Likelihood Score under Generalized Self-Concordance

Anonymous Author(s)
Affiliation
Address
email

Abstract

We show how, under a generalized self-concordance assumption and possible model misspecification, we can establish non-asymptotic bounds on the normalized likelihood score when using maximum likelihood or score matching. The tail behavior is governed by an effective dimension corresponding to the trace of the sandwich covariance. We also show how our non-asymptotic approach allows us to obtain confidence set for the estimator and analyze Rao's score test.

1 Introduction

The problem of statistical inference on learned parameters is regaining the importance it deserves as machine learning and data science are increasingly impacting humanity and society through a large range of successful applications from transportation to healthcare. The classical asymptotic theory of M-estimation is well established in a rather general setting under the assumption that the parametric model is well-specified, i.e., the underlying data distribution belongs to the parametric family. We mention here, among many of them, the monographs [18, 31, 30]. The main tool is the local asymptotic normality (LAN) condition introduced by Le Cam [21]. In many real problems, the parametric model is usually an approximation to the data distribution, so it is too restrictive to assume that the model is well-specified. To relax this restriction, model misspecification has been considered in the asymptotic regime; see, e.g., [16, 34, 13]. Another limitation of classical asymptotic theory is its asymptotic regime where $n \to \infty$ and the parameter dimension d is fixed. This is inapplicable in the modern context where the data are of high dimension involving a huge number of parameters.

The non-asymptotic viewpoint has been fruitful to address higher dimensional problems—the results are developed for all fixed n so that it also captures the asymptotic regime where the parameter dimension can grow with n. Early works in this line of research focus on specific models such as Gaussian models [8, 9, 20, 6], ridge regression [15], logistic regression [4], and robust M-estimation [37]; see [5] for a survey. Spokoiny [26] addressed the finite-sample regime in full generality in a spirit similar to the classical LAN theory. The approach of [26] relies on heavy empirical process machinery and requires strong global assumptions on the deviation of the empirical risk process. More recently, Ostrovskii and Bach [23] focused on risk bounds, specializing their discussion to linear models with (pseudo) self-concordant losses and obtained a more transparent analysis under neater assumptions. A critical characteristic shared by both works is that the neighborhood of the target parameter is defined by the so-called *Dikin ellipsoid*, a geometric object identified in the theory of convex optimization [22, 7, 10, 29, 11, 12].

The Dikin ellipsoid corresponds to the distance measured by the Euclidean distance weighted by the Hessian matrix at the optimum. This weighted Euclidean distance is adapted to the geometry near the target parameter and thus leads to sharper bounds which do not depend on the minimum eigenvalue of the Hessian. This important property has been used fruitfully in various problems of learning theory and mathematical statistics [36, 35, 14].

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

2 Problem formulation

We briefly recall the framework of statistical inference via empirical risk minimization. Let $(\mathbb{Z}, \mathcal{Z})$ be a measurable space. Let $Z \in \mathbb{Z}$ be a random element following some unknown distribution $\mathbb{P}$. Consider a parametric family of distributions $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ which may or may not contain $\mathbb{P}$. We are interested in finding the parameter $\theta_\star$ so that the model $P_{\theta_\star}$ best approximates the underlying distribution $\mathbb{P}$. For this purpose, we choose a *loss function* ℓ and minimize the *population risk* $L(\theta) := \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta; Z)]$. Through out this paper, we assume that

$$\theta_\star = \arg\min_{\theta\in\Theta} L(\theta)$$

uniquely exists and satisfies $\theta_\star \in \mathrm{int}(\Theta)$, $\nabla_\theta L(\theta_\star) = 0$, and $\nabla^2_\theta L(\theta_\star) \succ 0$.

Consistent loss function. We focus on loss functions that are consistent in the following sense: When the model is *well-specified*, i.e., there exists $\theta_0 \in \Theta$ such that $\mathbb{P} = P_{\theta_0}$, it holds that $\theta_0 = \theta_\star$. We give below two popular choices of consistent loss functions.

Example 1 (Maximum likelihood estimation)**.** A widely used loss function in statistical machine learning is the negative log-likelihood $\ell(\theta; z) := -\log p_\theta(z)$ where p_θ is the probability mass/density function for the discrete/continuous case. When $\mathbb{P} = P_{\theta_0}$ for some $\theta_0 \in \Theta$, we have $L(\theta) = \mathbb{E}[-\log p_\theta(Z)] = \mathrm{KL}(p_{\theta_0}\|p_\theta) - \mathbb{E}[\log p_{\theta_0}(Z)]$ where KL is the Kullback-Leibler divergence. As a result, $\theta_0 \in \arg\min_{\theta\in\Theta} \mathrm{KL}(p_{\theta_0}\|p_\theta) = \arg\min_{\theta\in\Theta} L(\theta)$. Moreover, if there is no θ such that $p_\theta \overset{\text{a.s.}}{=} p_{\theta_0}$, then θ_0 is the unique minimizer of L.

Example 2 (Score matching estimation)**.** Another important example appears in *score matching* [17]. Assume that $\mathbb{Z} = \mathbb{R}^p$ and $\mathbb{P}$ and P_θ have densities p and p_θ w.r.t the Lebesgue measure. Let $p_\theta(z) = q_\theta(z)/\Lambda(\theta)$ where $\Lambda(\theta)$ is an unknown normalizing constant. We can choose the loss

$$\ell(\theta; z) := \Delta_z \log q_\theta(z) + \frac{1}{2}\|\nabla_z \log q_\theta(z)\|^2 + \text{const.}$$

Here $\Delta := \sum_{k=1}^p \partial^2/\partial z_k^2$ is the Laplace operator. since [17, Thm. 1]

$$L(\theta) = \frac{1}{2}\mathbb{E}\left[\|\nabla_z q_\theta(z) - \nabla_z p(z)\|^2\right],$$

it follows that when $p = p_{\theta_0}$ we have $\theta_0 \in \arg\min_{\theta\in\Theta} L(\theta)$. In fact, when $q_\theta > 0$ and there is no θ such that $p_\theta =_{a.s.} p_{\theta_0}$, the true parameter θ_0 is the unique minimizer of L [17, Thm. 2].

Empirical risk minimizer and likelihood score. We assume we have an i.i.d. sample $\{Z_i\}_{i=1}^n$ from $\mathbb{P}$. To learn the parameter $\theta_\star$, we minimize the empirical risk to obtain the *empirical risk minimizer*

$$\theta_n \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i) =: \ell_n(\theta).$$

When ℓ is chosen as the negative log-likelihood, the negative gradient $-\nabla\ell_n(\theta)$ is known as the **likelihood score**. In Sec. 3, we will establish a non-asymptotic bound for the norm of the likelihood score at $\theta_\star$ normalized by the Hessian $\nabla^2\ell_n(\theta_\star)$, i.e., $\left\|\nabla^2\ell_n(\theta_\star)^{-1/2}\nabla\ell_n(\theta_\star)\right\|$. We show how this bound can be used to obtain a finite-sample confidence bound for $\theta_\star$ constructed from θ_n. That is, for any $\delta \in (0, 1)$, we will construct a confidence set $\mathcal{C}_n(\delta) \subset \Theta$ based on θ_n such that

$$\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \geq 1 - \delta.$$

In Fig. 1, we illustrate this result for a logistic regression model. Finally, we apply our approach to analyze Rao's score test for goodness-of-fit testing.

3 Main results

3.1 Preliminaries

Notation. We denote by $S(\theta; z) := \nabla_\theta \ell(\theta; z)$ the gradient of the loss at z and $H(\theta; z) := \nabla^2_\theta \ell(\theta; z)$ the Hessian at z. Their population versions are $S(\theta) := \mathbb{E}[S(\theta; Z)]$ and $H(\theta) := \mathbb{E}[H(\theta; Z)]$, respectively. We assume standard regularity assumptions so that $S(\theta) = \nabla_\theta L(\theta)$ and $H(\theta) =$

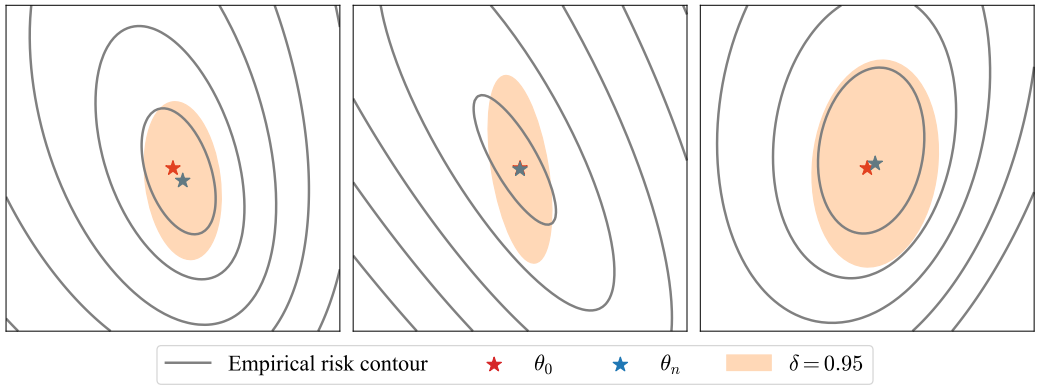

Figure 1: Confidence set in Thm. 4 under a logistic regression model with true parameter θ_0 and $X \sim \mathcal{N}(0, \Sigma)$. **Left:** $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$; **Middle:** $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$; **Right:** $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$.

