z_{ij}(M^*)| \leq B$ for some $B > 0$, the kernel weights involve Gaussian-type terms $\exp(-z^2/h^2)$, the noise vector w satisfies $\|w\| \leq \epsilon$ with high probability (e.g., at least $\mathbb{P}(\|w\| \leq \epsilon)$), the Hessian Lipschitz constant is L_2, and the spectral condition $\sigma_r(1 + \delta + \zeta_2 q) < G$ holds for some constants $\sigma_r, \zeta_2, q, G > 0$. Then for any bandwidth $h > 0$, the parameter δ satisfies the upper bound*

$$\delta < \sqrt{\frac{h^4\left(2 - \frac{G}{\sigma_r} - \frac{2\epsilon L_2 \exp\left(\frac{\epsilon^2}{h^2}\right)}{h^2}\right)}{8\exp\left(\frac{\epsilon^2}{h^2}\right)\left[h^2 + 2\epsilon^2 + 2\epsilon^2 \exp\left(\frac{\epsilon^2}{h^2}\right)\right]}} - 1, \tag{J.32}$$

with probability at least $\mathbb{P}(\|w\| \leq \epsilon)$.

The proof of the theorem is provided in Section J.3.

Lemma J.1 provides an explicit upper bound on the parameter δ in terms of the noise norm $\|w\| \leq \epsilon$ and the kernel bandwidth h, showing how the interaction between noise amplitude and kernel smoothing affects stability; specifically, δ decreases as ϵ becomes small, and the exponential terms in the bound reflect the noise-suppressing effect of the kernel, which ensures that the bound holds with high probability whenever $\|w\|$ is sufficiently controlled.

J.3. Proof of Lemma J.1

We already have:

$$\delta \leq \frac{B}{\sqrt{2\left(\exp\left(-\frac{B^2}{h^2}\right) + 2\right)}} - 1. \tag{J.33}$$

which implies:

$$\delta \leq \frac{1}{h^2}\left(\frac{B}{\sqrt{2(G_{\min}+2)}}-1\right). \tag{J.34}$$

with:

$$h^2 = \frac{2B^2}{G_{\min}}, \tag{J.35}$$

Thus,

$$\delta \leq \frac{1}{h^2}\left(\frac{B}{\frac{2\sqrt{B^2+h^2}}{h}}-1\right) = \frac{1}{h^2}\left(\frac{Bh}{2\sqrt{B^2+h^2}}-1\right). \tag{J.36}$$

So we have:

$$\delta \leq \frac{1}{h^2}\left(\frac{\epsilon}{\frac{2\sqrt{\epsilon^2+h^2}}{h}}-1\right) = \frac{1}{h^2}\left(\frac{\epsilon h}{2\sqrt{\epsilon^2+h^2}}-1\right). \tag{J.37}$$

with at least $\mathbb{P}(\|w\| \leq \epsilon)$. We then restart with the original form (J.18):

$$0 < \delta < \sqrt{\frac{2-\frac{G}{\sigma_r}-\frac{2BL_2}{G_{\min}h^2}}{\frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2}+\frac{16B^2}{G_{\min}^2h^4}}}-1. \tag{J.38}$$

Since $G_{\min}$ appears only in the denominators, replacing $G_{\min}$ by its lower bound (which is the worst-case scenario) yields an upper bound for δ.In the numerator we have

$$\frac{2BL_2}{G_{\min}h^2} \leq \frac{2BL_2\exp\left(\frac{B^2}{h^2}\right)}{h^2}. \tag{J.39}$$

In the denominator, the two terms become:

$$\frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2} \leq \frac{8\left(1+\frac{2B^2}{h^2}\right)\exp\left(\frac{B^2}{h^2}\right)}{h^2}, \tag{J.40}$$

Thus, the bound for δ becomes:

$$0 < \delta < \sqrt{\frac{2-\frac{G}{\sigma_r}-\frac{2BL_2\exp\left(\frac{B^2}{h^2}\right)}{h^2}}{\frac{8\left(1+\frac{2B^2}{h^2}\right)\exp\left(\frac{B^2}{h^2}\right)}{h^2}+\frac{16B^2\exp\left(\frac{2B^2}{h^2}\right)}{h^4}}}-1. \tag{J.41}$$

Then the denominator of Equation (J.41) becomes:

$$D = \frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2}+\frac{16B^2}{G_{\min}^2h^4} \leq \frac{8\left(1+\frac{2B^2}{h^2}\right)\exp\left(\frac{B^2}{h^2}\right)}{h^2}+\frac{16B^2\exp\left(\frac{2B^2}{h^2}\right)}{h^4}. \tag{J.42}$$

$$\frac{8\left(1+\frac{2B^2}{h^2}\right)\exp\left(\frac{B^2}{h^2}\right)}{h^2} = \frac{8\exp\left(\frac{B^2}{h^2}\right)}{h^4}\left(h^2+2B^2\right), \tag{J.43}$$

so that:

$$D \leq \frac{1}{h^4}\left[8\exp\left(\frac{B^2}{h^2}\right)(h^2+2B^2)+16B^2\exp\left(\frac{2B^2}{h^2}\right)\right]. \tag{J.44}$$

We may factor $\exp\left(\frac{B^2}{h^2}\right)$ from the terms in brackets:

$$D \leq \frac{8\exp\left(\frac{B^2}{h^2}\right)}{h^4}\left[h^2+2B^2+2B^2\exp\left(\frac{B^2}{h^2}\right)\right]. \tag{J.45}$$

Thus the overall bound (J.41) becomes:

$$\delta < \sqrt{\frac{N}{D}} - 1 = \sqrt{\frac{h^4\Big(2 - \frac{G}{\sigma_r} - \frac{2BL_2\exp\Big(\frac{B^2}{h^2}\Big)}{h^2}\Big)}{8\exp\Big(\frac{B^2}{h^2}\Big)\Big[h^2 + 2B^2 + 2B^2\exp\Big(\frac{B^2}{h^2}\Big)\Big]}} - 1. \tag{J.46}$$

So we have:

$$\delta < \sqrt{\frac{N}{D}} - 1 = \sqrt{\frac{h^4\Big(2 - \frac{G}{\sigma_r} - \frac{2\epsilon L_2\exp\Big(\frac{\epsilon^2}{h^2}\Big)}{h^2}\Big)}{8\exp\Big(\frac{\epsilon^2}{h^2}\Big)\Big[h^2 + 2\epsilon^2 + 2\epsilon^2\exp\Big(\frac{\epsilon^2}{h^2}\Big)\Big]}} - 1. \tag{J.47}$$

with at least $\mathbb{P}(\|w\| \leq \epsilon)$.

K. Proof of Theorem 6.1

Proof. With the loss function in (2.3) and the simplified notations in (C.1) and (C.2), we have:

$$\langle \nabla_M \hat{g}(M^*, w), \hat{X}U^\top + U\hat{X}^\top\rangle = -\int_0^1 \nabla^2_M \hat{g}(tM^* + (1-t)\hat{M})(\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top)dt. \tag{K.1}$$

$$\nabla_M \hat{g}(M, w) = -\frac{1}{n}\sum_{i=1}^n \frac{1}{G_i(M)}\nabla_M G_i(M). \tag{K.2}$$

Thus, at $M = M^*$ the inner product:

$$\langle \nabla_M \hat{g}(M^*, w), \hat{X}U^\top + U\hat{X}^\top\rangle = -\frac{1}{n}\sum_{i=1}^n \frac{1}{G_i(M^*)}\Big\langle \nabla_M G_i(M^*), \hat{X}U^\top + U\hat{X}^\top\Big\rangle. \tag{K.3}$$

On the other hand, a standard Taylor expansion applied to the function:

$$h(V) = \langle \nabla_M \hat{g}(V, w), \hat{X}U^\top + U\hat{X}^\top\rangle, \tag{K.4}$$

provides

$$h(\hat{M}) - h(M^*) = \int_0^1 \Big\langle \nabla h\Big(tM^* + (1-t)\hat{M}\Big), \hat{M} - M^*\Big\rangle dt. \tag{K.5}$$

$$h(\hat{M}) = \langle \nabla_M \hat{g}(\hat{M}, w), \hat{X}U^\top + U\hat{X}^\top\rangle = 0, \tag{K.6}$$

we obtain:

$$\langle \nabla_M \hat{g}(M^*, w), \hat{X}U^\top + U\hat{X}^\top\rangle = -\int_0^1 \Big\langle \nabla h\Big(tM^* + (1-t)\hat{M}\Big), \hat{M} - M^*\Big\rangle dt. \tag{K.7}$$

But note that:

$$\nabla h(V) = \nabla^2_M \hat{g}(V, w)(\cdot, \hat{X}U^\top + U\hat{X}^\top), \tag{K.8}$$

so that Equation (K.1) becomes:

$$\langle \nabla_M \hat{g}(M^*, w), \hat{X}U^\top + U\hat{X}^\top\rangle = -\int_0^1 \nabla^2_M \hat{g}\Big(tM^* + (1-t)\hat{M}\Big)\Big(\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top\Big)dt. \tag{K.9}$$

So we have:

$$\nabla^2_M \hat{g}(M) = \frac{1}{n}\sum_{i=1}^n \left[\frac{1}{G_i(M)^2}\nabla_M G_i(M) \otimes \nabla_M G_i(M) - \frac{1}{G_i(M)}\nabla^2_M G_i(M)\right], \tag{K.10}$$

with:

$$G_i(M) = \frac{1}{n}\sum_{j=1}^n \exp\left(-\frac{u_{ij}(M)^2}{h^2}\right), \tag{K.11}$$

and:

$$\nabla_M G_i(M) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)\left(-\frac{2u_{ij}(M)}{h^2}\right)\Big[-\nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i\Big],$$
$$= \frac{1}{n}\sum_{j=1}^{n} \exp\left(-\frac{u_{ij}(M)^2}{h^2}\right)\left[\left(\frac{4u_{ij}(M)^2}{h^4} - \frac{2}{h^2}\right)\nabla_M u_{ij}(M)\otimes\nabla_M u_{ij}(M) - \frac{2u_{ij}(M)}{h^2}\nabla_M^2 u_{ij}(M)\right], \tag{K.12}$$

and where:

$$u_{ij}(M) = \Big[(Y_j + w_j - \mathcal{A}(M)_j) - (Y_i + w_i - \mathcal{A}(M)_i)\Big], \tag{K.13}$$

and

$$\nabla_M u_{ij}(M) = -\nabla_M \mathcal{A}(M)_j + \nabla_M \mathcal{A}(M)_i. \tag{K.14}$$

Now, introduce the convex path:

$$M_t = tM^* + (1-t)\hat{M}, \quad t \in [0,1], \tag{K.15}$$

and denote for brevity

$$Q = \hat{X}U^\top + U\hat{X}^\top, \qquad H = \hat{M} - M^*. \tag{K.16}$$

Then the Hessian's bilinear form of (K.10) at M_t applied to (H, Q) is:

$$\nabla_M^2 \hat{g}(M_t)(H,Q) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{G_i(M_t)^2}\langle\nabla_M G_i(M_t), H\rangle\langle\nabla_M G_i(M_t), Q\rangle - \frac{1}{G_i(M_t)}\nabla_M^2 G_i(M_t)(H,Q)\Big]. \tag{K.17}$$

We notive that:

$$-\int_0^1 \nabla_M^2 \hat{g}\Big(tM^* + (1-t)\hat{M}\Big)\Big(\hat{M} - M^*,\, Q\Big)dt$$
$$= -\int_0^1 \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{G_i(M_t)^2}\langle\nabla_M G_i(M_t), \hat{M} - M^*\rangle\langle\nabla_M G_i(M_t), Q\rangle - \frac{1}{G_i(M_t)}\nabla_M^2 G_i(M_t)\Big(\hat{M} - M^*, Q\Big)\right\}dt, \tag{K.18}$$

with all quantities evaluated at $M_t = tM^* + (1-t)\hat{M}$. We take the simplified notations (K.15) and (K.16) into (K.18), we have:

$$H = \hat{M} - M^*, \quad Q = \hat{X}U^\top + U\hat{X}^\top, \quad \text{and} \quad M_t = tM^* + (1-t)\hat{M}. \tag{K.19}$$

Then the integral becomes:

$$-\int_0^1 \nabla_M^2 \hat{g}(M_t)(H,Q)dt = -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big[\frac{\langle\nabla_M G_i(M_t), H\rangle\langle\nabla_M G_i(M_t), Q\rangle}{G_i(M_t)^2} - \frac{\nabla_M^2 G_i(M_t)(H,Q)}{G_i(M_t)}\Big]dt, \tag{K.20}$$

In a compact form, we write. This is the simplified integrated form of the second-order term.

$$\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u = -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\{\frac{\langle\nabla_M G_i(tM^* + (1-t)\hat{M}), \hat{M} - M^*\rangle\langle\nabla_M G_i(tM^* + (1-t)\hat{M}), \hat{X}U^\top + U\hat{X}^\top\rangle}{\big[G_i(tM^* + (1-t)\hat{M})\big]^2}\right.$$
$$\left. - \frac{\nabla_M^2 G_i(tM^* + (1-t)\hat{M})(\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top)}{G_i(tM^* + (1-t)\hat{M})}\right\}dt. \tag{K.21}$$

$$\Big\|\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u\Big\| = \Big\|-\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\{\frac{\langle\nabla_M G_i(M_t), \hat{M} - M^*\rangle\langle\nabla_M G_i(M_t), \hat{X}U^\top + U\hat{X}^\top\rangle}{G_i(M_t)^2}\right.$$
$$\left. - \frac{\nabla_M^2 G_i(M_t)(\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top)}{G_i(M_t)}\right\}dt\Big\|, \tag{K.22}$$

To solve this integration, in a first step we take the absolute value inside the integral and sum (using the triangle inequality) to obtain

$$\left\|\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u\right\| \le \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left[\frac{\left\|\langle \nabla_M G_i(M_t), \hat{M}-M^*\rangle \langle \nabla_M G_i(M_t), \hat{X}U^\top + U\hat{X}^\top\rangle\right\|}{G_i(M_t)^2} + \frac{\left\|\nabla_M^2 G_i(M_t)(\hat{M}-M^*, \hat{X}U^\top + U\hat{X}^\top)\right\|}{G_i(M_t)}\right] dt. \tag{K.23}$$

Then we write:

$$\left\|\mathbf{e}^\top H\hat{X}u\right\| \le \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\|\frac{\langle \nabla_M G_i(M_t), H\rangle \langle \nabla_M G_i(M_t), Q\rangle}{G_i(M_t)^2}\right\| dt + \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\|\frac{\nabla_M^2 G_i(M_t)(H,Q)}{G_i(M_t)}\right\| dt. \tag{K.24}$$

A short calculation on Equation (K.11) shows that for each i and j we have:

$$\nabla_M \exp\Big(-\frac{u_{ij}(M)^2}{h^2}\Big) = -\frac{2u_{ij}(M)}{h^2}\exp\Big(-\frac{u_{ij}(M)^2}{h^2}\Big)\nabla_M u_{ij}(M). \tag{K.25}$$

Thus:

$$\nabla_M G_i(M_t) = \frac{1}{n}\sum_{j=1}^{n}\Big[-\frac{2u_{ij}(M_t)}{h^2}\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big)\nabla_M u_{ij}(M_t)\Big]. \tag{K.26}$$

Assume that the derivatives of u_{ij} are uniformly bounded (say, $\|\nabla_M u_{ij}(M_t)\| \le L_u$ for all i, j, t). Then:

$$\|\nabla_M G_i(M_t)\| \le \frac{1}{n}\sum_{j=1}^{n}\frac{2|u_{ij}(M_t)|}{h^2}\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big)L_u. \tag{K.27}$$

