First order expansion of convex regularized estimators

Pierre C Bellec,
Department of Statistics,
Rutgers University,
501 Hill Center,
Piscataway, NJ 08854, USA.
pierre.bellec@rutgers.edu

Arun K Kuchibhotla,
Department of Statistics,
The Wharton School,
University of Pennsylvania,
Philadelphia, PA 19104, USA.
arunku@upenn.edu

Abstract

We consider first order expansions of convex penalized estimators in high-dimensional regression problems with random designs. Our setting includes linear regression and logistic regression as special cases. For a given penalty function h and the corresponding penalized estimator $\hat{\beta}$, we construct a quantity η, the first order expansion of $\hat{\beta}$, such that the distance between $\hat{\beta}$ and η is an order of magnitude smaller than the estimation error $\|\hat{\beta} - \beta^*\|$. In this sense, the first order expansion η can be thought of as a generalization of influence functions from the mathematical statistics literature to regularized estimators in high-dimensions. Such first order expansion implies that the risk of $\hat{\beta}$ is asymptotically the same as the risk of η which leads to a precise characterization of the MSE of $\hat{\beta}$; this characterization takes a particularly simple form for isotropic design. Such first order expansion also leads to inference results based on $\hat{\beta}$. We provide sufficient conditions for the existence of such first order expansion for three regularizers: the Lasso in its constrained form, the lasso in its penalized form, and the Group-Lasso. The results apply to general loss functions under some conditions and those conditions are satisfied for the squared loss in linear regression and for the logistic loss in the logistic model.

Introduction. We consider learning problems where one observes observations $(X_1, Y_1), ..., (X_n, Y_n)$ with responses Y_i and feature vectors $X_i \in \mathbb{R}^p$. The literature of the past two decades has demonstrated the great success of regularized estimators that are commonly defined as solutions to regularized optimization problems of the form

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} n^{-1} \sum_{i=1}^n \ell(Y_i, X_i^T \beta) + h(\beta), \tag{1}$$

where $\ell(\cdot, \cdot)$ is referred to as the loss (e.g. squared loss, logistic loss) and $h : \mathbb{R}^p \to \mathbb{R}$ is a regularization penalty (e.g. the ℓ_1-norm for the Lasso, the $\ell_{2,1}$ norm for the Group-Lasso). All tuning parameters are included in $h(\cdot)$. The performance of such regularized estimators is measured in terms of prediction error or in terms estimation error $\|\hat{\beta} - \beta^*\|$ if the data comes from a model such as $Y = X\beta^* + \varepsilon$ for some noise random variable ε in linear regression or

$$\mathbb{P}(Y = 1 | X = x) = 1/(1 + \exp(x^T \beta^*)) = 1 - \mathbb{P}(Y = 0 | X = x)$$

in logistic regression, where β^* is the unknown coefficient vector. For instance, if $s = \|\beta^*\|_0$ is the sparsity of β^* in the above model, and $(X_i, Y_i)_{i=1,...,n}$ are iid observations with the same distribution as (X, Y), both the Lasso in linear regression and the logistic Lasso in logistic regression enjoy rate optimality: $\|\hat{\beta} - \beta^*\|^2 \le s \log(ep/s)/n$; see [35, 1] or the proof of Proposition 3.4 in Appendix F for self-contained proofs. The latter estimation bound is optimal in a minimax sense and cannot

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

be improved, and the minimax rate $s\log(ep/s)/n$ represents the scale below which uncertainty is unavoidable by information theoretic arguments, see for instance [36, Section 5].

We are interested in providing first order expansion of $\hat{\beta}$ at scales negligible compared to the minimax estimation rate, e.g. at scales negligible compared to $s\log(ep/s)/n$ in the aforementioned sparsity contexts. To be more precise, the results below will construct random first order expansion η such that η is measurable w.r.t. a much smaller sigma algebra than that generated by $(X_i, Y_i)_{i=1,\ldots,n}$, and

$$\|\eta - \hat{\beta}\|_K^2 = o_p(1)\|\hat{\beta} - \beta^*\|_K^2 \quad \text{for some norm } \|\cdot\|_K \text{ related to the problem at hand,} \quad (2)$$

where $o_p(1)$ is a quantity that converges to 0 in probability. In other words, we provide a first-order expansion of $\hat{\beta}$ similar to an influence function expansion, cf. Section 1. This allows for understanding bias and standard deviation of $\hat{\beta}$ at a finer scale than simply showing that $\hat{\beta} - \beta^*$ converge to zero at the minimax rate. The present paper intends to answer the two questions below regarding such first order expansion.

- **(Q1)** How to construct η such that (2) holds for a given convex regularized estimator such as (1)?
- **(Q2)** How are such first order expansions useful in high-dimensional learning problems where convex regularized estimators (1) are commonly used?

An expansion η satisfying (2) is interesting in and by itself because it describes phenomena at a finer scale than most of the literature in high-dimensional problems which focuses on minimax prediction and estimation bounds. More importantly, we will see in Section 4 that such first-order expansions lead to exact identities for the loss of estimators, and in Section 5 that such first-order expansions can be used for inference (i.e., uncertainty quantification) about the unknown coefficient vector β^*.

Notation. Throughout the paper, $C_1, C_2, C_3, ...$ denote positive absolute constants and we write $a \lesssim b$ if $a \leq Cb$ for some absolute constant $C > 0$. The Euclidean norm in $\mathbb{R}^p$ or in $\mathbb{R}^n$ is denoted by $\|\cdot\|$. For any positive definite matrix A, we write $\|u\|_A = \|A^{1/2}u\|$ for the matrix square-root A. For matrices, $\|\cdot\|_{op}$ and $\|\cdot\|_F$ denote the operator norm and Frobenius norm. For any real a, $a_+ = \max(0, a)$. If $S \subset \{1, ..., p\}$, $v \in \mathbb{R}^p, M \in \mathbb{R}^{p\times p}$ then v_S is the restriction $(v_j, j \in S)$ and $M_{S,S}$ is the square submatrix of M made of entries indexed in $S \times S$.

1 Influence functions and Construction of η

To answer **(Q1)**, we start with a recap of unregularized estimators that correspond to $h(\cdot) \equiv 0$, when p is fixed as $n \to +\infty$. In this case, it is well-known that certain smoothness assumptions on the loss such as twice differentiability [25, 19] or stochastic equicontinuity [42, 41] imply (for any norm, since all norms are equivalent in $\mathbb{R}^p$ for fixed p):

$$\left\|\hat{\beta} - \beta^* - \sum_{i=1}^n \frac{\psi(X_i, Y_i)}{n}\right\| = o_p(1)\|\hat{\beta} - \beta^*\| \Leftrightarrow (1 + o_p(1))\sqrt{n}(\hat{\beta} - \beta^*) = \sum_{i=1}^n \frac{\psi(X_i, Y_i)}{\sqrt{n}}, \quad (3)$$

for some target β^* and a mean zero function $\psi(\cdot, \cdot)$ sometimes referred to as the influence function. See [25, Theorem 3.1], [42, Page 52], [41, Theorem 6.17], [19, Lemma 5.4] for details. In this case we can take $\eta = \beta^* + \sum \psi(X_i, Y_i)/n$ in (2). This representation allows us to claim asymptotic unbiasedness and fluctuations of order $n^{-1/2}$ for $\hat{\beta}$ around β^*. It also shows that estimator $\hat{\beta}$ behaves like an average and hence allows transfer of results (e.g., central limit theorems) for averages to study of $\hat{\beta}$ in terms of variance estimation, confidence intervals, hypothesis testing and bootstrap.

A general study of such representation for regularized problems is lacking in the literature. [23] is the first work that analyzed linear regression lasso when the number of covariates p is fixed and does not change with the sample size n. In the more challenging regime where $p \geq n$, Theorem 5.1 of [22] provides a first order expansion allowing for p to diverge (almost exponentially) with n. In the present work, we simplify and present a unified derivation of such first order expansion result, generalizing [22, Theorem 5.1] beyond the squared loss, beyond the ℓ_1 penalty and beyond certain assumptions of [22] on $\mathbb{E}[X_i X_i^T]$. The derivation of (3) can be motivated by defining

$$\tilde{\eta} := \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\beta-\beta^*) + \frac{1}{2n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\{X_i^\top(\beta-\beta^*)\}^2 + h(\beta), \quad (4)$$

with $h(\cdot) \equiv 0$. Here and throughtout $\ell'(y,u)$ and $\ell''(y,u)$ represent (first and second) partial derivatives of ℓ with respect to u. The right hand side of (4) (with $h(\cdot) \equiv 0$) is the quadratic approximation of $\sum_{i=1}^n \ell(Y_i, X_i^\top \beta)/n$ around $\beta = \beta^*$ (without the term independent of β). The final first order expansion η is obtained by replacing the quadratic part of the approximation by its expectation as in the next display. Following the intuitive construction of η for the unregularized problem, we construct a first order expansion for the regularized problem as

$$\begin{aligned} \eta &:= \operatorname{argmin}_{\beta\in\mathbb{R}^p}\ n^{-1}\textstyle\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\beta-\beta^*) + \frac{1}{2}(\beta-\beta^*)^\top K(\beta-\beta^*) + h(\beta) \\ &:= \operatorname{argmin}_{\beta\in\mathbb{R}^p}\ \tfrac{1}{2}\left\|K^{1/2}\left(\beta-\beta^*-n^{-1}\textstyle\sum_{i=1}^n K^{-1}X_i\ell'(Y_i,X_i^\top\beta^*)\right)\right\|^2 + h(\beta). \end{aligned} \tag{5}$$

where $K := n^{-1}\sum_{i=1}^n \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)X_iX_i^\top\right]$. From this definition, we can write $\eta = \eta_K\left(\beta^* + n^{-1}\sum_{i=1}^n K^{-1}X_i\ell'(Y_i,X_i^\top\beta^*)\right)$, for a function $\eta_K(\cdot)$ (depending on $h(\cdot), K$). Our main results prove under some mild assumptions that

$$\|\hat\beta - \eta_K(\beta^* + \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i)\|_K \ = \ o_p(1)\|\hat\beta-\beta^*\|_K.$$

Comparing this with (3) we note that for the unregularized problem, $\eta_K(\beta) = \beta$ is the identity.

2 Main Results: Approximation Theorem

We introduce the notion of Gaussian complexity for the following results. For any set $T \subset \mathbb{R}^p$ and a covariance matrix Σ, the Gaussian complexity of T is given by

$$\gamma(T,\Sigma) := \mathbb{E}\left[\sup_{u\in T:\,\|\Sigma^{1/2}u\|=1}\ |g^\top\Sigma^{1/2}u|\right] = \mathbb{E}\left[\sup_{u\in T}\ \tfrac{|g^\top\Sigma^{1/2}u|}{\|\Sigma^{1/2}u\|}\right], \tag{6}$$

where the expectation is with respect to the standard normal vector $g \sim N(0, I_p)$. We also need the notion of L-subGaussianity. A random vector X is said to be L-subGaussian with respect to a (positive definite) matrix Σ if

$$\forall u \in \mathbb{R}^p, \quad \mathbb{E}[\exp(u^T X)] \le \exp(L^2\|u\|_\Sigma^2/2) \qquad \text{where } \|u\|_\Sigma = \|\Sigma^{1/2}u\|. \tag{7}$$

This implies $\sup_u \mathbb{P}(|u^\top X| \ge t\|u\|_\Sigma) \le 2\exp(-t^2/(2L^2))$. Recall that the scaled norm $\|\cdot\|_K$ is defined by $\|u\|_K^2 = n^{-1}\sum_{i=1}^n \mathbb{E}[\ell''(Y_i, X_i^\top\beta^*)(X_i^\top u)^2]$. Consider the following assumptions:

(A1) There exists constants $0 \le B, B_2, B_3 < \infty$ such that the loss satisfies $\forall u_1, u_2 \in \mathbb{R}, \forall y$,

$$\frac{|\ell''(y,u_1)-\ell''(y,u_2)|}{|u_1-u_2|} \le B, \qquad |\ell''(y,u_1)| \le B_2, \qquad \sup_{u\in\mathbb{R}^p}\frac{\|\Sigma^{1/2}u\|^2}{\|K^{1/2}u\|^2} \le B_3. \tag{8}$$

(A2) The observations $(X_1,Y_1),\dots,(X_n,Y_n)$ are iid. Further $X_1,\dots,X_n$ are mean zero and L-subGaussian with respect to their covariance Σ, i.e., (7) holds.

Note that L in **(A2)** is necessarily no smaller than one, i.e., $L \ge 1$. Define the error

$$\mathcal{E} := \|\hat\beta-\beta^*\|_K + \|\eta-\beta^*\|_K \qquad \text{where} \qquad \|\cdot\|_K \text{ is the norm } \|u\|_K = \|K^{1/2}u\|. \tag{9}$$

The quantity $\mathcal{E}$ quantifies the error made by $\hat\beta$ and η in estimating β^* with respect to the norm $\|\cdot\|_K$. Bounds on $\|\hat\beta-\beta^*\|_K$ and $\|\eta-\beta^*\|_K$ follow from the existing literature; see [35] or Proposition 3.4 and its proof in Appendix F.

Theorem 2.1. *Let $r_n := n^{-1/2}\gamma(T,\Sigma)$ and assume that $r_n \le 1$. Further assume **(A1)** and **(A2)** hold true. Then with probability at least $1 - 2e^{-C_4 n r_n^2} - 2e^{-C_5\log n}$ we have the following:*

1. *If $\{\hat\beta-\beta^*, \eta-\beta^*\} \subseteq T$ then $\|\hat\beta-\eta\|_K \lesssim LB_2B_3r_n^{1/2}\mathcal{E} + B^{1/2}(B_3L)^{3/2}(1+r_n^3\sqrt{n})\mathcal{E}^{3/2}$.*
2. *If $\{\hat\beta-\eta, \hat\beta-\beta^*, \eta-\beta^*\} \subseteq T$ then $\|\hat\beta-\eta\|_K \lesssim B_2B_3L^2r_n\mathcal{E} + BB_3^{3/2}L^3(1+r_n^3\sqrt{n})\mathcal{E}^2$.*

The set T mentioned in Theorem 2.1(1) are available in the literature for many convex penalties. In the following, we will find this for constrained lasso, penalized lasso, and group lasso (with non-overlapping groups) under sharp conditions. We refer to [7] for slope penalty, and Negahban et al. [35, Lemma 1] and van de Geer [39, Def. 4.4 and Theorem 4.1] where set T is presented for a general class of penalty functions including nuclear norm, group lasso (with overlapping groups).

Proofs of Theorem 2.1 and all following results are given in the supplement. An outline Theorem 2.1 is given in Section 6. Although Theorem 2.1 is stated under assumption **(A2)**, we present a deterministic version of the result (in Section 6) that replaces r_n by suprema of different stochastic processes.

Squared loss in the linear model. Consider $\ell(y,u) = (y-u)^2/2$ and n iid observations

$$Y_i = X_i^T \beta^* + \varepsilon_i, \text{ and } X_i \text{ is independent of } \varepsilon_i \text{ for } i = 1, \dots, n, \tag{10}$$

Then we have $K = \Sigma = \mathbb{E}[X_1 X_1^T]$ and the second derivative ℓ'' is constant. Hence condition (8) is satisfied with $B = 0$ and $B_2 = B_3 = 1$. The conclusions of the Theorem 2.1 can be rewritten as

$$\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subseteq T \quad \Rightarrow \quad \|\hat{\beta} - \eta\|_K \lesssim L r_n^{1/2} \mathcal{E}. \tag{11}$$

$$\{\hat{\beta} - \eta, \hat{\beta} - \beta^*, \eta - \beta^*\} \subseteq T \quad \Rightarrow \quad \|\hat{\beta} - \eta\|_K \lesssim L^2 r_n \mathcal{E}, \tag{12}$$

where $\mathcal{E} = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| + \|\Sigma^{1/2}(\eta - \beta^*)\|$. Since $r_n \leq 1$ (and typically $r_n \to 0$ while L stays bounded, as we will see in the examples below), the inequality in (12) is stronger than the inequality in (11). In the linear model, we thus refer to inequality (11) as the "slow rate" inequality, and to (12) as the "fast rate" one. The set T encodes the low-dimensional structure and characterizes the rate r_n through the Gaussian complexity $\gamma(T, \Sigma)$. The fast rate inequality is granted provided that T contains the difference $(\eta - \hat{\beta})$ additionally to the error vectors $\{\hat{\beta} - \beta^*, \eta - \beta^*\}$. Conditions that ensure the fast rate inequality will be made explicit in Section 3.2 for the Lasso.

Logistic loss in the logistic model. The following proposition shows that (8) is again satisfied.

Proposition 2.2. *Consider the logistic loss $\ell(y,u) = yu - \log(1+e^u)$ for $y \in \{0,1\}, u \in \mathbb{R}$. Assume that $(X_i, Y_i)_{i=1,\dots,n}$ are iid satisfying the logistic regression model*

$$\mathbb{P}(Y_i = 1 | X_i) = 1 - \mathbb{P}(Y_i = 0 | X_i) = 1/(1 + \exp({X_i}^\top \beta^*)),$$

for some $\beta^* \in \mathbb{R}^p$ with $\|\Sigma^{1/2}\beta^*\| \leq 1$.[1] *Assume **(A2)** holds. Then* (8) *holds with $B = 1/(6\sqrt{3})$, $B_2 = 1$ and an absolute constant $B_3 > 0$.*

In this logistic model, the conclusions of Theorem 2.1 present an extra term compared to the linear model with squared loss because the Lipschitz constant B in (8) is non-zero: Theorem 2.1 reads that with high probability

$$\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subset T \quad \Rightarrow \quad \|\hat{\beta} - \eta\|_K \lesssim L r_n^{1/2} \mathcal{E} + B^{1/2} L^{3/2} (1 + r_n^3 \sqrt{n}) \mathcal{E}^{3/2}, \tag{13}$$

$$\{\hat{\beta} - \eta, \hat{\beta} - \beta^*, \eta - \beta^*\} \subset T \quad \Rightarrow \quad \|\hat{\beta} - \eta\|_K \lesssim L^2 r_n \mathcal{E} + B L^3 (1 + r_n^3 \sqrt{n}) \mathcal{E}^2. \tag{14}$$

Similar to the case of squared loss, inequality (14) is stronger than inequality (13) when $\hat{\beta} - \eta$ belongs in T additionally to $\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subset T$.