$\nabla_\theta^2 L(\theta)$. The two optimality conditions then read $S(\theta_\star) = 0$ and $H(\theta_\star) \succ 0$. It follows that $\lambda_\star := \lambda_{\min}(H(\theta_\star)) > 0$ and $\lambda^\star := \lambda_{\max}(H(\theta_\star)) > 0$. Furthermore, we let $G(\theta) := \mathbb{C}\mathrm{ov}(S(\theta; Z))$ be the covariance matrix of the gradient. We define their empirical quantities as $\ell_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$, $S_n(\theta) := \frac{1}{n}\sum_{i=1}^n S(\theta; Z_i)$, $H_n(\theta) := \frac{1}{n}\sum_{i=1}^n H(\theta; Z_i)$, and

$$G_n(\theta) := \frac{1}{n}\sum_{i=1}^n [S(\theta; Z_i) - S(\theta)][S(\theta; Z_i) - S(\theta)]^\top.$$

Recall that $-S_n(\theta)$ is the likelihood score for maximum likelihood estimation. Our analysis of the estimator θ_n is local to a *Dikin ellipsoid* at $\theta_\star$ of radius r, i.e.,

$$\Theta_r(\theta_\star) := \left\{\theta \in \Theta : \|\theta - \theta_\star\|_{H(\theta_\star)} < r\right\},$$

where, given a positive semi-definite matrix J, we let $\|x\|_J := \left\|J^{1/2}x\right\|_2 = \sqrt{x^\top J x}$.

Effective dimension. A quantity that plays a central role in our analysis is the *effective dimension*:

$$d_\star := \mathbf{Tr}\left\{H(\theta_\star)^{-1/2} G(\theta_\star) H(\theta_\star)^{-1/2}\right\}. \tag{1}$$

The effective dimension appears recently in non-asymptotic analyses of (penalized) M-estimation [27, 23]. When the model is well-specified, it can be shown that $H(\theta_\star) = G(\theta_\star)$ and thus $d_\star = d$. When the model is misspecified, it can be much smaller than d depending on the spectrum of $H(\theta_\star)$ and $G(\theta_\star)$; see Appx. B for a through discussion. Moreover, it is closely connected to classical asymptotic theory of M-estimation under model misspecification.

Generalized self-concordance. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define $D_x f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x + tu)|_{t=0}$, $D_x^2 f(x)[u, v] := D_x(D_x f(x)[u])[v]$ for $x, u, v \in \mathbb{R}^d$, and $D_x^3 f(x)[u, v, w]$ similarly.

Definition 1 (Generalized self-concordance [28])**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be open and $f : \mathcal{X} \to \mathbb{R}$ be a closed convex function. For a constant $R > 0$ and an integer $\nu > 0$, we say f is (R, ν)-generalized self-concordant on $\mathcal{X}$ if

$$\left|D_x^3 f(x)[u, u, v]\right| \le R\, \|u\|^2_{\nabla^2 f(x)} \|v\|^{\nu-2}_{\nabla^2 f(x)} \|v\|_2^{3-\nu}$$

with the convention $0/0 = 0$ for the case $\nu < 2$ and $\nu > 3$. Recall that $\|u\|^2_{\nabla^2 f(x)} := u^\top \nabla^2 f(x) u$.

Remark 2. *When $\nu = 3$, this definition reduces to the standard self-concordance [22]. When $\nu = 2$, it recovers the pseudo self-concordance [4]. We give several examples in Appx. D.*

3.2 Normalized score

Our first result is a non-asymptotic bound for the normalized score at $\theta_\star$, i.e., $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$. Due to the space constraint, we state the precise assumptions in Appx. A. The problem-specific constants K_1, K_2, and σ_H are also defined in Appx. A. The proof is deferred to Appx. C.

Proposition 3. *Under Asmps. 2 and 3 with $r = 0$, it holds that, with probability at least $1 - \delta$,*

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \lesssim K_1^2 \log(e/\delta) d_\star / n$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2)\log(4d/\delta)$.

3.3 Confidence set

We then construct a finite-sample confidence set for $\theta_\star$ using the bound for the normalized score. We also provide a thorough discussion on this result in Appx. B and its proof in Appx. C. The quantities K_ν, $R_\nu^\star$, ω_ν are defined in Cor. 20, Eq. (7), Eq. (13) in the appendix, respectively.

Theorem 4. *Let $\nu \in [2, 3)$ and $\varepsilon \in (0, K_\nu]$. Suppose that Asmps 1,2, and 3 with $r = K_\nu / R_\nu^\star$ hold true. Let*

$$\mathcal{C}_n(\delta) := \theta_n + \left\{\theta \in \Theta : \theta^\top H_n(\theta_n)\theta \leq C K_1^2 \omega_\nu^2(\varepsilon) \log(4e/\delta) d_\star / n\right\}, \tag{2}$$

where C is an absolute constant. Then we have $\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \geq 1 - \delta$ whenever n satisfies

$$n \gtrsim \max\left\{(K_2^2 + \sigma_H^2)\left[\log(2d/\delta) + d\log(\omega_\nu(\varepsilon)n)\right], \left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log(e/\delta)}{\varepsilon^2}\right]^{1/(3-\nu)}\right\}. \tag{3}$$

Remark 5. *According to Huber [16], under suitable regularity assumptions, it holds that $\sqrt{n} H_n(\theta_n)^{1/2}(\theta_n - \theta_\star) \to_d L \sim \mathcal{N}(0, H(\theta_\star)^{-1/2} G(\theta_\star) H(\theta_\star)^{-1/2})$ which implies that*

$$n(\theta_n - \theta_\star)^\top H_n(\theta_n)(\theta_n - \theta_\star) \to_d L^\top L.$$

This induces an asymptotic confidence set with a similar form of (2) and radius $O(\mathbb{E}[L^\top L]/n) = O(d_\star/n)$. Our result characterizes the critical sample size *enough to enter the asymptotic regime.*

3.4 Rao's score test

We illustrate how our results can be used to analyze Rao's score test for goodness-of-fit testing. In this subsection, we will assume that the model is well-specified. We use $\theta_\star$ to denote the true parameter of $\mathbb{P}$ and reserve θ_0 for the parameter under the null hypothesis. The proof is deferred to Appx. D.

Given a subset $\Theta_0 \subset \Theta$, a goodness-of-fit testing problem is to test the hypotheses

$$\mathbf{H}_0 : \theta_\star \in \Theta_0 \leftrightarrow \mathbf{H}_1 : \theta_\star \notin \Theta_0.$$

We focus on a simple null hypothesis where $\Theta_0 := \{\theta_0\}$ is a singleton. A statistical test consists of a test statistic $T := T(Z_1, \ldots, Z_n)$ and a prescribed critical value t, and we reject the null hypothesis if $T > t$. Its performance is quantified by the *type I error rate* $\mathbb{P}(T > t \mid \mathbf{H}_0)$ and *statistical power* $\mathbb{P}(T > t \mid \mathbf{H}_1)$. The test statistic of Rao's score test is $T_{\text{Rao}} := \|S_n(\theta_0)\|^2_{H_n^{-1}(\theta_0)}$.

Theorem 6 (Rao's score statistic)**.** *Suppose that Asmps. 2 and 3 with $r = 0$ hold true. For an arbitrary $t > 0$, let $\Omega(\theta) := G(\theta)^{\frac{1}{2}} H(\theta)^{-1} G(\theta)^{\frac{1}{2}}$ and*

$$\tau_n := \left[\frac{2\|S(\theta_0)\|^2_{H^{-1}(\theta_0)}/3 - t}{4\|S(\theta_0)\|_{H^{-1}(\theta_0)}}\right]^2 - \frac{1}{n}\mathbf{Tr}(\Omega(\theta_0)).$$

(a) Under $\mathbf{H}_0$, we have, with probability at least $1 - \delta$,

$$T_{Rao} \lesssim K_1^2 \log(e/\delta)\frac{d}{n} \tag{4}$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2)\log(4d/\delta)$.

(b) Under $\mathbf{H}_1$, whenever $\tau_n \geq 0$, we have, with c being an absolute constant,

$$\begin{aligned}\mathbb{P}(T_{Rao} > t) \geq 1 - \exp\left(-c\min\left\{\frac{n^2\tau_n^2}{K_1^2\|\Omega(\theta_0)\|_2^2}, \frac{n\tau_n}{K_1\|\Omega(\theta_0)\|_\infty}\right\}\right) \\ - 2d\exp\left(-\frac{n}{4(K_2^2 + 2\sigma_H^2)}\right).\end{aligned} \tag{5}$$

Remark 7. *The bound (4) suggests that, for a fixed significance level $\alpha \in (0, 1)$, we can choose the critical value $t = t_n(\alpha) = O(d_\star/n)$ so that their type I error rates are below α. With this choice of $t_n(\alpha)$, we can then characterize the statistical power of Rao's score test under a fixed alternative hypothesis $\theta_\star \neq \theta_0$—they decay to zero exponentially fast as $n \to \infty$.*

References

[1] J. D. Abernethy, E. Hazan, and A. Rakhlin. An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory*, 2008.

[2] Anonymous Author(s). Title. In *Conference*, 2022.

