Linear Models are Most Favorable among Generalized Linear Models

Kuan-Yun Lee and Thomas A. Courtade
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{timkylee, courtade}@berkeley.edu

***Abstract*—We establish a nonasymptotic lower bound on the L_2 minimax risk for a class of generalized linear models. It is further shown that the minimax risk for the canonical linear model matches this lower bound up to a universal constant. Therefore, the canonical linear model may be regarded as most favorable among the considered class of generalized linear models (in terms of minimax risk). The proof makes use of an information-theoretic Bayesian Cramér-Rao bound for log-concave priors, established by Aras et al. (2019).**

I. Introduction and Main Results

As their name suggests, generalized linear models (GLMs) are a flexible class of parametric statistical models that generalize the class of linear models relating a random observation $X \in \mathbb{R}^n$ to a parameter $\theta \in \mathbb{R}^d$ via the linear relation

$$X = M\theta + Z, \tag{1}$$

where $M \in \mathbb{R}^{n\times d}$ is a known (fixed) design matrix, and $Z \in \mathbb{R}^n$ is a random noise vector. For a univariate GLM in canonical form with natural parameter $\eta \in \mathbb{R}$, the density of observation $X \in \mathbb{R}$ given η is expressed as the exponential family

$$f(x;\eta) = h(x)\exp\left(\frac{\eta x - \Phi(\eta)}{s(\sigma)}\right),$$

for known functions $h : \mathcal{X} \subseteq \mathbb{R} \to [0,\infty)$ (the *base measure*), $\Phi : \mathbb{R} \to \mathbb{R}$ (the *cumulant function*) and a scale parameter $s(\sigma) > 0$. For this general class of models, the question of central importance is how well one can estimate η from an observation $X \sim f(\cdot;\eta)$, where $f(\cdot;\eta)$ is understood to be a density on a probability space $(\mathcal{X} \subseteq \mathbb{R}, \mathcal{F})$ with respect to a dominating σ-finite measure λ. This class of models captures a wide variety of parametric models such as binomial, Gaussian, Poisson, etc. As a specific example, we can take $\mathcal{X} = \{0,1,2,\ldots\}$ equipped with the counting measure λ. For $h(x) = 1/x!$, $\Phi(t) = e^t$ and $s(\sigma) = 1$, the observation $X \sim f(\cdot;\eta)$ will be $\mathrm{Poisson}(e^\eta)$.

In this paper, we restrict our attention to multivariate GLMs of the form

$$f(x;\theta) = \prod_{i=1}^{n}\left\{h(x_i)\exp\left(\frac{x_i\langle m_i,\theta\rangle - \Phi(\langle m_i,\theta\rangle)}{s(\sigma)}\right)\right\}, \tag{2}$$

for a real parameter $\theta \in \mathbb{R}^d$ and a fixed design matrix $M \in \mathbb{R}^{n\times d}$, with rows given by the vectors $\{m_i\}_{i=1}^n \subset \mathbb{R}^d$. In words, the above model assumes each X_i is drawn from the same exponential family, with respective natural parameters $\langle m_i,\theta\rangle$, $i = 1,2,\ldots,n$. This captures the linear model (1) in the usual case where the noise vector Z is assumed to be standard normal on $\mathbb{R}^n$, but is also flexible enough to capture many other models of interest.

In terms of parameter estimation, a typical figure of merit is the constrained L_2 minimax risk, which corresponds to the worst-case L_2 estimation error, where θ is allowed to range over a constrained set Θ. For our purposes, we take Θ equal to the Euclidean ball in $\mathbb{R}^d$, denoted $\mathbb{B}_2^d(1) := \{v : v \in \mathbb{R}^d, \|v\|_2^2 \le 1\}$, which is a common choice in applications. More precisely, we make the following definition.

Definition 1. *For the generalized linear model* (2)*, we define the associated minimax risk via*

$$R^*(h,\Phi,M,s(\sigma)) := \inf_{\hat\theta}\sup_{\theta\in\mathbb{B}_2^d(1)}\mathbb{E}\|\theta-\hat\theta\|_2^2,$$

where the expectation is over $X \sim f(\cdot;\theta)$, and the infimum is over all $\mathbb{R}^d$-valued estimators $\hat\theta$ (i.e., measurable functions of the observation X).

