A local stochastic Lipschitz condition with application to Lasso for high dimensional generalized linear models

Short title: LSL and Lasso

Zhiyi Chi
Department of Statistics
215 Glenbrook Road, U-4120
Storrs, CT 06269, USA

September 7, 2010

Abstract

For regularized estimation, the upper tail behavior of the random Lipschitz coefficient associated with empirical loss functions is known to play an important role in the error bound of Lasso for high dimensional generalized linear models. The upper tail behavior is known for linear models but much less so for nonlinear models. We establish exponential type inequalities for the upper tail of the coefficient and illustrate an application of the results to Lasso likelihood estimation for high dimensional generalized linear models.

AMS 2010 subject classification. 62G08; 60E15.

Key words and phrases. Lasso, sparsity, measure concentration, generalized linear models, non-convex.

Acknowledgement. Research partially supported by NSF grant DMS-07-06048.

1 Introduction

Let (Y_1, Z_1), ..., (Y_N, Z_N) be independent random variables taking values in a product measurable space $\mathcal{Y} \times \mathcal{Z}$, with Y_i being regarded as response variables and Z_i as covariates. In order to cover both random designs and fixed designs, (Y_i, Z_i) are not necessarily identically distributed. A large class of Lasso type estimators for high dimensional generalized linear models can be formulated as

$$\widehat{\theta} = \operatorname*{arg\,min}_{v \in D_0} \left\{ \sum_{i \le N} [\gamma_i(h(Z_i)^\top v, Y_i) + b(v)] + \sum_{j \le p} \lambda_j |v_j| \right\}, \tag{1}$$

where $D_0 \neq \emptyset$ is a domain in $\mathbb{R}^p$, $\gamma_i(t, y)$ are a given set of real valued functions on $\mathbb{R} \times \mathcal{Y}$, oftentimes identical to each other, $h = (h_1, \ldots, h_p) : \mathcal{Z} \to \mathbb{R}^p$ and $b : D_0 \to \mathbb{R}$ are given functions, and $\lambda_1, \ldots, \lambda_p > 0$ are coefficients of the weighted ℓ_1 penalty on v. In this article, we only consider nonadaptive Lasso, in which $\lambda_1, \ldots, \lambda_N$ are fixed beforehand.

Under the setting of (1), for each $v \in D_0$, we have N loss functions, each defined as $(y, z) \to \gamma_i(h(z)^\top v, y) + b(v)$. The corresponding empirical losses are $\gamma_i(h(Z_i)^\top v, Y_i) + b(v)$, and the corresponding expected total loss is

$$L(v) = \sum_{i \le N} \mathsf{E}[\gamma_i(h(Z_i)^\top v, Y_i) + b(v)], \quad v \in D_0. \tag{2}$$

As the title suggests, the main interest of the article is the so called "local stochastic Lipschitz" (LSL) condition. By LSL we mean the following. For the time being, denote by

$$\tilde{L}(v) = \sum_{i \le N} [\gamma_i(h(Z_i)^\top v, Y_i) + b(v)] - L(v)$$

the fluctuation of the empirical total loss from its expectation at parameter value v. Let $\theta \in \mathbb{R}^p$ be fixed. Under smooth conditions for γ_i, it is easy to see $\tilde{L}(v)$ is differentiable with probability (w.p.) 1, which in general leads to Lipschitz continuity of $\tilde{L}(v)$ provided D_0 is compact. The LSL condition, on the other hand, refers to a bound on the upper tail probability of the random variable

$$\sup_{v \in D_0,\, v \neq \theta} \frac{|\tilde{L}(v) - \tilde{L}(\theta)|}{\sum_{j \leq p} \lambda_j |v_i - \theta_j|}. \tag{3}$$

Note that the LSL condition is with respect to a weighted ℓ_1 norm of $\mathbb{R}^p$. The condition is called "local" because θ is fixed, even though its value is typically unknown.

Although it might not be apparent at this point, the LSL condition is closely related to the issue of estimation error for Lasso. For linear regression with square loss function $(y - h(z)^\top v)^2$, this relationship is well known and has been regularly employed to obtain estimation error bounds [4, 3, 2, 1]. Indeed, in this case, due to linearity, the LSL condition is rather easy to establish. However, for other loss functions, the LSL condition is much less clear and, to my best knowledge, has not been fully explored. An alternative to the LSL condition is a convexity assumption, in which $\gamma_i(t, y)$ is convex in t and $b(v)$ is convex in v. The convexity assumption allows a linear interpolation technique to be employed to yield upper bounds for estimation error [12]. While the convexity assumption allows for nondifferentiable γ_i, it is not clear how the technique can be extended to nonconvex loss functions.

We shall establish the LSL condition for general loss functions. For differentiability, we only require that $\gamma_i(t, y)$ be first order differentiable in t with the partial derivative being Lipschitz. After getting various results on the LSL condition, we will then illustrate an application of the LSL condition to Lasso type nonlinear regression, by finding an upper bound for the ℓ_2 norm of estimation error.

Previously, in [6], the LSL condition was studied for loss functions of the form $(y - g_i(h(z)^\top v))^2$, $i \leq N$, where $g_i : \mathbb{R} \to \mathbb{R}$ are nonlinear. The condition was established under the assumptions that g_i are twice continuously differentiable and

$$Y_i = g_i(h(Z_i)^\top \theta) + \varepsilon_i, \tag{4}$$

where ε_i are uniformly bounded zero mean noise. In this article, we extend the result on two aspects. First, the LSL condition is established for general $\gamma_i(t, y)$, while still under the assumption of uniform boundedness. Second, it is established for (4) when ε_i are Gaussian. Whereas the bounds for general $\gamma_i(t, y)$ is of Bernstein type, the bounds for the Gaussian case is of Hoeffding type. In [6], a truncation argument was suggested for the Gaussian case. However, the LSL condition obtained in this way is not as tight as the one to be obtained here. The tools used to get the results on the LSL condition are various measure concentration and comparison inequalities in Probability [9, 8, 7].

Section 2 presents several results on the LSL condition. The discussion in the section is actually more general. It provides upper bounds on the tail probability of the remainder of the Taylor expansion of $\tilde{L}(v)$. The LSL condition is a simple consequence of these bounds.

In Section 3, we consider an application of the LSL condition to Lasso. Besides the LSL condition, Lasso involves another issue, that is, the amount of separation of v and θ based on the difference between $\gamma_i(h(Z_i)^\top v, Y_i)$ and $\gamma_i(h(Z_i)^\top \theta, Y_i)$. This issue is of different nature from the LSL condition, and its resolution in general requires further conditions on the matrix $[h_j(Z_i)]_{i \leq N, j \leq p}$. The issue has been studied in quite a few works [14, 2, 5, 1, 13]. For transparency, we will use a restricted eigenvalue condition in [1] for our purpose. We will consider an example of Lasso type MLE for high dimensional generalized linear model and apply the LSL condition to bound the ℓ_2 norm of the estimation error. Unfortunately, the method of the example gives no clue on model selection or more elaborate bounds similar to those obtained for linear models under square loss [13, 1, 4]. All the proofs are presented in Section 4.

1.1 Notation

For $q \in [1, \infty)$, denote by $\|a\|_q$ the ℓ_q norm of $a \in \mathbb{R}^d$. For two vectors $a = (a_1, \ldots, a_m)^\top$ and $b = (b_1, \ldots, b_n)^\top$, recall that their tensor product is

$$a \otimes b = (a_1 b^\top, \ldots, a_m b^\top)^\top = (a_1 b_1, \ldots, a_1 b_n, \ldots, a_n b_1, \ldots, a_m b_n)^\top \in \mathbb{R}^{mn}.$$

Denote by $v^{\otimes k}$ the tensor product of k copies of v.

If f is a function on a domain $\Omega \subset \mathbb{R}^d$, then it is Lipschitz (under the Euclidean norm) if

$$\|f\|_{\mathrm{Lip}} := \sup_{x \neq y \in \Omega} \frac{|f(x) - f(y)|}{\|x - y\|_2} < \infty.$$

Finally, for any random vector X, denote its deviation from mean by

$$[\![X]\!] = X - \mathsf{E}X.$$

By linearity of expectation, $[\![X + Y]\!] = [\![X]\!] + [\![Y]\!]$. By this notation,

$$\tilde{L}(v) = \sum_{i \leq N} \left[\![\gamma_i(h(Z_i)^\top v, Y_i)\right]\!] .$$

The right hand side is independent of $b(v)$ and at the same time better reveals the other quantities involved. We will discard the notation $\tilde{L}$ in favor of $[\![\cdot]\!]$ for the rest of the article.

