A scalable estimate of the out-of-sample prediction error via approximate leave-one-out

Kamiar Rahnama Rad *
Baruch College, City University of New York
and
Arian Maleki †
Columbia University

Abstract

The paper considers the problem of out-of-sample risk estimation under the high dimensional settings where standard techniques such as K-fold cross validation suffer from large biases. Motivated by the low bias of the leave-one-out cross validation (LO) method, we propose a computationally efficient closed-form approximate leave-one-out formula (ALO) for a large class of regularized estimators. Given the regularized estimate, calculating ALO requires minor computational overhead. With minor assumptions about the data generating process, we obtain a finite-sample upper bound for $|\text{LO} - \text{ALO}|$. Our theoretical analysis illustrates that $|\text{LO}-\text{ALO}| \to 0$ with overwhelming probability, when $n, p \to \infty$, where the dimension p of the feature vectors may be comparable with or even greater than the number of observations, n. Despite the high-dimensionality of the problem, our theoretical results do not require any sparsity assumption on the vector of regression coefficients. Our extensive numerical experiments show that $|\text{LO} - \text{ALO}|$ decreases as n, p increase, revealing the excellent finite sample performance of ALO. We further illustrate the usefulness of our proposed out-of-sample risk estimation method by an example of real recordings from spatially sensitive neurons (grid cells) in the medial entorhinal cortex of a rat.

Keywords: High-dimensional statistics, Regularized estimation, Out-of-sample risk estimation, Cross validation, Generalized linear models.

1 Introduction

1.1 Main objectives

Consider a dataset $\mathcal{D} = \{(y_1, \boldsymbol{x_1}), (y_2, \boldsymbol{x_2}), \ldots, (y_n, \boldsymbol{x_n})\}$ where $\boldsymbol{x_i} \in \mathrm{R}^p$ and $y_i \in \mathrm{R}$. In many applications, we model these observations as independent and identically distributed draws from some joint distribution

*K.R. is supported by the NSF DMS grant 1810888, and the Eugene M. Lang Junior Faculty Research Fellowship.
†A.M. gratefully acknowledges NSF DMS grant 1810888.

$q(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)p(\boldsymbol{x}_i)$ where the superscript $\top$ denotes the transpose of a vector. To estimate the parameter $\boldsymbol{\beta}^*$ in such models, researchers often use the optimization problem

$$\hat{\boldsymbol{\beta}} \triangleq \arg\min_{\boldsymbol{\beta} \in \mathrm{R}^p} \Big\{ \sum_{i=1}^n \ell(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \Big\}, \tag{1}$$

where ℓ is called the loss function, and is typically set to $-\log q(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta})$ when q is known, and $r(\boldsymbol{\beta})$ is called the regularizer. In many applications, such as parameter tuning or model selection, one would like to estimate the *out-of-sample prediction error*, defined as

$$\mathrm{Err}_{\mathrm{extra}} \triangleq \mathrm{E}[\phi(y_{\mathrm{new}}, \boldsymbol{x}_{\mathrm{new}}^\top \hat{\boldsymbol{\beta}})|\mathcal{D}], \tag{2}$$

where $(y_{\mathrm{new}}, \boldsymbol{x}_{\mathrm{new}})$ is a new sample from the distribution $q(y|\boldsymbol{x}^\top \boldsymbol{\beta}^*)p(\boldsymbol{x})$ independent of $\mathcal{D}$, and ϕ is a function that measures the closeness of y_{new} to $\boldsymbol{x}_{\mathrm{new}}^\top \hat{\boldsymbol{\beta}}$. A standard choice for ϕ is $-\log q(y|\boldsymbol{x}^\top \boldsymbol{\beta})$. However, in general we may use other functions too. Since $\mathrm{Err}_{\mathrm{extra}}$ depends on the rarely known joint distribution of $(y_i, \boldsymbol{x}_i)$, a core problem in model assessment is to estimate it from data.

This paper considers a computationally efficient approach to the problem of estimating $\mathrm{Err}_{\mathrm{extra}}$ under the high-dimensional setting, where both n and p are large, but n/p is a fixed number, possibly less than one. This high dimensional setting has received a lot of attention [El Karoui, 2018, El Karoui et al., 2013, Bean et al., 2013, Donoho and Montanari, 2016, Nevo and Ritov, 2016, Su et al., 2017, Dobriban and Wager, 2018]. But the problem of estimating $\mathrm{Err}_{\mathrm{extra}}$ has not been carefully studied in generality, and as a result the issues of the existing techniques and their remedies have not been explored. For instance, a popular technique in practice is the K-fold cross validation, where K is a small number, e.g. 3 or 5. Figure 1 compares the performance of the K-fold cross validation for 4 different values of K on a LASSO linear regression problem. This figure implies that in high-dimensional settings, K-fold cross validation suffers from a large bias, unless K is a large number. This bias is due to the fact that in high-dimensional settings the fold that is removed in the training phase, may have a major effect on the solution of (1). This claim can be easily seen for LASSO linear regression with an IID design matrix using phase transition diagrams [Donoho et al., 2011]. To summarize, as the number of folds increases, the bias of the estimates reduces at the expense of a higher computational complexity.

In this paper, we consider the most extreme form of cross validation, namely leave-one-out cross-validation (LO), which according to Figure 1 is the least biased cross validation based estimate of the out-of-sample error. We will use the fact that both n and p are large numbers to approximate LO for both smooth and non-smooth regularizers. Our estimate, called approximate leave-one-out (ALO), requires solving the optimization problem (1) once. Then, it uses $\hat{\boldsymbol{\beta}}$ to approximate LO without solving the optimization problem again. In addition to obtaining $\hat{\boldsymbol{\beta}}$, ALO requires a matrix inversion and two matrix-matrix multi-

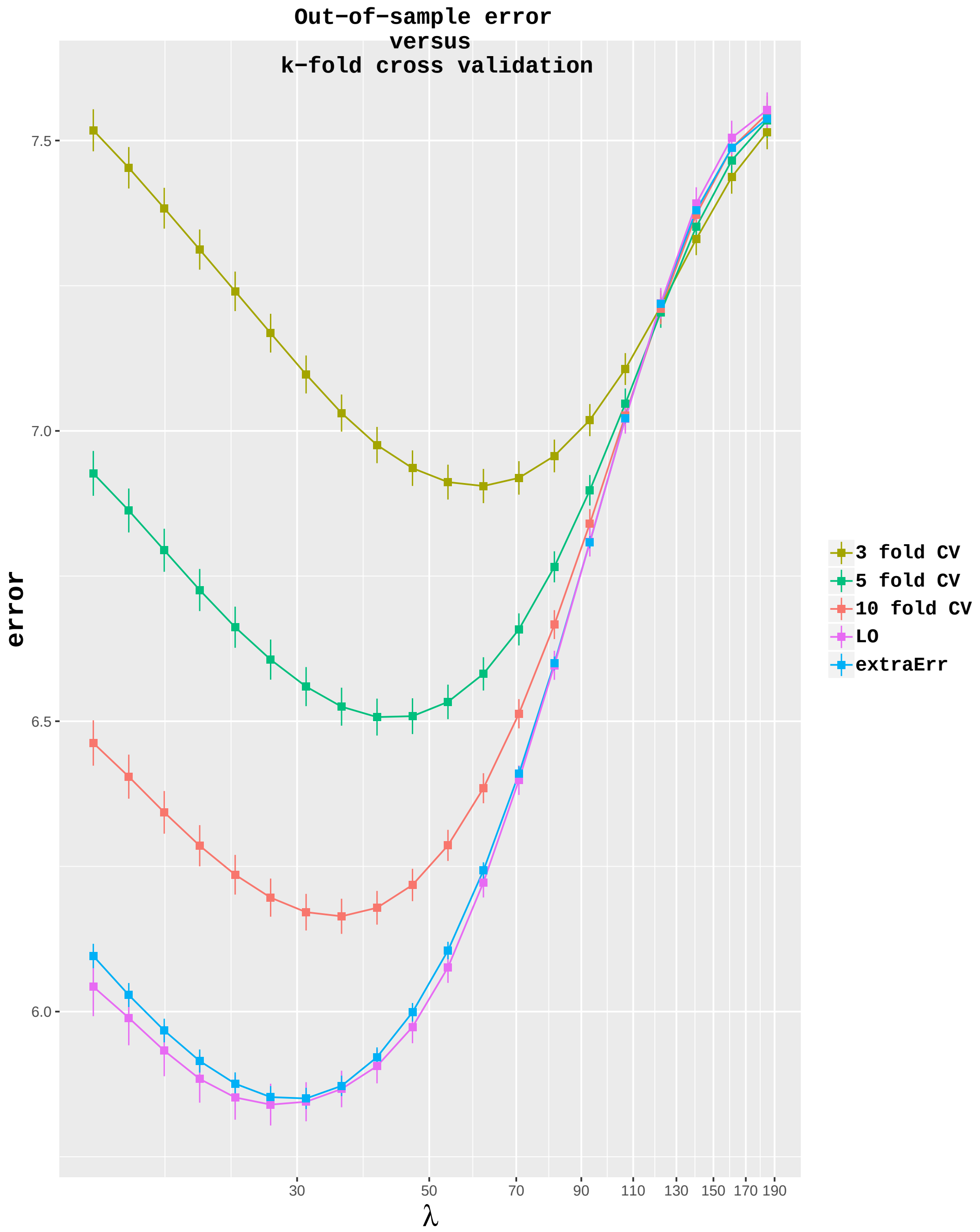
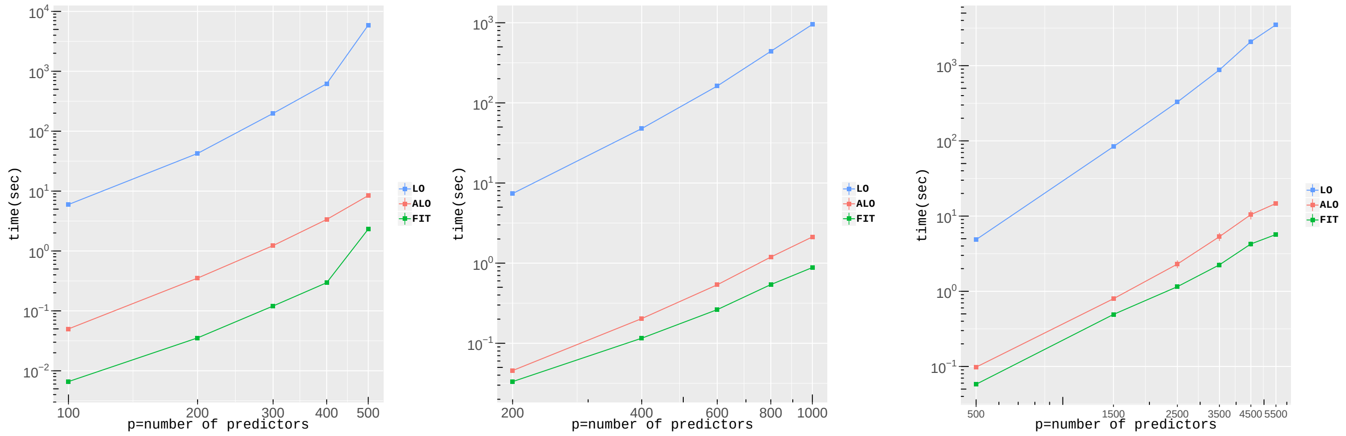


Figure 1: Comparison of K-fold cross validation (for $K = 3, 5, 10$) and leave-one-out cross validation with the true (oracle-based) out-of-sample error for the LASSO problem where $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top\boldsymbol{\beta})^2$ and $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. In high-dimensional settings the upward bias of K-fold CV clearly decreases as number of folds increase. Data is $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma^2\boldsymbol{I})$ where $\boldsymbol{X} \in \mathrm{R}^{p\times n}$. The number of nonzero elements of the true $\boldsymbol{\beta}^*$ is set to k and their values is set to $1/3$. Dimensions are $(p, n, k) = \big(1000, 250, 50\big)$ and $\sigma = 2$. The rows of $\boldsymbol{X}$ are independent $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. Extra-sample test data is $y_{\text{new}} \sim \mathrm{N}(\boldsymbol{x}_{\text{new}}^\top\boldsymbol{\beta}^*, \sigma^2)$ where $\boldsymbol{x}_{\text{new}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. The true (oracle-based) out-of-sample prediction error is $\mathrm{Err}_{\text{extra}} = \mathrm{E}[(y_{\text{new}} - \boldsymbol{x}_{\text{new}}^\top\hat{\boldsymbol{\beta}})^2|\boldsymbol{y}, \boldsymbol{X}] = \sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$. All depicted quantities are averages based on 500 random independent samples, and error bars depict one standard error.



(a) Elastic-net linear regression (section 5.2.1) for $\frac{n}{p} = 5$.

(b) LASSO logistic regression (section 5.2.2) for $\frac{n}{p} = 1$.

(c) Elastic-net Poisson regression (section 5.2.3) for $\frac{n}{p} = \frac{1}{10}$.

Figure 2: The time to compute ALO and LO. FIT refers to the time to fit $\hat{\boldsymbol{\beta}}$ and the ALO time includes computing $\hat{\boldsymbol{\beta}}$. Calculating LO takes orders of magnitude longer than ALO.

plications. Despite these extra steps ALO offers a significant computational saving compared to LO. This point is illustrated in Figure 2 by comparing the computational complexity of ALO with that of LO and a single fit as both n and p increase for various data shapes, that is $n > p$, $n = p$, and $n < p$. Details of this simulation are given in Section 5.2.4.

