Finite-Sample Performance of Gradient Descent in Logistic Regression with Gaussian Design

Junren Chen
Department of Statistics
Columbia University
New York, NY 10027
jc6315@columbia.edu

Arya Mazumdar
Halıcıoğlu Data Science Institute
University of California, San Diego
La Jolla, CA, 92093
arya@ucsd.edu

Abstract

We consider the parameter estimation problem in logistic regression with Gaussian design: the estimation of a fixed unknown parameter $\theta^* \in \mathbb{R}^d$ ($\|\theta^*\|_2 \geq 1$) from n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \sim N(0, I_d)$ and $y_i|x_i \sim \text{Bernoulli}(1/(1+\exp(-x_i^\top \theta^*)))$. Our main aim is to characterize the finite-sample estimation performance and convergence behavior of gradient descent (GD) on the maximum likelihood objective (i.e., the logistic loss). Under small $O(1)$ stepsize and 0 initialization, we show that GD linearly converges to a small neighborhood of θ^* achieving an ℓ_2 error of order $O(\sqrt{\|\theta^*\|_2^5 d/n})$. This substantially goes beyond existing theoretical results that lack non-asymptotic estimation error rate and exhibit much slower parameter convergence. We also establish a *faster* local linear convergence to the same statistical error under a large $\Theta(\|\theta^*\|_2)$ stepsize. The main technical component is to show that the gradient of the logistic loss satisfies a certain approximate invertibility condition (AIC). To that end, we uniformly control the deviation of the gradient from its population counterpart by covering and peeling arguments, and then show that the population GD is a contraction by a delicate analysis based on the eigenvalues of population Hessian matrices. Finally, we build upon the recent work [23] and devise a novel efficient estimator that attains a sharper rate in high dimensions. This indicates that the existing non-asymptotic guarantees exhibit sub-optimal dependence on $\|\theta^*\|_2$, and that in many regimes $\Theta(\sqrt{\|\theta^*\|_2 d/n})$ is the tight estimation error rate. Numerical examples are provided to corroborate our theoretical results.

1 Introduction

Logistic regression is a foundational model in statistics, optimization and machine learning. Given $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0,1\}$ with binary responses, it assumes the existence of some underlying parameter $\theta^* \in \mathbb{R}^d$ such that the data follows $y_i|x_i \sim \text{Bernoulli}(s(x_i^\top \theta^*))$, $i \in [n]$ with the sigmoid function $s(a) = 1/(1+e^{-a})$, meaning that $\mathbb{P}(y_i = 1) = 1/(1+e^{-x_i^\top \theta^*})$ and $\mathbb{P}(y_i = 0) = 1/(1+e^{x_i^\top \theta^*})$. A standard estimator of the true parameter θ^* is the maximum likelihood estimator (MLE) [24] that minimizes the logistic loss function [1]

$$L(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\log(e^{x_i^\top \theta}+1) - y_i x_i^\top \theta\big),$$

[1] Throughout this paper, we adopt the convention that the responses $\{y_i\}_{i=1}^n$ are $\{0,1\}$-valued. One can readily transfer our development to the (equivalent) convention of $\{-1,1\}$-valued responses $\{\tilde{y}_i\}_{i=1}^n$ (e.g., [13, 23, 38]) by letting $y_i = \frac{1}{2}(\tilde{y}_i+1)$, under which the logistic loss reads $L(\theta) = \frac{1}{n}\sum_{i=1}^n \log(1+e^{-\tilde{y}_i x_i^\top \theta})$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

given by $\hat{\theta}_{mle} = \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$. Since MLE does not admit a closed form, one typically can minimize $L(\theta)$ via gradient descent (GD) with some stepsize η,

$$\theta_{t+1} = \theta_t - \eta\nabla L(\theta_t), \quad t = 0, 1, \cdots, \tag{GD}$$

where $\nabla L(\theta) = \frac{1}{n}\sum_{i=1}^n (s(x_i^\top\theta) - y_i)x_i$. The problems of *estimation* and *optimization* have been widely studied in logistic regression.

The estimation problem is of interests to statisticians and concerned with the performance of a certain estimator in approximating the underlying θ^*. Perhaps a more fundamental question is to determine the minimax optimal error rate. To answer these questions with an explicit error rate depending on the model parameters $(n, d, \|\theta^*\|_2)$,[2] existing works assume i.i.d. samples where x_i's are drawn from a specific distribution [18, 7, 13, 23, 16].

From an optimization perspective, a central question is to understand the capacity of (GD) in minimizing $L(\theta)$. In light of the convexity of $L(\theta)$, traditional optimization theory [28, Sections 1, 2] yields the global convergence of (GD) under small enough η, guaranteeing that $L(\theta_t)$ decreases monotonically. The works [32, 15] provided fine-grained characterizations under small stepsizes. A more recent line of works studied large stepsize GD in logistic regression [38, 37, 40, 35], referred to as the edge of stability regime [10].

This paper is concerned with both the estimation and optimization in logistic regression. Following [18, 7, 13, 23] we assume that the covariates $x_1, \cdots, x_n$ are i.i.d. standard Gaussian vectors and focus on the estimation of the true parameter θ^*.

Assumption 1. *$\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples with $x_i \sim N(0, I_d)$, and when conditioning on x_i, $y_i \sim \mathrm{Bernoulli}(s(x_i^\top\theta^*))$ for some $\theta^* \in \mathbb{R}^d$.*

1.1 Related Works

As it is unrealistic to cover the vast literature of logistic regression, we shall focus on two lines of works that are closest to our paper.

Parameter estimation. The most relevant works are [18, 7] that derived the non-asymptotic error rates for MLE under Assumption 1. Kuchelmeister and Van de Geer [18] first established the finite-sample error rate of minimizing $L(\theta)$ under a Probit model. It was proved [18, Theorem 2.1.1] that a constrained variant of MLE achieves ℓ_2 error of $O(\sqrt{\|\theta^*\|_2^3 \frac{d\log n}{n}})$ whenever $n \gtrsim \|\theta^*\|_2 d\log n$. More recently, Chardon et al. [7] established similar guarantee for MLE with refinement, showing [7, Theorem 1, Equation (19)] that MLE exists and achieves

$$\|\hat{\theta}_{\mathrm{mle}} - \theta^*\|_2 \lesssim \sqrt{\frac{\|\theta^*\|_2^3 d}{n}} \tag{1}$$

whenever $n \gtrsim \|\theta^*\|_2 d$. On a relevant passing note, [6] derived the phase transitions for the existence of MLE in high-dimensional asymptotics.

Another two closely related works [13, 23] studied the estimation of parameter direction $r^* := \frac{\theta^*}{\|\theta^*\|_2}$ through algorithms other than MLE. Under $\|\theta^*\|_2 \gtrsim 1$, [13] determined the minimax error rate as $\tilde{\Theta}(\max\{\sqrt{\frac{d}{n\|\theta^*\|_2}}, \frac{d}{n}\})$. This notably exhibits a transition to the classical halfspace learning result [20] as $\|\theta^*\|_2 \to \infty$. While computational tractability was not a consideration in the upper bound in [13] (it involves nonconvex minimization routines), more recently [23] addressed this by showing that a fast iterative algorithm attains the near optimal direction estimation error rate.

We emphasize all the above results study Gaussian covariates (cf. Assumption 1) as a benchmark. Extension to sub-Gaussian covariates is highly challenging in general [7, 6].

Convergence of (GD). The optimization path of (GD) has been characterized [32, 15] under small stepsizes. More recently larger stepsizes have been extensively studied, with positive results under separable data [38, 37, 39, 41] while negative result under non-separable data [26, 25]. These results typically work with data obeying separability [32, 38, 37, 41, 3] and bounded covariates $\|x_i\|_2 \le 1$ [15, 38, 37]. Yet, such optimization works do not touch on the estimation problem, i.e., how well GD

[2] Here, $\|\theta^*\|_2$ indicates how likely y_i is flipped to $-\mathrm{sign}(x_i^\top\theta^*)$ and hence represents the signal-to-noise ratio (SNR).

estimates the true parameter θ^* (indeed some of these works do not assume existence of θ^*). Also, the convergence in parameter, if obtained, is often very slow, e.g., the direction convergence rates of [32, 15] are no faster than $O(\frac{1}{\log t})$.

[17, 16] studied algorithms analogous to perceptron, called GLM-tron and Isotron, with the GLM-tron reducing to (GD) with $\eta = 1$ when applied to logistic regression. Both works considered covariates independently drawn from the unit ℓ_2-ball. While these methods appear more versatile, the closest result [16, Theorem 1] requires $\tilde{O}(\sqrt{n})$ iterations to reach an error rate no faster than $O(n^{-1/4})$.

1.2 Our Contributions

Despite the huge literature, a few questions remain open. As a practically relevant algorithm, how well does (GD) estimate θ^*? Under Gaussian design, are there stronger convergence behaviors beyond existing results (e.g., [15, 16, 38])? Statistically, is the non-asymptotic error rate of the MLE in (1) optimal? If not, what is the minimax optimal error rate then?

We answer these questions by establishing the following results:

- Under a small stepsize $\eta \in (0, 8)$, (GD) with $\theta_0 = 0$ linearly converges to θ^* and achieves an error rate of $O(\sqrt{\|\theta^*\|_2^5 d/n})$ (cf. Theorem 1). Both the estimation error rate and fast linear convergence exceed existing results for GD in logistic regression (cf. Remarks 1, 2).
- Under a large stepsize $\eta \asymp \|\theta^*\|_2$, (GD) locally achieves an acceleration–a faster linear convergence with a sharper per-iteration contraction factor (cf. Theorem 2).
- A novel efficient estimator achieves an error rate that is sharper than the best known rate $O(\sqrt{\|\theta^*\|_2^3 d/n})$ achieved by MLE in high dimensions (cf. Theorem 3). This indicates that existing estimation rates are sub-optimal in terms of the dependence on $\|\theta^*\|_2$, and that the minimax optimal rate is $\Theta(\sqrt{\|\theta^*\|_2 d/n})$ in some regimes (cf. Remarks 4, 5).

Notation. We denote the gradient of the logistic loss by $h_{\theta^*}(\theta) := \frac{1}{n}\sum_{i=1}^n \big(s(x_i^\top \theta) - y_i\big)x_i = \nabla L(\theta)$. Let $\|u\|_2$ be the ℓ_2 norm of u. Define the unit sphere $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ and unit ball $\mathbb{B}_2^d = \{u \in \mathbb{R}^d : \|u\|_2 \le 1\}$. For $r > 0$ and $\theta_0 \in \mathbb{R}^d$, let $\mathbb{B}_2^d(r) = r\mathbb{B}_2^d$ and $\mathbb{B}_2^d(\theta_0; r) = \theta_0 + r\mathbb{B}_2^d$. Let $\|M\|_{\mathrm{op}}$ denote the operator norm of a matrix M. For a random variable X, define the sub-Gaussian and sub-exponential norms by $\|X\|_{\psi_2} := \inf\{K > 0 : \mathbb{E}[\exp(X^2/K^2)] \le 2\}$, $\|X\|_{\psi_1} := \inf\{K > 0 : \mathbb{E}[\exp(|X|/K)] \le 2\}$, respectively. We write $T_1 = O(T_2)$ (or $T_1 \lesssim T_2$) to denote $T_1 \le CT_2$, and write $T_1 = \Omega(T_2)$ (or $T_1 \gtrsim T_2$) to denote $T_1 \ge CT_2$. We also write $T_1 = \Theta(T_2)$ (or $T_1 \asymp T_2$) if both $T_1 = O(T_2)$ and $T_1 = \Omega(T_2)$ hold. We use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$ to hide log factor in the sample size n. More notation will be introduced as needed.

Overview. We provide the main results of GD and an overview of the proof techniques in Section 2. A novel estimator that achieves sharper statistical error rate is developed in Section 3. We provide numerical examples in Section 4 to corroborate our theory, and give concluding remarks to close the paper in Section 5. The complete proofs of Theorems 1, 2, 3 appear in Appendices A, B, C, with supporting lemmas provided in Appendix D.

2 Gradient Descent

We assume that $\|\theta^*\|_2 \ge 1$ throughout this paper (it can be relaxed to $\|\theta^*\|_2 \ge c$ for some small constant $c > 0$). Note that we cover the strong signal regime with $\|\theta^*\|_2 \gg 1$ as in recent works [13, 18, 7, 23].

Assumption 2. *$\theta^* \in \mathbb{R}^d$ is an unknown fixed parameter satisfying $\|\theta^*\|_2 \ge 1$.*

We formalize the gradient descent in the following Algorithm 1, which is also a specific instance of GLM-tron [16]. Due to the simplicity this has been one of the most commonly used algorithm in logistic regression, and we study under Gaussian designs the following question:

What are the finite-sample estimation performance and convergence behavior of Algorithm 1?

While a large body of convergence results exist, little is known about the estimation performance of Algorithm 1. We shall see that our results are novel to the literature in many ways.

Algorithm 1 Gradient Descent for Logistic Regression

Require: data $\{(x_i, y_i)\}_{i=1}^n$, initialization θ_0, stepsize η

1: **for** $t = 0, 1, 2, \cdots$ **do**

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{n} \sum_{i=1}^n (s(x_i^\top \theta_t) - y_i) x_i$$

2: **end for**

Output: $\{\theta_t\}_{t \geq 0}$

2.1 Small Stepsize

Traditional convex optimization theory [28] guarantees that Algorithm 1 converges globally under $\eta \leq 2/\beta$, where the smoothness parameter $\beta := \sup_{\theta \in \mathbb{R}^d} \|\nabla^2 L(\theta)\|_{\mathrm{op}} = \sup_{\theta \in \mathbb{R}^d} \|\frac{1}{n} \sum_{i=1}^n s'(x_i^\top \theta) x_i x_i^\top\|_{\mathrm{op}}$. By $s'(a) \leq s'(0) = \frac{1}{4}$ we obtain $\beta = \|\frac{1}{4n} \sum_{i=1}^n x_i x_i^\top\|_{\mathrm{op}}$, and under Gaussian designs with $n \gtrsim d$ one has $\beta \approx \frac{1}{4}$ (cf. Lemma 13). Therefore GD with any fixed $\eta \in (0, 8)$ converges globally and achieves a monotonically decreasing logistic loss, referred to as the stable regime.

But how close are the iterates to the true parameter θ^*? In the same stable regime we answer this question by the following result (Theorem 1). Note that we put $\eta \in [0.1, 7.9]$ to avoid the dependence of the contraction factor on η, but it can be observed from our proof that, under $\eta \in (0, 8)$, we obtain a contraction factor of $1 - \tilde{c}\|\theta^*\|_2^{-3} \eta(8 - \eta)$ for some universal constant $\tilde{c}$. Also, our result is essentially a *global* convergence (although we consider $\theta_0 = 0$ for concreteness). We show that $\{\theta_t\}_{t \geq 0}$ *linearly* converges to a small radius-$O(\sqrt{\|\theta^*\|_2^5 d/n})$ ℓ_2 ball surrounding θ^* within $O(\|\theta^*\|_2^3 \log \frac{n}{d})$ iterations, whenever $n = \tilde{\Omega}_{\|\theta^*\|_2}(d)$.

Theorem 1 (Small stepsize GD)**.** *Under Assumptions 1–2, for any $\eta \in [0.1, 7.9]$, there exist some universal constants $C, c, \tilde{C} > 0$ such that the following holds. If*

$$n \geq C\Big(\|\theta^*\|_2^6 d \log n + \|\theta^*\|_2^6 \log \|\theta^*\|_2\Big), \tag{2}$$

then with probability at least $1 - 2e^{-d}$, $\{\theta_t\}_{t \geq 0}$ generated by Algorithm 1 with $\theta_0 = 0$ and stepsize η satisfies

$$\|\theta_t - \theta^*\|_2 \leq \left(1 - \frac{c}{\|\theta^*\|_2^3}\right)^t \|\theta^*\|_2 + \tilde{C}\sqrt{\frac{\|\theta^*\|_2^5 d}{n}}, \quad \forall t = 0, 1, \cdots. \tag{3}$$

To position our result we provide two remarks from the angles of estimation and optimization.

Remark 1. *To our best knowledge, the only existing estimation guarantee of GD with logistic loss is [16, Theorem 1] for bounded designs (i.e., $\|x_i\|_2 \leq 1$), and our Theorem 1 provides the first non-asymptotic estimation error rate to GD in logistic regression with Gaussian designs. We believe that there are major differences between Gaussian designs and bounded designs, for instance, the bounded designs bypass the curse of dimensionality and as a result d does not appear in their error rate. Moreover, a naive comparison finds that our Theorem 1 offers the following improvements: (i) Our error decay rate $O(n^{-1/2})$ is faster than their $O(n^{-1/4})$; (ii) To attain the final error rates our iteration complexity is $\tilde{O}(\|\theta^*\|_2^3)$ while theirs reads $\tilde{O}(\|\theta^*\|_2 \sqrt{n})$.*

Remark 2. *Theorem 1 is also interesting from the optimization aspect in terms of the fast linear convergence in parameter. Specifically, recent works [32, 15, 27] also established the convergence of $\theta_t / \|\theta_t\|_2$ to a specific point, but the convergence rates, such as $O(\frac{\ln \ln t}{\ln t})$ from [15, Theorem 1], are much slower than ours. More are known about convergence in loss, such as the decreasing logistic loss with rate $E(\theta_t) := L(\theta_t) - \min_\theta L(\theta) = O(\frac{1}{t})$ [28], a comparable $E(\theta_t) = O(\frac{\log^2 t}{t})$ of recent work [15], and the more recent linear convergence $E(\theta_t) < 0.1 \min_\theta L(\theta) + \varepsilon$ whenever $t = \Omega(\log \frac{n}{\varepsilon})$ [3]. In light of a standard estimate $\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \leq \frac{1}{8}\|\theta - \theta^*\|_2^2$ (where $\mathcal{L}(\theta) = \mathbb{E}L(\theta)$ is the population risk), our Equation (3) immediately yields the linear convergence in excess risk*

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \leq \left(1 - \frac{c'}{\|\theta^*\|_2^3}\right)^t \|\theta^*\|_2^2 + C' \frac{\|\theta^*\|_2^5 d}{n}, \quad \forall t = 0, 1, \cdots. \tag{4}$$

We expect that $\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)$ is comparable to $L(\theta_t) - \min_\theta L(\theta)$ up to some concentration error terms and $\mathcal{L}(\theta_{mle}) - \mathcal{L}(\theta^*) = O(d/n)$ [7, Theorem 1]. Therefore, our linear convergence in risk is comparable to [3] and faster than [28, 15].

Continuing from Remark 2, we however emphasize that these convergence results are established under different assumptions and therefore a close comparison may not be possible. In particular, the aforementioned prior results do not rely on Gaussian data, but rather some of them assume bounded designs and/or separable data (possibly with a margin) [3, 15, 32, 27]. We believe that all these results are complementary to each other.