Notice that the noise w enters through the differences $(w_j - w_i)$. In the following we denote:

$$\delta := \|w\|_\infty. \tag{K.28}$$

Using the triangle inequality, we separately bound the two integrated terms:

$$\left\|\mathbf{e}^\top \mathbf{H}\hat{X}u\right\| \le \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\{\left\|\frac{\langle \nabla_M G_i(M_t), H\rangle \langle \nabla_M G_i(M_t), Q\rangle}{G_i(M_t)^2}\right\| + \left\|\frac{\nabla_M^2 G_i(M_t)(H,Q)}{G_i(M_t)}\right\|\right\} dt. \tag{K.29}$$

It is natural to expect that the $\nabla_M G_i$ and $\nabla_M^2 G_i$ terms inherit the exponential from

$$G_i(M_t) = \frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big). \tag{K.30}$$

A direct differentiation of Equation (K.30) shows that:

$$\nabla_M\Big(\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big)\Big) = -\frac{2u_{ij}(M_t)}{h^2}\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big)\nabla_M u_{ij}(M_t). \tag{K.31}$$

Taking the norm we may write:

$$\|\nabla_M \exp(-u_{ij}(M_t)^2/h^2)\|_F \le \frac{2|u_{ij}(M_t)|}{h^2}\exp\Big(-\frac{u_{ij}(M_t)^2}{h^2}\Big)\|\nabla_M u_{ij}(M_t)\|_F. \tag{K.32}$$

Since the noise enters u_{ij} as a difference (i.e. $w_j - w_i$), we have the bound:

$$|u_{ij}(M_t)| \le \delta, \tag{K.33}$$

so that for some constant L_1 (which also absorbs bounds on $\|\nabla_M u_{ij}(M_t)\|$) we may write:

$$\|\nabla_M G_i(M_t)\| \le \frac{L_1\delta}{h^2}\exp\Big(-\frac{u_{ij}^{\min}(M_t)^2}{h^2}\Big), \tag{K.34}$$

where we set:

$$u_{ij}^{\min}(M_t)^2 := \min_{1\le j\le n} u_{ij}(M_t)^2. \tag{K.35}$$

Similarly, one may bound the Hessian by:

$$\|\nabla_M^2 G_i(M_t)\| \le \frac{L_2\delta}{h^2}\exp\Big(-\frac{u_{ij}^{\min}(M_t)^2}{h^2}\Big), \tag{K.36}$$

for a constant $L_2 > 0$. In addition, we assume that the denominator satisfies $G_i(M_t) \ge G_{\min} > 0$. Substitute the above bounds into the two terms:

$$\begin{aligned}\left\|\frac{\langle\nabla_M G_i(M_t), H\rangle\langle\nabla_M G_i(M_t), Q\rangle}{G_i(M_t)^2}\right\| &\le \frac{\|\nabla_M G_i(M_t)\|^2}{G_i(M_t)^2}\|H\|\|Q\| \\ &\le \frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{ij}^{\min}(M_t)^2}{h^2}\Big)\|H\|\|Q\|.\end{aligned} \tag{K.37}$$

Similarly,

$$\left\|\frac{\nabla_M^2 G_i(M_t)(H,Q)}{G_i(M_t)}\right\| \le \frac{\|\nabla_M^2 G_i(M_t)\|}{G_i(M_t)}\|H\|\|Q\| \le \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{ij}^{\min}(M_t)^2}{h^2}\Big)\|H\|\|Q\|. \tag{K.38}$$

Thus, after summing over i and integrating over t (which introduces only constant factors), we obtain the bound:

$$\left\|\mathbf{e}^\top \mathbf{H}\hat{X}u\right\| \le \left(\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right)\|H\|\|Q\|, \tag{K.39}$$

where we set, for simplicity,

$$u_{\min}^2 = \min_{i,j,t} u_{ij}(M_t)^2. \tag{K.40}$$

$$\left\|\mathbf{e}^\top \mathbf{H}\hat{X}u\right\| \le C\frac{\|w\|_\infty}{h^2}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\|\hat{M}-M^*\|\|\hat{X}U^\top + U\hat{X}^\top\|. \tag{K.41}$$

We wish to maximize η subject to:

$$\begin{aligned}&\|\hat{\mathbf{X}}^\top\hat{\mathbf{H}}\mathbf{e}\| \le \left(\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right)\|H\|\|Q\|, \\ &\eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}.\end{aligned} \tag{K.42}$$

Notice that the eigenvalue constraint forces all eigenvalues of $\hat{\mathbf{H}}$ to lie between η and 1. Hence, in a best-case scenario we may choose:

$$\hat{\mathbf{H}} = I_{n^2}, \tag{K.43}$$

which implies that

$$\eta \le 1. \tag{K.44}$$

Plugging $\hat{\mathbf{H}} = I_{n^2}$ into the first constraint (K.42) gives:

$$\|\hat{\mathbf{X}}^\top\mathbf{e}\| \le \left(\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right)\|H\|\|Q\|. \tag{K.45}$$

In general, if this inequality is not satisfied, one may scale down $\hat{\mathbf{H}}$ so that the effective gain is reduced. In particular, if we set:

$$\hat{\mathbf{H}} = \eta I_{n^2} \quad (\text{with } \eta \le 1), \tag{K.46}$$

then

$$\|\hat{\mathbf{X}}^\top\hat{\mathbf{H}}\mathbf{e}\| = \eta\|\hat{\mathbf{X}}^\top\mathbf{e}\|, \tag{K.47}$$

and the constraint becomes:

$$\eta\|\hat{\mathbf{X}}^\top\mathbf{e}\| \le \left(\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right)\|H\|\|Q\|. \tag{K.48}$$

That is,

$$\eta \leq \frac{\|H\|\|Q\|}{\|\hat{\mathbf{X}}^\top \mathbf{e}\|} \left[\frac{L_1^2 \delta^2}{h^4 G_{\min}^2} \exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2 \delta}{h^2 G_{\min}} \exp\Big(-\frac{u_{\min}^2}{h^2}\Big) \right]. \tag{K.49}$$

We can define:

$$C := \frac{L_1^2 \delta^2}{h^4 G_{\min}^2} \exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2 \delta}{h^2 G_{\min}} \exp\Big(-\frac{u_{\min}^2}{h^2}\Big) \tag{K.50}$$

(with also the extra factors $\|H\|\|Q\|$ appearing). For clarity, define the second problem as:

$$\begin{aligned} \max_{\eta, \hat{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}} \mathbf{e}\| \leq C\|H\|\|Q\|, \\ & \eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}. \end{aligned} \tag{K.51}$$

A brief outline of the solution is : $y = \hat{\mathbf{H}}\mathbf{e}$. Since $\hat{H} \succeq \eta I$ (by $\eta I \preceq \hat{H}$), we have a gain in the sense that $\|y\| = \|\hat{H}e\| \geq \eta\|e\|$. (When e is normalized $\|e\| = 1$ the inequality is $\|y\| \geq \eta$.)

$$\|\hat{\mathbf{X}} y\|^2 \geq 2\lambda_{r^*}(\hat{X}\hat{X}^\top)\|y\|^2, \tag{K.52}$$

the bound on the left (when applied to $y = \hat{H}e$) yields:

$$\|\hat{\mathbf{X}}^\top \hat{\mathbf{H}} e\|^2 = \|\hat{\mathbf{X}} y\|^2 \geq 2\lambda_{r^*}(\hat{X}\hat{X}^\top)\|y\|^2. \tag{K.53}$$

$$\|\hat{\mathbf{X}}^\top \hat{\mathbf{H}} e\| \leq C\|H\|\|Q\|. \tag{K.54}$$

Thus, combining with the lower bound (K.52), we find:

$$2\lambda_{r^*}(\hat{X}\hat{X}^\top)\|y\|^2 \leq \big(C\|H\|\|Q\|\big)^2. \tag{K.55}$$

Using $\|y\| \geq \eta\|e\|$ and, if we normalize so that $\|e\| = 1$, we obtain $2\lambda_{r^*}(\hat{X}\hat{X}^\top)\eta^2 \leq \big(C\|H\|\|Q\|\big)^2$. That is, solving for η we have

$$\eta \leq \frac{C\|H\|\|Q\|}{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}. \tag{K.56}$$

Since the optimization is to maximize η subject to the constraint, the best (largest) value one may take is at most

$$\eta^* = \frac{C\|H\|\|Q\|}{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}, \tag{K.57}$$

or, writing C explicitly,

$$\eta^* = \frac{\|H\|\|Q\|}{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}} \left[\frac{L_1^2 \delta^2}{h^4 G_{\min}^2} \exp\left(-\frac{2u_{\min}^2}{h^2}\right) + \frac{L_2 \delta}{h^2 G_{\min}} \exp\left(-\frac{u_{\min}^2}{h^2}\right) \right]. \tag{K.58}$$

This is the solution to the second problem. We assumed here that $\|e\| = 1$ (or otherwise, the factor $\|e\|$ remains explicitly in the bound). The derivation uses the lower bound that came from the structure of the matrix $\hat{X}\hat{X}^\top$ and the fact that the eigenvalues of $\hat{H}$ satisfy $\lambda_i(\hat{H}) \geq \eta$. Thus, the optimal achievable η is given by the boxed expression above.

We now have: $y = \hat{\mathbf{H}}\mathbf{e}$ Since: $\eta I_{n^2} \preceq \hat{\mathbf{H}}$, it follows that for each vector (in particular, for the fixed $\mathbf{e}$) we have:

$$\|y\|^2 = \|\hat{\mathbf{H}}\mathbf{e}\|^2 \geq \eta\|\mathbf{e}\|^2. \tag{K.59}$$

Often the vector $\mathbf{e}$ is taken to be a unit-vector; hence from now on we assume $\|\mathbf{e}\| = 1$ so that $\|y\|^2 \geq \eta$. In the baseline problem it was shown that for any $y \in \mathbb{R}^{n^2}$, one has:

$$\|\hat{\mathbf{X}} y\|^2 \geq 2\lambda_{r^*}(\hat{X}\hat{X}^\top)\|y\|^2. \tag{K.60}$$

Hence, in our setting,

$$\|\hat{\mathbf{X}} y\|^2 \geq 2\lambda_{r^*}(\hat{X}\hat{X}^\top)\eta. \tag{K.61}$$

On the other hand the constraint (K.61) implies that:

$$\|\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\mathbf{e}\| = \|\hat{\mathbf{X}}^\top y\| \le C\|\hat{M} - M^*\|_F \|\hat{X}\|_2 \|w\|. \tag{K.62}$$

Since (by consistency of norms) one may relate the norm $\|\hat{\mathbf{X}}^\top y\|$ to $\|\hat{\mathbf{X}} y\|$ (when $\hat{X}$ has full column-rank) or simply use the fact that the bound holds on the action of y through $\hat{X}$, we square the inequality to obtain:

$$\|\hat{\mathbf{X}} y\|^2 \le \Big(C\|\hat{M} - M^*\|_F \|\hat{X}\|_2 \|w\|\Big)^2. \tag{K.63}$$

Combining this with the lower bound (K.52) gives

$$2\lambda_{r^*}(\hat{X}\hat{X}^\top)\eta \le \Big(C\|\hat{M} - M^*\|_F \|\hat{X}\|_2 \|w\|\Big)^2. \tag{K.64}$$

Solving for η we deduce that any feasible pair $(\eta, \hat{\mathbf{H}})$ must satisfy:

$$\eta \le \frac{C^2\|\hat{M} - M^*\|_F^2 \|\hat{X}\|_2^2 \|w\|^2}{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}. \tag{K.65}$$

That is, the maximal achievable value is

$$\eta^\star = \frac{C^2\|\hat{M} - M^*\|_F^2 \|\hat{X}\|_2^2 \|w\|^2}{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}. \tag{K.66}$$

(Recall that we assumed $\|\mathbf{e}\| = 1$. Should $\|\mathbf{e}\|$ differ from 1 the bound would include the factor $1/\|\mathbf{e}\|^2$.) Hence, the solution of the given maximization problem is:

$$\eta^\star = \frac{C^2\|\hat{M} - M^*\|_F^2 \|\hat{X}\|_2^2 \|w\|^2}{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}, \text{with} \quad C = 2\left(\frac{\zeta_2^2}{c^2} + \frac{\zeta_3}{c}\right) \tag{K.67}$$

Because the constraint:

$$(1 - \delta - \zeta_2 q)\|M\|_F^2 \le \mathbf{m}^\top \mathbf{H}\mathbf{m} \le (1 + \delta + \zeta_2 q)\|M\|_F^2, \quad \forall M : \operatorname{rank}(M) \le 2r, \tag{K.68}$$

must hold for every matrix M (with $m = \operatorname{vec}(M)$) the operator $\mathbf{H}$ is sandwiched over the set of these low–rank matrices. In other words, if we define $\eta := 1 - \delta - \zeta_2 q$, then we have that:

$$\mathbf{m}^\top \mathbf{H}\mathbf{m} \ge \eta\|M\|_F^2, \quad \forall M : \operatorname{rank}(M) \le 2r. \tag{K.69}$$

This is analogous to requiring $\mathbf{H} \succeq \eta I$ (on the low–rank subspace). In the previous analysis (where our decision variable was η), one shows that if one defines $y = \mathbf{H}\mathbf{e}$, then necessarily

$$\|y\|^2 = \|\mathbf{H}\mathbf{e}\|^2 \ge \eta\|\mathbf{e}\|^2. \tag{K.70}$$

Assuming without loss of generality that $\|\mathbf{e}\| = 1$ we deduce:

$$\|y\|^2 \ge \eta = 1 - \delta - \zeta_2 q. \tag{K.71}$$

$$\|\hat{\mathbf{X}} y\|^2 \ge 2\lambda_{r^*}(\hat{X}\hat{X}^\top)\|y\|^2 \ge 2\lambda_{r^*}(\hat{X}\hat{X}^\top)(1 - \delta - \zeta_2 q). \tag{K.72}$$

The constraint

$$\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| = \|\hat{\mathbf{X}}^\top y\| \le 2\zeta_1 q\|\hat{X}\|_2 \tag{K.73}$$

implies, after squaring,

$$\|\hat{\mathbf{X}} y\|^2 \le \Big(2\zeta_1 q\|\hat{X}\|_2\Big)^2. \tag{K.74}$$

Thus, combining with the lower bound (K.71) and (K.72), we have

$$2\lambda_{r^*}(\hat{X}\hat{X}^\top)(1 - \delta - \zeta_2 q) \le 4\zeta_1^2 q^2 \|\hat{X}\|_2^2. \tag{K.75}$$

We now solve for δ. Rearranging the previous inequality we obtain

$$1 - \delta - \zeta_2 q \le \frac{4\zeta_1^2 q^2 \|\hat{X}\|_2^2}{2\lambda_{r^*}(\hat{X}\hat{X}^\top)} = \frac{2\zeta_1^2 q^2 \|\hat{X}\|_2^2}{\lambda_{r^*}(\hat{X}\hat{X}^\top)}. \tag{K.76}$$

That is,

$$\delta \geq 1 - \zeta_2 q - \frac{2\zeta_1^2 q^2 \|\hat{X}\|_2^2}{\lambda_{r^*}(\hat{X}\hat{X}^\top)}. \tag{K.77}$$

Since we minimize δ it is best to take the smallest choice allowed (and note that we must have $\delta \geq 0$); hence,

$$\delta^\star = \max\left\{0, 1 - \zeta_2 q - \frac{2\zeta_1^2 q^2 \|\hat{X}\|_2^2}{\lambda_{r^*}(\hat{X}\hat{X}^\top)}\right\}. \tag{K.78}$$