The main algorithmic and theoretical contributions of this paper are as follows. First, our computational complexity comparison between LO and ALO, confirmed by extensive numerical experiments, show that ALO offers a major reduction in the computational complexity of estimating the out-of-sample risk. Moreover, with minor assumptions about the data generating process, we obtain a finite-sample upper bound for $|\text{LO} - \text{ALO}|$, proving that under the high-dimensional settings ALO presents a sensible approximation of LO for a large class of regularized estimation problems in the generalized linear family. Finally, we provide readily usable R implementation of ALO online; see `https://github.com/Francis-Hsu/alocv`, and we illustrate the usefulness of our proposed out-of-sample risk estimation in unexpected scenarios that fail to satisfy the assumptions of our theoretical framework. Specifically, we present a novel neuroscience example about the computationally efficient tuning of the spatial scale in estimating an inhomogeneous spatial point process.

1.2 Relevant work

The problem of estimating $\text{Err}_{\text{extra}}$ from $\mathcal{D}$ has been studied for (at least) the past 50 years. Methods such as cross validation (CV) [Stone, 1974, Geisser, 1975], Allen's PRESS statistic [Allen, 1974], generalized cross validation (GCV) [Craven and Wahba, 1979, Golub et al., 1979], and bootstrap [Efron, 1983] have been proposed for this purpose. In the high dimensional setting, employing LO or bootstrap is computationally expensive and the less computationally complex approaches such as 5-fold (or 10-fold) CV suffer from high bias as illustrated in Figure 1.

As for the computationally efficient approaches, extensions of Allen's PRESS [Allen, 1974], and generalized cross validation (GCV) [Craven and Wahba, 1979, Golub et al., 1979] to non-linear models and classifiers with ridge penalty are well known: smoothing splines for generalized linear models in [O'Sullivan et al., 1986], spline estimation of generalized additive models [Burman, 1990], ridge estimators in logistic regression in [Cessie and Houwelingen, 1992], smoothing splines with non-Gaussian data using various extensions of GCV in [Gu, 1992, Xiang and Wahba, 1996, Gu and Xiang, 2001], support vector machines [Opper and Winther, 2000], kernel logistic regression in [Cawley and Talbot, 2008], and Cox's proportional hazard model with a ridge penalty in [Meijer and Goeman, 2013]. Moreover, leave-one-out approximations for posterior means of Bayesian models with Gaussian process priors using the Laplace approximation and Expectation Propagation were introduced in [Vehtari et al., 2016], and extended in [Vehtari et al., 2017]. Despite the existence of this vast literature, the performance of such approximations in high-dimensional settings is unknown except for the straightforward linear ridge regression framework. Moreover, past heuristic approaches have only considered the ridge regularizer. The results of this paper include a much broader set of regularizers; examples include but are not limited to LASSO [Tibshirani, 1996], elastic net [Zou and Hastie, 2005] and bridge [Frank and Friedman, 1993], just to name a few.

More recently, a few papers have studied the problem of estimating $\text{Err}_{\text{extra}}$ under high-dimensional settings [Mousavi et al., 2018, Obuchi and Kabashima, 2016]. The approximate message passing framework introduced in [Maleki, 2011, Donoho et al., 2009] was used in [Mousavi et al., 2018] to obtain an estimate of $\text{Err}_{\text{extra}}$ for LASSO linear regression. In another related paper, [Obuchi and Kabashima, 2016] obtained similar results using approximations popular in statistical physics. The results of [Mousavi et al., 2018] and [Obuchi and Kabashima, 2016] are only valid for cases where the design matrix has IID entries and the empirical distribution of the regression coefficients converges weakly to a distribution with a bounded second moment. In this paper, our theoretical analysis includes correlated design matrices, and regularized estimators beyond LASSO linear regression.

In addition to these approaches, another contribution has been to study GCV and $\text{Err}_{\text{extra}}$ for restricted least-squares estimators of submodels of the overall model without regularization [Breiman and Freedman, 1983, Leeb, 2008, Leeb, 2009]. In [Leeb, 2008] it was shown that a variant of GCV converges to $\text{Err}_{\text{extra}}$ uniformly over a collection of candidate models provided that there are not too many candidate models, ruling out complete subset selection. Moreover, since restricted least-squares estimators are studied, the conclusions exclude the regularized problems considered in this paper.

Finally, it is worth mentioning that in another line of work, strategies have been proposed to obtain unbiased estimates of the in-sample error. In contrast to the out-of-sample error, the in-sample error is about the prediction of new responses for the same explanatory variables as in the training data. The literature of in-sample error estimation is too vast to be reviewed here. Mallow's C_p [Mallows, 1973], Akaike's Information Criterion (AIC) [Akaike, 1974, Hurvich and Tsai, 1989], Stein's Unbiased Risk Estimate (SURE) [Stein, 1981, Zou et al., 2007, Tibshirani and Taylor, 2012] and Efron's Covariance Penalty [Efron, 1986] are seminal examples of in-sample error estimators. When n is much larger than p, the in-sample prediction error is expected to be close to the out-of-sample prediction error. The problem is that in high-dimensional settings, where n is of the same order as (or even smaller than) p, the in-sample and out-of-sample errors are different.

The rest of the paper is organized as follows. After introducing the notations, we first present the approximate leave-one-out formula (ALO) for twice differentiable regularizers in Section 2.1. In Section 2.2, we show how ALO can be extended to nonsmooth regularizers such as LASSO using Theorem 1 and Theorem 2. In Section 3, we compare the computational complexity and memory requirements of ALO and LO. In Section 4, we present Theorem 3, illustrating with minor assumptions about the data generating process that $|\text{LO} - \text{ALO}| \to 0$ with overwhelming probability, when $n, p \to \infty$, where p may be comparable with or even greater than n. The numerical examples in Section 5 study the statistical accuracy and computational efficiency of the approximate leave-one-out approach. To illustrate the accuracy and computational efficiency of ALO we apply it to synthetic and real data in Section 5. We generate synthetic data, and compare ALO and LO for elastic-net linear regression in Section 5.2.1, LASSO logistic regression in Section 5.2.2, and elastic-net Poisson regression in Section 5.2.3. For real data we apply LASSO, elastic-net and ridge logistic regression to sonar returns from two undersea targets in Section 5.3.1, and we apply LASSO Poisson regression to real recordings from spatially sensitive neurons (grid cells) in Section 5.3.2. Our synthetic and real data examples cover various data shapes, that is $n > p$, $n = p$ and $n < p$. In Section 6 we discuss directions for future work. Technical proofs are collected in Section A, the appendix.

1.3 Notation

We first review the notations that will be used in the rest of the paper. Let $\boldsymbol{x}_{\boldsymbol{i}}^\top \in \mathrm{R}^{1\times p}$ stand for the ith row of $\boldsymbol{X} \in \mathrm{R}^{n\times p}$. $\boldsymbol{y}_{/i} \in \mathrm{R}^{(n-1)\times 1}$ and $\boldsymbol{X}_{/i} \in \mathrm{R}^{(n-1)\times p}$ stand for $\boldsymbol{y}$ and $\boldsymbol{X}$, excluding the ith entry y_i and the ith row $\boldsymbol{x}_{\boldsymbol{i}}^\top$, respectively. The vector $\boldsymbol{a} \odot \boldsymbol{b}$ stands for the entry-wise product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we use $\boldsymbol{a} < \boldsymbol{b}$ to indicate element-wise inequalities. Moreover, $|\boldsymbol{a}|$ stands for the vector obtained by applying the element-wise absolute value to every element of $\boldsymbol{a}$. For a set $S \subset \{1, 2, 3, \ldots, p\}$, let $\boldsymbol{X}_S$ stands for the submatrix of $\boldsymbol{X}$ restricted to *columns* indexed by S. Likewise, we let $\boldsymbol{x}_{i,S} \in \mathrm{R}^{|S|\times 1}$ stand for for subvector of $\boldsymbol{x}_i$ restricted to the entries indexed by S. For a vector $\boldsymbol{a}$, depending on which notation is easier to read, we may use $[\boldsymbol{a}]_i$ or a_i to denote the ith entry of $\boldsymbol{a}$. The diagonal matrix with elements of the vector $\boldsymbol{a}$ is referred to as diag$[\boldsymbol{a}]$. Moreover, define

$$\begin{aligned}
\dot{\phi}(y,z) &\triangleq \frac{\partial \phi(y,z)}{\partial z}, \quad \dot{\ell}_i(\boldsymbol{\beta}) \triangleq \frac{\partial \ell(y_i|z)}{\partial z}|_{z=\boldsymbol{x}_i^\top \boldsymbol{\beta}}, \quad \ddot{\ell}_i(\boldsymbol{\beta}) \triangleq \frac{\partial^2 \ell(y_i|z)}{\partial z^2}|_{z=\boldsymbol{x}_i^\top \boldsymbol{\beta}} \\
\dot{\boldsymbol{\ell}}_{/i}(.) &\triangleq [\dot{\ell}_1(.), \cdots, \dot{\ell}_{i-1}(.), \dot{\ell}_{i+1}(.), \cdots, \dot{\ell}_n(.)]^\top, \\
\ddot{\boldsymbol{\ell}}_{/i}(.) &\triangleq [\ddot{\ell}_1(.), \cdots, \ddot{\ell}_{i-1}(.), \ddot{\ell}_{i+1}(.), \cdots, \ddot{\ell}_n(.)]^\top.
\end{aligned}$$

The notation poly log n denotes polynomial of $\log n$ with a finite degree. Finally, let $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ stand for the largest and smallest singular values of $\boldsymbol{A}$, respectively.

2 Approximate leave-one-out

2.1 Twice differentiable losses and regularizers

The leave-one-out cross validation estimate is defined through the following formula:

$$\text{LO} \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}), \tag{3}$$

where

$$\hat{\boldsymbol{\beta}}_{/i} \triangleq \underset{\boldsymbol{\beta}\in \mathrm{R}^p}{\arg\min} \Big\{ \sum_{j\neq i} \ell(y_j|\boldsymbol{x}_j^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \Big\}, \tag{4}$$

is the leave-i-out estimate. If done naively, the calculation of LO asks for the optimization problem (4) to be solved n times, a computationally demanding task when p and n are large. To resolve this issue, we use the following simple strategy: Instead of solving (4) accurately, we use one step of the Newton method for solving (4) with initialization $\hat{\boldsymbol{\beta}}$. Note that this step requires both ℓ and r to be twice differentiable. We will explain how this limitation can be lifted in the next section. The Newton step leads to the following simple approximation of $\hat{\boldsymbol{\beta}}_{/i}$:[1]

$$\tilde{\boldsymbol{\beta}}_{/i} = \hat{\boldsymbol{\beta}} + \Big(\sum_{j\neq i} \boldsymbol{x_j}\boldsymbol{x_j}^\top \ddot{\ell}(y_j|\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}) + \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] \Big)^{-1} \boldsymbol{x_i} \dot{\ell}(y_i|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is defined in (1). Note that $\sum_{j\neq i} \boldsymbol{x_j}\boldsymbol{x_j}^\top \ddot{\ell}(y_j|\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}) + \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]$ is still dependent on the observation that is removed. Hence, the process of computing the inverse (or solving a linear equation) must be repeated n times. Standard methods for calculating inverses (or solving linear equations) require cubic time and quadratic space (see Appendix C.3 in [Boyd and Vandenberghe, 2004]), rendering them impractical for high-dimensional applications when repeated n times[2]. We use the Woodburry lemma to reduce the

[1] Note that in the rest of the paper for notational simplicity of our theoretical results we have assumed that $r(\boldsymbol{\beta}) = \sum_{i=1}^p r(\beta_i)$. However, the extension to non-separable regularizers is straightforward.

[2] A natural idea for reducing the computational burden involves exploiting structures (such as sparsity and banded-ness) of the involved matrices. However, in this paper we do not make any assumption regarding the structure of $\boldsymbol{X}$.

computational cost:

$$\Big(\sum_{j\neq i}\boldsymbol{x_j}\boldsymbol{x_j}^\top\ddot{\ell}(y_j|\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}})+\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]\Big)^{-1}=\boldsymbol{J}^{-1}+\frac{\boldsymbol{J}^{-1}\boldsymbol{x_i}\ddot{\ell}(y_i|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})\boldsymbol{x_i}^\top\boldsymbol{J}^{-1}}{1-\boldsymbol{x_i}^\top\boldsymbol{J}^{-1}\boldsymbol{x_i}\ddot{\ell}(y_i|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})},\tag{5}$$

where $\boldsymbol{J}=(\sum_{j=1}^n\boldsymbol{x_j}\boldsymbol{x_j}^\top\ddot{\ell}(y_j|\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}})+\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})])$. Following this approach we define ALO as

$$\mathrm{ALO}\ \triangleq\ \frac{1}{n}\sum_{i=1}^n\phi\left(y_i,\boldsymbol{x_i}^\top\tilde{\boldsymbol{\beta}}_{/i}\right)=\frac{1}{n}\sum_{i=1}^n\phi\left(y_i,\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}+\left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right),\tag{6}$$

where

$$\boldsymbol{H}\ \triangleq\ \boldsymbol{X}\left(\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]+\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})].\tag{7}$$

Algorithm 1 summarizes how one should obtain an ALO estimate of the $\mathrm{Err}_{\mathrm{extra}}$. We will show that under the high-dimensional settings one Newton step is sufficient for obtaining a good approximation of $\hat{\boldsymbol{\beta}}_{/i}$, and the difference $|\mathrm{ALO}-\mathrm{LO}|$ is small when either n or both n,p are large. However, before that we resolve the differentiability issue of the approach we discussed above.