[3] J. Ash, S. Goel, A. Krishnamurthy, and S. Kakade. Gone fishing: Neural active learning with Fisher embeddings. In *NeurIPS*, 2021.

[4] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 2010.

[5] F. Bach. *Learning Theory from First Principles*. Online version, 2021.

[6] Y. Baraud. Confidence balls in Gaussian regression. *The Annals of Statistics*, 32(2):528–551, 2004.

[7] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[8] R. Beran. Confidence sets centered at C_p-estimators. *Annals of the Institute of Statistical Mathematics*, 48(1):1–15, 1996.

[9] R. Beran and L. Dumbgen. Modulation of estimators and confidence sets. *Annals of Statistics*, pages 1826–1856, 1998.

[10] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] S. Bubeck and R. Eldan. The entropic barrier: Exponential families, log-concave geometry, and self-concordance. *Mathematics of Operations Research*, 44(1), 2019.

[12] S. Bubeck and Y. T. Lee. Black-box optimization with a politician. In *International Conference on Machine Learning*, pages 1624–1631. PMLR, 2016.

[13] A. P. Dawid, M. Musio, and L. Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1), 2016.

[14] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *ICML*. PMLR, 2020.

[15] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, 2012.

[16] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[17] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[18] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1 edition, 1981.

[19] F. Kunstner, P. Hennig, and L. Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *NeurIPS*, 2019.

[20] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5), 2000.

[21] L. M. Le Cam. *Locally Asymptotically Normal Families of Distributions: Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Press, 1960.

[22] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[23] D. M. Ostrovskii and F. Bach. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1), 2021.

[24] J. Pennington and P. Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In *Advances in neural information processing systems*, 2018.

[25] A. Soen and K. Sun. On the variance of the Fisher information for deep learning. In *NeurIPS*, 2021.

[26] V. Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6), 2012.

[27] V. Spokoiny. Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1), 2017.

[28] T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. *Mathematical Programming*, 178(1), 2019.

[29] L. Tunçel and A. Nemirovski. Self-concordant barriers for convex approximations of structured convex sets. *Foundations of Computational Mathematics*, 10(5):485–525, 2010.

[30] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge university press, 2009.

[31] A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

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

[34] J. Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, 2013.

[35] S. Yang and M. Mohri. Optimistic bandit convex optimization. In *NIPS*, 2016.

[36] Y. Zhang and X. Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In *ICML*. PMLR, 2015.

[37] W. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of statistics*, 46(5), 2018.

Appendix

Table of Contents

A Assumptions **8**

B Discussion **8**

C Proof of main results **10**

C.1 Localization 10

C.2 Proof of the main theorems 12

D Examples and applications **14**

D.1 Examples 14

D.2 Goodness-of-fit testing 15

E Technical tools **16**

E.1 Properties of generalized self-concordant functions 16

E.2 Concentration of random vectors and matrices 18

A Assumptions

Our key assumption is the generalized self-concordance of the loss function.

Assumption 1 (Generalized self-concordance)**.** For any $z \in \mathcal{Z}$, the scoring rule $\ell(\cdot; z)$ is (R, ν)-generalized self-concordant for some $R > 0$ and $\nu \geq 2$. Moreover, $L(\cdot)$ is also (R, ν)-generalized self-concordant.

Remark 8. *When $\nu = 2$, it is straightforward to check that the generalized self-concordance of $\ell(\cdot; z)$ implies the one of $L(\cdot)$.*

In order to control the empirical gradient $S_n(\theta)$, we assume that the norm of the normalized gradient at $\theta_\star$ has a light tail.

Assumption 2 (Sub-Gaussian gradient)**.** There exists a constant $K_1 > 0$ such that the normalized gradient at θ_* is sub-Gaussian with parameter K_1, i.e., $\left\|G(\theta_\star)^{-1/2} S(\theta_\star; Z)\right\|_{\psi_2} \leq K_1$. Here $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm whose definition is recalled in Appx. E.

Remark 9. *When the loss function is of the form $\ell(\theta; z) = \ell(y, \theta^\top x)$, we have $S(\theta; Z) = \ell'(Y, \theta^\top X) X$ where $\ell'(y, \bar{y}) = \mathrm{d}\ell(y, \bar{y})/\mathrm{d}\bar{y}$. As a result, Asmp. 2 holds true if (i) $\ell'(Y, \theta_\star^\top X)$ is sub-Gaussian and X is bounded or (ii) $\ell'(Y, \theta_\star^\top X)$ is bounded and X is sub-Gaussian. For least squares with $\ell(y, \theta^\top x) = \frac{1}{2}(y - \theta^\top x)^2$, the derivative $\ell'(Y, \theta_\star^\top X) = \theta_\star^\top X - Y$ is the negative residual. Asmp. 2 is guaranteed if the residual is sub-Gaussian and X is bounded. For logistic regression with $\ell(y, \theta^\top x) = -\log \sigma(y \cdot \theta^\top x)$ where $\sigma(u) = (1 + e^{-u})^{-1}$, the derivative $\ell'(Y, \theta_\star^\top X) = [\sigma(Y \cdot \theta_\star^\top X) - 1] Y \in [-1, 1]$ is bounded. Thus, Asmp. 2 is guaranteed if X is sub-Gaussian.*

In order to control the empirical Hessian, we assume that the Hessian of the loss function at a neighborhood of $\theta_\star$ satisfies the matrix Bernstein condition.

Assumption 3 (Matrix Bernstein of Hessian)**.** There exist constants $K_2, r > 0$ such that, for any $\theta \in \Theta_r(\theta_\star)$, the standardized Hessian

$$H(\theta)^{-1/2} H(\theta; Z) H(\theta)^{-1/2} - I_d$$

satisfies a Bernstein condition (defined in Appx. E) with parameter K_2. Moreover,

$$\sigma_H^2 := \sup_{\theta \in \Theta_r(\theta_\star)} \left| \mathbb{V}\mathrm{ar}\left(H(\theta)^{-1/2} H(\theta; Z) H(\theta)^{-1/2} \right) \right|_2 < \infty,$$

where, for a matrix $J \in \mathbb{R}^{d \times d}$, we define $|J|_2 := \max\{\lambda_{\max}(J), |\lambda_{\min}(J)|\}$ and $\mathbb{V}\mathrm{ar}(J) := \mathbb{E}[JJ^\top] - \mathbb{E}[J]\,\mathbb{E}[J]^\top$. By convention, we let $\Theta_0(\theta_\star) = \{\theta_\star\}$.

B Discussion

Fisher information and model misspecification. When the model is well-specified, the covariance matrix $G(\theta)$ coincides with the well-known Fisher information $\mathcal{I}(\theta) := \mathbb{E}_{Z \sim P_\theta}[S(\theta; Z)]$ at $\theta_\star$. The Fisher information plays a central role in mathematical statistics and, in particular, M-estimation; see [24, 19, 3, 25] for recent developments in this line of research. It quantifies the amount of information a random variable carries about the model parameter. Under a well-specified model, it also coincides with the Hessian matrix $H(\theta)$ at the optimum which captures the curvature of the population risk. When the model is misspecified, the Fisher information deviates from the Hessian matrix. In the asymptotic regime, this discrepancy is reflected in the limiting covariance of the weighted M-estimator which admits a sandwich form $H(\theta_\star)^{-1} G(\theta_\star) H(\theta_\star)^{-1}$; see, e.g., [16, Sec. 4].

Effective dimension. The counterpart of the sandwitch covariance in the non-asymptotic regime is the effective dimension $d_\star$; see, e.g., [27, 23]. Our bounds also enjoy the same merit—its dimension dependency is via the effective dimension. When the model is well-specified, the effective dimension reduces to d, recovering the same rate of convergence $O(d/n)$ as in classical linear regression; see, e.g., [5, Prop. 3.5]. When the model is misspecified, the effective dimension provides a characterization of the problem complexity which is adapted to both the data distribution and the loss function via the matrix $H(\theta_\star)^{-1/2} G(\theta_\star) H(\theta_\star)^{-1/2}$. To gain a better understanding on the effective dimension $d_\star$, we summarize it in Tab. 1 under different regimes of eigendecay, assuming that $G(\theta_\star)$

Table 1: Comparison between the effective dimension $d_\star$ and the parameter dimension d in different regimes of eigendecays of $G(\theta_\star)$ and $H(\theta_\star)$ assuming they share the same eigenvectors.

	Eigendecay		**Dimension Dependency**		**Ratio**
	$G(\theta_\star)$	$H(\theta_\star)$	$d_\star$	d	$d_\star/d$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$d^{(\beta-\alpha+1)\vee 0}$	d	$d^{(\beta-\alpha)\vee(-1)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$d^{1-\alpha}e^{\nu d}$	d	$d^{-\alpha}e^{\nu d}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	1	d	d^{-1}
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	d if $\mu = \nu$ 1 if $\mu > \nu$ $e^{(\nu-\mu)d}$ if $\mu < \nu$	d	1 if $\mu = \nu$ d^{-1} if $\mu > \nu$ $d^{-1}e^{(\nu-\mu)d}$ if $\mu < \nu$

and $H(\theta_\star)$ share the same eigenvectors. It is clear that, when the spectrum of $G(\theta_\star)$ decays faster than the one of $H(\theta_\star)$, the dimension dependency can be better than $O(d)$. In fact, it can be as good as $O(1)$ when the spectrum of $G(\theta_\star)$ and $H(\theta_\star)$ decay as $e^{\mu i}$ and $i^{-\beta}$, respectively.