2.2 Large Stepsize

Due to the ubiquitousness of large stepsize GD in machine learning optimization [10], large stepsize GD for logistic regression receives significant recent interests [38, 37, 41, 39]. Our second main result, the following Theorem 2, establishes the local linear convergence of GD with a stepsize that can be much larger than $2/\beta \approx 8$.

Theorem 2 (Large stepsize GD)**.** *Under Assumptions 1–2, there exist some universal constants $c_0, c, C, \tilde{C} > 0$ such that the following holds. Fix any $\eta > 0$ such that*

$$\frac{c}{\|\theta^*\|_2^2 q'(\|\theta^*\|_2)} < \eta < \frac{2 - c\|\theta^*\|_2^{-2}}{m(\|\theta^*\|_2)} \tag{5}$$

where $m(\tau) = \mathbb{E}_{g\sim N(0,1)}[s'(\tau g)]$, $q(\tau) = \mathbb{E}_{g\sim N(0,1)}[s(\tau g)g]$ and $q'(\tau) = \mathbb{E}_{g\sim N(0,1)}[s'(\tau g)g^2]$, if

$$n \geq C\bigg(\|\theta^*\|_2^6 d \log n + \|\theta^*\|_2^7 d\bigg), \tag{6}$$

then with probability at least $1 - 2e^{-d}$, $\{\theta_t\}_{t\geq 0}$ generated by Algorithm 1 with $\theta_0 \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$ and stepsize η satisfies

$$\|\theta_t - \theta^*\|_2 \leq \left(1 - \frac{c}{\|\theta^*\|_2^2}\right)^t \frac{c_0}{\|\theta^*\|_2} + \tilde{C}\sqrt{\frac{\|\theta^*\|_2^5 d}{n}}, \quad \forall t = 0, 1, \cdots. \tag{7}$$

Moreover, a specific stepsize $\eta = \frac{1}{m(\|\theta^*\|_2)}$ satisfies (5).

In light of $m(\tau) \asymp 1/\tau$ and $q'(\tau) \asymp 1/\tau^3$ for $\tau \geq 1$ (cf. Lemma 15), the stepsize in Equation (5) is of order $\|\theta^*\|_2$, which is hence much larger than $\eta \in (0, 8)$ (the stable regime) when $\|\theta^*\|_2 \gtrsim 1$ (e.g., the specific $\eta = 1/m(\|\theta^*\|_2)$ already exceeds 8 under $\|\theta^*\|_2 \geq 2$). Under $\|\theta^*\|_2 \gg 1$, the large stepsize offers a notable acceleration, achieving a per-iteration contraction factor $1 - \Theta(\|\theta^*\|_2^{-2})$ that improves on $1 - \Theta(\|\theta^*\|_2^{-3})$ in Theorem 1. Yet the costs are that an accurate θ_0 and a slightly higher sample complexity are needed, and that the stepsize η requires some tuning according to $\|\theta^*\|_2$. The estimation error rate remains the same.

Similar to Equation (4) in Remark 2, we also have a linear convergence in population risk

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \leq \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)^t \frac{c_0'}{\|\theta^*\|_2^2} + C'\frac{\|\theta^*\|_2^5 d}{n}, \quad \forall t = 0, 1, \cdots. \tag{8}$$

We now put our result in perspective.

Remark 3. *[38] first showed that large stepsize GD minimizes logistic loss and attains $E(\theta_t) = L(\theta_t) - \min_\theta L(\theta) = O(1/t)$, then [37] provided finer optimization-theoretic characterizations and established a faster convergence $E(\theta_t) = \tilde{O}(1/t^2)$ (cf. their Corollary 2). See also further developments using adaptive stepsizes [41] or in a regularized setting [39]. While the faster convergence is commonly uncovered as a benefit of large stepsize, we emphasize that the settings in this line of works and in our paper are essentially different: [38, 37, 41, 39] assume bounded covariates and separable data with margin γ (see, e.g., [37, Assumption 1]), while we treat Gaussian designs and, under the sample complexity (6), our data $\{(x_i, y_i)\}_{i=1}^n$ is provably non-separable (cf. [7, Theorem 1]). A naive comparison finds that our (8) is faster than [38, 37] and comparable to [41] using adaptive large stepsizes.*

Concerning large stepsizes, there is an essential difference between separable data and non-separable data [35, 39]: for linearly separable data the logistic loss is minimized at infinity, which enables arbitrarily large stepsize (e.g., [37]); in contrast, under non-separable data (and mild condition) logistic loss admits a finite, unique minimizer, thus the GD iterates are necessarily unstable if η exceeds some threshold [12]. Indeed the behaviors of large stepsize GD in non-separable data are much more intricate, for instance, under $d \geq 2$, it was shown [26, Theorem 3] that GD with $\eta = c\|\nabla^2 L(\theta_{mle})\|_{\mathrm{op}}^{-1}$, $c \in (0, 1]$ may converge to a stable cycle. Our Theorem 2 complements the negative results of [26] with a positive radius-$\Theta(\|\theta^*\|_2^{-1})$ local convergence under Gaussian designs.

Finally, while the different assumptions may forbid a close comparison, we again emphasize that the finite-sample statistical rate, $O_{\|\theta^*\|_2}(\sqrt{d/n})$, is unique to our paper (see also Remark 1).

2.3 Technical Overview

The proofs of Theorems 1, 2 essentially depart from the related works that we have compared with. Our major technical component is a structured condition — the *approximate invertibility condition* (AIC) of the gradient $h_{\theta^*}(u)$ — which implies the linear convergence of the GD iterates to a small neighborhood of the true parameter θ^*, hence simultaneously settling the estimation and convergence. We note in passing that recent works [22, 23, 8, 11, 1, 9] have leveraged similar structured conditions to analyze different nonlinear regression problems.

Specifically, our AIC in Theorem 1 states that, for $\eta \in [0.1, 7.9]$ and some universal constants c, C,

$$\|u - \theta^* - \eta \cdot h_{\theta^*}(u)\|_2 \leq \left(1 - \frac{c}{\|\theta^*\|_2^3}\right)\|u - \theta^*\|_2 + C\left(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}\right) \tag{AIC}$$

holds for all $u \in \mathbb{B}_2^d(2\|\theta^*\|_2)$. Note that the (AIC) uniformly controls how much the actual gradient descent step, $\eta \cdot h_{\theta^*}(u)$, deviates from the ideal invertibility step $u - \theta^*$.[3] As such the AIC alone is enough to yield the claimed guarantee in Theorem 1 (cf. Appendix A.3). Similarly, a "local" AIC over $\mathbb{B}_2^n(\theta^*; \frac{c_0}{\|\theta^*\|_2})$ yields the local convergence of Theorem 2. See (AIC2) in Appendix B.3.

The main technical work lies in the establishment of (AIC) under Gaussian designs. To that end we start with a decomposition

$$\|u - \theta^* - \eta \cdot h_{\theta^*}(u)\|_2 \leq \eta \cdot \underbrace{\|h_{\theta^*}(u) - \mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_c} + \underbrace{\|u - \theta^* - \eta\mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_e}.$$

Bounding T_c. We start with a further decomposition $T_c \leq T_{c1} + T_{c2}$, where $T_{c1} := \big\|\frac{1}{n}\sum_{i=1}^n \big(\big(s(x_i^\top\theta^*) - y_i\big)x_i - \mathbb{E}\big[\big(s(x_i^\top\theta^*) - y_i\big)x_i\big]\big)\big\|_2$ and $T_{c2} := \big\|\frac{1}{n}\sum_{i=1}^n \big(\big(s(x_i^\top u) - s(x_i^\top\theta^*)\big)x_i - \mathbb{E}\big[\big(s(x_i^\top u) - s(x_i^\top\theta^*)\big)x_i\big]\big)\big\|_2$. It turns out that establishing sharp bounds on the two terms requires rather different techniques. For T_{c1}, a direct application of the sub-Gaussianity of $\frac{1}{n}\sum_{i=1}^n\big(s(x_i^\top\theta^*) - y_i\big)x_i - \mathbb{E}\big[\big(s(x_i^\top\theta^*) - y_i\big)x_i\big]$ does not give the tight bound since this overlooks the fact that the variance of the multiplier, $s(x_i^\top\theta^*) - y_i$, decreases as $\|\theta^*\|_2$ increases. To capture the effect of $\|\theta^*\|_2$ we have to use the Bernstein's inequality under moment conditions along with a covering argument (cf. Appendix A.1.1). In contrast, the key feature of T_{c2} is that it becomes smaller for u being close to θ^*, thus requiring the technique of localization. A standard approach to this end is via symmetrization, Talagrand's contraction and concentration of empirical process (e.g., [19, 4]), while following this route leads to intricacy in the contraction due to the product process structure of T_{c2}. We instead return to a more elementary peeling argument and decompose $\mathbb{B}_2^d(2\|\theta^*\|_2)$ into a small number of shells, over each of which $\|u - \theta^*\|_2$ either roughly remains at a constant (up to a factor of 2) or is too small to have an impact. We then control T_{c2} over each shell via covering argument. See Appendix A.1.2 for details.

Bounding T_e. This part of analysis is more elementary but remains rather subtle. Setting the stage, we first (i) use Stein's identity [33, 34] to write the population gradient $\mathbb{E}[h_{\theta^*}(u)]$ as $F(u) - F(\theta^*)$ where $F(u) = m(\|u\|_2)u$, and $m(\tau) = \mathbb{E}_{g\sim N(0,1)}[s'(\tau g)]$ is defined as in Theorem 2, and then (ii) introduce a useful integral representation $F(u) - F(\theta^*) = \int_0^1 \nabla F(\theta^* + t(u - \theta^*))(u - \theta^*)\,dt$ (cf.

[3] Suppose we were at u, then taking a descent step of $u - \theta^*$ brings us to $u - (u - \theta^*) = \theta^*$, which is the desired parameter.

Lemma 18). Based on this, the main bulk of techniques lies an eigenvalue analysis of $\nabla F(z) = m(\|z\|_2)I_d + \frac{m'(\|z\|_2)}{\|z\|_2} zz^\top$, the population Hessian matrix, along with a careful account of how well the linear map $\nabla F(\theta^*)(u - \theta^*)$ approximates $F(u) - F(\theta^*)$. Note that the eigenvalues of $\nabla F(z)$ are $q'(\|z\|_2)$ in the direction of $z/\|z\|_2$ and $m(\|z\|_2)$ in the orthogonal space. Due to $m(\tau) \asymp \tau^{-1}$, $q'(\tau) \asymp \tau^{-3}$ for $\tau \geq 1$ (cf. Lemma 15), $\nabla F(\|\theta^*\|_2)$ is highly anisotropic when $\|\theta^*\|_2 \gg 1$, which is exactly what dictates the different convergence behaviors under small and large stepsizes (cf. Appendices A.2 and B.2).

3 Sharper Error Rate

In this section, we ask a more fundamental question: *What is the best possible error rate for the estimation of θ^*?* In particular, we explore to what extent existing non-asymptotic error rates — including (1) for MLE [18] and our $O(\sqrt{\|\theta^*\|_2^5 d/n})$ for GD — are (sub-)optimal. It was shown (e.g., [2]) that the scaling of (n, d), namely $\sqrt{d/n}$, is sharp. Therefore it remains to investigate the dependence on $\|\theta^*\|_2$, in which the MLE rate is sharper than our GD rate. We show that even the MLE rate $O(\sqrt{\|\theta^*\|_2^3 d/n})$ due to [7] exhibits sub-optimal dependence on $\|\theta^*\|_2$. In fact, in this section we devise a computationally efficient estimator that attains a sharper estimation error rate in high dimensions (cf. Remark 4). In some regimes, this error rate reduces to $\tilde{O}(\sqrt{\|\theta^*\|_2 d/n})$ and is nearly optimal in light of a matching lower bound (cf. Remark 5).

We shall start with a potential caveat of GD and MLE in parameter estimation: both algorithms simultaneously estimate the direction $\theta^*/\|\theta^*\|_2$ and the norm $\|\theta^*\|_2$. This could be problematic because when $\|\theta^*\|_2$ increases, direction estimation becomes easier while norm estimation becomes harder. In the extreme regime with $\|\theta^*\|_2 \to \infty$, logistic regression reduces to the halfspace learning problem where norm estimation is not possible [20, 14, 22, 13]. Therefore, under $\|\theta^*\|_2 \gg 1$, the norm estimation error dominates the direction estimation error and dictates the final error rate.

To overcome this limitation, our main idea is to estimate the direction and norm separately. For direction estimation we naturally invoke the recent near-optimal direction estimator due to [23], which *does not require knowledge of* $\|\theta^*\|_2$ and, intriguingly, is essentially a *(sub)gradient descent with a different loss function*, the ReLU loss; more details are provided in Appendix E.1. We restate a non-asymptotic direction estimation bound from [23] (see Appendix E.2), which shows an improved estimate under larger $\|\theta^*\|_2$. In this section, we use $\mathrm{polylog}(n)$ to denote a generic factor bounded by $(\log n)^C$ for some universal constant $C > 0$.

Lemma 1. *[23, Corollary 7 & Theorem 5] Under Assumptions 1–2 and fix some $T_0 \geq \log_2 \log_2 \frac{n}{d}$, if $n \geq Cd \operatorname{polylog} n$ with large enough universal constant C, then with probability at least $1 - e^{-n}$, $\hat{r}_{T_0}$ obtained by running $\tilde{r}_{t+1} = \hat{r}_t - \frac{\sqrt{2\pi}}{n} \sum_{i=1}^n \big(\frac{\mathrm{sign}(x_i^\top \hat{r}_t)+1}{2} - y_i\big)x_i$, $\hat{r}_{t+1} = \frac{\tilde{r}_{t+1}}{\|\tilde{r}_{t+1}\|_2}$, $t \geq 0$ (or equivalently, Equation (11) below) with an arbitrary $r_0 \in \mathbb{S}^{d-1}$ satisfies*

$$\left\| \hat{r}_{T_0} - \frac{\theta^*}{\|\theta^*\|_2} \right\|_2 \leq \mathrm{polylog}(n) \left(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n} \right). \tag{9}$$

Before moving on to the norm estimate, we mention that we utilize sample splitting in this estimator. Particularly, for some $\nu \in [0.1, 0.9]$ such that νn is an integer, we utilize $\{(x_i, y_i)\}_{i=1}^{\nu n}$ to obtain the direction estimator $\hat{\nu}_{T_0}$ and then we further estimate $\|\theta^*\|_2$ using $\{(x_i, y_i)\}_{i=\nu n+1}^n$ (and $\hat{\nu}_{T_0}$). This ensures some independence and facilitates the analysis, and has been applied to different problems [29, 21]. Yet we believe that the practical performance of the estimator does not rely on this.

To estimate the norm, we use the remaining samples to construct the so-called "average" estimator $\hat{v} = \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^n y_i x_i$ [30, 31] whose expectation is given by (cf. Lemma 23)

$$v := \mathbb{E}[\hat{v}] = \mathbb{E}[y_i x_i] = q(\|\theta^*\|_2)\theta^*/\|\theta^*\|_2 \tag{10}$$

where $q(\tau) = \mathbb{E}[s(\tau g)g] = \mathbb{E}[s'(\tau g)\tau]$ has appeared in Theorem 2. By Lemma 21, q has an inverse on $(0, \infty)$, denoted by $q^{-1} : (0, \frac{1}{\sqrt{2\pi}}) \to (0, \infty)$. Hence we seek an estimate of $q(\|\theta^*\|_2)$ and then apply q^{-1} to get an estimate of $\|\theta^*\|_2$. Due to $\|v\|_2 = q(\|\theta^*\|_2)$ from (10), a naive estimate of $q(\|\theta^*\|_2)$ is $\|\hat{v}\|_2$, yet the concentration of $\|\hat{v}\|_2$ about $\|v\|_2$ is under Euclidean norm and necessarily involves dimension d. It turns out that a better idea is based on the inner product $\langle v, \theta^*/\|\theta^*\|_2 \rangle$,

which leads us to using the optimal direction estimator, $\hat{r}_{T_0}$, as a surrogate of $\theta^*/\|\theta^*\|_2$. Hence we propose to use $\langle \hat{v}, \hat{r}_{T_0} \rangle$ to estimate $q(\|\theta^*\|_2)$, and in turn $q^{-1}(\langle \hat{v}, \hat{r}_{T_0} \rangle)$ is our final norm estimator.

Putting pieces together, we formalize the estimator in Algorithm 2. Our main result in this section provides a non-asymptotic ℓ_2 error rate achieved by this novel estimator.