1.2 Notes

The methods in Section 2 can be used with little change to deal with the following additive mixture of loss functions,

$$\sum_{i \leq N} \sum_{k \leq q} [\gamma_{ik}(h_k(Z_i)^\top v, Y_i) + b_k(v)]$$

where for each $k \leq q$ and $i \leq N$, $h_k = (h_{k1}, \ldots, h_{kp})$ is a function from $\mathcal{Z}$ to $\mathbb{R}^p$, and γ_{ik} is a loss function. For example

$$\sum_{i \leq N} \gamma_i(h(Z_i)^\top v, Y_i) + \sum_{i \leq N} \tilde{\gamma}_i(\tilde{h}(\tilde{Z}_i)^\top u, Y_i)$$

is a special case of additive mixture, where Z_i and $\tilde{Z}_i$ are covariates that may be identical or have completely different sets of coordinates. Due to identifiability issue in the context of parameter estimation, such mixtures will not further considered in the article.

2 Local stochastic Lipschitz condition

In this section, we present exponential bounds on the tail probability of the random local Lipschitz coefficient (3). As noted earlier, these bounds are consequences of more general results on the tail probability of remainders of Taylor expansion of random functions. Therefore, most of the discussion below will be on the latter and the results on the LSL condition will be given as corollaries.

2.1 General loss function

Suppose $\gamma_1, \ldots, \gamma_N$ satisfy the following regularity condition.

Assumption 1 (Regularity)**.** *There are $m \in \{0, 1, 2, \ldots\}$ and $-\infty \le a_i < b_i \le \infty$, $i \le N$, such that w.p. 1, each $\gamma_i(t, Y_i)$ as a function of t is m times differentiable on (a_i, b_i) with the m-th derivative being bounded and Lipschitz. Let F_m, F_{m+1} be constants such that w.p. 1,*

$$\begin{cases} \left| \dfrac{\partial^m \gamma_i(t, Y_i)}{\partial t^m} \right| \le F_m, \\ \left| \dfrac{\partial^m \gamma_i(t, Y_i)}{\partial t^m} - \dfrac{\partial^m \gamma_i(t', Y_i)}{\partial t^m} \right| \le F_{m+1} |t - t'|, \end{cases} \qquad \forall\, t, t' \in (a_i, b_i),\ i \le N.$$

Suppose h satisfies the following condition.

Assumption 2 (Boundedness)**.** *There are constants $d_1, \ldots, d_p \in (0, \infty)$, such that*

$$\mathsf{Pr}\left\{ \max_{i \le N} |h_j(Z_i)| \le d_j,\ \forall\, j \le p \right\} = 1.$$

Next, let $D_0 \ne \emptyset$ be a domain in $\mathbb{R}^p$.

Assumption 3 (Parameter Domain)**.** *For (a_i, b_i) as in Assumption 1 and h as in Assumption 2,*

$$\mathsf{Pr}\left\{ h(Z_i)^\top v \in (a_i, b_i), \forall\, v \in D_0,\ i \le N \right\} = 1.$$

From Assumption 1 and dominated convergence, differentiation and expectation can be exchanged for $\gamma_i(t, Y_i)$, i.e.,

$$\mathsf{E}\left[\frac{\partial^k \gamma_i(t, Y_i)}{\partial t^k} \right] = \frac{\partial^k \mathsf{E}[\gamma_i(t, Y_i)]}{\partial t^k}, \quad t \in (a_i, b_i),\ i \le N,\ k \le m.$$

By Assumption 2, $|h_j(Z_i)/d_j| \le 1$ w.p. 1. Therefore, d_j can be thought of as the "scales" of the functions h_j.

Theorem 2.1. *Under Assumptions 1 – 3, fix an arbitrary $\theta \in D_0$. Then for $v \in D_0$,*

$$\begin{aligned} &\sum_{i \le N} [\![\gamma_i(h(Z_i)^\top v, Y_i)]\!] \\ &= \sum_{k \le m} \frac{1}{k!} \sum_{i \le N} \left[\!\!\left[\frac{\partial^k \gamma_i(h(Z_i)^\top \theta, Y_i)}{\partial t^k} [h(Z_i)^\top (v - \theta)]^k \right]\!\!\right] + \xi(v) \left(\sum_{j \le p} d_j |v_j - \theta_j| \right)^m \qquad (5) \\ &= \sum_{k \le m} \frac{1}{k!} \sum_{i \le N} \left[\!\!\left[\frac{\partial^k \gamma_i(h(Z_i)^\top \theta, Y_i)}{\partial t^k} h(Z_i)^{\otimes k} \right]\!\!\right]^\top (v - \theta)^{\otimes k} + \xi(v) \left(\sum_{j \le p} d_j |v_j - \theta_j| \right)^m, \qquad (6) \end{aligned}$$

where $\{\xi(v), v \in D_0\}$ is a process that has the following upper tail property

$$\mathsf{Pr}\left\{ \sup_{v \in D_0} |\xi(v)| > A\sqrt{2\ln(2p)} + B\sqrt{2\ln(p^m/q)} + C\ln(p^m/q) \right\} \le q, \quad \forall\, q \in (0, 1)$$

with A, B, and C being set as follows. First, let

$$R = \sup_{u, v \in D_0} \sum_{j \le p} d_j |u_j - v_j|, \quad \phi = \min\left(\frac{2F_m}{m!}, \frac{F_{m+1} R}{(m+1)!} \right), \quad \psi = \begin{cases} F_{m+1}/m! & m \ne 1 \\ F_{m+1}/2 & m = 1. \end{cases} \qquad (7)$$

Then

$$A = 8\psi R \mathsf{E} \sqrt{\max_{j \le p} \sum_{i \le N} [h_j(Z_i)/d_j]^2}, \quad B = \phi \sqrt{\mathsf{E} \max_{j \le p} \sum_{i \le N} [h_j(Z_i)/d_j]^{2m}}, \quad C = 8\phi,$$

where in the definition of B the convention $x^0 \equiv 1$ is used for $m = 0$.

Note that if $F_{m+1} > 0$, then the above result is meaningful only when $R < \infty$, that is, D_0 is bounded. On the other hand, if w.p. 1, for $i \le N$, $\gamma_i(t, Y_i)$ is a linear function of t, then one can set $F_{m+1} = 0$. By Theorem 2.1, this yields $A = B = C = 0$, which implies $\xi(v) \equiv 0$. Of course, the last fact is easy to be seen by the linearity of $\gamma_i(t, Y_i)$.

Of particular interest is the case where $m = 1$. From Theorem 2.1, the following result obtains.

Corollary 2.2. *Under Assumptions 1 – 3 with $m = 1$, fix an arbitrary $\theta \in D_0$. Then for $v \in D_0$,*

$$\sum_{i \le N} [\![\gamma_i(h(Z_i)^\top v, Y_i)]\!] = \sum_{i \le N} [\![\gamma_i(h(Z_i)^\top \theta, Y_i)]\!] + [\xi_1 + \xi(v)] \sum_{j \le p} d_j |v_j - \theta_j| \tag{8}$$

where $\xi(v)$ is as in Theorem 2.1 and ξ_1 is a random variable with the following upper tail property

$$\mathsf{Pr}\left\{|\xi_1| > F_1 \sqrt{2N \ln(2p/q)}\right\} \le q, \quad \forall\, q \in (0, 1).$$

Since

$$|\xi_1| + \sup_{v \in D_0} |\xi(v)| \ge \sup_{v \in D_0,\, v \ne \theta} \frac{1}{\sum_{j \le p} \lambda_j |v_i - \theta_j|} \left| \sum_{i \le N} [\![\gamma_i(h(Z_i)^\top v, Y_i) - \gamma_i(h(Z_i)^\top \theta, Y_i)]\!] \right|,$$

from the result, we then get a desired form of the LSL condition. For any q, $q' \in (0, 1)$ not necessarily equal, one can find $M(q, q')$, such that w.p. at least $1 - q - q'$, the random local Lipschitz coefficient on the right hand side is no greater than $M(q, q')$. Moreover, one can set

$$M(q, q') = A\sqrt{2 \ln(2p)} + B\sqrt{2 \ln(p/q)} + C \ln(p/q) + F_1 \sqrt{2N \ln(2p/q')},$$

with A, B and C given as in Theorem 2.1 with $m = 1$.