Comparison to classical asymptotic theory. Classical asymptotic theory of M-estimation is usually based on two assumptions: (a) the model is well-specified and (b) the sample size n is much larger than the parameter dimension d. These assumptions prevent it from being applicable to many real applications where the parametric family is only an approximation to the unknown data distribution and the data is of high dimension involving a large amount of parameters. On the contrary, our results do not require a well-specified model and the dimension dependency is replaced by the effective dimension $d_\star$ which captures the complexity of the parameter space. Moreover, they are of non-asymptotic nature—they hold true for any n as long as it exceeds some constant factor of $d_\star$. This allows the number of parameters to potentially grow with the same size.

Comparison to recent non-asymptotic theory. Recently, Spokoiny [26] achieved a breakthrough on finite-sample analysis of parametric M-estimation. Although being fully general, their results require strong global assumptions on the deviation of empirical risk process and are built upon advanced tools from empirical process theory. Restricting ourselves to generalized self-concordant losses, we are able to provide a more transparent analysis with neat assumptions only at the optimum parameter $\theta_\star$. Moreover, our results maintain some generality, covering several interesting examples in statistical machine learning as provided in Appx. D.1.

Ostrovskii and Bach [23] also considered self-concordant losses for M-estimation. However, their results are limited to generalized linear models whose loss is (pseudo) self-concordant and admits the form $\ell(\theta; Z) := \ell(Y, \theta^\top X)$. While sharing the same rate $O(d_\star/n)$, our results are more general than theirs in two aspects. First, the loss need not be of the form $\ell(Y, \theta^\top X)$, encompassing the score matching loss in Ex. 5 below. Second, the notion of generalized self-concordance encompasses both the standard and pseudo self-concordance, allowing us to obtain a unified analysis rather than separate ones as in [23].

Pseudo self-concordant losses have also been considered for semi-parametric models [2]. They focus on bounding the excess risk rather than providing confidence sets. Moreover, their results require a localization assumption on θ_n while this is proved in Prop. 11 in this paper.

Regularization. Our results can also be applied to regularized empirical risk minimization by including the regularization term in the loss function. Let θ_n^λ and $\theta_\star^\lambda$ be the minimizer of the *regularized* empirical and population risk, respectively. Let $d_\star^\lambda := \mathbf{Tr}\left((H_\star^\lambda)^{-1/2} G_\star^\lambda (H_\star^\lambda)^{-1/2}\right)$ where $H_\star^\lambda$ and $G_\star^\lambda$ are the regularized Hessian and the covariance of the regularized gradient at $\theta_\star^\lambda$, respectively. Then our results characterize the concentration of θ_n^λ around $\theta_\star^\lambda$:

$$\left\|\theta_n^\lambda - \theta_\star^\lambda\right\|_{H_\star^\lambda}^2 \le O(d_\star^\lambda/n).$$

This result coincides with Spokoiny [27, Thm. 2.1]. If the goal is to estimate the unregularized population risk minimizer $\theta_\star$, then we need to pay an additional error $\left\|\theta_\star^\lambda - \theta_\star\right\|_{H_\star^\lambda}^2$ which is referred to as the modeling bias [27, Sec. 2.5].

For instance, let $Z := (X, Y)$ where $X \in \mathbb{R}^d$ with $\mathbb{E}[XX^\top] = I_d$ and $Y \in \mathbb{R}$. Consider the regularized squared loss

$$\ell^\lambda(\theta; z) := \frac{1}{2}(y - \theta^\top x)^2 + \frac{1}{2}\theta^\top \Lambda \theta,$$

where $\Lambda = \mathbf{diag}\{\lambda_1, \ldots, \lambda_d\}$. The regularized effective dimension is of order [27, Sec. 2.1]

$$O\left(\sum_{k=1}^{d} \frac{1}{1+\lambda_i}\right)$$

which can be much smaller than d if $\{\lambda_k\}$ is increasing.

C Proof of main results

Our proof techniques rely on a self-concordance property to localize the estimator and control the Hessian and related quantities. This property was, up to our knowledge, first put to use in machine learning by Abernethy et al. [1] in the context of sequential allocation of experiments and multi-armed bandits. The key observation is that, within the Dikin ellipsoid, the variation of the Hessian can be easily controlled. More recently, Ostrovskii and Bach [23] obtained risk bounds for generalized linear models based on this observation. Our results and proof techniques also rely on this observation. We show how to leverage this observation to obtain confidence sets for a broad class of statistical models under a generalized self-concordance assumption owing to the use of the matrix Bernstein inequality. For instance, we obtain confidence bounds for parameter estimation using score matching and generalized linear statistical models under possible model misspecification as provided in Appx. D.1.

Our proofs are inspired by Ostrovskii and Bach [23]. However, there are two key differences. First, since they focus on loss functions of the form $\ell(Y, \theta^\top X)$, the Hessian is $\ell''(Y, \theta^\top X)XX^\top$ where $\ell''(y, \bar{y}) := \mathrm{d}^2\ell(y, \bar{y})/\mathrm{d}\bar{y}^2$. As a result, they can control the deviation of the empirical Hessian using inequalities for sample second-moment matrices of sub-Gaussian random vectors [23, Thm. A.2]. In contrast, we use matrix Bernstein inequality which allows us to work with a larger class of loss functions. Second, we extend their localization result from pseudo self-concordant losses to generalized self-concordant losses (Prop. 11). This is enabled by a new property on the existence of a unique minimizer for generalized self-concordant functions (Prop. 21).

In the remaining of this section, we first prove a localization result Prop. 11 and the score bound Prop. 3 in Appx. C.1. It not only guarantees the existence and uniqueness of θ_n but also localizes it. We then, in Appx. C.2, control the empirical Hessian at θ_n using a covering number argument. Finally, we prove Thm. 4.

C.1 Localization

We start by showing that the empirical risk ℓ_n is generalized self-concordant.

Lemma 10. *Under Asmp. 1, the empirical risk ℓ_n is $(n^{\nu/2-1}R, \nu)$-generalized self-concordant.*

Proof. By Asmp. 1, the loss $\ell(\cdot; Z_i)$ is (R, ν)-generalized self-concordant for every $i \in [n] := \{1, \ldots, n\}$. Note that ℓ_n is the empirical average of $\{\ell(\cdot; Z_i)\}_{i=1}^n$. Hence, it follows from [28, Prop. 1] that ℓ_n is $(n^{\nu/2-1}R, \nu)$-generalized self-concordant □

Applying Prop. 21 to ℓ_n leads to the localization result. Let $\lambda_{n,\star} := \lambda_{\min}(H_n(\theta_\star))$ and $\lambda_n^\star := \lambda_{\min}(H_n(\theta_\star))$. Recall K_ν from Cor. 20. Define

$$R_{n,\nu}^\star := \begin{cases} \lambda_{n,\star}^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_{n,\star}^{(\nu-3)/2} n^{\nu/2-1} R & \text{if } \nu \in (2, 3] \\ (\nu/2-1)(\lambda_n^\star)^{(\nu-3)/2} n^{\nu/2-1} R & \text{if } \nu > 3. \end{cases} \tag{6}$$

Proposition 11. *Under Asmp. 1, whenever $R_{n,\nu}^\star \left\|S_n(\theta_\star)\right\|_{H_n^{-1}(\theta_\star)} \leq K_\nu$, the estimator θ_n uniquely exists and satisfies*

$$\left\|\theta_n - \theta_\star\right\|_{H_n(\theta_\star)} \leq 4 \left\|S_n(\theta_\star)\right\|_{H_n^{-1}(\theta_\star)}.$$

Proof. The claim follows directly from Lem. 10 and Prop. 21. □

Prop. 11 implies that the empirical risk minimizer uniquely exists if $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$ is small. Hence, it remains to bound $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$, which can be achieved by controlling $\|S_n(\theta_\star)\|_{H^{-1}(\theta_\star)}$ and $H_n(\theta_\star)$. Recall $d_\star$ from (1).

Lemma 12. *Under Asmp. 2, it holds that, with probability at least* $1-\delta$*,*

$$\|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d_\star.$$

Proof. By the first order optimality condition, we have $S(\theta_\star) = 0$. As a result,

$$X := \sqrt{n} G^{-1/2}(\theta_\star) S_n(\theta_\star; Z)$$

is an isotropic random vector. Moreover, it follows from Lem. 26 that $\|X\|_{\psi_2} \lesssim K_1$. Define $J := G^{1/2}(\theta_\star) H^{-1}(\theta_\star) G^{1/2}(\theta_\star)/n$. Then we have

$$\|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} = \|X\|_J^2 .$$

Invoking Thm. 27 yields the claim. □

Lemma 13. *Under Asmp. 3 with* $r = 0$*, it holds that, with probability at least* $1-\delta$*,*

$$\frac{1}{2} H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2} H(\theta_\star)$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2) \log(2d/\delta)$.