Algorithm 1 Risk estimation with ALO for twice differentiable losses and regularizers

Input. $(\boldsymbol{x_1},y_1),(\boldsymbol{x_2},y_2),\ldots,(\boldsymbol{x_n},y_n)$.
Output. $\mathrm{Err}_{\mathrm{extra}}$ estimate.

1. Calculate $\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}\in\mathrm{R}^p}{\arg\min}\Big\{\sum_{i=1}^n\ell(y_i|\boldsymbol{x_i}^\top\boldsymbol{\beta})+\lambda r(\boldsymbol{\beta})\Big\}$.
2. Obtain $\boldsymbol{H}=\boldsymbol{X}\left(\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]+\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$.
3. The estimate of $\mathrm{Err}_{\mathrm{extra}}$ is given by $\frac{1}{n}\sum_{i=1}^n\phi\left(y_i,\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}+\left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)$.

2.2 Nonsmooth regularizers

The Newton step, used in the derivation of ALO, requires the twice differentiability of the loss function and regularizer. However, in many modern applications non-smooth regularizers, such as LASSO, are preferable. In this section, we explain how ALO can be used for non-smooth regularizers. We start with the ℓ_1-regularizer, and then extend it to the other bridge estimators. A similar approach can be used for other non-smooth regularizers. Consider

$$\hat{\boldsymbol{\beta}}\triangleq\underset{\boldsymbol{\beta}\in\mathrm{R}^p}{\arg\min}\Big\{\sum_{i=1}^n\ell(y_i|\boldsymbol{x_i}^\top\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1\Big\}.\tag{8}$$

Let $\hat{\boldsymbol{g}}$ be a subgradient of $\|\boldsymbol{\beta}\|_1$ at $\hat{\boldsymbol{\beta}}$, denoted by $\hat{\boldsymbol{g}} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$. Then, the pair $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{g}})$ must satisfy the zero-subgradient condition

$$\sum_{i=1}^{n} \boldsymbol{x_i} \dot{\ell}(y_i|\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}) + \lambda \hat{\boldsymbol{g}} = 0.$$

As a starting point we use a smooth approximation of the function $\|\boldsymbol{\beta}\|_1$ in our ALO formula. For instance, we can use the following approximation introduced in [Schmidt et al., 2007]:

$$r^\alpha(\boldsymbol{\beta}) = \sum_{i=1}^{p} \frac{1}{\alpha}\Big(\log(1+e^{\alpha\beta_i}) + \log(1+e^{-\alpha\beta_i})\Big).$$

Since $\lim_{\alpha\to\infty} r^\alpha(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, we can use

$$\hat{\boldsymbol{\beta}}^\alpha \triangleq \operatorname*{arg\,min}_{\boldsymbol{\beta}\in \mathrm{R}^p} \Big\{ \sum_{i=1}^{n} \ell(y_i|\boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} r^\alpha(\beta_i) \Big\}, \tag{9}$$

to obtain the following formula for ALO:

$$\mathrm{ALO}^\alpha \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}^\alpha + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha)}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha)}\right)\left(\frac{H^\alpha_{ii}}{1-H^\alpha_{ii}}\right)\right) \tag{10}$$

where $\boldsymbol{H}^\alpha \triangleq \boldsymbol{X}\left(\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^\alpha)] + \boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]$. Note that $\|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2 \to 0$ as $\alpha \to \infty$, according to Lemma 15 in Section A.2. Therefore, we take the $\alpha \to \infty$ limit in (10), yielding a simplification of ALO^α in this limit. To prove this claim, we denote the active set of $\hat{\boldsymbol{\beta}}$ with S, and we suppose the following:

Assumption 1. *$\hat{\boldsymbol{\beta}}$ is the unique global minimizer of (1).*

Assumption 2. *$\hat{\boldsymbol{\beta}}^\alpha$ is the unique global minimizer of (9) for every value of α.*

Assumption 3. *$\ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$.*

Assumption 4. *The strict dual feasibility condition $\|\hat{\boldsymbol{g}}_{S^c}\|_\infty < 1$ holds.*

Theorem 1. *If Assumptions 1,2,3 and 4 hold, then*

$$\lim_{\alpha\to\infty} \mathrm{ALO}^\alpha = \frac{1}{n}\sum_{i=1}^{n} \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right), \tag{11}$$

where $\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$.

The proof of this theorem is presented in Section A.2. For the rest of the paper, the right hand side of (11) is the ALO formula we use as an approximation of LO for LASSO problems. In the simulation section, we show that the formula we obtain in Theorem 1 offers an accurate estimate of the out-of-sample prediction error. For instance, in the standard LASSO problem, where $\ell(u,v) = (u-v)^2/2$ and $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, Theorem

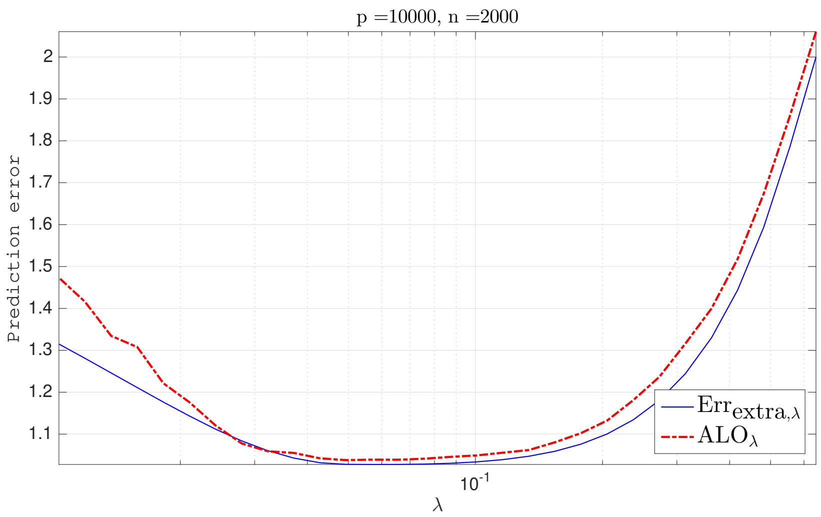


Figure 3: Out-of-sample prediction error versus ALO. Data is $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma^2 \boldsymbol{I})$ where $\sigma^2 = 1$ and $\boldsymbol{X} \in \mathrm{R}^{p\times n}$ with $p = 10000$ and $n = 2000$. The number of nonzero elements of the true $\boldsymbol{\beta}^*$ is set to $k = 400$ and their values is set to 1. The rows $\boldsymbol{x}_i^\top$ of the predictor matrix are generated randomly as $\mathrm{N}(0, \boldsymbol{\Sigma})$ with correlation structure $\mathrm{cor}(X_{ij}, X_{ij'}) = 0.3$ for all $i = 1, \ldots, n$ and $j, j' = 1, \cdots, p$. The covariance matrix $\boldsymbol{\Sigma}$ is scaled such the signal variance $\mathrm{var}(\boldsymbol{x}^\top \boldsymbol{\beta}^*) = 1$. Out-of-sample test data is $y_{\text{new}} \sim \mathrm{N}(\boldsymbol{x}_{\text{new}}^\top \boldsymbol{\beta}^*, \sigma^2)$ where $\boldsymbol{x}_{\text{new}} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$. Out-of-sample error is calculated as $\mathrm{E}_{(y_{\text{new}}, \boldsymbol{x}_{\text{new}})}[(y_{\text{new}} - \boldsymbol{x}_{\text{new}}^\top \hat{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}] = \sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ and ALO is calculated using equation (12).

1 gives the following estimate of the out-of-sample prediction error:

$$\lim_{\alpha\to\infty} \mathrm{ALO}^\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})^2}{(1 - H_{ii})^2}, \tag{12}$$

where $\boldsymbol{H} = \boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top$. Figure 3 compares this estimate with the oracle estimate of the out-of-sample prediction error on a LASSO example. More extensive simulations are reported in Section 5.

Note that Assumptions 1,2 and 3 hold for most of the practical problems. For instance, to study the conditions under which Assumption 1 holds refer to [Tibshirani et al., 2013]. Moreover, for $\ell(u, v) = (u - v)^2/2$, Assumption 1 is a consequence of Assumption 4 [Wainwright, 2009]. Assumption 4 also holds in many cases with probability one with respect to the randomness of the dataset [Wainwright, 2009, Tibshirani and Taylor, 2012]. Even if this assumption is violated in a specific problem (note that checking this assumption is straightforward), we can use the following theorem to evaluate the accuracy of the ALO formula in Theorem 1.

Theorem 2. *Let S and T denote the active set of $\hat{\boldsymbol{\beta}}$, and the set of zero coefficients at which the subgradient vector is equal to 1 or -1. Then,*

$$\boldsymbol{x}_{i,S}^\top \left(\boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1} \boldsymbol{x}_{i,S} \ddot{\ell}_i(\hat{\boldsymbol{\beta}}) < \lim\inf_{\alpha\to\infty} H_{ii}^\alpha$$
$$\lim\sup_{\alpha\to\infty} H_{ii}^\alpha < \boldsymbol{x}_{i,S\cup T}^\top \left(\boldsymbol{X}_{S\cup T}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_{S\cup T}\right)^{-1} \boldsymbol{x}_{i,S\cup T} \ddot{\ell}_i(\hat{\boldsymbol{\beta}})$$

This Theorem is proved in A.3. A simple implication of this theorem is that

$$\limsup_{\alpha\to\infty} \text{ALO}^{\alpha} \leq \frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^{\top}\hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}^{h}}{1-H_{ii}^{h}}\right)\right), \tag{13}$$

and

$$\liminf_{\alpha\to\infty} \text{ALO}^{\alpha} \geq \frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^{\top}\hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}^{l}}{1-H_{ii}^{l}}\right)\right), \tag{14}$$

where

$$\begin{aligned}\boldsymbol{H}^{l} &= \boldsymbol{X}_S\left(\boldsymbol{X}_S^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1}\boldsymbol{X}_S^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})],\\ \boldsymbol{H}^{h} &= \boldsymbol{X}_{S\cup T}\left(\boldsymbol{X}_{S\cup T}^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_{S\cup T}\right)^{-1}\boldsymbol{X}_{S\cup T}^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})].\end{aligned} \tag{15}$$

By comparing (13) and (14) we can evaluate the error in our simple formula of the risk, presented in Theorem 1. The approach we proposed above can be extended to other non-differentiable regularizers too. Below we consider two other popular classes of estimators: (i) bridge and (ii) elastic net, and show how we can derive ALO formulas for each estimator.

Bridge estimators: Consider the class of bridge estimators

$$\hat{\boldsymbol{\beta}} \triangleq \underset{\boldsymbol{\beta}\in \mathrm{R}^p}{\arg\min}\Big\{\sum_{i=1}^{n}\ell(y_i|\boldsymbol{x}_{\boldsymbol{i}}^{\top}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_q^q\Big\}, \tag{16}$$

where q is a number between $(1, 2)$. Note that these regularizers are only one time differentiable at zero. Hence, the Newton method introduced in Section 2.1 is not directly applicable. One can argue intuitively that since the regularizer is differentiable at zero, none of the regression coefficients will be zero. Hence, the regularizer is locally twice differentiable and formula (6) works well. While this argument is often correct, we can again use the idea introduced above for LASSO to obtain the following ALO formula that can be used even when an estimate of 0 is observed:

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^{\top}\hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right), \tag{17}$$

where if we define $S \triangleq \{i : \beta_i \neq 0\}$ and for $u \neq 0$, $\ddot{r}^q(u) \triangleq q(q-1)|u|^{q-2}$, then

$$\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + \lambda\text{diag}[\ddot{\boldsymbol{r}}_S^{q}(\hat{\boldsymbol{\beta}})]\right)^{-1}\boldsymbol{X}_S^{\top}\,\text{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]. \tag{18}$$

This formula is derived in Section A.4.

Elastic-net: Finally, we consider the following elastic-net estimator

$$\hat{\boldsymbol{\beta}} \triangleq \underset{\boldsymbol{\beta}\in \mathrm{R}^p}{\arg\min}\Big\{\sum_{i=1}^{n}\ell(y_i|\boldsymbol{x}_{\boldsymbol{i}}^{\top}\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1\Big\}. \tag{19}$$

Again by smoothing the ℓ_1-regularizer (similar to what we did for LASSO) we obtain the following ALO formula for the out-of-sample predictor error:

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, \boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right),$$

where $S = \{i : \hat{\beta}_i \neq 0\}$, and

$$\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + 2\lambda_1\boldsymbol{I}\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]. \tag{20}$$

We do not derive this formula, since it follows exactly the same lines as those of LASSO and bridge. Algorithm 2 summarizes all the calculations required for the calculation of ALO for elastic-net.

Algorithm 2 Risk estimation with ALO for elastic-net regularizer

Input. $(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)$.