3 What is the low-dimensional set T? Application to Lasso and Group-Lasso

We now provide applications of the above result to three different penalty functions commonly used in high-dimensional settings. Throughout this section, for any cone $T \subseteq \mathbb{R}^p$, let $\phi(T)$ be the smallest singular value of $\Sigma^{1/2}$ restricted to T, i.e., $\phi(T) = \min_{u \in T : \|u\| = 1} \|\Sigma^{1/2} u\|$. Further consider

(N1) The features are normalized such that $\Sigma_{jj} \leq 1$ for all $1 \leq j \leq p$.

3.1 Constrained Lasso

Let $R > 0$ be a fixed parameter. Our first example studies the constrained lasso penalty [38]

$$h(\beta) = +\infty \quad \text{if } \|\beta\|_1 > R \qquad \text{and} \qquad h(\beta) = 0 \quad \text{if } \|\beta\|_1 \leq R, \tag{15}$$

i.e., h is the convex indicator function of the ℓ_1-ball of radius $R > 0$. Applying the above result requires two ingredients: proving that the error vectors $\{\hat{\beta} - \beta^*, \eta - \beta^*\}$ belong to some set T with high probability, and proving that $r_n = n^{-1/2}\gamma(T, \Sigma)$ is small. Define for any real $k \geq 1$,

$$T_{\texttt{lasso}}(k) := \{u \in \mathbb{R}^p : \|u\|_1 \leq \sqrt{k}\|u\|\}. \tag{16}$$

The parameter k above will typically be a constant times $s = \|\beta^*\|_0$, the sparsity of β^*. If $R = \|\beta^*\|_1$, then the triangle inequality reveals that the error vectors of $\hat{\beta}$ and η satisfy

$$\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subseteq T := \{u \in \mathbb{R}^p : \|u_{S^c}\|_1 \leq \|u_S\|_1\}, \tag{17}$$

where $S = \{j = 1, ..., p : \beta_j^* \neq 0\}$ is the support of the true β^* and u_S is the restriction of u to S. By the Cauchy-Schwarz inequality $\|u_S\|_1 \leq \sqrt{s}\|u_S\|_2$, thus T in (17) satisfies $T \subset T_{\texttt{lasso}}(4s)$.

[1] The constant 1 can be replaced by another absolute constant; this will only change B_3 to a different constant.

Lemma 3.1. *If **(N1)** holds and $k \geq 1$, then we have $\gamma(T, \Sigma) \lesssim \phi(T)^{-1}\sqrt{k \log(2p/k)}$ for any cone $T \subset T_{lasso}(k)$ where $T_{lasso}(k)$ is defined in* (16)*.*

Hence under **(N1)** and by setting $k = 4s$ and $r_n = \phi(T)^{-1}\sqrt{s \log(ep/s)/n}$ we have in the linear model with squared loss that, with high probability,

$$\|\Sigma^{1/2}(\eta - \hat{\beta})\| \lesssim L\phi(T)^{-1/2}(s \log(ep/s)/n)^{1/4}(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| + \|\Sigma^{1/2}(\eta - \beta^*)\|) \quad (18)$$

and we have established that η is a first order expansion of $\hat{\beta}$ with respect to the norm $\|\cdot\|_\Sigma$ if $s \log(ep/s)/n \to 0$. It is informative to study the order of magnitude of the right hand side in (18). For that purpose, the following Lemma gives explicit bounds on $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|$ and $\|\Sigma^{1/2}(\eta - \beta^*)\|$.

Lemma 3.2. *Consider the linear model with squared loss* (10) *and assume **(A2)**. Let $\hat{\beta}, \eta$ in* (1) *and* (5) *with penalty* (15)*. Then if $R = \|\beta^*\|_1$, we have with probability at least $1 - 2e^{-nr_n^2}$,*

$$\|\Sigma^{1/2}(\eta - \beta^*)\| \lesssim L\sigma^* r_n, \quad \text{and} \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| \lesssim L\sigma^* r_n (1 - C_6 L^2 r_n)^{-1}, \quad (19)$$

where $r_n = \phi(T)^{-1}\sqrt{s \log(ep/s)/n}$ and $(\sigma^*)^2 = (\varepsilon_1^2 + ... + \varepsilon_n^2)/n$.

The above lemma provides a slight improvement in the rate compared to [17, Theorem 11.1(a)]. Combined with inequality (18), we have established that $\|\Sigma^{1/2}(\hat{\beta} - \eta)\| \lesssim L^2\sigma^* r_n^{3/2}$. If $r_n \to 0$ (e.g., if $s \log(ep/s)/n \to 0$ while $\phi(T)$ stays bounded away from 0), this means that the distance $\|\Sigma^{1/2}(\hat{\beta} - \eta)\|$ between $\hat{\beta}$ and η is an order of magnitude smaller than the risk bounds in (19).

Inclusion (17) is granted regardless of the loss ℓ, as soon as β^* lies on the boundary of $\{\beta \in \mathbb{R}^p : \|\beta\|_1 = R\}$. In logistic regression, i.e., the setting of Proposition 2.2 with the constrained Lasso penalty (15), inequality (13) yields that with high probability, $\|\eta - \hat{\beta}\|_K \lesssim L[r_n^{1/2} + L^{1/2}(1 + r_n^3\sqrt{n})\mathcal{E}^{1/2}]\mathcal{E}$. An extra term appears compared to the squared loss. In order to obtain a first-order expansion as in (2) requires $r_n \to 0$ as well as $(1 + r_n^3\sqrt{n})\mathcal{E}^{1/2} \to 0$. These conditions can be obtained if risk bounds such as (19) are available, see [35, 1] or Proposition 3.4 and its proof in Appendix F for applicable general techniques. A more detailed discussion of Logistic Lasso is given in the next subsection.

3.2 Penalized Lasso

We now consider the ℓ_1-norm penalty

$$h(\beta) = \lambda\|\beta\|_1 \qquad \text{for some} \qquad \lambda \geq 0. \quad (20)$$

Here, the fact that $\hat{\beta} - \beta^*, \eta - \beta^* \in T$ for some low-dimensional cone T is not granted almost surely, in that regard the situation differs from the constrained Lasso case in (17). We may find such low-dimensional cone T simultaneously for $\hat{\beta}, \eta$ for both the squared loss and logistic loss as follows, using ideas from [35, 11]. Let f_n be the convex function so that the objective in (1) is equal to $f_n(\beta) + h(\beta)$ and let g_n be the convex function so that the objective in (5) is $g_n(\beta) + h(\beta)$. Since $\hat{\beta}$ and η are solutions of the corresponding optimization problems (1) and (5),

$$\begin{aligned} h(\hat{\beta}) - h(\beta^*) &\leq f_n(\beta^*) - f_n(\hat{\beta}) \leq \nabla f_n(\beta^*)^T(\beta^* - \hat{\beta}), \\ h(\eta) - h(\beta^*) &\leq g_n(\beta^*) - g_n(\eta) \leq \nabla g_n(\beta^*)^T(\beta^* - \eta). \end{aligned} \quad (21)$$

Since $\nabla g_n(\beta^*) = \nabla f_n(\beta^*)$, both η and $\hat{\beta}$ belong to the set $\hat{T} = \{b \in \mathbb{R}^p : h(b) - h(\beta^*) \leq \nabla f_n(\beta^*)^T(b - \beta^*)\}$. Next, for both the squared loss and the logistic loss, $\nabla f_n(\beta^*)$ has subGaussian coordinates under **(A2)**. Combining these remarks, we obtain the following, proved in supplement.

Lemma 3.3. *Let h be as in* (20)*. Consider the linear model* (10) *and assume **(A2)**, **(N1)**. Let $\xi > 0$ be a constant and let $\lambda = L\sigma^*(1 + 3\xi)\sqrt{2 \log(p/s)/n}$ where $(\sigma^*)^2 = (\varepsilon_1^2 + \ldots + \varepsilon_n^2)/n$ and $\|\beta^*\|_0 = s$. Then*

$$\mathbb{P}\left[\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subset T\right] \geq 1 - \frac{2}{\xi^2 \log(p/s)(p/s)^\xi} \text{ where } T = T_{lasso}\left(s(6 + 2\xi^{-1})^2\right). \quad (22)$$

If instead the logistic regression model and assumptions of Proposition 2.2 are fulfilled and $\lambda = (L/2)(1 + 3\xi)\sqrt{2 \log(p/s)/n}$, then the previous display (22) *also holds.*

The set T above is the set $T_{\texttt{lasso}}(k)$ in (16) with $k = s(6 + 2\xi^{-1})^2$. Eq. (22) defines a low-dimensional cone T that contains both error vectors $\hat{\beta} - \beta^*, \eta - \beta^*$ for the squared loss and the logistic loss. The Gaussian width of the set T in (18) is already bounded in Lemma 3.1. Hence the Gaussian width of T in the previous lemma is bounded from above as in the previous section, i.e., $\gamma(\Sigma, T) \lesssim \phi(T)^{-1}(6 + 2\xi^{-1})\sqrt{s\log(2p/s)}$ by Lemma 3.1, and the "slow rate" inequality (18) again holds with high probability, where $\phi(T)$ denotes the restricted eigenvalue of the set T of the previous lemma. Risk bounds similar to (19) are given below. We emphasize here the fact that the error vectors of the Lasso belong to the cone (22) with high probability is not new: this is a powerful technique used throughout the literature on high-dimensional statistics starting from [11, 35]. The novelty of our results are inequalities such as (18) which shows that the distance $\|\Sigma^{1/2}(\hat{\beta} - \eta)\|$ is an order of magnitude faster than the minimax risk $\sqrt{s\log(ep/s)/n}$. We will now state a result similar to Lemma 3.2 for linear and logistic lasso.

Proposition 3.4. *Consider the penalized lasso estimator $\hat{\beta}$ given by*

$$\hat{\beta} := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) + \lambda\|\beta\|_1,$$

*where ℓ is either the squared or logistic loss and λ is chosen as in Lemma 3.3 for some $\xi > 0$. Assume **(A1)**, **(A2)**. With T defined in* (22)*, assume that $\exists\theta > 0$ s.t. for all $u \in T$ with $\|u\|_K \le 1$,*

$$\theta^2\|u\|_K^2 \le \tfrac{1}{n}\textstyle\sum_{i=1}^n \left\{\ell(Y_i, X_i^\top\beta^* + X_i^\top u) - \ell(Y_i, X_i^\top\beta^*) - u^\top X_i\ell'(Y_i, X_i^\top\beta^*)\right\}, \tag{23}$$

as well as

$$L(2 + 5\xi)\sqrt{2s\log(p/s)/n} \le B_3^{1/2}\phi(T)\theta^2 \times \begin{cases} 1/\sigma^*, & \text{for } \ell\text{, the squared loss,} \\ 2, & \text{for } \ell\text{, the logistic loss.} \end{cases} \tag{24}$$

Then with probability at least $1 - 2/(\xi^2\log(p/s)(p/s)^\xi)$,

$$\|\hat{\beta} - \beta^*\|_K \le \frac{L(2 + 5\xi)}{B_3^{1/2}\phi(T)\theta^2}\sqrt{\frac{2s\log(p/s)}{n}} \times \begin{cases} \sigma^*, & \text{for } \ell\text{, the squared loss,} \\ 0.5, & \text{for } \ell\text{, the logistic loss.} \end{cases} \tag{25}$$

The proof is given Appendix F. Assumption (23) is the classical restricted strong convexity condition and we verify this for linear and logistic loss in Proposition F.1. Results similar to Proposition 3.4 are known in the literature [35] but the main novelty of our result is that the tuning parameter λ is of order $\sqrt{\log(p/s)/n}$ and not $\sqrt{\log(p)/n}$ which proves the minimax optimal rate.

Faster rates for the penalized lasso. Fast rates for the Lasso can be obtained using the second inequality of Theorem 2.1, which when specialized to the squared loss gives (12). To verify the main additional assumption of $\hat{\beta} - \eta \in T$, we prove sparsity of η and $\hat{\beta}$. Since $\hat{\beta}, \eta$ are defined through a penalized quadratic problem, we can leverage existing results in the literature that imply that $\eta, \hat{\beta}$ satisfies $\|\eta\|_0 \vee \|\hat{\beta}\|_0 \le \tilde{C}s$ under suitable conditions on the design and as long as $s\log(ep/s)/n$ is small enough, for some constant $\tilde{C}$ that depends on the restricted singular values of Σ; cf., e.g.,[44, Lemma 1], [9, Theorem 3] [22, Lemma 3.5], [4, Lemma 6.1]. We prove such as result for the Group-Lasso in Proposition 3.7 below. Now we define the cones T_0 and T as the sets

$$T_0 := \{u \in \mathbb{R}^p : \|u\|_0 \le (2\tilde{C} + 1)s\} \subset T = \{u \in \mathbb{R}^p : \|u\|_1 \le (2\tilde{C} + 1)^{1/2}\sqrt{s}\|u\|\}. \tag{26}$$

where the inclusion is obtained thanks to the Cauchy-Schwarz inequality. Then $\{\eta - \hat{\beta}, \hat{\beta} - \beta^*, \eta - \beta^*\} \subset T$ with high probability, the Gaussian width $\gamma(T, \Sigma)$ is bounded by Lemma 3.1 and the second inequality of Theorem 2.1 yields

$$\|\Sigma^{1/2}(\eta - \hat{\beta})\| \lesssim L^2 r_n\mathcal{E}, \quad \text{where} \quad r_n = \phi(T)^{-1}(s\log(ep/s)/n)^{1/2}.$$

Since $\mathcal{E} \lesssim r_n$ with high probability by known prediction bounds for the Lasso (see Proposition 3.4 and its proof in Appendix F for rates with squared and logistic loss), we obtain that with high probability,

$$\|\Sigma^{1/2}(\eta - \hat{\beta})\| \lesssim L^3\phi^{-2}(T)s\log(ep/s)/n = L^3 r_n^2, \tag{27}$$

a rate that is the square of the minimax rate r_n, hence much smaller. For squared loss, this rate is also faster than the rate obtained in (18) which is of order $r_n^{3/2}$. This faster rate is obtained

thanks to the inclusion $\{\hat{\beta}-\eta, \hat{\beta}-\beta^*, \eta-\beta^*\} \subset T$, whereas in the setting of (18) we only had $\{\hat{\beta}-\beta^*, \eta-\beta^*\} \subset T$ but not $\hat{\beta}-\eta \in T$. To our knowledge, the only result in the literature similar to the above bounds is given by [22, Theorem 5.1]. This result from [22] shows that (27) holds for squared loss, provided that the covariance Σ satisfies (a) the minimal singular value of Σ is at least $c_3 > 0$, (b) the maximal singular value of Σ is at most c_4, and (c) the covariance matrix Σ satisfies

$$\max_{A\subset[p]:|A|\leq c_5 s} \max_{j\in A} \sum_{j\in A^c} |\Sigma_{ij}| \leq c_6. \tag{28}$$

Our results show that a first order expansion for the Lasso can be obtained using the slow rate bound (11) without the requirement that the spectral norm of Σ is bounded, and for the fast rate without the stringent assumption (28) on the correlations of Σ. Not only do our results generalize Theorem 5.1 from [22] to more general Σ, Theorem 2.1 shows how to obtain first-order expansion η beyond the squared loss (e.g. logistic loss) and beyond the ℓ_1-penalty of the lasso: the previous subsection tackles the constrained Lasso penalty (15) and the next subsection tackles the Group-Lasso penalty.

Sparsity of η for any general loss function is proved in Proposition 3.7. This alone does not imply inclusion of $\eta-\hat{\beta}$ in a low-dimensional set without sparsity of $\hat{\beta}$. Sparsity of $\hat{\beta}$ for general loss function is not well-studied but for logistic loss function Section D.4 of the supplement of [10] proves a sparsity bound of the form $\|\hat{\beta}\|_0 \leq \tilde{C}s$, similar to the squared loss. Unfortunately the proof there requires $\lambda \gtrsim \sqrt{\log p/n}$ instead of condition $\lambda \gtrsim \sqrt{\log(p/s)/n}$ used in Lemma 3.3 above and in [26, 37, 7, 4, 2].

3.3 Group-Lasso

Consider now a partition of $\{1, ..., p\}$ into M groups $G_1, ..., G_M$. For simplicity, we assume that the groups have the same size $d = p/M$, which is typically the case in multitask learning with d tasks and M shared features. The Group-Lasso penalty studied in this subsection is

$$h(\hat{\beta}) = \lambda \sum_{k=1}^M \|\beta_{G_k}\| \qquad \text{where } \beta_{G_k} \in \mathbb{R}^{|G_k|} \text{ is the restriction } (\beta_j, j \in G_k). \tag{29}$$

In both the linear model with squared loss and in logistic regression with the logistic loss, we now show that $\hat{\beta}-\beta^*$ and $\eta-\beta^*$ belong to a low-dimensional cone (Lemma 3.5), and that the Gaussian width of this cone is bounded from above by $\sqrt{s}(\sqrt{d}+\sqrt{2\log(M/s)})$ where s is the number of groups with $\beta^*_{G_k} \neq 0$ (Lemma 3.6).