Algorithm 2 An Estimator with Improved Error Rate

Require: data $\{(x_i, y_i)\}_{i=1}^n$, iteration number T_0, initialization $\hat{r}_0 \in \mathbb{S}^{d-1}$, parameter $\nu \in (0,1)$ (such that νn is an integer)

1: **Step 1: Direction estimator**

2: **for** $t = 0, 1, 2, \cdots, T_0 - 1$ **do**

$$\hat{r}_{t+1} = \frac{\hat{r}_t - \frac{\sqrt{2\pi}}{\nu n} \sum_{i=1}^{\nu n} \big(\frac{\mathrm{sign}(x_i^\top \hat{r}_t)+1}{2} - y_i \big) x_i}{\|\hat{r}_t - \frac{\sqrt{2\pi}}{\nu n} \sum_{i=1}^{\nu n} \big(\frac{\mathrm{sign}(x_i^\top \hat{r}_t)+1}{2} - y_i \big) x_i\|_2} \tag{11}$$

3: **end for**

4: **Step 2: Norm estimator**

5: Compute $\hat{v} = \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i x_i$

6: Compute $\hat{\lambda} = q^{-1}(\langle \hat{r}_{T_0}, \hat{v} \rangle)$

Output: $\hat{\theta} = \hat{\lambda} \cdot \hat{r}_{T_0}$

Theorem 3 (Performance of Algorithm 2)**.** *Under Assumptions 1–2, if* $n \geq \mathrm{polylog}\, n \big(\|\theta^*\|_2 d + \|\theta^*\|_2^4 \big)$ *then with probability at least* $1 - n^{-1}$*,* $\hat{\theta}$ *obtained by Algorithm 2 with* $T_0 \geq \log_2 \log_2 \frac{n}{d}$*,* $\hat{r}_0 = (1, 0, \cdots, 0)^\top$ *and* $\nu \in [0.1, 0.9]$ *satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \mathrm{polylog}(n) \cdot \left(\|\theta^*\|_2^3 \sqrt{\frac{1}{n}} + \sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2^2 d}{n} \right). \tag{12}$$

Remark 4. *Note that the rate (12) improves on the best known rate, the* $O(\sqrt{\|\theta^*\|_2^3 d/n})$ *for MLE [18, 7], as long as* $\mathrm{polylog}(n)\|\theta^*\|_2^3 \frac{1}{\sqrt{n}} \leq \sqrt{\|\theta^*\|_2^3 d/n}$*, or equivalently,* $d \geq \|\theta^*\|_2^3 \,\mathrm{polylog}(n)$*, which covers most high-dimensional regimes of interest. Similarly, it improves on our GD rate,* $O(\sqrt{\|\theta^*\|_2^5 d/n})$*, whenever* $d \geq \|\theta^*\|_2 \,\mathrm{polylog}(n)$*.*

Remark 5. *Equation (12) reduces to* $\tilde{O}(\sqrt{\|\theta^*\|_2 d/n})$ *whenever* $\|\theta^*\|_2^3 \sqrt{\frac{1}{n}} + \frac{\|\theta^*\|_2^2 d}{n} \lesssim \sqrt{\frac{\|\theta^*\|_2 d}{n}}$*, or equivalently,* $n \gtrsim \|\theta^*\|_2^3 d + \|\theta^*\|_2^5$*. In the regime, the rate* $\tilde{O}(\sqrt{\|\theta^*\|_2 d/n})$ *achieved by Algorithm 2 is nearly sharp in light of a matching lower bound* $\Omega(\sqrt{\|\theta^*\|_2 d/n})$*. In fact, suppose that* $\tilde{\theta}$ *is an estimator for* θ^**, then* $\tilde{\theta}/\|\tilde{\theta}\|_2$ *is an estimator for* $\theta^*/\|\theta^*\|_2$ *and therefore performs no better than an existing lower bound* $\|\frac{\tilde{\theta}}{\|\tilde{\theta}\|_2} - \frac{\theta^*}{\|\theta^*\|_2}\|_2 \gtrsim \sqrt{\frac{d}{\|\theta^*\|_2 n}}$ *[13, Theorem 1]; combining with a known bound* $\|\frac{\tilde{\theta}}{\|\tilde{\theta}\|_2} - \frac{\theta^*}{\|\theta^*\|_2}\|_2 \leq \frac{2\|\tilde{\theta} - \theta^*\|_2}{\|\theta^*\|_2}$ *(e.g., [23, Fact 13]), we obtain* $\|\tilde{\theta} - \theta^*\|_2 \gtrsim \sqrt{\frac{\|\theta^*\|_2 d}{n}}$*. Therefore, we conclude that the tight error rate in the regime of* $n \gtrsim \|\theta^*\|_2^3 d + \|\theta^*\|_2^5$ *scales as* $\sqrt{\|\theta^*\|_2 d/n}$*. Nonetheless, the fundamental question of the optimal rate under* $n \lesssim \|\theta^*\|_2^3 d + \|\theta^*\|_2^5$ *remains open.*

Remark 6. *We mention that Algorithm 2 generalizes to high-dimensional sparse case where the true parameter* θ^* *is* k*-sparse. To that end, we only need to incorporate a hard thresholding operator into (11); see [23, Algorithm 1]. Under* $n = \tilde{\Omega}(\|\theta^*\|_2 k + \|\theta^*\|_2^4)$*, this sparse counterpart of Algorithm 2 achieves* $\tilde{O}(\frac{\|\theta^*\|_2}{\sqrt{n}} + \sqrt{\frac{\|\theta^*\|_2 k}{n}} + \frac{\|\theta^*\|_2^2 k}{n})$*. In contrast, it appears to be much harder to extend Theorems 1, 2 to the sparse case.*

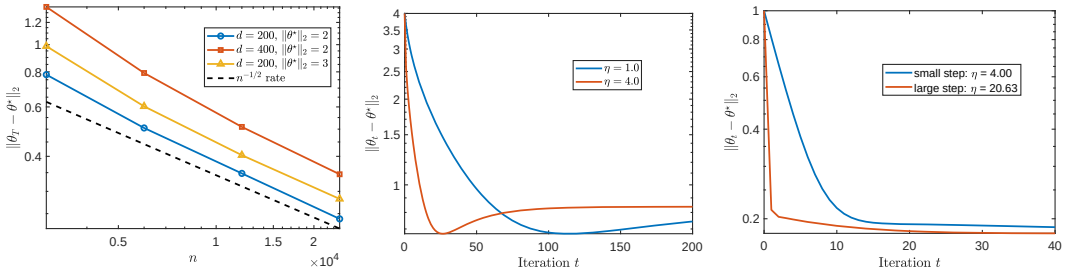
4 Simulations

We provide numerical examples to corroborate our theoretical results. We follow Assumptions 1, 2 to generate $\{(x_i, y_i)\}_{i=1}^n$, and set $\theta^* = \|\theta^*\|_2 r^*$ where r^* is uniformly distributed over $\mathbb{S}^{n-1}$. Each result is averaged over 50 independent trials. All experiments were implemented using Matlab R2022a on a laptop with an Intel CPU up to 2.5 GHz and 32 GB RAM.

Estimation performance. We first validate the theoretical estimation error rate $O(\sqrt{\|\theta^*\|_2^5 d/n})$ of GD. We run Algorithm 1 with $\theta_0 = 0$, $\eta = 4$, and use θ_{100} as estimator. We test $n \in \{3000, 6000, 12000, 24000\}$ under three settings of $(d, \|\theta^*\|_2) = (200, 2), (400, 2), (200, 3)$. The log-log plots in Figure 1a validate the predicted $O(n^{-1/2})$ decay rate and confirm that the estimation error increases with both d and $\|\theta^*\|_2$.

Convergence under $\eta \in (0, 8)$. To illustrate the linear convergence predicted by Theorem 1 for constant stepsizes, we run Algorithm 1 from $\theta_0 = 0$ with $\eta = 1$ and $\eta = 4$ under the setting $(n, d, \|\theta^*\|_2) = (5000, 200, 4)$. Figure 1b reports the estimation error $\|\theta_t - \theta^*\|_2$ over the first 200 iterations on a logarithmic scale. The approximately straight-line decay in the early iterations is consistent with the linear convergence behavior established in Theorem 1.

Large stepsize. To illustrate the local acceleration predicted by Theorem 2, we run Algorithm 1 in a high-signal setting $(n, d, \|\theta^*\|_2) = (80000, 100, 8)$. We initialize the iterates locally around θ^* by setting $\theta_0 = \theta^* + u$, where $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. We compare the constant stepsize $\eta = 4$ with the large stepsize $\eta = 1/m(8) \approx 20.63$. Figure 1c reports the estimation error over the first 40 iterations. The large-stepsize trajectory displays a much steeper initial decay and reaches the statistical error plateau within only a few iterations, consistent with the accelerated local contraction in Theorem 2.



(a) Estimation performance (b) Linear conver. $\eta = 1,\ 4$ (c) $\eta = 4$ v.s. $\eta = 20.63$

Figure 1: Estimation and convergence of GD under Gaussian designs.

Performance of Algorithm 2. While the aim of Algorithm 2 in this paper is primarily to understand the minimax optimal rate, we also evaluate its numerical performance. We compare Algorithm 2 (without sample splitting) and GD with $\eta = 4$ and sufficiently many iterations, which converges to the MLE if it exists. Our numerical results suggest that Algorithm 2 outperforms GD in high dimensions under a moderately large n. Due to page limit we defer the discussion to Appendix E.3.

5 Conclusion

We provide a finite-sample analysis to gradient descent (GD) for logistic regression with Gaussian designs. We show that GD with $O(1)$ stepsize and 0 initialization linearly converges to the true parameter θ^* and achieves error rate $O(\sqrt{\|\theta^*\|_2^5 d/n})$. We also show that GD with $\Theta(\|\theta^*\|_2)$ stepsize and a good enough initialization achieves a faster linear convergence and the same estimation performance. The two results are among the first estimation performance bounds to GD in logistic regression and exhibit a convergence rate faster than most existing results. We also devise a novel estimator whose estimation error rate improves on MLE, GD in high dimensions and is nearly sharp when $n \gtrsim \|\theta^*\|_2^3 d + \|\theta^*\|_2^5$. There remain many interesting problems worth further investigations, including extension of our GD results to sparse logistic regression, whether similar results hold under correlated Gaussian designs or even non-Gaussian designs, the minimax optimal rate under $n \lesssim \|\theta^*\|_2^3 d + \|\theta^*\|_2^5$ (Remark 5), a finite-sample analysis of GD with adaptive stepsizes, estimator with error rate sharper than Algorithm 2, among many others.

Acknowledgment

The work was supported by NSF awards 2217058 and 2112665.

References

[1] Pedro Abdalla, Radu Balan, and Junren Chen. Uniform recovery of structured signals from nonlinear observations: Improved error rates. *arXiv preprint arXiv:2604.20075*, 2026.

[2] Felix Abramovich and Vadim Grinshtein. Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory*, 62(6):3721–3730, 2016.

[3] Kyriakos Axiotis and Maxim Sviridenko. Gradient descent converges linearly for logistic regression on separable data. In *International Conference on Machine Learning*, pages 1302–1319. PMLR, 2023.

[4] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013.

[6] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[7] Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv:2411.02137*, 2024.

[8] Junren Chen, Lijun Ding, Dong Xia, and Ming Yuan. Low-rank tensor estimation from nonlinear observations: A unified framework. *arXiv preprint arXiv:2510.16965*, 2025.

[9] Junren Chen and Ming Yuan. One-bit phase retrieval: Optimal rates and efficient algorithms. *IEEE Transactions on Information Theory*, 2026.

[10] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[11] Michael P Friedlander, Halyun Jeong, Yaniv Plan, and Özgür Yılmaz. Nbiht: An efficient algorithm for 1-bit compressed sensing with optimal error decay rate. *IEEE Transactions on Information Theory*, 68(2):1157–1177, 2021.

[12] Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2013.

[13] Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437. PMLR, 2024.

[14] Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

[15] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 2019.

[16] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

[17] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, volume 1, page 9, 2009.

[18] Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *Sankhya A*, pages 1–70, 2024.

[19] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer.

[20] Philip M Long. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[21] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

[22] Namiko Matsumoto and Arya Mazumdar. Binary iterative hard thresholding converges with optimal number of measurements for 1-bit compressed sensing. *Journal of the ACM*, 71(5):1–64, 2024.

[23] Namiko Matsumoto and Arya Mazumdar. Learning sparse generalized linear models with binary outcomes via iterative hard thresholding. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3933–4032. PMLR, 30 Jun–04 Jul 2025.

[24] Peter McCullagh and John A. Nelder. *Generalized linear models*. Routledge, 2019.

[25] Si Yi Meng, Baptiste Goujaud, Antonio Orvieto, and Christopher De Sa. Gradient descent on logistic regression: Do large step-sizes work with data on the sphere? *arXiv preprint arXiv:2507.11228*, 2025.

[26] Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024.

[27] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[28] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[29] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems*, 26, 2013.

[30] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[31] Rocco A Servedio. On pac learning using winnow, perceptron, and a perceptron-like algorithm. In *Proceedings of the Twelfth Annual Conference on Computational learning theory*, pages 296–307, 1999.

[32] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[33] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press, 1972.

[34] Charles Stein, Persi Diaconis, Susan Holmes, and Gesine Reinert. Use of exchangeable pairs in the analysis of simulations. *Lecture Notes-Monograph Series*, pages 1–26, 2004.

[35] Alexander Tyurin. From logistic regression to the perceptron algorithm: Exploring gradient descent with large step sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20938–20946, 2025.

[36] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[37] Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024.

[38] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36:74229–74256, 2023.

[39] Jingfeng Wu, Pierre Marion, and Peter Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. *arXiv preprint arXiv:2506.02336*, 2025.

[40] Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. Gradient descent converges arbitrarily fast for logistic regression via large and adaptive stepsizes. In *Forty-second International Conference on Machine Learning*, 2025.

[41] Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.

A Proof of Theorem 1 (GD with Small Stepsize)

The main ingredient is an approximate invertibility condition (AIC) that controls

$$\|u-\theta^*-\eta\cdot h_{\theta^*}(u)\|_2, \quad \forall u\in\mathbb{B}_2^d(2\|\theta^*\|_2).$$

We start with a decomposition

$$\|u-\theta^*-\eta\cdot h_{\theta^*}(u)\|_2 \le \eta\cdot \underbrace{\|h_{\theta^*}(u)-\mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_c} + \underbrace{\|u-\theta^*-\eta\mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_e}, \tag{13}$$

where T_c is a concentration term and T_e is a bias term.

A.1 Bounding T_c (concentration term)

By substituting $h_{\theta^*}(u)=\frac{1}{n}\sum_{i=1}^n(s(x_i^\top u)-y_i)x_i$ and a further decomposition,

$$\begin{aligned}
T_c :&= \left\|\frac{1}{n}\sum_{i=1}^n\left(\big(s(x_i^\top u)-y_i\big)x_i-\mathbb{E}\Big[\big(s(x_i^\top u)-y_i\big)x_i\Big]\right)\right\|_2\\
&\le \left\|\frac{1}{n}\sum_{i=1}^n\left(\big(s(x_i^\top \theta^*)-y_i\big)x_i-\mathbb{E}\Big[\big(s(x_i^\top \theta^*)-y_i\big)x_i\Big]\right)\right\|_2\\
&+\left\|\frac{1}{n}\sum_{i=1}^n\left(\big(s(x_i^\top u)-s(x_i^\top \theta^*)\big)x_i-\mathbb{E}\Big[\big(s(x_i^\top u)-s(x_i^\top \theta^*)\big)x_i\Big]\right)\right\|_2\\
&:= T_{c1}+T_{c2}.
\end{aligned}\tag{14}$$

Conditioning on x_i, Assumption 1 ensures $\mathbb{E}[s(x_i^\top\theta^*)-y_i]=0$, hence T_{c1} can be written as

$$T_{c1}=\left\|\frac{1}{n}\sum_{i=1}^n\big(s(x_i^\top\theta^*)-y_i\big)x_i\right\|_2$$

A.1.1 Bounding T_{c1}

We shall establish that

$$T_{c1}\lesssim\sqrt{\frac{d}{n\|\theta^*\|_2}}+\frac{d}{n} \tag{15}$$

holds with probability at least $1-e^{-d}$. This is a consequence of the following result, where we bound the multiplier process $T_{c1}=\sup_{v\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^n(s(x_i^\top\theta^*)-y_i)x_i^\top v$ via Bernstein's inequality (cf. Lemma 11) along with a covering argument.

Lemma 2 (Bounding T_{c1} via covering)**.** *There exist universal constants $C,c>0$ such that, for every $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n\big(s(x_i^\top\theta^*)-y_i\big)x_i\right\|_2\le C\left[\sqrt{\frac{d+\log(1/\delta)}{n\|\theta^*\|_2}}+\frac{d+\log(1/\delta)}{n}\right].$$

Proof. For convenience we let $R:=\|\theta^*\|_2$ and $\xi_i=s(x_i^\top\theta^*)-y_i$, then we seek to bound $\|\frac{1}{n}\sum_{i=1}^n\xi_i x_i\|_2$. We now construct a $\frac{1}{2}$-net of $\mathbb{S}^{d-1}$, denoted by $N_{1/2}$, such that its cardinality is bounded by $|N_{1/2}|\le 5^d$ (e.g., [36, Corollary 4.2.13]). By a standard discretization argument (see, e.g., [36, Exercise 4.4.2]),

$$\left\|\frac{1}{n}\sum_{i=1}^n\xi_i x_i\right\|_2\le 2\sup_{v\in N_{1/2}}\frac{1}{n}\sum_{i=1}^n\xi_i x_i^\top v. \tag{16}$$

Fix any unit vector $v\in\mathbb{S}^{d-1}$ we introduce a shorthand $Y_i(v):=\xi_i\,x_i^\top v$. In light of (16), all that remains is to (i) bound $\frac{1}{n}\sum_{i=1}^n Y_i(v)$ for a fixed $v\in\mathbb{S}^{d-1}$ and then (ii) take a union bound over $N_{1/2}$.

Bounding $\frac{1}{n}\sum_{i=1}^n Y_i(v)$. Fix $v \in \mathbb{S}^{d-1}$. Recall that $\mathbb{E}Y_i(v) = 0$. Moreover, conditional on x_i,

$$\begin{aligned}\mathbb{E}[\xi_i^2 \mid x_i] &= s(x_i^T\theta^*)^2 + s(x_i^T\theta^*) - 2s(x_i^T\theta^*)^2 \\ &= s(x_i^\top\theta^*)\big(1 - s(x_i^\top\theta^*)\big) = s'(x_i^\top\theta^*),\end{aligned} \tag{17}$$

where the last equality is due to Lemma 14. Hence

$$\mathbb{E}[Y_i(v)^2] = \mathbb{E}\big[s'(x_i^\top\theta^*)(x_i^\top v)^2\big].$$

We now bound this variance. Let $e = \theta^*/\|\theta^*\|_2$ and $g_1 := x_i^\top e \sim N(0,1)$, which along with $R = \|\theta^*\|_2$ deliver $x_i^\top\theta^* = Rg_1$. By rotational invariance we can decompose v along $e = \theta^*/\|\theta^*\|_2$ and the orthogonal directions, that is

$$v = (v^\top e)e + [v - (v^\top e)e], \quad \text{where } \|v - (v^\top e)e\|_2 = \sqrt{1 - (v^\top e)^2},$$

and then establish

$$x_i^\top v = v^\top e \cdot g_1 + \sqrt{1 - (v^\top e)^2}\cdot g_2, \quad \text{where } g_2 \sim N(0,1) \text{ is independent of } g_1 \tag{18}$$

Then

$$\begin{aligned}&\mathbb{E}[Y_i(v)^2] = \mathbb{E}\big[s'(x_i^\top\theta^*)(x_i^\top v)^2\big] \\ &= \mathbb{E}\Big[s'(Rg_1)\big\{v^\top e\cdot g_1 + \sqrt{1-(v^\top e)^2}\cdot g_2\big\}^2\Big] \\ &\quad \blacktriangleright \text{ by (18)} \\ &= (v^\top e)^2\mathbb{E}[s'(Rg_1)g_1^2] + \big(1-(v^\top e)^2\big)\mathbb{E}[s'(Rg_1)] \\ &\le \mathbb{E}[s'(Rg_1)] \\ &\quad \blacktriangleright \text{ by Lemma 16} \\ &\lesssim \frac{1}{R}\end{aligned} \tag{19}$$

$\blacktriangleright$ by Lemma 15 and $R = \|\theta^*\|_2 \ge 1$ (cf. Assumption 2)

We proceed to estimate higher-order moments, $\mathbb{E}|Y_i(v)|^q$ for integer $q \ge 3$. Similarly to (19), for any integer $q \ge 3$,

$$\begin{aligned}&\mathbb{E}|Y_i(v)|^q = \mathbb{E}|\xi_i x_i^\top v|^q \\ &\le \mathbb{E}|\xi_i|^2|x_i^\top v|^q \\ &\quad \blacktriangleright \text{ by } |\xi_i| = |s(x_i^\top\theta^*) - y_i| \le 1 \\ &= \mathbb{E}\big(s'(x_i^\top\theta^*)|x_i^\top v|^q\big) \\ &\quad \blacktriangleright \text{ by Equation (17)} \\ &= \mathbb{E}\Big(s'(Rg_1)\Big|(v^\top e)g_1 + \sqrt{1-(v^\top e)^2}\cdot g_2\Big|^q\Big) \\ &\quad \blacktriangleright \text{ by (18) and } g_1 = x_i^\top e \text{ with } e = \theta^*/\|\theta^*\|_2 \\ &\le 2^q\bigg(\mathbb{E}\Big[s'(Rg_1)|g_1|^q\Big] + \mathbb{E}\Big[s'(Rg_1)|g_2|^q\Big]\bigg) \\ &\quad \blacktriangleright \text{ by } (T_1+T_2)^q \le 2^q(|T_1|^q + |T_2|^q),\ |v^\top e| \le 1 \\ &\le 2^q\bigg(\frac{2q!}{\sqrt{2\pi}}\frac{1}{R^{q+1}} + q!\cdot\mathbb{E}[s'(Rg_1)]\bigg) \\ &\quad \blacktriangleright \text{ by Lemma 17, } g_1 \perp g_2, \text{ and } \mathbb{E}|g_2|^q \le q! \\ &\lesssim \frac{2^q q!}{R}\end{aligned} \tag{20}$$

$\blacktriangleright$ by Lemma 15 and $\|\theta^*\|_2 \ge 1$

Combining (19) and (20), we have validated the conditions in Bernstein's inequality (cf. Lemma 11) with $X_i = \frac{Y_i(v)}{n}$, $v \lesssim \frac{1}{nR}$ and $c \lesssim \frac{1}{n}$, and as a result we obtain

$$\mathbb{P}\bigg(\bigg|\frac{1}{n}\sum_{i=1}^n Y_i(v)\bigg| \lesssim \sqrt{\frac{t}{nR}} + \frac{t}{n}\bigg) \ge 1 - 2\exp(-t), \quad \forall t > 0.$$

Union bound. We now take a union bound over $v \in N_{1/2}$, yielding

$$\mathbb{P}\Big(\sup_{v \in N_{1/2}} \Big| \frac{1}{n} \sum_{i=1}^{n} Y_i(v) \Big| \lesssim \sqrt{\frac{t}{nR}} + \frac{t}{n} \Big) \geq 1 - 2\exp(d\log(5) - t), \quad \forall t > 0.$$

Setting $t = d\log(5) + 2\log(1/\delta)$ proves the claim. □

A.1.2 Bounding T_{c2}

We seek to bound

$$T_{c2} := \Big\| \frac{1}{n} \sum_{i=1}^{n} \Big(\big(s(x_i^\top u) - s(x_i^\top \theta^*)\big) x_i - \mathbb{E}\Big[\big(s(x_i^\top u) - s(x_i^\top \theta^*)\big) x_i \Big] \Big) \Big\|_2$$

uniformly for all $u \in \mathbb{B}_2^d(2\|\theta^*\|_2)$. Due to the multiplier $s(x_i^\top u) - s(x_i^\top \theta^*)$, one expects a tighter bound for u being close to θ^*. We achieve such "localization" via a peeling argument, along with a covering argument over each shell. This establishes the following Lemma 3.