2.2 Gaussian case

Suppose $Z_1, \ldots, Z_N$ are fixed and

$$Y_i = \mu_i - \omega_i$$

where μ_i are some unknown constants, and $\omega_1, \ldots, \omega_N$ are independent square-integrable random variables with mean 0. Let $f_1, \ldots, f_N : \mathbb{R} \to \mathbb{R}$ be a set of transforms specified beforehand, and $h = (h_1, \ldots, h_p) : \mathcal{Z} \to \mathbb{R}^p$ a measurable function. Suppose the goal is to use $f_i(h(Z_i)^\top v)$ to approximate μ_i under the square loss functions

$$\gamma_i(t, Y_i) = (Y_i - f_i(t))^2/2. \tag{9}$$

For any v, provided that $h(Z_i)^\top v$ is in the domain of f_i for all $i \le N$,

$$\begin{aligned} [\![\gamma_i(h(Z_i)^\top v, Y_i)]\!] &= \frac{1}{2}(\mu_i - \omega_i - f_i(h(Z_i)^\top v))^2 - \frac{1}{2}\mathsf{E}[(\mu_i - \omega_i - f_i(h(Z_i)^\top v))^2] \\ &= \omega_i[f_i(h(Z_i)^\top v) - \mu_i] + \frac{1}{2}[\omega_i^2 - \mathsf{Var}(\omega_i)]. \end{aligned}$$

Thus, for any θ, provided that $h(Z_i)^\top \theta$ is in the domain of f_i for all $i \le N$ as well

$$\sum_{i \le N} [\![\gamma_i(h(Z_i)^\top v, Y_i)]\!] - \sum_{i \le N} [\![\gamma_i(h(Z_i)^\top \theta, Y_i)]\!] = \sum_{i \le N} \omega_i \left[f_i(h(Z_i)^\top v) - f_i(h(Z_i)^\top \theta)\right].$$

As a result, we will focus on the expansion of the random function

$$v \to \sum_{i \le N} \omega_i f_i(h(Z_i)^\top v)$$

around any fixed $\theta \in D_0$.

Assumption 4 (Regularity)**.** *There are $m \in \{0, 1, 2, \ldots\}$ and $-\infty \le a_i < b_i \le \infty$, $i \le N$, such that each f_i is m times differentiable on (a_i, b_i) with the m-th derivative being bounded and Lipschitz. Let*

$$F_m = \max_{i \le N} \sup_{t \in (a_i, b_i)} |f_i^{(m)}(t)|, \quad F_{m+1} = \max_{i \le N} \left\| f_i^{(m)} \right\|_{\mathrm{Lip}}.$$

Since Z_i are fixed, Assumption 2 is no longer needed. Instead, simply define

$$d_j = \max_{i \le N} |h_j(Z_i)|.$$

Also, modify Assumption 3 as follows.

Assumption 5 (Parameter Domain)**.** *The domain $D_0 \neq \emptyset$ of candidate parameter values satisfies $h(Z_i)^\top v \in (a_i, b_i)$, $\forall v \in D_0$, $i \le N$.*

In [6], the case where ω_i are uniformly bounded is considered. Here we shall deal with the following situation.

Assumption 6 (Gaussian)**.** *$\omega_1, \ldots, \omega_N$ are independent Gaussian variables with $\mathsf{Var}(\omega_i) \le \sigma_0^2$, $i \le N$, where $\sigma_0 \in (0, \infty)$ is a constant.*

Theorem 2.3. *Let the loss functions $\gamma_1, \ldots, \gamma_N$ be as in (9). Under Assumptions 4 – 6, fix an arbitrary $\theta \in D_0$. Then for $v \in D_0$,*

$$\begin{aligned}
&\sum_{i \le N} \omega_i f_i(h(Z_i)^\top v) \\
&= \sum_{k \le m} \frac{1}{k!} \left(\sum_{i \le N} \omega_i f_i^{(k)}(h(Z_i)^\top \theta)[h(Z_i)^\top (v - \theta)]^k \right) + \xi(v) \left(\sum_{j \le p} d_j |v_j - \theta_j| \right)^m \qquad (10) \\
&= \sum_{k \le m} \frac{1}{k!} \left(\sum_{i \le N} \omega_i f_i^{(k)}(h(Z_i)^\top \theta) h(Z_i)^{\otimes k} \right)^\top (v - \theta)^{\otimes k} + \xi(v) \left(\sum_{j \le p} d_j |v_j - \theta_j| \right)^m, \qquad (11)
\end{aligned}$$

where $\{\xi(v), v \in D_0\}$ is a process that has the following upper tail property

$$\mathsf{Pr}\left\{ \sup_{v \in D_0} |\xi(v)| > \sigma_0 (A\sqrt{\ln(2p)} + B\sqrt{2\ln(p^m/q)}) \right\} \le q, \quad \forall\, q \in (0, 1)$$

with A and B being set as follows. First, set R, ϕ and ψ as in (7). Then

$$A = 8\psi R \sqrt{\max_{j \le p} \sum_{i \le N} [h_j(Z_i)/d_j]^2}, \quad B = \phi \sqrt{\max_{j \le p} \sum_{i \le N} [h_j(Z_i)/d_j]^{2m}},$$

where in the definition of B the convention $x^0 \equiv 1$ is used for $m = 0$.

Comparing to Theorem 2.1, the above upper tail bound does not have a term of the form $C \ln(p^m/q)$. This is because in the Gaussian case, we can get a Hoeffding type inequality for the upper tail instead of a Bernstein type inequality.

From Theorem 2.3, the following result for the case $m = 1$ obtains. Note that the result is not entirely the same as Corollary 2.2.

Corollary 2.4. *Under Assumptions 4 – 6 with $m = 1$, fix an arbitrary $\theta \in D_0$. Define positive constants $w_1, \ldots, w_p$ as*

$$w_j^2 = \sigma_0^{-2} \sum_{i \le N} \mathsf{Var}(\omega_i) h_j(Z_i)^2. \qquad (12)$$

Then for $v \in D_0$,

$$\sum_{i\le N} \omega_i f_i(h(Z_i)^\top v) = \sum_{i\le N} \omega_i f_i(h(Z_i)^\top \theta) + \sigma_0 F_1 \xi_1 \sum_{j\le p} w_j |v_j - \theta_j| + \xi(v) \sum_{j\le p} d_j |v_j - \theta_j| \tag{13}$$

where $\{\xi(v) : v \in D_0\}$ is as in Theorem 2.3 and ξ_1 is a random variable with the following upper tail property

$$\mathsf{Pr}\left\{|\xi_1| > \sqrt{2\ln(p/q)}\right\} \le q, \quad \forall\, q \in (0,1).$$

Similar to Corollary 2.2, the above result can be used to get the LSL condition. For example, for any q, $q' \in (0,1)$, one can set

$$M(q,q') = \sigma_0 \left[A\sqrt{\ln(2p)} + B\sqrt{2\ln(p/q)} + F_1\sqrt{2\ln(p/q')}\right],$$

with A and B given as in Theorem 2.3 with $m = 1$, such that w.p. at least $1 - q - q'$, the following random local Lipschitz coefficient

$$\sup_{v\in D_0,\, v\ne\theta} \frac{1}{\sum_{j\le p}\lambda_j |v_i - \theta_j|} \left| \sum_{i\le N} \omega_i [f_i(h(Z_i)^\top v) - f_i(h(Z_i)^\top \theta)] \right|$$

is no greater than $M(q,q')$, where $\lambda_j = \max(w_j, d_j)$.

3 An application to high dimensional Lasso

Under Assumptions 1 – 3, we consider the case where $Z_1, \ldots, Z_N$ are fixed. For simplicity, assume $d_1 = \ldots = d_N = d$ in Assumption 2. Consider the following Lasso functional

$$\widehat{\theta} = \arg\min_{v\in D_0} \left\{ \sum_{i\le N} \gamma_i(h(Z_i)^\top v, Y_i) + \lambda d \|v\|_1 \right\}, \tag{14}$$

where $\lambda > 0$ is the tuning parameter. Suppose D_0 is compact so that the minimum is always obtained. The goal is to have $\widehat{\theta}$ approximate to θ, where

$$\theta = \arg\min_{v\in D_0} \sum_{i\le N} \mathsf{E}[\gamma_i(h(Z_i)^\top v, Y_i)].$$

We next consider applying Corollary 2.2 to bound $\|\widehat{\theta} - \theta\|_2$. Denote $X_i = h(Z_i)$ and X the $N \times p$ matrix with $X_i^\top$ as the i-th row vector. The total expected loss function now can be written as

$$L(v) = \sum_{i\le N} \mathsf{E}[\gamma_i(X_i^\top v, Y_i)], \quad v \in D_0.$$

Denote by $\mathrm{spt}(v) = \{j \le p : v_j \ne 0\}$ and by $\|v\|_0$ the cardinality of the set. In general, in order to bound $\|\widehat{\theta} - \theta\|_2$, some conditions on X are needed in order to get a bound in terms of the ℓ_2 norm of $v - \theta$ (cf. [13, 4, 1]). For transparency, we use a "restricted eigenvalue" condition formulated in [1], which says that for some $1 \le s \le p$ and $c > 0$,

$$\kappa(s,K) := \min\left\{ \frac{\|Xv\|_2}{\sqrt{N}\|v_J\|_2} : 1 \le |J| \le s,\ v \ne 0, \|v_{J^c}\|_1 \le K\|v_J\|_1 \right\} > 0.$$

To see where the LSL condition is to be used, we first summarize an argument that has been more or less used for special cases of Lasso (cf. [12, 1]). Note that the argument does not lead to model selection or more elaborate bounds that have been obtained especially for linear models under square loss [5, 13, 1, 4].