Before we state our main results, we make the following assumption throughout:

Assumption 1. *We assume the cumulant function $\Phi : \mathbb{R} \to \mathbb{R}$ in* (2) *is twice-differentiable, with second derivative uniformly bounded as $\Phi'' \le L$, for some $L > 0$.*

Remark 2. *This assumption is standard in the literature on minimax estimation for GLMs, and is equivalent to the map $\theta \longmapsto \mathbb{E}_{X\sim f(\cdot;\theta)}[X]$ being L-Lipschitz. See, for example, [1]–[4].*

Our first main result is a general lower bound on the minimax risk for the class of GLMs introduced above.

Theorem 3. *The L_2 minimax risk for the class of models* (2) *is lower bounded according to*

$$R^*(h,\Phi,M,s(\sigma)) \gtrsim \min\left(\frac{s(\sigma)}{L}\ \mathrm{Tr}\left((M^\top M)^{-1}\right), 1\right), \tag{3}$$

where $\gtrsim$ denotes inequality, up to a universal constant.

Remark 4. *In case $M^\top M$ is not invertible, we adopt the convention that $\mathrm{Tr}\left((M^\top M)^{-1}\right) = +\infty$. This situation occurs when M is not full rank, in which case θ is not identifiable in the null space of M and constant error is unavoidable.*

Remark 5. *In fact, with minor modification, Theorem 3 holds for the more general class of GLMs with observations drawn from densities of the form*

$$f(x;\theta) = \prod_{i=1}^{n} \left\{ h_i(x_i) \exp\left(\frac{x_i \langle m_i, \theta\rangle - \Phi_i(\langle m_i, \theta\rangle)}{s_i(\sigma)} \right) \right\}.$$

See Section II-C for further remarks.

Remark 6. *Since minimax risk is generally characterized modulo universal constants, the statement* (3) *in terms of $\gtrsim$ is sufficient for our purposes. However, a careful analysis of our arguments reveals that $\gtrsim$ can be replaced with $\geq$ at the expense of including a modest constant in the RHS of* (3) *(e.g., $1/(\pi e^3)$).*

Most interestingly, the minimax bound (3) holds uniformly over the class of GLMs given by (2), and is of the correct order for the canonical linear model (1). Indeed, under the linear model $X = LM\theta + Z$, where Z is standard Gaussian with covariance $\sigma^2 L \cdot I$ and the design matrix M is full rank, the maximum likelihood estimator (MLE) estimator $\hat{\theta}_{\text{MLE}}$ is given by

$$\hat{\theta}_{\text{MLE}} = L^{-1}(M^\top M)^{-1} M^\top X.$$

One can explicitly calculate the L_2 error as

$$\begin{aligned} \mathbb{E}\|\theta - \hat{\theta}_{\text{MLE}}\|_2^2 &= \mathbb{E}\|\theta - L^{-1}(M^\top M)^{-1}M^\top X\|_2^2 \\ &= \frac{1}{L^2}\mathbb{E}\|(M^\top M)^{-1}M^\top Z\|_2^2 \\ &= \frac{\sigma^2}{L}\operatorname{Tr}((M^\top M)^{-1}). \end{aligned} \tag{4}$$

The linear model in this case corresponds to $h(x) = e^{-x^2/(2L\sigma^2)}$, $s(\sigma) = \sigma^2$, and $\Phi(t) = Lt^2/2$ in (2).

Comparing (4) to Theorem 3, we find that our minimax lower bound is achieved (up to a universal constant) for linear models of the above form. To summarize, we have the following:

Corollary 7. *Fix a design matrix M, scale parameter $s(\sigma)$ and $L > 0$. Among those generalized linear models in* (2) *with $\Phi'' \leq L$, linear models are most favorable in terms of minimax risk. More precisely, among this class of models,*

$$R^*(h, \Phi, M, s(\sigma)) \gtrsim R^*(e^{-(\cdot)^2/(2Ls(\sigma))}, (\cdot)^2 L/2, M, s(\sigma)).$$

Roughly speaking, the above asserts that linear models are most favorable among a broad class of GLMs, giving this paper its name.