Then, we want to use:

$$\begin{aligned} \left\|\lambda_{r^*}(\hat{X}\hat{X}^\top) - \lambda_{r^*}(M^*)\right\| &\leq \|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*), \\ \left\|\lambda_1(\hat{X}\hat{X}^\top) - \lambda_1(M^*)\right\| &\leq \|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*). \end{aligned} \tag{K.79}$$

From

$$\left\|\lambda_{r^*}(\hat{X}\hat{X}^\top) - \lambda_{r^*}(M^*)\right\| \leq \tau\lambda_r(M^*), \tag{K.80}$$

we obtain the lower bound

$$\lambda_{r^*}(\hat{X}\hat{X}^\top) \geq \lambda_{r^*}(M^*) - \tau\lambda_r(M^*). \tag{K.81}$$

Similarly, the second inequality in Equation (K.79) gives the upper bound

$$\lambda_1(\hat{X}\hat{X}^\top) \leq \lambda_1(M^*) + \tau\lambda_r(M^*). \tag{K.82}$$

Thus we have:

$$\eta^\star = \frac{C^2\|\hat{M} - M^*\|_F^2 \|\hat{X}\|_2^2 \|w\|^2}{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}. \tag{K.83}$$

Using the lower bound Equation (K.81), we may further bound

$$\eta^\star \leq \frac{C^2\|\hat{M} - M^*\|_F^2 \|\hat{X}\|_2^2 \|w\|^2}{2\left(\lambda_{r^*}(M^*) - \tau\lambda_r(M^*)\right)}. \tag{K.84}$$

This way the error is expressed in terms of the eigenvalues of M^* and the perturbation level τ. Similarly we have:

$$\eta^* \leq \frac{\|H\|\|Q\|}{\sqrt{2\Big(\lambda_{r^*}(M^*) - \tau\lambda_r(M^*)\Big)}} \left[\frac{L_1^2\delta^2}{h^4 G_{\min}^2} \exp\left(-\frac{2u_{\min}^2}{h^2}\right) + \frac{L_2\delta}{h^2 G_{\min}} \exp\left(-\frac{u_{\min}^2}{h^2}\right)\right]. \tag{K.85}$$

Solve for $\|\hat{M} - M^*\|_F^2$ starting from the definition:

$$\|\hat{M} - M^*\|_F^2 = \frac{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}{C^2\|\hat{X}\|_2^2\|w\|^2}\eta_f^*(\hat{X}). \tag{K.86}$$

We now want to calculate lower bound for $\lambda_{r^*}(\hat{X}\hat{X}^\top)$. By the Wielandt–Hoffman theorem [36] we have:

$$\left|\lambda_{r^*}(\hat{X}\hat{X}^\top) - \lambda_{r^*}(M^*)\right| \leq \|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*). \tag{K.87}$$

This yields

$$\lambda_{r^*}(\hat{X}\hat{X}^\top) \geq \lambda_{r^*}(M^*) - \tau\lambda_r(M^*). \tag{K.88}$$

Replacing the denominator in the Equation (K.86) for $\|\hat{M} - M^*\|_F^2$ gives:

$$\|\hat{M} - M^*\|_F^2 \leq \frac{2\Big(\lambda_{r^*}(M^*) - \tau\lambda_r(M^*)\Big)}{C^2\|\hat{X}\|_2^2\|w\|^2}\eta_f^*(\hat{X}). \tag{K.89}$$

Thus, taking square roots on both sides, we obtain the final upper bound:

$$\|\hat{M} - M^*\|_F \leq \sqrt{\frac{2\Big(\lambda_{r^*}(M^*) - \tau\lambda_r(M^*)\Big)}{C^2\|\hat{X}\|_2^2\|w\|^2}\eta_f^*(\hat{X})}. \tag{K.90}$$

Similarly from Equation (K.85), we have:

$$\|\hat{M} - M^*\|_F \leq \frac{\sqrt{2\Big(\lambda_{r^*}(M^*) - \tau\lambda_r(M^*)\Big)}}{\|Q\|\left[\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\big(-\frac{2u_{\min}^2}{h^2}\big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\big(-\frac{u_{\min}^2}{h^2}\big)\right]}\eta^*. \tag{K.91}$$

This expression provides the desired upper bound for the Frobenius norm of the difference $\hat{M} - M^*$ in terms of the feasibility error $\eta_f^*(\hat{X})$ and the spectral properties of the matrices involved. We will show that under suitable assumptions one may bound

$$\lambda_{r^*}(M^*) - \tau\lambda_r(M^*), \tag{K.92}$$

in terms of the error $\|\hat{M} - M^*\|_F^2$. One common tool is Weyl's inequality [37]. We assume that $\hat{M}$ denotes an estimator of M^* (for instance, $\hat{M} = \hat{X}\hat{X}^\top$) so that the error in the eigenvalues obeys:

$$\big|\lambda_i(\hat{M}) - \lambda_i(M^*)\big| \leq \|\hat{M} - M^*\|_2 \leq \|\hat{M} - M^*\|_F, \quad \text{for each } i. \tag{K.93}$$

In particular, for the index $i = r^*$ we have:

$$\lambda_{r^*}(M^*) - \lambda_{r^*}(\hat{M}) \leq \|\hat{M} - M^*\|_F. \tag{K.94}$$

Now we begin to calculate the upper bound for $\|M - M^*\|_F$: We have:

$$\|\hat{M} - M^*\|_F \leq \min\left\{\sqrt{\frac{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}{C^2\|\hat{X}\|_2^2\|w\|^2}}\eta_f^*(\hat{X}), \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|\left[\frac{L_1^2\delta^2}{h^4 G_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2 G_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right]}\eta^*\right\}. \tag{K.95}$$

For the first part of Equation (K.95), We start with the bound:

$$T_1 \leq \sqrt{\frac{2\eta_f^*(\hat{X})}{C^2\|w\|^2}}, \tag{K.96}$$

and the lower bound from Equation (K.57) to compute for the function approximation error

$$\eta_f^*(\hat{X}) \geq \frac{1 - \delta_f^*(\hat{X}) - \zeta_2 q}{1 + \delta_f^*(\hat{X}) + \zeta_2 q} \geq \frac{1 - \delta - \zeta_2 q}{1 + \delta + \zeta_2 q}. \tag{K.97}$$

Substituting this lower bound (K.97) into the expression for T_1 (noting that a smaller $\eta_f^*(\hat{X})$ gives a larger overall upper bound for the error) we obtain

$$T_1 \leq \sqrt{\frac{2}{C^2\|w\|^2}\cdot\frac{1 - \delta - \zeta_2 q}{1 + \delta + \zeta_2 q}}. \tag{K.98}$$

Since T_1 is an upper bound on $\|M - M^*\|_F$, we then have:

$$\|M - M^*\|_F \leq \sqrt{\frac{2}{C^2\|w\|^2}\cdot\frac{1 - \delta - \zeta_2 q}{1 + \delta + \zeta_2 q}}. \tag{K.99}$$

This is the desired upper bound for $\|M - M^*\|_F$ in terms of δ, ζ_2, q, C, and $\|w\|$. Now we want to solve for the upper bound. Multiply numerator and denominator by $(2\zeta_1 + 3\zeta_2 X)$ to obtain

$$\frac{1 - \delta - \zeta_2 q}{1 + \delta + \zeta_2 q} = \frac{(1 - \delta)(2\zeta_1 + 3\zeta_2 X) - \zeta_2 X(1 - 3\delta)}{(1 + \delta)(2\zeta_1 + 3\zeta_2 X) + \zeta_2 X(1 - 3\delta)}. \tag{K.100}$$

A short calculation of Equation (K.100) shows:

$$(1 - \delta)(2\zeta_1 + 3\zeta_2 X) - \zeta_2 X(1 - 3\delta) = 2\zeta_1(1 - \delta) + 2\zeta_2 X, \tag{K.101}$$

$$(1 + \delta)(2\zeta_1 + 3\zeta_2 X) + \zeta_2 X(1 - 3\delta) = 2\zeta_1(1 + \delta) + 4\zeta_2 X. \tag{K.102}$$

Thus,

$$X^2 = \frac{2}{C^2\|w\|^2} \cdot \frac{\zeta_1(1-\delta)+\zeta_2 X}{\zeta_1(1+\delta)+2\zeta_2 X}. \tag{K.103}$$

Multiplying by $C^2\|w\|^2$ we arrive at the cubic equation in X:

$$2\zeta_2 C^2\|w\|^2 X^3 + \zeta_1(1+\delta)C^2\|w\|^2 X^2 - 2\zeta_2 X - 2\zeta_1(1-\delta) = 0. \tag{K.104}$$

Thus, the Equation (K.99) is given by:

$$\|M-M^*\|_F \le \frac{-\zeta_1(1+\delta)C^2\|w\|^2 + \sqrt{\zeta_1^2(1+\delta)^2C^4\|w\|^4 + 16\zeta_2 C^2\|w\|^2\zeta_1(1-\delta)}}{4\zeta_2}. \tag{K.105}$$

For the other part, We start with the two facts:

$$\|M-M^*\|_F \le \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|}\left[\frac{L_1^2\delta^2}{h^4\Gamma_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2\Gamma_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right] \times \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}. \tag{K.106}$$

Returning to the bound (K.99) for $\|M-M^*\|_F$ and substituting our expression (with $X = \|\hat{M}-M^*\|_F$), we obtain:

$$\begin{aligned}\|M-M^*\|_F \le \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|}&\left[\frac{L_1^2\delta^2}{h^4\Gamma_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2\Gamma_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right] \\ &\cdot \frac{\zeta_1(1-\delta)+\zeta_2\|\hat{M}-M^*\|_F}{\zeta_1(1+\delta)+2\zeta_2\|\hat{M}-M^*\|_F}.\end{aligned} \tag{K.107}$$

We start with the inequality (K.107) (after eliminating q)

$$\|M-M^*\|_F \le B\epsilon^2 \cdot \frac{\zeta_1(1-\delta)+\zeta_2\|\hat{M}-M^*\|_F}{\zeta_1(1+\delta)+2\zeta_2\|\hat{M}-M^*\|_F}, \tag{K.108}$$

where

$$B := \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|}\left[\frac{L_1^2\delta^2}{h^4\Gamma_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2\Gamma_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right]. \tag{K.109}$$

we now denote:

$$y := \|M-M^*\|_F. \tag{K.110}$$

Thus, the inequality (K.107) becomes:

$$y \le B\epsilon^2 \cdot \frac{\zeta_1(1-\delta)+\zeta_2 y}{\zeta_1(1+\delta)+2\zeta_2 y}. \tag{K.111}$$

Our aim is to move all occurrences of y to one side and solve for an upper bound on y. Multiply both sides by the denominator (assuming it is positive)

$$y\Big[\zeta_1(1+\delta)+2\zeta_2 y\Big] \le B\epsilon^2\Big[\zeta_1(1-\delta)+\zeta_2 y\Big]. \tag{K.112}$$

Write the left-side and right-side explicitly, and bring all terms to one side of the inequality:

$$2\zeta_2 y^2 + \Big[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\Big]y - B\epsilon^2\zeta_1(1-\delta) \le 0. \tag{K.113}$$

Solve the quadratic equation and using the quadratic formula:

$$y = \frac{-\Big[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\Big] \pm \sqrt{\Big[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\Big]^2 + 8B\epsilon^2\zeta_1\zeta_2(1-\delta)}}{4\zeta_2}. \tag{K.114}$$

Since $y = \|M-M^*\|_F \ge 0$, we take the positive root. Thus the upper bound of $\|M-M^*\|_F$ is:

$$\|M-M^*\|_F \le \frac{-\Big[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\Big] + \sqrt{\Big[\zeta_1(1+\delta) - B\epsilon^2\zeta_2\Big]^2 + 8B\epsilon^2\zeta_1\zeta_2(1-\delta)}}{4\zeta_2}. \tag{K.115}$$

Recall that:

$$B = \frac{\sqrt{2\lambda_{r^*}(\hat{X}\hat{X}^\top)}}{\|Q\|}\left[\frac{L_1^2\delta^2}{h^4\Gamma_{\min}^2}\exp\Big(-\frac{2u_{\min}^2}{h^2}\Big) + \frac{L_2\delta}{h^2\Gamma_{\min}}\exp\Big(-\frac{u_{\min}^2}{h^2}\Big)\right], \tag{K.116}$$

with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$.

□

L. Convergence Theory of The New Loss Problem

Given that is satisfied, then if this inequality holds:

$$\|XX^\top - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|X_0X_0^\top - M^w\|_F, \tag{L.1}$$

because in [32], we have:

$$\sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|X_0X_0^\top - M^w\|_F \leq C_w^2\sqrt{1-(\delta+\zeta_{2q})^2} - C_wD_r. \tag{L.2}$$

Thus, for the remainder of the proof, we aim to certify that starting from X_0, if we apply the gradient descent algorithm, above will be satisfied every step along this trajectory. In order to do so, we use Taylor's expansion to obtain:

$$\hat{g}(M,w) - \hat{g}(M^w,w) = \frac{[\nabla^2\hat{g}(N,w)](M-M^w, M-M^w)}{2}, \tag{L.3}$$

where N is some convex combination of M and M^w, and $M \in \mathbb{R}^{n\times n}$ is any matrix of rank at most r. In light of the RIP property (2.6) of the function and (2.8), one can write:

$$\frac{1-\delta-\zeta_{2q}}{2}\|M-M^w\|_F^2 \leq \hat{g}(M,w) - \hat{g}(M^w,w) \leq \frac{1+\delta+\zeta_{2q}}{2}\|M-M^w\|_F^2. \tag{L.4}$$

This means that if $M_1, M_2 \in \mathbb{R}^{n\times n}$ are two matrices of rank at most r with $\hat{g}(M_1,w) \leq \hat{g}(M_2,w)$, then:

$$\|M_1 - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|M_2 - M^w\|_F, \tag{L.5}$$

because $\hat{g}(M_1,w) - \hat{g}(M^w,w) \leq \hat{g}(M_2,w) - \hat{g}(M^w,w)$. Thus, one can conclude that $\hat{g}(X_tX_t^\top,w) \leq \hat{g}(X_0X_0^\top,w)\ \forall t$, where X_t denotes the t^{th} step of the gradient descent algorithm starting from X_0.