Proof. Due to Asmp. 3, the standardized Hessian at $\theta_\star$

$$H(\theta_\star)^{-1/2} H(\theta_\star; Z) H(\theta_\star)^{-1/2} - I_d$$

satisfies a Bernstein condition with parameter K_2. It then follows from Thm. 28 that

$$\mathbb{P}\left(\left| H(\theta_\star)^{-1/2} H_n(\theta_\star) H(\theta_\star)^{-1/2} - I_d \right|_2 \geq \frac{1}{2} \right) \leq 2d \exp\left\{ -\frac{n}{4(2\sigma_H^2 + K_2)} \right\}.$$

As a result, whenever $n \geq 4(K_2^2 + 2\sigma_H^2) \log(2d/\delta)$, we have

$$\frac{1}{2} I_d \preceq H(\theta_\star)^{-1/2} H_n(\theta_\star) H(\theta_\star)^{-1/2} \preceq \frac{3}{2} I_d,$$

or equivalently

$$\frac{1}{2} H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2} H(\theta_\star).$$

□

Proof of Prop. 3. Define two events

$$\mathcal{A} := \left\{ \|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} \lesssim \frac{1}{n} K_1^2 \log(2e/\delta) d_\star \right\} \text{ and } \mathcal{B} := \left\{ \frac{1}{2} H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2} H(\theta_\star) \right\}.$$

According to Lems. 12 and 13, whenever $n \geq 4(K_2^2 + 2\sigma_H^2) \log(4d/\delta)$, we have $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/2$ and $\mathbb{P}(\mathcal{B}) \geq 1 - \delta/2$. On the event $\mathcal{AB}$, we have

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \leq 2 \|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} \lesssim \frac{2}{n} K_1^2 \log(2e/\delta) d_\star \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d_\star.$$

Since $\mathbb{P}(\mathcal{AB}) \geq 1 - \mathbb{P}(\mathcal{A}^c) - \mathbb{P}(\mathcal{B}^c) \geq 1 - \delta$, we have, with probability at least $1-\delta$,

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d_\star$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2) \log(4d/\delta)$. □

C.2 Proof of the main theorems

Before we prove the main theorem, we control the empirical Hessian. Recall that $\lambda_\star := \lambda_{\min}(H(\theta_\star))$ and $\lambda^\star := \lambda_{\max}(H(\theta_\star))$. Let

$$R_\nu^\star := \begin{cases} \lambda_\star^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_\star^{(\nu-3)/2} R & \text{if } \nu \in (2,3] \\ (\nu/2-1)(\lambda^\star)^{(\nu-3)/2} R & \text{if } \nu > 3. \end{cases} \tag{7}$$

Proposition 14. *Fix $\varepsilon \in (0, K_\nu]$. Under Asmps. 1 and 3 with $r = K_\nu / R_\nu^\star$, it holds that, with probability at least $1-\delta$,*

$$\frac{1}{2\omega_\nu^2(\varepsilon)} H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2}\omega_\nu^2(\varepsilon) H(\theta_\star), \quad \textit{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star),$$

whenever whenever $n \geq 4(K_2^2 + 2\sigma_H^2)\left\{\log(2d/\delta) + d\log\left[3(1.5\omega_\nu(\varepsilon)n)^{(\nu/2-1)}\right]\right\}$.

Proof. We prove the result in the following steps.

Step 1. Take a τ-covering and relate $H_n(\theta)$ to $H_n(\bar\theta)$ for some $\bar\theta$ in the covering. Let $\tau := \varepsilon / R_\nu^\star [1.5\omega_\nu(\varepsilon)n]^{\nu/2-1}$. Take an τ-covering $\mathcal{N}_\tau$ of $\Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$ w.r.t. $\|\cdot\|_{H(\theta_\star)}$, and let $\pi(\theta)$ be the projection of θ onto $\mathcal{N}_\tau$. Let

$$d_{n,\nu}(\theta_1, \theta_2) := \begin{cases} n^{\nu/2-1} R \|\theta_2 - \theta_1\|_2 & \text{if } \nu = 2 \\ (\nu/2-1) n^{(\nu/2-1)} R \|\theta_2 - \theta_1\|_2^{3-\nu} \|\theta_2 - \theta_1\|_{H_n(\theta_1)}^{\nu-2} & \text{otherwise.} \end{cases}$$

By Lem. 10 and Prop. 16, we have, for all $\theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$,

$$\frac{1}{\omega_\nu(d_{n,\nu}(\pi(\theta), \theta))} H_n(\pi(\theta)) \preceq H_n(\theta) \preceq \omega_\nu(d_{n,\nu}(\pi(\theta), \theta)) H_n(\pi(\theta)), \tag{8}$$

where it holds if $d_{n,\nu}(\pi(\theta), \theta) < 1$ for the case $\nu > 2$.

Step 2. Related $H_n(\theta)$ to $H(\theta_\star)$ for all θ in the covering. Fix an arbitrary $\theta \in \mathcal{N}_\tau$. Following the same argument as Lem. 13, we have, with probability at least $1-\delta$,

$$\frac{1}{2} H(\theta) \preceq H_n(\theta) \preceq \frac{3}{2} H(\theta)$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2)\log(2d/\delta)$. It then follows from Asmp. 1 and Lem. 18 that

$$\frac{1}{\omega_\nu(R_\nu^\star \|\theta - \theta_\star\|_{H(\theta_\star)})} H(\theta_\star) \preceq H(\theta) \preceq \omega_\nu(R_\nu^\star \|\theta - \theta_\star\|_{H(\theta_\star)}) H(\theta_\star), \tag{9}$$

since $R_\nu^\star \|\theta - \theta_\star\|_{H(\theta_\star)} \leq \varepsilon < 1$. By the monotonicity of ω_ν, we get

$$\frac{1}{\omega_\nu(\varepsilon)} H(\theta_\star) \preceq H(\theta) \preceq \omega_\nu(\varepsilon) H(\theta_\star),$$

and thus, with probability at least $1-\delta$,

$$\frac{1}{2\omega_\nu(\varepsilon)} H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2}\omega_\nu(\varepsilon) H(\theta_\star)$$

whenever $n \geq 4(K_2^2 + 2\sigma_H^2)\log(2d/\delta)$. Define the event

$$\mathcal{A} := \left\{ \frac{1}{2\omega_\nu(\varepsilon)} H(\theta_\star) \preceq H_n(\pi(\theta)) \preceq \frac{3}{2}\omega_\nu(\varepsilon) H(\theta_\star), \quad \text{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star) \right\}.$$

Since $|\mathcal{N}_\tau| \leq (3\varepsilon/\tau R_\nu^\star)^d$ [23], by a union bound, we have $\mathbb{P}(\mathcal{A}) \geq 1-\delta$ whenever

$$n \geq 4(K_2^2 + 2\sigma_H^2)\left\{\log(2d/\delta) + d\log\left[3(1.5\omega_\nu(\varepsilon)n)^{(\nu/2-1)}\right]\right\}.$$

Step 3. Combine the previous two steps. On the event $\mathcal{A}$, we have $H_n(\pi(\theta)) \preceq \frac{3}{2}\omega_\nu(\varepsilon)H(\theta_\star)$ for all $\theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$. A similar argument as Lem. 18 shows that

$$d_{n,\nu}(\pi(\theta),\theta) \le \begin{cases} \lambda_\star^{-1/2} R\tau & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_\star^{(\nu-3)/2}[3\omega_\nu(\varepsilon)/2]^{(\nu-2)/2} n^{\nu/2-1} R\tau & \text{if } \nu \in (2,3] \\ (\nu/2-1)(\lambda^\star)^{(\nu-3)/2}[3\omega_\nu(\varepsilon)/2]^{(\nu-2)/2} n^{\nu/2-1} R\tau & \text{otherwise.} \end{cases}$$

Substituting τ gives $d_{n,\nu}(\pi(\theta),\theta) \le R_\nu^\star \frac{\varepsilon}{R_\nu^\star} = \varepsilon$. Hence, by (8), we obtain

$$\frac{1}{2\omega_\nu^2(\varepsilon)} H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2}\omega_\nu^2(\varepsilon) H(\theta_\star), \quad \text{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star).$$

on the event $\mathcal{A}$. □

Theorem 15. *Let $\nu \in [2,3)$ and $\varepsilon \in (0, K_\nu]$. Under Asmps. 1 to 3 with $r = 0$, whenever*[1]

$$n \gtrsim \max\left\{ (K_2^2 + \sigma_H^2)\log(2d/\delta), \left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log(e/\delta)}{\varepsilon^2}\right]^{1/(3-\nu)} \right\},$$

the empirical risk minimizer θ_n uniquely exists and satisfies, with probability at least $1-\delta$,

$$\|\theta_n - \theta_\star\|_{H(\theta_\star)}^2 \le C K_1^2 \omega_\nu^2(\varepsilon) \log(e/\delta) \frac{d_\star}{n}. \tag{10}$$

Here C is an absolute constant.