We shall pause to announce the final bound on T_{c2}. Setting

$$\delta = \exp(-d), \qquad \rho = \frac{d}{n}$$

and applying a relaxation $\log\log_2(\frac{6\|\theta^*\|_2}{\rho}) \leq \log \frac{Cn\|\theta^*\|_2}{d}$, we reach the following: if $n \gtrsim d\log n + \log\|\theta^*\|_2$, then with probability at least $1 - e^{-d}$,

$$T_{c2} \lesssim \sqrt{\frac{d\log n + \log\|\theta^*\|_2}{n}} \|u - \theta^*\|_2 + \frac{d}{n}, \quad \forall u \in \mathbb{B}_2^d(2\|\theta^*\|_2). \tag{21}$$

Lemma 3 (Bounding T_{c2} via peeling)**.** *For $u \in \mathbb{R}^d$, define*

$$Z_n(u) := \frac{1}{n} \sum_{i=1}^{n} \big(s(x_i^\top u) - s(x_i^\top \theta^*)\big) x_i - \mathbb{E}\big[\big(s(x^\top u) - s(x^\top \theta^*)\big) x\big].$$

Fix $0 < \rho \leq 3\|\theta^*\|_2$. Then there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$,

$$\|Z_n(u)\|_2 \leq C \left[\sqrt{\frac{d\log(Cn) + \log\Big(\frac{\log_2(6\|\theta^*\|_2/\rho)}{\delta}\Big)}{n}} + \frac{d\log(Cn) + \log\Big(\frac{\log_2(6\|\theta^*\|_2/\rho)}{\delta}\Big)}{n} \right] \max\{\|u-\theta^*\|_2, \rho\}$$

holds uniformly over all u satisfying $\|u\|_2 \leq 2\|\theta^*\|_2$.

Proof. Let $R_0 = 3\|\theta^*\|_2$ and write $h := u - \theta^*$. In light of

$$\|u - \theta^*\|_2 \leq \|u\|_2 + \|\theta^*\|_2 \leq 3\|\theta^*\|_2 = R_0,$$

we shall restrict our attention to

$$\|h\|_2 \leq R_0.$$

For $w \in \mathbb{S}^{d-1}$, define the scalar process $Z_n(h, w) := \langle w, Z_n(\theta^* + h)\rangle$, i.e.,

$$Z_n(h,w) = \frac{1}{n} \sum_{i=1}^{n} \big[\big(s(x_i^\top(\theta^* + h)) - s(x_i^\top \theta^*)\big)(x_i^\top w) - \mathbb{E}\big(s(x^\top(\theta^* + h)) - s(x^\top \theta^*)\big)(x^\top w)\big],$$

and note that

$$\|Z_n(u)\|_2 = \sup_{w \in \mathbb{S}^{d-1}} Z_n(u - \theta^*, w).$$

Non-uniform bound. We first establish a fixed-(h, w) concentration bound for $Z_n(h, w)$ for $w \in \mathbb{S}^{d-1}$. Since $\sup_{a\in\mathbb{R}} |s'(a)| \le 1/4$, we have $\left|s(x_i^\top(\theta^* + h)) - s(x_i^\top\theta^*)\right| \le \frac{1}{4}|x_i^\top h|$, and therefore

$$\left|\big(s(x_i^\top(\theta^* + h)) - s(x_i^\top\theta^*)\big)(x_i^\top w)\right| \le \frac{1}{4}|x_i^\top h|\,|x_i^\top w|.$$

Suppose that $\|h\|_2 \le 2\tilde{r}$ and $\|a\|_2 = 1$, then $x_i^\top h$ is sub-Gaussian with ψ_2 norm $O(\tilde{r})$. Also, $x_i^\top a$ is sub-Gaussian with ψ_2 norm $O(1)$. Hence, the product is sub-exponential with ψ_1 norm $O(\tilde{r})$:

$$\|\big(s(x_i^\top(\theta^* + h)) - s(x_i^\top\theta^*)\big)(x_i^\top a)\|_{\psi_1} \lesssim \|(x_i^T h)(x_i^T w)\|_{\psi_1} \le \|x_i^T h\|_{\psi_2}\|x_i^T w\|_{\psi_2} \lesssim \tilde{r}.$$

See [36, Section 2] for details. By Bernstein's inequality for sub-exponential variables (cf. Lemma 12), for every fixed pair (h, w) with $\|h\|_2 \le 2\tilde{r}$ and $w \in \mathbb{S}^{d-1}$,

$$\mathbb{P}\left(|Z_n(h, w)| \gtrsim t\right) \le 2\exp\left(-cn\min\left\{(\frac{t}{\tilde{r}})^2, \frac{t}{\tilde{r}}\right\}\right), \quad \forall t > 0. \tag{22}$$

$$\Longrightarrow \mathbb{P}\left(|Z_n(h, w)| \gtrsim \tilde{r}\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right)\right) \le 2e^{-t}, \quad \forall t > 0. \tag{23}$$

Peeling argument: $\|h\|_2 \ge \rho$. We now apply a peeling argument. For the fixed $0 < \rho \le R_0 := 3\|\theta^*\|_2$ and

$$j = 0, 1, \ldots, J, \qquad J := \lceil \log_2(R_0/\rho)\rceil,$$

define

$$r_j := 2^j\rho, \qquad \mathcal{S}_j := \{h : \ r_j \le \|h\|_2 \le 2r_j\}.$$

Let $\mathcal{N}_j$ be an (r_j/n)-net of $\mathcal{S}_j = \{h : \ r_j \le \|h\|_2 \le 2r_j\}$, and we may choose $|\mathcal{N}_j| \le (Cn)^d$ [36, Corollary 4.2.13]. Also let $\mathcal{M}$ be a $1/2$-net of $\mathbb{S}^{d-1}$ with $|\mathcal{M}| \le 5^d$ [36, Corollary 4.2.13]. Applying the fixed-(h, w) Bernstein bound in (23) to all pairs

$$(h_0, w_0) \in \mathcal{N}_j \times \mathcal{M}$$

and taking a union bound over $j = 0, \ldots, J$, along with a suitable choice of t to offset the effect of the union bound, we obtain that

$$|Z_n(h_0, w_0)| \le Cr_j \underbrace{\left(\sqrt{\frac{d\log(Cn) + \log\left(\frac{J+1}{\delta}\right)}{n}} + \frac{d\log(Cn) + \log\left(\frac{J+1}{\delta}\right)}{n}\right)}_{:=A_n}, \tag{24}$$

$$\forall h_0 \in \mathcal{N}_j, \ \forall w_0 \in \mathcal{M}, \forall j = 0, \cdots, J \tag{25}$$

holds with probability at least $1 - \delta/2$.

It remains to pass from the nets to all points in each shell. For any $h, h' \in \mathbb{R}^d$ and any $w \in \mathbb{S}^{d-1}$, using again $\sup_{a\in\mathbb{R}} |s'(a)| \le 1/4$, we obtain

$$|Z_n(h, w) - Z_n(h', w)| \le \frac{1}{4}\left[\frac{1}{n}\sum_{i=1}^n |x_i^\top(h - h')|\,|x_i^\top w| + \mathbb{E}|x^\top(h - h')|\,|x^\top w|\right].$$

By Cauchy–Schwarz,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |x_i^\top(h - h')|\,|x_i^\top w| &\le \left(\frac{1}{n}\sum_{i=1}^n (x_i^\top(h - h'))^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n (x_i^\top w)^2\right)^{1/2} \\ &= \left((h - h')^\top\widehat{\Sigma}(h - h')\right)^{1/2}\left(w^\top\widehat{\Sigma}w\right)^{1/2},\end{aligned}$$

where $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top = \frac{1}{n}XX^\top$ $(X = [x_1, ..., x_n])$ is the sample covariance. Therefore,

$$\frac{1}{n}\sum_{i=1}^n |x_i^\top(h - h')|\,|x_i^\top w| \le \|\widehat{\Sigma}\|_{\mathrm{op}}\|h - h'\|_2\|w\|_2 = \|\widehat{\Sigma}\|_{\mathrm{op}}\|h - h'\|_2 = \frac{\|X\|_{\mathrm{op}}^2}{n}\|h - h'\|_2.$$

Similarly, since $x \sim N(0, I_d)$,

$$\mathbb{E}|x^\top(h-h')|\,|x^\top w| \le \big(\mathbb{E}(x^\top(h-h'))^2\big)^{1/2}\big(\mathbb{E}(x^\top w)^2\big)^{1/2} = \|h-h'\|_2\|w\|_2 = \|h-h'\|_2.$$

Also,

$$\mathbb{E}|x^\top(h-h')|\,|x^\top w| \le \|h-h'\|_2.$$

Thus

$$|Z_n(h,w) - Z_n(h',w)| \le C\Big(1+\frac{\|X\|_{\mathrm{op}}^2}{n}\Big)\|h-h'\|_2.$$

By Lemma 13, with probability at least $1-\delta/2$,

$$\|X\|_{\mathrm{op}} \lesssim C\sqrt{n+d+\log(1/\delta)}.$$

On the intersection of the these high-probability events, fix any $h \in \mathcal{S}_j$ and choose $h_0 \in \mathcal{N}_j$ with $\|h-h_0\|_2 \le \frac{r_j}{n}$. Then uniformly over $w \in \mathbb{S}^{d-1}$,

$$|Z_n(h,w) - Z_n(h_0,w)| \lesssim \left(1+\frac{n+d+\log(1/\delta)}{n}\right)\frac{r_j}{n} \lesssim r_j A_n,$$

This extends the bound in (24) to $\forall h_0 \in \mathcal{S}_j$, $a_0 \in \mathcal{M}$, $j \in 0, \cdots, J$, yielding

$$|Z_n(h,w_0)| \le C r_j A_j, \quad \forall h \in \mathcal{S}_j,\ \forall w_0 \in \mathcal{M},\ \forall j = 0, \cdots, J, \tag{26}$$

so long as the constant C is increased accordingly.

It remains to show that $w_0 \in \mathcal{M}$ is sufficient for bounding

$$\|Z_n(\theta^*+h)\|_2 = \sup_{w\in\mathbb{S}^{d-1}} Z_n(h,w).$$

Indeed, a standard discretization argument (see, e.g., [36, Exercise 4.4.2]) yields

$$\|Z_n(\theta^*+h)\|_2 \le 2\sup_{w_0\in\mathcal{M}}|\langle w_0, Z_n(\theta^*+h)\rangle| = 2\sup_{w_0\in\mathcal{M}}|Z_n(h,w_0)|.$$

Therefore, by the bound in (26), for every $h \in \mathcal{S}_j$,

$$\|Z_n(\theta^*+h)\|_2 \le C r_j A_n.$$

Since $\|h\|_2 \ge r_j$ on $\mathcal{S}_j$ by definition, we get

$$\|Z_n(\theta^*+h)\|_2 \le C A_n \|h\|_2, \quad \forall \rho \le \|h\|_2 \le R_0$$

Handling $\|h\|_2 \le \rho$. It remains to handle the small ball $\|h\|_2 \le \rho$. The same argument as above, applied with $\tilde{r} = \rho$ in Equation (23) and the (ρ/n)-net of the ball $\{h : \|h\|_2 \le \rho\}$, gives

$$\|Z_n(\theta^*+h)\|_2 \lesssim A_n\rho, \qquad \forall h \ \text{ obeying } \ \|h\|_2 \le \rho.$$

We omit further details for this step.

Combining the shell bound and the small-ball bound, we obtain uniformly over all $\|u\|_2 \le 2\|\theta^*\|_2$,

$$\|Z_n(u)\|_2 \le C A_n \max\{\|u-\theta^*\|_2, \rho\}.$$

Since $J+1 \le C\log_2(2R_0/\rho)$, this proves the claim. □

A.1.3 Final bound on T_c

We now substitute the bounds in Equations (15) and (21) into (14) to obtain our final bound on the concentration term T_c: if $n \gtrsim d\log n + \log\|\theta^*\|_2$, then with probability at least $1-2e^{-d}$, we have

$$\|h_{\theta^*}(u) - \mathbb{E}[h_{\theta^*}(u)]\|_2 \lesssim \sqrt{\frac{d\log n + \log\|\theta^*\|_2}{n}}\|u-\theta^*\|_2 + \sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}, \ \ \forall u \in \mathbb{B}_2^d(2\|\theta^*\|_2). \tag{27}$$

A.2 Bounding T_e (bias term)

We now control the bias term

$$T_e := \left\|u - \theta^* - \eta \cdot \mathbb{E}[h_{\theta^*}(u)]\right\|_2,$$

where by $h_{\theta^*}(u) = \frac{1}{n}\sum_{i=1}^n (s(x_i^\top u) - y_i)x_i$ and $\mathbb{E}[y_i|x_i] = s(x_i^\top \theta^*)$

$$\mathbb{E}[h_{\theta^*}(u)] = \mathbb{E}\big[\big(s(x_i^\top u) - s(x_i^\top \theta^*)\big)x_i\big].$$

Now an important step is to apply stein's identity,

$$\begin{aligned}
\mathbb{E}[h_{\theta^*}(u)] &= \mathbb{E}[x_i s(x_i^\top u)] - \mathbb{E}[x_i s(x_i^\top \theta^*)] \\
&= \mathbb{E}[s'(x_i^\top u)]u - \mathbb{E}[s'(x_i^\top \theta^*)]\theta^* \\
&\quad \blacktriangleright \text{ by Equation (91) in Lemma 10} \\
&= F(u) - F(\theta^*). \\
&\quad \blacktriangleright \text{ we introduce } F(u) := \mathbb{E}[s'(x_i^\top u)]u = m(\|u\|_2)u
\end{aligned}$$

Recall that $m(\tau) := \mathbb{E}_{g\sim N(0,1)} s'(\tau g)$. The main work of the analysis on this bias term is to identify η such that

$$T_e := \left\|u - \theta^* - \eta \cdot \mathbb{E}[h_{\theta^*}(u)]\right\|_2 = \left\|u - \theta^* - \eta(F(u) - F(\theta^*))\right\|_2 \tag{28}$$

is a contraction. Under the small stepsize $\eta \in (0, 8)$, we establish the following result by the main idea of approximating $F(u) - F(\theta^*)$ via a linear map on $u - \theta^*$.

We shall pause to see the implication on T_e. By (28), we shall set $v = \theta^*$ in Lemma 4 to establish that, under $\eta \in [0.1, 7.9]$ (as chosen in Theorem 1),

$$T_e \le \left(1 - \frac{c}{\|\theta^*\|_2^3}\right)\|u - \theta^*\|_2, \quad \forall u \in \mathbb{B}_2^d(2\|\theta^*\|_2). \tag{29}$$

holds for some universal constant $c > 0$.