Output. $\text{Err}_{\text{extra}}$ estimate.

1. Calculate $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{\sum_{i=1}^{n}\ell(y_i|\boldsymbol{x_i}^\top\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1\right\}$.
2. Calculate $S = \{i : \hat{\beta}_i \neq 0\}$.
3. Obtain $\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + 2\lambda_1\boldsymbol{I}\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$, where $\boldsymbol{X}_S$ only includes the columns of $\boldsymbol{X}$ that are in S.
4. The estimate of $\text{Err}_{\text{extra}}$ is given by $\frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, \boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)$.

3 Computational complexity and memory requirements of ALO

Counting the number of floating point operations algorithms require is a standard approach for comparing their computational complexities. In this section, we calculate and compare the number of operations required by ALO and LO. We first start with Algorithm 1 and then discuss Algorithm 2.

Algorithm 1

Before we start the calculations, we should warn the reader that in many cases the specific structure of the loss and/or the regularizer enables more efficient implementation of the formulas. However, here we consider the worst case scenario. Furthermore, the calculations below are concerned with the implementation of ALO and LO on a single computer, and we have not explored their parallel or distributed implementations.

The first step of Algorithm 1 requires solving an optimization problem. Several different methods exist for solving this optimization problem. Here, we discuss the interior point method and the accelerated gradient descent algorithm. Suppose that our goal is to reach accuracy ϵ. Then, interior point method requires $O(\log(1/\epsilon))$ iterations to reach this accuracy, while accelerated gradient descent requires $O(\frac{1}{\sqrt{\epsilon}})$

iterations [Nesterov, 2013]. Furthermore, each iteration of the accelerated gradient descent requires $O(np)$ operations, while each iteration of the interior point method requires $O(p^3)$ operations.

Regarding the memory usage of these two algorithms, note that in the accelerated gradient descent algorithm the memory is mainly used for storing matrix $\boldsymbol{X}$. Hence, the amount of memory that is required by this algorithm is $O(np)$. On the other hand, interior point method uses $O(p^3)$ of memory.

The second step of Algorithm 1 is to calculate the matrix $\boldsymbol{H}$. This requires inverting the matrix $\left(\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}$. In general, this inversion requires $O(p^3)$ (e.g. by using Cholesky factorization). However, if n is much smaller than p, then one can use a better trick for performing the matrix inversion; suppose that both ℓ and r are strongly convex at $\hat{\boldsymbol{\beta}}$ and define $\boldsymbol{\Gamma} \triangleq (\operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})])^{\frac{1}{2}}$, and $\boldsymbol{\Lambda} \triangleq \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]$. Then, from the matrix inversion lemma we have

$$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Gamma}^2\boldsymbol{X} + \boldsymbol{\Lambda})^{-1}\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^\top - \boldsymbol{X}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma}(\boldsymbol{I} + \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^\top. \tag{21}$$

The inversion $(\boldsymbol{I} + \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma})^{-1}$ requires $O(n^3)$ operations and $O(np)$ of memory (the main memory usage is for storing $\boldsymbol{X}$). Also, the other matrix-matrix multiplications require $O(n^2p+n^3)$ operations. Hence, overall if we use the matrix inversion lemma, then the calculation of $\boldsymbol{H}$ requires $O(n^3 + n^2p)$ operations. In summary, the calculation of $\boldsymbol{H}$ requires $O(\min(p^3 + np^2, n^3 + n^2p))$. Also, the amount of memory that is required by the algorithm is $O(np)$. The last step of ALO, i.e. Step 3 in Algorithm 1, requires only $O(np)$ operations. Hence, the calculations of ALO in Algorithm 1 requires

1. Through interior point method: $O(\min(p^3 \log(1/\epsilon) + p^3 + np^2, p^3 \log(1/\epsilon) + n^3 + n^2p))$
2. Through accelerated gradient descent: $O(\min(np\frac{1}{\sqrt{\epsilon}} + p^3 + np^2, np\frac{1}{\sqrt{\epsilon}} + n^3 + n^2p))$

Similarly, the calculation of the LO requires solving n optimization problem of the form (4). Hence, the number of floating point operations that are required for LO are:

1. Through interior point method: $O(np^3 \log(1/\epsilon))$.
2. Through accelerated gradient descent: $O(n^2p\frac{1}{\sqrt{\epsilon}})$.

Algorithm 2

Note that in Algorithm 2, we have used the specific form of the regularizer and simplified the form of $\boldsymbol{H}$. Hence, this allows for faster calculation of $\boldsymbol{H}$ and equivalently faster calculation of the ALO estimate. Again the first step of calculating ALO is to solve the optimization problem. Solving this optimization problem by the interior point method or accelerated proximal gradient descent requires $O(p^3 \log(1/\epsilon))$ and $O(np\frac{1}{\sqrt{\epsilon}})$ floating point operations respectively. The next step is to calculate $\boldsymbol{H}$. If $\hat{\boldsymbol{\beta}}$ is s-sparse, i.e., has only s non-zero coefficients, then the calculation of $\boldsymbol{H}$ requires $O(s^3 + ns^2)$ floating point operations. Also, the amount of memory required for this inversion is $O(s^2)$. Finally, the last step requires $O(np)$ operations. Hence, calculating an ALO estimate of the risk requires:

1. Through interior point method: $O(p^3 \log(1/\epsilon) + s^3 + ns^2 + np)$.

2. Through accelerated proximal gradient descent: $O(np\frac{1}{\sqrt{\epsilon}} + s^3 + ns^2 + np)$

The calculations of LO in the worst case is similar to what we had in the previous section:[3]

1. Through interior point method: $O(np^3 \log(1/\epsilon))$.

2. Through accelerated proximal gradient descent: $O(n^2 p \frac{1}{\sqrt{\epsilon}})$.

In this section, we used the number of floating point operations to compare the computational complexity of ALO and LO . However, since this approach is based on the worst case scenarios and is not capable of capturing the constants, it is less accurate than comparing the timing of algorithms through simulations. Hence, Section 5 compares the performance of ALO and LO through simulations.

Memory usage

First, we discuss Algorithm 1. We only consider the accelerated gradient descent algorithm. As discussed above, the amount of memory that is required for Step 1 of ALO is $O(np)$ (the main memory usage is for storing matrix X). For the second step, direct inversion of $\left(\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}$ requires $O(p^2)$ of memory. However, by using the formula derived in (21) the memory usage reduces to $O(n^2)$ (for inverting $(\boldsymbol{I} + \boldsymbol{\Gamma X \Lambda}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma})^{-1}$). Hence, the total amount of memory required for the second step of Algorithm 1 is $O(\min(np+n^2, np+p^2))$: np for storing $\boldsymbol{X}$ and n^2 or p^2 for calculating $\left(\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}$. The last step of ALO requires negligible amount of memory. Hence, the total amount of memory ALO requires especially when $n < p$, is $O(np + n^2)$, which is the same as $O(np)$. Note that the amount of memory required by LO is also $O(np)$, since it requires to store $\boldsymbol{X}$.

The situation is even more favorable for ALO in Algorithm 2; all the memory requirements are the same as before, except that the amount of memory that is required for the calculation and storing of $\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + 2\lambda_1 \boldsymbol{I}\right)^{-1}$ is $O(s^2)$.

4 Theoretical Results in High Dimensions

4.1 Assumptions

In this section, we introduce assumptions later used in our theoretical results. The assumptions and theoretical results that follow are presented for finite sample sizes. However, the final conclusions of this paper

[3]It is known that after a finite number of iterations the estimates of proximal gradient descent becomes sparse, and hence the iterations require less operations. Hence, in practice the sparsity can reduce the computational complexity of calculating LO even though this gain is not captured in the worst case analysis of this section.

are focused on the high-dimensional asymptotic setting in which $n, p \to \infty$ and $n/p \to \delta_o$, where δ_o is a finite number bounded away from zero. Hence, if we write a constant as $c(n)$, it may be the case that the constant depends on both n and p, but since $p \sim n/\delta_o$, we drop the dependance on p. We use this simplification for the sake of brevity and clarity of presentation. Since our major theorem involves finite sample sizes it is straightforward to go beyond this high-dimensional asymptotic setting and obtain more general results useful for other asymptotic settings.

Assumption 5. *The rows of $\boldsymbol{X} \in \mathrm{R}^{n\times p}$ are independent zero mean Gaussian vectors with covariance $\boldsymbol{\Sigma}$. Let $\rho_{\max}$ denote the largest eigenvalue of $\boldsymbol{\Sigma}$.*

As we mentioned earlier, in our asymptotic setting, we assume that $n/p \to \delta_o$ for some δ_o bounded away from zero. Furthermore, we assume that the rows of $\boldsymbol{X}$ are scaled in a way that $\rho_{max} = \Theta(\frac{1}{n})$ to ensure that $\boldsymbol{x}_i^\top \boldsymbol{\beta} = O_p(1)$ and $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = O(1)$, assuming that each β_i is $O(1)$. Under this scaling the signal-to-noise ratio in each observation remains fixed as n, p grow.[4] For more information on this asymptotic setting and scaling, the reader may refer to [El Karoui, 2018, Donoho and Montanari, 2016, Donoho et al., 2011, Bayati and Montanari, 2012, Weng et al., 2018, Dobriban and Wager, 2018].

Assumption 6. *There exist finite constants $c_1(n)$ and $c_2(n)$, and $q_n \to 0$ all functions of n, such that with probability at least $1 - q_n$ for all $i = 1, \ldots, n$*

$$
\begin{aligned}
c_1(n) &> \|\dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})\|_\infty, && (22)\\
c_2(n) &> \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2}. && (23)\\
c_2(n) &> \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{r}}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2}. && (24)
\end{aligned}
$$

In what follows, for various regularizers and regression methods, by explicitly quantifying constants $c_1(n)$ and $c_2(n)$, we discuss conditions (22), (23), and (24) in Assumption 6. We consider the ridge regularizer in Lemma 1 and the smoothed-ℓ_1 (and elastic-net) regularizer in Lemma 2. Concerning various regression methods, we consider logistic (Lemma 3), robust regression (Lemma 4), least-squares (Lemmas 6 and 7), and Poisson (Lemmas 8 and 9) regression. The results below show that under mild assumptions, for the cases mentioned above, $c_1(n)$ and $c_2(n)$ are polynomial functions of $\log n$, a result that plays a key role in our main theoretical result presented in Section 4.2.

Lemma 1. *For the ridge regularizer $r(z) = z^2$, we have*

$$
\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{r}}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} = 0.
$$

[4]Furthermore, under this scaling of the optimal value of λ will be $O_p(1)$ [Mousavi et al., 2018].

Due to simplicity we skip the proof. As mentioned in Section 2.2, a standard smooth approximation of the ℓ_1-norm is given by

$$r^{\alpha}(z) = \sum_{i=1}^{p} \frac{1}{\alpha}\Big(\log(1+e^{\alpha z}) + \log(1+e^{-\alpha z})\Big).$$

Lemma 2. *For the smoothed-ℓ_1 regularizer we have*

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{r}}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq 4\alpha^2.$$

We present the proof of this result in Section A.5.6. Note that as a consequence of Lemma 2, for the smoothed elastic-net regularizer, defined as $r(z) = \gamma z^2 + (1-\gamma) r^{\alpha}(z)$ for $\gamma \in [0,1]$, we have

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{r}}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq 4(1-\gamma)\alpha^2.$$

Lemma 3. *In the generalized linear model family, for the negative logistic regression log-likelihood $\ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}) = -y\boldsymbol{x}^\top\boldsymbol{\beta} + \log(1+e^{\boldsymbol{x}^\top\boldsymbol{\beta}})$, where $y \in \{0,1\}$, we have*

$$\begin{aligned} \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} &\leq \sqrt{\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}, \\ \|\dot{\ell}(\boldsymbol{\beta})\|_\infty &\leq 1. \end{aligned}$$

We present the proof of this result in Section A.5.1. Our next example is about a smooth approximation of the Huber loss used in robust estimation, known as the pseudo-Huber loss:

$$f_H(z) = \gamma^2 \left(\sqrt{1 + (\frac{z}{\gamma})^2} - 1 \right),$$

where $\gamma > 0$ is a fixed number.

Lemma 4. *For the pseudo-Huber loss function $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = f_H(y - \boldsymbol{x}^\top\boldsymbol{\beta})$, we have*

$$\begin{aligned} \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} &\leq \frac{3}{\gamma}\sqrt{\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}, \\ \|\dot{\ell}(\boldsymbol{\beta})\|_\infty &\leq \gamma. \end{aligned}$$

The proof of this result is presented in Section A.5.4.

Lemma 5. *If Assumption 5 holds with $\rho_{\max} = c/n$, and $\delta_0 = n/p$, then*

$$\Pr\left(\sigma_{max}(\boldsymbol{X}^\top\boldsymbol{X}) \geq c\Big(1 + 3\frac{1}{\sqrt{\delta_0}}\Big)^2 \right) \leq \mathrm{e}^{-p}.$$

The proof of this Lemma presented in Section A. Putting together Lemmas 1, 2, 3, 4 and 5, we conclude that for ridge/smoothed-ℓ_1 regularized robust/logistic regression we have $c_1(n) = O(1)$ and $c_2(n) = O(1)$.