L.1. For MSE Loss Function

Theorem L.1. *With MSE loss function, the vanilla gradient descent method applied to* (3) *under Assumptions 1–4 converges to $\mathcal{P}_r(M^w)$, the best rank-r approximation of M^w, linearly up to a difference D_r if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < C_w^2(1-\delta-\zeta_{2\epsilon}) - C_w\sqrt{\frac{1-\delta-\zeta_{2\epsilon}}{1+\delta+\zeta_{2\epsilon}}}D_r, \tag{L.6}$$

meaning that vanilla gradient descent will reach a point $\widetilde{M}$ linearly with $\|\widetilde{M} - \mathcal{P}_r(M^w)\|_F \geq D_r$, where:

$$D_r = \|M^w - \mathcal{P}_r(M^w)\|_F, \quad C_w = \sqrt{2(\sqrt{2}-1)}\sigma_r(M^w). \tag{L.7}$$

The linear convergence is also contingent on the fixed step size η satisfying:

$$\eta \leq \left(12\rho^{1/2}\left(C\sqrt{(1-(\delta+\zeta_{2\epsilon}))^2 + \|M^w\|_F}\right)\right)^{-1}. \tag{L.8}$$

The proofs and detailed math are provided in Section M.1. We can summarize as:

$$\eta \leq \left(12\rho^{1/2}C_0\right)^{-1}. \tag{L.9}$$

L.2. Our Kernel Loss Function

Theorem L.2. *With the kernel loss function, the vanilla gradient descent method applied to 2.3 and 2.1 converges to $\mathcal{P}_r(M^w)$, the best rank-r approximation of M^w, linearly up to a difference D_r if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < C_w^2(1-\delta-\zeta_2\epsilon) - C_w\sqrt{\frac{1-\delta-\zeta_2\epsilon}{1+\delta+\zeta_2\epsilon}}D_r, \tag{L.10}$$

meaning that vanilla gradient descent will reach a point $\widetilde{M}$ linearly with $\|\widetilde{M} - \mathcal{P}_r(M^w)\|_F \geq D_r$, where

$$D_r = \|M^w - \mathcal{P}_r(M^w)\|_F, \quad C_w = \sqrt{2(\sqrt{2}-1)}\sigma_r(M^w). \tag{L.11}$$

The linear convergence is also contingent on the fixed step size η satisfying:

$$\eta \leq h^2\left(12\rho^{1/2}\left(C\sqrt{(1-(\delta+\zeta_2\epsilon))^2 + \|M^w\|_F}\right)\right)^{-1}. \tag{L.12}$$

The proofs and detailed math are provided in Section M.2. We can summarize as:

$$\eta \leq h^2\left(12\rho^{1/2}C_0\right)^{-1}. \tag{L.13}$$

In conclusion, the MSE loss shares the same convergence theorem with the new loss. However, the new loss should pay more attention to the η because it should mind the step length of h.

M. Convergence Proof

M.1. Proof of Theorem L.1

Proof. the gradient descent update:

$$M_{t+1} = M_t - \eta\nabla\hat{g}(M_t, w), \tag{M.1}$$

satisfies:

$$\hat{g}(M_{t+1}, w) \leq \hat{g}(M_t, w) \quad \text{for all } t \geq 0. \tag{M.2}$$

We further assume that the gradient $\nabla\hat{g}(M, w)$ is Lipschitz continuous with constant L. In our setting [32] shows that:

$$L \leq 12\rho\sqrt{r}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right), \tag{M.3}$$

where:

$$\rho = \sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}, \tag{M.4}$$

and r is a rank parameter (which appears when one controls the norm of the factorized gradient). A standard result (the descent lemma) for any function $\hat{g}$ with L-Lipschitz gradient tells us that for any two matrices M and N,

$$\hat{g}(N, w) \leq \hat{g}(M, w) + \langle\nabla\hat{g}(M, w), N - M\rangle + \frac{L}{2}\|N - M\|_F^2. \tag{M.5}$$

Let

$$N = M_{t+1} = M_t - \eta\nabla\hat{g}(M_t, w), \tag{M.6}$$

we obtain:

$$\hat{g}(M_{t+1}, w) \leq \hat{g}(M_t, w) - \eta\|\nabla\hat{g}(M_t, w)\|_F^2 + \frac{L\eta^2}{2}\|\nabla\hat{g}(M_t, w)\|_F^2. \tag{M.7}$$

Rearrange the right-hand side of Equation (M.7) as:

$$\hat{g}(M_{t+1}, w) \leq \hat{g}(M_t, w) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla\hat{g}(M_t, w)\|_F^2. \tag{M.8}$$

Thus, if we choose η so that:

$$1 - \frac{L\eta}{2} > 0 \quad \iff \quad \eta < \frac{2}{L}, \tag{M.9}$$

then the term $\eta\left(1 - \frac{L\eta}{2}\right)\|\nabla\hat{g}(M_t, w)\|_F^2$ is positive. In other words, as long as $\eta < \frac{2}{L}$, every gradient descent step will produce a decrease in $\hat{g}(M, w)$. In our model the structure of $\hat{g}(M, w)$ (a log–sum–exp function) together with the properties of $\mathcal{A}(M)$ imply (after some detailed technical estimates) that the Lipschitz constant L can be upper bounded by:

$$L \leq 12\rho\sqrt{r}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right). \tag{M.10}$$

Thus, a sufficient condition for descent is to assume:

$$\eta < \frac{2}{12\rho\sqrt{r}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right)}. \tag{M.11}$$

We ensure that the iterates satisfy:

$$\hat{g}(X_{t+1}X_{t+1}^\top, w) \leq \hat{g}(X_tX_t^\top, w) \tag{M.12}$$

for every $t \geq 0$. This monotonic decrease is a key ingredient in proving the convergence of the vanilla gradient descent procedure to a global minimizer of $\hat{g}(\cdot, w)$. Conveniently, above inshows that $\hat{g}(X_tX_t^\top, 0) \leq \hat{g}(X_{t-1}X_{t-1}^\top, 0)$ for all $t \geq 0$. However, this result can be extended to:

$$\hat{g}(X_tX_t^\top, w) \leq \hat{g}(X_{t-1}X_{t-1}^\top, w), \tag{M.13}$$

by making:

$$1/\eta \geq 12\rho r^{(1/2)}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right), \tag{M.14}$$

since $\nabla\hat{g}(\cdot, w)$ is now a ρ-Lipschitz continuous function. Given, a sufficient condition to the above inequality is that:

$$\eta \leq \left(12\rho r^{(1/2)}\left(2(\sqrt{2}-1)\sqrt{(1-(\delta+\zeta_2 q)^2} + \|M^w\|_F\right)\right)^{-1}, \tag{M.15}$$

for such η, the vanilla gradient descent can converge to the global minima.

□

M.2. Proof of Theorem L.2

Proof. One may expect a similar monotonicity in the descent of the objective (M.14), provided that the step-size satisfies a condition of the form:

$$\frac{1}{\eta} \geq 12\tilde{\rho}\sqrt{r}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right), \tag{M.16}$$

where $\tilde{\rho}$ is the Lipschitz constant of $\nabla\hat{g}(\cdot, w)$. (In the MSE case the corresponding Lipschitz constant was denoted by ρ.) Notice that the log-sum-exp function is smooth and it is well known that if each loss term (here a squared difference divided by h^2) has a Hessian bounded by a constant then the gradient of the log-sum-exp (which is a soft-max of these terms) is Lipschitz with a slightly larger constant. In many cases one can obtain an inequality of the form:

$$\tilde{\rho} \leq \frac{C}{h^2}\rho, \tag{M.17}$$

for some universal constant C (in many applications, one may take $C = 1$ up to a harmless constant). A sufficient condition is then that:

$$\eta \leq \left(12\tilde{\rho}\sqrt{r}\left(2(\sqrt{2}-1)\sqrt{1-(\delta+\zeta_2 q)^2} + \|M^w\|_F\right)\right)^{-1}. \tag{M.18}$$

Replacing $\tilde{\rho}$ with ρ/h^2 (up to a multiplicative constant) we obtain:

$$\eta \leq \left(\frac{12\rho}{h^2} \sqrt{r} \Big(2(\sqrt{2}-1)\sqrt{1-(\delta+\zeta_2 q)^2} + \|M^w\|_F \Big) \right)^{-1}. \tag{M.19}$$

Thus, compared to the MSE case, the new step-size condition has an extra dependence on $1/h^2$ in the Lipschitz term. In other words, in order for vanilla gradient descent to converge to the global minimizer when using the above $\hat{g}(M, w)$, the step-size η must be chosen small enough so that (M.19) holds.

□

M.3. Proof of the Lower Bound

M.4. Proof of Theorem 7.1

Proof. for MSE loss, we start with the inequality:

$$\|M - M^*\|_F^2 \geq \frac{2}{L} \left[(1+\delta)\|M - M^*\|_F^2 + 2\sqrt{1+\delta}\|w\|\|M - M^*\|_F \right]. \tag{M.20}$$

For convenience, define:

$$x = \|M - M^*\|_F \quad \text{with } x \geq 0. \tag{M.21}$$

Then the inequality (M.20) becomes:

$$x^2 \geq \frac{2}{L} \Big[(1+\delta)x^2 + 2\sqrt{1+\delta}\|w\|x \Big]. \tag{M.22}$$

Multiply both sides by L (assuming $L > 0$) and factor x^2 from the first two terms:

$$\Big[L - 2(1+\delta) \Big] x^2 - 4\sqrt{1+\delta}\|w\|x \geq 0. \tag{M.23}$$

Since $x \geq 0$, a solution is $x = 0$ (which corresponds to perfect recovery). For a nonzero x we require:

$$\Big[L - 2(1+\delta) \Big] x - 4\sqrt{1+\delta}\, \|w\| \geq 0. \tag{M.24}$$

Assuming a sufficiently large L so that:

$$L - 2(1+\delta) > 0, \tag{M.25}$$

we can solve for x:

$$x \geq \frac{4\sqrt{1+\delta}\|w\|}{L - 2(1+\delta)}. \tag{M.26}$$

Thus, the nontrivial lower bound for $\|M - M^*\|_F$ is

$$\|M - M^*\|_F \geq \frac{4\sqrt{1+\delta}\epsilon}{L - 2(1+\delta)}. \tag{M.27}$$

with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$. □

M.5. Proof of Theorem 7.2

Proof. For the kernel loss function, we are given the inequality in Section H:

$$\|M - M^*\| \geq \sqrt{\frac{1}{L} \cdot \frac{1-\delta}{2}\|M - M^*\|_F^2 - (1+\delta)R(w)\|M - M^*\|_F}. \tag{M.28}$$

For clarity we introduce the notation

$$x = \|M - M^*\|_F \quad (x \geq 0). \tag{M.29}$$

Then the inequality reads:

$$\|M - M^*\|_F \geq \sqrt{\frac{1-\delta}{2L}x^2 - (1+\delta)R(w)x}. \tag{M.30}$$

Because the square root is defined only when its argument is nonnegative we must assume that

$$\frac{1-\delta}{2L}x^2 - (1+\delta)R(w)x \geq 0. \tag{M.31}$$

Factor out $x \geq 0$:

$$x\left(\frac{1-\delta}{2L}x - (1+\delta)R(w)\right) \geq 0. \tag{M.32}$$

Thus (apart from the trivial case $x = 0$) we require

$$\frac{1-\delta}{2L}x - (1+\delta)R(w) \geq 0 \quad \Longrightarrow \quad x \geq \frac{2L(1+\delta)R(w)}{1-\delta}. \tag{M.33}$$

Under that condition the radicand is nonnegative, and the inequality becomes:

$$\|M - M^*\|_F \geq \sqrt{x\left(\frac{1-\delta}{2L}x - (1+\delta)R(w)\right)}. \tag{M.34}$$

Thus, the lower bound (expressed in terms of the Frobenius norm $x = \|M - M^*\|_F$ and the given quantities) is:

$$\|M - M^*\|_F \geq \frac{2L(1+\delta)R(\epsilon)}{1-\delta}, \tag{M.35}$$

with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$. This is the desired lower bound for $\|M - M^*\|_F$. □

N. Compared with Ma's General Loss Result

N.1. Optimization Landscape Result with $\delta \leq 1/3$

Lemma N.1. [32] *Assume that the objective function satisfies Assumptions 2.3 and 2.1, and that $f(M, 0)$ satisfies the RIP property with some δ-RIP$_{2r,2r}$ constant such that $\delta < 1/3$. For every $\epsilon \in \left[0, \frac{1/3-\delta}{\zeta_2}\right]$, with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, every local minimizer $\hat{X}$ satisfies:*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{2\zeta_1\epsilon}{1 - 3(\delta + \zeta_2\epsilon)}. \tag{N.1}$$

The lemma proof is in O.

Lemma N.1 (adapted from [32]) establishes a global-to-local optimality guarantee under noise, assuming the objective satisfies standard smoothness and RIP conditions. Specifically, if the noise vector satisfies $\|w\|_2 \leq \epsilon$ with high probability and the noiseless function $f(M, 0)$ satisfies the δ-RIP$_{2r,2r}$ condition with $\delta < 1/3$, then for any $\epsilon \in [0, \frac{1/3-\delta}{\zeta_2}]$, all local minimizers $\hat{X}$ satisfy the error bound $\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{2\zeta_1\epsilon}{1-3(\delta+\zeta_2\epsilon)}$, highlighting that under controlled noise, local solutions remain close to the ground truth with a bound that degrades gracefully in ϵ.

From the above lemma, we can know that, for MSE loss:

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \mathcal{O}(\epsilon). \tag{N.2}$$

still holds, which means MSE is better than the general loss result. And according to the kernel loss 4.1, our robust loss is better. The results indicate that our condition (J.18), which is:

$$0 < \delta < \sqrt{\frac{2 - \frac{G}{\sigma_r} - \frac{2BL_2}{G_{\min}h^2}}{\frac{8\left(1+\frac{2B^2}{h^2}\right)}{G_{\min}h^2} + \frac{16B^2}{G_{\min}^2h^4}}} - 1. \tag{N.3}$$

should be smaller than $1/3$. For larger δ situations, according to Ma's general loss we see:

N.2. Optimization Landscape Result with $\delta \geq 1/3$

Lemma N.2. [32] *Assume that the objective function satisfies assumptions 2.3 and 2.1 with $f(M,0)$ satisfying the δ-RIP property for a constant $\delta \in (0,1)$. Consider an arbitrary number $\tau \in (0, 1-\delta^2)$. Every local minimizer $\hat{X} \in \mathbb{R}^{n\times r}$ satisfying:*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*), \tag{N.4}$$

will also satisfy the following inequality with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$:

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\epsilon(1+\delta+\zeta_2\epsilon)\zeta_1 C(\tau, M^*)}{\sqrt{1-\tau-\zeta_2\epsilon-\delta}} \tag{N.5}$$

for all $\epsilon < \frac{\sqrt{1-\tau-\delta}}{\zeta_2}$, where

$$C(\tau, M^*) = \sqrt{\frac{2(\lambda_1(M^*) + \tau\lambda_r(M^*))}{(1-\tau)\lambda_r(M^*)}}. \tag{N.6}$$

Lemma N.2 (from [32]) provides a refined local error bound for low-rank matrix estimation under noise, assuming the objective satisfies standard smoothness and RIP conditions with RIP constant $\delta \in (0,1)$. If a local minimizer $\hat{X}$ is sufficiently close to the ground truth—specifically, within a $\tau\lambda_r(M^*)$ neighborhood in Frobenius norm—then, with high probability (at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$), the error remains controlled and satisfies

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\epsilon(1+\delta+\zeta_2\epsilon)\zeta_1 C(\tau, M^*)}{\sqrt{1-\tau-\zeta_2\epsilon-\delta}}, \tag{N.7}$$

for all $\epsilon < \frac{\sqrt{1-\tau-\delta}}{\zeta_2}$. The constant $C(\tau, M^*)$ captures the curvature of the problem via the spectrum of M^*, and the result quantifies how local proximity and structured noise jointly lead to provable closeness to the ground truth.

The proof is provided in Section P. We noticed that when $\delta \geq 1/3$,

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \mathcal{O}\Big(\frac{\epsilon(1+\epsilon)}{\sqrt{1-\epsilon}}\Big) \tag{N.8}$$

, which is not $\|\hat{X}\hat{X}^\top - M^*\|_F \leq \mathcal{O}(\epsilon)$.