Lemma 4 (T_2 is a contraction under small stepsize)**.** *Let $F(z) = m(\|z\|_2)z$ and $m(\tau) = \mathbb{E}_{g\sim N(0,1)}[s'(\tau g)]$. Fix $0 < \eta < 8$. Let $v \in \mathbb{R}^d$ satisfy $\|v\|_2 = R \ge 1$. For any $u \in \mathbb{R}^d$ satisfying $\|u\|_2 \le 2R$, we have that*

$$\|\eta(F(u) - F(v)) - (u - v)\|_2 \le \left(1 - \frac{c\eta\left(2 - \frac{\eta}{4}\right)}{R^3}\right)\|u - v\|_2.$$

Proof. Let $h := u - v$, then by Lemma 18, we have

$$F(u) - F(v) = \int_0^1 \nabla F(v + th)h\, dt$$

where

$$\nabla F(z) = m(\|z\|_2)I + \frac{m'(\|z\|_2)}{\|z\|_2} zz^\top, \;\; z \neq 0 \quad \text{with the convention } \nabla F(0) = m(0)I_d. \tag{30}$$

Define the matrix

$$A := \int_0^1 \nabla F(v + th)\, dt, \tag{31}$$

then

$$F(u) - F(v) = Ah. \tag{32}$$

Spectrum of A. For every $t \in [0,1]$, since $\|u\|_2 \le 2R$ and $\|v\|_2 = R$, and $v + th = (1-t)v + tu$, we have

$$\|v + th\|_2 \le (1-t)\|v\|_2 + t\|u\|_2 \le 2R.$$

In light of (30), if $z \neq 0$, then the eigenvalues of $\nabla F(z)$ are $m(\|z\|_2)$ in directions orthogonal to z, and

$$m(\|z\|_2) + \|z\|_2 m'(\|z\|_2) = q'(\|z\|_2) \tag{33}$$

in the direction of $\frac{z}{\|z\|_2}$; we mention that this extends to $z = 0$. Note that in (33), we introduce

$$q(\tau) = \tau m(\tau)$$

as in Lemma 15, so that

$$q'(\|z\|_2) = m(\|z\|_2) + \|z\|_2 m'(\|z\|_2) = \mathbb{E}[s'(\|z\|_2 g) g^2] \le \frac{1}{4}\mathbb{E}[g^2] = \frac{1}{4}. \tag{34}$$

By Lemma 15, we have $m(\tau) \asymp \frac{1}{1+\tau},\ q'(\tau) \asymp \frac{1}{(1+\tau)^3}$ for $\tau > 0$. Combining with $\|v + th\|_2 \le 2R$ and $R \ge 1$, we obtain that all the eigenvalues of $\nabla F(v + th)$, for any $t \in [0,1]$, are of order $\Omega(R^{-3})$. This means that

$$\nabla F(v + th) \succeq \frac{c}{R^3} I \quad \text{for some absolute constant } c. \tag{35}$$

and therefore for any $w \in \mathbb{R}^d$,

$$\langle w, Aw \rangle = \int_0^1 w^\top \nabla F(v + th) w \, dt \ge \frac{c\|w\|_2^2}{R^3} \tag{36}$$

▶ by Equations (31) and (35)

On the other hand, since $s'(a) \le 1/4$, in light of (33) and (34) we have

$$\|\nabla F(v + th)\|_{\mathrm{op}} \le m(\|v + th\|_2) \le \frac{1}{4}$$

which then yields

$$\langle w, Aw \rangle = \int_0^1 w^\top \nabla F(v + th) w \, dt \le \frac{1}{4}\|w\|_2^2, \quad \forall w \in \mathbb{R}^d. \tag{37}$$

Combining (36) and (37), we conclude that the positive definite matrix A satisfies

$$\frac{c}{R^3} I_d \preceq A \preceq \frac{1}{4} I_d,$$

and we can taken an eigenvalue decomposition of A as

$$A = O^\top \Lambda O \tag{38}$$

for some orthonormal matrix O and a diagonal matrix

$$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d), \quad \lambda_i \in \Big[\frac{c_1}{R^3}, \frac{1}{4}\Big],\ i \in [d].$$

Now using (32) and $h = u - v$,

$$\eta(F(u) - F(v)) - (u - v) = (\eta A - I)h.$$

Hence

$$\|\eta(F(u) - F(v)) - (u - v)\|_2^2 = \|(\eta A - I)h\|_2^2 = \|h\|_2^2 - 2\eta\langle h, Ah\rangle + \eta^2 \langle h, A^2 h\rangle. \tag{39}$$

To bound $\langle h, A^2 h\rangle$, we shall use the eigenvalue decomposition of A in (38) to obtain

$$h^\top \Big(A^2 - \frac{1}{4}A\Big)h = (Oh)^\top \Big(\Lambda^2 - \frac{\Lambda}{4}\Big)(Oh) \le 0 \implies h^\top A^2 h \le \frac{1}{4} h^\top A h. \tag{40}$$

To lower bound $\langle h, Ah\rangle$, Equation (36) yields

$$\langle h, Ah \rangle \ge \frac{c\|h\|_2^2}{R^3}. \tag{41}$$

Substituting (40) and (41) into (39), we obtain

$$\|\eta(F(u) - F(v)) - (u - v)\|_2^2 \le \|h\|_2^2 - (2\eta - \frac{\eta^2}{4}) h^\top A h \le \left(1 - \frac{c\eta(2 - \eta/4)}{R^3}\right)\|h\|_2^2$$

Taking square roots and using $\sqrt{1-a} \le 1 - \frac{a}{2}$ for $a \in [0,1)$, we arrive at

$$\|\eta(F(u) - F(v)) - (u - v)\|_2 \le \sqrt{1 - \frac{c\eta(2 - \eta/4)}{R^3}}\|u - v\|_2 \le \left(1 - \frac{c\eta(2 - \eta/4)}{2R^3}\right)\|u - v\|_2.$$

The result follows. □

A.3 AIC and Convergence

We are now ready to put pieces together to establish our main structure condition, the approximate invertibility condition (AIC). Under $m \gtrsim d\log n + \log\|\theta^*\|_2$, the bound (27) on T_c holds with probability at least $1-2e^{-d}$, and also under $\eta \in [0.1, 7.9]$ the bound (29) on T_e holds deterministically. We now substitute these two bounds into (13) to obtain

$$\|u-\theta^*-\eta\cdot h_{\theta^*}(u)\|_2 \le \left(1-\frac{c}{\|\theta^*\|_2^3}+C\sqrt{\frac{d\log n+\log\|\theta^*\|_2}{n}}\right)\|u-\theta^*\|_2 + C'\left(\sqrt{\frac{d}{n\|\theta^*\|_2}}+\frac{d}{n}\right), \qquad \forall u\in\mathbb{B}_2^d(2\|\theta^*\|_2). \tag{42}$$

Following recent works [22, 23, 8, 11, 1], (42) is referred to as an *approximate invertibility condition* (AIC) in the sense that it essentially controls the difference between the ideal descent step $u-\theta^*$ and the actual gradient step $\eta\cdot h_{\theta^*}(u)$. To imply the desired convergence, we need the linear term, $(1-\frac{c}{\|\theta^*\|_2^3}+\sqrt{\frac{d\log n+\log\|\theta^*\|_2}{n}})\cdot\|u-\theta^*\|_2$, to be a contraction. To that end we enforce the sample complexity

$$n \gtrsim \|\theta^*\|_2^6 d\log n + \|\theta^*\|_2^6\log\|\theta^*\|_2, \tag{43}$$

under which the AIC (42) simplifies to

$$\|u-\theta^*-\eta\cdot h_{\theta^*}(u)\|_2 \le \left(1-\frac{c'}{\|\theta^*\|_2^3}\right)\|u-\theta^*\|_2 + C'\left(\sqrt{\frac{d}{n\|\theta^*\|_2}}+\frac{d}{n}\right),\ \ \forall u\in\mathbb{B}_2^d(2\|\theta^*\|_2). \tag{AIC1}$$

To complete the proof of Theorem 1, all that remains is to show that the above (AIC1) implies the claimed linear convergence. We address this via the following lemma by an argument similar to [8, Theorem 3.1], which is to control $\{\|\theta_t-\theta^*\|_2\}_{t=0}^\infty$ by a sequence defined iteratively.

Lemma 5 ((AIC1) implies convergence)**.** *Under (43) and (AIC1), $\{\theta_t\}_{t\ge0}$ generated by Algorithm 1 with $\theta_0=0$ (and the η in (AIC1)) satisfies*

$$\|\theta_t-\theta^*\|_2 \le \left(1-\frac{c'}{\|\theta^*\|_2^3}\right)^t\|\theta^*\|_2+\tilde{C}\sqrt{\frac{\|\theta^*\|_2^5 d}{n}}, \qquad \forall t=0,1,\cdots.$$

Proof. We define a sequence by $\{f_t\}_{t=0}^\infty$ by the recurrence and the initial value

$$f_{t+1}=\left(1-\frac{c'}{\|\theta^*\|_2^3}\right)f_t+C'\left(\sqrt{\frac{d}{n\|\theta^*\|_2}}+\frac{d}{n}\right),\ t\ge0, \quad \text{and} \quad f_0=\|\theta^*\|_2.$$

It is not hard to observe the formula

$$f_t=\left(1-\frac{c'}{\|\theta^*\|_2^3}\right)^t\|\theta^*\|_2+\left[1-\left(1-\frac{c'}{\|\theta^*\|_2^3}\right)^t\right]\frac{C'}{c'}\left(\sqrt{\frac{d\|\theta^*\|_2^5}{n}}+\frac{d\|\theta^*\|_2^3}{n}\right), \quad \forall t\ge0. \tag{44}$$

Hence, under (43) and $\|\theta^*\|_2\ge1$, we have that

$$f_t \le \left(1-\frac{c'}{\|\theta^*\|_2^3}\right)^t\|\theta^*\|_2+O\left(\sqrt{\frac{\|\theta^*\|_2^5 d}{n}}\right), \quad \forall t\ge0,$$

and in turn we only need to prove $\|\theta_t-\theta^*\|_2\le f_t$ for any $t\ge0$. Before that, we further observe that (43) and $\|\theta^*\|_2\ge1$ imply

$$\frac{C'}{c'}\left(\sqrt{\frac{d\|\theta^*\|_2^5}{n}}+\frac{d\|\theta^*\|_2^3}{n}\right)\le\|\theta^*\|_2,$$

hence (44) also yields $f_t\le\|\theta^*\|_2$ for any $t\ge0$.

It remains to prove $\|\theta_t-\theta^*\|_2\le f_t$ for any $t\ge0$. We accomplish this by induction. Since $\theta_0=0$ and $f_0=\|\theta^*\|_2$, $\|\theta_0-\theta^*\|_2\le f_0$ holds trivially. Now suppose that $\|\theta_t-\theta^*\|_2\le f_t$, and we seek to prove

$\|\theta_{t+1} - \theta^*\|_2 \le f_{t+1}$. First by $f_t \le \|\theta^*\|_2$, the hypothesis yields $\|\theta_t\|_2 \le f_t + \|\theta^*\|_2 \le 2\|\theta^*\|_2$. Therefore we can apply (AIC1) with $u = \theta_t$, yielding

$$\|\theta_t - \theta^* - \eta \cdot h_{\theta^*}(\theta_t)\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^3}\right)\|\theta_t - \theta^*\|_2 + C'\left(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}\right).$$

Since $\theta_{t+1} = \theta_t - \eta \cdot h_{\theta^*}(\theta_t)$, and also using the hypothesis $\|\theta_t - \theta^*\|_2 \le f_t$, the above display yields

$$\|\theta_{t+1} - \theta^*\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^3}\right) f_t + C'\left(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}\right) = f_{t+1}.$$

This completes the induction. □

B Proof of Theorem 2 (GD with Large Stepsize)

The overall proof architecture remains the same as Theorem 1, that is to establish an AIC and then use it to imply the desired convergence. The first main difference is that the AIC for Theorem 2 only holds "locally" over a small radius-$O(\|\theta^*\|_2^{-1})$ neighborhood of θ^*. Again we start with (13), i.e., the decomposition

$$\|u - \theta^* - \eta \cdot h_{\theta^*}(u)\|_2 \le \eta \cdot \underbrace{\|h_{\theta^*}(u) - \mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_c} + \underbrace{\|u - \theta^* - \eta\mathbb{E}[h_{\theta^*}(u)]\|_2}_{:=T_e}, \tag{45}$$

and we shall restrict to $u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$ for some small enough constant c_0. The major technical difference arises in the analysis of the bias term T_e.

B.1 Bounding T_c (concentration term)

While re-iterating the arguments in Appendix A.1 (especially Appendix A.1.2) over a smaller range of u, namely $\mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$, may lead to slightly tighter bound, such improvement is marginal and can only refine log factors. Therefore, we directly reuse the previous bound, which we restate here (cf. Appendix A.1.3): if $n \gtrsim d\log n + \log\|\theta^*\|_2$, then with probability at least $1 - 2e^{-d}$, we have

$$\|h_{\theta^*}(u) - \mathbb{E}[h_{\theta^*}(u)]\|_2 \lesssim \sqrt{\frac{d\log n + \log\|\theta^*\|_2}{n}}\|u - \theta^*\|_2 + \sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}, \ \ \forall u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2}). \tag{46}$$

B.2 Bounding T_e (bias term)

We shall now bound T_e over $u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$. The main ingredient is the following Lemma 6, which involves two functions that are recurring in our analysis:

$$m(\tau) = \mathbb{E}[s'(\tau g)] \text{ and } q'(\tau) = \mathbb{E}[s'(\tau g)g^2]$$

See,e.g., Theorem 2 and Lemma 15. An important starting point is, again, the transform in (28) based on stein's identity:

$$T_e := \big\|u - \theta^* - \eta \cdot \mathbb{E}[h_{\theta^*}(u)]\big\|_2 = \big\|u - \theta^* - \eta(F(u) - F(\theta^*))\big\|_2, \tag{47}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is defined as $F(u) = m(\|u\|_2)u$.

Let us pause to see the implied bound of Lemma 6 on T_e. We set $v = \theta^*$ in Lemma 6, then (49) is also the stepsize assumption in Theorem 2, and the conclusion

$$\|\eta(F(u) - F(\theta^*)) - (u - \theta^*)\|_2 \le \left(1 - \frac{c_1}{R^2}\right)\|u - \theta^*\|_2, \quad \forall u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2}),$$

in light of (47), yields

$$T_e \le \left(1 - \frac{c_1}{\|\theta^*\|_2^2}\right)\|u - \theta^*\|_2, \quad \forall u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2}). \tag{48}$$

This is final bound on the bias term T_e.

Lemma 6. *There exist universal constants $c, c_0, c_1 > 0$ such that the following holds. Let $v \in \mathbb{R}^d$ satisfy $R := \|v\|_2 \geq 1$. If*

$$\frac{c}{R^2 q'(R)} < \eta < \frac{2 - cR^{-2}}{m(R)}, \tag{49}$$

then for any $\|u - v\|_2 \leq \frac{c_0}{R}$ we have

$$\|\eta(F(u) - F(v)) - (u - v)\|_2 \leq \Big(1 - \frac{c_1}{R^2}\Big) \|u - v\|_2.$$

Moreover, $\eta = \frac{1}{m(R)}$ is a specific choice satisfying (49).

Proof. Let $h := u - v$. If $h = 0$, the result is trivial, so we assume $h \neq 0$. Let $e := \frac{v}{\|v\|_2}$, and we shall decompose h along e and an orthogonal direction $\frac{w}{\|w\|_2}$ (here, $\{e, \frac{w}{\|w\|_2}\}$ is orthonormal basis for $\mathrm{span}(u, v)$, and in the case of $u = \lambda v$ for some $\lambda \in \mathbb{R}$ we can choose $w/\|w\|_2$ as an arbitrary unit vector orthogonal to v):

$$h = \alpha e + w, \qquad \text{for some } w \perp e. \tag{50}$$

With these conventions and $R = \|v\|_2 \geq 1$ we have

$$u = v + h = (R + \alpha)e + w, \tag{51}$$

and hence $\rho := \|u\|_2$ equals

$$\rho := \|u\|_2 = \sqrt{(R + \alpha)^2 + \|w\|_2^2}. \tag{52}$$

Since $\|h\|_2 \leq \frac{c_0}{R}$ and $R \geq 1$, choosing sufficiently small c_0 (together with triangle inequality) ensures

$$\frac{R}{2} \leq R - \|h\|_2 \leq \rho \leq R + \|h\|_2 \leq 2R \tag{53}$$

The above conventions, along with $q(\tau) = \tau m(\tau)$ (cf. Lemma 15), also allow us to write

$$F(v) = m(\|v\|_2)v = m(R)Re = q(R)e,$$

and in light of (51)

$$F(u) = m(\|u\|_2)u = m(\rho)\big((R + \alpha)e + w\big).$$

Hence

$$F(u) - F(v) = \Big(m(\rho)(R + \alpha) - q(R)\Big)e + m(\rho)w. \tag{54}$$

Approximating $F(u) - F(v)$ via linear map. We now come to the most technical step in this proof, which approximates $F(u) - F(v)$ via a linear map on h. We introduce

$$J_v = q'(R)ee^\top + m(R)(I - ee^\top). \tag{55}$$

Then for $h = \alpha e + w$ with $w \perp e$ (cf. Equation (50)), we have

$$ee^\top h = \alpha e, \qquad (I - ee^\top)h = w,$$

and hence

$$J_v h = q'(R)ee^\top h + m(R)(I - ee^\top)h = q'(R)\alpha e + m(R)w. \tag{56}$$

We want to use $J_v h$ to approximate $F(u) - F(v) = F(v + h) - F(v)$, with the intuition behind being that J_v in (55) is just the Jacobian matrix $\nabla F(v)$ (see Lemma 18). Our main claim is that

$$F(u) - F(v) = J_v h + \mathcal{R}, \quad \text{with } \|\mathcal{R}\|_2 \leq C\frac{\|h\|_2^2}{R^2}. \tag{57}$$

Proof of (57). In light of (54) and (56),

$$F(u) - F(v) - J_v h = \underbrace{\Big(m(\rho)(R+\alpha) - q(R) - q'(R)\alpha\Big)}_{:=T_1} e + \underbrace{\big(m(\rho) - m(R)\big)}_{T_2} w \tag{58}$$

and we shall treat the components of e and of w separately.