Lemma 6. *For the loss function $\ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2$, we have*

$$\begin{aligned} \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} &= 0, \\ \|\dot{\ell}(\hat{\boldsymbol{\beta}})\|_\infty &\leq \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_\infty. \end{aligned}$$

We skip the proof of this lemma because it is straightforward.

Lemma 7. *Assume $\boldsymbol{y} \sim N(\boldsymbol{X\beta}^*, \sigma_\epsilon^2 \boldsymbol{I})$, and $\ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2$. Let Assumption 5 hold with $\rho_{\max} = c/n$. Finally, let $n/p = \delta_0$ and $\frac{1}{n}\|\boldsymbol{\beta}^*\|_2^2 = \tilde{c}$. If $r(\beta) = \gamma\beta^2 + (1-\gamma)r^\alpha(\beta)$, and $0 < \gamma < 1$, then*

$$\Pr\left(\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_\infty > \tilde{\zeta}\sqrt{\log n}\right) \leq \frac{10}{n} + 2n\mathrm{e}^{-n+1} + n\mathrm{e}^{-p},$$

where $\tilde{\zeta}$ is a constant that only depends on $\sigma_\epsilon, \alpha, c, \tilde{c}, \lambda, \delta_0$ and γ (and is free of n and p).

We present the proof of this result in Section A.5.5. Putting together Lemmas 1, 2, 6, and 7, we conclude that for smoothed elastic-net regularized least squares regression we have $c_1(n) = O(\sqrt{\log n})$ and $c_2(n) = O(1)$.

Lemma 8. *In the generalized linear model family, for the negative Poisson regression log-likelihood $\ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}) = -f(\boldsymbol{x}^\top \boldsymbol{\beta}) + y \log f(\boldsymbol{x}^\top \boldsymbol{\beta}) - \log y!$ with the conditional mean $\mathrm{E}[y|\boldsymbol{x}, \boldsymbol{\beta}] = f(\boldsymbol{x}^\top \boldsymbol{\beta})$ where $f(z) = \log(1 + e^z)$ (known as a soft-rectifying nonlinearity[5]), we have*

$$\begin{aligned} \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} &\leq (1 + 6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})} \\ \|\dot{\ell}(\boldsymbol{\beta})\|_\infty &\leq 1 + \|\boldsymbol{y}\|_\infty. \end{aligned}$$

We present the proof of this result in Section A.5.2.

Lemma 9. *Assume that $y_i \sim \mathrm{Poisson}\left(f(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)\right)$ where $f(z) = \log(1 + e^z)$. Let Assumption 5 hold with $\rho_{\max} = c/n$. Finally, let $n/p = \delta_0$ and $\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma\beta}^* = \tilde{c}$. Then, for large enough n, we have*

$$\begin{aligned} \Pr\left((1 + 6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})} \geq \zeta_1 \log^{3/2} n\right) &\leq n^{1-\log\log n} + \frac{2}{n} + \mathrm{e}^{-n\log(\frac{1}{\mathbb{P}(Z\leq 1)})} + \mathrm{e}^{-p} \\ \Pr\left(\|\boldsymbol{y}\|_\infty \geq 6\sqrt{\tilde{c}}\log^{3/2} n\right) &\leq n^{1-\log\log n} + \frac{2}{n} + \mathrm{e}^{-n\log(\frac{1}{\mathbb{P}(Z\leq 1)})} \end{aligned}$$

[5]The "soft-rectifying" nonlinearity $f(z) = \log(1 + e^z)$ behaves linearly for large z, and decays exponentially on its left tail. Owing to the convexity and log-concavity of this nonlinearity the log-likelihood is concave [Paninski, 2004], leading to a convex estimation problem. Since the actual nonlinearity of neural systems is often sub-exponential, the "soft-rectifying" nonlinearity is popular in analyzing neural data (see [Pillow, 2007, Park et al., 2014, Alison and Pillow, 2017, Zolrowski and Pillow, 2018] and references therein).

where $Z \sim N(0, \tilde{c})$ and ζ_1 is a constant that only depends on $c, \tilde{c}$, and δ_0 (and is free of n and p).

The proof of this result is presented in Section A.5.3. Putting together Lemmas 1, 2, 8, and 9, we conclude that for ridge/smoothed elastic-net regularized Poisson regression we have $c_2(n) = O(\log^{3/2}(n))$ and $c_1(n) = O(\log^{3/2}(n))$.

In summary, in the high-dimensional asymptotic setting, for all the examples we have discussed so far, $c_1(n) = O(\log^{3/2}(n))$ and $c_2(n) = O(\log^{3/2}(n))$. Hence, in the results that we will see in the next section we assume that both $c_1(n)$ and $c_2(n)$ are polynomial functions of $\log(n)$. Finally, we assume that the curvatures of the optimization problems involved in (1) and (4) have a lower bound:

Assumption 7. *There exists a constant $\nu > 0$, and a sequence $\tilde{q}_n \to 0$ such that for all $i = 1, \ldots, n$*

$$\inf_{t \in [0,1]} \sigma_{\min}\left(\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{/i})] + \boldsymbol{X}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{/i})]\boldsymbol{X}_{/i}\right) \geq \nu \tag{25}$$

with probability at least $1 - \tilde{q}_n$. Here, $\sigma_{\min}(\boldsymbol{A})$ stands for the smallest singular value of $\boldsymbol{A}$.

Assumption 7 means that optimization problems (1) and (4) are strongly convex, and strong convexity is a standard assumption made in the analysis of high dimensional problems, eg. [Van de Geer, 2008,Negahban et al., 2012]. Moreover, if $r(\beta) = \gamma\beta^2 + (1-\gamma)r^{\alpha}(\beta)$, and $0 < \gamma < 1$, then $\nu = 2\gamma$.

Before we mention our main result, we should also mention that Assumptions 7, 5, and 6 can be weakened at the expense of making our final result look more complicated. For instance, the Gaussianity of the rows of $\boldsymbol{X}$ can be replaced with the subgaussianity assumption with minor changes in our final result. We expect our results (or slightly weaker ones) to hold even when the rows of $\boldsymbol{X}$ have heavier tails. However, for the sake of brevity we do not study such matrices in the current paper. Furthermore, the smoothness of the second derivatives of the loss function and the regularizer that is assumed in (23) and (24) can be weakened at the expense of slower convergence in Theorem 3. We will clarify this point in a footnote after (142) in the proof.

4.2 Main theoretical result

Now based on these results we bound the difference $|\text{ALO} - \text{LO}|$. The proof is given in Section A.6.

Theorem 3. *Let $n/p = \delta_0$ and Assumption 5 hold with $\rho_{\max} = c/p$. Moreover, suppose that Assumptions 6, and 7 are satisfied, and that n is large enough such that $q_n + \tilde{q}_n < 0.5$. Then with probability at least $1 - 4ne^{-p} - \frac{8n}{p^3} - \frac{8n}{(n-1)^3} - q_n - \tilde{q}_n$ the following bound is valid:*

$$\max_{1 \leq i \leq n} \left| \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1 - H_{ii}}\right) \right| \leq \frac{C_o}{\sqrt{p}}, \tag{26}$$

where

$$C_o \triangleq (\frac{72c^{3/2}}{\nu^3})\left(1+\sqrt{\delta_0}(\sqrt{\delta_0}+3)^2\frac{c\log n}{\log p}\right)\left(c_1^2(n)c_2(n)+c_1^3(n)c_2^2(n)\frac{5(c^{1/2}+c^{3/2}(\sqrt{\delta_0}+3)^2)}{\nu^2}\right). \quad (27)$$

Recall that in Section 4.1 we proved that for many regularized regression problems in the generalized linear family both $c_1(n) = O(\text{PolyLog}(n))$ and $c_2(n) = O(\text{PolyLog}(n))$, where the notation $\text{PolyLog}(n)$ denotes a polynomial in $\log(n)$. These examples included ridge and smoothed-ℓ_1 (and elastic-net) regularizers and logistic, robust, least-squares, and Poisson regression. More specifically, the maximum degree we observed for the logarithm was 3/2, which happened for the Poisson regression. Furthermore, as mentioned in the last section, in the high-dimensional asymptotic setting in which $n, p \to \infty$ and $n/p \to \delta_o$, where δ_o is a finite number bounded away from zero, to keep the signal-to-noise ratio fixed in each observation (as p and n grow), we considered the scaling that $n\rho_{\max} = O(1)$. Combining these, it is straightforward to see that $C_0(n) = O(c_1^3(n)c_2^2(n)) = O(\text{PolyLog}(n))$. Therefore, the difference $\max_{1\le i\le n}\left|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}} - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right| = O_p(\frac{\text{PolyLog}(n)}{\sqrt{n}})$. Theorem 3 proves the accuracy of the approximation of the leave-one-out estimate of the regression coefficients. As a simple corollary of this result we can also prove the accuracy of our approximation of LO.

Corollary 1. *Suppose that all the assumptions used in Theorem 3 hold. Moreover, suppose that*

$$\max_{i=1,2,\dots,n}\sup_{|b_i|<\frac{C_o}{\sqrt{p}}}\left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}_{/i}+b_i\right)\right| \le c_3(n)$$

with probability r_n. Then, with probability at least $1 - 4ne^{-p} - \frac{8n}{p^3} - \frac{8n}{(n-1)^3} - q_n - \tilde{q}_n - r_n$

$$|\text{ALO} - \text{LO}| \le \frac{c_3(n)C_o}{\sqrt{p}}, \quad (28)$$

where C_o is the constant defined in Theorem 3.

The proof of this result can be found in Section A.8. As we discussed before, in all the examples we have seen so far $\frac{C_o}{\sqrt{p}}$ is $O(\frac{\text{PolyLog}(n)}{\sqrt{n}})$. Hence, to obtain the convergence rate of ALO to LO we only need to find an upper bound for $c_3(n)$. Note that usually the loss function ℓ that is used in the optimization problem is also used as the function ϕ to measure the prediction error. Hence, assuming $\phi(\cdot,\cdot) = \ell(\cdot,\cdot)$, we study the value of $c_3(n)$ for the examples we discussed in Section 4.1.

1. If ϕ is the loss function of Lemma 3, then $\left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top\beta\right)\right| \le 2$, leading to $c_3(n) = 2$.
2. If ϕ is the loss function of Lemma 8, then $\left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top\beta\right)\right| \le 1 + \|\boldsymbol{y}\|_\infty$. Furthermore, we proved in Lemma 9, that under the data generating mechanism described there, with high probability $\|\boldsymbol{y}\|_\infty < 6\sqrt{\tilde{c}\log^3(n)}$, leading to $c_3(n) = 1 + 6\sqrt{\tilde{c}\log^3(n)}$.

3. For the pseudo-Huber loss described in Lemma 4, we have $\left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top \beta\right)\right| \leq \gamma$, leading to $c_3(n) = \gamma$.

4. For the square loss $\left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i} + b_i\right)\right| \leq |y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i}| + |b_i| \leq |y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i}| + \frac{C_o}{\sqrt{p}}$. Hence, in order to obtain a proper upper bound we require more information about the estimate $\hat{\boldsymbol{\beta}}_{/i}$. Suppose that our estimates are obtained from the optimization problem we discussed in Lemma 7. Then, based on (94) and (97) in the proof of Lemma 7 in Appendix A.5.5

$$\max_i |y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i}| \leq \max_i |y_i| + \max_i |\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i}| \leq 2\sqrt{(c\tilde{c} + \sigma_\epsilon^2)\log n} + \sqrt{\frac{10c(c\tilde{c} + \sigma_\epsilon^2)\log n}{\lambda\gamma}}.$$

with probability at most $\frac{4}{n} + n\mathrm{e}^{-n+1}$, leading to $c_3(n) = 2\sqrt{(c\tilde{c} + \sigma_\epsilon^2)\log n} + \sqrt{\frac{20c(c\tilde{c}+\sigma_\epsilon^2)\log n}{\lambda\gamma}} + \frac{C_o}{\sqrt{p}}$.

In summary, in the high-dimensional asymptotic setting, for regularized regression methods introduced in Section 4.1, such as least-squares, logistic, Poisson and robust regression, with $r(\beta) = \gamma\beta^2 + (1-\gamma)r^\alpha(\beta)$, and $0 < \gamma < 1$, and assuming $\phi(\cdot,\cdot) = \ell(\cdot,\cdot)$, we have $c_3(n) = O(\mathrm{PolyLog(n)})$, leading to $|\mathrm{ALO} - \mathrm{LO}| = O_p(\frac{\mathrm{PolyLog(n)}}{\sqrt{n}})$. In short, these examples show that ALO offers a consistent estimate of LO.

Finally, note that in the p fixed, $n \to \infty$ regime, Theorem 3 fails to yield $|\mathrm{ALO} - \mathrm{LO}| = o_p(1)$. This is just an artifact of our proof. In Theorem 6, presented in Section A.9 we prove that under mild regularity conditions, error between ALO and LO is $o_p(1/n)$ when $n \to \infty$ and p is fixed. For the sake of brevity details are presented in Section A.9.

5 Numerical Experiments

5.1 Summary

To illustrate the accuracy and computational efficiency of ALO we apply it to synthetic and real data. We generate synthetic data, and compare ALO and LO for elastic-net linear regression in Section 5.2.1, LASSO logistic regression in Section 5.2.2, and elastic-net Poisson regression in Section 5.2.3. We should emphasize that our simulations are performed on a single personal computer, and we have not considered the impact of parallelization on the performance of ALO and LO. In other words, the simulation results reported for LO are based on its sequential implementation on a single personal computer. For real data, we apply LASSO, elastic-net and ridge logistic regression to sonar returns from two undersea targets in Section 5.3.1, and we apply LASSO Poisson regression to real recordings from spatially sensitive neurons in Section 5.3.2. Our synthetic and real data examples cover various data shapes where $n > p$, $n = p$ and $n < p$.