N.3. Convergence Theorem for Our Robust Loss

Lemma N.3. (*from* [32]) *The vanilla gradient descent method applied to (2.4) under Assumptions 3 and 4 converges to $\mathcal{P}_r(M^w)$, the best rank-r approximation of M^w, linearly up to a difference D_r if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < C_w^2(1-\delta-\zeta_2\epsilon) - C_w\sqrt{\frac{1-\delta-\zeta_2\epsilon}{1+\delta+\zeta_2\epsilon}}D_r, \tag{N.9}$$

meaning that vanilla gradient descent will reach a point $\tilde{M}$ linearly with $\|\tilde{M} - \mathcal{P}_r(M^w)\|_F \geq D_r$, where

$$D_r = \|M^w - \mathcal{P}_r(M^w)\|_F, \;\; C_w = \sqrt{2(\sqrt{2}-1)\sigma_r(M^w)}. \tag{N.10}$$

The linear convergence is also contingent on the fixed step size η satisfying:

$$\eta \leq \left(12\rho r^{(1/2)}\left(C\sqrt{(1-(\delta+\zeta_2\epsilon)^2} + \|M^w\|_F\right)\right)^{-1}, \tag{N.11}$$

for all $\epsilon < \frac{1-\delta}{\zeta_2}$ with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, where $C = 2(\sqrt{2}-1)$.

The proof is provided in Section 3.2 of [32]. Lemma N.3 establishes a linear convergence guarantee for vanilla gradient descent applied to the noisy low-rank matrix recovery problem, under Assumptions 3 and 4. It shows that if the initialization X_0 is sufficiently close to the noise-perturbed ground truth M^w, specifically satisfying the condition in (N.9), then the iterates converge linearly to a point $\tilde{M}$

satisfying $\|\tilde{M}-\mathcal{P}_r(M^w)\|_F \geq D_r$, where D_r quantifies the residual energy outside the top-r spectrum. The convergence rate and neighborhood depend on the spectral structure of M^w, the RIP constant δ, and the noise level ϵ, with high-probability guarantees when $\|w\|_2 \leq \epsilon$. Furthermore, convergence holds under a fixed step size η subject to a bound inversely proportional to the smoothness parameter ρ, the rank r, and the Frobenius norm of M^w, thereby ensuring that stability and convergence are preserved in the presence of bounded noise.

This is very hard to satisfy. Because the kernel loss function is not symmetric. With new assumption 5, we are hard to have:

Corollary N.4. (*from* [32]) *The vanilla gradient descent method applied to* (2.4) *under Assumptions 3,4,5 converges to M^w linearly if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < 2(\sqrt{2}-1)(1-\delta-\zeta_2\epsilon)\sigma_r(M^w), \tag{N.12}$$

with fixed step size η satisfying:

$$\eta \leq \left(12\rho r^{(1/2)}\left(C\sqrt{(1-(\delta+\zeta_2\epsilon)^2}+\|M^w\|_F\right)\right)^{-1}, \tag{N.13}$$

for all $\epsilon < \frac{1-\delta}{\zeta_2}$ with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, where $C = 2(\sqrt{2}-1)$.

Corollary N.4 refines the linear convergence result of Lemma N.3 by incorporating Assumption 5 and establishing conditions under which vanilla gradient descent converges directly to M^w, the noise-perturbed ground truth, rather than just its best rank-r approximation. Specifically, if the initialization X_0 satisfies the proximity condition $\|X_0X_0^\top - M^w\|_F < 2(\sqrt{2}-1)(1-\delta-\zeta_2\epsilon)\sigma_r(M^w)$, then the iterates converge linearly to M^w with high probability over the noise realization. The step size η must again be chosen appropriately based on the problem's smoothness constant ρ, rank r, and signal strength $\|M^w\|_F$, ensuring that the descent dynamics remain stable even in the presence of bounded noise.

Lemma N.5. (*from* [32]) *Suppose that the objective function of* (2.4) *satisfies assumptions 3 with a δ-RIP$_{2r,2r}$ constant of $\delta < 1/3$ in the noiseless case. Consider the ground truth solution M^* which is of rank r. For a given constant $\alpha > 0$, there exists a finite constant $\xi > 0$ such that at least one of the following three conditions holds for any $X \in \mathbb{R}^{n\times r}$:*

$$\begin{aligned} &\operatorname{dist}(X, M^*) \leq \alpha,\ \|\nabla_X \hat{g}(X,w)\|_F \geq \xi, \\ &\lambda_{\min}(\nabla^2_X \hat{g}(X,w)) \leq -2\xi, \end{aligned} \tag{N.14}$$

with probability at least $\mathbb{P}(\|w\|_2 \leq \frac{1/3-\delta}{\zeta_2+2\zeta_\alpha/3})$, where $\zeta_\alpha := \zeta_1/(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)$.

The background together with the detailed proof are provided in Section [32] proof of section 4.

Lemma N.5 establishes a strict saddle property for the noisy optimization problem (2.4) under a δ-RIP condition with $\delta < 1/3$ in the noiseless setting. It guarantees that for any point $X \in \mathbb{R}^{n\times r}$, at least one of the following must hold: (i) X is within distance α of the ground truth M^*, (ii) the gradient norm is large, $\|\nabla_X \hat{g}(X,w)\|_F \geq \xi$, or (iii) the Hessian has a strictly negative eigenvalue, $\lambda_{\min}(\nabla^2_X \hat{g}(X,w)) \leq -2\xi$. This ensures that saddle points are unstable and can be escaped by first-order methods. The result holds with high probability when the noise is bounded as $\|w\|_2 \leq \frac{1/3-\delta}{\zeta_2+2\zeta_\alpha/3}$, where ζ_α depends on the spectral properties of M^* and the target accuracy α, confirming that the robust landscape of $\hat{g}(X,w)$ admits no spurious stable critical points far from the true solution.

O. Proof of Lemma N.1

O.1. Proof of Lemma O.1

Lemma O.1. *If $\hat{X}$ is a local minimum of* (2.4) *with $\hat{M} = \hat{X}\hat{X}^\top$, then*

$$\sigma_r^2(\hat{M}) \geq \frac{G^2}{(1+\delta+\zeta_2 q)^2}, \quad \text{and} \quad G^2 \leq \sigma_r^2\rho^2, \tag{O.1}$$

where $G = -\sigma_{\min}(\nabla_M \hat{g}(\hat{M}, w))$.

Proof of Lemma O.1. Based on Ma [32]'s result, consider the case where $\operatorname{rank}(\hat{M}) = r$. Under this assumption, consider the singular value decomposition (SVD) of $\hat{M}$: $\hat{M} = \sum_{i=1}^{r} \sigma_i u_i u_i^\top$, where σ_i's are eigenvalues and u_i's are unit eigenvectors. Let u_G be a unit eigenvector of $\nabla f(\hat{M}, w)$ such that $u_G^\top \nabla f(\hat{M}, w) u_G = -G$. Furthermore, for a constant $p \in [0, 1]$, define:

$$M_p = \sum_{i=1}^{r-1} \sigma_i u_i u_i^\top + \sigma_r (p u_G + \sqrt{1-p^2} u_r)(p u_G + \sqrt{1-p^2} u_r)^\top. \tag{O.2}$$

We expand the term $\|M_p - M\|_F^2$: $\|M_p - \hat{M}\|_F^2 = 2\sigma_r^2 p^2$, which means that

$$\langle \nabla_M f(\hat{M}, w), M_p - \hat{M} \rangle = -\frac{G}{2\sigma_r} \|M_p - \hat{M}\|_F^2. \tag{O.3}$$

Computing the Taylor expansion of $\hat{g}(M, w)$ in terms of M at the point $\hat{M}$ with the mean-value theorem gives:

$$\hat{g}(M_p, w) = \hat{g}(\hat{M}, w) + \langle \nabla_M \hat{g}(\hat{M}, w), M_p - \hat{M} \rangle + \frac{1}{2}[\nabla^2 \hat{g}(\tilde{M}, w)](M_p - \hat{M}, M_p - \hat{M}), \tag{O.4}$$

for some matrix $\tilde{M}$ that is a convex combination of M_p and $\hat{M}$. Due to(2.2), we have:

$$\|\nabla_M \hat{g}(M) - \nabla_M \hat{g}(M')\|_F \le \rho \|M - M'\|_F. \tag{O.5}$$

This inequality implies that the Hessian $\nabla^2 \hat{g}(M, w)$ is bounded by ρ in the sense that:

$$\|\nabla^2 \hat{g}(M, w)\|_F \le \rho. \tag{O.6}$$

Thus, we get the following bound for $\hat{g}(M_p, w)$:

$$\hat{g}(M_p, w) \le \hat{g}(\hat{M}, w) + \langle \nabla_M \hat{g}(\hat{M}, w), M_p - \hat{M} \rangle + \frac{1}{2}\rho \|M_p - \hat{M}\|_F^2. \tag{O.7}$$

So we have:

$$G^2 \le \sigma_r^2 \rho^2. \tag{O.8}$$

Based on our the kernel loss (2.3), we still have the original result: $\sigma_r^2(\hat{M}) \ge \frac{G^2}{(1+\delta+\zeta_2 q)^2}$, and$G^2 \le \sigma_r^2 \rho^2$.

□

P. Proof of Theorem N.2

Given a matrix $\hat{X}$, we aim to find the smallest δ such that there is an instance of the problem with this RIP constant for which $\hat{X}$ is a local minimizer that is not associated with the ground truth. For notational convenience, we denote this optimal value as $\delta^*(\hat{X})$. Namely, $\delta^*(\hat{X})$ is the optimal value to the following optimization problem:

$$\begin{aligned} \min_{\delta, f(\cdot,w)} \quad & \delta \\ \text{s.t.} \quad & \hat{X} \text{ is a local minimizer of } f(\cdot, w), \\ & f(\cdot, 0) \text{ satisfies the } \delta\text{-RIP}_{2r} \text{ property}. \end{aligned} \tag{P.1}$$

By the above optimization problem, we know that $\delta \ge \delta^*(\hat{X})$ for all local minimizers $\hat{X}$ of $f(\cdot, w)$, where δ is the best RIP constant of the problem. Since (P.1) is difficult to analyze, we replace its two constraints with some necessary conditions, thus forming a relaxation of the original problem with its optimal value being a lower bound on $\delta^*(\hat{X})$.

To find a necessary condition replacing the two constraints, we introduce the following lemma. This is the first lemma that captures the necessary conditions of a critical point of (2.4), a problem where random noise is considered.

Lemma P.1. *Assume that the objective function $f(M, w)$ of (2.4) satisfies all assumptions in Section, and that $\hat{X}$ is a first-order critical point of (2.4). Then, $\hat{X}$ must satisfy the following conditions for some symmetric matrix $\mathbf{H} \in \mathbb{R}^{n^2 \times n^2}$:*

1. $\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2$.
2. $\mathbf{H}$ *satisfies the $(\delta + \zeta_2 q)$-RIP$_{2r,2r}$ property, which means that the inequality*
$$(1 - \delta - \zeta_2 q)\|M\|_F^2 \leq \mathbf{m}^\top \mathbf{H}\mathbf{m} \leq (1 + \delta + \zeta_2 q)\|M\|_F^2 \tag{P.2}$$
holds for every matrix $M \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(M) \leq 2r$, where$\mathbf{m} = \operatorname{vec}(M)$ and $\mathbf{e} = \operatorname{vec}(\hat{X}\hat{X}^\top - M^*)$, $\hat{\mathbf{X}}$ is defined as per Section.

Given Lemma P.1, we can obtain a relaxation of problem (P.1), namely the following optimization problem:

$$\begin{aligned} \min_{\delta, \mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2, \\ & (1 - \delta - \zeta_2 q)\|M\|_F^2 \leq \mathbf{m}^\top \mathbf{H}\mathbf{m} \leq \\ & (1 + \delta + \zeta_2 q)\|M\|_F^2, \ \ \forall M : \operatorname{rank}(M) \leq 2r. \end{aligned} \tag{P.3}$$

where $\mathbf{m} = \operatorname{vec}(M)$. Note that since the second constraint is hard to deal with, so we solve the following problem that has the same optimal value:

$$\begin{aligned} \min_{\delta, \mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2, \\ & (1 - \delta - \zeta_2 q) I_{n^2} \preceq \mathbf{H} \preceq (1 + \delta + \zeta_2 q) I_{n^2}. \end{aligned} \tag{P.4}$$

If the optimal value of (P.4) is denoted as $\delta_f^*(\hat{X})$, then we know that $\delta_f^*(\hat{X}) \leq \delta^*(\hat{X}) \leq \delta$ due to (P.3) being a relaxation of (P.1). By further lower-bounding $\delta_f^*(\hat{X})$ with an expression in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, we can obtain an upper bound on $\|\hat{X}\hat{X}^\top - M^*\|_F$.

Lemma P.2. *Let $\hat{X}$ be a first-order critical point of (2.4), and suppose that $f(M, w)$ satisfies all assumptions stated in 2.3. Then there exists a symmetric matrix $\mathbf{H} \in \mathbb{R}^{n^2 \times n^2}$ such that:*

- $\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2$.
- $\mathbf{H}$ *satisfies the $(\delta + \zeta_2 q)$-RIP$_{2r,2r}$ property:*
$$(1 - \delta - \zeta_2 q)\|M\|_F^2 \leq \mathbf{m}^\top \mathbf{H}\mathbf{m} \leq (1 + \delta + \zeta_2 q)\|M\|_F^2, \tag{P.5}$$
for every $M \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(M) \leq 2r$, where $\mathbf{m} = \operatorname{vec}(M)$ and $\mathbf{e} = \operatorname{vec}\big(\hat{X}\hat{X}^\top - M^*\big)$.

From these conditions, one can form the relaxation of (P.1) .

Let $\delta_f^*(\hat{X})$ be the optimal value of that relaxation. It follows that:

$$\delta_f^*(\hat{X}) \leq \delta^*(\hat{X}) \leq \delta. \tag{P.6}$$

By bounding $\delta_f^*(\hat{X})$ from below in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, one obtains an upper bound on $\|\hat{X}\hat{X}^\top - M^*\|_F$.

Proof of Lemma P.2. Similar to the last section, we first define $\hat{M} = \hat{X}\hat{X}^\top$. Since $\hat{X}$ is a first-order critical point, it follows from that $\nabla_X h(\hat{X}, w) = 0$. Thus,

$$0 = \langle \nabla_X h(\hat{X}, w), U \rangle = \langle \nabla_M f(\hat{M}, w), \hat{X}U^\top + U\hat{X}^\top \rangle, \tag{P.7}$$

for an arbitrary $U \in \mathbb{R}^{n \times r}$. Let $u = \operatorname{vec}(U)$. Next, we define the function $g(\cdot) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}$:

$$g(V) = \langle \nabla_M f(V, w), \hat{X}U^\top + U\hat{X}^\top \rangle, \tag{P.8}$$

for all $V \in \mathbb{R}^{n\times n}$. Then, $g(\hat{M}) = 0$ due to (P.7). By the mean-value theorem (MTV), we have:

$$\begin{aligned} g(\hat{M}) - g(M^*) &= \int_0^1 \langle \nabla g(tM^* + (1-t)\hat{M}), \hat{M} - M^* \rangle \mathrm{d}t \\ &= \int_0^1 [\nabla^2_M f(tM^* + (1-t)\hat{M})](\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top)\mathrm{d}t \\ &= \mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u \end{aligned} \tag{P.9}$$

where $\mathbf{H} \in \mathbb{R}^{n^2\times n^2}$ is a symmetric matrix that is independent of U and satisfies:

$$\mathrm{vec}(K)^\top \mathbf{H}\,\mathrm{vec}(L) = \int_0^1 [\nabla^2_M f(tM^* + (1-t)\hat{M})](K, L)\mathrm{d}t \tag{P.10}$$

for all $K, L \in \mathbb{R}^{n\times n}$. This means:

$$\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u = g(\hat{M}) - g(M^*). \tag{P.11}$$

Taking the absolute value of both sides and upper-bounding the right-hand side gives:

$$\begin{aligned} |\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u| &= \|g(\hat{M}) - g(M^*)\| \le \|g(M^*)\| \\ &\le \zeta_1 q \|\hat{X}U^\top + U\hat{X}^\top\|_F \\ &\le 2\zeta_1 q \|\hat{X}U^\top\|_F \\ &= 2\zeta_1 q \sqrt{\mathrm{tr}(\hat{X}\hat{X}^\top UU^\top)} \\ &\le 2\zeta_1 q \|\hat{X}\|_2 \|u\|, \end{aligned} \tag{P.12}$$

where the second line follows from combining and (2.7), and the fourth line follows from the cyclic property of trace operators. Choosing $u = \hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}$ can simplify the above inequality to

$$\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \le 2\zeta_1 q \|\hat{X}\|_2. \tag{P.13}$$

Furthermore, the δ-RIP$_{2r,2r}$ property of the objective function means that:

$$(1-\delta)\|M\|_F^2 \le [\nabla^2 f(\xi, 0)](M, M) \le (1+\delta)\|M\|_F^2 \tag{P.14}$$

for all M with $\mathrm{rank}(M) \le 2r$. Combining with the fact that

$$\|\,\mathrm{vec}(M)^\top \mathbf{H}\,\mathrm{vec}(M) - [\nabla^2 f(\xi, 0)](M, M)\| \le \zeta_2 q \|M\|_F^2, \tag{P.15}$$

gives (P.2). □

Proof of the Theorem. The proof is provided in [32]'s proof od section 3.2. There is some difference, so we provide a slightly different proof here.