Theorem 3.1. *Suppose the following conditions are satisfied.*

1) For some $K > 1$,

$$\kappa := \kappa(2\|\theta\|_0, K) > 0 \tag{15}$$

2) For some $C_\gamma > 0$,

$$L(v) - L(\theta) \geq C_\gamma \|X(v-\theta)\|_2^2, \quad \forall v \in D_0. \tag{16}$$

3) Given $q \in (0,1)$, suppose there is $M_q > 0$, such that w.p. at least $1-q$,

$$\left| \sum_{i \leq N} \left[\!\left[\gamma_i(X_i^\top \widehat{\theta}, Y_i) - \gamma_i(X_i^\top \theta, Y_i) \right]\!\right] \right| \leq M_q d \|\widehat{\theta} - \theta\|_1. \tag{17}$$

Then, by setting

$$\lambda = \frac{(K+1)M_q d}{K-1} \tag{18}$$

in the Lasso functional (14)*, on the event that* (17) *holds,*

$$\|\widehat{\theta} - \theta\|_2 \leq \frac{M_q \sqrt{\|\theta\|_0}}{N} \times \frac{2\sqrt{2+K^2} K d}{C_\gamma \kappa^2 (K-1)} \tag{19}$$

Theorem 3.1 has three conditions. The first one is the aforementioned restricted eigenvalue condition. In some cases, the second condition is easy to establish. The third condition is the LSL condition. By Corollaries 2.2 and 2.4, M_q can be set reasonably small, ideally of order $\sqrt{N}$ or even smaller.

Example 3.1. Let $\mathcal{Y}$ be a Euclidean space and $\mathcal{F} = \{f(y \,|\, t) : y \in \mathcal{Y}, t \in [a,b]\}$ a family of densities on $\mathcal{Y}$, where $-\infty < a < b < \infty$. Suppose given Z_i, the density of Y_i is

$$f(y \,|\, X_i^\top \theta)$$

where θ is the parameter and X_i again is $h(Z_i)$. Suppose it is known that $\theta \in D_0$, where $D_0 \subset \mathbb{R}^p$ is an open bounded region such that for $v \in D_0$, $X_i^\top v \in [a,b]$ for each $i \leq N$. Then any solution $\widehat{\theta}$ to (14) with

$$\gamma_i(t,y) = -\ln f(y \,|\, t) := \ell(t,y), \quad i \leq N \tag{20}$$

is an ℓ_1 regularized MLE of θ. Suppose X satisfies (15). We next find some conditions in order for (16) to hold. Let $I(t)$ denote the Fisher information of $\mathcal{F}$ at t and

$$D(t,s) = \int f(y \,|\, t) \ln \frac{f(y \,|\, t)}{f(y \,|\, s)} \, \mathrm{d}y = \mathsf{E}[\ell(s,Y)] - \mathsf{E}[\ell(t,Y)], \quad Y \sim f(y \,|\, t),$$

the Kullback-Leibler distance from $f(y \,|\, s)$ to $f(y \,|\, t)$. For $\mathcal{F}$ with enough regularity, it is not hard to show D has the following properties:

1) D, $\partial D/\partial s$, $\partial^2 D/\partial s^2$ are continuous in (t,s);

2) $D(t,t) = (\partial D/\partial s)(t,t) = 0$, $I(t) = (\partial^2 D/\partial s^2)(t,t) > 0$;

3) every $t \in [a,b]$ is identifiable in $\mathcal{F}$; and

4) as $h \to 0$, $D(t, t+h)/h^2 \to I(t)$ uniformly for $t \in [a, b]$.

Property 2) implies that for s in a neighborhood of t, $D(t, s) \geq I(t)(t-s)^2/2$. Together with the other three properties and the compactness of $[a, b] \times [a, b]$, for some $C_{\mathcal{F}} > 0$, $D(t, s) \geq C_{\mathcal{F}}(t-s)^2$ for all t, s. Now for $i \leq N$ and $v \in D_0$, since Y_i has density $f(y \mid X_i^\top \theta)$,

$$\mathsf{E}[\gamma_i(X_i^\top v, Y_i)] - \mathsf{E}[\gamma_i(X_i^\top \theta, Y_i)] = D(X_i^\top \theta, X_i^\top v) \geq C_{\mathcal{F}} |X_i^\top \theta - X_i^\top v|^2.$$

Then by the definition of $L(v)$,

$$L(v) - L(\theta) \geq C_{\mathcal{F}} \sum_{i \leq N} |X_i^\top (v - \theta)|^2 = C_{\mathcal{F}} \|X(v - \theta)\|_2^2,$$

so (16) is satisfied.

Finally, if γ_i defined in (20) satisfies Assumptions 1 – 3, then by Corollary 2.2 and Theorem 3.1, given $q_1, q_2 \in (0, 1)$ with $q_1 + q_2 < 1$, the following bound

$$\|\widehat{\theta} - \theta\|_2 \leq \frac{(M_1 + M_2)\sqrt{\|\theta\|_0}}{N} \times \frac{2\sqrt{2 + K^2} K d}{C_{\mathcal{F}} \kappa^2 (K - 1)}$$

holds with probability at least $1 - q_1 - q_2$, where M_1 and M_2 are as follows. Denote by $V_1, \ldots, V_p$ the column vectors of X and

$$\Delta = \sup_{u, v \in D_0} \|u - v\|_1.$$

Denote

$$F_1 = \operatorname{ess\,sup} \left(\sup_t \max_{i \leq N} \left| \dot{\ell}(t, Y_i) \right| \right), \quad F_2 = \operatorname{ess\,sup} \left(\max_{i \leq N} \left\| \dot{\ell}(\cdot, Y_i) \right\|_{\mathrm{Lip}} \right).$$

Note $d_1 = \ldots = d_N = d$. Then

$$M_1 = A\sqrt{2\ln(2p)} + B\sqrt{2\ln(p/q_1)} + 8\phi \ln(p/q_1), \quad M_2 = F_1 \sqrt{2N \ln(2p/q_2)},$$

where

$$A = 4F_2 \Delta \max_{j \leq p} \|V_j\|_2, \quad B = (F_2/2)\Delta \max_{j \leq p} \|V_j\|_2, \quad \phi = \min(2F_1, F_2 d \Delta/2).$$

Up to a factor of $\sqrt{\ln(p/q_2)}$, $M_2 = O(\sqrt{N})$. Typically, for well designed X, $\max_{j \leq p} \|V_j\|_2 = O(\sqrt{N})$. Therefore, $M_1 = O(\sqrt{N})$ up to a multiplicative factor $\sqrt{\ln(p/q_1)}$ and an additive remainder of order $\ln(p/q_1)$. As a result, $\|\widehat{\theta} - \theta\|_2$ is of order $\sqrt{\|\theta\|_0/N}$ up to factors much smaller than $\sqrt{N}$ unless p is extremely large.