Figures 4, 5, 6, 7, and the middle-lower panel of Figure 10 reveal that ALO offers a reasonably accurate estimate of LO for a large range of λ. These figures show that ALO deteriorates for extremely small values of λ, specially when $p > n$. This is not a serious issue because the λs minimizing LO and ALO tend to be far from those small values.

The real data example in Section 5.3.1, illustrating ALO and LO in Figure 7, is about classifying sonar returns from two undersea targets using penalized logistic regression. The neuroscience example in Section 5.3.2 is about estimating an inhomogeneous spatial point process using an over-complete basis from a sparsely sampled two-dimensional space. Given the spatial nature of the problem, the design matrix $\boldsymbol{X}$ is very sparse, which fails to satisfy the dense Gaussian design assumption we made in Theorem 3. Nevertheless, the lower middle panel of Figure 10 illustrates the excellent performance of ALO in approximating LO in an example where $p = 10000$ and $n = 3133$.

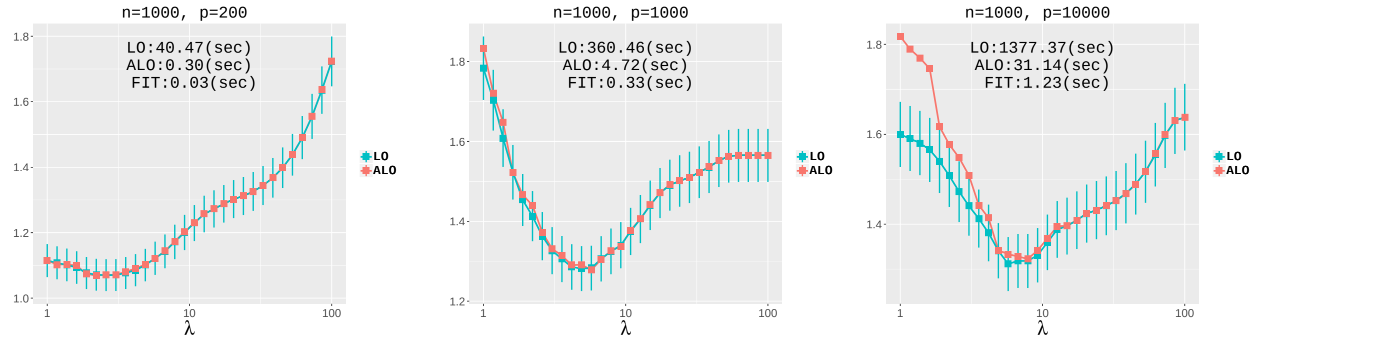
Figure 2 compares the computational complexity(time) of a single fit, ALO and LO, as we increase p while we keep the ratio $\frac{n}{p}$ fixed. We consider various data shapes, models, and penalties. Figure 2a shows time versus p for elastic-net linear regression when $\frac{n}{p} = 5$. Figure 2b shows time versus p for LASSO logistic regression when $\frac{n}{p} = 1$. Figure 2c shows time versus p for elastic-net Poisson regression when $\frac{n}{p} = \frac{1}{10}$. Finally, the middle-lower panel of Figure 10 shows that for the neuroscience example ALO takes 7 seconds in comparison to the 60428 seconds required by LO. All these numerical experiments illustrate the significant computational saving offered by ALO. As it pertains to the reported run times, all fittings in this paper were performed using a 3.1 GHz Intel Core i7 MacBook Pro with 16 GB of memory. All the codes for the figures presented in this paper are available here `https://github.com/RahnamaRad/ALO`.

5.2 Simulations

In all the examples in this section (5.2.1, 5.2.2, 5.2.3 and 5.2.4), we let the true unknown parameter vector $\boldsymbol{\beta}^* \in \mathrm{R}^p$ to have $k = n/10$ non-zero coefficients. The k non-zero coefficients are randomly selected, and their values are independently drawn from a zero mean unit variance Laplace distribution. The rows $\boldsymbol{x_1}^\top, \cdots, \boldsymbol{x_n}^\top$ of the design matrix $\boldsymbol{X}$ are independently drawn from $\mathrm{N}(0, \boldsymbol{\Sigma})$. We consider two correlation structures: 1) *Spiked*: $cor(X_{ij}, X_{ij'}) = 0.5$, and 2) *Toeplitz*: $cor(X_{ij}, X_{ij'}) = 0.9^{|j'-j|}$. $\boldsymbol{\Sigma}$ is scaled such that the signal variance $\mathrm{var}(\boldsymbol{x_i}^\top \boldsymbol{\beta}^*) = 1$ regardless of the problem dimension. In this section, all the fittings and calculations of LO (and the one standard error interval of LO) were computed using the `glmnet` package in R [Friedman et al., 2010], and ALO was computed using the `alocv` package in R [He et al., 2018].

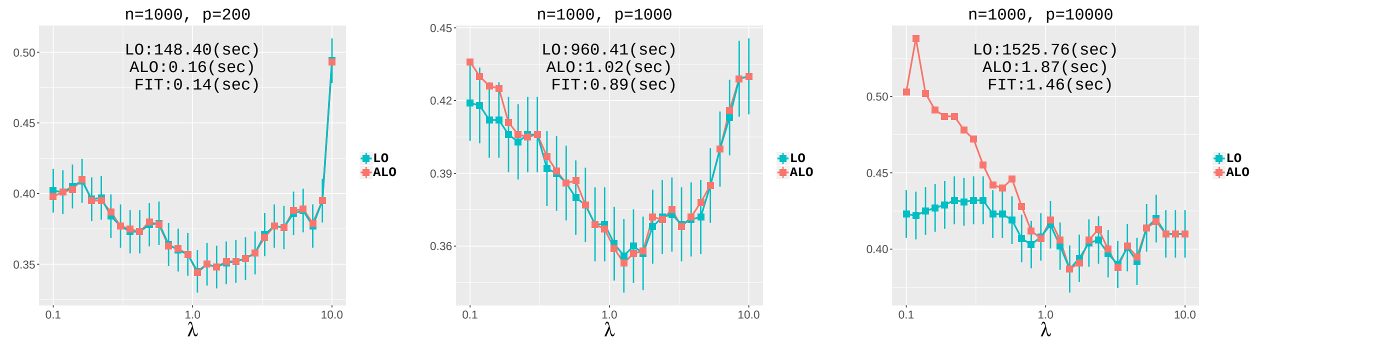
5.2.1 Linear regression with elastic-net penalty

We set $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top\boldsymbol{\beta})^2$, $r(\boldsymbol{\beta}) = \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1$ and $\alpha = 0.5$. We let the rows $\boldsymbol{x_1}^\top, \cdots, \boldsymbol{x_n}^\top$ of $\boldsymbol{X}$ to have a *Spiked* covariance and to generate data, we sample $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{I})$. Moreover, $\phi(y, \boldsymbol{x}^\top\boldsymbol{\beta}) = \left(y - \boldsymbol{x}^\top\boldsymbol{\beta}\right)^2$ so that ALO $= \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}}{1-H_{ii}}\right)^2$ with $\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top\boldsymbol{X}_S + \lambda(1-\alpha)\boldsymbol{I}\right)^{-1}\boldsymbol{X}_{\boldsymbol{S}}^\top$. For various data shapes, that is $\frac{n}{p} \in \{5,\ 1,\ \frac{1}{10}\}$, we depict results in Figure 4 where reported times refer to the required time to fit the model, compute ALO and LO for a sequence of 30 logarithmically spaced tuning parameters from 1 to 100.



(a) $n > p$ (b) $n = p$ (c) $n < p$

Figure 4: The ALO and LO mean square error for elastic-net linear regression. The red error bars identify the one standard error interval of LO.



(a) $n > p$ (b) $n = p$ (c) $n < p$

Figure 5: The ALO and LO misclassification errors (as a function of λ) for LASSO logistic regression. The red error bars identify the one standard error interval of LO.

5.2.2 Logistic regression with LASSO penalty

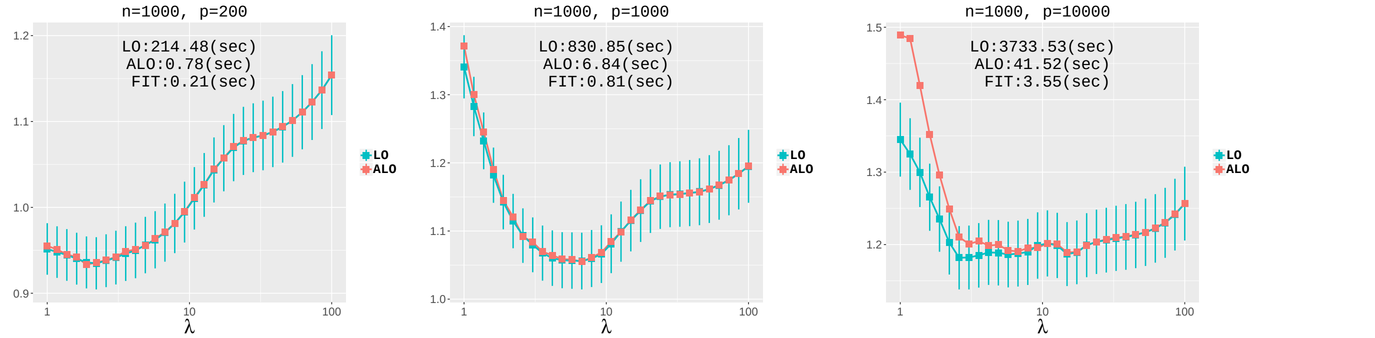
We set $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = -y\boldsymbol{x}^\top\boldsymbol{\beta} + \log(1 + e^{\boldsymbol{x}^\top\boldsymbol{\beta}})$ (the negative logistic log-likelihood) and $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. We let the rows $\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_n^\top$ of $\boldsymbol{X}$ to have a *Toeplitz* covariance and to generate data, we sample $y_i \sim$ *Binomial* $\left(\frac{e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}^*}}{1+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}^*}}\right)$. We take the misclassification rate as our measure of error, and $1_{\{\boldsymbol{x}^\top\boldsymbol{\beta}>0\}}$ as prediction, where $1_{\{\cdot\}}$ is the indicator function, so that

$$\text{ALO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - 1_{\{\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} + \frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\frac{H_{ii}}{1-H_{ii}}>0\}}\right|$$

where $\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$, $\dot{\ell}_i(\hat{\boldsymbol{\beta}}) = \left(1 + e^{-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}}\right)^{-1} - y_i$ and $\ddot{\ell}_i(\hat{\boldsymbol{\beta}}) = e^{\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}}\left(1 + e^{\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}}\right)^{-2}$. For various data shapes, that is $\frac{n}{p} \in \{5,\ 1,\ \frac{1}{10}\}$, we depict results in Figure 5 where reported times refer to the required time to fit the model, compute ALO and LO for a sequence of 30 logarithmically spaced tuning parameters from 0.1 to 10.

5.2.3 Poisson regression with elastic-net penalty

We set $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = e^{y\boldsymbol{x}^\top\boldsymbol{\beta}} - y\boldsymbol{x}^\top\boldsymbol{\beta}$ (the negative Poisson log-likelihood), $r(\boldsymbol{\beta}) = \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1$ and $\alpha = 0.5$. We let the rows $\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_n^\top$ of $\boldsymbol{X}$ to have a *Spiked* covariance and to generate data, we sample



(a) $n > p$ (b) $n = p$ (c) $n < p$

Figure 6: The ALO and LO mean absolute errors (as a function of λ) for elastic-net Poisson regression. The red error bars identify the one standard error interval of LO.

$y_i \sim Poisson\left(e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}^*}\right)$. We use the mean absolute error as our measure of error, and $e^{\boldsymbol{x}^\top \boldsymbol{\beta}}$ as prediction, so that

$$\text{ALO} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\left|y_i - e^{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\frac{H_{ii}}{1-H_{ii}}}\right|$$

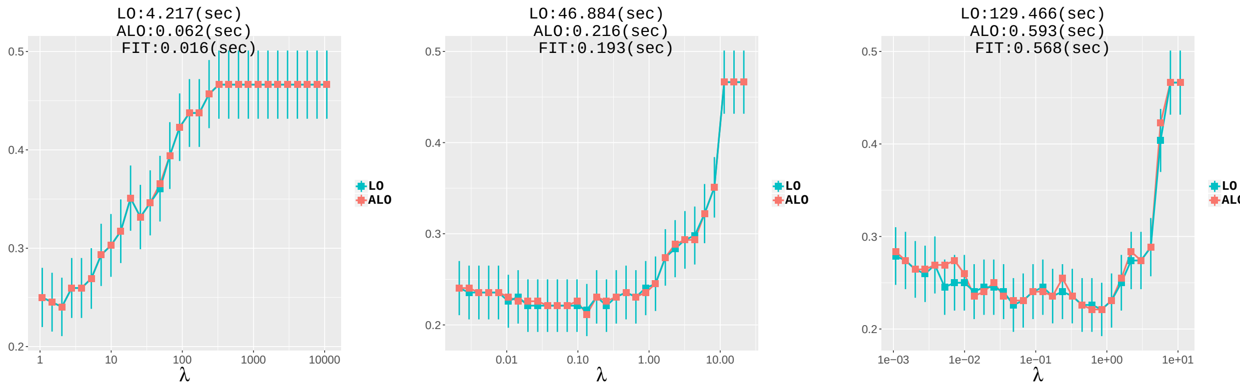
where $\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + \lambda(1-\alpha)\boldsymbol{I}\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$, $\dot{\ell}_i(\hat{\boldsymbol{\beta}}) = e^{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}} - y_i$, and $\ddot{\ell}_i(\hat{\boldsymbol{\beta}}) = e^{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}$. For various data shapes, that is $\frac{n}{p} \in \{5,\ 1,\ \frac{1}{10}\}$, we depict results in Figure 6 where reported times refer to the required time to fit the model, compute ALO and LO for a sequence of 30 logarithmically spaced tuning parameters from 1 to 100.