A. Related Work

Perhaps most closely related to our work is that of Abramovich and Grinshtein [1], albeit for a slightly different setup. In particular, Abramovich and Grinshtein provide minimax lower bounds for the Kullback-Leibler divergence between the vector $M\theta$ and any estimator $\widehat{M\theta}$ under a k-sparse setting $\|\theta\|_0 \leq k$, with the parameter θ constrained to have at most k non-zero entries. When the cumulant function Φ is strongly convex with $0 < R \leq \Phi'' \leq L$ for some fixed constants R, L, we can adapt the arguments of [1] to obtain the following L_2 minimax lower bound

$$\inf_{\widehat{M\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \|M\theta - \widehat{M\theta}\|_2^2 \gtrsim \frac{ds(\sigma)R}{L^2} \cdot \frac{\lambda_{\min}(M^\top M)}{\lambda_{\max}(M^\top M)},$$

where M is assumed to be full rank and $\lambda_{\min}$ and $\lambda_{\max}$ denote smallest and largest eigenvalues, respectively. The minimax lower bound for estimating $M\theta$ is not directly comparable to our result, where the goal is estimation of θ. Nevertheless, using the operator norm inequality $\|M(\theta - \hat{\theta})\|_2^2 \leq \lambda_{\max}(M^\top M)\|\theta - \hat{\theta}\|_2^2$, we may conclude

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \|\theta - \hat{\theta}\|_2^2 \gtrsim \frac{ds(\sigma)R}{L^2} \cdot \frac{\lambda_{\min}(M^\top M)}{\lambda_{\max}^2(M^\top M)}.$$

A direct computation shows that (3) is sharper than the above L_2 minimax estimate since

$$\frac{d\,\lambda_{\min}(M^\top M)}{\lambda_{\max}^2(M^\top M)} \leq \frac{d}{\lambda_{\max}(M^\top M)} \leq \operatorname{Tr}\left(\left(M^\top M\right)^{-1}\right).$$

As for a general theory, apart from the gaussian linear model, the minimax estimator for the GLM does not have a closed form, but the Maximum Likelihood Estimator (MLE) can be approximated by iterative weighted linear regression [5]. A variety of estimators such as aggregate estimators [6], robust estimators [7] and GLM with Lasso [9] have been proposed to solve different settings of the GLM. We refer interested readers to [8] for the theory of GLMs.

Another line of related work explores models with stochastic design matrix M. Duchi, Jordan and Wainwright [10] consider inference of a parameter θ under privacy constraints. Negahban et al. [3] and Loh et al. [4] provide consistency and convergence rates for M-estimators in GLMs with low-dimensional structure under high-dimensional scaling.

Separate from the minimax problems considered here, model selection is another line of popular work. Model selection in linear regression dates back to the seventies and has regained popularity over the past decade, due to the increase in need of data exploration for high dimensional data; see [11]–[13] and many other works for the history. More recently, tools in model selection for linear regression have been adapted for the GLM; see [1] for a brief discussion.

B. Organization

The remainder of this paper is organized as follows. Preliminaries for the derivation of our minimax lower bounds are introduced in Section II-A. The proof of Theorem 3 is given in Section II-B, with further remarks in Section II-C.

II. Derivation of Minimax Bound for the GLM

The following notation is used throughout: upper-case letters (e.g., X, Y) denote random variables or matrices, and lower-case letters (e.g., x, y) denote realizations of random variables or vectors. We use subscript notation v_i to denote the i-th component of a vector $v = (v_1, v_2, \ldots, v_d)$, and we define the leave-one-out vector $v^{(j)} := (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)$.

A. Preliminaries

In the general framework of parametric statistics, let $(\mathcal{X}, \mathcal{F}, P_\theta; \theta \in \mathbb{R}^d)$ be a dominated family of probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$ with dominating σ-finite measure λ. To each P_θ, we associate a density $f(\cdot;\theta)$ with respect to λ according to

$$dP_\theta(x) = f(x;\theta)d\lambda(x). \tag{5}$$

Assuming the maps $\theta \longmapsto f(x;\theta)$, $x \in \mathcal{X}$, are differentiable, the Fisher information matrix associated to observation $X \sim f(\cdot;\theta)$ and parameter $\theta \in \mathbb{R}^d$ is defined as the matrix-valued map $\theta \longmapsto \mathcal{I}_X(\theta)$ with components

$$[\mathcal{I}_X(\theta)]_{ij} = \mathbb{E}\left[\frac{\partial \log f(X;\theta)}{\partial \theta_i}\frac{\partial \log f(X;\theta)}{\partial \theta_j}\right], \quad \theta \in \mathbb{R}^d.$$

Here and throughout, log denotes the natural logarithm. The following regularity assumption is standard when dealing with Fisher information.

Assumption 2. *The densities $f(\cdot;\theta)$ are sufficiently regular to permit the following exchange of integration and differentiation:*

$$\int_{\mathcal{X}} \nabla_\theta f(x;\theta)d\lambda(x) = 0; \quad \theta \in \mathbb{R}^d. \tag{6}$$

Here, ∇_θ denotes the gradient with respect to θ.