Lemma 3.5. *Consider the linear model* (10) *and assume that* $\max_{k=1,...,M} \|\Sigma_{G_k,G_k}\|_{op} \leq 1$ *and that each group has the same size* $|G_k| = d = p/M$. *Let* $\xi > 0$ *and set* $\lambda = L\sigma^*(1+\xi)[\sqrt{d}+(1+2\xi)\sqrt{2\log(M/s)}]$ *where* $(\sigma^*)^2 = (\sum_{i=1}^n \varepsilon_i^2)/n$ *and* s *is the number of groups with* $\beta^*_{G_k} \neq 0$. *Then*

$$\mathbb{P}\left(\{\hat{\beta}-\beta^*, \eta-\beta^*\} \subset T\right) \geq 1 - 2/\left(2\xi^2 \log(M/s)(M/s)^{\xi}\right). \tag{30}$$

for $T = \{\delta \in \mathbb{R}^p : \sum_{k=1}^M \|\delta_{G_k}\| \leq \sqrt{s}\|\delta\|_2(2+3\xi^{-1})\}$. *If instead the logistic regression model and assumptions of Proposition 2.2 are fulfilled and* λ *is as above with* $\sigma^* = 1/2$, *then* (30) *also holds.*

The fact that the Group-Lasso belongs with high probability to a low-dimensional cone has been used before to prove risk bounds, e.g., [31, 5]. However the tuning parameter in the above lemma is smaller than that used in these works and using such cones to prove first expansion as in the present paper are, to our knowledge, novel.

Lemma 3.6. *Assume that* $\max_{k=1,...,M} \|\Sigma_{G_k,G_k}\|_{op} \leq 1$ *and that each group has the same size* $|G_k| = d = p/M$. *The set* T *defined in the previous lemma satisfies* $\gamma(T,\Sigma) \lesssim C(\xi)\phi(T)^{-1}\sqrt{sd + s\log(M/s)}$ *for some constant* $C(\xi)$ *that depends only on* ξ.

Hence if the number of groups M, the group-sparsity s (number of groups such that $\beta^*_{G_k} \neq 0$) and the group size $d = p/M$ satisfy $(sd + s\log(M/s))/n \to 0$ while $\phi(T)$ is bounded away from 0, the above Lemmas combined with Theorem 2.1 imply that η is a first-order expansion of $\hat{\beta}$ for both the squared loss in linear regression and logistic loss in the logistic model. We leverage this result to obtain an exact risk identity for the Group-Lasso in the next section.

Proposition 3.7. ***(A1), (A2).*** *Let the setting of Lemma 3.6 be fulfilled. Fix* λ *as in Lemma 3.5 for both squared and logistic loss for some* $\xi > 0$ *and* T *be the cone defined in Lemma 3.5. If* $\|K\|_{op} \leq C_{\max} < \infty$ *and the assumptions of Proposition 3.4 hold, then*

$$\mathbb{P}\left(|\{k \in [M] : \eta_{G_k} \neq 0\}| \leq s\tilde{C}\right) \geq 1 - 2/(\xi^2 \log(M/s)(M/s)^{\xi}),$$

where $\tilde{C} := 1 + C_{\max}\{2(3+\xi)(1+\xi^{-1})\}^2 B_3^2 \phi(T)^{-2}$. *For the squared loss, the same holds for* $\hat{\beta}$ *with* $\tilde{C}$ *replaced by* $(1+o(1))\tilde{C}$ *provided* $\phi(T)^{-1}\sqrt{sd + s\log(M/s)}/\sqrt{n} \to 0$.

The proof is given in Appendix G. For the Lasso the assumption of $\|K\|_{op} \leq C_{\max}$ can be relaxed to a bound on the sparse maximal eigenvalue of K using devices from [44, Lemma 1], [47, Corollary 2], [8, Lemma 3] or [4, Proposition 7.4]. See also [31, Theorem 3.1] and [29, Lemma 6] for similar results for the Group-Lasso, although with a larger tuning parameter than in Proposition 3.7.

For the squared loss, if the condition number of Σ stays bounded then $C_{\max}/\phi(T)^{-2}$ is also bounded. Then if $r_n = \sqrt{sd + s\log(M/s)}/\sqrt{n} \to 0$, Proposition 3.7 yields that both $\hat{\beta} - \eta$ belongs to the cone $\{\delta \in \mathbb{R}^p : \sum_{k=1}^M \|\delta_{G_k}\| \leq (1+o(1))(2\tilde{C}s)^{1/2}\|\delta\|_2\}$, which yields the "fast rate" bound (12).

4 Application to exact risk identities

In the linear model with the squared loss and identity covariance ($\Sigma = I_p$), the expansion η in (5) is particularly simple: η becomes the proximal operator of the penalty h at the point $z = \beta^* + n^{-1/2}\sum_{i=1}^n \varepsilon_i X_i$, i.e, $\eta = \mathsf{prox}_h(z)$ where $\mathsf{prox}_h(x) = \mathrm{argmin}_{b\in\mathbb{R}^p} \|x-b\|^2/2 + h(b)$. Hence the loss $\|\eta - \beta^*\|$ of η has a simple form and if a first-order expansion (2) is available, for instance for the Lasso or Group-Lasso as a consequence of the Lemmas of the previous section, then the loss $\|\hat{\beta} - \beta^*\|$ is exactly the loss of $\mathsf{prox}(z)$ up to a smaller order term. Let us emphasize that the next result and following discussion provide exact risk identities for the loss $\|\hat{\beta} - \beta^*\|$ (as in (32) below), and not only upper bounds up to multiplicative constants.

Theorem 4.1. *[Exact Risk Identity] Consider the linear model* (10) *and the regularized problem* (1) *with an arbitrary proper convex function* $h(\cdot)$. *Assume that* $X_1, \ldots, X_n$ *are iid* $N(0, I_p)$ *independent of* $\varepsilon_1, ..., \varepsilon_n$ *and set* $\sigma^* = (\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2)^{1/2}$. *Then with probability at least* $1 - 2\exp(-t^2/2)$,

$$\left| \|\hat{\beta} - \beta^*\| - \mathbb{E}_Z\left[\|\beta^* - \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)\|^2\right]^{1/2} \right| \leq \frac{\sigma^*(t+1)}{n^{1/2}} + \|\hat{\beta} - \eta\| \tag{31}$$

where $Z = \frac{1}{n^{1/2}\sigma^*}\sum_{i=1}^n \varepsilon_i X_i \sim N(0, I_p)$ *and* $\mathbb{E}_Z$ *denotes the expectation with respect* Z.

Theorem 4.1 is a generalization of Corollary 5.2 of [22] where the result is stated for $h(\beta) = \lambda\|\beta\|_1$ with $\lambda \gtrsim \sigma^*\sqrt{2\log(p)/n}$. For the case of lasso, either in its constrained form with tuning parameter chosen as in Lemma 3.3 or the penalized Lasso with tuning parameter as in Lemma 3.3, inequality (18) holds thanks to (17) and Lemma 3.1 for the constrained lasso, and thanks to Lemmas 3.1 and 3.3 for the penalized lasso. Hence for both the constrained and penalized lasso, if $\Sigma = I_p$ with Gaussian design, the second term on the right hand side of (31) is $O_p(\sigma^*/\sqrt{n}) + O_p(s\log(ep/s)/n)^{1/4})(\|\eta - \beta^*\| + \|\hat{\beta} - \beta^*\|)$. Hence if $s, n, p \to +\infty$ with $s\log(ep/s)/n \to 0$ and $s/p \to 0$ then (31) implies

$$\|\hat{\beta} - \beta^*\| = (1 + o_p(1))\mathbb{E}_Z[\|\beta^* - \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)\|^2]^{1/2}. \tag{32}$$

For the penalized lasso, since η represents a soft-thresholding operator which can be written in closed form, Theorem 4.1 allows a refined study of the risk of $\hat{\beta}$; see [14, Theorem 5.1]. Similarly for the group lasso, we have from Lemmas 3.5 and 3.6 that $\|\eta - \hat{\beta}\| = O_p((sd + s\log(M/s))^{1/4}/n^{1/4})\|\hat{\beta} - \beta^*\|$ (slow rate) which is again negligible relative to $\|\hat{\beta} - \beta^*\|$ if $(sd + s\log(M/s))/n \to 0, s/M \to 0$. Thus, (32) again holds true. For the group lasso $\eta = \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)$ represents the Block James-Stein estimator in the sequence model; see [13, Section 2.1].

Extending Corollary 5.2 of [22] to more general loss/penalty functions, the above device lets us characterize the risk $\|\hat{\beta} - \beta^*\|$: Up to a multiplicative constant of order $1 + o_p(1)$, the risk is the same as the risk of the proximal of h in the Gaussian sequence model where one observes $N(\beta^*, (\sigma^*)^2/n)$.

5 Application to inference

The second application we wish to mention is related to confidence intervals in the linear model when the squared loss is used and $X_1, ..., X_n$ are iid Gaussian $N(0, \Sigma)$. Assume that one is interested in constructing a confidence interval for a specific linear combination $a^T\beta^*$ for some $a \in \mathbb{R}^p$.

Further assume, for simplicity, that Σ is known and that a is normalized with $\|\Sigma^{-1/2}a\| = 1$. Then previous works on *de-biasing* [45, 46, 20, 21, 40, 22, 4] suggests, given an estimator $\hat{\beta}$ that may be biased, to consider the bias-corrected estimator $\hat{\theta}$ defined by $\hat{\theta} = a^T\hat{\beta} + \|z_a\|^{-2}z_a^T(y - \mathsf{X}\hat{\beta})$, where $y = (Y_1, ..., Y_n)$ is the response vector and X is the design matrix with rows $X_1, ..., X_n$ and $z_a = \mathsf{X}\Sigma^{-1}a \sim N(0, I_n)$ is sometimes referred to as a score vector for the estimation of $a^T\beta^*$.

Proposition 5.1. *Assume that $X_1, ..., X_n$ are iid $N(0, \Sigma)$ and is independent of $\varepsilon = (\varepsilon_1, ..., \varepsilon_n) \sim N(0, I_n)$. Assume that for some cone T and $r_n = \gamma(T, \Sigma)/\sqrt{n}$ we have*

$$\mathbb{P}(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| + \|\Sigma^{1/2}(\eta - \beta^*)\| \leq C_7 r_n, \{\eta - \hat{\beta}, \eta - \beta^*, \hat{\beta} - \beta^*\} \subset T) \geq 1 - \alpha. \quad (33)$$

Then for some T_n with the t-distribution with n degrees-of-freedom, with probability $1-\alpha-4e^{-nr_n^2/2}$,

$$\sqrt{n}(\hat{\theta} - a^T\beta^*) - T_n = O_p((1 + r_n))\|\Sigma^{1/2}(\eta - \beta^*)\| + O_p(\sqrt{n}r_n)\|\Sigma^{1/2}(\eta - \hat{\beta})\|, \quad (34)$$

$$= O_p(r_n(1 + r_n)) + O_p(\sqrt{n}r_n^3). \quad (35)$$

Because T_n has t distribution with n degrees of freedom, asymptotically $\mathbb{P}(|T_n| \leq 1.96) \to 0.95$ and hence from (35), we get that $\mathbb{P}(n^{1/2}|\hat{\theta} - a^\top\beta^*| \leq 1.96) \to 0.95$ if $r_n^3\sqrt{n} \to 0$. Therefore, $[\hat{\theta} - 1.96/n^{1/2}, \hat{\theta} + 1.96/n^{1/2}]$ represents a 95% confidence interval for $a^\top\beta^*$. Conclusion (34) is a consequence of Theorem 2.1.

Lasso. Eq. (33) is satisfied for the penalized Lasso for $r_n = \sqrt{s\log(ep/s)/n}$ and the cone T in (26), in situations where $\|\hat{\beta}\|_0 \leq \tilde{C}s$ with high probability as explained in the discussion surrounding (26). In order to construct confidence interval based on (34), the right hand side of (35) needs to converges to 0. This is the case if $r_n \to 0$ and $\sqrt{n}r_n^3 \to 0$. For the Lasso with $r_n = s\log(ep/s)/n$, this translates to the sparsity condition $s^3\log(ep/s)^3/n^2 \to 0$, i.e., $s = o(n^{2/3})$ up to logarithmic factors. Hence the first order expansion results of the present paper lets us derive de-biasing results for the Lasso beyond the condition $s \lesssim \sqrt{n}$ required in the early results [46, 20, 40] on de-biasing (other recent approaches, [22, 4] also allow to prove such result beyond $s \lesssim \sqrt{n}$). Moreover, the above proposition is general and apply to any regularized estimator such that (33) holds, with suitable bounds on the Gaussian complexity $\gamma(T, \Sigma)$. For $s \ggg n^{2/3}$, the estimator $\hat{\theta}$ requires an adjustment for asymptotic normality in the form a degree-of-freedom adjustment [4].

Group-Lasso. If s is the number of non-zero groups, $r_n = \sqrt{sd + s\log(M/s)}/\sqrt{n}$ and the condition number of Σ is bounded, then (33) holds thanks to Proposition 3.7, the last paragraph of Section 3.3 and the risk bound (86). Here, (35) is $o(1)$ if and only if $(sd + s\log(M/s))/n^{2/3} \to 0$. This improves the sample size requirement of [34], although Σ is assumed known in Proposition 5.1.

6 Proof sketch of Theorem 2.1 **(Detailed proofs are given in Appendix E)**

Theorem 6.1. *Define $\hat{K} := n^{-1}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)X_iX_i^\top$. Under assumption **(A1)**, we have*

(i) If $\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subseteq T$ then $\|\hat{\beta} - \eta\|_K \lesssim Q_{n,1}^{1/2}\mathcal{E} + B^{1/2}Z_n^{1/2}\mathcal{E}^{3/2}$.

(ii) If $\{\hat{\beta} - \eta, \hat{\beta} - \beta^*, \eta - \beta^*\} \subseteq T$ then $\|\hat{\beta} - \eta\|_K \lesssim Q_{n,2}\mathcal{E} + BZ_n\mathcal{E}^2$,

where

$$Q_{n,1} = \sup_{u\in T}\left|\frac{u^\top\hat{K}u}{\|u\|_K^2} - 1\right|,\ Q_{n,2} = \sup_{u,v\in T}\frac{|u^\top(\hat{K} - K)v|}{\|u\|_K\|v\|_K}\ \text{ and }\ Z_n = \sup_{u\in T}\frac{1}{n}\sum_{i=1}^n\frac{|X_i^\top u|^3}{\|u\|_K^3}.$$

Theorem 6.1 follows from the strong convexity of the objective function of η with respect to the norm $\|\cdot\|_K$ (cf. for instance, Lemma 1 of [6]) combined with Taylor expansions of the loss ℓ. Next, to prove Theorem 2.1, it remains to bound $Q_{n,1}(T), Q_{n,2}(T)$ and $Z_n(T)$. The quadratic processes $Q_{n,1}(T), Q_{n,2}(T)$ and cubic process $Z_n(T)$ can be bounded in terms of $\gamma(T, \Sigma)$ using generic chaining results, Theorem 1.13 of Mendelson [33] and Eq. (3.9) of [32], as follows.

Proposition 6.2. *[Control of $Q_{n,1}, Q_{n,2}$ and Z_n] Under assumptions **(A1)** and **(A2)**, we have*

(i) With probability $1 - 2\exp(-C_8t^2\gamma^2(T, \Sigma))$,

$$\max\{Q_{n,1}(T), Q_{n,2}(T)\} \leq C_9B_2B_3L^2\left(tn^{-1/2}\gamma(T, \Sigma) + t^2n^{-1}\gamma^2(T, \Sigma)\right).$$

(ii) With probability $1 - 2\exp(-C_{10}t\log n)$, $\quad Z_n(T) \leq C_{11}B_3^{3/2}L^3\left(1 + n^{-1}\gamma^3(T, \Sigma)\right)t^3$.

References

[1] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *The Annals of Statistics*, 47 (4):2117–2144, 2019.

[2] Pierre C Bellec. The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv:1804.01230*, 2018. URL https://arxiv.org/pdf/1804.01230.pdf.

[3] Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

[4] Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *arXiv:1902.08885*, 2019. URL https://arxiv.org/pdf/1902.08885.pdf.

[5] Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Towards the study of least squares estimators with convex penalty. In *Seminaire et Congres, to appear*, number 39. Societe mathematique de France, 2017. URL https://arxiv.org/pdf/1701.09120.pdf.

[6] Pierre C Bellec, Arnak S Dalalyan, Edwin Grappin, and Quentin Paris. On the prediction loss of the lasso in the partially labeled setting. *Electronic Journal of Statistics*, 12(2):3443–3472, 2018.

[7] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets lasso: Improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1670. URL https://arxiv.org/pdf/1605.08651.pdf.

[8] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[9] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. URL http://dx.doi.org/10.1214/14-AOS1204.

[10] Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619, 2016.

[11] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009. doi: 10.1214/08-AOS620. URL http://dx.doi.org/10.1214/08-AOS620.

[12] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[13] T Tony Cai and Harrison H Zhou. A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, 37(2):569–595, 2009.

[14] Emmanuel J Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15: 257–325, 2006.

[15] Antoine Dedieu. Error bounds for sparse classifiers in high-dimensions. *arXiv preprint arXiv:1810.03081*, 2018.

[16] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015.

[17] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[18] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 1–6, 2012. doi: 10.1214/ECP.v17-2079. URL http://ecp.ejpecp.org/article/view/2079.

[19] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

[20] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[21] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[22] Adel Javanmard and Andrea Montanari. De-biasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics, to appear*, 2015.

[23] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[24] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 45, pages 7–57. Institut Henri Poincaré, 2009.

[25] A. K. Kuchibhotla. Deterministic Inequalities for Smooth M-estimators. *ArXiv e-prints:1809.05172*, September 2018.