5.2.4 Timing simulations

To compare the timing of ALO with that of LO, we consider the following scenarios:

- Elastic-net linear regression, with rows of the design matrix having a *Spiked* covariance, data generated as described in Sections 5.2 and 5.2.1, and considered for a sequence of 10 logarithmically spaced tuning parameters from 1 to 100. We let $\frac{n}{p} = 5$.
- LASSO logistic regression, with rows of the design matrix having a *Toeplitz* covariance, data generated as described in Sections 5.2 and 5.2.2, and considered for a sequence of 10 logarithmically spaced tuning parameters from 0.1 to 10. We let $\frac{n}{p} = 1$.
- Elastic-net Poisson regression, with rows of the design matrix having a *Spiked* covariance, data generated as described in Sections 5.2 and 5.2.3, and considered for a sequence of 10 logarithmically spaced tuning parameters from 1 to 100. We let $\frac{n}{p} = \frac{1}{10}$.

The timings of a single fit, ALO and LO versus model complexity p are illustrated in Figure 2. The reported timings are obtained by recording the time required to find a single fit and LO using the `glmnet` package in R [Friedman et al., 2010], and to find ALO using the `alocv` package in R [He et al., 2018], all along the tuning parameters above. This process is repeated 5 times to obtain the average timing.



(a) Ridge (b) Elastic-net (c) LASSO

Figure 7: The ALO and LO deviances (as a function of λ) for penalized logistic regression applied to the sonar data (Section 5.3.1) where $n = 208$ and $p = 60$. The red error bars identify the one standard error interval of LO.

5.3 Real Data

5.3.1 Sonar data

Here we use ridge, elastic-net and LASSO logistic regression to classify sonar returns collected from a metal cylinder and a cylindrically shaped rock positioned on a sandy ocean floor. The data consists of a set of $n = 208$ returns, 111 cylinder returns and 97 rock returns, and $p = 60$ spectral features extracted from the returning signals [Gorman and Sejnowski, 1988]. We use the misclassification rate as our measure of error. Numerical results comparing ALO and LO for ridge, elastic-net and LASSO logistic regression are depicted in Figure 7. The single fit and LO (and the one standard error interval of LO) were computed using the `glmnet` package in R [Friedman et al., 2010], and ALO was computed using the `alocv` package in R [He et al., 2018]. The values of the tuning parameters are a sequence of 30 logarithmically spaced tuning parameters between two value automatically selected by the `glmnet` package.

5.3.2 Spatial point process smoothing of grid cells: a neuroscience application

In this section, we compare ALO with LO on a real dataset. This dataset includes electrical recordings of single neurons in the entorhinal cortex, an area in the brain found to be particularly responsible for the navigation and perception of space in mammals [Moser et al., 2008]. The entorhinal cortex is also one of the areas pathologically affected in the early stages of Alzheimer's disease, causing symptoms of spatial disorientation [Khan et al., 2014]. Moreover, the entorhinal cortex provides input to another area, the Hippocampus, which is involved in the cognition of space and the formation of episodic memory [Buzsaki and Moser, 2013].

Electrical recordings of single neurons in the medial domain of the entorhinal cortex (MEC) of freely moving rodents have revealed spatially modulated neurons, called grid cells, firing action potentials only

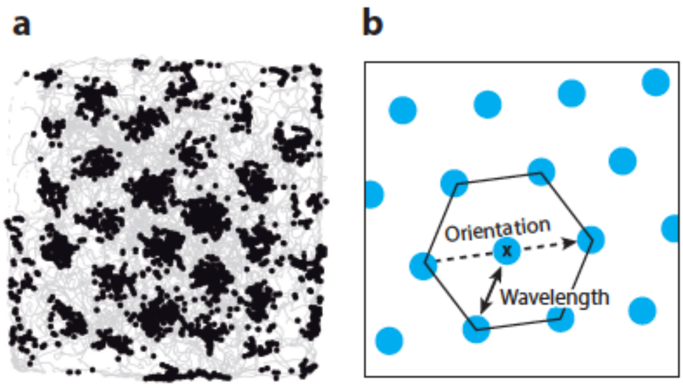


Figure 8: Left: Spike locations (black) are superimposed on the animal's trajectory(grey). Firing fields are areas covered by a cluster of action potentials. Right: The firing fields of a grid cell form a periodic triangular matrix tiling the entire environment available to the animal. Figure is adapted from [Moser et al., 2014].

around the vertices of two dimensional hexagonal lattices covering the environment in which the animal navigates. The hexagonal firing pattern of a single grid cell is illustrated in the left panel of Figure 8. These grid cells can be categorized according to the orientation of their triangular grid, the wavelength (distance between the vertices), and the phase (shift of the whole lattice). See the right panel of Figure 8 for an illustration of the orientation and wavelength of a single grid cell.

The data we analyze here consists of extra cellular recordings of several grid cells, and the simultaneously recorded location of the rat within a 300cm $\times$ 300cm box for roughly 20 minutes[6]. Since the number of spikes fired by a grid cell depends mainly on the location of the animal, regardless of the animal's speed and running direction [Hafting et al., 2005], it is reasonable to summarize this spatial dependency in terms of a rate map $\eta(\boldsymbol{r})$, where $\eta(\boldsymbol{r})dt$ is the expected number of spikes emitted by the grid cell in a fixed time interval dt, given that the animal is located at position $\boldsymbol{r}$ during this time interval [Rahnama Rad and Paninski, 2010, Pnevmatikakis et al., 2014, Dunn et al., 2015]. In other words, if the rat passes the same location again, we again expect the grid cell to fire at more or less the same rate[7], specifically according to a Poisson distribution with mean $\eta(\boldsymbol{r})dt$. For each grid cell, the estimation of the rate map $\eta(\boldsymbol{r})$ is a first step toward understanding the cortical circuitry underlying spatial cognition [Rowland et al., 2016]. Consequently, the estimation of firing fields without contamination from measurement noise or bias from overs-smoothing will help to clarify important questions about neuronal algorithms underlying navigation in real and mental spaces [Buzsaki and Moser, 2013].

To be concrete, we discretize the two dimensional space into an $m \times m$ grid, and discretize time into bins with width dt. In this example, dt is 0.4 seconds and m is 50. The experiment is 1252.9 seconds long, and therefore we have $\lceil \frac{1252.9}{0.4} \rceil = 3133$ time bins. In other words, $n = 3133$. We use $y_i \in \{0, 1, 2, 3, \cdots\}$ to denote the number of action potentials observed in time interval $[(i-1)dt, idt)$, where $i = 1, \cdots, n$. Moreover, we

[6]The source of the data is [Stensola et al., 2012]. For a video of a single grid cell recorded in the MEC see the clip https://www.youtube.com/watch?v=i9GiLBXWAHI.

[7]It is known that these rate maps can in some cases change with time but in most cases it is reasonable to assume them to be constant. Moreover, the two dimensional surface represented by $\eta(\boldsymbol{r})$ is not the same for different grid cells.

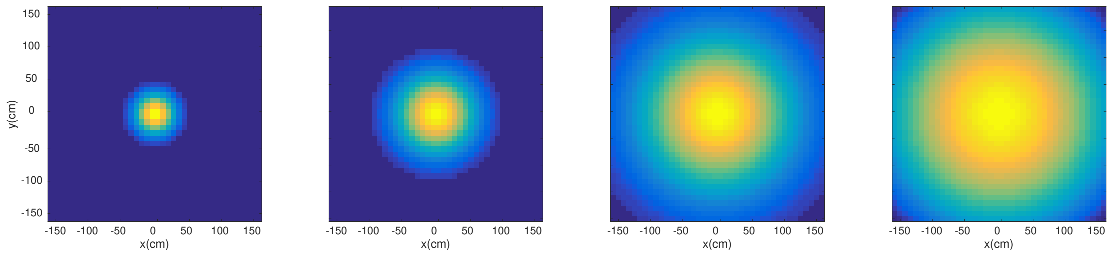


Figure 9: The four truncated Gaussian bumps

use $\boldsymbol{r_i} \in \mathrm{R}^{m^2}$ to denote a vector composed of zeros except for a single +1 at the entry corresponding to the animal's location within the $m \times m$ grid during the time interval $[(i-1)dt, idt)$. We assume a log-linear model $\log \eta(\boldsymbol{r}) = \boldsymbol{r}^\top \boldsymbol{z}$, relating the firing rate at location $\boldsymbol{r} \in \mathrm{R}^{m^2}$ to the latent vector $\boldsymbol{z}$ where the $m \times m$ latent spatial process responsible for the observed spiking activity is unraveled into $\boldsymbol{z} \in \mathrm{R}^{m^2}$. The firing rate can be written as $\eta(\boldsymbol{r_i}) = \exp\left(\boldsymbol{r_i}^\top \boldsymbol{z}\right)$. Due to this notation, $\boldsymbol{r_i}^\top \boldsymbol{z}$ is the value of $\boldsymbol{z}$ at the animal's location during the time interval $[(i-1)dt, idt)$. In this vein, the distribution of observed spiking activity can be written as

$$p(y_i|\boldsymbol{r_i}) \quad = \quad \frac{e^{-\eta(\boldsymbol{r_i})}\eta(\boldsymbol{r_i})^{y_i}}{y_i!}. \tag{29}$$

As mentioned earlier, the main goal is to estimate the two dimensional rate map $\eta(\cdot)$, and a large body of work has addressed the problem of estimating a smooth rate map from neural data [DiMatteo et al., 2001, Gao et al., 2002, Kass et al., 2005, Cunningham et al., 2008, Czanner et al., 2008, Cunningham et al., 2009, Paninski et al., 2010, Rahnama Rad and Paninski, 2010, Macke et al., 2011, Pnevmatikakis et al., 2014]. Here we employ an over-complete basis to account for the spatially localized sensitivity of grid cells. Since it is known that the rate map of any single grid cell consists of bumps of elevated firing rates, located at various points in the two dimensional space, as illustrated in the left panel of Figure 8, it is reasonable to represent $\boldsymbol{z}$ as a linear combination of $\{\boldsymbol{\psi_1}, \ldots, \boldsymbol{\psi_p}\}$, an over-complete basis in R^p [Brown et al., 2001, Pnevmatikakis et al., 2014, Dunn et al., 2015]. We compose the over-complete basis using truncated Gaussian bumps with various scales, distributed at all pixels. The four basic Gaussian bumps we use are depicted in Figure 9. Since we use four truncated Gaussian bumps for each pixel, in this example, we have a total of $p = 4m^2 = 10000$ basis functions. We employ the truncated Gaussian bumps $e^{-\frac{1}{2\sigma^2}(u_x^2+u_y^2)}1_{\left\{\exp\left(-\frac{1}{2\sigma^2}(u_x^2+u_y^2)\right)>0.05\right\}}$ where u_x and u_y are the horizontal and vertical coordinates. Define $\boldsymbol{\Psi} \in \mathrm{R}^{m^2 \times p}$ as a matrix composed of columns $\{\boldsymbol{\psi_1}, \ldots, \boldsymbol{\psi_p}\}$. Furthermore, define $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \in \mathrm{R}^p$ as $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \triangleq \boldsymbol{\Psi}^\top \boldsymbol{r_i}$, and define $\tilde{\boldsymbol{X}} \in \mathrm{R}^{n \times p}$ as a matrix composed of rows $\{\tilde{\boldsymbol{x}}_{\boldsymbol{1}}^\top, \ldots, \tilde{\boldsymbol{x}}_{\boldsymbol{n}}^\top\}$. We normalize the columns of $\tilde{\boldsymbol{X}}$, calling the resulting matrix $\boldsymbol{X}$. The columns of $\boldsymbol{X} \in \mathrm{R}^{n \times p}$ are unit normed. Formally, $\boldsymbol{X} = \tilde{\boldsymbol{X}}\boldsymbol{\Gamma}^{-1}$ where $\boldsymbol{\Gamma} \in \mathrm{R}^{p \times p}$ is a diagonal matrix filled with the column-norms of $\tilde{\boldsymbol{X}}$. We use $\{\boldsymbol{x}_{\boldsymbol{1}}^\top, \ldots, \boldsymbol{x}_{\boldsymbol{n}}^\top\}$ to refer to the rows of $\boldsymbol{X}$, yielding $\eta(\boldsymbol{r_i}) = \exp\left(\boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\beta}\right)$. Note that due to the above mentioned rescaling, we have the following relationship between the latent map $\boldsymbol{z}$ and $\boldsymbol{\beta}$: $\boldsymbol{z} = \boldsymbol{\Psi}\boldsymbol{\Gamma}\boldsymbol{\beta}$. Sparsity of $\boldsymbol{\beta}$ refers to our prior understanding that the rate map of a grid cells consists

