THE RAO, WALD, AND LIKELIHOOD-RATIO TESTS UNDER GENERALIZED SELF-CONCORDANCE

Lang Liu   *Zaid Harchaoui*

Department of Statistics, University of Washington

ABSTRACT

Three classical approaches to goodness-of-fit testing are Rao's test, Wald's test, and the likelihood-ratio test. The asymptotic equivalence of these three tests under the null hypothesis is a famous connection in statistical detection theory. We revisit these three likelihood-related tests from a non-asymptotic viewpoint under self-concordance assumptions. We recover the equivalence of the three tests and characterize the critical sample size beyond which the equivalence holds asymptotically. We also investigate their behavior under local alternatives. Along the way, we establish an estimation bound that matches the misspecified Cramér-Rao lower bound. We illustrate the interest of our results using generalized linear models and score matching with exponential families.

Index Terms— score test, self-concordance, statistical detection.

1. INTRODUCTION

Likelihood-based statistical inference is at the core of statistical theory, detection theory, and recent developments on computational and statistical trade-offs in learning theory [3, 7, 8]. Three classical approaches to statistical inference have been put forward independently, and have been found after their introduction to share essential connections. Rao's score test, Wald's parameter test, and the likelihood-ratio test have been known to be equivalent under the null hypothesis of goodness-of-fit testing under classical large n small d asymptotics. We mention here, among many of them, the general monographs [6, 15, 16]; a modern treatment of this equivalence under general assumptions is yet hard to find in monographs.

In higher dimensional settings, the non-asymptotic viewpoint has been fruitful in tackling estimation and prediction problems – the results are developed for all fixed n so that it also captures the asymptotic regime where d grows with n. Early works in this line of research focus on specific models such as least-squares regression [4], logistic regression [1], and robust M-estimation [20]. The paper [12] addressed the finite-sample regime in full generality in a spirit similar to the classical local asymptotic normality theory. The approach of [12] relies on heavy empirical process machinery and requires strong global assumptions on the deviation of the empirical risk process. More recently, the work in [11] focused on risk bounds, specializing the discussion to linear models with (pseudo) self-concordant losses and obtained a more transparent analysis under neater assumptions.

A critical tool arising from this line of research is the so-called *Dikin ellipsoid*, a geometric object identified in the theory of convex optimization [10]. 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 that 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. We show that the concordance of the Hessians near the optimum with the Hessian at the optimum has interesting implications for statistical inference. The aforementioned three tests achieve the same Type I error under the null hypothesis, and their Type II error under local alternatives can also be characterized. In fact, under generalized self-concordance, the empirical Hessian concentrates around the population Hessian within a Dikin ellipsoid, and their distance can be precisely controlled.

We first recall in Sec. 2 the three tests, Rao's score test, Wald's parameter test, and the likelihood-ratio test, and their classical asymptotic equivalence under the null hypothesis. We then present our main results in Sec. 3—the control of Type I and Type II errors of the three tests under generalized self-concordance. We illustrate in Sec. 4 the interest of the results using generalized linear models and score matching with exponential families. The proofs, additional discussions, and a complete bibliography can be found in [9].

2. THE THREE TESTS UNDER CLASSICAL ASYMPTOTICS

We briefly recall the framework of goodness-of-fit testing. 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

This work was supported by NSF DMS-2023166, CCF-2019844, DMS-2052239, DMS- 2134012, DMS-2133244, NIH, CIFAR-LMB, and faculty research awards. Part of this work was done while Zaid Harchaoui was visiting the Simons Institute for the Theory of Computing, and while Lang Liu was at the University of Washington.

family of distributions $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ such that there exists a unique $\theta_\star$ with $\mathbb{P} = P_{\theta_\star}$. Given an i.i.d. sample $\{Z_i\}_{i=1}^n$ from the data distribution $P_{\theta_\star}$, we are interested in inferring properties of the parameter $\theta_\star$.