For the e-component, by $q(\rho) = \rho m(\rho)$ we have

$$T_1 = \underbrace{q(\rho) - q(R) - q'(R)(\rho - R)}_{:=T_1'} + \underbrace{\big(q'(R) - m(\rho)\big)(\rho - R - \alpha)}_{T_1''}.$$

To bound T_1', we have

$$\begin{aligned}
|T_1'| &= \big|q(\rho) - q(R) - q'(R)(\rho - R)\big| \\
&\lesssim \Big(\sup_{\tau \geq \frac{R}{2}} |q''(\tau)|\Big) \cdot |\rho - R|^2 \\
&\qquad \blacktriangleright \text{ by Taylor's theorem and Equation (53)} \\
&\lesssim \frac{\|h\|_2^2}{R^4} \\
&\qquad \blacktriangleright \text{ by } q''(\tau) \asymp \frac{1}{\tau^4} \text{ and } |\rho - R| \leq \|h\|_2
\end{aligned} \tag{59}$$

We now bound $|T_1''| = |q'(R) - m(\rho)||\rho - R - \alpha|$. The bounds in Lemma 15 give

$$|q'(R) - m(\rho)| \leq |q'(R)| + |m(\rho)| \lesssim \frac{1}{R}.$$

Combining with

$$|\rho - R - \alpha| = \frac{\rho^2 - (R+\alpha)^2}{\rho + R + \alpha} \overset{(52)}{=} \frac{\|w\|_2^2}{\rho + R + \alpha} \overset{(50)}{\lesssim} \frac{\|w\|_2^2}{R} \overset{(50)}{\leq} \frac{\|h\|_2^2}{R},$$

we obtain

$$|T_1''| \lesssim \frac{1}{R} \cdot \frac{\|h\|_2^2}{R} = \frac{\|h\|_2^2}{R^2}. \tag{60}$$

Combining (59) and (60) yields

$$|T_1| = O\left(\frac{\|h\|_2^2}{R^2}\right). \tag{61}$$

For T_2, by using $|m'(\tau)| \lesssim \tau^{-2}$ for $\tau \geq \frac{1}{2}$ (cf. Lemma 19), we obtain

$$|T_2| = |m(\rho) - m(R)| \overset{(53)}{\leq} \sup_{\tau \geq R/2} |m'(\tau)| \cdot |\rho - R| \lesssim \frac{|\rho - R|}{R^2} \overset{(53)}{\leq} \frac{\|h\|_2}{R^2}.$$

Together with $\|w\|_2 \leq \|h\|_2$ from (50), we reach

$$\|T_2 w\|_2 \lesssim \frac{\|h\|_2}{R^2} \|w\|_2 \leq \frac{\|h\|_2^2}{R^2}. \tag{62}$$

We are now ready to establish the claimed approximation guarantee in (57):

$$\begin{aligned}
\big\|F(u) - F(v) - J_v h\big\|_2 &\overset{(58)}{\leq} |T_1| + |T_2|\|w\|_2 \\
&\leq \frac{\|h\|_2^2}{R^2}. \\
&\qquad \blacktriangleright \text{ by Equations (61) and (62)}
\end{aligned}$$

Concluding the proof. In view of (55), J_v has eigenvalue $\lambda_\| := q'(R)$ in the direction e and eigenvalue $\lambda_\perp = m(R)$ on the orthogonal subspace. Hence Lemma 16 yields $\lambda_\| < \lambda_\perp$. Note that our choice of stepsize in (49) is equivalent to

$$-\left(1 - \frac{c}{R^2}\right) < 1 - \eta\lambda_\perp < 1 - \eta\lambda_\| < 1 - \frac{c}{R^2},$$

which then also ensures

$$\|\eta J_v - I_d\|_{\mathrm{op}} \le 1 - \frac{c}{R^2}. \tag{63}$$

Using (57), we get

$$\begin{aligned}
&\|\eta(F(u) - F(v)) - (u - v)\|_2 \\
&= \|\eta J_v h + \mathcal{R} - h\|_2 \\
&\le \|(\eta J_v - I_d)h\|_2 + \eta\|\mathcal{R}\|_2 \\
&\le \left(1 - \frac{c}{R^2}\right)\|h\|_2 + \eta C\frac{\|h\|_2^2}{R^2} \\
&\le \left(1 - \frac{c}{R^2}\right)\|h\|_2 + C'\frac{\|h\|_2^2}{R} \\
&\quad \blacktriangleright \text{ By } m(R) \asymp \frac{1}{R} \text{ from Lemma 15, we have } \eta \lesssim R \\
&\le \left(1 - \frac{c}{R^2}\right)\|h\|_2 + \frac{C'c_0\|h\|_2}{R^2} \\
&\quad \blacktriangleright \text{ by } \|h\|_2 \le \frac{c_0}{R} \text{ as assumed} \\
&\le \left(1 - \frac{c}{2R^2}\right)\|h\|_2. \\
&\quad \blacktriangleright \text{ Taking } c_0 \text{ sufficiently small}
\end{aligned}$$

This proves the claim.

Specific stepsize $\eta = \frac{1}{m(R)}$. Finally, we prove the "moreover" part of the statement. We claim that the simpler choice $\eta = \frac{1}{m(R)}$ also belongs to the feasible interval $\left[\frac{c}{R^2 q'(R)}, \frac{2 - cR^{-2}}{m(R)}\right]$ provided $c > 0$ is sufficiently small. Indeed, the upper bound is immediate. Since $R \ge 1$, if $c \le 1$, then $1 \le 2 - cR^{-2}$, and hence $\frac{1}{m(R)} \le \frac{2 - cR^{-2}}{m(R)}$. For the lower bound, we need $\frac{c}{R^2 q'(R)} \le \frac{1}{m(R)}$. Since $m(R), q'(R) > 0$, this is equivalent to $c\, m(R) \le R^2 q'(R)$. Using the estimates from Lemma 15, we have

$$\frac{R^2 q'(R)}{m(R)} \gtrsim \frac{R^2 R^{-3}}{R^{-1}} \gtrsim 1 \implies \inf_{R \ge 1} \frac{R^2 q'(R)}{m(R)} > 0.$$

Choosing $c > 0$ smaller than this infimum and also $c \le 1$ yields $\frac{1}{m(R)} \in \left[\frac{c}{R^2 q'(R)}, \frac{2 - cR^{-2}}{m(R)}\right]$ for all $R \ge 1$. □

B.3 AIC and Convergence

We are in a position to establish the AIC. Substituting (46) and (47) into (45), along with $\eta \le \frac{2}{m(\|\theta^*\|_2)} \asymp \|\theta^*\|_2$ (cf. Lemma 15), it follows that

$$\begin{aligned}
\|u - \theta^* - \eta \cdot h_{\theta^*}(u)\|_2 &\le C'\|\theta^*\|_2\sqrt{\frac{d\log n + \log\|\theta^*\|_2}{n}}\|u - \theta^*\|_2 \\
&+ \left(1 - \frac{c_1}{\|\theta^*\|_2^2}\right)\|u - \theta^*\|_2 + C''\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right), \quad \forall u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2}).
\end{aligned} \tag{64}$$

To ensure that the pre-factor of $\|u - \theta^*\|_2$ is bounded away from 1, i.e.,

$$1 - \frac{c_1}{\|\theta^*\|_2^2} + C'\|\theta^*\|_2\sqrt{\frac{d\log n + \log\|\theta^*\|_2}{n}} < 1 - \frac{c_1}{2\|\theta^*\|_2^2},$$

we enforce the sample complexity

$$n \gtrsim \|\theta^*\|_2^6 d \log n + \|\theta^*\|_2^6 \log \|\theta^*\|_2, \tag{65}$$

which is (surprisingly) identical to (43). Then, the AIC in (64) leads to the following local AIC:

$$\|u - \theta^* - \eta \cdot h_{\theta^*}(u)\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)\|u - \theta^*\|_2 + C'\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right), \ \ \forall u \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2}) \tag{AIC2}$$

for some universal constants c', C', c_0 (after updating the value of C').

To conclude the proof of Theorem 2, it remains to show (AIC2) leads to the claimed convergence. An argument similar to Lemma 5 works, but since (AIC2) only holds locally over $\mathbb{B}_2^d(\theta^*; \frac{c}{\|\theta^*\|_2})$, there are two differences of (i) requiring a warm initialization and (ii) requiring a sample complexity slightly higher than (65).

Lemma 7 ((AIC2) implies convergence)**.** *Under (AIC2) and*

$$n \gtrsim \|\theta^*\|_2^7 d \quad \textit{with large enough hidden constant} \tag{66}$$

$\{\theta_0\}_{t\ge 0}$ *generated by Algorithm 1 with* $\theta_0 \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$ *(and the* η *in (AIC2)) satisfies*

$$\|\theta_t - \theta^*\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)^t \frac{c_0}{\|\theta^*\|_2} + \tilde{C}\sqrt{\frac{\|\theta^*\|_2^5 d}{n}}, \quad \forall t = 0, 1, \cdots.$$

Proof. We introduce a sequence $\{f_t\}_{t\ge 0}$ by the recurrence and the initial value

$$f_{t+1} = \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right) f_t + C'\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right), \quad t \ge 0, \quad \text{and} \quad f_0 = \frac{c_0}{\|\theta^*\|_2}. \tag{67}$$

One can find the closed-form expression

$$f_t = \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)^t \frac{c_0}{\|\theta^*\|_2} + \left[1 - \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)^t\right] \frac{C'}{c'}\left(\sqrt{\frac{\|\theta^*\|_2^5 d}{n}} + \frac{\|\theta^*\|_2^3 d}{n}\right), \quad t \ge 0. \tag{68}$$

Under (66), we have $\frac{\|\theta^*\|_2^3 d}{n} \le \sqrt{\frac{\|\theta^*\|_2^5 d}{n}}$, hence it is sufficient to prove $\|\theta_t - \theta^*\|_2 \le f_t$ for any $t \ge 0$. Also under (66) we have

$$\frac{C'}{c'}\left(\sqrt{\frac{\|\theta^*\|_2^5 d}{n}} + \frac{\|\theta^*\|_2^3 d}{n}\right) \le \frac{c_0}{\|\theta^*\|_2},$$

and as such (68) gives $f_t \le \frac{c_0}{\|\theta^*\|_2}$ for any $t \ge 0$.

All that remains is to prove $\|\theta_t - \theta^*\|_2 \le f_t$, $t \ge 0$, and we shall achieve this by induction. Due to $\theta_0 \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$, this holds trivially for $t = 0$. Suppose that $\|\theta_t - \theta^*\|_2 \le f_t$, we then have $\|\theta_t - \theta^*\|_2 \le \frac{c_0}{\|\theta^*\|_2}$, thus $\theta_t \in \mathbb{B}_2^d(\theta^*; \frac{c_0}{\|\theta^*\|_2})$. Therefore we can set $u = \theta_t$ in (AIC2), yielding

$$\|\theta_t - \theta^* - \eta \cdot h_{\theta^*}(\theta_t)\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right)\|\theta_t - \theta^*\|_2 + C'\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right).$$

By $\theta_{t+1} = \theta_t - \eta \cdot h_{\theta^*}(\theta_t)$ and $\|\theta_t - \theta^*\|_2 \le f_t$, the above display gives

$$\|\theta_{t+1} - \theta^*\|_2 \le \left(1 - \frac{c'}{\|\theta^*\|_2^2}\right) f_t + C'\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right) \overset{(67)}{=} f_{t+1},$$

thus completing the induction. The claim is proved. □

C Proof of Theorem 3 (Performance of Algorithm 2)

In this proof, for simplicity in notation, we denote $\hat{r}_{T_0}$ by $\hat{r}$ and in turn $\hat{\theta} = \hat{\lambda} \cdot \hat{r}$. We start with a decomposition

$$\begin{aligned}\|\hat{\theta} - \theta^*\|_2 &\leq \left\|\hat{\lambda}\hat{r} - \|\theta^*\|_2 \hat{r}\right\|_2 + \left\|\|\theta^*\|_2 \hat{r} - \theta^*\right\|_2 \\ &= \left|\hat{\lambda} - \|\theta^*\|_2\right| + \|\theta^*\|_2 \left\|\hat{r} - \frac{\theta^*}{\|\theta^*\|_2}\right\|_2, \end{aligned} \tag{69}$$

where the first term captures the norm estimation error while the second term accounts for the direction estimation error.

Due to Lemma 1, we immediately obtain the bound on $\|\theta^*\|_2 \big\|\hat{r} - \frac{\theta^*}{\|\theta^*\|_2}\big\|_2$. In fact, by assuming the bound in Lemma 1 holds (this is within the probability promised in Theorem 3), then we have

$$\|\theta^*\|_2 \left\|\hat{r} - \frac{\theta^*}{\|\theta^*\|_2}\right\|_2 \lesssim \mathrm{polylog}(n)\left(\sqrt{\frac{\|\theta^*\|_2 d}{n}} + \frac{\|\theta^*\|_2 d}{n}\right). \tag{70}$$

In the remainder of this appendix, we focus on bounding the norm estimation error, $|\hat{\lambda} - \|\theta^*\|_2|$.

C.1 Bounding $|\hat{\lambda} - \|\theta^*\|_2|$ (norm estimation error)

In order to bound $|\hat{\lambda} - \|\theta^*\|_2| = \left|q^{-1}(\hat{v}^\top \hat{r}) - \|\theta^*\|_2\right|$, we shall start with bounding

$$\left|\hat{v}^\top \hat{r} - q(\|\theta^*\|_2)\right|.$$

Lemma 8. *If $n \gtrsim d$ and $\|\theta^*\|_2 \geq 1$, then it holds with probability at least $1 - n^{-2}$ that*

$$\left|\hat{v}^\top \hat{r} - q(\|\theta^*\|_2)\right| \leq \mathrm{polylog}(n)\left(\frac{1}{\sqrt{n}} + \frac{d}{n\|\theta^*\|_2} + \frac{d^2}{n^2}\right). \tag{71}$$

Proof. We shall start with the decomposition

$$\left|\langle \hat{v}, \hat{r}\rangle - q(\|\theta^*\|_2)\right| \leq \underbrace{\left|\left\langle \hat{v}, \frac{\theta^*}{\|\theta^*\|_2}\right\rangle - q(\|\theta^*\|_2)\right|}_{I_1} + \underbrace{\left|\left\langle \hat{v}, \hat{r} - \frac{\theta^*}{\|\theta^*\|_2}\right\rangle\right|}_{:=I_2}.$$

Bounding I_1. Due to $\hat{v} = \frac{1}{(1-\nu)n}\sum_{i=\nu n+1}^{n} y_i x_i$ and (10) we can write

$$I_1 := \left|\frac{1}{(1-\nu)n}\sum_{i=\nu n+1}^{n} y_i \left\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\right\rangle - \mathbb{E}\Big(y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle\Big)\right|.$$

By $|y_i| \leq 1$,

$$\left\|y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle\right\|_{\psi_2} \lesssim \left\|\Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle\right\|_{\psi_2} \lesssim 1,$$

and by centering [36, Lemma 2.6.8],

$$\left\|y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle - \mathbb{E}\Big(y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle\Big)\right\|_{\psi_2} \lesssim 1, \tag{72}$$

and moreover by property of sum of independent, zero-mean sub-Gaussian variables [36, Proposition 2.6.1],

$$\left\|\frac{1}{(1-\nu)n}\sum_{i=\nu n+1}^{n} y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle - \mathbb{E}\Big(y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2}\Big\rangle\Big)\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}}. \tag{73}$$

In turn, a standard sub-Gaussian tail bounds yields that

$$\mathbb{P}\left(I_1 \lesssim \sqrt{\frac{\log n}{n}}\right) \geq 1 - \frac{1}{n^2}. \tag{74}$$

Bounding I_2. We first substitute $\hat{v}$ to write

$$I_2 = \left| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \Big\langle x_i, \hat{r} - \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \right|$$

We now let $e = \hat{r} - \frac{\theta^*}{\|\theta^*\|_2}$ and decompose it along $\frac{\theta^*}{\|\theta^*\|_2}$ and orthogonal space,

$$e = \Big\langle e, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \frac{\theta^*}{\|\theta^*\|_2} + \underbrace{e - \Big\langle e, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \frac{\theta^*}{\|\theta^*\|_2}}_{e_{\theta^*}^{\perp}}.$$

Therefore,

$$\begin{aligned} I_2 &= \left| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \langle x_i, e \rangle \right| \\ &\le \underbrace{\left| \frac{1}{(1-\nu)n} \sum_{i=\nu n}^{n} y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \right| \left| \Big\langle e, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \right|}_{:=I_{21}} + \underbrace{\left| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \langle x_i, e_{\theta^*}^{\perp} \rangle \right|}_{:=I_{22}}. \end{aligned} \tag{75}$$

To bound I_{11}, observe that

$$\left| \Big\langle e, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \right| = \left| \Big\langle \hat{r}, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle - 1 \right| = \frac{\|\hat{r} - \frac{\theta^*}{\|\theta^*\|_2}\|_2^2}{2}$$

Also, by (74) and Lemma 21,

$$\left| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \Big\langle x_i, \frac{\theta^*}{\|\theta^*\|_2} \Big\rangle \right| \le q(\|\theta^*\|_2) + O\Big(\sqrt{\frac{\log n}{n}}\Big) \lesssim 1.$$

Therefore, by the direction error bound in Lemma 1,

$$I_{21} \lesssim \left\| \hat{r} - \frac{\theta^*}{\|\theta^*\|_2} \right\|_2^2 \le \mathrm{polylog}(n) \max\left\{ \frac{d}{n\|\theta^*\|_2}, \frac{d^2}{n^2} \right\}. \tag{76}$$

To bound I_{12}, notice that it is zero-mean and satisfies

$$\left\| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \langle x_i, e_{\theta^*}^{\perp} \rangle \right\|_{\psi_2} = \left\| \Big\langle \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} \{ y_i x_i - \mathbb{E}(y_i x_i) \}, e_{\theta^*}^{\perp} \Big\rangle \right\|_{\psi_2} \lesssim \frac{\|e\|_2}{\sqrt{n}}$$

by a reasoning similar to Equations (72)–(73). Combining with the bound on $\|e\|_2$ in Lemma 1, we reach

$$\left\| \frac{1}{(1-\nu)n} \sum_{i=\nu n+1}^{n} y_i \langle x_i, e_{\theta^*}^{\perp} \rangle \right\|_{\psi_2} \lesssim \sqrt{\frac{\mathrm{polylog}\, n}{n}} \left(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n} \right).$$

Hence, by a sub-Gaussian tail bound similarly to (74),

$$I_{22} \lesssim \sqrt{\frac{\mathrm{polylog}\, n}{n}} \max\left\{ \sqrt{\frac{d}{n\|\theta^*\|_2}}, \frac{d}{n} \right\} \tag{77}$$

holds with probability at least $1 - \frac{1}{n^2}$. Substituting (76) and (77) into (75) yields

$$I_2 \le \mathrm{polylog}(n) \left(\frac{\sqrt{d}}{n\sqrt{\|\theta^*\|_2}} + \frac{d}{n^{3/2}} + \frac{d^2}{n^2} + \frac{d}{n\|\theta^*\|_2} \right) \tag{78}$$

Combining (74) and (78), under $n \gtrsim d$ and $\|\theta^*\|_2 \ge 1$, we obtain

$$\left| \langle \hat{v}, \hat{r} \rangle - q(\|\theta^*\|_2) \right| \le \mathrm{polylog}(n) \left(\frac{1}{\sqrt{n}} + \frac{d}{n\|\theta^*\|_2} + \frac{d^2}{n^2} \right)$$

as claimed. □

We now examine the effect of applying q^{-1} and seek to bound $|\hat{\lambda} - \|\theta^*\|_2| = |q^{-1}(\hat{v}^\top \hat{r}) - q^{-1}(q(\|\theta^*\|_2))|$.

Lemma 9. *Assume that the event (71) in Lemma 8 holds, i.e.,*

$$\left|\hat{v}^\top \hat{r} - q(\|\theta^*\|_2)\right| \le \mathrm{polylog}(n)\left(\frac{1}{\sqrt{n}} + \frac{d}{n\|\theta^*\|_2} + \frac{d^2}{n^2}\right). \tag{79}$$

Under

$$n \ge \mathrm{polylog}(n)\left(\|\theta^*\|_2^4 + \|\theta^*\|_2 d\right) \tag{80}$$

we have

$$|\hat{\lambda} - \|\theta^*\|_2| \le \mathrm{polylog}(n)\left(\|\theta^*\|_2^3\sqrt{\frac{1}{n}} + \frac{\|\theta^*\|_2^2 d}{n}\right) \tag{81}$$

Proof. Letting $q(\infty) = \frac{1}{\sqrt{2\pi}}$, a key fact in this proof is the order of the asymptotic gap (cf. Lemma 22)

$$G(\tau) := q(\infty) - q(\tau) \asymp \frac{1}{\tau^2}, \qquad \forall \tau \ge \frac{1}{2}, \tag{82}$$

which implies

$$q(\infty) - q(\|\theta^*\|_2) \asymp \frac{1}{\|\theta^*\|_2^2}. \tag{83}$$

On the other hand, (80) and (79) together ensure that

$$|\hat{v}^\top \hat{r} - q(\|\theta^*\|_2)| \le \frac{c_*}{\|\theta^*\|_2^2} \quad \text{for some small enough } c_*. \tag{84}$$

Combining (83) and (84) gives

$$\hat{v}^\top \hat{r} < q(\infty) - \frac{c_1}{\|\theta^*\|_2^2} \quad \text{for some universal constant } c_1,$$

hence it is meaningful to apply q^{-1} to $\hat{v}^\top \hat{r}$.