[26] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 04 2018. doi: 10.1214/17-AOS1562. URL `https://doi.org/10.1214/17-AOS1562`.

[27] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[28] Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.

[29] Han Liu and Jian Zhang. Estimation consistency of the group lasso and its applications. In *Artificial Intelligence and Statistics*, pages 376–383, 2009.

[30] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

[31] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 08 2011. doi: 10.1214/11-AOS896. URL `http://dx.doi.org/10.1214/11-AOS896`.

[32] Shahar Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geometric and Functional Analysis*, 20(4):988–1027, 2010.

[33] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, 2016. doi: 10.1016/j.spa.2016.04.028. URL `https://ideas.repec.org/a/eee/spapps/v126y2016i12p3652-3680.html`.

[34] Ritwik Mitra and Cun-Hui Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2):1829–1873, 2016.

[35] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[36] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[37] Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[39] Sara van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86, 2014.

[40] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

[41] Aad van der Vaart. Part iii: Semiparameric statistics. *Lectures on Probability Theory and Statistics*, pages 331–457, 2002.

[42] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

[44] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

[45] Cun-Hui Zhang. Statistical inference for high-dimensional data. *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report*, (48):28–31, 2011.

[46] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[47] Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 11 2012. doi: 10.1214/12-STS399. URL https://doi.org/10.1214/12-STS399.

SUPPLEMENT

	Lasso	Group-Lasso, M groups of size $d = p/M$
Tuning parameter	$\lambda \gtrsim [\frac{2}{n} \log \frac{p}{s}]^{\frac{1}{2}}$ in (20)	$\lambda \gtrsim [d + \frac{2}{n} \log \frac{M}{s}]^{\frac{1}{2}}$ in (29)
Minimax rate r_n	$\|\hat{\beta} - \beta^*\|_\Sigma \lesssim r_n$ $r_n = [\frac{2s}{n} \log \frac{p}{s}]^{\frac{1}{2}}$	$\|\hat{\beta} - \beta^*\|_\Sigma \lesssim r_n$ $[\frac{d}{n} + \frac{s}{n} \log \frac{M}{s}]^{\frac{1}{2}}$
Gaussian width bound	$\gamma(T, \Sigma) \lesssim [s \log \frac{p}{s}]^{1/2}$	$\gamma(T, \Sigma) \lesssim [sd + s \log \frac{M}{s}]^{1/2}$ by Lemma 3.6
Restricted Eigenvalue (RE)	$\|\eta - \hat{\beta}\|_\Sigma \lesssim r_n^{3/2}$ by (11) and Lemma 3.3 (Lasso) or Lemma 3.5 (GL)	
$\|\Sigma\|_{op} \vee \|\Sigma^{-1}\|_{op} \leq C$	$\|\eta - \hat{\beta}\|_\Sigma \lesssim r_n^2$ by (12), Proposition 3.7	

Table 1: Summary of rates for $\|\hat{\beta} - \beta^*\|_\Sigma$ and $\|\eta - \hat{\beta}\|_\Sigma$ for the squared loss. For the Lasso, s is the sparsity of β^* while for the Group-Lasso s is the number of non-zero groups in β^*. If $r_n \to 0$ then $\|\eta - \hat{\beta}\|_\Sigma$ is an order of magnitude smaller than $\|\hat{\beta} - \beta^*\|_\Sigma$ and the minimax rate. In this table $\gtrsim$ may hide constants depending on the subgaussian parameter L as well as restricted eigenvalues of Σ, denoted by $\phi(T)$ in the paper.

	Lasso	Group-Lasso, M groups of size $d = \frac{p}{M}$
Tuning parameter	$\lambda \gtrsim [\frac{2}{n} \log \frac{p}{s}]^{\frac{1}{2}}$ in (20)	$\lambda \gtrsim [d + \frac{2}{n} \log \frac{M}{s}]^{\frac{1}{2}}$ in (29)
Minimax rate r_n	$\|\hat{\beta} - \beta^*\|_K \lesssim r_n$ $r_n = [\frac{2s}{n} \log \frac{p}{s}]^{\frac{1}{2}}$ (Prop. 3.4)	$\|\hat{\beta} - \beta^*\|_K \lesssim r_n$ $[\frac{d}{n} + \frac{s}{n} \log \frac{M}{s}]^{\frac{1}{2}}$
Gaussian width bound	$\gamma(T, \Sigma) \lesssim [s \log \frac{p}{s}]^{1/2}$	$\gamma(T, \Sigma) \lesssim [sd + s \log \frac{M}{s}]^{1/2}$ by Lemma 3.6
RSC (cf. Appendix F)	$\|\eta - \hat{\beta}\|_K \lesssim r_n^{3/2}(1 + r_n^3\sqrt{n})$ by (13) and Lemma 3.3 or Lemma 3.5	
$\|K\|_{op} \vee \|K^{-1}\|_{op} \leq C$	$\|\eta - \hat{\beta}\|_K \lesssim r_n^2(1 + r_n^3\sqrt{n})$ by (14) and Proposition 3.7	

Table 2: Summary of rates for $\|\hat{\beta} - \beta^*\|_K$ and $\|\eta - \hat{\beta}\|_K$ for the logistic loss. For the Lasso, s is the sparsity of β^* while for the Group-Lasso s is the number of non-zero groups in β^*. If $r_n \to 0$ as well as $r_n^3\sqrt{n} \to 0$ then $\|\eta - \hat{\beta}\|_K$ is an order of magnitude smaller than $\|\hat{\beta} - \beta^*\|_K$ and the minimax rate. In this table $\gtrsim$ may hide constants depending on the subgaussian parameter L, the constants B_3 and Restricted Strong Convexity (RSC) constants.

Proofs. All Theorems, Lemmas and Propositions from the submission are proved in the present supplement. The results are restated before their proofs for convenience.

A Proofs of Section 3

Lemma 3.1. *If **(N1)** holds and $k \geq 1$, then we have $\gamma(T, \Sigma) \lesssim \phi(T)^{-1}\sqrt{k \log(2p/k)}$ for any cone $T \subset T_{\texttt{lasso}}(k)$ where $T_{\texttt{lasso}}(k)$ is defined in* (16)*.*

Proof of Lemma 3.1. This is a consequence of Lemma 3.6 proved below, by taking p groups of size $d = 1$, i.e., the groups are $G_j = \{j\}$ for each $j = 1, ..., p$ and $M = p$. The condition $\max_{k=1,...,M} \|\Sigma_{G_k,G_k}\|_{op} \leq 1$ necessary to apply Lemma 3.5 is equivalent to the normalization **(N1)**. □

Lemma 3.2. *Consider the linear model with squared loss* (10) *and assume **(A2)**. Let $\hat{\beta}, \eta$ in* (1) *and* (5) *with penalty* (15)*. Then if $R = \|\beta^*\|_1$, we have with probability at least $1 - 2e^{-nr_n^2}$,*

$$\|\Sigma^{1/2}(\eta - \beta^*)\| \lesssim L\sigma^* r_n, \quad \textit{and} \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| \lesssim L\sigma^* r_n(1 - C_{12}L^2 r_n)^{-1}, \tag{19}$$

where $r_n = \phi(T)^{-1}\sqrt{s \log(ep/s)/n}$ and $(\sigma^*)^2 = (\varepsilon_1^2 + ... + \varepsilon_n^2)/n$.

Proof of Lemma 3.2. Since $R = \|\beta^*\|_1$, the inclusion (17) holds by the triangle inequality, i.e., we have $\hat{\beta} - \beta^* \in T$ as well as $\eta - \beta^* \in T$.

Next, we first bound the loss of η. The optimization problem (5) for the squared loss for the penalty (15) can be rewritten as

$$\eta = \underset{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq R}{\operatorname{argmin}} \frac{1}{2}\|\Sigma^{1/2}(\beta - \beta^*) - n^{-1/2}Z\|^2, \qquad \text{where} \qquad Z = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \Sigma^{-1/2} X_i.$$

By optimality of η for the above optimization problem, we have (see, e.g., the properties of convex projections in [3]) that

$$\|\Sigma^{1/2}(\eta - \beta^*)\| \leq \frac{1}{\sqrt{n}} \frac{(\Sigma^{1/2}(\eta - \beta^*))^T Z}{\|\Sigma^{1/2}(\eta - \beta^*)\|} \leq \frac{\sigma^*}{\sqrt{n}} \sup_{u \in T : \|\Sigma^{1/2}u\| = 1} u^T \Sigma^{1/2} Z / \sigma^*.$$

Next, notice that Z/σ^* is L-subgaussian because for any $u \in \mathbb{R}^p$, by independence,

$$\mathbb{E}[\exp(u^T Z)] = \prod_{i=1}^n \mathbb{E}[e^{n^{-1/2}\varepsilon_i X_i^T \Sigma^{-1/2} u}] \leq \prod_{i=1}^n e^{n^{-1}\varepsilon_i^2 L^2 \|u\|^2/2} = e^{(\sigma^*)^2 L^2 \|u\|^2/2}. \tag{36}$$

By a tail bound on suprema of subGaussian processes, we obtain that with probability at least $1 - e^{-t^2}$, inequality $\sup_{u \in T : \|\Sigma^{1/2}u\| = 1} u^T \Sigma^{1/2} Z/\sigma^* \leq C_{13}(\gamma(T, \Sigma) + t)$ holds for some absolute constant C_{13}. We have proved in Lemma 3.1 that $\gamma(T, \Sigma) \leq C_{14}\phi(T)^{-1}\sqrt{s\log(2p/s)}$ for another absolute constant. The choice $t = \phi(T)^{-1}\sqrt{s\log(2p/s)} = r_n\sqrt{n}$ completes the proof for η.

We now prove the bound for $\hat{\beta}$. For the squared loss in the linear model,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq R}{\operatorname{argmin}} \|\mathsf{X}(\beta - \beta^*) - \varepsilon\|^2/(2n)$$

where X is the design matrix with rows $X_1, ..., X_n$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$. The optimality conditions of the above optimization problem yields that

$$\frac{\|\mathsf{X}(\hat{\beta} - \beta^*)\|^2}{n\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|} \leq \frac{\varepsilon^T \mathsf{X}(\hat{\beta} - \beta^*)}{n\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|} = \frac{1}{\sqrt{n}} \frac{(\Sigma^{1/2}(\hat{\beta} - \beta^*))^T Z}{\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|} \leq \frac{\sigma^*}{\sqrt{n}} \sup_{u \in T : \|\Sigma^{1/2}u\| = 1} \frac{u^T \Sigma^{1/2} Z}{\sigma^*}.$$

We have already bounded in the previous paragraph the supremum in the right hand side with probability at least $1 - e^{-nr_n^2}$. It remains to show that the left hand side is larger than $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|(1 - C_{15}r_n)$ with high probability. Since $\hat{\beta} - \beta^* \in T$, an application of [28] to the set $(\Sigma^{1/2}T) \cap \{v \in \mathbb{R}^p : \|v\| = 1\}$ yields that, with probability at least $1 - 2e^{-r_n^2 n}$,

$$\Big| \frac{\|\mathsf{X}(\hat{\beta} - \beta^*)\|}{\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|} - \sqrt{n} \Big| \leq C_{16}(\gamma(T, \Sigma) + \sqrt{n}r_n) \leq C_{17}\sqrt{n}r_n.$$

In the same event, we have $\|\mathsf{X}(\hat{\beta} - \beta^*)\|^2/n \geq (1 - C_{18}r_n)^2 \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2$ and the proof is complete. □

The following Lemma will be useful.

Lemma A.1. *The following, (i) in the linear model and (ii) in the logistic model, hold for any convex penalty h.*

(i) Consider the linear model (10)*, assume **(A2)** and assume that $(\varepsilon_1, ..., \varepsilon_n)$ is independent of $(X_1, ..., X_n)$ and let $(\sigma^*)^2 = (1/n)\sum_{i=1}^n \varepsilon_i^2$. Then almost surely,*

$$\{\eta, \hat{\beta}\} \subset \hat{T} = \{b \in \mathbb{R}^p : \sqrt{n}(h(b) - h(\beta^*)) \leq Z^T \Sigma^{1/2}(b - \beta^*)\} \tag{37}$$

where Z is an $L(\sigma^*)$-subgaussian vector in the sense that $\mathbb{E}\exp(u^T Z) \leq \exp(L^2(\sigma^*)^2\|u\|^2/2)$.

*(ii) Consider the logistic model and assume **(A2)**. Then almost surely*

$$\{\eta, \hat{\beta}\} \subset \tilde{T} = \{b \in \mathbb{R}^p : \sqrt{n}(h(b) - h(\beta^*)) \leq \tilde{Z}^T \Sigma^{1/2}(b - \beta^*)\}$$

where $\tilde{Z}$ is an $L/2$-subgaussian vector in the sense that $\mathbb{E}\exp(u^T \tilde{Z}) \leq \exp((L/2)^2\|u\|^2/2)$.

Proof. (i) We first prove the result in the linear model for the squared loss. Here (21) holds with g_n and f_n defined before (21), so that

$$\sqrt{n}\nabla f_n(\beta^*) = \sqrt{n}\nabla g_n(\beta^*) = \Sigma^{1/2}Z \qquad \text{where} \qquad Z \triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \Sigma^{-1/2}X_i \tag{38}$$

in the linear model for the squared loss. Let $\hat{T} = \{b \in \mathbb{R}^p : h(b) - h(\beta^*) \le Z^T\Sigma^{1/2}(b-\beta^*)\}$. Then both η and $\hat{\beta}$ belong to $\hat{T}$ by (21). We already proved that Z is is $L\sigma^*$-subgaussian in the sense that $\mathbb{E}[\exp(Z^Tu)] \le \exp(L^2(\sigma^*)^2\|u\|^2/2)$ for all $u \in \mathbb{R}^p$ in (36); this completes the proof of (i).

(ii) In logistic regression with the logistic loss, (21) again, holds, i.e., $\{\eta, \hat{\beta}\}$ belong to $\tilde{T} = \{b \in \mathbb{R}^p : h(b) - h(\beta^*)) \le \tilde{Z}^T\Sigma^{1/2}(b-\beta^*)\}$ where

$$\tilde{Z} \triangleq \sqrt{n}\Sigma^{-1/2}\nabla f_n(\beta^*) = \sqrt{n}\Sigma^{-1/2}\nabla g_n(\beta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(Y_i - \frac{1}{1+e^{X_i^T\beta^*}}\right)\Sigma^{-1/2}X_i \tag{39}$$

where g_n and f_n are defined before (21). We now show that $\tilde{Z} \in \mathbb{R}^p$ is a subgaussian vector. Note that $\mathbb{E}[Y_i|X_i] = 1/(1+e^{X_i^T\beta^*})$ so that $\mathbb{E}[\tilde{Z}|X_1,...,X_n] = 0$ and $\mathbb{E}[\tilde{Z}] = 0$. If B is Bernoulli with parameter p, then $\mathbb{E}[e^{t(B-p)}] = pe^{t(1-p)} + (1-p)e^{t(-p)}$ which is maximized at $p = 1/2$, hence $\mathbb{E}[e^{t(B-p)}] \le (e^{t/2} + e^{-t/2})/2$. For any $u \in \mathbb{R}^p$, set $v = \Sigma^{-1/2}u$ and notice that by independence and the law of total expectation,

$$\mathbb{E}[e^{\tilde{Z}^Tu}] = \prod_{i=1}^n \mathbb{E}[e^{(Y_i - \mathbb{E}[Y_i|X_i])\frac{X_i^Tv}{\sqrt{n}}}] \le \prod_{i=1}^n \frac{\mathbb{E}e^{\frac{X_i^Tv}{2\sqrt{n}}} + \mathbb{E}e^{\frac{-X_i^Tv}{2\sqrt{n}}}}{2} \le e^{L^2\|\Sigma^{1/2}v\|^2/8} = e^{L^2\|u\|^2/8},$$

where for the last inequality we use that each X_i is L-subgaussian. Hence $\tilde{Z}$ is $L/2$-subgaussian for the logistic loss. □

Lemma 3.3. *Let h be as in* (20)*. Consider the linear model* (10) *and assume* ***(A2), (N1)****. Let $\xi > 0$ be a constant and let $\lambda = L\sigma^*(1+3\xi)\sqrt{2\log(p/s)/n}$ where $(\sigma^*)^2 = (\varepsilon_1^2 + \ldots + \varepsilon_n^2)/n$ and $\|\beta^*\|_0 = s$. Then*

$$\mathbb{P}\left[\{\hat{\beta} - \beta^*, \eta - \beta^*\} \subset T\right] \ge 1 - \frac{2}{\xi^2\log(p/s)(p/s)^\xi} \text{ where } T = T_{lasso}\left(s(6+2\xi^{-1})^2\right). \tag{22}$$

If instead the logistic regression model and assumptions of Proposition 2.2 are fulfilled and $\lambda = (L/2)(1+3\xi)\sqrt{2\log(p/s)/n}$, then the previous display (22) *also holds.*

Proof of Lemma 3.3. We will apply the previous lemma, but first let us derive some properties of subgaussian vectors.

Let U be a random vector valued in $\mathbb{R}^p$ such that each component U_j is 1-subgaussian in the sense that $\mathbb{E}[e^{tU_j^2}] \le e^{t^2/2}$, for each $j = 1, ..., p$. Let $\mu > 0$ be a deterministic real.