Let $\Theta_0 \subset \Theta$ be a subset of parameters. A goodness-of-fit testing problem is to test the null hypothesis that $\theta_\star \in \Theta_0$ versus the alternative hypothesis that $\theta_\star \notin \Theta_0$. In this paper, we focus on a simple null hypothesis where $\Theta_0 := \{\theta_0\}$ is a singleton. In other words,

$$\mathcal{H}_0 : \theta_\star = \theta_0 \leftrightarrow \mathcal{H}_1 : \theta_\star \neq \theta_0.$$

A statistical test consists of a *test statistic* $T := T(Z_1, \ldots, Z_n)$ and a prescribed *critical value* t_n, and we reject the null hypothesis if $T > t_n$. Its performance is quantified by the *type I error rate* $\Pr(T > t_n \mid \mathcal{H}_0)$ and *statistical power* $\Pr(T > t_n \mid \mathcal{H}_1)$.

Classical goodness-of-fit tests include the Rao (score) test, the Wald (parameter) test, and the likelihood-ratio test (LRT). Moreover, the three tests are known – under classical large n fixed d asymptotics – to be asymptotically equivalent under the null hypothesis. In the following, we introduce each of them under standard regularity conditions and give intuition on their asymptotically equivalence.

Notation. Let $\ell(\theta; z) := -\log P_\theta(z)$ and $L(\theta) := \mathbb{E}[-\log P_\theta(Z)]$. Following the terminology in statistical learning theory, we refer to ℓ as the *loss function* and L as the *population risk*. The *empirical risk* is defined as $L_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$. We denote by $S(\theta; z) := \nabla_\theta \ell(\theta; z)$ the gradient of the loss at z and $H(\theta; z) := \nabla_\theta^2 \ell(\theta; z)$ 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 $H(\theta) = \nabla_\theta^2 L(\theta)$. Note that the two optimality conditions then read $S(\theta_\star) = 0$ and $H(\theta_\star) \succ 0$. Furthermore, we let $G(\theta; z) := S(\theta; z)S(\theta; z)^\top$ and $G(\theta) := \mathbb{E}[S(\theta; Z)S(\theta; Z)^\top]$ be the autocorrelation matrices of the gradient. We define their empirical quantities as $S_n(\theta) := n^{-1}\sum_{i=1}^n S(\theta; Z_i)$, $H_n(\theta) := n^{-1}\sum_{i=1}^n H(\theta; Z_i)$, and $G_n(\theta) := n^{-1}\sum_{i=1}^n G(\theta; Z_i)$. For simplicity of the notation, we let $G_\star := G(\theta_\star)$ and $H_\star := H(\theta_\star)$.

2.1. The Rao test

To motivate the Rao test, we assume the null hypothesis is true, that is, $\theta_\star = \theta_0$. As a result, it holds that $S(\theta_0) = 0$. The Rao test is based on $S_n(\theta_0)$—an empirical estimator of $S(\theta_0)$. It is unbiased since $\mathbb{E}[S_n(\theta_0)] = S(\theta_0) = 0$. By the central limit theorem, it holds that

$$\sqrt{n} S_n(\theta_0) \to_d \mathcal{N}_d(0, G(\theta_0)).$$

Since the model is well-specified, i.e., $\mathbb{P} \in \mathcal{P}_\Theta$, it can be shown that $G(\theta_0) = H(\theta_0)$. As a result, $G(\theta_0)$ can be estimated by $H_n(\theta_0)$ and thus, by Slutsky's lemma,

$$n \left\|S_n(\theta_0)\right\|^2_{H_n(\theta_0)^{-1}} \to_d \chi^2_d, \quad \text{under } \mathcal{H}_0, \tag{1}$$

where $\|u\|_A^2 := u^\top A u$ for a vector $u \in \mathbb{R}^d$ and a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$. The Rao statistic is exactly $T_{\text{Rao}} := \|S_n(\theta_0)\|^2_{H_n(\theta_0)^{-1}}$, and the larger it is the less likely the null hypothesis is true.