While the Fisher information is one notion of *information* that an observation $X \sim f(\cdot;\theta)$ reveals about the unknown parameter θ, it also makes sense to consider the usual mutual information $I(X;\theta)$ under the further assumption that θ is distributed according to a known prior distribution π (a probability measure on $\mathbb{R}^d$). Recent results by the authors together with Aras and Pananjady establish a quantitative relation between these two notions of information [14]. To state the result precisely, recall that a probability measure $d\mu = e^{-V}dx$ on $\mathbb{R}^d$ is said to be log-concave if the potential $V : \mathbb{R}^d \to \mathbb{R}$ is a convex function.

Lemma 8 ([14, Theorem 2])**.** *Let $\theta \sim \pi$, where π is log-concave on $\mathbb{R}^d$, and given θ let $X \sim f(\cdot;\theta)$. If Assumption 2 holds, then*

$$I(X;\theta) \le d \cdot \phi\left(\frac{\operatorname{Tr}(\operatorname{Cov}(\theta)) \cdot \operatorname{Tr}(\mathbb{E}\,\mathcal{I}_X(\theta))}{d^2}\right), \tag{7}$$

where

$$\phi(x) := \begin{cases} \sqrt{x} & \text{if } 0 \le x \le 1 \\ 1 + \frac{1}{2}\log x & \text{if } x \ge 1. \end{cases}$$

As discussed extensively in [14], the above result is related to the van Trees inequality [15], [16], and its entropic improvement due to Efroimovich [17]. The crucial feature of (7) compared to these other results is that it does not depend on the (information theorist's version of) Fisher information of the prior π, commonly denoted $\mathcal{J}(\pi)$. This is what is gained via the assumption of log-concavity, and is important for our analysis where we introduce (log-concave) priors with arbitrarily large Fisher information.

B. Proof of Theorem 3

Recall that the design matrix M has as its rows $\{m_i\}_{i=1}^n \subset \mathbb{R}^d$. Writing the matrix M in terms of its SVD $M = U\Sigma V^\top$ and defining u_i as the i-th column of the matrix $U^\top$, we have

$$\langle m_i, \theta\rangle = \langle \underbrace{\Sigma u_i}_{\bar{m}_i}, \underbrace{V^\top\theta}_{\bar{\theta}}\rangle = \langle \bar{m}_i, \bar{\theta}\rangle, \tag{8}$$

where we defined the variables $\bar{m}_i := \Sigma u_i$ and $\bar{\theta} := V^\top\theta$. Since V is an orthogonal matrix by definition, it follows by rotation invariance of the L_2 ball $\mathbb{B}_2^d(1)$ that the estimation problem can be equivalently formulated under the reparametrization $(\theta, M) \longrightarrow (\bar{\theta}, \bar{M})$, where $\bar{M} := MV = U\Sigma$. More specifically, the minimax risk for θ over the set of estimators for estimating $\theta \in \mathbb{B}_2^d(1)$ is equal to the minimax risk for estimating $\bar{\theta} \in \mathbb{B}_2^d(1)$. More precisely,

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \inf_{\hat{\bar{\theta}}} \sup_{\bar{\theta} \in \mathbb{B}_2^d(1)} \mathbb{E}\|\bar{\theta} - \hat{\bar{\theta}}\|_2^2.$$

As a result, we may assume without loss of generality that $M^\top M$ is a diagonal matrix.

By definition, minimax risk is lower bounded by the Bayes risk, when θ is assumed to be distributed according to a prior π, defined on the L_2 ball $\mathbb{B}_2^d(1)$. Hence, our task is to judiciously select a prior π that yields the desired lower bound. Toward this end, we will let π be the uniform measure on the rectangle $\prod_{i=1}^d [-\epsilon_i/2, \epsilon_i/2]$ for values $(\epsilon_i)_{i=1,2,\ldots,d}$ to be determined below satisfying

$$\sum_{i=1}^d \epsilon_i^2 \le 4. \tag{9}$$

In other words, our construction implies θ has independent components, with the i-th coordinate θ_i uniform on the interval $[-\epsilon_i/2, \epsilon_i/2]$. The interval lengths will, in general, be chosen to exploit the structure of the design matrix M.