Similar conclusions can be made if $f(y \mid X_i^\top \theta)$ is the density of $N(X_i^\top \theta, \sigma_0^2)$. In this case, we can use Corollary 2.4. For brevity, the detail is omitted. □

4 Proofs

In this section we give proofs for the results in previous sections. First, recall that for $q \in [1, \infty)$,

$$\|a \otimes b\|_q^q = \|a\|_q^q \|b\|_q^q, \tag{21}$$

and for $a_1, a_2 \in \mathbb{R}^m$, $b_1, b_2 \in \mathbb{R}^n$, $(a_1^\top a_2)(b_1^\top b_2) = (a_1 \otimes b_1)^\top (a_2 \otimes b_2)$, giving

$$(a_1^\top a_2)^k = (a_1^{\otimes k})^\top (a_2^{\otimes k}). \tag{22}$$

4.1 Proofs for Section 2

Proof of Theorem 2.1. By (22), (5) and (6) are equivalent. For notational brevity, we shall avoid explicit use of d_j. For this reason, the domain D_0 is not the one to be directly worked on. Rather, we shall consider

$$D = \{(d_1 v_1, \ldots, d_p v_p)^\top : v \in D_0\}. \tag{23}$$

In other words, D is the image of D_0 under the 1-1 transform $T : v \to (d_1 v_1, \ldots, d_p v_p)^\top$. We shall use the ℓ_1 norm on D. Note that the norm induces a weighted ℓ_1 norm on D_0 as

$$\|u - v\| = \|Tu - Tv\|_1 = \sum_{j \le p} d_j |u_j - v_j|,$$

which is the reason why $\sum_{j \le p} d_j |u_j - \theta_j|$ appears in the expansions (5) and (6). Moreover, R in (7) can be expressed as the diameter of D under ℓ_1,

$$R = \sup_{u,v \in D} \|u - v\|_1.$$

Based on the same consideration as (23), denote for $i \le N$, $j \le p$,

$$X_{ij} = h_j(Z_i)/d_j, \quad X_i = (X_{i1}, \ldots, X_{ip})^\top, \quad V_j = (X_{1j}, \ldots, X_{Nj})^\top. \tag{24}$$

Then Assumption 2 on the boundedness of $h_j(Z_i)$ implies

$$\mathsf{Pr}\{|X_{ij}| \le 1, \forall i \le N,\, j \le p\} = 1. \tag{25}$$

Furthermore, for $v \in D$, let $u \in D_0$ such that $Tu = v$. Then $X_i^\top v = h(Z_i)^\top u$, so we can easily translate an expansion in terms of $X_i^\top v$ into one in terms of $h(Z_i)^\top u$. Therefore, until the end of the proof, we will focus on D.

For brevity, for each $i \le N$, denote

$$f_i(t) = \gamma_i(t, Y_i), \quad f_i^{(k)}(t) = \frac{\partial^k \gamma_i(t, Y_i)}{\partial t^k}, \quad k \le m+1.$$

Fix $\theta \in D$. For $i \le N$ and v, define random vectors $c = (c_1, \ldots, c_N)$ and $t = (t_1, \ldots, t_N)$ with

$$c_i = X_i^\top \theta, \quad t_i = X_i^\top (v - \theta).$$

For $i \le N$, let φ_i be the following random function on $\mathbb{R}$,

$$\varphi_i(t) = \begin{cases} t^{-m} \left[f_i(c_i + t) - \sum_{k \le m} \frac{f_i^{(k)}(c_i)}{k!} t^k \right], & t \ne 0; \\ 0, & t = 0. \end{cases} \tag{26}$$

We need the following property of φ_i.

Lemma 4.1. *W.p. 1, each $\varphi_i \in C(a_i - c_i, b_i - c_i)$, and*

$$|\varphi_i(t)| \le \min\left(\frac{2F_m}{m!}, \frac{F_{m+1}|t|}{(m+1)!} \right) \tag{27}$$

and $\|\varphi_i\|_{\mathrm{Lip}} \le \psi$, where

$$\psi = \begin{cases} F_{m+1}/m! & m \ne 1 \\ F_{m+1}/2 & m = 1. \end{cases}$$

Lemma 4.1 will be proved later. Clearly,

$$\sum_{i\le N}\gamma_i(X_i^\top v, Y_i) = \sum_{i\le N} f_i(c_i+t_i) = \sum_{i\le N}\left(\sum_{k\le m}\frac{f_i^{(k)}(c_i)}{k!}t_i^k + \varphi_i(t_i)t_i^m\right)$$
$$= \sum_{k\le m}\frac{1}{k!}\left(\sum_{i\le N} f_i^{(k)}(c_i)t_i^k\right) + \sum_{i\le N}\varphi_i(t_i)t_i^m,$$

where, by Assumption 2, w.p. 1, $t_i = X_i^\top(v-\theta)\in(a_i-c_i, b_i-c_i)$, $\forall i\le N$, $v\in D$. Then by (22),

$$\sum_{i\le N}\gamma_i(X_i^\top v, Y_i) = \sum_{k\le m}\frac{1}{k!}\left(\sum_{i\le N} f_i^{(k)}(c_i)X_i^{\otimes k}\right)^\top (v-\theta)^{\otimes k} + \left(\sum_{i\le N}\varphi_i(t_i)X_i^{\otimes m}\right)^\top (v-\theta)^{\otimes m}.$$

Therefore,

$$\sum_{i\le N}[\![\gamma_i(X_i^\top v, Y_i)]\!] = \sum_{k=1}^{m}\frac{1}{k!}\sum_{i\le N}\left[\!\!\left[f_i^{(k)}(c_i)X_i^{\otimes k}\right]\!\!\right]^\top (v-\theta)^{\otimes k} + \sum_{i\le N}[\![\varphi_i(t_i)X_i^{\otimes m}]\!]^\top (v-\theta)^{\otimes m}. \quad (28)$$

By Hölder inequality and (21),

$$\left|\sum_{i\le N}[\![\varphi_i(t_i)X_i^{\otimes m}]\!]^\top (v-\theta)^{\otimes m}\right| \le \left\|\sum_{i\le N}[\![\varphi_i(t_i)X_i^{\otimes m}]\!]\right\|_\infty \|v-\theta\|_1^m. \quad (29)$$

For each $\jmath = (j_1,\ldots,j_p)$ with $j_s\le p$, denote

$$X_{i\jmath} = X_{ij_1}\cdots X_{ij_m},$$

where the product on the right hand side is defined to be 1 if $m=0$. Then the coordinates of $X_i^{\otimes m}$ can be written as $X_{i\jmath}$, with $\jmath$ sorted, say, in the dictionary order. Let

$$Z_\jmath = \sup_{v\in D}\left|\sum_{i\le N}[\![\varphi_i(t_i)X_{i\jmath}]\!]\right|.$$

Then from (29),

$$\left|\sum_{i\le N}[\![\varphi_i(t_i)X_i^{\otimes m}]\!]^\top (v-\theta)^{\otimes m}\right| \le \|v-\theta\|_1^m \max_\jmath\left|\sum_{i\le N}[\![\varphi_i(t_i)X_{i\jmath}]\!]\right| \le \|v-\theta\|_1^m \max_\jmath Z_\jmath. \quad (30)$$

By (25), w.p. 1, $|X_{ij}|\le 1$, $i\le N$, $j\le p$, and so $|t_i| = |X_i^\top(v-\theta)|\le\|v-\theta\|_1\le R$. Then by Lemma 4.1,

$$|\varphi_i(t_i)| \le \min\left(\frac{2F_m}{m!}, \frac{F_{m+1}R}{(m+1)!}\right) = \phi.$$

It follows that

$$|\varphi_i(t_i)X_{i\jmath}|\le\phi, \quad |[\![\varphi_i(t_i)X_{i\jmath}]\!]|\le 2\phi := M_0,\ \forall\jmath, \quad \text{w.p. 1.} \quad (31)$$

Observe that given $v\in D$, for each $i\le N$, $\varphi_i(t_i)X_{i\jmath}$ is a function only in (Y_i, Z_i). Therefore, by independence, for $m\ge 0$ and $v\in D$,

$$\mathsf{Var}\left(\sum_{i\le N}\varphi_i(t_i)X_{i\jmath}\right) = \sum_{i\le N}\mathsf{Var}\left(\varphi_i(t_i)X_{i\jmath}\right) \le \sum_{i\le N}\mathsf{E}\left[\varphi_i(t_i)^2X_{i\jmath}^2\right] \le \phi^2\mathsf{E}\left[\sum_{i\le N}X_{i\jmath}^2\right].$$

If $m = 0$, then the right hand side is $N\phi^2$. If $m \geq 1$, by Young inequality,

$$\sum_{i\leq N} X_{ij}^2 = \sum_{i\leq N} X_{ij_1}^2 \cdots X_{ij_m}^2 \leq \prod_{s\leq m} \left(\sum_{i\leq N} X_{ij_s}^{2m} \right)^{1/m} = \prod_{s\leq m} \|V_{j_s}\|_{2m}^2 \leq \max_{j\leq p} \|V_j\|_{2m}^{2m}.$$

Therefore,

$$\mathsf{Var}\left(\sum_{i\leq N} \varphi_i(t_i) X_{ij} \right) \leq S_0^2 := \begin{cases} \phi^2 N & m = 0 \\ \phi^2 \mathsf{E}\left[\max_{j\leq p} \|V_j\|_{2m}^{2m}\right] & m \geq 1. \end{cases} \tag{32}$$