2.2. The Wald test

Let $\theta_n := \arg\min_{\theta\in\Theta} L_n(\theta)$ be (assumed to be unique and exist) the *empirical risk minimizer* (or *maximum likelihood estimator*). The Wald statistic is defined as $T_{\text{Wald}} := \|\theta_n - \theta_0\|^2_{H_n(\theta_n)}$, i.e., the squared norm of the difference $\theta_n - \theta_0$ weighted by $H_n(\theta_n)$. It is related to the Rao test via the Taylor expansion: there exists $\bar\theta_n$, a convex combination of θ_0 and θ_n, such that

$$S_n(\theta_0) = S_n(\theta_0) - S_n(\theta_n) = H_n(\bar\theta_n)(\theta_0 - \theta_n).$$

Since it is usually true that $\theta_n = \theta_0 + o_p(1)$ as $n \to \infty$, we have

$$\begin{aligned} T_{\text{Rao}} &= \left\|H_n(\bar\theta_n)(\theta_n - \theta_0)\right\|^2_{H_n(\theta_0)^{-1}} \\ &= \|\theta_n - \theta_0\|^2_{H_n(\theta_0)} + o_p(1) = T_{\text{Wald}} + o_p(1). \end{aligned}$$

Moreover, it can be shown that $H_n(\theta_n) \to_p H(\theta_0)$ and $H_n(\bar\theta_n) \to_p H(\theta_0)$, which implies

$$nT_{\text{Wald}} = n \left\|\theta_n - \theta_0\right\|^2_{H_n(\theta_n)} \to_d \chi^2_d, \quad \text{under } \mathcal{H}_0. \tag{2}$$

2.3. The likelihood-ratio test

The LRT statistic is defined as $T_{\text{LR}} := 2[L_n(\theta_0) - L_n(\theta_n)]$. Since $-nL_n(\theta)$ is the log-likelihood of the data $\{Z_i\}_{i=1}^n$ under the model P_θ, the LRT statistic can be written as the log-likelihood ratio of the data under P_{θ_n} and the one under P_{θ_0}. It is related to the Wald statistic (and thus the Rao statistic) via another Taylor expansion: there exists $\tilde\theta_n$, a convex combination of θ_0 and θ_n, such that

$$L_n(\theta_0) - L_n(\theta_n) = \frac{1}{2}\left\|\theta_n - \theta_0\right\|^2_{H_n(\tilde\theta_n)},$$

where we have used $S_n(\theta_n) = 0$. Following the argument of the Wald statistic, we have $T_{\text{LR}} = T_{\text{Wald}} + o_p(1)$ and

$$nT_{\text{LR}} = 2n[L_n(\theta_0) - L_n(\theta_n)] \to_d \chi^2_d, \quad \text{under } \mathcal{H}_0. \tag{3}$$

This is known as the Wilks theorem [19].

2.4. Equivalence of the three tests

According to the limiting behavior in (1), (2), and (3), it is clear that the Rao, Wald, and likelihood-ratio tests are asymptotically equivalent under the null. However, due to its asymptotic nature, it is unclear how large n should be in order for the equivalence to be valid.

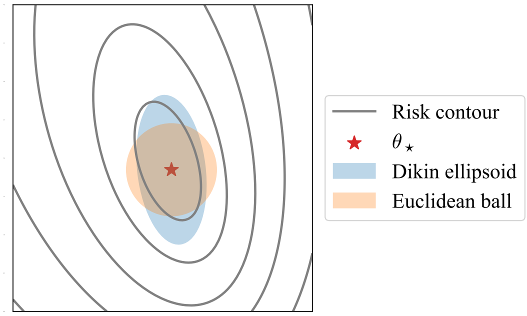

Fig. 1: Dikin ellipsoid and Euclidean ball.

3. MAIN RESULTS

After introducing some notation and definitions in Sec. 3.1, we present in Sec. 3.2 our main results characterizing the type I error rate and the power of the Rao, Wald, and LR tests under generalized self-concordance.