Proof. Similar to the proof of Prop. 3, we define two events

$$\mathcal{A} := \left\{ \|S_n(\theta_\star)\|_{H^{-1}(\theta_\star)}^2 \lesssim \frac{1}{n} K_1^2 \log(2e/\delta) d_\star \right\} \text{ and } \mathcal{B} := \left\{ \frac{1}{2} H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2} H(\theta_\star) \right\}.$$

In the following, we let

$$n \gtrsim \max\left\{ 4(K_2^2 + 2\sigma_H^2)\log(4d/\delta), \left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log(e/\delta)}{\varepsilon^2}\right]^{1/(3-\nu)} \right\}.$$

Following the same argument as Prop. 3, we have $\mathbb{P}(\mathcal{AB}) \ge 1-\delta$ and

$$\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}^2 \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d_\star.$$

Now, it suffices to prove, on the event $\mathcal{AB}$,

$$\|\theta_n - \theta_\star\|_{H(\theta_\star)}^2 \lesssim K_1^2 \log(e/\delta) \frac{d_\star}{n}.$$

Recall $R_{n,\nu}^\star$ and $R_\nu^\star$ from (6) and (7). It is straightforward to check that $R_{n,\nu}^\star \le \sqrt{2} n^{\nu/2-1} R_\nu^\star$ for all $\nu \in [2,3]$. Consequently, it holds that

$$R_{n,\nu}^\star \|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} \lesssim R_\nu^\star n^{(\nu-3)/2} \sqrt{K_1^2 \log(e/\delta) d_\star} \le \varepsilon \le K_\nu$$

since $n^{3-\nu} \gtrsim (R_\nu^\star)^2 K_1^2 \log(e/\delta) d_\star / \varepsilon^2$. As a result, by Prop. 11, we have that θ_n uniquely exists and satisfies

$$\|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \le 4 \|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)},$$

and thus, using the event $\mathcal{B}$,

$$\|\theta_n - \theta_\star\|_{H(\theta_\star)}^2 \lesssim \|\theta_n - \theta_\star\|_{H_n(\theta_\star)}^2 \lesssim K_1^2 \log(e/\delta) \frac{d_\star}{n}.$$

□

[1] Here $\gtrsim$ hides an absolute constant.

Proof of Thm. 4. We start by defining some events. Let

$$\begin{aligned}
\mathcal{A} &:= \left\{ \|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} \lesssim \frac{1}{n} K_1^2 \log(3e/\delta) d_\star \right\} \\
\mathcal{B} &:= \left\{ \frac{1}{2} H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2} H(\theta_\star) \right\} \\
\mathcal{C} &:= \left\{ \frac{1}{2\omega_\nu^2(\varepsilon)} H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2} \omega_\nu^2(\varepsilon) H(\theta_\star), \quad \text{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star) \right\}.
\end{aligned} \tag{11}$$

In the following, we let

$$n^{3-\nu} \gtrsim \max\left\{ \left\{ 4(K_2^2 + 2\sigma_H^2) \left[\log \frac{6d}{\delta} + d \log \left[3(1.5\omega_\nu(K_\nu)n)^{\frac{\nu-2}{2}} \right] \right] \right\}^{3-\nu}, \frac{(R_\nu^\star)^2 K_1^2 d_\star \log \frac{e}{\delta}}{\varepsilon^2} \right\}.$$

According to Lem. 12, Lem. 13, and Prop. 14, it holds that $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/3$, $\mathbb{P}(\mathcal{B}) \geq 1 - \delta/3$, and $\mathbb{P}(\mathcal{C}) \geq 1 - \delta/3$. This implies that $\mathbb{P}(\mathcal{ABC}) \geq 1 - \delta$. Now, it suffices to prove, on the event $\mathcal{ABC}$,

$$\|\theta_n - \theta_\star\|^2_{H_n(\theta_n)} \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log(e/\delta) \frac{d_\star}{n}.$$

Following the same argument as Thm. 15, we obtain

$$\|\theta_n - \theta_\star\|_{H(\theta_\star)} \lesssim \|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \lesssim n^{-1/2} \sqrt{K_1^2 \log(e/\delta) d_\star} \leq \varepsilon / R_\nu^\star.$$

Therefore, using the event $\mathcal{C}$, we have

$$\|\theta_n - \theta_\star\|^2_{H_n(\theta_n)} \lesssim \omega_\nu^2(\varepsilon) \|\theta_n - \theta_\star\|^2_{H(\theta_\star)} \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log(e/\delta) \frac{d_\star}{n}.$$

As a result, $\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \geq 1 - \delta$ whenever n satisfies (3). □

D Examples and applications

We give several examples from statistical machine learning and prove the results for goodness-of-fit testing in Sec. 3.4.

D.1 Examples

We begin with several standard examples in the literature of M-estimation.

Example 3 (Linear regression)**.** Let $Z := (X, Y)$ be a pair of input and output, where $X \in \mathbb{R}^d$ with $\mathbb{E}[XX^\top] \succ 0$ and $Y \in \mathbb{R}$. Consider the statistical model

$$Y - \theta^\top X \sim \mathcal{N}(0, \sigma^2).$$

Its log-likelihood is given by $-(Y - \theta^\top X)^2/(2\sigma^2) + C\sigma$ leading to the squared loss

$$\ell(\theta; z) := \frac{1}{2}(y - \theta^\top x)^2.$$

It is clear that $\ell(\cdot; z)$ is convex. Moreover, since $D_\theta^3 \ell(\theta; z)[u, u, v] = 0$ for all $u, v \in \mathbb{R}^d$, the loss ℓ is generalized self-concordant for all $\nu \geq 2$ and $R \geq 0$.

Example 4 (Logistic regression)**.** Let $Z := (X, Y)$ be a pair of input and label, where $X \in \mathbb{R}^d$ with $\|X\| \leq M$ and $Y \in \{-1, 1\}$. Consider the statistical model

$$Y \mid X \sim \text{Bernoulli}\left(\sigma(\theta^\top X)\right),$$

where $\sigma(u) := (1 + e^{-u})^{-1}$. Its log-likelihood is given by $\mathbb{1}\{Y = 1\} \log \sigma(\theta^\top X) + \mathbb{1}\{Y = -1\} \log \sigma(-\theta^\top X)$ leading the loss

$$\ell(\theta; z) := -\log\left(\sigma(Y \cdot \theta^\top X)\right).$$

Take any $u, v \in \mathbb{R}^d$. It can be shown that

$$\left|D_\theta^3 \ell(\theta; z)[u, u, v]\right| \leq \left|1 - 2\sigma(\theta^\top x)\right| \left|v^\top x\right| D_\theta^2 \ell(\theta; z)[u, u].$$

Note that $|1 - 2\sigma(u)| \leq 1$ for all $u \in \mathbb{R}$. It then follows from the Cauchy-Schwarz that ℓ is $(M, 2)$-generalized self-concordant.

Derivation of Ex. 4. Recall that the loss is given by

$$\ell(\theta; z) := -\log\big(\sigma(Y \cdot \theta^\top X)\big).$$

Take any $u, v \in \mathbb{R}^d$. It can be computed that

$$\begin{aligned} D_\theta^2 \ell(\theta; z)[u, u] &= \sigma(y \cdot \theta^\top x)[1 - \sigma(y \cdot \theta^\top x)](u^\top x)^2 \\ D_\theta^3 \ell(\theta; z)[u, u, v] &= \sigma(y \cdot \theta^\top x)[1 - \sigma(y \cdot \theta^\top x)][1 - 2\sigma(y \cdot \theta^\top x)]y(u^\top x)^2(v^\top x). \end{aligned}$$

Note that $\big|[1 - 2\sigma(y \cdot \theta^\top x)]y\big| \le 1$. It then follows that

$$\big|D_\theta^3 \ell(\theta; z)[u, u, v]\big| \le \big\|v^\top x\big\| \, D_\theta^2 \ell(\theta; z)[u, u] \le M \, \|v\| \, \|u\|^2_{\nabla^2 \ell(\theta; z)} \,,$$

and thus ℓ is $(M, 2)$-generalized self-concordant. □

We then give a popular loss function from score matching whose statistical behavior is less well-known.

Example 5 (Score matching with exponential families)**.** Assume that $\mathbb{Z} = \mathbb{R}^p$. Consider an exponential family on $\mathbb{R}^d$ with densities

$$\log p_\theta(z) = \theta^\top t(z) + h(z) - \Lambda(\theta).$$

The non-normalized density q_θ then reads $\log q_\theta(z) = \theta^\top t(z) + h(z)$. As a result, the score matching loss becomes

$$\ell(\theta; z) = \frac{1}{2}\theta^\top A(z)\theta - b(z)^\top \theta + c(z) + \text{const},$$

where $A(z) := \sum_{k=1}^p \frac{\partial t(z)}{\partial z_k}\big(\frac{\partial t(z)}{\partial z_k}\big)^\top$ is p.s.d, $b(z) := \sum_{k=1}^p \Big[\frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial h(z)}{\partial z_k}\frac{\partial t(z)}{\partial z_k}\Big]$, and $c(z) := \sum_{k=1}^p \Big[\frac{\partial^2 h(z)}{\partial z_k^2} + \big(\frac{\partial h(z)}{\partial z_k}\big)^2\Big]$. Therefore, the score matching loss $\ell(\theta; z)$ is convex. Moreover, since the third derivatives of $\ell(\cdot; z)$ is zero, the score matching loss is generalized self-concordant for all $\nu \ge 2$ and $R \ge 0$.

D.2 Goodness-of-fit testing

Proof of Thm. 6. (a) Under $\mathbf{H}_0$, we have $\theta_\star = \theta_0$. It then follows from Prop. 3 that, with probability at least $1 - \delta$,

$$T_{\text{Rao}} := \|S_n(\theta_0)\|^2_{H_n^{-1}(\theta_0)} \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d$$

whenever $n \ge 4(K_2^2 + 2\sigma_H^2)\log(4d/\delta)$.