We now describe our construction of the sequence $(\epsilon_i)_{i=1,2,\ldots,d}$. We start with the simple case, in which the matrix M does not have full (column) rank. In this case, there exists an eigenvalue $\lambda_k(M^\top M) = 0$. For this index k, we set $\epsilon_i = 2\delta_{ik}$, $i = 1, 2, \ldots, d$, where δ_{ij} is the Kronecker delta function. Now, we may bound

$$\begin{aligned}
\mathbb{E}\|\theta - \hat{\theta}\|_2^2 &\ge \operatorname{Var}(\theta_k - \hat{\theta}_k) \\
&\overset{(a)}{\ge} \frac{1}{2\pi e} e^{2h(\theta_k - \hat{\theta}_k)} \overset{(b)}{\ge} \frac{1}{2\pi e} e^{2h(\theta_k|\hat{\theta}_k)} \\
&= \frac{1}{2\pi e} e^{2h(\theta_k) - 2I(\hat{\theta}_k;\theta_k)} \overset{(c)}{\ge} \frac{1}{2\pi e} e^{2h(\theta_k) - 2I(X;\theta_k)} \\
&\overset{(d)}{=} \frac{2}{\pi e}
\end{aligned}$$

where (a) follows from the max-entropy property of gaussians; (b) follows since conditioning reduces entropy: $h(\theta_k - \hat{\theta}_k) \ge h(\theta_k - \hat{\theta}_k|\hat{\theta}_k) = h(\theta_k|\hat{\theta}_k)$; (c) follows from the data processing inequality since $\theta_k \to X \to \hat{\theta}_k$ forms a Markov chain; and (d) follows since $\theta_k \sim \operatorname{Unif}(-1, 1)$ and $I(X;\theta_k) = 0$, since π is supported in the kernel of M by construction.

Having shown the minimax risk is lower bounded by a constant when M does not have full (column) rank, we assume henceforth that M has full rank.

Note that under our assumptions, the pair (X, θ) has a joint distribution, and therefore so does the pair (X, θ_i). Consistent with the previously introduced notation, we write $\mathcal{I}_X(\theta_i)$ to denote the Fisher information of X drawn according to the conditional law of X given θ_i. With this notation in hand, the next lemma provides a comparison between the expected Fisher information conditioned on a single component θ_i of the parameter θ and the i-th diagonal entry of the expected Fisher information matrix conditioned for parameter θ.

Lemma 9. *When the components of parameter $\theta \sim \pi$, $\theta \in \mathbb{R}^d$ are independent and $X \sim f(\cdot;\theta)$ is generated by the GLM* (2)*, we have*

$$\mathbb{E}\left[\mathcal{I}_X(\theta)\right]_{ii} \geq \mathbb{E}\,\mathcal{I}_X(\theta_i) \quad i = 1, 2, \ldots, d.$$

Proof. The desired estimate is obtained by observing

$$\begin{aligned}
\mathbb{E}[\mathcal{I}_X(\theta)]_{ii} &= \mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i} f(X;\theta)\right)^2}{f(X;\theta)^2}\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[\frac{\left(\mathbb{E}\left[\frac{\partial}{\partial\theta_i} f(X;\theta)\middle|\, \theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i}\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right] = \mathbb{E}\,\mathcal{I}_X(\theta_i).
\end{aligned}$$

In the above, (a) is due to Cauchy-Schwarz. Indeed, let π_i and $\pi^{(i)}$ denote the marginal laws of θ_i and $\theta^{(i)}$, respectively. Using independence of θ_i and $\theta^{(i)}$, note that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i} f(X;\theta)\right)^2}{f(X;\theta)^2}\right] \\
&= \int_{\mathbb{R}}\int_{\mathcal{X}}\int_{\mathbb{R}^{d-1}} \frac{\left(\frac{\partial}{\partial\theta_i} f(x;\theta)\right)^2}{f(x;\theta)} d\pi^{(i)}(\theta^{(i)}) d\lambda(x) d\pi_i(\theta_i) \\
&\geq \int_{\mathbb{R}}\int_{\mathcal{X}} \frac{\left(\int_{\mathbb{R}^{d-1}} \frac{\partial}{\partial\theta_i} f(x;\theta) d\pi^{(i)}(\theta^{(i)})\right)^2}{\int_{\mathbb{R}^{d-1}} f(x;\theta) d\pi^{(i)}(\theta^{(i)})} d\lambda(x) d\pi_i(\theta_i) \\
&= \mathbb{E}\left[\frac{\left(\mathbb{E}\left[\frac{\partial}{\partial\theta_i} f(X;\theta)\middle|\,\theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right],
\end{aligned}$$

where the last line follows since

$$x \longmapsto \mathbb{E}\left[f(X;\theta)|\,\theta_i, X = x\right] = \int_{\mathbb{R}^{d-1}} f(x;\theta) d\pi^{(i)}(\theta^{(i)})$$

is the density (w.r.t. λ) of X given θ_i.