3.1. Preliminaries

Dikin ellipsoid. In our analysis of the Wald test and LRT, the first step is to localize the estimator to a *Dikin ellipsoid* at $\theta_\star$ of radius r, i.e.,

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

The key difference between Dikin ellipsoids and Euclidean balls is that the shape of a Dikin ellipsoid is adapted to the geometry near the optimum whereas the shape of the Euclidean ball is always the same no matter which population risk is used. This is illustrated in Fig. 1.

Generalized self-concordance. We will use the notion of *self-concordance* from convex optimization in our analysis. Self-concordance originated from the analysis of the interior-point and Newton-type convex optimization methods [10]. Bach [1] adapted it to statistical models by defining the notion of pseudo self-concordance, and derived finite-sample generalization bounds for logistic regression. Recently, Sun and Tran-Dinh [14] proposed the *generalized self-concordance* which unifies self-concordance and pseudo 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)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be open and $f : \mathcal{X} \to \mathbb{R}$ be a closed convex function. For $R > 0$ and $\nu > 0$, 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 $\|u\|^2_{\nabla^2 f(x)} := u^\top \nabla^2 f(x) u$.

In contrast to strong convexity which imposes a gross lower bound on the Hessian, generalized self-concordance specifies the rate at which the Hessian can vary, leading to a finer control on the Hessian. As we shall see in Sec. 3.2, owing to the generalized self-concordance, we are able to remove the direct dependency on $\lambda_\star := \lambda_{\min}(H_\star)$ in our bounds.

Concentration of Hessian. One key result towards deriving our bounds is the concentration of empirical Hessian, i.e., $(1 - c_n(\delta))H(\theta) \preceq H_n(\theta) \preceq (1 + c_n(\delta))H(\theta)$ with probability at least $1 - \delta$. When the loss function is of the form $\ell(\theta; z) := \ell(y, \theta^\top x)$ (e.g., generalized linear models), the empirical Hessian is of the form of a sample covariance. Assuming X to be sub-Gaussian, Ostrovskii and Bach [11] obtained a concentration bound for $H_n(\theta_\star)$ with $c_n(\delta) = O(\sqrt{(d + \log(1/\delta))/n})$ via the concentration bound for sample covariance [17, Thm. 5.39]. For general loss functions, such a special structure cannot be exploited. We overcame this challenge by the matrix Bernstein inequality [18, Thm. 6.17], obtaining a tighter concentration bound with $c_n(\delta) := O(\sqrt{\log(d/\delta)/n})$.

Model misspecification and effective dimension. Even though the framework of the goodness-of-fit testing assumes a well-specified model, our intermediate analysis also holds under model misspecification, i.e., $\mathbb{P} \notin \mathcal{P}_\Theta$, assuming that $\theta_\star := \arg\min_{\theta \in \Theta} L(\theta)$ uniquely exists. Examples include score matching [5]. Under model misspecification, a quantity that plays a central role is the *effective dimension*.

Definition 2. *We define the effective dimension to be*

$$d_\star := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2}). \tag{4}$$

The effective dimension appears in non-asymptotic analyses of (penalized) M-estimation recently; see, e.g., [13]. The quantity provides a characterization of the complexity of the parameter space Θ which is adapted to both the data distribution and the loss function. When the model is well-specified, it can be shown that $H_\star = G_\star$ and thus $d_\star = d$. When the model is misspecified, $d_\star$ can be much smaller than d depending on the spectra of $H_\star$ and $G_\star$. To illustrate this, we can compare $d_\star$ with d under different regimes of eigendecay; see [9, Tab. 3].

3.2. The three tests in the non-asymptotic setting

We now give simplified versions of our main theorems. We use C to represent a constant that does not directly depend on $n, d, R, \lambda_\star$ and may change from line to line. We use $\lesssim$ and $\gtrsim$ to hide such constants. The precise versions, including assumptions, can be found in [9]. Recall that $\lambda_\star := \lambda_{\min}(H_\star)$.