Fix one $\jmath = (j_1, \ldots, j_p)$. We next combine (31) and (32) with measure concentration to bound the upper tail of $Z_\jmath$. Again, note that given v, $\varphi_i(t_i)X_{i\jmath}$ is a function only in (Y_i, Z_i), with $t_i = X_i^\top(v-\theta)$. Let

$$\mathcal{T} = \{\tau = (\tau_{v,a}^1, \ldots, \tau_{v,a}^N) : v \in D, \quad a \in \{-1, 1\}\},$$

be a collection of functions parameterized by $D \times \{-1, 1\}$ mapping $(\mathcal{Y} \times \mathcal{Z})^N$ into $\mathbb{R}^N$, such that

$$\tau_{v,a}^i(Y_i, Z_i) = aM_0^{-1} [\![\varphi_i(t_i)X_{i\jmath}]\!], \quad i \leq N.$$

Then $Z_\jmath = M_0\tilde{Z}$, $S_0^2 = M_0^2\tilde{S}^2$, with

$$\tilde{Z} = \sup_{\tau\in\mathcal{T}} \sum_{i\leq N} \tau^i(Y_i, Z_i), \quad \tilde{S}^2 = \sup_{\tau\in\mathcal{T}} \mathsf{Var}\left(\sum_{i\leq N} \tau^i(Y_i, Z_i) \right).$$

From (31), for $v \in D$ and $a = \pm 1$, $\tau_{v,a}^i \in [-1, 1]$. Clearly, $\mathsf{E}\tau_{v,a}^i(Y_i, Z_i) = 0$. Furthermore, w.p. 1, $\tau_{v,a}^i(Y_i, Z_i)$ is continuous in v. Therefore, by dominated convergence argument, Theorem 1.1 in [7] can be applied to $\tilde{Z}$. Let $w = 2\mathsf{E}\tilde{Z} + \tilde{S}^2 = 2\mathsf{E}Z_\jmath/M_0 + S_0^2/M_0^2$. Then by [7],

$$\mathsf{Pr}\{Z_\jmath > \mathsf{E}Z_\jmath + M_0 a\} = \mathsf{Pr}\left\{\tilde{Z} > \mathsf{E}\tilde{Z} + a\right\} \leq \exp\left\{-\frac{a^2}{2w + 3a}\right\}, \quad \forall a > 0.$$

For $s > 0$, $a = (1/2)(3s + \sqrt{9s^2 + 8sw})$ is the unique positive solution to $a^2/(2w + 3a) = s$. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq a + b$,

$$\begin{aligned} \mathsf{E}Z_\jmath + M_0 a &\leq \mathsf{E}Z_\jmath + (M_0/2)(3s + \sqrt{9s^2} + \sqrt{8sw}) \\ &= \mathsf{E}Z_\jmath + M_0\left(3s + \sqrt{2s(2\mathsf{E}Z_\jmath/M_0 + S_0^2/M_0^2)}\right) \\ &\leq \mathsf{E}Z_\jmath + M_0\left(3s + \sqrt{4s\mathsf{E}Z_\jmath/M_0} + \sqrt{2sS_0^2/M_0^2}\right) \\ &\leq \mathsf{E}Z_\jmath + M_0(4s + \mathsf{E}Z_\jmath/M_0 + (S_0/M_0)\sqrt{2s}). \end{aligned}$$

Then

$$\mathsf{Pr}\left\{Z_\jmath > 2\mathsf{E}Z_\jmath + S_0\sqrt{2s} + 4M_0 s\right\} \leq e^{-s}. \tag{33}$$

To find an upper bound for $\mathsf{E}Z_\jmath$, let $\varepsilon_1, \ldots, \varepsilon_N$ be a Rademacher sequence independent of (Y_i, Z_i). By symmetrization inequality (cf. [9], Lemma 6.3)

$$\mathsf{E}Z_\jmath \leq 2\mathsf{E}\sup_{v\in D}\left|\sum_{i\leq N} \varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|. \tag{34}$$

By Fubini Theorem, the expectation on the right hand side is

$$\mathsf{E}_{X,Y}\mathsf{E}_{\varepsilon}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|,$$

where $\mathsf{E}_{X,Y}$ denotes the expectation only with respect to the (marginal) distribution of (X_1,Y_1), ..., (X_N,Y_N), and similarly for E_{ε}.

From (26), $\varphi_i(0)=0$. Assume $\psi>0$ first. Given (X_1,Y_1), ..., (X_N,Y_N), by Lemma 4.1 and (25),

$$t\to\varphi_i(t)X_{i\jmath}/\psi$$

is a contraction for each $i\le N$. Meanwhile, we can write

$$\mathsf{E}_{\varepsilon}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|=\mathsf{E}_{\varepsilon}\sup_{t\in T}\left|\sum_{i\le N}\varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|,$$

with $T=T(X_1,\ldots,X_N)=\{(t_1,\ldots,t_N):t_i=X_i^{\top}(v-\theta),\ v\in D\}$. Then by a comparison inequality (cf. Theorem 4.12 in [9]),

$$\mathsf{E}_{\varepsilon}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|\le 2\psi\mathsf{E}_{\varepsilon}\sup_{t\in T}\left|\sum_{i\le N}\varepsilon_i t_i\right|.$$

Using $t_i=X_i^{\top}(v-\theta)$ and by the same argument for (29)

$$\begin{aligned}\mathsf{E}_{\varepsilon}\sup_{t\in T}\left|\sum_{i\le N}\varepsilon_i t_i\right|&=\mathsf{E}_{\varepsilon}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i X_i^{\top}(v-\theta)\right|\\&\le\mathsf{E}_{\varepsilon}\sup_{\jmath\le p,\,v\in D}\left|\sum_{i\le N}\varepsilon_i X_{i\jmath}\right|\|v-\theta\|_1\le R\mathsf{E}_{\varepsilon}\max_{\jmath\le p}\left|\varepsilon^{\top}V_{\jmath}\right|,\end{aligned}$$

where $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_N)^{\top}$. With (X_i,Y_i) being fixed, by a result in [10] (Lemma 5.2),

$$\mathsf{E}_{\varepsilon}\max_{\jmath\le p}\left|\varepsilon^{\top}V_{\jmath}\right|\le\sqrt{2\ln(2p)}\max_{\jmath\le p}\|V_{\jmath}\|_2.$$

Combining the inequalities and taking expectation with respect to (X_i,Y_i),

$$\mathsf{E}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i\varphi_i(t_i)X_{i\jmath}\right|\le 2\sqrt{2}\psi R\sqrt{\ln(2p)}\mathsf{E}\left[\max_{\jmath\le p}\|V_{\jmath}\|_2\right].\tag{35}$$

If $\psi=0$, then $\varphi_i\equiv 0$ and the above inequality holds trivially. Combining (33) – (35) yields

$$\mathsf{Pr}\left\{Z_{\jmath}>M_1\sqrt{2\ln(2p)}+\sqrt{2}S_0\sqrt{s}+4M_0 s\right\}\le e^{-s}.$$

where

$$M_1=8\psi R\mathsf{E}\left[\max_{\jmath\le p}\|V_{\jmath}\|_2\right].\tag{36}$$

Finally, since there are p^m different values of $\jmath$, by union-sum inequality,

$$\mathsf{Pr}\left\{\max_{\jmath}Z_{\jmath}>M_1\sqrt{\ln(2p)}+\sqrt{2}S_0\sqrt{\ln(p^m/q)}+4M_0\ln(p^m/q)\right\}\le q,\quad\forall\,q\in(0,1).\tag{37}$$

Note M_1, S_0^2 and $4M_0$ are exactly A, B and C in Theorem 2.1. Then by (30), the proof is complete. □

Proof of Corollary 2.2. As in the proof of Theorem 2.1, we consider the domain D in (23) and X_i in (24). Still denote $c_i = X_i^\top \theta$. For $m = 1$, by Theorem 2.1, for θ, $v \in D$

$$\sum_{i \le N} [\![f_i(X_i^\top v)]\!] = \sum_{i \le N} [\![f_i(c_i)]\!] + \sum_{i \le N} [\![f_i'(c_i) X_i]\!]^\top (v - \theta) + \xi(v) \|v - \theta\|_1.$$

By Hölder inequality,

$$\left| \sum_{i \le N} [\![f_i'(c_i) X_i]\!]^\top (v - \theta) \right| \le \|v - \theta\|_1 \max_{j \le p} \left| \sum_{i \le N} [\![f_i'(c_i) X_{ij}]\!] \right|.$$

Given $j \le p$, $[\![f_i'(c_i) X_{ij}]\!]$ are independent with mean 0, and each $|f_i'(c_i) X_{ij}| \le F_1$. Therefore, by Hoeffding inequality ([11], p. 191) and union-sum inequality,

$$\mathsf{Pr} \left\{ \max_{j \le p} \left| \sum_{i \le N} [\![f_i'(c_i) X_{ij}]\!] \right| \ge t \right\} \le 2p \exp \left\{ -\frac{t^2}{2NF_1^2} \right\}.$$