(b) Define three events

$$\begin{aligned} \mathcal{A} &:= \{T_{\text{Rao}} > t\} \\ \mathcal{B} &:= \left\{\frac{1}{2}H(\theta_0) \preceq H_n(\theta_0) \preceq \frac{3}{2}H(\theta_0)\right\} \\ \mathcal{C} &:= \left\{-4\,\|S(\theta_0)\|_{H^{-1}(\theta_0)}\,\|S_n(\theta_0) - S(\theta_0)\|_{H^{-1}(\theta_0)} + \frac{2}{3}\,\|S(\theta_0)\|^2_{H^{-1}(\theta_0)} > t_n\right\}. \end{aligned}$$

On the event $\mathcal{B}$, it holds that

$$\begin{aligned} T_{\text{Rao}} &\ge 2S(\theta_0)^\top H_n^{-1}(\theta_0)[S_n(\theta_0) - S(\theta_0)] + S(\theta_0)^\top H_n^{-1}(\theta_0)S(\theta_0) \\ &\ge -2\left\|H(\theta_0)^{1/2}H_n^{-1}(\theta_0)H(\theta_0)^{1/2}\right\|_2 \|S(\theta_0)\|_{H^{-1}(\theta_0)}\,\|S_n(\theta_0) - S(\theta_0)\|_{H^{-1}(\theta_0)} \\ &\quad + \frac{2}{3}\,\|S(\theta_0)\|^2_{H^{-1}(\theta_0)} \\ &\ge -4\,\|S(\theta_0)\|_{H^{-1}(\theta_0)}\,\|S_n(\theta_0) - S(\theta_0)\|_{H^{-1}(\theta_0)} + \frac{2}{3}\,\|S(\theta_0)\|^2_{H^{-1}(\theta_0)}\,. \end{aligned}$$

This implies $\mathcal{CB} \subset \mathcal{AB}$, and thus

$$\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{AB}) \geq \mathbb{P}(\mathcal{CB}) \geq 1 - \mathbb{P}(\mathcal{B}^c) - \mathbb{P}(\mathcal{C}^c).$$

Note that $\mathcal{C}^c = \{\|S_n(\theta_0) - S(\theta_0)\|^2_{H^{-1}(\theta_0)} - \mathbf{Tr}(\Omega(\theta_0))/n \geq \tau_n\}$ where

$$\tau_n = \left[\frac{2\|S(\theta_0)\|^2_{H^{-1}(\theta_0)}/3 - t}{4\|S(\theta_0)\|_{H^{-1}(\theta_0)}}\right]^2 - \frac{1}{n}\mathbf{Tr}(\Omega(\theta_0)) = \frac{\|S(\theta_0)\|^2_{H^{-1}(\theta_0)}}{36} + O(n^{-1}).$$

It follows from Thm. 27 that, whenever $\tau_n \geq 0$,

$$\mathbb{P}(\mathcal{C}^c) \leq \exp\left(-c\min\left\{\frac{n^2\tau_n^2}{K_1^2\|\Omega(\theta_0)\|_2^2}, \frac{n\tau_n}{K_1\|\Omega(\theta_0)\|_\infty}\right\}\right).$$

Moreover, due to Thm. 28, we have

$$\mathbb{P}(\mathcal{B}^c) \leq 2d\exp\left(-\frac{n}{4(K_2^2 + 2\sigma_H^2)}\right),$$

and this proves the claim. □

E Technical tools

In this section, we first recall and prove some key properties of generalized self-concordance. We then review some key results regarding the concentration of random vectors and matrices.

E.1 Properties of generalized self-concordant functions

Throughout this section, we let $f : \mathbb{R}^d \to \mathbb{R}$ be (R, ν)-generalized self-concordant as in Definition 1, where $R > 0$ and $\nu \geq 2$. For simplicity of the notation, we denote $\|\cdot\|_x := \|\cdot\|_{\nabla^2 f(x)}$. Let

$$d_\nu(x, y) := \begin{cases} R\|y - x\|_2 & \text{if } \nu = 2 \\ (\nu/2 - 1)R\|y - x\|_2^{3-\nu}\|y - x\|_x^{\nu-2} & \text{if } \nu > 2 \end{cases} \tag{12}$$

and

$$\omega_\nu(\tau) := \begin{cases} (1-\tau)^{-2/(\nu-2)} & \text{if } \nu > 2 \\ e^\tau & \text{if } \nu = 2 \end{cases} \tag{13}$$

with $\mathrm{dom}(\omega_\nu) = \mathbb{R}$ if $\nu = 2$ and $\mathrm{dom}(\omega_\nu) = (-\infty, 1)$ if $\nu > 2$.

The next proposition gives bounds for the Hessian of f.

Proposition 16 (Sun and Tran-Dinh [28], Prop. 8)**.** *For any $x, y \in \mathrm{dom}(f)$, we have*

$$\frac{1}{\omega_\nu(d_\nu(x, y))}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \omega_\nu(d_\nu(x, y))\nabla^2 f(x),$$

where it holds if $d_\nu(x, y) < 1$ for the case $\nu > 2$.

We then give the bounds for function values. Define two functions

$$\bar{\omega}_\nu(\tau) := \int_0^1 \omega_\nu(t\tau)\mathrm{d}t = \begin{cases} \tau^{-1}(e^\tau - 1) & \text{if } \nu = 2 \\ -\tau^{-1}\log(1-\tau) & \text{if } \nu = 4 \\ \frac{\nu-2}{\nu-4}\frac{1-(1-\tau)^{(\nu-4)/(\nu-2)}}{\tau} & \text{otherwise} \end{cases} \tag{14}$$

and

$$\bar{\bar{\omega}}_\nu(\tau) := \int_0^1 t\bar{\omega}_\nu(t\tau)\mathrm{d}t = \begin{cases} \tau^{-2}(e^\tau - \tau - 1) & \text{if } \nu = 2 \\ -\tau^{-2}[\tau + \log(1-\tau)] & \text{if } \nu = 3 \\ \tau^{-2}[(1-\tau)\log(1-\tau) + \tau] & \text{if } \nu = 4 \\ \frac{\nu-2}{\nu-4}\frac{1}{\tau}\left[\frac{\nu-2}{2(3-\nu)\tau}\left((1-\tau)^{2(3-\nu)/(2-\nu)} - 1\right) - 1\right] & \text{otherwise.} \end{cases} \tag{15}$$

Proposition 17 (Sun and Tran-Dinh [28], Prop. 10)**.** *For any $x, y \in \mathrm{dom}(f)$, we have*

$$\bar{\bar{\omega}}_\nu(-d_\nu(x,y))\,\|y-x\|_x^2 \le f(y) - f(x) - \langle \nabla f(x), y-x\rangle \le \bar{\bar{\omega}}_\nu(d_\nu(x,y))\,\|y-x\|_x^2\,,$$

where it holds if $d_\nu(x,y) < 1$ for the case $\nu > 2$.

In the following, we fix $x \in \mathrm{dom}(f)$ and assume $\nabla^2 f(x) \succ 0$. We denote $\lambda_{\min} := \lambda_{\min}(\nabla^2 f(x))$ and $\lambda_{\max} := \lambda_{\max}(\nabla^2 f(x))$. The next lemma bounds $d_\nu(x,y)$ with the local norm $\|y-x\|_x$. Let

$$R_\nu := \begin{cases} \lambda_{\min}^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_{\min}^{(\nu-3)/2} R & \text{if } \nu \in (2,3] \\ (\nu/2-1)\lambda_{\max}^{(\nu-3)/2} R & \text{if } \nu > 3. \end{cases} \tag{16}$$

Lemma 18. *For any $\nu \ge 2$ and $y \in \mathrm{dom}(f)$, we have*

$$d_\nu(x,y) \le R_\nu\,\|y-x\|_x\,. \tag{17}$$

Moreover, it holds that

$$\frac{1}{\omega_\nu(R_\nu\,\|y-x\|_x)}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \omega_\nu(R_\nu\,\|y-x\|_x)\nabla^2 f(x),$$

where it holds if $R_\nu\,\|y-x\|_x < 1$ for the case $\nu > 2$.

Proof. Recall the definition of d_ν in (12). If $\nu = 2$, then, by the Cauchy-Schwarz inequality,

$$d_\nu(x,y) = R\,\|y-x\|_2 \le \left\|[\nabla^2 f(x)]^{-1/2}\right\|_2 R\,\|y-x\|_x \le \lambda_{\min}^{-1/2} R\,\|y-x\|_x\,.$$

The case $\nu > 2$ can be proved similarly. □

We then prove some useful properties for the function $\bar{\bar{\omega}}$.

Lemma 19. *For any $\nu \ge 2$, the following statements hold true:*

(a) The function $\varphi(\tau) := \bar{\bar{\omega}}_\nu(-\tau)$ is strictly decreasing on $[0,\infty)$ with $\varphi(0) = 1/2$ and $\varphi(\tau) \ge 0$ for all $\tau \ge 0$.

(b) The function $\psi(\tau) := \bar{\bar{\omega}}_\nu(-\tau)\tau$ is strictly increasing on $[0,\infty)$ with $\psi(0) = 0$.