Now we write the right-hand side of (79) as E, which satisfies $E \le \frac{c_*}{\|\theta^*\|_2^2}$, then we have

$$q(\|\theta^*\|_2) - E \le \hat{v}^\top \hat{r} \le q(\|\theta^*\|_2) + E$$

and moreover

$$\begin{aligned} |\hat{\lambda} - \|\theta^*\|_2| &= \left|q^{-1}(\hat{v}^\top \hat{r}) - q^{-1}(q(\|\theta^*\|_2))\right| \\ &\le \max\left\{q^{-1}\Big(q(\|\theta^*\|_2) + E\Big) - \|\theta^*\|_2, \|\theta^*\|_2 - q^{-1}\Big(q(\|\theta^*\|_2) - E\Big)\right\} \end{aligned} \tag{85}$$

▶ by the monotonicity of q and q^{-1} (cf. Lemma 21)

We shall first bound

$$\hat{\lambda}_{\max} := q^{-1}\Big(q(\|\theta^*\|_2) + E\Big)$$

which satisfies

$$q(\hat{\lambda}_{\max}) = q(\|\theta^*\|_2) + E \tag{86}$$

The asymptotic gap is given by

$$G(\hat{\lambda}_{\max}) = q(\infty) - q(\hat{\lambda}_{\max}) = q(\infty) - q(\|\theta^*\|_2) - E = G(\|\theta^*\|_2) - E. \tag{87}$$

Now using (82) and $E \le \frac{c_*}{\|\theta^*\|_2^2}$ for some small enough c_*,

$$\frac{1}{\hat{\lambda}_{\max}^2} \asymp G(\hat{\lambda}_{\max}) = G(\|\theta^*\|_2) - E \gtrsim \frac{1}{\|\theta^*\|_2^2} \implies \hat{\lambda}_{\max} \lesssim \|\theta^*\|_2 \tag{88}$$

By the mean value theorem and the derivative for an inverse function,

$$
\begin{aligned}
&q^{-1}\Big(q(\|\theta^*\|_2)+E\Big)-\|\theta^*\|_2\\
&=q^{-1}\Big(q(\|\theta^*\|_2)+E\Big)-q^{-1}\big(q(\|\theta^*\|_2)\big)\\
&=q^{-1}(q(\hat{\lambda}_{\max}))-q^{-1}\big(q(\|\theta^*\|_2)\big)\\
&\quad\blacktriangleright \text{ by Equation (86)}\\
&\le (q^{-1})'(\xi_0)\big(q(\hat{\lambda}_{\max})-q(\|\theta^*\|_2)\big)\\
&\quad\blacktriangleright \text{ by mean value theorem, } \xi_0\in[q(\|\theta^*\|_2),q(\hat{\lambda}_{\max})]\\
&=\frac{E}{q'(\hat{\xi}_0)}\\
&\quad\blacktriangleright \text{ by derivative of inverse function, } \hat{\xi}_0\in[\|\theta^*\|_2,\hat{\lambda}_{\max}], \text{ and (86)}\\
&\lesssim \|\theta^*\|_2^3 E.\\
&\quad\blacktriangleright \text{ By (88) and Lemma 15}
\end{aligned}
$$

We now similarly bound the second term in (85),

$$
\begin{aligned}
&\|\theta^*\|_2-q^{-1}\Big(q(\|\theta^*\|_2)-E\Big)\\
&=q^{-1}(q(\|\theta^*\|_2))-q^{-1}\Big(q(\|\theta^*\|_2)-E\Big)\\
&=(q^{-1})'(\xi_1)\cdot E\\
&\quad\blacktriangleright \text{ by mean value theorem, } \xi_1\in[q(\|\theta^*\|_2)-E,q(\|\theta^*\|_2)]\\
&=\frac{1}{q'(\hat{\xi}_1)}E\\
&\quad\blacktriangleright \text{ by derivative of inverse function, } \hat{\xi}_1<\|\theta^*\|_2\\
&\lesssim \|\theta^*\|_2^3 E.\\
&\quad\blacktriangleright \text{ By Lemma 15}
\end{aligned}
$$

Substituting the past two displays and $E=\mathrm{polylog}(n)\big(\frac{1}{\sqrt{n}}+\frac{d}{n\|\theta^*\|_2}+\frac{d^2}{n^2}\big)$ into (85) gives

$$
|\hat{\lambda}-\|\theta^*\|_2|\le \mathrm{polylog}(n)\bigg(\frac{\|\theta^*\|_2^3}{\sqrt{n}}+\frac{\|\theta^*\|_2^2 d}{n}+\frac{\|\theta^*\|_2^3 d^2}{n^2}\bigg)
$$

which together with $\frac{\|\theta^*\|_2^3 d^2}{n^2}\le\frac{\|\theta^*\|_2^2 d}{n}$ under $n\gtrsim\|\theta^*\|_2 d$ concludes the proof. □

We now substitute (70) and (81) into (69) to complete the proof of Theorem 3.

D Technical Lemmas

Lemma 10 (Stein identities [33, 34])**.** *Let $g\sim N(0,1)$ and let $\phi:\mathbb{R}\to\mathbb{R}$ be absolutely continuous with $\mathbb{E}|\phi'(g)|<\infty$, $\mathbb{E}|g\phi(g)|<\infty$. Then*

$$
\mathbb{E}[g\phi(g)]=\mathbb{E}[\phi'(g)]. \tag{89}
$$

Moreover, let $x\sim N(0,I_d)$ and let $\Phi:\mathbb{R}^d\to\mathbb{R}$ be continuously differentiable with $\mathbb{E}\|\nabla\Phi(x)\|_2<\infty$, $\mathbb{E}\|x\Phi(x)\|_2<\infty$. Then

$$
\mathbb{E}[x\Phi(x)]=\mathbb{E}[\nabla\Phi(x)]. \tag{90}
$$

In particular, for $\Phi(x)=f(x^\top\theta)$,

$$
\mathbb{E}[xf(x^\top\theta)]=\mathbb{E}[f'(x^\top\theta)]\theta, \tag{91}
$$

provided both sides are well-defined.

Lemma 11. *[5, Theorem 2.10] Let $X_1, ..., X_n$ be independent random variables, and assume that for some $v, c > 0$,*

$$\sum_{i=1}^{n} \mathbb{E}X_i^2 \le v,$$
$$\sum_{i=1}^{n} \mathbb{E}|X_i|^q \le \frac{q!}{2} v c^{q-2}, \quad \textit{for all integers } q \ge 3,$$

then we have

$$\mathbb{P}\bigg(\Big|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\Big| \ge \sqrt{2vt} + ct\bigg) \le 2\exp(-t), \quad \forall t > 0.$$

Lemma 12. *[36, Theorem 2.8.1] Let $X_1, ..., X_N$ be independent, zero-mean, sub-exponential random variables. Then for every $t \ge 0$, for some absolute constant c we have*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{N} X_i\Big| \ge t\right) \le 2\exp\left(-c\min\Big\{\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1\le i\le N}\|X_i\|_{\psi_1}}\Big\}\right)$$

Lemma 13. *[36, Theorem 7.3.1 & Corollary 7.3.3] Let $X \in R^{n\times d}$ be a matrix with independent $N(0,1)$ entries. Then $\mathbb{E}\|X\|_{\rm op} \le \sqrt{n} + \sqrt{d}$ and*

$$\mathbb{P}\bigg(\|X\|_{\rm op} \ge \sqrt{n} + \sqrt{d} + t\bigg) \le 2\exp(-ct^2), \quad \forall t \ge 0.$$

D.1 Functions Arising in the Proofs

Lemma 14. *For $s(a) = \frac{1}{1+e^{-a}}$, we have*

$$s'(a) = \frac{e^{-a}}{(1+e^{-a})^2} = s(a)\big(1 - s(a)\big)$$

and $\sup_{a\in\mathbb{R}} |s'(a)| \le \frac{1}{4}$.

Proof. It can be shown by direct differentiation. □

Lemma 15 (Order of $m(\tau)$ and $q'(\tau)$)**.** *Let $g \sim N(0,1)$ and*

$$m(\tau) = \mathbb{E}[s'(\tau g)],$$
$$q(\tau) = \tau m(\tau) = \mathbb{E}[\tau s'(\tau g)] = \mathbb{E}[s(\tau g)g].$$

Then, for all $\tau \ge 0$,

$$m(\tau) \asymp \frac{1}{1+\tau}, \qquad q'(\tau) = \mathbb{E}[s'(\tau g)g^2] \asymp \frac{1}{(1+\tau)^3}.$$

Proof. We first note that $s'(a) = \frac{1}{(e^{a/2}+e^{-a/2})^2}$, hence there exist universal constants $c, C > 0$ such that

$$ce^{-|a|} \le s'(a) \le Ce^{-|a|}, \qquad \forall a \in \mathbb{R}. \tag{92}$$

Proof of $m(\tau) \asymp \frac{1}{1+\tau}$. In view of (92),

$$m(\tau) = \mathbb{E}[s'(\tau g)] \asymp \mathbb{E}[e^{-\tau|g|}].$$

To show $m(\tau) \asymp \frac{1}{1+\tau}$, all that remains is to establish $\mathbb{E}[e^{-\tau|g|}] \asymp \frac{1}{1+\tau}$. To show this, note that

$$\mathbb{E}[e^{-\tau|g|}] = \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\tau x}e^{-x^2/2}\,dx.$$

For the upper bound, since $e^{-x^2/2} \le 1$,

$$\int_0^\infty e^{-\tau x} e^{-x^2/2}\, dx \le \int_0^\infty e^{-\tau x}\, dx = \frac{1}{\tau}$$

for $\tau > 0$. While for $0 \le \tau \le 1$ the integral is bounded by an absolute constant. Taken collectively,

$$\mathbb{E}[e^{-\tau|g|}] \lesssim \frac{1}{1+\tau}.$$

For the lower bound, integrate over $0 \le x \le (1+\tau)^{-1}$. On this interval, we have

$$\tau x \le 1, \qquad x^2/2 \le 1/2,$$

and in turn $e^{-\tau x} e^{-x^2/2} \ge e^{-3/2}$, which yields

$$\int_0^\infty e^{-\tau x} e^{-x^2/2}\, dx \ge \int_0^{(1+\tau)^{-1}} e^{-\tau x} e^{-x^2/2}\, dx \gtrsim \frac{1}{1+\tau}.$$

This proves $m(\tau) \asymp \frac{1}{1+\tau}$.

Proof of $q'(\tau) \asymp \frac{1}{(1+\tau)^3}$. By Stein's identity (cf. Equation (89) of Lemma 10),

$$q(\tau) = \tau \mathbb{E}[s'(\tau g)] = \mathbb{E}[s(\tau g) g],$$

and differentiating gives $q'(\tau) = \mathbb{E}[s'(\tau g) g^2]$. Using again $s'(a) \asymp e^{-|a|}$ from (92), we get

$$q'(\tau) \asymp \mathbb{E}[g^2 e^{-\tau|g|}].$$

Now

$$\mathbb{E}[g^2 e^{-\tau|g|}] = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^2 e^{-\tau x} e^{-x^2/2}\, dx.$$

For the upper bound, since $e^{-x^2/2} \le 1$,

$$\int_0^\infty x^2 e^{-\tau x} e^{-x^2/2}\, dx \le \int_0^\infty x^2 e^{-\tau x}\, dx = \frac{2}{\tau^3}$$

for $\tau > 0$. For $0 \le \tau \le 1$, the expectation is bounded by an absolute constant. Hence

$$\mathbb{E}[g^2 e^{-\tau|g|}] \lesssim \frac{1}{(1+\tau)^3}.$$

For the lower bound, we shall integrate over $0 \le x \le (1+\tau)^{-1}$. On this interval, we have $e^{-\tau x} e^{-x^2/2} \ge e^{-3/2}$. Thus

$$\int_0^\infty x^2 e^{-\tau x} e^{-x^2/2}\, dx \ge \int_0^{(1+\tau)^{-1}} x^2 e^{-\tau x} e^{-x^2/2}\, dx \gtrsim \int_0^{(1+\tau)^{-1}} x^2\, dx \asymp \frac{1}{(1+\tau)^3}.$$

Therefore

$$q'(\tau) = \mathbb{E}[s'(\tau g) g^2] \asymp \frac{1}{(1+\tau)^3}.$$

This completes the proof. □

Lemma 16 (Comparing $m(\tau)$ and $q'(\tau)$)**.** *Define $m(\tau)$ and $q'(\tau)$ as in Lemma 15. We have*

$$m(\tau) > q'(\tau), \quad \forall \tau > 0.$$

Proof. In view of Lemma 14, s' is an even function and is strictly decreasing in $|a|$. Let $X := g^2$. Then $X \ge 0$ and $\mathbb{E}X = 1$. For each fixed $\tau \ge 0$, define

$$\varphi_\tau(x) := s'(\tau\sqrt{x}), \qquad x \ge 0.$$

Since $s'(a)$ is decreasing in $|a|$, the function φ_τ is nonincreasing on $[0,\infty)$. Moreover, if $\tau > 0$, then φ_τ is strictly decreasing. With this notation,

$$m(\tau) = \mathbb{E}[\varphi_\tau(X)], \qquad q'(\tau) = \mathbb{E}[X\varphi_\tau(X)].$$

Therefore

$$q'(\tau) - m(\tau) = \mathbb{E}[X\varphi_\tau(X)] - \mathbb{E}[X]\mathbb{E}[\varphi_\tau(X)] = \mathrm{Cov}(X, \varphi_\tau(X)). \tag{93}$$

We claim this covariance is not greater than 0. Indeed, let X' be an independent copy of X. Then

$$\mathrm{Cov}(X, \varphi_\tau(X)) = \frac{1}{2}\mathbb{E}\Big[(X - X')(\varphi_\tau(X) - \varphi_\tau(X'))\Big].$$

Since φ_τ is nonincreasing, we have $(X - X')(\varphi_\tau(X) - \varphi_\tau(X')) \le 0$, which yields

$$\mathrm{Cov}(X, \varphi_\tau(X)) \le 0.$$

In light of (93),

$$q'(\tau) - m(\tau) \le 0,$$

which gives $m(\tau) \ge q'(\tau)$. If $\tau > 0$, then φ_τ is strictly decreasing and $X = g^2$ is nondegenerate, so the covariance is strictly negative. Therefore

$$m(\tau) > q'(\tau), \qquad \forall \tau > 0,$$

as claimed. □

Lemma 17 (Upper bound on $\mathbb{E}(s'(\tau g)|g|^q)$)**.** *For every integer $q \ge 3$ and every $\tau \ge 1$,*

$$\mathbb{E}_{g\sim N(0,1)}\big[s'(\tau g)|g|^q\big] \le \frac{2q!}{\sqrt{2\pi}}\frac{1}{\tau^{q+1}}.$$

Proof. Let $H_q(\tau) := \mathbb{E}_{g\sim N(0,1)}\big[s'(\tau g)|g|^q\big]$ for convenience. In light of (92),

$$H_q(\tau) \le \mathbb{E}\big[e^{-\tau|g|}|g|^q\big] = \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\tau x}x^q e^{-x^2/2}\,dx.$$

Since $e^{-x^2/2} \le 1$, we have

$$\mathbb{E}\big[e^{-\tau|g|}|g|^q\big] \le \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\tau x}x^q\,dx.$$

By the change of variables $u = \tau x$,

$$\int_0^\infty e^{-\tau x}x^q\,dx = \frac{1}{\tau^{q+1}}\int_0^\infty e^{-u}u^q\,du = \frac{\Gamma(q+1)}{\tau^{q+1}} = \frac{q!}{\tau^{q+1}}.$$

Therefore,

$$H_q(\tau) \le \frac{2q!}{\sqrt{2\pi}}\frac{1}{\tau^{q+1}}.$$

This proves the claim. □

Lemma 18 (Map $F : \mathbb{R}^d \to \mathbb{R}^d$ in Equation (28))**.** *Let $F(z) = m(\|z\|_2)z$ for $z \in \mathbb{R}^d$ where $m(\tau) = \mathbb{E}_{g\sim(0,1)}(s'(\tau g))$. Then we have:*

- *The Jacobian of F at $z \ne 0$ is given by*

$$\nabla F(z) = m(\|z\|_2)I_d + \frac{m'(\|z\|_2)}{\|z\|_2}zz^\top. \tag{94}$$

- *With the convention $\nabla F(0) := m(0)I_d$, we have the integral representation*

$$F(u) - F(v) = \int_0^1 \nabla F(v + th)h\,dt \tag{95}$$

where $h = u - v$.