Theorem 1 (Type I error rate)**.** *Under $\mathcal{H}_0$, we have, with probability at least $1 - \delta$,*

$$T_{\mathit{Rao}} \lesssim \frac{d}{n} + \frac{1}{n} \log \frac{e}{\delta}$$

whenever $n \gtrsim \log(2d/\delta)$. *Additionally,*

$$T_{\mathit{LR}}, T_{\mathit{Wald}} \lesssim \frac{d}{n} + \frac{1}{n}\log\frac{e}{\delta}$$

whenever

$$n \gtrsim \log\frac{2d}{\delta} + d_\star \frac{R^2}{\lambda_\star^{3-\nu}}\left(\log\frac{2nd}{\delta}\right)^{\nu-2}\log\frac{e}{\delta}. \tag{5}$$

This result shows, from a *non-asymptotic viewpoint*, that the three test statistics have a tail behavior that is governed by a χ_d^2 distribution. We also characterize the critical sample size in (5) enough to enter the asymptotic regime. In terms of power, let us look at their tail behavior under *local alternatives* when $\theta_\star \to \theta_0$ as $n \to \infty$. Let $\Omega(\theta) := G(\theta)^{1/2}H(\theta)^{-1}G(\theta)^{1/2}$ and $h(\tau) := \min\{\tau^2, \tau\}$.

Theorem 2 (Statistical power)**.** *Let $\theta_\star \neq \theta_0$ that may depend on n. The following statements are true for sufficiently large n.*

(a) If $\|\theta_\star - \theta_0\|_{H(\theta_0)} = O(n^{-1/2})$, we have

$$Pr(T_{\mathit{Rao}} > t_n(\alpha)) \leq 2de^{-Cn} + e^{-Ch(\|\Omega(\theta_0)\|_2^{-1})}.$$

If $\|\theta_* - \theta_0\|_{H(\theta_0)} = \omega(n^{-1/2})$, we have

$$Pr(T_{\mathit{Rao}} > t_n(\alpha)) \geq 1 - 2de^{-Cn} - e^{-Ch\left(\frac{n\bar{\tau}_n}{\|\Omega(\theta_0)\|_2}\right)},$$

where $\bar{\tau}_n \asymp \|\theta_\star - \theta_0\|_{H(\theta_0)}^2$.

(b) If $\|\theta_\star - \theta_0\|_{H(\theta_0)} = O(n^{-1/2})$, we have

$$\begin{aligned}&Pr(T_{\mathit{Wald}} > t_n(\alpha))\\ &\leq 2nde^{-C(\frac{\lambda_\star^{3-\nu}n}{R^2d})^{1/(\nu-1)}} + e^{-Ch\left(\|\Omega(\theta_\star)\|_2^{-1}\right)}.\end{aligned}$$

If $\|\theta_* - \theta_0\|_{H(\theta_0)} = \omega(n^{-1/2})$, we have

$$\begin{aligned}&Pr(T_{\mathit{Wald}} > t_n(\alpha))\\ &\geq 1 - 2nde^{-C(\frac{\lambda_\star^{3-\nu}n}{R^2d})^{\frac{1}{\nu-1}}} - e^{-Ch(\frac{n\bar{\tau}'_n}{\|\Omega(\theta_\star)\|_2})},\end{aligned}$$

where $\bar{\tau}'_n \asymp \|\theta_\star - \theta_0\|_{H(\theta_0)}^2$.

(c) The same statements in (b) hold for T_{LR}.

Recall that $\theta_\star$ depends on n under the local alternatives. When $\|\theta_\star - \theta_0\|_{H(\theta_0)} = O(n^{-1/2})$ and n is sufficiently large, we have $\|\Omega(\theta_0)\|_2 \approx \|\Omega(\theta_\star)\|_2 = 1$. Hence, according to Thm. 2, the powers of the three tests are asymptotically upper bounded by a constant. When $\|\theta_\star - \theta_0\|_{H(\theta_0)} = \omega(n^{-1/2})$ and $\|\theta_\star - \theta_0\|_{H(\theta_0)} = O(n^{-(\nu-2)/(2\nu-2)})$, the powers of the three tests tend to one at rate $O\big(\exp(-n\|\theta_\star - \theta_0\|_{H(\theta_0)}^2)\big)$.