To analyze the local condition in Theorem 4.2, it is helpful to replace the parameter δ in problem (P.4) with a new scalar η and consider the alternative optimization program

$$\begin{aligned} \max_{\eta, \widehat{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \|\widehat{\mathbf{X}}^\top \widehat{\mathbf{H}}\mathbf{e}\| \le 2\zeta_1 q \|\widehat{X}\|_2, \\ & \eta I_{n^2} \preceq \widehat{\mathbf{H}} \preceq I_{n^2}. \end{aligned} \tag{P.16}$$

Any feasible pair $(\delta, \mathbf{H})$ for (P.4) immediately generates a feasible point for (P.16) through

$$\eta = \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}, \qquad \widehat{\mathbf{H}} = \frac{1}{1+\delta+\zeta_2 q}\mathbf{H}. \tag{P.17}$$

If $\eta_f^*(\widehat{X})$ denotes the optimal objective value of (P.16), then

$$\eta_f^*(\widehat{X}) \ge \frac{1-\delta_f^*(\widehat{X})-\zeta_2 q}{1+\delta_f^*(\widehat{X})+\zeta_2 q} \ge \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}, \tag{P.18}$$

because every local minimizer satisfies $\delta_f^*(\widehat{X}) \le \delta^*(\widehat{X}) \le \delta$. Thus the remaining task is to obtain an upper bound on $\eta_f^*(\widehat{X})$.

To achieve this, we examine the dual of (P.16), which takes the form

$$\begin{aligned} \min_{U_1,U_2,G,\lambda,y} \quad & \operatorname{tr}(U_2) + 4\zeta_1^2 q^2 \|\widehat{X}\|_2^2 \lambda + \operatorname{tr}(G) \\ \text{s.t.} \quad & \operatorname{tr}(U_1) = 1, \\ & (\widehat{\mathbf{X}} y)\mathbf{e}^\top + \mathbf{e}(\widehat{\mathbf{X}} y)^\top = U_1 - U_2, \\ & \begin{pmatrix} G & -y \\ -y^\top & \lambda \end{pmatrix} \succeq 0, \\ & U_1 \succeq 0, \quad U_2 \succeq 0. \end{aligned} \tag{P.19}$$

Define

$$M = (\widehat{\mathbf{X}} y)\mathbf{e}^\top + \mathbf{e}(\widehat{\mathbf{X}} y)^\top, \tag{P.20}$$

and decompose M as $M = [M]_+ - [M]_-$, where both parts are positive semidefinite. A feasible dual solution can be constructed by choosing

$$y^* = \frac{y}{\operatorname{tr}([M]_+)}, \qquad U_1^* = \frac{[M]_+}{\operatorname{tr}([M]_+)}, \qquad U_2^* = \frac{[M]_-}{\operatorname{tr}([M]_+)}, \tag{P.21}$$

and

$$\lambda^* = \frac{\|y^*\|}{2\zeta_1 q \|\widehat{X}\|_2}, \qquad G^* = \frac{y^*(y^*)^\top}{\lambda^*}. \tag{P.22}$$

Evaluating the dual objective under this choice gives

$$\frac{\operatorname{tr}([M]_-) + 4\zeta_1 q \|\widehat{X}\|_2 \|y\|}{\operatorname{tr}([M]_+)}. \tag{P.23}$$

Assume $\widehat{X}$ satisfies $\|\widehat{X}\widehat{X}^\top - M^*\|_F \le \tau \lambda_r(M^*)$. Then $\widehat{X} \neq 0$, and for any nonzero y satisfying that $\widehat{X}^\top \operatorname{mat}(y)$ is symmetric, one has

$$\|\widehat{\mathbf{X}} y\|^2 \ge 2\lambda_{r^*}(\widehat{X}\widehat{X}^\top)\|y\|^2. \tag{P.24}$$

Perturbation bounds imply

$$|\lambda_{r^*}(\widehat{X}\widehat{X}^\top) - \lambda_{r^*}(M^*)| \le \tau \lambda_r(M^*), \qquad |\lambda_1(\widehat{X}\widehat{X}^\top) - \lambda_1(M^*)| \le \tau \lambda_r(M^*). \tag{P.25}$$

Combining these with (P.24), we obtain

$$\frac{2\|\widehat{X}\|_2 \|y\|}{\|\widehat{\mathbf{X}} y\|} \le \sqrt{\frac{2(\lambda_1(M^*) + \tau\lambda_r(M^*))}{(1-\tau)\lambda_r(M^*)}} \equiv C(\tau, M^*). \tag{P.26}$$

Let θ denote the angle between $\widehat{\mathbf{X}} y$ and $\mathbf{e}$. Then

$$\operatorname{tr}([M]_+) = \|\widehat{\mathbf{X}} y\|\|\mathbf{e}\|(1+\cos\theta), \qquad \operatorname{tr}([M]_-) = \|\widehat{\mathbf{X}} y\|\|\mathbf{e}\|(1-\cos\theta). \tag{P.27}$$

Substituting these into (P.23) together with (P.26) gives

$$\eta_f^*(\widehat{X}) \le \frac{1 - \cos\theta + 2\zeta_1 q\, C(\tau, M^*)/\|\mathbf{e}\|}{1+\cos\theta}. \tag{P.28}$$

Combining this with (P.18) yields

$$\|\mathbf{e}\| \le \frac{(1+\delta+\zeta_2 q)\zeta_1 q\, C(\tau, M^*)}{\cos\theta - \zeta_2 q - \delta}. \tag{P.29}$$

To bound $\cos\theta$, note that bounding $\sin^2\theta$ suffices. Following the standard orthogonal decomposition of $\widehat{X}$ and Z, one eventually arrives at the estimate

$$\sin^2\theta \leq \frac{\tau}{2-\tau} \leq \tau. \tag{P.30}$$

Since $\tau < 1$, this implies

$$\cos\theta \geq \sqrt{1-\tau}. \tag{P.31}$$

Substituting this into (P.29) completes the argument.

□

Q. Proof of Theorem N.3

First and foremost, we restate this lemma from :

Lemma Q.1. *For any matrix $X \in \mathbb{R}^{n\times r}$, given a positive semidefinite matrix $M \in \mathbb{R}^{n\times n}$ of rank r, we have:*

$$\|XX^\top - M\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r(M)(\mathrm{dist}(X,M))^2. \tag{Q.1}$$

Also, given Assumption 5, we have

$$\nabla_M f(M^w, w) = 0. \tag{Q.2}$$

First, we establish that the PL inequality holds in a neighborhood of the global minimizer.

Lemma Q.2. *Consider the global minimizer M^w of (2.4). There exists a constant $\mu > 0$ such that the PL inequality:*

$$\frac{1}{2}\|\nabla_X h(X,w)\|_F^2 \geq \mu(h(X,w) - f(\mathcal{P}_r(M^w), w)), \tag{Q.3}$$

holds for all $X \in \mathbb{R}^{n\times r}$ satisfying:

$$\mathrm{dist}(X, M^w) < \max\{\sqrt{2(\sqrt{2}-1)}\sqrt{1-(\delta+\zeta_2 q)^2}(\sigma_r(M^w))^{1/2} - D_r, 0\} \tag{Q.4}$$

and

$$D_r \leq \mathrm{dist}(X, \mathcal{P}_r(M^w)), \tag{Q.5}$$

for $q < (1-\delta)/\zeta_2$.

Proof of Lemma Q.2. We prove the Lemma when $C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r > 0$, since otherwise it is trivial. Denote $M \coloneqq XX^\top$. First, we fix a constant $\tilde{C}$ such that:

$$\mathrm{dist}(X, M^w) \leq \tilde{C} < C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r. \tag{Q.6}$$

Then, we define q_1 and q_2 as follows:

$$q_1 = \sqrt{1 - \frac{\tilde{C}^2}{2(\sqrt{2}-1)\sigma_r(M^w)}}, q_2 = \frac{\sqrt{2}\mu'}{\sigma_r(M^w)^{1/2} - \tilde{C}}. \tag{Q.7}$$

Now, both q_1 and q_2 are nonnegative resulting from the assumption above. Furthermore, we know that $\delta + \zeta_2 q < \sqrt{1 - \frac{\tilde{C}^2}{2(\sqrt{2}-1)\sigma_r(M^w)}}$ from (Q.6), then

$$\frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q} > \frac{1-q_1+q_2}{1+q_1}, \tag{Q.8}$$

for some small enough μ'. Define $\mu = (\mu')^2/(1+\delta+\zeta_2 q+2\rho)$. First, we make the assumption that:

$$\frac{1}{2}\|\nabla_X h(X,w)\|_F^2 < \mu(h(X,w) - f(\mathcal{P}_r(M^w), w)). \tag{Q.9}$$

From this assumption, we have:

$$\begin{aligned}
&\mu(h(X,w) - f(\mathcal{P}_r(M^w),w)) \\
&\leq \mu\left(\langle \nabla_M f(\mathcal{P}_r(M^w),w), M - \mathcal{P}_r(M^w)\rangle + \frac{1+\delta+\zeta_2 q}{2}\|M-\mathcal{P}_r(M^w)\|_F^2\right) \\
&\leq \mu\left(\rho\|M^w - \mathcal{P}_r(M^w)\|_F\|M-\mathcal{P}_r(M^w)\|_F + \frac{1+\delta+\zeta_2 q}{2}\|M-\mathcal{P}_r(M^w)\|_F^2\right) \\
&\leq \mu\left(\rho\|M-\mathcal{P}_r(M^w)\|_F^2 + \frac{1+\delta+\zeta_2 q}{2}\|M-\mathcal{P}_r(M^w)\|_F^2\right),
\end{aligned} \tag{Q.10}$$

due to Taylor's theorem and (2.8). So then (Q.9) leads to:

$$\frac{1}{2}\|\nabla h(X,w)\|_F^2 < \mu(\frac{(1+\delta+\zeta_2 q)}{2} + \rho)\|M-\mathcal{P}_r(M^w)\|_F^2. \tag{Q.11}$$

Therefore,

$$\|\nabla h(X,w)\|_F \leq \mu'\|M-\mathcal{P}_r(M^w)\|_F. \tag{Q.12}$$

Then consider the following optimization problem:

$$\begin{aligned}
\min_{\delta,\mathbf{H}\in\mathbb{S}^{n^2}} \quad & \delta \\
\text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq \mu'\|\mathbf{e}\|, \\
& \mathbf{H} \text{ satisfies the } (\delta+\zeta_2 q)\text{-RIP}_{2r} \text{ property},
\end{aligned} \tag{Q.13}$$

where $\mathbf{e} = \text{vec}(XX^\top - \mathcal{P}_r(M^w))$. If we denote the optimal value of (Q.13) as $\delta_f^*(X,\mu')$, then $\delta_f^*(X,\mu') \leq \delta$ because the constraints of (Q.13) are necessary conditions for (Q.9), according to Lemma 12 of . Therefore,

$$\frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q} \leq \frac{1-\delta_f^*(X,\mu')-\zeta_2 q}{1+\delta_f^*(X,\mu')+\zeta_2 q}. \tag{Q.14}$$

and:

$$\eta_f^*(\hat{X}) \geq \frac{1-\delta_f^*(\hat{X})-\zeta_2 q}{1+\delta_f^*(\hat{X})+\zeta_2 q} \geq \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}, \tag{Q.15}$$

Moreover, by the same logic of (Q.15), we know that $\eta_f^*(X,\mu') \geq \frac{1-\delta_f^*(X,\mu')-\zeta_2 q}{1+\delta_f^*(X,\mu')+\zeta_2 q}$, where $\eta_f^*(X,\mu')$ is the optimal value of the optimization problem:

$$\begin{aligned}
\max_{\eta,\hat{\mathbf{H}}} \quad & \eta \\
\text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\mathbf{e}\| \leq \mu'\|\mathbf{e}\|, \\
& \eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}.
\end{aligned} \tag{Q.16}$$

Gives:

$$\eta_f^*(X,\mu') \leq \frac{1-q_1+q_2}{1+q_1}, \tag{Q.17}$$

therefore making a contradiction to (Q.8), subsequently proving (Q.3). □

Proof. If we certify that:

$$\frac{\|XX^\top - M^w\|_F}{C_w} < C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r \tag{Q.18}$$

for any given $X \in \mathbb{R}^{n\times r}$, then a direct substitution can certify that (Q.4) holds for X, since by Lemma Q.1,

$$\text{dist}(X,M) \leq \frac{\|XX^\top - M^w\|_F}{C_w}. \tag{Q.19}$$

Therefore, the certification of (Q.18) means that the PL inequality (Q.3) holds for this given X. Given that is satisfied, then if this inequality holds:

$$\|XX^\top - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|X_0X_0^\top - M^w\|_F, \tag{Q.20}$$

(Q.18) will also hold, because:

$$\sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|X_0X_0^\top - M^w\|_F \leq C_w^2\sqrt{1-(\delta+\zeta_{2q})^2} - C_wD_r. \tag{Q.21}$$

Thus, for the remainder of the proof, we aim to certify that starting from X_0, if we apply the gradient descent algorithm, (L.1) will be satisfied every step along this trajectory. In order to do so, we use Taylor's expansion and (Q.2) to obtain

$$f(M,w) - f(M^w,w) = \frac{[\nabla^2 f(N,w)](M-M^w, M-M^w)}{2}, \tag{Q.22}$$

where N is some convex combination of M and M^w, and $M \in \mathbb{R}^{n\times n}$ is any matrix of rank at most r. In light of the RIP property of the function and Equation (2.8), one can write:

$$\frac{1-\delta-\zeta_{2q}}{2}\|M-M^w\|_F^2 \leq f(M,w) - f(M^w,w) \leq \frac{1+\delta+\zeta_{2q}}{2}\|M-M^w\|_F^2. \tag{Q.23}$$

This means that if $M_1, M_2 \in \mathbb{R}^{n\times n}$ are two matrices of rank at most r with $f(M_1,w) \leq f(M_2,w)$, then:

$$\|M_1 - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|M_2 - M^w\|_F, \tag{Q.24}$$

because $f(M_1,w) - f(M^w,w) \leq f(M_2,w) - f(M^w,w)$. Thus, one can conclude that $f(X_tX_t^\top, w) \leq f(X_0X_0^\top, w)\ \forall t$, where X_t denotes the t^{th} step of the gradient descent algorithm starting from X_0. Hence, (L.1) follows for all X_t. Conveniently, Lemma 11 inshows that $f(X_tX_t^\top, 0) \leq f(X_{t-1}X_{t-1}^\top, 0)$ for all $t \geq 0$. However, this result can be extended to:

$$f(X_tX_t^\top, w) \leq f(X_{t-1}X_{t-1}^\top, w), \tag{Q.25}$$

by making

$$1/\eta \geq 12\rho r^{(1/2)}\left(\sqrt{\frac{1+\delta+\zeta_{2q}}{1-\delta-\zeta_{2q}}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right), \tag{Q.26}$$

since $\nabla f(\cdot, w)$ is now a ρ-Lipschitz continuous function. Given, a sufficient condition to the above inequality is that:

$$\eta \leq \left(12\rho r^{(1/2)}\left(2(\sqrt{2}-1)\sqrt{(1-(\delta+\zeta_{2q})^2} + \|M^w\|_F\right)\right)^{-1}. \tag{Q.27}$$

This finally means that the PL inequality (Q.3) is established for the entire trajectory starting from X_0. Now, applying Theorem 1 ingives:

$$h(X_t,w) - f(\mathcal{P}_r(M^w),w) \leq (1-\mu\eta)^\top(h(X_0,w) - f(\mathcal{P}_r(M^w),w)), \tag{Q.28}$$

which implies a linear convergence as desired.