Proof. We first compute the Jacobian (94). Write $\rho := \|z\|_2 = \left(\sum_{k=1}^d z_k^2\right)^{1/2}$. The i-th component of F is $F_i(z) = m(\rho)z_i$. For $z \neq 0$, we have $\frac{\partial \rho}{\partial z_j} = \frac{z_j}{\rho}$. Therefore, by the product rule and the chain rule,

$$\frac{\partial F_i(z)}{\partial z_j} = \frac{\partial}{\partial z_j}\big(m(\rho)z_i\big) = m'(\rho)\frac{\partial \rho}{\partial z_j}z_i + m(\rho)\frac{\partial z_i}{\partial z_j} = m'(\rho)\frac{z_j}{\rho}z_i + m(\rho)\mathbf{1}_{\{i=j\}}.$$

Thus the (i,j)-th entry of the Jacobian matrix is

$$(\nabla F(z))_{ij} = m(\rho)\mathbf{1}_{\{i=j\}} + \frac{m'(\rho)}{\rho}z_i z_j.$$

In matrix form, this is exactly

$$\nabla F(z) = m(\rho)I_d + \frac{m'(\rho)}{\rho}zz^\top = m(\|z\|_2)I_d + \frac{m'(\|z\|_2)}{\|z\|_2}zz^\top.$$

We next prove the integral representation (95). Let $h := u - v$ and define the line segment $\gamma(t) := v + th,\ \ t \in [0,1]$. We shall assume that $\gamma(t) \neq 0$ for $t \in [0,1]$; we can cover the remaining cases by a separate treatment (or an approximation argument) and the convention $\nabla F(0) = m(0)I_d$ does not affect the value of the integral. For each coordinate $i = 1, \ldots, d$, define the one-dimensional function $\psi_i(t) := F_i(\gamma(t)) = F_i(v + th)$. By the chain rule,

$$\psi_i'(t) = \sum_{j=1}^d \frac{\partial F_i(\gamma(t))}{\partial z_j}h_j.$$

This is the i-th component of the vector $\nabla F(\gamma(t))h$. Hence, by the fundamental theorem of calculus,

$$F_i(u) - F_i(v) = \psi_i(1) - \psi_i(0) = \int_0^1 \psi_i'(t)\,dt = \int_0^1 \big(\nabla F(v+th)h\big)_i\,dt.$$

Since this holds for every coordinate i, we obtain the vector identity in (95). This completes the proof. □

Lemma 19 (Upper bound on $m'(\tau)$)**.** *For $m(\tau) := \mathbb{E}_{g\sim N(0,1)}[s'(\tau g)]$, we have*

$$|m'(\tau)| \lesssim \frac{1}{\tau^2}, \qquad \tau \geq \frac{1}{2}.$$

Proof. Differentiating under the expectation gives

$$m'(\tau) = \mathbb{E}[s''(\tau g)g].$$

We now use the explicit formula for s''. Since $s'(a) = \frac{1}{(e^{a/2}+e^{-a/2})^2}$, we have

$$s''(a) = -\frac{e^{a/2} - e^{-a/2}}{(e^{a/2}+e^{-a/2})^3} = \frac{e^{-a/2} - e^{a/2}}{(e^{a/2}+e^{-a/2})^3}.$$

Therefore,

$$|s''(a)| = \frac{|e^{a/2} - e^{-a/2}|}{(e^{a/2}+e^{-a/2})^3}.$$

If $a \geq 0$, then

$$|s''(a)| \leq \frac{e^{a/2}}{e^{3a/2}} = e^{-a}.$$

If $a < 0$, then

$$|s''(a)| \leq \frac{e^{-a/2}}{e^{-3a/2}} = e^{a} = e^{-|a|}.$$

Hence, for all $a \in \mathbb{R}$,

$$|s''(a)| \leq e^{-|a|}. \tag{96}$$

It follows that, for $\tau \geq 1/2$,

$$|m'(\tau)| \leq \mathbb{E}[|s''(\tau g)||g|] \leq \mathbb{E}[e^{-\tau|g|}|g|].$$

Using the Gaussian density,

$$\mathbb{E}[e^{-\tau|g|}|g|] = \frac{2}{\sqrt{2\pi}} \int_0^\infty x e^{-\tau x} e^{-x^2/2}\, dx.$$

Since $e^{-x^2/2} \leq 1$, we obtain

$$\mathbb{E}[e^{-\tau|g|}|g|] \leq C \int_0^\infty x e^{-\tau x}\, dx.$$

Finally,

$$\int_0^\infty x e^{-\tau x}\, dx = \frac{1}{\tau^2} \int_0^\infty u e^{-u}\, du = \frac{1}{\tau^2},$$

where we used the change of variables $u = \tau x$. Therefore,

$$|m'(\tau)| \leq \frac{C}{\tau^2}, \qquad \tau \geq \frac{1}{2}.$$

This proves the claim. □

Lemma 20 (Upper bound on $q''(\tau)$)**.** *For $q(\tau)$ defined in Lemma 15, we have*

$$|q''(\tau)| \lesssim \frac{1}{\tau^4}, \qquad \tau \geq \frac{1}{2}.$$

Proof. Recall that $q'(\tau) = \mathbb{E}[s'(\tau g) g^2]$. Differentiating under the expectation gives

$$q''(\tau) = \mathbb{E}[s''(\tau g) g^3].$$

Recall from Equation (96) in the proof of Theorem 19 that, for all $a \in \mathbb{R}$,

$$|s''(a)| \leq e^{-|a|}.$$

It follows that, for $\tau \geq 1/2$,

$$|q''(\tau)| \leq \mathbb{E}[|s''(\tau g)||g|^3] \leq \mathbb{E}[e^{-\tau|g|}|g|^3].$$

Using the Gaussian density,

$$\mathbb{E}[e^{-\tau|g|}|g|^3] = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^3 e^{-\tau x} e^{-x^2/2}\, dx.$$

Since $e^{-x^2/2} \leq 1$, we get

$$\mathbb{E}[e^{-\tau|g|}|g|^3] \leq C \int_0^\infty x^3 e^{-\tau x}\, dx.$$

Finally,

$$\int_0^\infty x^3 e^{-\tau x}\, dx = \frac{1}{\tau^4} \int_0^\infty u^3 e^{-u}\, du = \frac{6}{\tau^4},$$

where we used the change of variables $u = \tau x$. Therefore,

$$|q''(\tau)| \leq \frac{C}{\tau^4}, \qquad \tau \geq \frac{1}{2}.$$

This proves the claim. □

Lemma 21. *Let $q(\tau)$ be defined as in Lemma 15. $q(\tau)$ is strictly increasing on $[0, \infty)$ and satisfies*

$$q(0) = 0, \qquad \lim_{\tau \to \infty} q(\tau) = \frac{1}{\sqrt{2\pi}}.$$

Consequently, q admits an inverse on $(0, \frac{1}{\sqrt{2\pi}})$, which we denote by $q^{-1} : (0, \frac{1}{\sqrt{2\pi}}) \to (0, \infty)$.

Proof. By Lemma 15, for every $\tau \geq 0$,

$$q(\tau) = \tau \mathbb{E}[s'(\tau g)] = \mathbb{E}[s(\tau g) g].$$

We first prove monotonicity. Differentiating $q(\tau) = \mathbb{E}[s(\tau g) g]$ gives

$$q'(\tau) = \mathbb{E}[s'(\tau g) g^2].$$

Since $s'(a) = \frac{1}{(e^{a/2} + e^{-a/2})^2} > 0$ for every $a \in \mathbb{R}$, and since $\mathbb{P}(g \neq 0) = 1$, we have

$$q'(\tau) = \mathbb{E}[s'(\tau g) g^2] > 0$$

for every $\tau \geq 0$. Hence q is strictly increasing on $[0, \infty)$.

Next,

$$q(0) = 0 \cdot m(0) = 0.$$

For the limit as $\tau \to \infty$, using again the representation

$$q(\tau) = \mathbb{E}[s(\tau g) g],$$

we note that for every $g \neq 0$,

$$s(\tau g) \to \mathbf{1}_{\{g > 0\}} \qquad \text{as } \tau \to \infty.$$

Moreover, we have $|s(\tau g) g| \leq |g|$ and $\mathbb{E}|g| < \infty$. Thus, by dominated convergence,

$$\lim_{\tau \to \infty} q(\tau) = \mathbb{E}[g \mathbf{1}_{\{g > 0\}}] = \frac{1}{\sqrt{2\pi}} \int_0^\infty x e^{-x^2/2} \, dx = \frac{1}{\sqrt{2\pi}}.$$

Since q is continuous and strictly increasing from 0 to $1/\sqrt{2\pi}$, it admits an inverse on $\left(0, \frac{1}{\sqrt{2\pi}}\right)$. □

Lemma 22. *Let $q(\tau)$ be defined as in Lemma 15. For every $\tau \geq 1/2$,*

$$\frac{1}{\sqrt{2\pi}} - q(\tau) \asymp \frac{1}{\tau^2}.$$

Proof. In light of Lemma 15,

$$q(\tau) = \tau \mathbb{E}[s'(\tau g)] = \mathbb{E}[s(\tau g) g].$$

Therefore,

$$\frac{1}{\sqrt{2\pi}} - q(\tau) = \mathbb{E}[g \mathbf{1}_{\{g > 0\}}] - \mathbb{E}[s(\tau g) g].$$

Using the symmetry of the standard Gaussian distribution, we can write

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} - q(\tau) &= \int_0^\infty x \phi(x) \, dx - \int_{\mathbb{R}} s(\tau x) x \phi(x) \, dx \\ &= \int_0^\infty x \phi(x) \, dx - \int_0^\infty x \phi(x) \big(s(\tau x) - s(-\tau x)\big) \, dx, \end{aligned}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is the density function of $N(0,1)$. Since $s(a) - s(-a) = 2s(a) - 1$, we get

$$1 - \big(s(\tau x) - s(-\tau x)\big) = 1 - (2s(\tau x) - 1) = 2(1 - s(\tau x)) = \frac{2}{1 + e^{\tau x}}.$$

Hence

$$\frac{1}{\sqrt{2\pi}} - q(\tau) = 2 \int_0^\infty \frac{x \phi(x)}{1 + e^{\tau x}} \, dx.$$

Changing variables $u = \tau x$ gives

$$\frac{1}{\sqrt{2\pi}} - q(\tau) = \frac{2}{\tau^2} \int_0^\infty \frac{u \phi(u/\tau)}{1 + e^u} \, du.$$

We now prove matching upper and lower bounds. For the upper bound, since

$$\phi(u/\tau) \le \phi(0) = \frac{1}{\sqrt{2\pi}},$$

we have

$$\frac{1}{\sqrt{2\pi}} - q(\tau) \le \frac{2\phi(0)}{\tau^2} \int_0^\infty \frac{u}{1+e^u}\, du \le \frac{C}{\tau^2}.$$

For the lower bound, restrict the integral to $u \in [0,1]$. Since $\tau \ge 1/2$, we have $u/\tau \le 2$ for $u \in [0,1]$. Thus

$$\phi(u/\tau) \ge \phi(2) > 0 \qquad \text{for } 0 \le u \le 1.$$

Therefore,

$$\frac{1}{\sqrt{2\pi}} - q(\tau) \ge \frac{2\phi(2)}{\tau^2} \int_0^1 \frac{u}{1+e^u}\, du \ge \frac{c}{\tau^2}.$$

Combining the two bounds proves the claim. □

Lemma 23. *Under Assumptions 1–2, we have*

$$v = \mathbb{E}[y_i x_i] = q(\|\theta^*\|_2)\frac{\theta^*}{\|\theta^*\|_2},$$

where $q(\tau) := \mathbb{E}_{g\sim N(0,1)}[s(\tau g)g]$. Note that by Stein's identity $q(\tau) = \mathbb{E}[\tau s'(\tau g)]$ and thus coincides with the definition in Lemma 15.

Proof. By conditioning on x_i,

$$\mathbb{E}[y_i x_i] = \mathbb{E}[s(x_i^\top \theta^*)x_i].$$

If $\theta^* = 0$, the claim is immediate since the expectation is zero. Assume $\theta^* \neq 0$ and set $e := \frac{\theta^*}{\|\theta^*\|_2}$. By rotational invariance of the standard Gaussian distribution, $\mathbb{E}[s(x_i^\top \theta^*)x_i]$ must be parallel to θ^*, and its component along e is

$$\big\langle \mathbb{E}[s(x_i^\top \theta^*)x_i], e\big\rangle = \mathbb{E}[s(\|\theta^*\|_2 g)g] = q(\|\theta^*\|_2),$$

where $g = x_i^\top e \sim N(0,1)$. Hence $\mathbb{E}[y_i x_i] = q(\|\theta^*\|_2)e = q(\|\theta^*\|_2)\frac{\theta^*}{\|\theta^*\|_2}$. □

E Estimator in [23] and Algorithm 2

In this small appendix we provide more details on the estimator of [23], in order to support Equation (11) and Lemma 1 in the main text.

E.1 Sub-Gradient Descent

We first explain that Equation (11), or equivalently

$$\tilde{r}_{t+1} = \hat{r}_t - \frac{\sqrt{2\pi}}{n}\sum_{i=1}^n \big(\frac{\mathrm{sign}(x_i^\top \hat{r}_t)+1}{2} - y_i\big)x_i, \tag{97}$$

$$\hat{r}_{t+1} = \frac{\tilde{r}_{t+1}}{\|\tilde{r}_{t+1}\|_2} \tag{98}$$

as written in Lemma 1, can be viewed as a (sub-)gradient descent algorithm with respect to the ReLU loss. To see this, we transition to the $\{-1,1\}$-valued responses $\tilde{y}_i = 2y_i - 1$, consider the Hamming distance loss

$$L_{\mathrm{ham}}(u) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\big(\tilde{y}_i \neq \mathrm{sign}(x_i^\top u)\big) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\big(-\tilde{y}_i x_i^\top u \ge 0\big)$$

and then consider its convex relaxation, the ReLU loss obtained by replacing $\mathbf{1}(a>0)$ to $\mathrm{ReLU}(a) = \max\{a,0\} = \frac{a+|a|}{2}$,

$$L_{\mathrm{ReLU}}(u) = \frac{1}{2n}\sum_{i=1}^n \Big(|x_i^\top u| - \tilde{y}_i x_i^\top u\Big).$$

It is easy to see that a sub-gradient of the ReLU loss is given by

$$\partial L_{\mathrm{ReLU}}(u) = \frac{1}{2n}\sum_{i=1}^{n}\Big(\operatorname{sign}(x_i^\top u) - \tilde{y}_i\Big)x_i = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\operatorname{sign}(x_i^\top u)+1}{2} - y_i\Big)x_i.$$

As such, (97) is exactly the sub-gradient descent with ReLU loss. Moreover, the additional normalization step, (98), is natural since the desired true direction $\theta^*/\|\theta^*\|_2$ resides in $\mathbb{S}^{d-1}$.

E.2 Derivation of Lemma 1

We show the derivation of Lemma 1 from Corollary 7 and Theorem 5 of [23] with the sparsity level k therein set as d. To start, we note that β in [23] denotes the signal norm $\|\theta^*\|_2$. Hence $\beta \geq 1$ under Assumption 2.

We first show that (9) follows from the error rate of [23, Corollary 7]. In particular, Equation (17) therein shows that, to achieve ϵ estimation error, we require a sample size of

$$n = \tilde{\Omega}\bigg(\frac{d}{\epsilon^2\|\theta^*\|_2}\bigg), \qquad \text{if } 1 \leq \|\theta^*\|_2 \leq \frac{C_0}{\epsilon}; \tag{99}$$

$$n = \tilde{\Omega}\bigg(\frac{d}{\epsilon}\bigg), \qquad \text{if } \|\theta^*\|_2 > \frac{C_0}{\epsilon}. \tag{100}$$

This implies an error rate of

$$\epsilon_0 = \tilde{O}\bigg(\sqrt{\frac{d}{n\|\theta^*\|_2}} + \frac{d}{n}\bigg)$$

under any $\|\theta^*\|_2 \geq 1$, as with our Equation (9).

We then show the probability of $1 - e^{-n}$ and the iteration complexity of $\log_2\log_2\frac{n}{d}$, as stated in our Lemma 1. In fact, we can set ρ of [23, Theorem 5], the probability parameter therein, as e^{-n}. In view of Equation (14) therein, this can only add factor of $\mathrm{polylog}(n)$ to the sample complexity, which remains at $n \gtrsim d\,\mathrm{polylog}\,n$, as assumed in our Lemma 1. Moreover, in view of Equation (16) of [23], the sequence of direction estimates $\{\hat{r}_t\}_{t\geq 0}$ achieves a very fast convergence rate,

$$\bigg\|\hat{r}_t - \frac{\theta^*}{\|\theta^*\|_2}\bigg\|_2 \leq 2^{2^{-t}}\epsilon_0^{1-2^{-t}}, \qquad \forall t \geq 0.$$

A direct calculation then finds that $\hat{r}_t$ with $t \geq \log_2\log_2\frac{n}{d}$ already achieves the desired direction estimation error of $\tilde{O}(\epsilon_0)$.

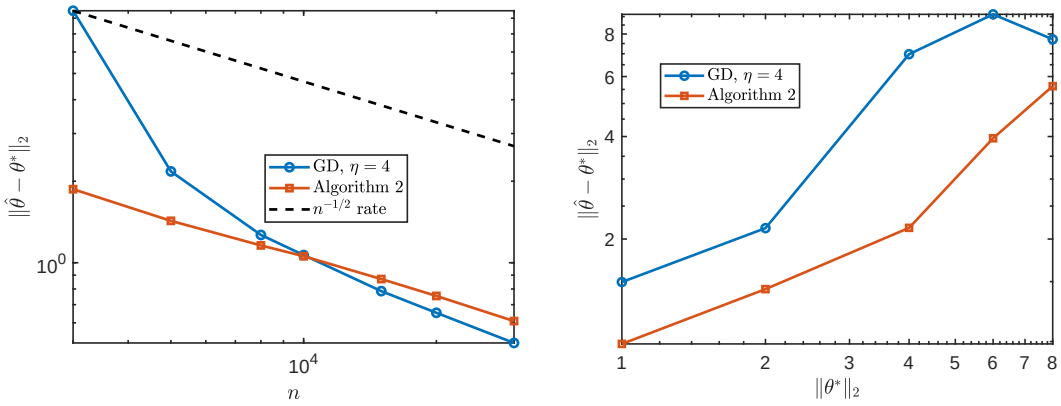
E.3 Numerical Performance of Algorithm 2

We compare the estimation performance of Algorithm 2 with GD under the following settings:

- We set $T_0 = 30$ in Algorithm 2 and run it without sample splitting — i.e., both Step 1 and Step 2 use all the n samples.
- We run GD in Algorithm 1 with $\theta_0 = 0$, $\eta = 4$ for 400 iterations. Suppose that the MLE exists, then traditional optimization theory [28] guarantees that the GD iterates converge to MLE, and as we run sufficiently many iterations we expect that the estimator, θ_{400}, well approximates MLE.

Consistent with Remark 4, our numerical simulations suggest that Algorithm 2 outperforms GD, and possibly also MLE, in high-dimensional setting, especially when n is only moderately large. We provide two numerical examples to illustrate this phenomenon:

- We first test both algorithms in a high-dimensional regime with $(d, \|\theta^*\|_2) = (1000, 2)$ and $n \in \{3000, 5000, 8000, 10000, 15000, 20000, 30000\}$. The log-log plots in Figure 2a suggest that Algorithm 2 yields more accurate estimates than GD when $n < 10000$, while GD becomes slightly better when $n > 10000$. The reference dashed line further confirms the $O(n^{-1/2})$ error decay rate of Algorithm 2.
- We next consider the setting $(n, d) = (5000, 1000)$ and vary $\|\theta^*\|_2$ over $\{1, 2, 4, 6, 8\}$. The results in Figure 2b again show that Algorithm 2 can substantially outperform GD in high-dimensional, moderate-sample regimes.



(a) Error v.s. n

(b) Error v.s. $\|\theta^*\|_2$

Figure 2: Comparing Algorithm 2 with small stepsize GD.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We write the abstract and introduction based on the contributions and scope of this paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention the limitations of our work in Remarks 5 and 6, and leave them as open questions in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full set of assumptions is given in Assumptions 1–2, and the complete and correct proofs of our theorems are provided in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose all the information needed to reproduce the experimental results of the paper in Section 4. The code will also be open sourced after acceptance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use synthetic data produced by certain random seed and the code will be open sourced after acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We disclose all the relevant information, such as choices of parameters, in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have reported the estimation error and convergence rate in Section 4 that are averaged over a sufficient number of independent trials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources needed to reproduce the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We checked the NeurIPS Code of Ethics and confirm that research conducted in the paper conform it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is primarily a theoretical work and we are not aware of potential positive societal impacts and negative societal impacts of this work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: As far as we can tell, this paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing and editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.