Since U_j is 1-subgaussian, $\mathbb{P}(U_j > \sqrt{2x}) \le e^{-x}$ by a Chernoff bound, hence $(U_j)_+^2/2$ is stochastically dominated by an exponential random variable τ with parameter 1 and there exists a probability space on which both U_j and τ are defined such that $U_j \le \sqrt{2\tau}$ holds almost surely. Hence using $|\sqrt{a} - \sqrt{b}|^2 \le a^2 - b^2$ for any $a > b > 0$,

$$\mathbb{E}[(U_j-\mu)_+^2] \le \mathbb{E}[(\sqrt{2\tau}-\mu)_+^2] \le \int_{\mu^2/2}^\infty (\sqrt{2t}-\mu)_+^2 e^{-t}dt \le 2\int_{\mu^2/2}^\infty (t-\mu^2/2)_+e^{-t}dt = 2e^{-\mu^2/2}.$$

The same holds with U_j replaced by $-U_j$. For $\mu = (1+\xi)\sqrt{2\log(p/s)}$ for $\xi \ge 0$, this shows that

$$\mathbb{E}\sum_{j=1}^p (|U_j| - \mu)_+^2 \le \mathbb{E}\sum_{j=1}^p (U_j-\mu)_+^2 + (-U_j-\mu)_+^2 \le 4pe^{-\mu^2/2} = 4s/(p/s)^\xi.$$

By Markov's inequality, for this value of μ and $\xi > 0$,

$$\mathbb{P}\left(\frac{1}{s}\sum_{j=1}^p (|U_j| - \mu)_+^2 \le 2\xi^2\log(p/s)\right) \ge 1 - \frac{4}{2\xi^2\log(p/s)(p/s)^\xi}. \tag{40}$$

By the triangle inequality, on this event, we also have

$$\max_{A\subset\{1,\dots,p\}:|A|=s}\left(\frac{1}{s}\sum_{j\in A}U_j^2\right)^{1/2}\leq \mu+\xi\sqrt{2\log(p/s)}=(1+2\xi)\sqrt{2\log(p/s)}.$$

Furthermore, if $\hat{A}$ denotes the subset achieving the maximum in the left hand side above, any value U_j^2 with $j\notin\hat{A}$ is smaller than the average of the values in $\hat{A}$ and

$$\max_{j\notin\hat{A}}|U_j|\leq \max_{A\subset\{1,\dots,p\}:|A|=s}\left(\frac{1}{s}\sum_{j\in A}U_j^2\right)^{1/2}\leq(1+2\xi)\sqrt{2\log(p/s)}. \tag{41}$$

In linear regression with the squared loss, the vector Z is $L\sigma^*$-subgaussian in the sense that $\mathbb{E}\exp(u^TZ)\leq\exp(L^2(\sigma^*)^2\|u\|^2/2)$ holds. Hence, since Σ satisfies the normalization (**(N1)**), the random vector $U=\Sigma^{1/2}Z/(L\sigma^*)$ satisfies for all $j=1,...,p$ that $\mathbb{E}[\exp(tU_j)]\leq e^{-t^2/2}$ and the bound (41) holds on an event of probability at least equal to the right hand side of (40).

For any $\xi>0$, if $\lambda=L\sigma^*(1+3\xi)\sqrt{2\log(p/s)/n}$ then any $b\in\hat{T}$ where $\hat{T}$ is defined in (37) satisfies

$$0\leq Z^T\Sigma^{1/2}(b-\beta^*)-\lambda\sqrt{n}(\|b\|_1-\|\beta^*\|_1)=L\sigma^*\left(U^T(b-\beta^*)-\lambda\sqrt{n}(L\sigma^*)^{-1}(\|b\|_1-\|\beta^*\|_1)\right).$$

This implies, by replacing λ by its value and using the Cauchy-Schwarz inequality on the support of β^*, that

$$0\leq U^T\delta+(1+3\xi)\sqrt{2\log(p/s)}(\sqrt{s}\|\delta\|_2-\|b_{S^c}\|_1)$$

where $\delta=b-\beta^*$ and $S=\mathrm{supp}(\beta^*)$. Then each component of U satisfies $\mathbb{E}[\exp(tU_j)]\leq e^{-t^2/2}$ as explained above, the bound (40) applies. Hereafter, assume that event (41) holds. On event (41), $\sum_{j\in S}U_j\delta_j\leq(1+2\xi)\sqrt{2\log(p/s)}\|\delta\|_2$ by the Cauchy-Schwarz inequality. If $\hat{A}\subset\{1,...,p\}$ contains the indices of the s largest coefficients of U in absolute value, then $\sum_{j\in\hat{A}}U_j\delta_j\leq(1+2\xi)\sqrt{2\log(p/s)}\|\delta\|_2$ again by the Cauchy-Schwarz inequality. Finally, $\sum_{j\notin S\cup\hat{A}}U_j\delta_j\leq(1+2\xi)\sqrt{2\log(p/s)}\|\delta_{S^c}\|_1$ because $\max_{j\notin S\cup\hat{A}}|U_j|$ is bounded from above as in (41). Combining the above inequalities, on the event (41) we have

$$0\leq\sqrt{s}\|\delta\|_2(2+5\xi)\sqrt{2\log(p/s)}-\xi\sqrt{2\log(p/s)}\|\delta_{S^c}\|_1.$$

This implies that $\|\delta_{S^c}\|_1\leq\sqrt{s}\|\delta\|_2(2\xi^{-1}+5)$ and $\|\delta\|_1\leq\sqrt{s}\|\delta\|_2(6+2\xi^{-1})$.

The proof in logistic regression is the same up to a different scaling due to $\tilde{Z}$ from Lemma A.1(ii) being $L/2$-subgaussian, while in linear regression with the squared loss we had Z being $L\sigma^*$-subgaussian. □

B Group-Lasso

Lemma 3.5. *Consider the linear model* (10) *and assume that* $\max_{k=1,\dots,M}\|\Sigma_{G_k,G_k}\|_{op}\leq 1$ *and that each group has the same size* $|G_k|=d=p/M$. *Let* $\xi>0$ *and set* $\lambda=L\sigma^*(1+\xi)[\sqrt{d}+(1+2\xi)\sqrt{2\log(M/s)}]$ *where* $(\sigma^*)^2=(\sum_{i=1}^n\varepsilon_i^2)/n$ *and* s *is the number of groups with* $\beta^*_{G_k}\neq 0$. *Then*

$$\mathbb{P}\left(\{\hat{\beta}-\beta^*,\eta-\beta^*\}\subset T\right)\geq 1-2/\left(2\xi^2\log(M/s)(M/s)^{\xi}\right). \tag{30}$$

for $T=\{\delta\in\mathbb{R}^p:\sum_{k=1}^M\|\delta_{G_k}\|\leq\sqrt{s}\|\delta\|_2(2+3\xi^{-1})\}$. *If instead the logistic regression model and assumptions of Proposition 2.2 are fulfilled and* λ *is as above with* $\sigma^*=1/2$, *then* (30) *also holds.*

Proof of Lemma 3.5. Define $U=\Sigma^{1/2}Z/(L\sigma^*)$ where Z is as in Lemma A.1(i). Then $\mathbb{E}[\exp(v^TZ/(L\sigma^*))]\leq\exp(\|v\|^2/2)$ by the properties of Z stated in Lemma A.1(i). We wish to study the restriction U_{G_k} of U to group G_k. Let M_k be the matrix with $|G_k|$ rows and p columns

made of the rows of $\Sigma^{1/2}$ indexed in G_k. Then $U_{G_k} = M_k Z/(L\sigma^*)$ and by applying the concentration inequality in [18] to the subgaussian vector $Z/(L\sigma^*)$ and the matrix $M_k^T M_k \in \mathbb{R}^{p\times p}$,

$$\|U_{G_k}\|_2^2 \le \mathsf{trace}(M_k^T M_k) + 2\sqrt{x}\mathsf{trace}(M_k^T M_k M_k^T M_k)^{1/2} + 2x\|M_k^T M_k\|_{op}.$$

with probability at least $1-e^{-x}$. By properties of the trace, $\mathsf{trace}(M_k^T M_k) = \mathsf{trace}(M_k M_k^T) = \mathsf{trace}(\Sigma_{G_k,G_k})$. Similarly for the second term, $\mathsf{trace}(M_k^T M_k M_k^T M_k)^{1/2} = \|\Sigma_{G_k,G_k}\|_F$. Finally, $\|M_k^T M_k\|_{op} = \|M_k M_k^T\|_{op} = \|\Sigma_{G_k,G_k}\|_{op}$ so that the previous display reads

$$\|U_{G_k}\|_2^2 \le \mathsf{trace}(\Sigma_{G_k,G_k}) + 2\sqrt{x}\|\Sigma_{G_k,G_k}\|_F + 2x\|\Sigma_{G_k,G_k}\|_{op} \tag{42}$$

$$\le d + 2\sqrt{xd} + 2x \le (\sqrt{d}+\sqrt{2x})^2 \tag{43}$$

where for the second inequality we used that $\mathsf{trace}(\Sigma_{G_k,G_k}) \le |G_k| = d$ and $\|\Sigma_{G_k,G_k}\|_F^2 \le \sqrt{|G_k|} = \sqrt{d}$ using the assumption $\|\Sigma_{G_k,G_j}\|_{op} \le 1$. Hence $W_k = (\|U_{G_k}\|_2 - \sqrt{d})_+$ is 1-subgaussian for every group $k = 1, ..., M$, in the sense that $\mathbb{P}(W_k > \sqrt{2x}) \le e^{-x}$. As previously for the lasso, W_k is thus stochastically dominated by $\sqrt{2\tau}$ where τ is an exponential random variable with parameter 1, and

$$\mathbb{E}[(\|U_{G_k}\|_2 - \sqrt{d} - \mu)_+^2] \le \mathbb{E}[(\sqrt{2\tau}-\mu)_+^2] \le 2\int_0^\infty (t-\mu^2/2)_+^2 e^{-t}dt = 2e^{-\mu^2/2}. \tag{44}$$

Define $W \ge 0$ by $W^2 = \sum_{k=1}^M (\|U_{G_k}\|_2 - \sqrt{d} - \mu)_+^2$. We have thus proved that $\mathbb{E}[W^2] \le 2Me^{-\mu^2/2}$ and for $\mu = (1+\xi)\sqrt{2\log(M/s)}$ we obtain $\mathbb{E}[W^2] \le 2s/(M/s)^\xi$, and by Markov's inequality

$$\mathbb{P}\left(\tfrac{1}{s}W^2 \le \xi^2 2\log(M/s)\right) \ge 1 - 2/(\xi^2\log(M/s)(M/s)^\xi). \tag{45}$$

Furthermore, on the above event (45), by the triangle inequality we have

$$\max_{A\subset\{1,\dots,M\}:|A|=s}\left(\frac{1}{s}\sum_{k\in A}\|U_{G_k}\|_2^2\right)^{1/2} \le \sqrt{d}+\mu+\xi\sqrt{2\log(M/s)} = \sqrt{d}+(1+2\xi)\sqrt{2\log(M/s)} \triangleq \lambda_0.$$

Let λ_0 be defined as the right hand side of the previous display and notice that $\sqrt{n}\lambda/(L\sigma^*) = (1+\xi)\lambda_0$, so that if $\hat{A}$ is the subset of $[M]$ with the indices k with largest $\|U_{G_k}\|$ (i.e., a subset attaining the maximum in the previous display), we have proved that

$$\max_{k\in\hat{A}^c}\|(\Sigma^{1/2}Z)_{G_k}\| \le \Big(\frac{1}{s}\sum_{k\in\hat{A}}\|(\Sigma^{1/2}Z)_{G_k}\|_2^2\Big)^{1/2} \le (1+\xi)^{-1}\sqrt{n}\lambda. \tag{46}$$

By Lemma A.1(i), and the Cauchy-Schwarz inequality on the groups in S,

$$0 \le Z^T\Sigma^{1/2}\delta + \sqrt{n}\lambda\sum_{k=1}^M(\|\beta^*_{G_k}\| - \|b_{G_k}\|) \le Z^T\Sigma^{1/2}\delta + \sqrt{n}\lambda\left(\sqrt{s}\|\delta\| - \sum_{k\notin S}\|b_{G_k}\|\right)$$

$$= L\sigma^*\left[\delta^T U + (1+\xi)\lambda_0\left(\sqrt{s}\|\delta\| - \sum_{k\notin S}\|b_{G_k}\|\right)\right],$$

where $\delta = (b-\beta^*)$ and $U = \Sigma^{1/2}Z/(L\sigma^*)$. We now bound $U^T\delta$ which appears on the previous display. On the above event, (45) we have $\sum_{k\in S}{\delta_{G_k}}^T U_{G_k} \le (\sum_{k\in S}\|U_{G_k}\|^2)^{1/2}\|\delta\| \le \sqrt{s}(\sqrt{d} + (1+2\xi)\sqrt{2\log(M/s)})\|\delta\| = \sqrt{s}\lambda_0\|\delta\|$. Similarly, if $\hat{A}$ contains the indices of the s groups with the largest $\|U_{G_k}\|$ then on the above event (45), $\sum_{k\in\hat{A}}{\delta_{G_k}}^T U_{G_k} \le (\sum_{k\in\hat{A}}\|U_{G_k}\|^2)^{1/2}\|\delta\| \le \sqrt{s}(\sqrt{d}+(1+2\xi)\sqrt{2\log(M/s)})\|\delta\| = \sqrt{s}\lambda_0\|\delta\|$. For any group G_k with $k \notin S\cup\hat{A}$, we have $\|U_{G_k}\| \le \lambda_0$. Combining these bounds with the fact that $\lambda = L\sigma^*(1+\xi)\lambda_0$, we have established that on event (45),

$$0 \le (3+\xi)\sqrt{s}\lambda_0\|\delta\| - \xi\lambda_0\sum_{k\notin S}\|\delta_{G_k}\|, \qquad \sum_{k\notin S}\|\delta_{G_k}\| \le \sqrt{s}(1+3/\xi)\|\delta\|. \tag{47}$$

On the groups indexed in S, by the Cauchy-Schwarz inequality we have $\sum_{k\in S}\|\delta_{G_k}\| \le \sqrt{s}\|\delta\|$. Hence $\delta = b-\beta^*$ belongs to the cone defined in the statement of the Lemma. □

Lemma 3.6. *Assume that* $\max_{k=1,\dots,M} \|\Sigma_{G_k,G_k}\|_{op} \le 1$ *and that each group has the same size* $|G_k| = d = p/M$. *The set* T *defined in the previous lemma satisfies* $\gamma(T,\Sigma) \lesssim C(\xi)\phi(T)^{-1}\sqrt{sd + s\log(M/s)}$ *for some constant* $C(\xi)$ *that depends only on* ξ.

Proof of Lemma 3.6. By definition of the restricted eigenvalue $\phi(T)$, for any $u \in T$ with $\|\Sigma^{1/2}u\| = 1$ we have $\|u\| \le \phi(T)^{-1}$. Let $g \sim N(0, I_p)$; we wish to bound the expectation $\mathbb{E}\sup_{u\in T:\|\Sigma^{1/2}u\|=1} |g^T\Sigma^{1/2}u|$. Let $U = \Sigma^{1/2}g$. Let $\mu = \sqrt{2\log(M/s)}$. We have for any $u \in T$ with $\|\Sigma^{1/2}u\| = 1$,

$$u^T U \le \sum_{k=1}^M \|u_{G_k}\|\|U_{G_k}\| = \sum_{k=1}^M \|u_{G_k}\|(\|U_{G_k}\| - \mu - \sqrt{d}) + (\mu + \sqrt{d})\sum_{k=1}^M \|u_{G_k}\|.$$

For the second term, since $u \in T$, inequality $\sum_{k=1}^M \|u_{G_k}\| \le \sqrt{s}(2 + 3/\xi)\|u\|$ hence the second term is bounded from above as follows, $(\mu + \sqrt{d})\sum_{k=1}^M \|u_{G_k}\| \le (\mu + \sqrt{d})\sqrt{s}(2 + 3/\xi)\phi(T)^{-1}$. It remains to bound the first term. By the Cauchy-Schwarz inequality,

$$\sum_{k=1}^M \|u_{G_k}\|(\|U_{G_k}\| - \mu - \sqrt{d}) \le \|u\| \left(\sum_{k=1}^M (\|U_{G_k}\| - \mu - \sqrt{d})_+^2\right)^{1/2}.$$

Finally, $\|u\| \le \phi(T)^{-1}$ and the expectation bound (44) show that the previous display is bounded from above by $\phi(T)^{-1}(M2e^{-\mu^2/2})^{1/2} = \phi(T)^{-1}\sqrt{2s}$.

□

C Proofs of Section 4

Theorem 4.1. *[Exact Risk Identity] Consider the linear model* (10) *and the regularized problem* (1) *with an arbitrary proper convex function* $h(\cdot)$. *Assume that* $X_1, \dots, X_n$ *are iid* $N(0, I_p)$ *independent of* $\varepsilon_1, ..., \varepsilon_n$ *and set* $\sigma^* = (\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2)^{1/2}$. *Then with probability at least* $1 - 2\exp(-t^2/2)$,

$$\left| \|\hat\beta - \beta^*\| - \mathbb{E}_Z \left[\|\beta^* - \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)\|^2 \right]^{1/2} \right| \le \frac{\sigma^*(t+1)}{n^{1/2}} + \|\hat\beta - \eta\| \tag{31}$$

where $Z = \frac{1}{n^{1/2}\sigma^*}\sum_{i=1}^n \varepsilon_i X_i \sim N(0, I_p)$ *and* $\mathbb{E}_Z$ *denotes the expectation with respect* Z.