Given $q \in (0, 1)$, let $t = \sqrt{N} F_1 \sqrt{2 \ln(2p/q)}$ to get the right hand side no greater than q. Combining this with the bound for $\xi(v)$, the proof is complete. □

Proof of Theorem 2.3. The proof is similar to that of Theorem 2.1, so we will be brief. Define domain D as (23) and X_{ij}, X_i, V_j as in (24). Let $c = (c_1, \ldots, c_N)$ and $t = (t_1, \ldots, t_N)$ with

$$c_i = X_i^\top \theta, \quad t_i = X_i^\top (v - \theta).$$

Define φ_i as in (26), however, note that the meaning of f_i is different here. In particular, f_i are nonrandom and hence φ_i are nonrandom as well. In spite of this, Lemma 4.1 still holds. Corresponding to (28),

$$\begin{aligned} &\sum_{i \le N} \omega_i f_i(h(Z_i)^\top v) \\ &= \sum_{k \le m} \frac{1}{k!} \left(\sum_{i \le N} \omega_i f_i^{(k)}(c_i) X_i^{\otimes k} \right)^\top (v - \theta)^{\otimes k} + \left(\sum_{i \le N} \omega_i \varphi_i(t_i) X_i^{\otimes m} \right)^\top (v - \theta)^{\otimes m}. \end{aligned}$$

The next step is to bound the upper tail probability of $\max_{\jmath} Z_{\jmath}$, where for $\jmath = (j_1, \ldots, j_m)$,

$$Z_{\jmath} = \sup_{v \in D} \left| \sum_{i \le N} \omega_i \varphi_i(t_i) X_{i\jmath} \right|, \quad \omega = (\omega_1, \ldots, \omega_N)^\top.$$

Write $\omega_i = \sigma_i \varepsilon_i$, where $\sigma_i^2 = \mathsf{Var}(\omega_i) \le \sigma_0^2$ and $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. $\sim N(0, 1)$. Fix one $\jmath$. Then

$$Z_{\jmath} = Z(\varepsilon) = \sup_{v \in D} \left| \sum_{i \le N} \varepsilon_i \sigma_i \varphi_i(t_i) X_{i\jmath} \right|, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^\top.$$

The function Z is Lipschitz on $\mathbb{R}^N$ under the Euclidean norm (ℓ_2 norm), because for a, $b \in \mathbb{R}^N$,

$$|Z(a) - Z(b)| \le \sup_{v \in D} \left| \sum_{i \le N} (a_i - b_i) \sigma_i \varphi_i(t_i) X_{i\jmath} \right| \le \|a - b\|_2 \sigma_0 S_0,$$

where, as in (32),

$$S_0^2 = \begin{cases} \phi^2 N & m = 0 \\ \phi^2 \max_{j \le p} \|V_j\|_{2m}^{2m} & m \ge 1. \end{cases}$$

Now by a concentration inequality for Gaussian measure ([8], p. 41)

$$\mathsf{Pr}\left\{Z(\varepsilon) \ge \mathsf{E}Z(\varepsilon) + r\sigma_0 S_0\right\} \le \exp(-r^2/2), \quad \forall r > 0. \tag{38}$$

By Lemma 4.1 and $|X_{ij}| \le 1$, $t \to \varphi_i(t)X_{ij}/\psi$ is a contraction with 0 being mapped to 0. Then by a comparison result for Gaussian process ([9], Corollary 3.17 and (3.13))

$$\mathsf{E}Z(\varepsilon) \le 4\sigma_0\psi\mathsf{E}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i t_i\right| = 4\sigma_0\psi\mathsf{E}\sup_{v\in D}\left|\sum_{i\le N}\varepsilon_i X_i^\top(v-\theta)\right| \le 4\sigma_0 R\psi\mathsf{E}\max_{j\le p}\left|\varepsilon^\top V_j\right|,$$

and

$$\mathsf{E}\max_{j\le p}\left|\varepsilon^\top V_j\right| \le 3\sqrt{\ln p}\max_{j\le p}\sqrt{\mathsf{Var}(\varepsilon^\top V_j)} = 3\sqrt{\ln p}\max_{j\le p}\|V_j\|_2.$$

Using an argument in [10], one can get a bound for the expectation that is tighter for large p.

Lemma 4.2. *There is*

$$\mathsf{E}\max_{j\le p}\left|\varepsilon^\top V_j\right| \le 2\sqrt{\ln(2p)}\max_{j\le p}\|V_j\|_2.$$

Now (38) can be written in terms of $Z_\jmath$. Then, as in (37), for $q \in (0,1)$,

$$\mathsf{Pr}\left\{\max_{\jmath} Z_\jmath > \sigma_0\left(M_1\sqrt{\ln(2p)} + \sqrt{2\ln(p^m/q)}S_0\right)\right\} \le q, \tag{39}$$

where

$$M_1 = 8R\psi\max_{j\le p}\|V_j\|_2.$$

This then finishes the proof. □

Proof of Corollary 2.4. From Theorem 2.3, it is seen that

$$\sum_{i\le N}\omega_i f_i(h(Z_i)^\top v) = \sum_{i\le N}\omega_i f_i(h(Z_i)^\top\theta) + \zeta + \xi(v)\sum_{j\le p} d_j|v_j - \theta_j|,$$

where

$$\zeta = \sum_{j\le p}\left(\sum_{i\le N}\omega_i f_i'(h(Z_i)^\top\theta)h_j(Z_i)\right)(v_j - \theta_j).$$

Therefore, with w_j being defined as in (12),

$$|\zeta| \le \sigma_0 F_1 \sum_{j\le p} w_j|v_j - \theta_j| \times \max_{j\le p}|W_j|,$$

with

$$W_j = \frac{1}{\sigma_0 F_1 w_j}\sum_{i\le N}\omega_i f_i'(h(Z_i)^\top\theta)h_j(Z_i).$$

It is easy to see that each W_j is Gaussian with mean 0 and variance no greater than 1. As a result,

$$\mathsf{Pr}\left\{\max_{j\le p}|W_j| \ge t\right\} \le p\exp(-t^2/2), \quad t \ge 0.$$

Given $q \in (0,1)$, letting $t = \sqrt{2\ln(p/q)}$ then finishes the proof. □

4.2 Proof for Section 3

Proof of Theorem 3.1. For $A \subset \{1, \ldots, p\}$ and $v \in \mathbb{R}^p$, denote by v_A the vector $u \in \mathbb{R}^p$ with $u_i = v_i \mathbf{1}\{i \in A\}$. By definition of $\widehat{\theta}$,

$$L(\widehat{\theta}) - L(\theta) \le \sum_{i \le N} \left[\!\left[\gamma_i(X_i^\top \theta, Y_i) - \gamma_i(X_i^\top \widehat{\theta}, Y_i) \right]\!\right] + \lambda d(\|\theta\|_1 - \|\widehat{\theta}\|_1).$$

Let $\lambda = (1 + 1/c) M_q d$, where $c > 0$ is to be determined. Then, writing $r = 1/c$, on the event that (17) holds,

$$L(\widehat{\theta}) - L(\theta) \le M_q d \left[\|\widehat{\theta} - \theta\|_1 + (1 + r)(\|\theta\|_1 - \|\widehat{\theta}\|_1) \right].$$

Fix any J containing $\mathrm{spt}(\theta)$. Then

$$\begin{aligned}
&\|\widehat{\theta} - \theta\|_1 + (1 + r)(\|\theta\|_1 - \|\widehat{\theta}\|_1) \\
&= \sum_{i \in J} |\widehat{\theta}_i - \theta_i| + \sum_{i \notin J} |\widehat{\theta}_i| + (1 + r) \left(\sum_{i \in J} |\theta_i| - \sum_{i \in J} |\widehat{\theta}_i| - \sum_{i \notin J} |\widehat{\theta}_i| \right) \\
&= \sum_{i \in J} \left[|\widehat{\theta}_i - \theta_i| + (1 + r)(|\theta_i| - |\widehat{\theta}_i|) \right] - r \sum_{i \notin J} |\widehat{\theta}_i| \\
&\le (2 + r) \|\widehat{\theta}_J - \theta\|_1 - r \|\widehat{\theta}_{J^c}\|_1.
\end{aligned}$$

On the one hand, the above inequalities yield

$$L(\widehat{\theta}) - L(\theta) \le M_q d (2 + 1/c) \|\widehat{\theta}_J - \theta\|_1, \tag{40}$$

and on the other, since by definition of θ, $L(\widehat{\theta}) \ge L(\theta)$,

$$\|\widehat{\theta}_{J^c}\|_1 \le (1 + 2c) \|\widehat{\theta}_J - \theta\|_1. \tag{41}$$