Equality (b) follows from independence between θ_i and $\theta^{(i)}$ and the Leibniz integral rule. Application of the latter can be justified by the assumed regularity of Φ and compactness of $\mathbb{B}_2^d(1)$. □

Next, fix $\epsilon_i > 0$. Since $\theta_i \sim \mathrm{Unif}(-\epsilon_i/2, \epsilon_i/2)$ has log-concave distribution, and the GLM (2) satisfies Assumption 2 (a consequence of Assumption 1 and the Leibniz integral rule, justified by regularity of Φ), we can apply Lemmas 8 and 9 to conclude

$$\begin{aligned}
e^{2h(\theta_i|\hat{\theta}_i)} &\geq e^{2h(\theta_i) - 2I(X;\theta_i)} \\
&\geq e^{2h(\theta_i) - 2\phi(\mathrm{Var}(\theta_i)\cdot\mathbb{E}\,\mathcal{I}_X(\theta_i))} \\
&\geq \epsilon_i^2 e^{-2\phi\left(\frac{\epsilon_i^2}{12}\mathbb{E}\left[\mathcal{I}_X(\theta)\right]_{ii}\right)}. \qquad (10)
\end{aligned}$$

Note that the last inequality used the identities $\mathrm{Var}(\theta_i) = \frac{\epsilon_i^2}{12}$ and $h(\theta_i) = \log(\epsilon_i)$, holding by construction.

Next, recall the following well-known identities associated with exponential families of the form we consider.

Lemma 10 ([8, Page 29])**.** *Fix m and θ, and consider a density $f(x;\theta) = h(x)\exp\left(\frac{x\langle m,\theta\rangle - \Phi(\langle m,\theta\rangle)}{s(\sigma)}\right)$ with respect to λ. A random observation $X \sim f(\cdot;\theta)$ has mean $\Phi'(\langle m,\theta\rangle)$ and variance $s(\sigma)\cdot\Phi''(\langle m,\theta\rangle)$.*

Combining the above with our assumption that $\Phi'' \leq L$, we have for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
[\mathcal{I}_X(\theta)]_{ii} &= \mathbb{E}_{X\sim f(\cdot;\theta)}\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)^2 \\
&= \frac{1}{s^2(\sigma)}\mathbb{E}_{X\sim f(\cdot;\theta)}\left(\sum_{j=1}^n M_{ji}\left(X_j - \Phi'(\langle m_j,\theta\rangle)\right)\right)^2 \\
&= \frac{1}{s^2(\sigma)}\sum_{j=1}^n \left(M_{ji}^2\,\mathrm{Var}(X_j)\right) \\
&\leq \frac{1}{s(\sigma)}\sum_{j=1}^n \left(M_{ji}^2\, L\right) \\
&= \frac{L}{s(\sigma)}[M^\top M]_{ii}. \qquad (11)
\end{aligned}$$

Putting (10) and (11) together, we conclude for any choice of $\epsilon_i > 0$,

$$e^{2h(\theta_i|\hat{\theta}_i)} \geq \epsilon_i^2 \exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^\top M]_{ii}\right)\right]. \qquad (12)$$

In case $\epsilon_i = 0$, we have the trivial equality $e^{2h(\theta_i|\hat{\theta}_i)} = 0$, which is consistent with the RHS of (12) evaluated at $\epsilon_i = 0$. Hence, the estimate (12) holds for all $\epsilon_i \geq 0$.

Summing (12) from $i = 1, 2, \ldots, d$, for parameter $\theta \sim \pi = \prod_{i=1}^d \mathrm{Unif}(-\epsilon_i/2, \epsilon_i/2)$ and any measurable function $\hat{\theta}$ of $X \sim f(\cdot;\theta)$, we have the following lower bound on the Bayesian L_2 risk,

$$\begin{aligned}
\mathbb{E}\|\theta - \hat{\theta}\|_2^2 &\geq \sum_{i=1}^d \mathrm{Var}(\theta_i - \hat{\theta}_i) \\
&\geq \frac{1}{2\pi e}\sum_{i=1}^d e^{2h(\theta_i|\hat{\theta}_i)}
\end{aligned}$$

$$\geq \frac{1}{2\pi e}\sum_{i=1}^{d}\epsilon_i^2 \exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^\top M]_{ii}\right)\right]. \tag{13}$$

It remains to choose an appropriate sequence $(\epsilon_i)_{i=1,2,\ldots,d}$ to obtain the desired lower bound. Toward this end, we consider two cases:

Case 1: $\mathrm{Tr}((M^\top M)^{-1}) \leq \frac{1}{3}\frac{L}{s(\sigma)}$.