Remark. A key result we establish towards proving Thms. 1 and 2 is

$$\|\theta_n - \theta_\star\|_{H_n(\theta_n)}^2 \lesssim \frac{d_\star}{n} + \frac{\|\Omega(\theta_\star)\|_2}{n}\log\frac{e}{\delta}$$

whenever n satisfies (5). This bound matches the misspecified Cramér-Rao lower bound [e.g., 2, Thm. 1] up to a constant factor. The bound also yields the Wald confidence set for the estimator θ_n under model misspecification.

4. EXAMPLES

To illustrate the generality of our results, we instantiate them on two familiar examples: generalized linear modeling using maximum likelihood and density estimation using score matching. Numerical examples can be found in [9, Sec. 5].

Example 1 (Generalized linear models)**.** *Let $Z := (X, Y)$ be a pair of input and output, where $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. Let $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and μ be a measure on $\mathcal{Y}$. Consider the statistical model*

$$p_\theta(y \mid x) \sim \frac{\exp[\theta^\top t(x,y) + h(x,y)]}{\int \exp[\theta^\top t(x,\bar{y}) + h(x,\bar{y})]\mathrm{d}\mu(\bar{y})}\mathrm{d}\mu(y)$$

with $\|t(X,Y)\|_2 \leq_{a.s.} M$. It induces the loss function

$$\begin{aligned}\ell(\theta; z) :=& -\theta^\top t(x,y) - h(x,y)\\ &+ \log\int \exp[\theta^\top t(x,\bar{y}) + h(x,\bar{y})]\mathrm{d}\mu(\bar{y}),\end{aligned}$$

which is $(2M, 2)$-generalized self-concordant. Hence, our bounds from Sec. 3.2 hold with $\nu = 2$ and $R = 2M$.

Example 2 (Score matching with exponential families)**.** *Assume that $\mathbb{Z} = \mathbb{R}^p$. Consider an exponential family on $\mathbb{R}^d$ with probability density*

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

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

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

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 positive semi-definite and

$$b(z) := \sum_{k=1}^p \left[\frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial h(z)}{\partial z_k}\frac{\partial t(z)}{\partial z_k}\right]$$

$$c(z) := \sum_{k=1}^p \left[\frac{\partial^2 h(z)}{\partial z_k^2} + \big(\frac{\partial h(z)}{\partial z_k}\big)^2\right].$$

The score matching loss $\ell(\theta; z)$ is convex. Moreover, since the third derivative of $\ell(\cdot; z)$ is zero, the score matching loss is generalized self-concordant for all $\nu \geq 2$ and $R \geq 0$. Therefore, our bounds from Sec. 3.2 hold with $\nu = 2$ and $R = 0$.

References

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

[2] S. Fortunati, F. Gini, and M. S. Greco. The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions. *IEEE Transactions on Signal Processing*, 64(9), 2016.

[3] S. Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

[4] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *COLT*, 2012.

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

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

[7] D. Kunisky. Hypothesis testing with low-degree polynomials in the morris class of exponential families. In *Conference on Learning Theory*, pages 2822–2848. PMLR, 2021.

[8] D. Kunisky, A. S. Wein, and A. S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *Mathematical Analysis, its Applications and Computation: ISAAC 2019, Aveiro, Portugal, July 29–August 2*, pages 1–50. Springer, 2022.

[9] L. Liu and Z. Harchaoui. Confidence sets under generalized self-concordance. *arXiv preprint*, 2022.

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

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

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

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

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

[15] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2009.

[16] A. W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[17] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 2010.

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

[19] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 1938.

[20] 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.