Set $c = (K - 1)/2$. Then $\lambda = (1 + 1/c) M_q d$ is as in (18). By (15), (16) and (40), for any $J \supset \mathrm{spt}(\theta)$ with $|J| \le 2\|\theta\|_0$,

$$N C_\gamma \kappa^2 \|\widehat{\theta}_J - \theta\|_2^2 \le L(\widehat{\theta}) - L(\theta) \le \frac{2 M_q K d}{K - 1} \|\widehat{\theta}_J - \theta\|_1.$$

Since $\|\widehat{\theta}_J - \theta\|_1 \le \sqrt{|J|} \|\widehat{\theta}_J - \theta\|_2$, it follows that

$$\|\widehat{\theta}_J - \theta\|_2 \le b \sqrt{|J|} \quad \text{with} \quad b = \frac{M_q}{N} \times \frac{2 K d}{C_\gamma \kappa^2 (K - 1)}. \tag{42}$$

Let A be the set of indices $i \notin \mathrm{spt}(\theta)$ corresponding to the $\|\theta\|_0$ largest $|\widehat{\theta}_i|$. Then (42) holds for both $J_0 = \mathrm{spt}(\theta)$ and $J_1 = \mathrm{spt}(\theta) \cup A$. It is well known that (cf. [5])

$$\|\widehat{\theta}_{J_1^c}\|_2^2 \le \frac{\|\widehat{\theta}_{J_0^c}\|_1^2}{\|\theta\|_0}.$$

By (41) and Cauchy-Schwartz inequality followed by (42),

$$\|\widehat{\theta}_{J_1^c}\|_2^2 \le \frac{K^2 \|\widehat{\theta}_{J_0} - \theta\|_1^2}{\|\theta\|_0} \le K^2 \|\widehat{\theta}_{J_0} - \theta\|_2^2 \le K^2 b^2 \|\theta\|_0.$$

Combining this with (42) applied to $J = J_1$,

$$\|\widehat{\theta} - \theta\|_2^2 = \|\widehat{\theta}_{J_1} - \theta\|_2^2 + \|\widehat{\theta}_{J_1^c}\|_2^2 \le b^2 |J_1| + K^2 b^2 \|\theta\|_0 = (2 + K^2) b^2 \|\theta\|_0.$$

So we finally arrive at (19). □

4.3 Proof of Lemmas

Proof of Lemma 4.1. If $m = 0$, then $\varphi_i(t) = f_i(c_i + t) - f_i(c_i)$. From Assumptions 1 and 3, the result is straightforward.

Let $m \geq 1$. For $t > 0$, by Taylor expansion with an integral remainder,

$$f_i(c_i + t) - \sum_{k \leq m} \frac{f_i^{(k)}(c_i)}{k!} t^k = \frac{1}{(m-1)!} \int_0^t (t-s)^{m-1} [f_i^{(m)}(c_i + s) - f_i^{(m)}(c_i)] \, \mathrm{d}s, \tag{43}$$

yielding

$$\varphi_i(t) = \frac{t^{-m}}{(m-1)!} \int_0^t (t-s)^{m-1} [f_i^{(m)}(c_i + s) - f_i^{(m)}(c_i)] \, \mathrm{d}s.$$

Therefore, by Assumption 1, on the one hand,

$$|\varphi_i(t)| \leq \frac{t^{-m}}{(m-1)!} \int_0^t (2F_m)(t-s)^{m-1} \, \mathrm{d}s = \frac{2F_m}{m!},$$

and on the other,

$$|\varphi_i(t)| \leq \frac{t^{-m}}{(m-1)!} \int_0^t (t-s)^{m-1} (F_{m+1} s) \, \mathrm{d}s = \frac{F_{m+1}|t|}{(m+1)!}.$$

The inequalities hold likewise for $t < 0$. Therefore, (27) holds. The above inequality also implies that φ_i is continuous at 0. It is clear that $\varphi_i(t)$ is continuous at $t \neq 0$. Thus $\varphi_i \in C(a_i - c_i, b_i - c_i)$.

It remains to show $\|\varphi_i\|_{\mathrm{Lip}} \leq \psi$. Since φ_i is differentiable at $t \neq 0$, it is enough to show $|\varphi_i'(t)| \leq \psi$ for $t \neq 0$. First, let $m = 1$. For $t \neq 0$,

$$\varphi_i'(t) = t^{-2}[f_i(c_i) - f_i(c_i + t) + t f_i'(c_i + t)] = t^{-2} \int_0^t [f_i'(c_i + t) - f_i'(c_i + t - s)] \, \mathrm{d}s.$$

By Assumption 1, $|f_i'(c_i + t) - f_i'(c_i + t - s)| \leq F_2 |s|$. Consequently $|\varphi_i'(t)| \leq F_2/2 = \psi$.

Finally, let $m \geq 2$. Define $g(t) = m f_i(c_i + t) - t f_i'(c_i + t)$. Then for $k < m$,

$$g^{(k)}(t) = (m-k) f_i^{(k)}(c_i + t) - t f_i^{(k+1)}(c_i + t)$$

and then

$$\begin{aligned}
\varphi_i'(t) &= t^{-m} \left(f_i'(c_i + t) - \sum_{k=1}^{m} \frac{f_i^{(k)}(c_i) t^{k-1}}{(k-1)!} \right) - m t^{-m-1} \left(f_i(c_i + t) - \sum_{k=0}^{m} \frac{f_i^{(k)}(c_i) t^k}{k!} \right) \\
&= -t^{-m-1} \left(m f_i(c_i + t) - t f_i'(c_i + t) - \sum_{k=0}^{m-1} \frac{(m-k) f_i^{(k)}(c_i) t^k}{k!} \right) \\
&= -t^{-m-1} \left(g(t) - \sum_{k=0}^{m-1} \frac{g^{(k)}(0) t^k}{k!} \right) \\
&= -\frac{t^{-m-1}}{(m-2)!} \int_0^t (t-s)^{m-2} [g^{(m-1)}(s) - g^{(m-1)}(0)] \, \mathrm{d}s,
\end{aligned}$$

where the last equality is by similar Taylor expansion as (43), now applied to g with order $m - 1$. For each s,

$$\begin{aligned}
g^{(m-1)}(s) - g^{(m-1)}(0) &= f_i^{(m-1)}(c_i + s) - s f_i^{(m)}(c_i + s) - f_i^{(m-1)}(c_i) \\
&= \int_0^s [f_i^{(m)}(c_i + s - u) - f_i^{(m)}(c_i + s)] \, \mathrm{d}u,
\end{aligned}$$

giving $|g^{(m-1)}(s) - g^{(m-1)}(0)| \le F_{m+1}s^2/2$. Then

$$|\varphi_i'(t)| \le \frac{t^{-m-1}F_{m+1}}{2(m-2)!}\int_0^t (t-s)^{m-2}s^2\,\mathrm{d}s = \frac{F_{m+1}}{m!} = \psi.$$

This finishes the proof. □

Proof of Lemma 4.2. Let $x = \mathsf{E}\max_{j\le p}|\varepsilon^\top V_j|$. By Jensen inequality, for any $t > 0$,

$$\exp(tx) \le \mathsf{E}\left[\exp\left(t\max_{j\le p}|\varepsilon^\top V_j|\right)\right] = \mathsf{E}\left[\max_{j\le p}\exp(t|\varepsilon^\top V_j|)\right] \le \sum_{j\le p}\mathsf{E}[\exp(t|\varepsilon^\top V_j|)].$$

Since $\varepsilon^\top V_j \sim N(0, \|V_j\|_2^2)$,

$$\mathsf{E}[\exp(t|\varepsilon^\top V_j|)] \le \mathsf{E}[\exp(t\varepsilon^\top V_j)] + \mathsf{E}[\exp(-t\varepsilon^\top V_j)] = 2\exp(t\|V_j\|_2^2).$$

Then

$$\exp(tx) \le 2p\exp\left(t^2\max_{j\le p}\|V_j\|_2^2\right).$$

The proof is finished by letting $t = x/(2\max_{j\le p}\|V_j\|_2^2)$. □

References

[1] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[2] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[3] Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. Spades and mixture models. *Ann. Statist.*, 38(4):2525–2558, 2010.

[4] Emmanuel J. Candès and Yaniv Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.

[5] Emmanuel J. Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[6] Zhiyi Chi. A hybrid estimator for high-dimensional generalized linear models with non-convex loss. Technical Report 10-27, University of Connecticut, Department of Statistics, 2010.

[7] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

[8] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[9] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[10] Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory.

[11] David Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

[12] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[13] Tong Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.

[14] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.