Proof. **(a)**. By definition, ω_ν is strictly increasing on $(-\infty, 1)$. As a result, for any $\tau \in (-\infty, 1)$,

$$\bar{\omega}_\nu'(\tau) = \int_0^1 t\omega_\nu'(t\tau)\mathrm{d}t > 0.$$

It then follows that, for any $\tau \ge 0$,

$$\varphi'(\tau) = -\bar{\bar{\omega}}_\nu'(-\tau) = -\int_0^1 t^2\bar{\omega}_\nu'(-t\tau)\mathrm{d}t < 0,$$

and thus φ is strictly decreasing on $[0,\infty)$. Note that $\omega_\nu(0) = 1$ and $\omega_\nu(\tau) > 0$ for all $\tau \in (-\infty, 1)$. It is straightforward to check that $\varphi(0) = 1/2$ and $\varphi(\tau) > 0$ for all $\tau \ge 0$.

(b) Due to (14), it is clear that $\tau \mapsto \tau\bar{\omega}_\nu(-\tau)$ is strictly increasing on $[0,\infty)$ and equals 0 at $\tau = 0$. Note that, for any $\tau \ge 0$,

$$\psi(\tau) = \int_0^1 t\tau\bar{\omega}_\nu(-t\tau)\mathrm{d}t = \frac{1}{\tau}\int_0^\tau t\bar{\omega}_\nu(-t)\mathrm{d}t.$$

We get

$$\psi'(\tau) = \frac{1}{\tau^2}\left[\tau^2\bar{\omega}_\nu(-\tau) - \int_0^\tau t\bar{\omega}_\nu(-t)\mathrm{d}t\right].$$

By the monotonicity of $\tau \mapsto \tau\bar{\omega}_\nu(-\tau)$, it follows that $\psi'(\tau) > 0$. □

Corollary 20. *Let $\tau \geq 0$. For any $\nu \geq 2$, there exists $K_\nu \in (0, 1/2]$ such that*

$$\bar{\bar{\omega}}_\nu(-\tau)\tau \leq K_\nu \Rightarrow \tau < 1 + \mathbb{1}\{\nu = 2\} \text{ and } \bar{\bar{\omega}}_\nu(-\tau) \geq 1/4.$$

In particular, $K_\nu = 1/2$ if $\nu = 2$ and $K_\nu = 1/4$ if $\nu = 3$.

Proof. The existence of K_ν follows directly from the strict monotonicity of φ and ψ shown in Lem. 19. For $\nu = 2$,

$$\bar{\bar{\omega}}_\nu(-\tau)\tau = \frac{e^{-\tau} + \tau - 1}{\tau} \leq 1/2 \Rightarrow \tau < 2.$$

As a result, we have $\bar{\bar{\omega}}_\nu(-\tau) \geq 1/4$. The case for $\nu = 3$ can be proved similarly. □

The next result shows that the local distance between the minimizer of f and x only depends on the geometry at x. It can be used to localize the empirical risk minimizer as in Prop. 11.

Proposition 21. *Whenever $R_\nu \|\nabla f(x)\|_{H^{-1}(x)} \leq K_\nu$, the function f has a unique minimizer $\bar{x}$ and*

$$\|\bar{x} - x\|_x \leq 4 \|\nabla f(x)\|_{H^{-1}(x)}.$$

Proof. Consider the level set

$$\mathcal{L}_f(f(x)) := \{y \in \mathcal{X} : f(y) \leq f(x)\} \neq \emptyset.$$

Take an arbitrary $y \in \mathcal{L}_f(f(x))$. According to Prop. 17, we have

$$0 \geq f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + \bar{\bar{\omega}}_\nu(-d_\nu(x, y)) \|y - x\|_x^2.$$

By the Cauchy-Schwarz inequality and Lems. 18 and 19, we get

$$\bar{\bar{\omega}}_\nu(-R_\nu \|y - x\|_x) \|y - x\|_2^2 \leq \|\nabla f(x)\|_{H^{-1}(x)} \|y - x\|_x$$

This implies

$$\bar{\bar{\omega}}_\nu(-R_\nu \|y - x\|_x) R_\nu \|y - x\|_2 \leq R_\nu \|\nabla f(x)\|_{H^{-1}(x)} \leq K_\nu.$$

Due to Cor. 20, it holds that $R_\nu \|y - x\|_x < 1 + \mathbb{1}\{\nu = 2\}$ and $\bar{\bar{\omega}}_\nu(-R_\nu \|y - x\|_x) \geq 1/4$. It follows that $d_\nu(x, y) < 1 + \mathbb{1}\{\nu = 2\}$ and

$$\|y - x\|_x \leq 4 \|\nabla f(x)\|_{H^{-1}(x)}.$$

Hence, the level set $\mathcal{L}_f(f(x))$ is compact so that f has a minimizer $\bar{x}$. Moreover, by Prop. 16 and $\nabla^2 f(x) \succ 0$, we obtain $\nabla^2 f(y) \succ 0$ for all $y \in \mathcal{L}_f(f(x))$. This yields that $\bar{x}$ is the unique minimizer of f and it satisfies

$$\|\bar{x} - x\|_x \leq 4 \|\nabla f(x)\|_{H^{-1}(x)}.$$

□

Remark 22. *A similar result also appears in [23, Prop. B.4]. We extend their result from $\nu \in \{2, 3\}$ to $\nu \geq 2$.*

E.2 Concentration of random vectors and matrices

We start with the precise definitions of sub-Gaussian random vectors [32, Chapter 3.4] and the matrix Bernstein condition [33, Chapter 6.4].

Definition 23 (Sub-Gaussian vector)**.** Let $S \in \mathbb{R}^d$ be a random vector. We say S is sub-Gaussian if $\langle S, s\rangle$ is sub-Gaussian for every $s \in \mathbb{R}^d$. Moreover, we define the sub-Gaussian norm of S as

$$\|S\|_{\psi_2} := \sup_{\|s\|_2 = 1} \|\langle S, s\rangle\|_{\psi_2}.$$

Note that $\|\cdot\|_{\psi_2}$ is a norm and satisfies, e.g., the triangle inequality.

Remark 24. *When S is not mean-zero, we have*

$$\|S - \mathbb{E}[S]\|_{\psi_2} = \sup_{\|s\|_2=1} \|\langle S - \mathbb{E}[S], s\rangle\|_{\psi_2} = \sup_{\|s\|_2=1} \left\|s^\top S - \mathbb{E}[s^\top S]\right\|_{\psi_2}.$$

According to Vershynin [32, Lemma 2.6.8], we obtain

$$\|S - \mathbb{E}[S]\|_{\psi_2} \leq C \sup_{\|s\|_2=1} \left\|s^\top S\right\|_{\psi_2} = C \|S\|_{\psi_2},$$

where C is an absolute constant.

Definition 25 (Matrix Bernstein condition)**.** Let $H \in \mathbb{R}^{d\times d}$ be a zero-mean symmetric random matrix. We say H satisfies a Bernstein condition with parameter $b > 0$ if, for all $j \geq 3$,

$$\mathbb{E}[H^j] \preceq \frac{1}{2} j! b^{j-2} \, \mathbb{V}\mathrm{ar}(H).$$

The sum of i.i.d. sub-Gaussian vectors is also sub-Gaussian according to the following lemma.

Lemma 26 (Vershynin [32], Lemma 5.9)**.** *Let $S_1, \ldots, S_n$ be i.i.d. random vectors, then we have $\|\sum_{i=1}^n S_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|S_i\|_{\psi_2}^2$.*

We call a random vector $S \in \mathbb{R}^d$ isotropic if $\mathbb{E}[S] = 0$ and $\mathbb{E}[SS^\top] = I_d$. The following theorem is a tail bound for quadratic forms of isotropic sub-Gaussian random vectors.

Theorem 27 (Ostrovskii and Bach [23], Theorem A.1)**.** *Let $S \in \mathbb{R}^d$ be an isotropic random vector with $\|S\|_{\psi_2} \leq K$, and let $J \in \mathbb{R}^{d\times d}$ be positive semi-definite. Then,*

$$\mathbb{P}(\|S\|_J^2 - \mathit{Tr}(J) \geq t) \leq \exp\left(-c \min\left\{\frac{t^2}{K^2 \|J\|_2^2}, \frac{t}{K \|J\|_\infty}\right\}\right).$$

In other words, with probability at least $1 - \delta$, it holds that

$$\|S\|_J^2 - \mathit{Tr}(J) \lesssim K^2 \left(\|J\|_2 \sqrt{\log(e/\delta)} + \|J\|_\infty \log(1/\delta)\right). \tag{18}$$

A zero-mean symmetric random matrix H is said to be sub-Gaussian with parameter V if $\mathbb{E}[e^{\lambda H}] \preceq e^{\lambda^2 V/2}$ for all $\lambda \in \mathbb{R}$. The next theorem is the Bernstein bound for random matrices.

Theorem 28 (Wainwright [33], Theorem 6.17)**.** *Let $\{H_i\}_{i=1}^n$ be a sequence of zero-mean independent symmetric random matrices that satisfies the Bernstein condition with parameter $b > 0$. Then, for all $\delta > 0$, it holds that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n H_i\right|_2 \geq \delta\right) \leq 2\,\mathbf{Rank}\left(\sum_{i=1}^n \mathbb{V}\mathrm{ar}(H_i)\right) \exp\left\{-\frac{n\delta^2}{2(\sigma^2 + b\delta)}\right\}, \tag{19}$$

where $\sigma^2 := \frac{1}{n}\left|\sum_{i=1}^n \mathbb{V}\mathrm{ar}(H_i)\right|_2$.