Proof of Theorem 4.1. Since $(X_i)_{i=1}^n$ and $(\varepsilon_i)_{i=1}^n$ are independent, we will condition throughout on $\varepsilon_1, \dots, \varepsilon_n$. The proximal operator is Lipschitz for any convex function h, in the sense that

$$\|\mathsf{prox}_h(\beta^* + z) - \mathsf{prox}_h(\beta^* + z')\| \le \|z - z'\| \quad \text{for all} \quad z, z' \in \mathbb{R}^p.$$

see Definition 2.3 and the following discussion in [27]. If $\Sigma = I_p$, the first order expansion η in (5) is given by $\eta = \mathsf{prox}_h\left(\beta^* + n^{-1}\sum_{i=1}^n X_i\varepsilon_i\right)$. This means $\eta = \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)$ for $Z = (\sum_{i=1}^n \varepsilon_i^2)^{-1/2}\sum_{i=1}^n \varepsilon_i X_i$ and $Z \sim N(0, I_p)$ if $X_1, ..., X_n$ are iid $N(0, I_p)$ independent of $\varepsilon_1, ..., \varepsilon_n$. Note that in this case, Z is independent of σ^*. Thus η is a $n^{-1/2}\sigma^*$-Lipschitz function of Z, and by the triangle inequality $\|\eta - \beta^*\|$ is also a $n^{-1/2}\sigma^*$-Lipschitz function of Z. By the Gaussian concentration inequality [12, Theorem 5.6], with probability $1 - 2\exp(-t^2/2)$ we have

$$|\|\eta - \beta^*\| - \mathbb{E}_Z[\|\eta - \beta^*\|]| \le n^{-1/2}\sigma^* t.$$

The Gaussian Poincaré inequality [12, Theorem 3.20] implies that $(\mathrm{Var}(\|\eta - \beta^*\|))^{1/2}$ is bounded by the Lipschitz constant and hence $|\mathbb{E}_Z[\|\eta - \beta^*\|] - (\mathbb{E}_Z[\|\eta - \beta^*\|^2])^{1/2}| \le n^{-1/2}\sigma^*$. Combining these inequalities above, we get with probability $1 - 2\exp(-t^2/2)$

$$\left| \|\eta - \beta^*\| - (\mathbb{E}[\|\beta^* - \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)\|^2])^{1/2} \right| \le n^{-1/2}\sigma^*(t+1).$$

Therefore, by triangle inequality, on the same event we have

$$\left| \|\hat\beta - \beta^*\| - (\mathbb{E}[\|\beta^* - \mathsf{prox}_h(\beta^* + n^{-1/2}\sigma^* Z)\|^2])^{1/2} \right| \le n^{-1/2}\sigma^*(t+1) + \|\eta - \hat\beta\|$$

which completes the proof. □

D Proofs of Section 5

Proposition 5.1. *Assume that $X_1, ..., X_n$ are iid $N(0, \Sigma)$ and is independent of $\varepsilon = (\varepsilon_1, ..., \varepsilon_n) \sim N(0, I_n)$. Assume that for some cone T and $r_n = \gamma(T, \Sigma)/\sqrt{n}$ we have*

$$\mathbb{P}(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| + \|\Sigma^{1/2}(\eta - \beta^*)\| \leq C_{19} r_n, \{\eta - \hat{\beta}, \eta - \beta^*, \hat{\beta} - \beta^*\} \subset T) \geq 1 - \alpha. \quad (33)$$

Then for some T_n with the t-distribution with n degrees-of-freedom, with probability $1 - \alpha - 4e^{-nr_n^2/2}$,

$$\sqrt{n}(\hat{\theta} - a^T\beta^*) - T_n = O_p((1 + r_n))\|\Sigma^{1/2}(\eta - \beta^*)\| + O_p(\sqrt{n}r_n)\|\Sigma^{1/2}(\eta - \hat{\beta})\|, \quad (34)$$

$$= O_p(r_n(1 + r_n)) + O_p(\sqrt{n}r_n^3). \quad (35)$$

Proof of Proposition 5.1. Some of the argument below is borrowed from [4, Section 6]. Let $T_n = \sqrt{n}\|z_a\|^{-2} z_a^T \varepsilon$, and let $Q_a = I_p - \Sigma^{-1} a a^T$; notice that T_n has the t-distribution with n degrees-of-freedom. Also note that $\Sigma^{1/2} Q_a \Sigma^{-1/2} = I - (\Sigma^{-1/2}a)(\Sigma^{-1/2}a)^\top$ and $z_a = X\Sigma^{-1}a = X\Sigma^{-1/2}\Sigma^{-1/2}a$; this implies that $\mathsf{X}Q_a$ is independent of z_a because $(z_a, \mathsf{X}Q_a)$ are jointly normal and uncorrelated. By definition of $\hat{\theta}$, simple algebra yields that

$$\sqrt{n}(\hat{\theta} - a^T\beta^*) - T_n = -\sqrt{n}\|z_a\|^{-2} z_a^T \mathsf{X} Q_a (\hat{\beta} - \beta^*) \quad (48)$$

$$= -\sqrt{n}\|z_a\|^{-2} z_a^T \mathsf{X} Q_a (\eta - \beta^*) + \sqrt{n}\|z_a\|^{-2} z_a^T \mathsf{X} Q_a (\eta - \hat{\beta}). \quad (49)$$

Note that η only depends on X through $\varepsilon^T\mathsf{X}$ hence η is independent of $(P_\varepsilon^\perp \mathsf{X}, P_\varepsilon^\perp z_a)$ where $P_\varepsilon^\perp = I_n - \|\varepsilon\|^{-2}\varepsilon\varepsilon^T$ is an orthogonal projection; set also $P_\varepsilon = \|\varepsilon\|^{-2}\varepsilon\varepsilon^T$ for the complementary projection. We further split the first term above, so that $\sqrt{n}(\hat{\theta} - a^T\beta^*) - T_n$ is equal to

$$\sqrt{n}\|z_a\|^{-2}\left(-\left[z_a^T P_\varepsilon^\perp \mathsf{X} Q_a(\eta - \beta^*)\right] - \left[z_a^T P_\varepsilon \mathsf{X} Q_a(\eta - \beta^*)\right] + \left[z_a^T \mathsf{X} Q_a(\eta - \hat{\beta})\right]\right)$$

Hereafter, assume that the event $\{\eta - \beta^*, \eta - \hat{\beta}\} \subset T$ holds (this holds with probability at least $1 - \alpha$ by assumption). For the first bracket inside the large parenthesis, $P_\varepsilon^\perp z_a$ is independent of $(\eta, \mathsf{X}Q_a)$ conditionally on ε, so that $z_a^T P_\varepsilon^\perp \mathsf{X} Q_a(\eta - \beta^*) = O_p(1)\|\mathsf{X}Q_a(\eta - \beta^*)\|$. For the second bracket, since P_ε is rank 1, $\|P_\varepsilon z_a\| = O_p(1)$ and the second bracket is also $O_p(1)\|\mathsf{X}Q_a(\eta - \beta^*)\|$ by the Cauchy-Schwarz inequality. Since $\eta - \beta^* \in T$ and $r_n = \gamma(T, \Sigma)/\sqrt{n}$ we have $\mathsf{X} = \mathsf{X}Q_a + z_a a^T$ so that by the triangle inequality,

$$\|\mathsf{X}Q_a(\eta - \beta^*)\| \leq \|\mathsf{X}(\eta - \beta^*)\| + \|z_a\||a^T(\eta - \beta^*)| \leq \|\mathsf{X}(\eta - \beta^*)\| + \|z_a\|\|\Sigma^{1/2}(\eta - \beta^*)\|.$$

By an application of [28], $\sup_{u \in T: \|\Sigma^{1/2}u\| = 1} |\|\mathsf{X}u\| - \sqrt{n}\|\Sigma^{1/2}u\|| \leq C_{20}(\gamma(T, \Sigma) + t)$ with probability at least $1 - 2e^{-t^2/2}$, since $\mathsf{X}\Sigma^{-1/2}$ has iid $N(0, 1)$ entries. We take $t = \gamma(T, \Sigma) = r_n\sqrt{n}$. Since $\|z_a\| = O_p(\sqrt{n})$, the first two brackets above are $O_p(\sqrt{n}(1 + r_n))\|\Sigma^{1/2}(\eta - \beta^*)\|$.

We now focus on the third bracket. Since $\eta - \hat{\beta} \in T$,

$$z_a^T \mathsf{X} Q_a (\eta - \hat{\beta}) \leq \|\Sigma^{1/2}(\eta - \hat{\beta})\|\|z_a\| \sup_{u \in T: \|\Sigma^{1/2}u\| = 1} \|z_a\|^{-1} z_a^T \mathsf{X} Q_a \Sigma^{-1/2}\Sigma^{1/2} u. \quad (50)$$

Since z_a and $\mathsf{X}Q_a$ are independent, conditionally on z_a, the random vector $\|z_a\|^{-1} z_a^T \mathsf{X} Q_a \Sigma^{-1/2}$ is normal with covariance matrix $\Sigma^{1/2} Q_a \Sigma^{-1/2}$. Since $\Sigma^{1/2} Q_a \Sigma^{-1/2}$ is a projection matrix in $\mathbb{R}^p$, $\Sigma^{1/2} Q_a \Sigma^{-1/2} \preceq I_p$ in the sense of positive semi-definite matrices. By the Sudakov-Fernique's inequality [43, Theorem 7.2.11], conditionally on z_a we have

$$\mathbb{E}\left[\sup_{u \in T: \|\Sigma^{1/2}u\| = 1} \|z_a\|^{-1} z_a^T \mathsf{X} Q_a \Sigma^{-1/2}\Sigma^{1/2} u \;\middle|\; z_a\right] \leq \mathbb{E}\left[\sup_{u \in T: \|\Sigma^{1/2}u\| = 1} g^T \Sigma^{1/2} u\right] = \gamma(T, \Sigma)$$

for some $g \sim N(0, I_p)$. Furthermore, by Gaussian concentration [12, Theorem 5.8] again conditionally on z_a we have $\sup_{u \in T: \|\Sigma^{1/2}u\| = 1} \|z_a\|^{-1} z_a^T \mathsf{X} Q_a \Sigma^{-1/2}\Sigma^{1/2} u \leq \gamma(T, \Sigma) + t$ with probability at least $1 - e^{-t^2/2}$. Taking $t = \gamma(T, \Sigma) = r_n\sqrt{n}$, we obtain that the right hand side of (50) is bounded from above on an event of probability at least $1 - e^{-t^2/2}$ by $\|\Sigma^{1/2}(\eta - \hat{\beta})\|\|z_a\|\sqrt{n}r_n$. Given that $\|z_a\|/\sqrt{n} = O_p(1)$ and $\sqrt{n}/\|z_a\| = O_P(1)$, the proof is complete. □

E Proofs of Results in 6

Theorem 6.1. *Define* $\hat{K} := n^{-1}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)X_iX_i^\top$. *Under assumption* ***(A1)****, we have*

(i) If $\{\hat{\beta}-\beta^*, \eta-\beta^*\} \subseteq T$ *then* $\|\hat{\beta}-\eta\|_K \lesssim Q_{n,1}^{1/2}\mathcal{E} + B^{1/2}Z_n^{1/2}\mathcal{E}^{3/2}$.

(ii) If $\{\hat{\beta}-\eta, \hat{\beta}-\beta^*, \eta-\beta^*\} \subseteq T$ *then* $\|\hat{\beta}-\eta\|_K \lesssim Q_{n,2}\mathcal{E} + BZ_n\mathcal{E}^2$,

where

$$Q_{n,1} = \sup_{u\in T}\left|\frac{u^\top\hat{K}u}{\|u\|_K^2} - 1\right|,\ Q_{n,2} = \sup_{u,v\in T}\frac{|u^\top(\hat{K}-K)v|}{\|u\|_K\|v\|_K}\ \text{ and }\ Z_n = \sup_{u\in T}\frac{1}{n}\sum_{i=1}^n\frac{|X_i^\top u|^3}{\|u\|_K^3}.$$

Proof of Theorem 6.1. By strong convexity of the objective function of η with respect to the norm $\|\cdot\|_K$, or alternatively by application of for instance [6, Lemma 1] or or [7, Lemma A.2],

$$\begin{aligned}\frac{1}{2}\|K^{1/2}(\hat{\beta}-\eta)\|^2 \le \frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\hat{\beta}-\beta^*) + \frac{1}{2}\|K^{1/2}(\hat{\beta}-\beta^*)\|^2 + h(\hat{\beta}) \\ - \left[\frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\eta-\beta^*) + \frac{1}{2}\|K^{1/2}(\eta-\beta^*)\|^2 + h(\eta)\right].\end{aligned} \tag{51}$$

From the definition (1) of $\hat{\beta}$, we get

$$0 \le \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\eta) + h(\eta) - \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\hat{\beta}) - h(\hat{\beta}). \tag{52}$$

Adding inequalities (51) and (52), the terms $h(\hat{\beta})$ and $h(\eta)$ cancel out and we obtain

$$\begin{aligned}&\frac{1}{2}\|K^{1/2}(\hat{\beta}-\eta)\|^2 \\ &\quad\le \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\eta) - \frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\eta-\beta^*) - \frac{1}{2}\|K^{1/2}(\eta-\beta^*)\|^2 \\ &\quad\quad - \left[\frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\hat{\beta}) - \frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\hat{\beta}-\beta^*) - \frac{1}{2}\|K^{1/2}(\hat{\beta}-\beta^*)\|^2\right].\end{aligned} \tag{53}$$

The terms on the right hand side resemble the remainder terms from a second order Taylor series expansion except for K instead of $\hat{K}$. Using the Taylor expansion above, we get for any $\beta\in\mathbb{R}^p$

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) - \frac{1}{n}\sum_{i=1}^n \ell'(Y_i, X_i^\top\beta^*)X_i^\top(\beta-\beta^*) - \frac{1}{2}\|K^{1/2}(\beta-\beta^*)\|^2 \tag{54}$$

$$= \frac{1}{2n}\sum_{i=1}^n |X_i^\top(\beta-\beta^*)|^2 a_i(\beta) + \frac{1}{2}(\beta-\beta^*)^\top(\hat{K}-K)(\beta-\beta^*), \tag{55}$$

where

$$a_i(\beta) := \int_0^1 \{\ell''(Y_i, X_i^\top\beta^* + tX_i^\top(\beta-\beta^*)) - \ell''(Y_i, X_i^\top\beta^*)\}dt. \tag{56}$$

(Note that for the squared loss, $\ell''=1$ is constant and $a_i(\cdot)=0$ which leads to a mucher simple analysis; the reader only interested in squared loss may skip the analysis of $a_i(\cdot)$). Substituting this in (53), we get

$$\|K^{1/2}(\hat{\beta}-\eta)\|^2 \le (\eta-\beta^*)^\top(\hat{K}-K)(\eta-\beta^*) - (\hat{\beta}-\beta^*)^\top(\hat{K}-K)(\hat{\beta}-\beta^*) \tag{57}$$

$$+ \frac{1}{n}\sum_{i=1}^n |X_i^\top(\eta-\beta^*)|^2 a_i(\eta) - \frac{1}{n}\sum_{i=1}^n |X_i^\top(\hat{\beta}-\beta^*)|^2 a_i(\hat{\beta}). \tag{58}$$

From the Lipschitz condition on $\ell''(\cdot,\cdot)$, we get $|a_i(\beta)| \le B|X_i^\top(\beta-\beta^*)|$, and hence part 1 of the result follows.

In part 1, we did not use any information about $\hat\beta - \eta$. For part 2, we will control the right hand side "quadratic forms" in (57) in a more refined way. By simple algebra and the definition of $Q_{n,2}(\cdot)$,

$$\begin{aligned}
&(\eta-\beta^*)^\top(\hat K - K)(\eta-\beta^*) - (\hat\beta-\beta^*)^\top(\hat K-K)(\hat\beta-\beta^*)\\
&\quad= ((\eta-\beta^*)-(\hat\beta-\beta^*))^\top(\hat K-K)((\eta-\beta^*)+(\hat\beta-\beta^*))\\
&\quad= (\eta-\hat\beta)^\top(\hat K-K)(\eta-\beta^*)\\
&\qquad\quad + (\eta-\hat\beta)^\top(\hat K-K)(\hat\beta-\beta^*)\\
&\quad\le Q_{n,2}(T)\|K^{1/2}(\eta-\hat\beta)\|\left[\|K^{1/2}(\hat\beta-\beta^*)\| + \|K^{1/2}(\eta-\beta^*)\|\right].
\end{aligned} \tag{59}$$

This completes the control of first difference in (57). For the second difference in (58), observe that

$$|a_i(\beta)| \le B|X_i^\top(\beta-\beta^*)| \quad\text{and}\quad |a_i(\beta)-a_i(\alpha)| \le B|X_i^\top(\beta-\alpha)|,$$

and hence

$$\begin{aligned}
&\frac1n\sum_{i=1}^n |X_i^\top(\eta-\beta^*)|^2 a_i(\eta) - \frac1n\sum_{i=1}^n |X_i^\top(\hat\beta-\beta^*)|^2 a_i(\hat\beta)\\
&\quad= \frac1n\sum_{i=1}^n a_i(\eta)\left[|X_i^\top(\eta-\beta^*)|^2 - |X_i^\top(\hat\beta-\beta^*)|^2\right] + \frac1n\sum_{i=1}^n |X_i^\top(\hat\beta-\beta^*)|^2\left[a_i(\eta)-a_i(\hat\beta)\right]\\
&\quad\le \frac{B}{n}\sum_{i=1}^n |X_i^\top(\eta-\beta^*)|\times|X_i^\top((\eta-\beta^*)-(\hat\beta-\beta^*))|\times|X_i^\top((\eta-\beta^*)+(\hat\beta-\beta^*))|\\
&\qquad+\frac{B}{n}\sum_{i=1}^n |X_i^\top(\hat\beta-\beta^*)|^2\times|X_i^\top(\eta-\hat\beta)|\\
&\quad\le \frac{B}{n}\sum_{i=1}^n |X_i^\top(\eta-\hat\beta)|\times\left[|X_i^\top(\eta-\beta^*)|^2+|X_i^\top(\hat\beta-\beta^*)|^2\right]\\
&\qquad+\frac{B}{n}\sum_{i=1}^n |X_i^\top(\eta-\hat\beta)|\times|X_i^\top(\eta-\beta^*)|\times|X_i^\top(\hat\beta-\beta^*)|.
\end{aligned}$$

The right hand side is trivially bounded by

$$3B\|\eta-\hat\beta\|_K\left[\|(\eta-\beta^*)\|_K^2+\|(\hat\beta-\beta^*)\|_K^2\right]\sup_{u,v,w\in\tilde T}\frac1n\sum_{i=1}^n\frac{|X_i^\top u|}{\|K^{1/2}u\|}\times\frac{|X_i^\top v|}{\|K^{1/2}v\|}\times\frac{|X_i^\top w|}{\|K^{1/2}w\|}.$$

Using $3abc \le a^3+b^3+c^3$ for any positive $\{a,b,c\}$, the previous display is bounded from above by

$$B\|K^{1/2}(\eta-\hat\beta)\|\left[\|K^{1/2}(\eta-\beta^*)\|^2+\|K^{1/2}(\hat\beta-\beta^*)\|^2\right]Z_n(T).$$

Substituting these bounds in (58), we get the result. □

Proposition 6.2. *[Control of $Q_{n,1}, Q_{n,2}$ and Z_n] Under assumptions **(A1)** and **(A2)**, we have*

(i) With probability $1-2\exp(-C_{21}t^2\gamma^2(T,\Sigma))$,

$$\max\{Q_{n,1}(T),Q_{n,2}(T)\} \le C_{22}B_2B_3L^2\left(tn^{-1/2}\gamma(T,\Sigma)+t^2n^{-1}\gamma^2(T,\Sigma)\right).$$

(ii) With probability $1-2\exp(-C_{23}t\log n)$, $\quad Z_n(T) \le C_{24}B_3^{3/2}L^3\left(1+n^{-1}\gamma^3(T,\Sigma)\right)t^3$.