□

R. Proof for Theorem N.5

First and foremost, we replace δ with $\delta + \zeta_{2q}$ in all of the proofs since in our noisy formulation, the problem is $(\delta+\zeta_{2q})$-RIP$_{2r,2r}$ instead. Then, we introduce the following Lemma in $\nabla_M f(M^*,w) \neq 0$ in the noisy formulation:

Lemma R.1. *Given a constant $\epsilon > 0$, an arbitrary $X \in \mathbb{R}^{n\times r}$, and the ground truth solution $M^* \in \mathbb{R}^{n\times n}$ of (2.4), if*

$$\|XX^\top\|_F^2 \geq \max\left\{\frac{2(1+\delta+\zeta_2 q)}{1-\delta-(\zeta_2+\zeta_D)q}\|M^*\|_F^2, (\frac{2\lambda\sqrt{r}}{1-\delta-(\zeta_2+\zeta_D)q})^{4/3}\right\}, \tag{R.1}$$

then

$$\|\nabla_X h(X,w)\|_F \geq \lambda, \tag{R.2}$$

where $\zeta_D = \zeta_1/D$ and D is a constant such that

$$D^2 \leq (\frac{2\lambda\sqrt{r}}{1-\delta-(\zeta_2+\zeta_D)q})^{4/3}. \tag{R.3}$$

Note that such D exists since we first require that $1-\delta-(\zeta_2+\zeta_D)q \geq 0$, meaning that $\frac{q\zeta_1}{1-\delta-q\zeta_2} \leq D$. Moreover, a sufficient condition to (R.3) is that $D \leq (2\lambda\sqrt{r})^{2/3}$, which can be simultaneously satisfied when λ is chosen properly. The introduction of the lower bound D will not affect the remainder of the proof of Theorem N.5, since in the later steps, we only require the existence of a constant C such that $\|XX^\top\|_F \leq C^2$ when $\|\nabla_X h(X,w)\|_F \leq \lambda$. Therefore, Lemma R.1 perfectly fits this role.

Proof of Lemma R.1. Denote $M \coloneqq XX^\top$. Using the RIP property and (2.7), we have:

$$\begin{aligned}
\langle \nabla_M f(M), M\rangle &= \int_0^1 [\nabla^2 f(M^* + s(M-M^*), w)][M-M^*, M]\mathrm{d}s + \langle \nabla_M f(M^*, w), M\rangle \\
&\geq (1-\delta-\zeta_2 q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F - \zeta_1 q\|M\|_F \\
&= (1-\delta-\zeta_2 q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F - \zeta_D q D\|M\|_F \\
&\geq (1-\delta-(\zeta_2+\zeta_D)q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F \\
&\geq \frac{1-\delta-(\zeta_2+\zeta_D)q}{2}\|M\|_F^2,
\end{aligned} \tag{R.4}$$

where the second last inequality results from (R.3), which implies that $D \leq \|M\|_F$; and the last inequality follows from (R.1). Then combining the fact that $\|X\|_F \leq \sqrt{r}\|M\|_F^{1/2}$, and $\|\nabla_X h(X,w)\|_F \geq \frac{\langle \nabla h(X,w), X\rangle}{\|X\|_F}$ yields the desired fact that

$$\begin{aligned}
\|\nabla_X h(X,w)\|_F \geq \frac{\langle \nabla h(X,w), X\rangle}{\|X\|_F} &= \frac{\langle \nabla_M f(M), M\rangle}{\|X\|_F} \\
&\geq \frac{(1-\delta-(\zeta_2+\zeta_D)q)\|M\|_F^2}{2\sqrt{r}\|M\|_F^{1/2}} \\
&= \frac{1-\delta-(\zeta_2+\zeta_D)q}{2\sqrt{r}}\|M\|_F^{3/2} \\
&\geq \lambda.
\end{aligned} \tag{R.5}$$

□

Then, utilizing Lemma R.1, we can prove in the same fashion to obtain:

$$\begin{aligned}
&\langle \nabla_M f(M,w), M^* - M\rangle \\
&\leq -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 - \langle \nabla_M f(M^*,w), M-M^*\rangle \\
&\leq -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 + \zeta_1 q\|M-M^*\|_F \\
&\leq -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 + \zeta_\alpha q(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)\|M-M^*\|_F \\
&\leq -(1-\delta-(\zeta_2-\zeta_\alpha)q)\|M-M^*\|_F^2
\end{aligned} \tag{R.6}$$

for any $M \in \mathbb{R}^{n\times n}$ that satisfies the requirements in Lemma 7 of . This is because $\|M-M^*\|_F \geq (\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)$ by the assumption of α and Lemma Q.1.

The above change will only affect the constant c in Lemma 7, and the new c will become

$$c = (\sqrt{r}\|M^*\|_F)^{-1}(\sqrt{2}-1)(1-\delta-(\zeta_2-\zeta_\alpha)q)\sigma_r(M^*). \tag{R.7}$$

Since the exact value of c is irrelevant and we only need to prove its existence, the rest of the proof follows from the existing procedure. Note that $c > 0$ is guaranteed by the assumption of noise in Theorem (N.5). Then, we proceed to show that it also be proved similarly, except for one key difference, which is:

$$K \coloneqq (1-3\delta-(3\zeta_2+2\zeta_\alpha)q)(\sqrt{2}-1)\sigma_r(M^*)\alpha^2. \tag{R.8}$$

To verify this statement, we leverage the inequality:

$$-\phi(\bar{M}) \geq f(M,w) - f(M^*,w) - (\delta+\zeta_2 q)\|M-M^*\|_F^2, \tag{R.9}$$

and furthermore we now have that:

$$\begin{aligned}
&f(M,w) - f(M^*,w) \\
&\geq \langle \nabla_M f(M^*,w), M-M^*\rangle + \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 \\
&\geq \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_1 q\|M-M^*\|_F^2 \\
&\geq \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_\alpha q(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)\|M-M^*\|_F^2 \\
&\geq \frac{1-\delta-(\zeta_2+2\zeta_\alpha)q}{2}\|M-M^*\|_F^2
\end{aligned} \tag{R.10}$$

for the same reason elaborated above. Combining the above two inequalities leads to:

$$-\phi(\bar{M}) \geq \frac{1-3\delta-(3\zeta_2+2\zeta_\alpha)q}{2}\|M-M^*\|_F^2 \geq K. \tag{R.11}$$

As assumed in Theorem N.5, since $q < \frac{1/3-\delta}{\zeta_2+2\zeta_\alpha/3}$, we know that $K > 0$. Finally, we choose $C = (\frac{2(1+\delta+\zeta_2\epsilon)}{1-\delta-(\zeta_2+\zeta_D)\epsilon}\|M^*\|_F^2)^{1/4}$ and invoke Lemmas 6-8 to complete the proof of Theorem N.5. Note the ϵ here is the same ϵ appeared in the statement of Theorem N.5.

S. Theoretical Study of The Combined Loss

S.1. Theoretical Study

Theorem S.1 (Combined Loss Bound)**.** *Let $r_i = Y_i - \mathcal{A}(XX^\top)_i$ for $i = 1,\dots,n$, and assume there is a constant $B > 0$ such that $r_i^2 \leq B$ for all i. Define the combined loss:*

$$L_{\text{combined}}(X) = (1-\lambda)\cdot\frac{1}{n}\sum_{i=1}^{n}\Big[-\log\Big(\frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{\big((r_j-r_i)^2\big)}{h^2}\Big)\Big)\Big] + \lambda\cdot\frac{1}{n}\sum_{i=1}^{n}r_i^2. \tag{S.1}$$

Then consider a matrix sensing or low-rank recovery problem where $Y = \mathcal{A}(M^*) + w$, and $\mathcal{A}$ satisfies a restricted isometry property (RIP)–like condition: there exists $\delta \in (0,1)$ such that, for every symmetric matrix E of rank at most $2r$, then with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$ we have:

$$\|XX^\top - M^*\|_F \leq \frac{1}{\sqrt{1-\delta}}\sqrt{\frac{L^*_{\text{combined}}}{\lambda} + \frac{1}{n}\epsilon^2}. \tag{S.2}$$

The detailed proof and simplified analysis are provided in Appendix S.2. We can use similar results as above to show the result. More detailed analysis will be provided in the later version.

S.2. Proof of Theorem S.1

Proof. The composite loss with random noise is given in (8.1):

$$\begin{aligned}&\mathbf{L}_{\text{combined}}(X)\\&= (1-\lambda)\cdot n^{-1}\sum_{i=1}^{n}\left(-\log\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{((Y_j-(\mathcal{A}(XX^\top)_j))-(Y_i-\mathcal{A}(XX^\top)_i))^2}{h^2}\right)\right)\\&+\lambda\cdot\frac{1}{n}\sum_{i=1}^{n}\big(Y_i-\mathcal{A}(XX^\top)_i\big)^2.\end{aligned}\tag{S.3}$$

The composite loss is more robust to different setting. For example, when the $\|w\|$ is small (MSE is better, or when $\|w\|$ is medium, (ours is large).

$$r_i = Y_i - \mathcal{A}(XX^\top)_i, \quad i = 1,\dots,n.\tag{S.4}$$

For convenience, assume that there exists a constant $B>0$ such that $r_i^2 \le B, \quad$ for all $i=1,\dots,n$. We recall that the loss is defined as:

$$\begin{aligned}L_{\text{combined}}(X) = (1-\lambda)\cdot\frac{1}{n}\sum_{i=1}^{n}\Big[-\log\Big(\frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{\big((r_j-r_i)^2\big)}{h^2}\Big)\Big)\Big]\\+\lambda\cdot\frac{1}{n}\sum_{i=1}^{n}r_i^2.\end{aligned}\tag{S.5}$$

We now bound each term separately. Since $\frac{1}{n}\sum_{i=1}^{n}r_i^2 \le B$, it follows that $\lambda\cdot\frac{1}{n}\sum_{i=1}^{n}r_i^2 \le \lambda B$. For any i and j, because $r_i^2\le B$ we have $\|r_j-r_i\| \le \|r_j\|+\|r_i\| \le \sqrt{B}+\sqrt{B} = 2\sqrt{B}$. $(r_j-r_i)^2 \le (2\sqrt{B})^2 = 4B$. Then, for every i and j, $\exp\left(-\frac{(r_j-r_i)^2}{h^2}\right) \ge \exp\left(-\frac{4B}{h^2}\right)$. Hence, for each fixed i we obtain:

$$\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{(r_j-r_i)^2}{h^2}\right) \ge \frac{1}{n}\cdot n\exp\left(-\frac{4B}{h^2}\right) = \exp\left(-\frac{4B}{h^2}\right).\tag{S.6}$$

Taking the negative logarithm yields:

$$-\log\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{(r_j-r_i)^2}{h^2}\right)\right) \le -\log\left(\exp\left(-\frac{4B}{h^2}\right)\right) = \frac{4B}{h^2}.\tag{S.7}$$

Since this bound holds for each i, averaging over i gives:

$$\frac{1}{n}\sum_{i=1}^{n}\left[-\log\left(\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{(r_j-r_i)^2}{h^2}\right)\right)\right] \le \frac{4B}{h^2}.\tag{S.8}$$

Putting the bounds (5.1) into the combined loss (S.5), we conclude that:

$$L_{\text{combined}}(X) \le (1-\lambda)\frac{4B}{h^2} + \lambda B.\tag{S.9}$$

Assuming that we are in a matrix sensing/low-rank recovery setting. In such settings one assumes Assumption 2.3 and 2.1. The observed vector is modeled as $Y = \mathcal{A}(M^*) + w$, where w is a noise vector. The combined loss is given by

$$\begin{aligned}L_{\text{combined}}(X) = (1-\lambda)\cdot\frac{1}{n}\sum_{i=1}^{n}\left[-\log\Big(\frac{1}{n}\sum_{j=1}^{n}\exp\Big(-\frac{\big((Y_j-\mathcal{A}(XX^\top)_j)-(Y_i-\mathcal{A}(XX^\top)_i)\big)^2}{h^2}\Big)\Big)\right]\\+\lambda\cdot\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i-\mathcal{A}(XX^\top)_i\Big)^2.\end{aligned}\tag{S.10}$$

Further assume that $\mathcal{A}$ satisfies an RIP–like property; that is, there exists $\delta \in (0,1)$ such that for every symmetric matrix E of rank at most $2r$ one has $(1-\delta)\|E\|_F^2 \le \frac{1}{n}\|\mathcal{A}(E)\|_2^2 \le (1+\delta)\|E\|_F^2$. In our setting we set $E = XX^\top - M^*$. Denote the quadratic term in the loss by:

$$f_{\text{quad}}(X) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \mathcal{A}(XX^\top)_i\Big)^2 = \frac{1}{n}\|\mathcal{A}(XX^\top - M^*) - w\|_2^2. \tag{S.11}$$

For simplicity (and without loss of generality for a bound) imagine that the optimization is such that the loss is nearly achieved; then the quadratic term is small. In particular, note that:

$$\frac{1}{n}\|\mathcal{A}(XX^\top - M^*)\|_2^2 \le f_{\text{quad}}(X) + \frac{1}{n}\|w\|_2^2. \tag{S.12}$$

If the noise is controlled (or even absent) and if the combined loss is small near a stationary point (say, bounded above by L^*_{combined}), then in particular

$$\frac{1}{n}\|\mathcal{A}(XX^\top - M^*)\|_2^2 \le \frac{L^*_{\text{combined}}}{\lambda} + \frac{1}{n}\|w\|_2^2. \tag{S.13}$$

(In the combined loss the quadratic part appears weighted by λ.) For clarity we define

$$\varepsilon^2 := \frac{L^*_{\text{combined}}}{\lambda} + \frac{1}{n}\|w\|_2^2. \tag{S.14}$$

Thus,

$$\frac{1}{n}\|\mathcal{A}(XX^\top - M^*)\|_2^2 \le \varepsilon^2. \tag{S.15}$$

The RIP property (2.6) guarantees that:

$$(1-\delta)\|XX^\top - M^*\|_F^2 \le \frac{1}{n}\|\mathcal{A}(XX^\top - M^*)\|_2^2 \le \varepsilon^2. \tag{S.16}$$

Recalling the definition of ε, this yields

$$\|XX^\top - M^*\|_F \le \frac{1}{\sqrt{1-\delta}}\sqrt{\frac{L^*_{\text{combined}}}{\lambda} + \frac{1}{n}\|w\|_2^2}. \tag{S.17}$$

So we have:

$$\|XX^\top - M^*\|_F \le \frac{1}{\sqrt{1-\delta}}\sqrt{\frac{L^*_{\text{combined}}}{\lambda} + \frac{1}{n}\epsilon^2}. \tag{S.18}$$

with probability at least $\mathbb{P}(\|w\|_2 \le \epsilon)$. □

T. Detailed Experimental Results

In this section as above Section 9, we provide a detailed example to the theoretical results. Assume that $w \in \mathbb{R}^m$ is a $0.05/\sqrt{m}$-sub-Gaussian vector. According to Lemma 1 in [38], this choice of w satisfies:

$$1 - 2e^{-\frac{\epsilon^2}{16m\sigma^2}} \le \mathbb{P}(\|w\|_2 \le \epsilon), \tag{T.1}$$

with $\sigma = 0.05$. We consider the problem of 1-bit Matrix Completion, which is a low-rank matrix optimization task commonly appearing in recommendation systems with binary inputs [34, 35]. The objective function is given by 2.3. It is straightforward to verify that the new loss satisfies the assumptions in 2.3 and 2.1 with $\zeta_1 = 1$ and $\zeta_2 = 0$. In Figures 6 to 11, we numerically demonstrate and compare the bounds in Theorem 4.1 and Theorem 4.3, with parameters $n = 40$, $r = 5$.