In this case, we choose $\epsilon_i^2 = 12\frac{s(\sigma)}{L}\left([M^\top M]_{ii}\right)^{-1}$ for $i = 1, 2, \ldots, d$. Note that by our assumption that $M^\top M$ is diagonal,

$$\sum_{i=1}^{d}\epsilon_i^2 = 12\frac{s(\sigma)}{L}\mathrm{Tr}((M^\top M)^{-1}) \leq 4,$$

so that (9) is satisfied. By an application of (13), we have

$$\begin{aligned}\mathbb{E}\|\theta-\hat{\theta}\|_2^2 &\gtrsim \sum_{i=1}^{d}\epsilon_i^2\exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^\top M]_{ii}\right)\right]\\ &= \frac{12}{e^2}\frac{s(\sigma)}{L}\sum_{i=1}^{d}\frac{1}{[M^\top M]_{ii}}\\ &\gtrsim \frac{s(\sigma)}{L}\mathrm{Tr}((M^\top M)^{-1}).\end{aligned}$$

Case 2: $\mathrm{Tr}((M^\top M)^{-1}) > \frac{1}{3}\frac{L}{s(\sigma)}$.

This case is the more difficult of the two. We shall make use of the following technical Lemma.

Lemma 11. *Let $(a_i)_{i=1,2,\ldots,d}$ be any positive sequence satisfying $\sum_{i=1}^{d} a_i^{-1} > 4$. Then, there exists a non-negative sequence $(\epsilon_i)_{i=1,2,\ldots,d}$ such that $\sum_{i=1}^{d}\epsilon_i^2 \leq 4$ and $\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} \geq 2e^{-2}$.*

Proof. Without loss of generality, assume that $a_1 \geq a_2 \geq \cdots \geq a_d > 0$. If $a_1 \leq 1/4$, then taking $(\epsilon_1, \epsilon_2, \ldots, \epsilon_d) = (2, 0, 0, \ldots, 0)$, and noticing that ϕ is an increasing function, we conclude

$$\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_1)} = 4e^{-2\phi(4a_1)} \geq 4e^{-2\phi(1)} > 2e^{-2}.$$

Now, in the following we assume that $a_1 > 1/4$. Let t denote the largest integer $k \in \{1, 2, \ldots, d\}$ satisfying $\sum_{i=1}^{k} a_i^{-1} \leq 4$. By the assumption that $\sum_{i=1}^{d} a_i^{-1} > 4$, we know that there always exists such a t, and t will satisfy $t < d$. We set

$$\epsilon_i = \begin{cases} a_i^{-1/2} & \text{if } 1 \leq i \leq t \\ 0 & \text{otherwise}\end{cases} \qquad i = 1, 2, \ldots, d. \tag{14}$$

By definition, $\sum_{i=1}^{d}\epsilon_i^2 = \sum_{i=1}^{t} a_i^{-1} \leq 4$ satisfies (9). This procedure results in

$$\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} = e^{-2}\sum_{i=1}^{t}\frac{1}{a_i}.$$

If $\sum_{i=1}^{t} a_i^{-1} \geq 2$, we can immediately see from the above and (14) that $\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} \geq 2e^{-2}$.

On the other hand, if $\sum_{i=1}^{t} a_i^{-1} < 2$, this implies that $a_{t+1}^{-1} \geq 2$. In this case, we simply take $\epsilon_{t+1} = 2$, and take $\epsilon_i = 0$ for $i \neq t+1$. With this choice, we have

$$\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} = 4e^{-2\phi(4a_{t+1})} \geq 4e^{-2\phi(2)} = 2e^{-2}.$$

The above discussion concludes the proof of Lemma 11. □

By considering the values $a_i = \frac{L}{12s(\sigma)}[M^\top M]_{ii}$, Lemma 11 ensures the existence of $(\epsilon_i)_{i=1,2,\ldots,d}$ satisfying (9) and, together with (13), gives $\mathbb{E}\|\theta-\hat{\theta}\|_2^2 \gtrsim 1$. This completes the proof of Theorem 3.