Proof of Proposition 6.2. Define the function classes F and H as

$$\begin{aligned}
F &:= \left\{(x,y)\mapsto x^\top u/\|u\|_K : u\in T\right\},\\
H &:= \left\{(x,y)\mapsto \ell''(y,x^\top\beta^*)x^\top u/\|u\|_K : u\in T\right\}.
\end{aligned}$$

It is then clear that

$$\max\{Q_{n,1}(T), Q_{n,2}(T)\} \leq \sup_{f\in F, h\in H} \left| \frac{1}{n}\sum_{i=1}^{n} \{f(X_i,Y_i)h(X_i,Y_i) - \mathbb{E}[f(X_i)h(X_i)]\} \right|,$$

$$Z_n(T) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|^3.$$

We now apply Theorem 1.13 of [33] with $2^{s_0/2} = \gamma(T,\Sigma), q = 5$. Hence, we get for any $t \geq 8$ with probability at least $1 - 2\exp(-c_1 t^2 \gamma^2(T,\Sigma))$,

$$\begin{aligned} &\sup_{f\in F, h\in H} \left| \frac{1}{n}\sum_{i=1}^{n} \{f(X_i,Y_i)h(X_i,Y_i) - \mathbb{E}[f(X_i,Y_i)h(X_i,Y_i)]\} \right| \qquad (60)\\ &\quad \leq \frac{c_2 t^2}{n}(\gamma(F) + 2^{s_0/2}\text{diam}(F))(\gamma(H) + 2^{s_0/2}\text{diam}(H))\\ &\qquad + \frac{c_2 t}{\sqrt{n}}(\text{diam}(F)(\gamma(H) + 2^{s_0/2}\text{diam}(H)) + \text{diam}(H)(\gamma(F) + 2^{s_0/2}\text{diam}(F))), \end{aligned}$$

where

$$\gamma(F) := L\mathbb{E}\left[\sup_{u\in T} \frac{|g^\top \Sigma^{1/2} u|}{\|K^{1/2}u\|}\right] \leq L \sup_{u\in T} \frac{\|\Sigma^{1/2}u\|}{\|K^{1/2}u\|}\gamma(T,\Sigma) \leq B_3^{1/2} L\gamma(T,\Sigma),$$

$$\gamma(H) := B_2 L\mathbb{E}\left[\sup_{u\in T} \frac{|g^\top \Sigma^{1/2} u|}{\|K^{1/2}u\|}\right] \leq B_2 B_3^{1/2} L\gamma(T,\Sigma),$$

and

$$\text{diam}(F) := L \sup_{u\in T} \frac{\|\Sigma^{1/2}u\|}{\|K^{1/2}u\|} \leq B_3^{1/2} L,$$

$$\text{diam}(H) := B_2 L \sup_{u\in T} \frac{\|\Sigma^{1/2}u\|}{\|K^{1/2}u\|} \leq B_2 B_3^{1/2} L.$$

Substituting these quantities in (60), part 1 of the result follows. Alternatively, one could apply Theorem 1.12 of [33] if $\ell(\cdot,\cdot)$ is assumed to be convex in the second argument (implying $\ell''(y, x^\top\beta^*) \geq 0$).

To prove part 2, we apply Equation (3.9) of [32] with $|I| = n$. Hence, we have with probability $1 - 2\exp(-c_1 t \log n)$

$$Z_n(T) = \sup_{f\in F} \frac{1}{n}\sum_{i=1}^{n} |f(X_i)^3| \leq \frac{ct^3}{n}\left(\gamma(F) + n^{1/3}\text{diam}(F)\right)^3, \qquad (61)$$

where $c > 0$ is an absolute constant and

$$\gamma(F) := L\mathbb{E}\left[\sup_{u\in T} \frac{|g^\top \Sigma^{1/2} u|}{\|K^{1/2}u\|}\right] \leq B_3^{1/2} L\gamma(T,\Sigma),$$

and

$$\text{diam}(F) := L \sup_{u\in T} \frac{\|\Sigma^{1/2}u\|}{\|K^{1/2}u\|} \leq B_3^{1/2} L.$$

Hence part 2 follows. □

E.1 Verification of Assumption (A1) for Logistic Loss

Proposition 2.2. *Consider the logistic loss $\ell(y,u) = yu - \log(1+e^u)$ for $y \in \{0,1\}, u \in \mathbb{R}$. Assume that $(X_i, Y_i)_{i=1,\ldots,n}$ are iid satisfying the logistic regression model*

$$\mathbb{P}(Y_i = 1|X_i) = 1 - \mathbb{P}(Y_i = 0|X_i) = 1/(1+\exp({X_i}^\top\beta^*)),$$

*for some $\beta^* \in \mathbb{R}^p$ with $\|\Sigma^{1/2}\beta^*\| \leq 1$.[2] Assume **(A2)** holds. Then* (8) *holds with $B = 1/(6\sqrt{3})$, $B_2 = 1$ and an absolute constant $B_3 > 0$.*

[2]The constant 1 can be replaced by another absolute constant; this will only change B_3 to a different constant.

Lipschitzness and boundedness of $\ell''(y,u)$ for logistic loss is straightforward. These parts do not require $\|\Sigma^{1/2}\beta^*\| \leq 1$. In order to prove the third part, we prove the following general result for general loss function with a lower second-order curvature.

Define for any $t > 0$

$$\alpha(t) := \inf_{y\in\mathbb{R}} \inf_{|u|\leq t} \ell''(y,u).$$

Note that $\alpha(\cdot)$ is non-increasing. This is called the lower curvature function and appears in the works Loh [30, Section 3.2] and Koltchinskii [24, Section 3].

Proposition E.1. *Suppose $X_1, \ldots, X_n$ are iid L-subGaussian random vectors with covariance Σ. Then there exists absolute constants $c_1, c_2 > 0$ such that for any $u \in \mathbb{R}^p$, we have*

$$\frac{u^\top K_n u}{u^\top \Sigma u} \geq \sup_{\tau>0} \alpha(\tau)\left\{1 - c_1 L(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2}\right\} \tag{62}$$

$$\geq \alpha(2L\sqrt{\log(c_2 L)}\|\Sigma^{1/2}\beta^*\|_2)/2. \tag{63}$$

Proof. Fix a number $\tau > 0$. It is clear that

$$u^\top K u = \mathbb{E}\left[\ell''(Y_1, X_1^\top\beta^*)(X_1^\top u)^2\right] \geq \alpha(\tau)\mathbb{E}\left[(X_1^\top u)^2 \mathbb{1}\{|X_1^\top\beta^*| \leq \tau\}\right]. \tag{64}$$

Observe now that for any $u \in \mathbb{R}^p$, we have

$$\begin{aligned}
0 &\leq \mathbb{E}[(X_1^\top u)^2] - \mathbb{E}[(X_1^\top u)^2 \mathbb{1}\{|X_1^\top\beta^*| \leq \tau\}] \\
&= \mathbb{E}[(X_1^\top u)^2 \mathbb{1}\{|X_1^\top\beta^*| > \tau\}] \\
&\leq (\mathbb{E}[(X_1^\top u)^4])^{1/2}(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2} \\
&\leq cL\|\Sigma^{1/2}u\|^2(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2},
\end{aligned}$$

for some absolute constant $c > 0$. This implies that

$$\mathbb{E}\left[(X_1^\top u)^2 \mathbb{1}\{|X_1^\top\beta^*| \leq \tau\}\right] \geq \|\Sigma^{1/2}u\|_2^2\left\{1 - cL(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2}\right\}.$$

Combining this with (64), we get

$$\frac{u^\top K u}{u^\top \Sigma u} \geq \sup_{\tau>0} \alpha(\tau)\left\{1 - cL(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2}\right\}.$$

This proves the first inequality. To prove the second inequality, take $\tau = \rho\|\Sigma^{1/2}\beta^*\|_2$ for some $\rho > 0$ (to be determined later). For this choice, we have from L-subGaussianity

$$\mathbb{P}\left(|X_1^\top\beta^*| \geq \tau\right) = \mathbb{P}(|X_1^\top\beta^*| \geq \rho\|\Sigma^{1/2}\beta^*\|_2) \leq 2\exp\left(-\frac{\rho^2}{2L^2}\right).$$

Hence, if $\rho = 2L(\log(\sqrt{8}cL))^{1/2}$ then

$$1 - cL(\mathbb{P}(|X_1^\top\beta^*| > \tau))^{1/2} \geq 1 - \sqrt{2}cL\exp\left(-\frac{\rho^2}{4L^2}\right) \tag{65}$$

$$= 1 - \exp\left(-\frac{\rho^2}{4L^2} + \log(\sqrt{2}cL)\right) \geq \frac{1}{2}. \tag{66}$$

Therefore, for any $u \in \mathbb{R}^p$,

$$\frac{u^\top K u}{u^\top \Sigma u} \geq \frac{1}{2}\alpha(2L(\log(\sqrt{8}cL))^{1/2}\|\Sigma^{1/2}\beta^*\|_2).$$

This completes the proof. □

F Verification of Restricted Strong Convexity and Rates for Logistic Lasso

In the main paper, we proved/stated bounds for $\|\hat{\beta} - \beta^*\|_K$ and $\|\eta - \beta^*\|_K$ for squared loss with different penalties. These proofs can be extended to the case of logistic loss once restricted strong convexity (RSC) condition is verified; see Proposition 3.4 which is restated and proved below. Also, see [35] where the RSC was introduced. We present the following result that proves RSC for general cones T.

Proposition F.1. *Fix any cone $T \subset \mathbb{R}^p$. Assume **(A1)** and **(A2)** holds. If the loss function satisfies*

$$\sup_{|s-t|\le u} \frac{\ell''(y,s)}{\ell''(y,t)} \le \exp(3u), \tag{67}$$

then for any $u \in T$ satisfying $\|u\|_K \le 1$, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta^* + X_i^\top u) - \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta^*) - \frac{1}{n}\sum_{i=1}^n u^\top X_i \ell'(Y_i, X_i^\top\beta^*) \\ \ge \|u\|_K^2 \frac{B_2^{-1}[0.5B_3^{-1/2} - \tilde{Q}_n(T)]}{\exp(6\sqrt{2}L(\log(4B_3^{1/2}B_2L^2))^{1/2})}, \end{aligned} \tag{68}$$

for a random quantity $\tilde{Q}_n(T)$ which satisfies the following: there exists a universal constant $C > 0$ such that for any $t \ge 0$, we have with probability $1 - \exp(-t)$,

$$Q_n(T) \le C\left(\frac{B_2L\gamma(T,\Sigma)}{\sqrt{n}} + \frac{B_2\gamma^2(T,\Sigma)}{n} + \frac{t^{1/2}B_2L^2}{\sqrt{n}} + \frac{tB_2L^2}{n}\right).$$

Assumption (67) can be verified for logistic regression easily. Therefore, if $\gamma(T,\Sigma)/\sqrt{n} \to 0$ then for all $u \in T$, $\|u\|_K \le 1$, we get

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta^* + X_i^\top u) - \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta^*) - \frac{1}{n}\sum_{i=1}^n u^\top X_i \ell'(Y_i, X_i^\top\beta^*) \ge \mathfrak{C}\|u\|_\Sigma^2,$$

for a constant $\mathfrak{C}$ depending on L, B, B_2, B_3.

Proof. For notational convenience, let

$$f_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta).$$

Define for β^* and any $u \in T$,

$$\Delta_2 f_n(\beta^*, u) := f_n(\beta^* + u) - f_n(\beta^*) - u^\top \nabla f_n(\beta^*).$$

Note that

$$\Delta_2 f_n(\beta^*, u) = \frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top(\beta^* + su))\langle X_i, u\rangle^2,$$

for some $s \in [0,1]$. From the stability property (67) of $\ell''(\cdot,\cdot)$, we get that

$$\ell''(Y_i, X_i^\top(\beta^* + su))\langle X_i, u\rangle^2 \ge \exp(-3|\langle u, X_i\rangle|)\ell''(Y_i, X_i^\top\beta^*)\langle X_i, u\rangle^2.$$

This implies that

$$\begin{aligned} \Delta_2 f_n(\beta^*, u) &= \frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\exp(-3|\langle u, X_i\rangle|)\langle X_i, u\rangle^2 \\ &= \|u\|_\Sigma^2 \frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\exp(-3|\langle u, X_i\rangle|)\left(\frac{\langle u, X_i\rangle}{\|u\|_\Sigma}\right)^2 \end{aligned} \tag{69}$$

Now define the function

$$\varphi_\tau(t) = \begin{cases} |t|, & \text{if } |t| \le \tau/2, \\ \tau - |t|, & \text{if } \tau/2 \le |t| \le \tau, \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that

$$\frac{\langle X_i, u\rangle^2}{\|u\|_\Sigma^2} \;\ge\; \varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right).$$

Further $\varphi_\tau(\cdot)$ is a 1-Lipschitz function. Using these properties, we get

$$\begin{aligned} \Delta_2 f_n(\beta^*, u) &\ge \|u\|_\Sigma^2 \frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\exp(-3|\langle u, X_i\rangle|)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right) \\ &\ge \|u\|_\Sigma^2 \exp(-3\tau\|u\|_\Sigma)\frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right). \end{aligned} \tag{70}$$

To complete the proof note that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right) \\ &\quad= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)\right] \\ &\qquad+ \frac{1}{n}\sum_{i=1}^n \left\{\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right) - \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)\right]\right\}. \end{aligned} \tag{71}$$

We now control the first term by noting that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\left\{\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)^2 - \varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)\right\}\right] \\ &\quad\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)^2 \mathbb{1}\left\{|\langle u, X_i\rangle| \ge \tau\|u\|_\Sigma/2\right\}\right] \\ &\quad\le B_2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\langle u, X_i\rangle^2}{\|u\|_\Sigma^2}\mathbb{1}\{|\langle u, X_i\rangle| \ge \tau\|u\|_\Sigma/2\}\right] \\ &\quad\le B_2\int_{\tau/2}^\infty 2t\exp\left(-\frac{t^2}{2L^2}\right)dt = 2B_2L^2\exp\left(-\frac{\tau^2}{8L^2}\right), \end{aligned} \tag{72}$$

using the boundedness of ℓ'' and L-sub-Gaussianity of X_i. This implies that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)\right] \ge \frac{\|u\|_{K_n}}{\|u\|_\Sigma} - 2B_2L^2\exp\left(-\frac{\tau^2}{8L^2}\right)dt \ge \frac{1}{2B_3^{1/2}}, \tag{73}$$

for $\tau := 2\sqrt{2}L(\log(4B_3^{1/2}B_2L^2))^{1/2}$.