Table 2: $\delta < 1/3$: Real vs. Numerical Errors for the new loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.5	0.131347	0.245299	67.189408	62.946586
0.6	0.010286	0.249747	67.795918	63.546119
0.7	0.010633	0.267247	70.130878	66.388579
0.8	0.220458	0.272450	70.810290	69.285569
0.9	0.127725	0.239321	66.365657	63.077514

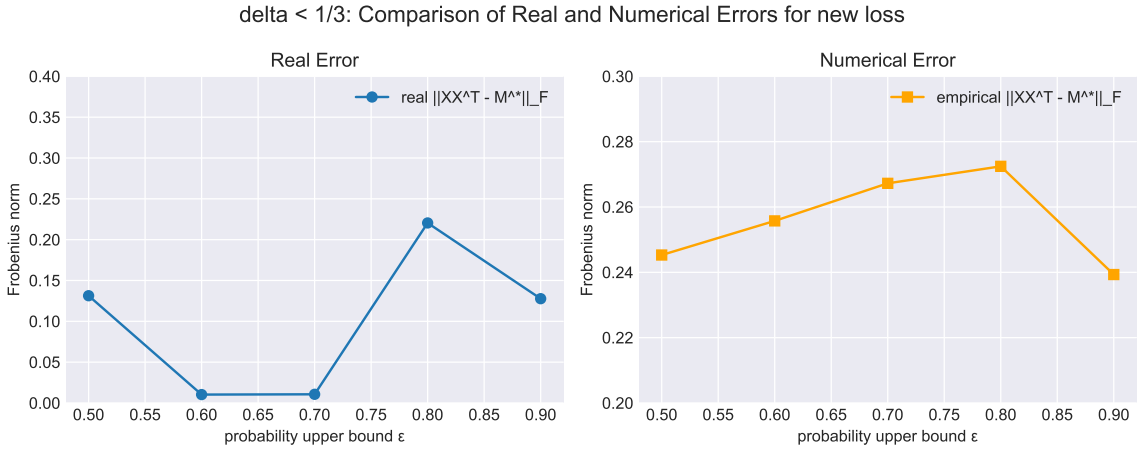

Figure 6: $\delta < 1/3$: Comparison of Real and Numerical Errors for new loss

Table 3: $\delta < 1/3$: Real vs. Numerical Errors for the MSE loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.6	0.152224	0.047391	112.673616	12.670269
0.65	0.117224	0.068880	111.240372	12.566897
0.7	0.135116	0.095493	111.873168	12.837108
0.75	0.199956	0.127546	121.697785	13.048538
0.8	0.173894	0.165297	115.219867	12.904689
0.85	0.175881	0.208955	109.081901	13.065476
0.9	0.144509	0.258679	112.820543	13.490991
0.95	0.164552	0.314585	114.623283	13.444282

Table 4: $\delta < 1/3$: Real vs. Empirical Errors for the composite loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.6	0.252242	0.542251	29.886229	21.379952
0.65	0.136733	0.126368	14.427459	13.312156
0.7	0.152275	0.121407	14.141430	13.780662
0.75	0.137491	0.122621	14.211950	12.999850
0.8	0.161383	0.123097	14.239475	12.907365
0.85	0.171834	0.116507	13.853113	12.747327
0.9	0.159968	0.827296	36.914869	18.623831
0.95	0.161281	0.124161	14.300877	13.291003

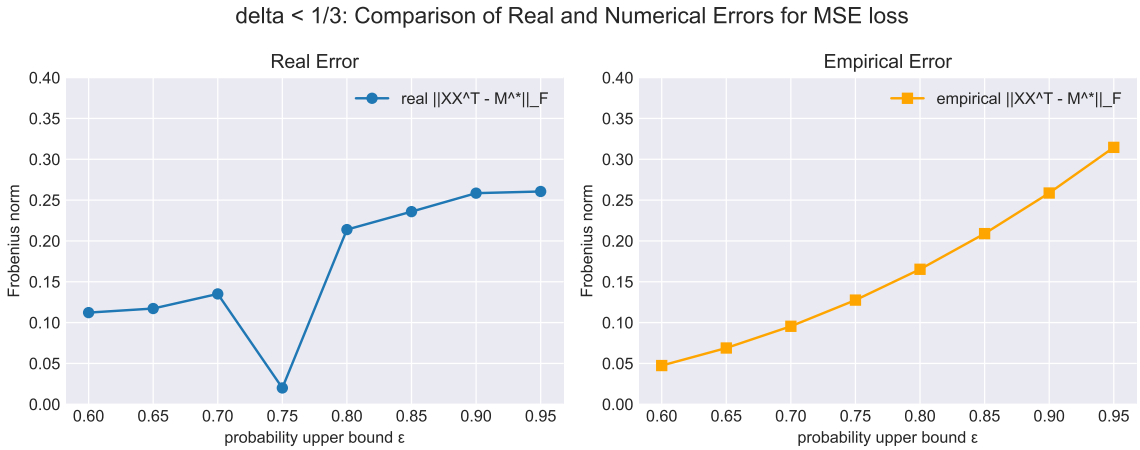

Figure 7: $\delta < 1/3$: Comparison of Real and Numerical Errors for MSE loss

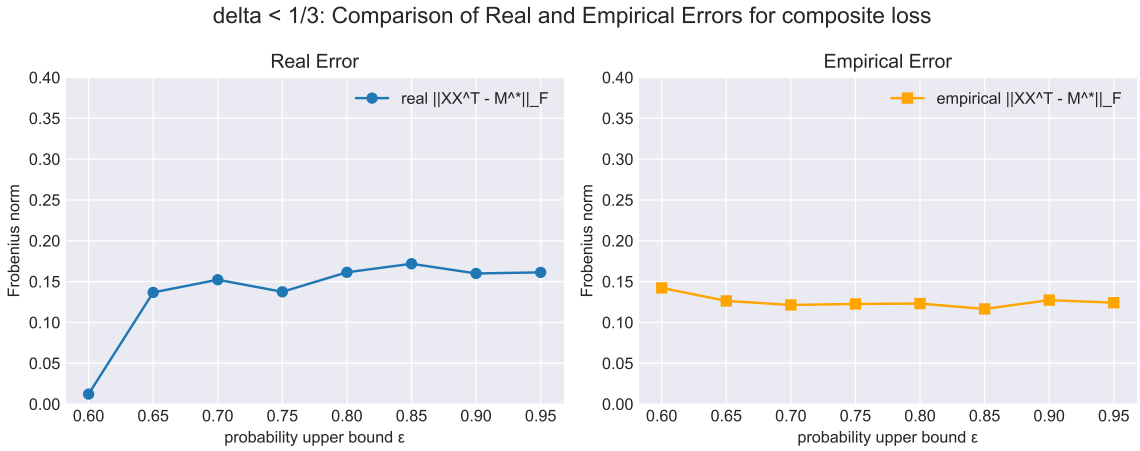

Figure 8: $\delta < 1/3$: Comparison of Real and Empirical Errors for composite loss

Table 5: $\delta > 1/3$: Real vs. Empirical Errors for the new loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.5	7.102596	0.311050	7566.029696	4847.329349
0.6	2.009672	0.345135	7969.801460	5300.450313
0.7	0.469574	0.233222	6551.454917	4318.232124
0.8	0.074324	0.287211	7270.322014	4534.227116
0.9	0.601774	0.262751	6953.848751	4461.188565

Table 6: $\delta > 1/3$: Real vs. Empirical Errors for MSE loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.5	0.52011	0.275556	22.519488	34.472289
0.6	0.60282	0.292817	23.214081	34.340328
0.7	0.70392	0.307916	23.805072	34.244936
0.8	0.86332	0.297505	23.399164	34.555703
0.9	0.75957	0.311414	23.939913	34.760302

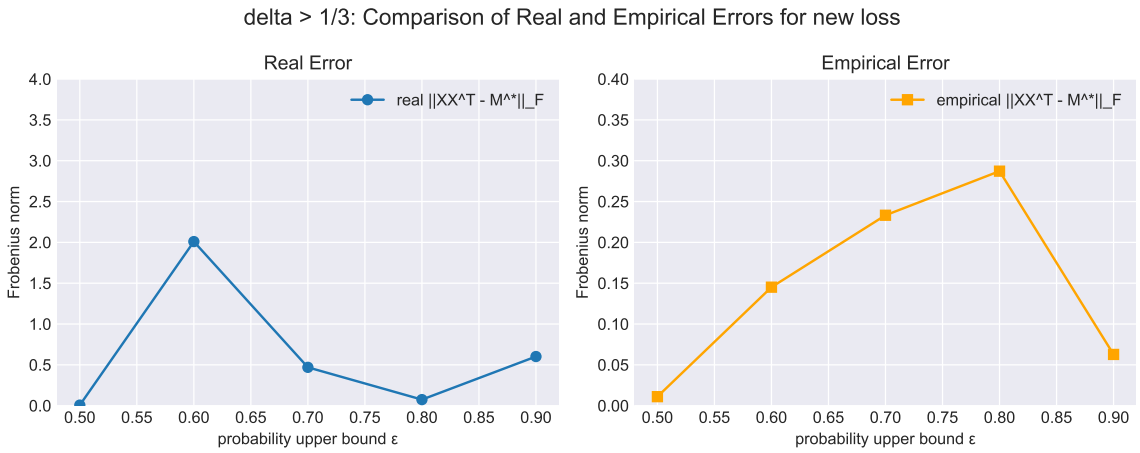

Figure 9: $\delta > 1/3$: Comparison of Real and Empirical Errors for new loss

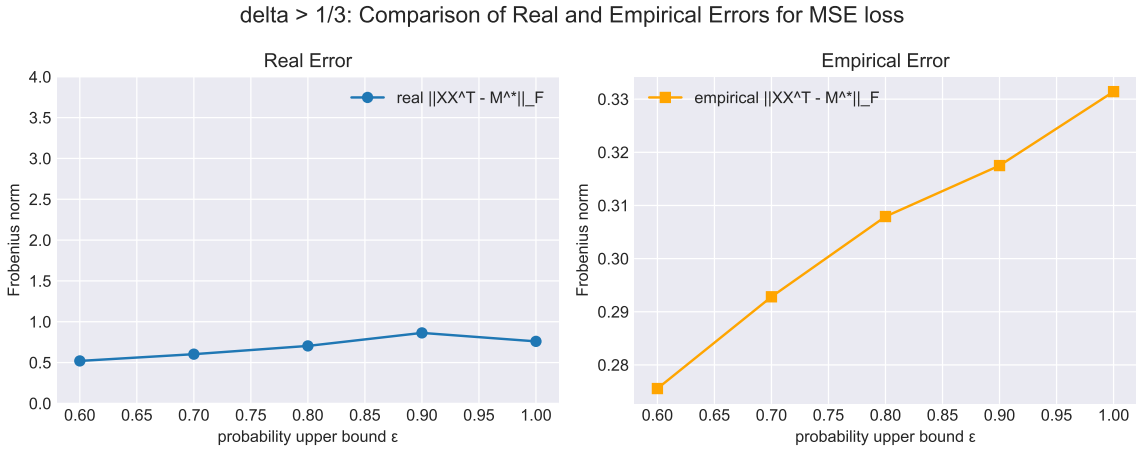

Figure 10: $\delta > 1/3$: Comparison of Real and Numerical Errors for MSE loss

Table 7: $\delta > 1/3$: Real vs. Empirical Errors for composite loss. Symbols: ϵ = probability upper bound, $E_{\text{real}} = \|XX^\top - M^*\|_F$, E_{emp} = empirical error, L = noise Lipschitz constant, H = noise Hessian constant.

ϵ	E_{real}	E_{emp}	L	H
0.5	0.100042	0.376493	26.322789	34.088333
0.6	0.086220	0.420021	28.553900	34.980704
0.7	0.391496	0.445266	28.626139	34.820591
0.8	0.140990	0.452202	28.398186	34.680517
0.9	0.340462	0.456870	28.996758	34.111456

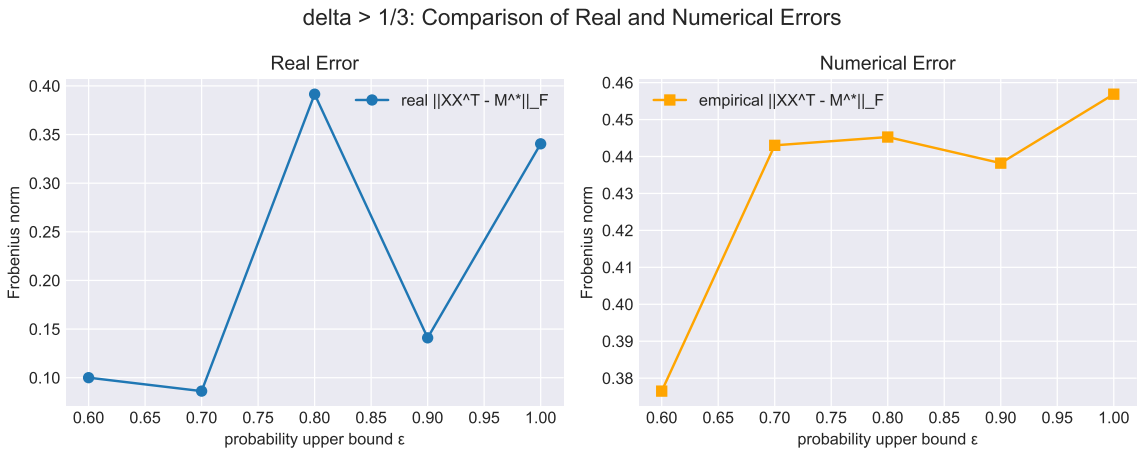

Figure 11: $\delta > 1/3$: Comparison of Real and Numerical Errors for composite loss