C. Remarks

A few remarks are in order. First, we note that the argument in the previous subsection actually yields the stronger entropic inequality,

$$\inf_{\hat{\theta}}\sup_{\theta\sim\pi}\sum_{i=1}^{d} e^{2h(\theta_i|\hat{\theta}_i)} \gtrsim \min\left(\frac{s(\sigma)}{L}\mathrm{Tr}\left((M^\top M)^{-1}\right), 1\right)$$

which improves Theorem 3 (seen by the max-entropy property of gaussians). Here, the supremum is taken over all distributions π supported on the L_2 ball $\mathbb{B}_2^d(1)$.

Second, we remark that our analysis is flexible enough for generalizations to other forms of the GLM. For example, consider observation X drawn from the density

$$f(x;\theta) = \prod_{i=1}^{n}\left\{h_i(x_i)\exp\left(\frac{x_i\langle m_i,\theta\rangle - \Phi_i(\langle m_i,\theta\rangle)}{s_i(\sigma)}\right)\right\}.$$

Suppose Assumption 1 holds for each cumulant function Φ_i (i.e., $\Phi_i'' \leq L$ for each $i = 1, \ldots, n$). Then, a slight modification in (11) yields

$$[\mathcal{I}_X(\theta)]_{ii} \leq \frac{L}{s^*(\sigma)}[M^\top M]_{ii}$$

where $s^*(\sigma) = \min_{i=1,2,\ldots,n} s_i(\sigma)$. Following (13) and the same choice of $(\epsilon_i)_{i=1,2,\ldots,d}$ in Section II-B with the argument $s(\sigma)$ replaced by $s^*(\sigma)$, we obtain minimax lower bound

$$\inf_{\hat{\theta}}\sup_{\theta\in\mathbb{B}_2^d(1)}\mathbb{E}\|\theta-\hat{\theta}\|_2^2 \gtrsim \min\left(\frac{s^*(\sigma)}{L}\mathrm{Tr}\left((M^\top M)^{-1}\right), 1\right).$$

In the special case where $s_1(\sigma) = \ldots = s_n(\sigma)$, the same minimax lower bound as Theorem 3 is recovered.

ACKNOWLEDGEMENTS

The authors thank Ashwin Pananjady for useful discussions. This work was supported in part by NSF grants CCF-1704967, CCF-0939370 and CCF-1750430.

REFERENCES

[1] F. Abramovich and V. Grinshtein, "Model Selection and Minimax Estimation in Generalized Linear Models," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3721–3730, 2016.

[2] H.-G. Müller and U. Stadtmüller, "Generalized Functional Linear Models," *the Annals of Statistics*, vol. 33, no. 2, pp. 774–805, 2005.

[3] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A Unified Framework for High-Dimensional Analysis of M-estimators with Decomposable Regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[4] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 559–616, 2015.

[5] J. A. Nelder and R. W. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[6] P. Rigollet, "Kullback–Leibler Aggregation and Misspecified Generalized Linear Models," *The Annals of Statistics*, vol. 40, no. 2, pp. 639–665, 2012.

[7] E. Cantoni and E. Ronchetti, "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1022–1030, 2001.

[8] P. McCullagh, *Generalized Linear Models*. Routledge, 2019.

[9] S. A. Van de Geer, "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[10] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax Optimal Procedures for Locally Private Estimation," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–201, 2018.

[11] H. Akaike, "Information Theory and An Extension of the Maximum Likelihood Principle," in *Selected papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.

[12] L. Birgé and P. Massart, "Minimal Penalties for Gaussian Model Selection," *Probability theory and related fields*, vol. 138, no. 1-2, pp. 33–73, 2007.

[13] N. Verzelen, "Minimax Risks for Sparse Regressions: Ultra-High Dimensional Phenomenons," *Electronic Journal of Statistics*, vol. 6, pp. 38–90, 2012.

[14] E. Aras, K.-Y. Lee, A. Pananjady, and T. A. Courtade, "A Family of Bayesian Cramér-Rao Bounds, and Consequences for Log-Concave Priors," *ISIT*, 2019.

[15] R. D. Gill and B. Y. Levit, "Applications of the van Trees Inequality: a Bayesian Cramér-Rao Bound," *Bernoulli*, vol. 1, no. 1-2, pp. 59–79, 1995.

[16] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, 1968.

[17] S. Y. Efroimovich, "Information Contained in a Sequence of Observations," *Problems in Information Transmission*, vol. 15, pp. 24–39, 1980.