Combining this with the lower bound on $\Delta_2 f_n(\beta^*, u)$, we get for all $u \in T$,

$$\Delta_2 f_n(\beta^*, u) \ge \|u\|_\Sigma^2\exp(-3\tau\|u\|_\Sigma)\left[0.5B_3^{-1/2} - \tilde{Q}_n(T)\right],$$

where

$$\tilde{Q}_n(T) := \max_{u\in T}\left|\frac{1}{n}\sum_{i=1}^n \left\{\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right) - \mathbb{E}\left[\ell''(Y_i, X_i^\top\beta^*)\varphi_\tau^2\left(\frac{\langle X_i, u\rangle}{\|u\|_\Sigma}\right)\right]\right\}\right|.$$

Before bounding $\tilde{Q}_n(T)$, note from assumption **(A1)** that

$$\|u\|_K^2 \le B_2\|u\|_\Sigma^2 \le B_2B_3\|u\|_K,$$

and hence for all $u \in T$

$$\Delta_2 f_n(\beta^*, u) \geq B_2^{-1}\|u\|_K^2 \exp(-3B_3^{1/2}\tau\|u\|_K)\left[0.5B_3^{-1/2} - \tilde{Q}_n(T)\right], \tag{74}$$

We now bound $\tilde{Q}_n(T)$. Define the function class

$$F := \{(x,y) \mapsto (\ell''(y, x^\top\beta^*))^{1/2}\varphi_\tau(\langle u, x\rangle/\|u\|_\Sigma) : u \in T\}.$$

From this definition, it follows that

$$\tilde{Q}_n(T) = \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n \left\{f^2(X_i,Y_i) - \mathbb{E}[f^2(X_i,Y_i)]\right\}\right|.$$

We now apply Theorem 5.5 of [16] to get with probability $1-\exp(-t)$,

$$\tilde{Q}_n(T) \leq C\left(\frac{\gamma(F)\text{diam}(F)}{\sqrt{n}} + \frac{\gamma^2(F)}{n} + \frac{t^{1/2}\text{diam}^2(F)}{\sqrt{n}} + \frac{t\text{diam}^2(F)}{n}\right),$$

for some constant $C > 0$ where

$$\begin{aligned}\text{diam}(F) &:= \sup_{f\in F}\|f(X,Y)\|_{\psi_2} \leq \sup_{u\in T}\frac{\|(\ell''(Y,X^\top\beta^*))^{1/2}\varphi_\tau(\langle u,X\rangle/\|u\|_\Sigma)\|_{\psi_2}}{\|u\|_\Sigma} \\ &\leq B_2^{1/2}\sup_{u\in T}\frac{\|\langle u,X\rangle\|_{\psi_2}}{\|u\|_\Sigma} \leq B_2^{1/2}L,\end{aligned}$$

and

$$\gamma(F) := B_2^{1/2}\mathbb{E}\left[\max_{u\in T}\frac{|\langle\Sigma^{1/2}u, g\rangle|}{\|u\|_\Sigma}\right] = B_2^{1/2}\gamma(T,\Sigma).$$

Substituting these quantities in (74) for $\|u\|_K \leq 1$ implies the result. □

The following result proves a rate result for linear and logistic regression with $h(\beta) = \lambda\|\beta\|_1$ (based on the restricted strong convexity result above). Define for the loss $\ell(\cdot,\cdot)$,

$$f_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) \quad \text{and} \quad \hat{\beta} := \operatorname*{argmin}_{\beta\in\mathbb{R}^p} f_n(\beta) + \lambda\|\beta\|_1$$

Recall from Lemma 3.3 that if $\lambda = L\sigma^*(1+3\xi)\sqrt{2\log(p/s)/n}$ in case of squared loss and $\lambda = (L/2)(1+3\xi)\sqrt{2\log(p/s)/n}$ in case of logistic loss, for some $\xi > 0$ then on the event (41),

$$\hat{\beta} - \beta^* \in T_{\texttt{lasso}}(s(6+2\xi^{-1})^2) := \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leq \sqrt{s}\|\delta\|(6+2\xi^{-1})\}.$$

This holds for both the linear and logistic lasso case.

Proposition 3.4. *Consider the penalized lasso estimator $\hat{\beta}$ given by*

$$\hat{\beta} := \text{argmin}_{\beta\in\mathbb{R}^p}\ \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) + \lambda\|\beta\|_1,$$

*where ℓ is either the squared or logistic loss and λ is chosen as in Lemma 3.3 for some $\xi > 0$. Assume **(A1)**, **(A2)**. With T defined in* (22)*, assume that $\exists\theta > 0$ s.t. for all $u \in T$ with $\|u\|_K \leq 1$,*

$$\theta^2\|u\|_K^2 \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \left\{\ell(Y_i, X_i^\top\beta^* + X_i^\top u) - \ell(Y_i, X_i^\top\beta^*) - u^\top X_i\ell'(Y_i, X_i^\top\beta^*)\right\}, \tag{23}$$

as well as

$$L(2+5\xi)\sqrt{2s\log(p/s)/n} \leq B_3^{1/2}\phi(T)\theta^2 \times \begin{cases}1/\sigma^*, & \textit{for } \ell\textit{, the squared loss,}\\ 2, & \textit{for } \ell\textit{, the logistic loss.}\end{cases} \tag{24}$$

Then with probability at least $1 - 2/(\xi^2\log(p/s)(p/s)^\xi)$,

$$\|\hat{\beta} - \beta^*\|_K \leq \frac{L(2+5\xi)}{B_3^{1/2}\phi(T)\theta^2}\sqrt{\frac{2s\log(p/s)}{n}} \times \begin{cases}\sigma^*, & \textit{for } \ell\textit{, the squared loss,}\\ 0.5, & \textit{for } \ell\textit{, the logistic loss.}\end{cases} \tag{25}$$

Assumption (23) is verified in Proposition F.1 and the assumptions of this proposition are satisfied for both squared and logistic loss. Proposition above proves the rate for logistic lasso with $\sqrt{\log(p/s)}$ instead of (usually seen) $\sqrt{\log(p)}$. See [15] for a similar result that requires more stringent conditions on Σ.

The argument in Proposition 3.4 is not specific to the Lasso. The same argument yields a similar bound for the Group-Lasso with rate $\sqrt{sd + \log(M/s)}/n\sqrt{n}$ under the setting of Lemma 3.5 using the Gaussian-width bound from Lemma 3.6.

Proof of Proposition 3.4. Set $\delta = \hat{\beta} - \beta^*$. From the definition of $\hat{\beta}$ we get that

$$\begin{aligned} f_n(\beta^* + \delta) + \lambda\|\beta^* + \delta\|_1 &\leq f_n(\beta^*) + \lambda\|\beta^*\|_1 \\ f_n(\beta^* + \delta) - f_n(\beta^*) &\leq \lambda(\|\beta^*\|_1 - \|\beta^* + \delta\|_1). \end{aligned} \tag{75}$$

Adding $-\delta^\top \nabla f_n(\beta^*)$ from both sides, we get

$$f_n(\beta^* + \delta) - f_n(\beta^*) - \delta^\top \nabla f_n(\beta^*) \leq -\delta^\top \nabla f_n(\beta^*) + \lambda(\|\beta^*\|_1 - \|\beta^* + \delta\|_1). \tag{76}$$

Applying triangle inequality and then Hölder's inequality on S (the support of β^*), we obtain

$$\begin{aligned} f_n(\beta^* + \delta) - f_n(\beta^*) - \delta^\top \nabla f_n(\beta^*) &\leq -\delta^\top \nabla f_n(\beta^*) + \lambda(\|\delta_S\|_1 - \|\delta_{S^c}\|_1) \\ &\leq -\delta^\top \nabla f_n(\beta^*) + \lambda(\sqrt{s}\|\delta\|_2 - \|\delta_{S^c}\|_1). \end{aligned} \tag{77}$$

From Lemma A.1, take $Z := n^{1/2}\Sigma^{-1/2}\nabla f_n(\beta^*)$ (this will be $\tilde{Z}$ for logistic loss). Define $U = \Sigma^{1/2}Z/(L\sigma*)$ for squared loss and $U = 2\Sigma^{1/2}\tilde{Z}/L$ for logistic loss (as in the proof of Lemma 3.3), we get on event (41)

$$f_n(\beta^* + \delta) - f_n(\beta^*) - \delta^\top \nabla f_n(\beta^*) \leq \begin{cases} \sigma^* L n^{-1/2}\left[U^\top\delta + \sigma^* L^{-1}\lambda n^{1/2}\|\delta\|\sqrt{s}\right], & \text{for squared loss} \\ 0.5 L n^{-1/2}\left[U^\top\delta + 2L^{-1}\lambda n^{1/2}\|\delta\|\sqrt{s}\right], & \text{for logistic loss} \end{cases} \tag{78}$$

We will now complete the proof for squared loss and the result for logistic loss follows by replacing σ^* by $1/2$.

$$\begin{aligned} f_n(\beta^* + \delta) - f_n(\beta^*) - \delta^\top \nabla f_n(\beta^*) &\leq \sigma^* L\|\delta\|(2 + 5\xi)\sqrt{2s\log(p/s)/n} \\ &\leq \frac{\sigma^* L\|\delta\|_K(2 + 5\xi)}{B_3^{1/2}\phi(T)}\sqrt{2s\log(p/s)/n}. \end{aligned} \tag{79}$$

The last inequality above follows from the fact that $\|u\|_K \geq B_3^{-1/2}\|u\|_\Sigma \geq B_3^{-1/2}\phi(T)^{-1}\|u\|$. Now set $t = \min\{1, \|\delta\|_K^{-1}\sigma^* L(2 + 5\xi)\sqrt{s\log(p/s)/n}/(2B_3^{1/2}\phi(T)\theta^2)\}$ so that $\|t\delta\|_K \leq 1$ by assumption (24). Combining (23) with $u = t\delta$ and (79) yields

$$\begin{aligned} \theta^2 t^2\|\delta\|_K^2 &\leq f_n(\beta^* + u) - f_n(\beta^*) - u^\top \nabla f_n(\beta^*) \\ &\leq t[f_n(\beta^* + \delta) - f_n(\beta^*) - \delta^\top \nabla f_n(\beta^*)] \leq t\frac{\sigma^* L\|\delta\|_K(2 + 5\xi)}{B_3^{1/2}\phi(T)}\sqrt{2s\log(p/s)/n}. \end{aligned} \tag{80}$$

Rearranging this yields,

$$t \leq \frac{\sigma^* L(2 + 5\xi)}{B_3^{1/2}\phi(T)\theta^2\|\delta\|_K}\sqrt{2s\log(p/s)/n}.$$

By definition of t, this inequality implies $t = 1$ and hence

$$\frac{\sigma^* L(2 + 5\xi)}{B_3^{1/2}\phi(T)\theta^2\|\delta\|_K}\sqrt{2s\log(p/s)/n} \geq 1 \quad \Rightarrow \quad \|\delta\|_K \leq \frac{\sigma^* L(2 + 5\xi)}{B_3^{1/2}\phi(T)\theta^2}\sqrt{2s\log(p/s)/n}.$$

□

G Proof of sparsity of η

Proposition 3.7. *Assume (A1), (A2). Let the setting of Lemma 3.6 be fulfilled. Fix λ as in Lemma 3.5 for both squared and logistic loss for some $\xi > 0$ and T be the cone defined in Lemma 3.5. If $\|K\|_{op} \leq C_{\max} < \infty$ and the assumptions of Proposition 3.4 hold, then*

$$\mathbb{P}\left(|\{k \in [M] : \eta_{G_k} \neq 0\}| \leq s\tilde{C}\right) \geq 1 - 2/(\xi^2 \log(M/s)(M/s)^{\xi}),$$

where $\tilde{C} := 1 + C_{\max}\{2(3+\xi)(1+\xi^{-1})\}^2 B_3^2 \phi(T)^{-2}$. For the squared loss, the same holds for $\hat{\beta}$ with $\tilde{C}$ replaced by $(1+o(1))\tilde{C}$ provided $\phi(T)^{-1}\sqrt{sd + s\log(M/s)}/\sqrt{n} \to 0$.

Proof of Proposition 3.7. The estimator η is defined by

$$\eta := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\left\|K^{1/2}(\beta - \beta^*) - n^{-1/2}Z\right\|^2 + h(\beta)$$

where h is the Group-Lasso penalty given by (29) for some tuning parameter $\lambda > 0$, and

$$Z := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} K^{-1/2} X_i \ell'(Y_i, X_i^\top \beta^*).$$

For the squared loss, Z is also given in (38) while Z for the logistic loss is given by $\tilde{Z}$ in (39). By the KKT conditions, we get

$$0 \in K(\eta - \beta^*) - n^{-1/2}K^{1/2}Z + \partial h(\eta)$$

where $\partial h(\eta)$ is the sub-differential of h at η. This implies that for any group k such that $\eta_{G_k} \neq 0$,

$$\left\|\left(K(\eta - \beta^*) - n^{-1/2}K^{1/2}Z\right)_{G_k}\right\| = \lambda. \tag{81}$$

Let $\hat{A} \subset [M]$ be the set of s largest values of $\|(K^{1/2}Z)_{G_k}\|$. Let $\text{supp}_G(\eta) = \{k \in [M] : \eta_{G_k} \neq 0\}$ and let $\mathcal{B}$ be a subset of $\text{supp}_G(\eta)$ such that $\hat{A} \cap \mathcal{B} = \emptyset$ (or equivalently $\mathcal{B} = \hat{A}^c \cap \text{supp}_G(\eta)$). Note that

$$\begin{aligned} |\text{supp}_G(\eta)| &= |\text{supp}_G(\eta) \cap \hat{A}| + |\text{supp}_G(\eta) \cap \hat{A}^c| = |\text{supp}_G(\eta) \cap \hat{A}| + |\mathcal{B}| \\ &\leq |\hat{A}| + |\mathcal{B}| = s + |\mathcal{B}|. \end{aligned} \tag{82}$$

Hence it is enough to bound $|\mathcal{B}|$. Set $\hat{\lambda} = n^{-1/2}\max_{k \in \hat{A}^c}\|(K^{1/2}Z)_{G_k}\|$. Summing the squares of the KKT condition (81) above for $j \in \mathcal{B}$ yields

$$\begin{aligned} |\mathcal{B}|\lambda^2 &= \sum_{k \in \mathcal{B}}\left(K(\eta - \beta^*) - n^{-1/2}K^{1/2}Z\right)_{G_k}^2 \\ &= (n^{-1/2}Z - K^{1/2}(\eta - \beta^*))^\top \left(\sum_{k \in \mathcal{B}}\sum_{j \in G_k} K^{1/2}e_j e_j^\top K^{1/2}\right)(n^{-1/2}Z - K^{1/2}(\eta - \beta^*)) \\ &= (n^{-1/2}Z - K^{1/2}(\eta - \beta^*))^\top M (n^{-1/2}Z - K^{1/2}(\eta - \beta^*)), \end{aligned} \tag{83}$$

where $M = \sum_{k \in \mathcal{B}}\sum_{j \in G_k} K^{1/2}e_j e_j^\top K^{1/2}$. Taking square root and using triangle inequality, we get

$$\begin{aligned} \sqrt{|\mathcal{B}|}\lambda &\leq \sqrt{\sum_{k \in \mathcal{B}}\|n^{-1/2}(K^{1/2}Z)_{G_k}\|^2} + \|M^{1/2}K^{1/2}(\eta - \beta^*)\| \\ &\leq \sqrt{|\mathcal{B}|}\hat{\lambda} + \|M^{1/2}K^{1/2}(\eta - \beta^*)\|. \end{aligned} \tag{84}$$

The second inequality above follows from the fact that $\mathcal{B} \subset \hat{A}^c$ and the definition of $\hat{A}$. By (46), with probability at least given by the right hand side of (45), $\hat{\lambda} \leq (1+\xi)^{-1}\lambda$ and hence

$$\sqrt{|\mathcal{B}|} \;\leq\; (1 + \xi^{-1})\|M^{1/2}K^{1/2}(\eta - \beta^*)\|/\lambda \tag{85}$$

Therefore, $\sqrt{|\mathcal{B}|} \leq (1+\xi^{-1})\|M\|_{op}^{1/2}\|K^{1/2}(\eta-\beta^*)\|/\lambda$. By strong convexity of the quadratic program (5) (cf., e.g. [6, Lemma 1] or [7, Lemma A.2]) we have

$$\tfrac{1}{2}\|K^{1/2}(\eta-\beta^*)\|^2 \leq n^{-1/2}Z^T\Sigma^{1/2}(\eta-\beta^*) + \lambda\sum_{k=1}^{M}(\|\beta^*_{G_k}\| - \|\eta_{G_k}\|).$$

On event (45), by the rightmost inequality of (47) in the proof of Lemma 3.5 we have $\eta-\beta^* \in T$ for the set T defined in Lemma 3.6, and by the inequalities of (47), the previous display yields

$$\tfrac{1}{2}\|K^{1/2}(\eta-\beta^*)\|^2 \leq (3+\xi)\sqrt{s}\lambda\|\eta-\beta^*\| \leq (3+\xi)\sqrt{s}\lambda\phi(T)^{-1}B_3\|K^{1/2}(\eta-\beta^*)\| \tag{86}$$

and $\|K^{1/2}(\eta-\beta^*)\| \leq 2(3+\xi)\sqrt{s}\lambda\phi(T)^{-1}B_3$. Plugging this bound back in (85) we obtain

$$\sqrt{|\mathcal{B}|} \leq \sqrt{s}\ \|K_{\bar{G},\bar{G}}\|_{op}^{1/2}\ 2(3+\xi)(1+\xi^{-1})B_3\phi(T)^{-1} \tag{87}$$

where $\bar{G} = \cup_{k\in\mathcal{B}}G_k$. Hence we obtain $|\mathcal{B}| \lesssim s$ as required for any $\mathcal{B}$ such that the ratio $\|K_{\bar{G},\bar{G}}\|_{op}^{1/2}/\phi(T)$ is bounded.

The proof for $\hat{\beta}$ (in the squared loss case) follows the same argument. The only major difference is that we have the empirical Gram matrix $\mathsf{X}^T\mathsf{X}/n$ instead of Σ (where X is the design matrix with rows $X_1, ..., X_n$), and we need to bound the quantities $\|(\mathsf{X}^T\mathsf{X}/n)_{\bar{G},\bar{G}}\|_{op}$ and $\|\mathsf{X}(\hat{\beta}-\beta^*)\|/\sqrt{n}$. It is enough to notice that $\|\mathsf{X}(\hat{\beta}-\beta^*)\|/\sqrt{n} = \|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|(1+o(1))$ and $\|\mathsf{X}^T\mathsf{X}/n)_{\bar{G},\bar{G}}\|_{op} \leq (1+o(1))\|\Sigma_{\bar{G},\bar{G}}\|_{op}$ by an application of [28] with the Gaussian-width bound given in Lemma 3.6. $\square$