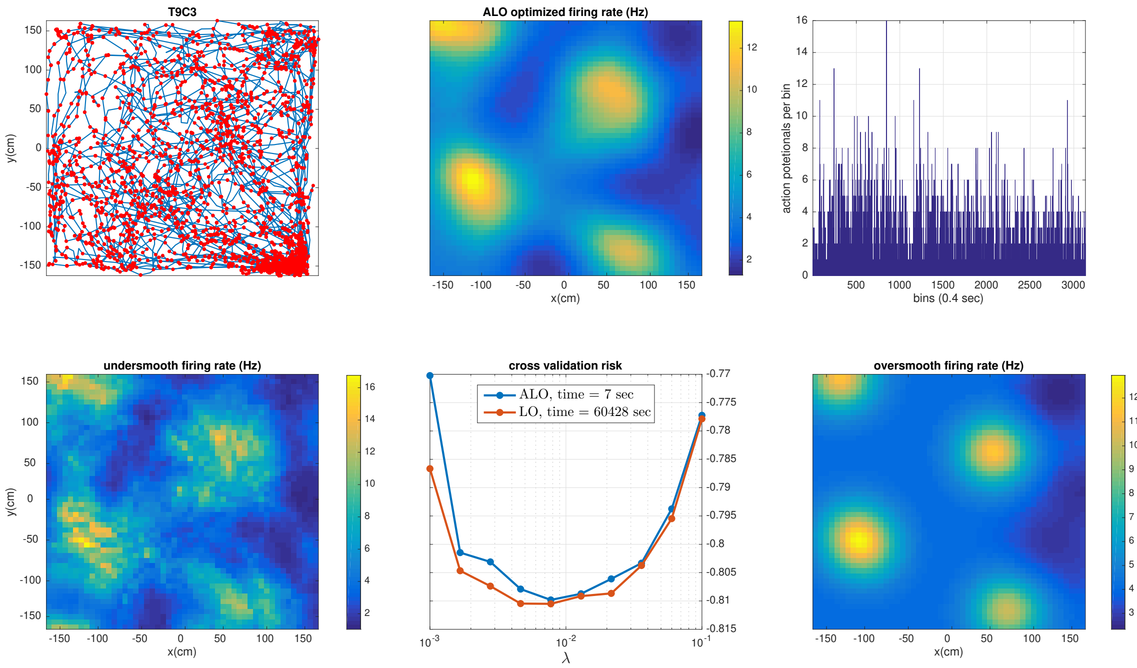


Figure 10: Top left: Spike locations (red) are superimposed on the animal's trajectory(black). Firing fields are areas covered by a cluster of action potentials. The firing fields of a grid cell form a periodic triangular matrix tiling the entire environment available to the animal. Top middle: ALO-based firing rate. Top right: LO-based firing rate. Bottom left: $\lambda = 0.001$-based firing rate. Bottom middle: ALO and LO over a wide range of λs. Bottom right: $\lambda = 0.1$-based firing rate.

of bumps of elevated firing rates, located at various points in the two dimensional space, and therefore, our estimation problem is as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &\triangleq \arg\min_{\boldsymbol{\beta}\in\mathrm{R}^p}\Big\{\sum_{i=1}^{n}[\eta(\boldsymbol{r_i}) - y_i\log\eta(\boldsymbol{r_i})] + \lambda\|\boldsymbol{\beta}\|_1\Big\},\\ &= \arg\min_{\boldsymbol{\beta}\in\mathrm{R}^p}\Big\{\sum_{i=1}^{n}\Big[\exp(\boldsymbol{x_i}^\top\boldsymbol{\beta}) - y_i\boldsymbol{x_i}^\top\boldsymbol{\beta}\Big] + \lambda\|\boldsymbol{\beta}\|_1\Big\}.\end{aligned}$$

Here we use the negative log-likelihood in equation (29) as the cost function, that is, $\phi(y, \boldsymbol{x}^\top\boldsymbol{\beta}) = y\boldsymbol{x}^\top\boldsymbol{\beta} - \exp(\boldsymbol{x}^\top\boldsymbol{\beta}) + \log y!$. We remind the reader that we will use ALO formula that was obtained in Theorem 1. Figures 10 illustrate that ALO is reasonable approximation of LO, allowing computationally efficient tuning of λ. To see the effect of λ of the rate map, we also present the maps resulting from small and large values of λ, leading to under and over smooth rate maps, respectively. As it pertains to the reported run times, all fittings in this section were performed using the `glmnet` package [Qian et al., 2013] in MATLAB.

6 Concluding Remarks

Leave-one-out cross validation (LO) is an intuitive and conceptually simple risk estimation technique. Despite its low bias in estimating the extra-sample prediction error, the high computational complexity of LO has limited its applications for high-dimensional problems. In this paper, by combining a single step of the Newton method with low-rank matrix identities, we obtained an approximate formula for LO, called ALO. We showed how ALO can be applied to popular non-differentiable regularizers, such as LASSO. With the aid of theoretical results and numerical experiments, we showed that ALO offers a computationally efficient and statistically accurate estimate of the extra-sample prediction error in high-dimensions.

Important directions for future work involve various approximations that further reduce the computational complexity. The computational bottleneck of ALO is the inversion of the large generalized hat matrix $\boldsymbol{H}$. This can make the application of ALO to ultra high dimensional problems computationally challenging. Since the diagonals of our $\boldsymbol{H}$ matrix can be represented as leverage scores of an augmented $\boldsymbol{X}$ matrix, scalable methods to approximately compute the leverage score may offer a promising avenue for future work. For example [Drineas et al., 2012] offers a randomized method to estimate the leverage scores. However, the randomized algorithm presented in [Drineas et al., 2012] applies to the $p \ll n$ case, making it challenging to apply these methods to high-dimensional settings where p is also very large. Nevertheless this is certainly a promising direction for speeding up ALO.

In another line of work, the generalized cross-validation approach [Craven and Wahba, 1979, Golub et al., 1979] approximates the diagonal elements of $\boldsymbol{H}$ with $\text{tr}(\boldsymbol{H})/n$. Computationally efficient randomized estimates of $\text{tr}(\boldsymbol{H})$ can be produced without having any explicit calculations of this matrix [Deshpande and Girard, 1991, Wahba et al., 1995, Girard, 1998, Lin et al., 2000]. The theoretical study of the additional errors introduced by these randomized approximations, and the scalable implementations of them is another promising avenue for future work.

Acknowledgements

We would like to thank the referees for carefully reading the original manuscript and raising important issues which improved the paper significantly. We are also grateful to the laboratory of Edvard Moser, at the Kavli Institute for Systems Neuroscience, at the Norwegian University of Science and Technology in Trondheim, for providing the grid cell data presented in section 5.3.2. K.R. is grateful to L. Paninski, E.A. Pnevmatikakis and Y. Roudi for fruitful conversations.

References

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[Alison and Pillow, 2017] Alison, W. and Pillow, J. (2017). Capturing the dynamical repertoire of single neurons with generalized linear models. *Neural Computation*, 29(12):3260–3289.

[Allen, 1974] Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.

[Bayati and Montanari, 2012] Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

[Bean et al., 2013] Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration ineqaulities: A nonasymptotic theory of independence*. Oxford University Press.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Oxford University Press.

[Breiman and Freedman, 1983] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of American Statistical Association*, 78(381):131–136.

[Brown et al., 2001] Brown, E., Nguyen, D., Frank, L., Wilson, M., and Solo, V. (2001). An analysis of neural receptive field plasticity by point process adaptive filtering. *PNAS*, 98:12261–12266.

[Burman, 1990] Burman, P. (1990). Estimation of generalized additive models. *Journal of Multivariate Analysis*, 32:230–255.

[Buzsaki and Moser, 2013] Buzsaki, G. and Moser, E. (2013). Memory, navigation and theta rythm in the hippocampal-entorhinal system. *Nat Neurosci*, 16(2):130–138.

[Cawley and Talbot, 2008] Cawley, G. and Talbot, N. (2008). Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71:243–264.

[Cessie and Houwelingen, 1992] Cessie, S. and Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

[Craven and Wahba, 1979] Craven, P. and Wahba, G. (1979). Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.

[Cunningham et al., 2009] Cunningham, J., Gilja, V., Ryu, S., and Shenoy, K. (2009). Methods for estimating neural firing rates, and their application to brain-machine interface. *Neural Networks*, 22(9):1235–1246.

[Cunningham et al., 2008] Cunningham, J., Yu, B., Shenoy, K., and Sahani, M. (2008). Inferring neural firing rates from spike trains using Gaussian processes. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc.

[Czanner et al., 2008] Czanner, G., Eden, U., Wirth, S., Yanike, M., Suzuki, W., and Brown, E. (2008). Analysis of between-trial and within-trial neural spiking dynamics. *Journal of Neurophysiology*, 99(5):2672–93.

[Deshpande and Girard, 1991] Deshpande, L. and Girard, D. (1991). Fast computation of cross-validated robust Splines and other non-linear smoothing Splines. *Curves and Surfaces*, pages 143–148.

[DiMatteo et al., 2001] DiMatteo, I., Genovese, C., and Kass, R. (2001). Bayesian curve fitting with free-knot splines. *Biometrika*, 88:1055–1073.

[Dobriban and Wager, 2018] Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.*, 46(1):247–279.

[Donoho et al., 2011] Donoho, D., Maleki, A., and Montanari, A. (2011). Noise sensitivity phase transition. *IEEE Trans. Inform. Theory*, 57(10).

[Donoho and Montanari, 2016] Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab Theory Relat Fields*, 166(3-4):935–969.

[Donoho et al., 2009] Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919.

[Drineas et al., 2012] Drineas, P., Magdon-Ismail, M., Mahoney, M., and Woodruff, D. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506.

[Dunn et al., 2015] Dunn, B., Morreaunet, M., and Roudi, Y. (2015). Correlations and functional connections in a population of grid cells. *PLOS Computational Biology*, 11(2):e1004052.

[Efron, 1983] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *JASA*, 78(382):316–331.

[Efron, 1986] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *JASA*, 81:461–470.

[El Karoui, 2018] El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields*, 170:95–175.

[El Karoui et al., 2013] El Karoui, N., Bean, D., Bickel, P., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *PNAS*, 110(36):14557–14562.

[Frank and Friedman, 1993] Frank, I. and Friedman, J. (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics*, 35:109–148.

[Friedman et al., 2010] Friedman, F., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[Gao et al., 2002] Gao, Y., Black, M., Bienenstock, E., Shoham, S., and Donoghue, J. (2002). Probabilistic inference of arm motion from neural activity in motor cortex. In Thomas G. Dietterich, Suzanna Becker, Z. G., editor, *Advances in Neural Information Processing Systems 14*, pages 213–220.

[Geisser, 1975] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of American Statistical Association*, 70(350):320–328.

[Girard, 1998] Girard, D. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.*, 26(1):315–334.

[Golub et al., 1979] Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

[Gorman and Sejnowski, 1988] Gorman, R. and Sejnowski, T. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89.

[Gu, 1992] Gu, C. (1992). Cross-validating non-gaussian data. *J. Comp. Graph. Stat.*, 1(2):169–179.

[Gu and Xiang, 2001] Gu, C. and Xiang, D. (2001). Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *J. Comp. Graph. Stat.*, 10(3):581–591.

[Hafting et al., 2005] Hafting, T., Fyhn, M., Molden, S., Moser, M., and Moser, E. (2005). Microstructure of a spatial map in the enthorhinal cortex. *Nature*, 436:801–806.

[He et al., 2018] He, L., Qin, W., Xu, P., and Zhou, Y. (2018). *alocv: Approximate Leave-One-Out Risk Estimation*. R package version 0.02.

[Hurvich and Tsai, 1989] Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2).

[Kass et al., 2005] Kass, R. E., Ventura, V., and Brown, E. N. (2005). Statistical issues in the analysis of neuronal data. *J Neurophysiol*, 94:8–25.

[Khan et al., 2014] Khan, U., Liu, L., Provenzano, F. A., Berman, D., Profacia, C., Sloa, R., Mayeux, R., Duff, K., and Small, S. (2014). Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease. *Nat. Neurosci.*, 17:304–311.

[Leeb, 2008] Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, pages 661–690.

[Leeb, 2009] Leeb, H. (2009). Conditional predictive inference post model selection. *Ann. Stat.*, 37(5B):2838–2876.

[Lin et al., 2000] Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Stat.*, 28(6):1570–1600.

[Macke et al., 2011] Macke, J., Gerwinn, S., White, L., Kaschube, M., and Bethge, M. (2011). Gaussian process methods for estimating cortical maps. *Neuroimage*, 56(2):570–581.

[Maleki, 2011] Maleki, A. (2011). Approximate message passing algorithm for compressed sensing. *Stanford University PhD Thesis*.

[Mallows, 1973] Mallows, C. (1973). Some comments on c_p. *Technometrics*, 15:661–675.

[Meijer and Goeman, 2013] Meijer, R. and Goeman, J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155.

[Moser et al., 2008] Moser, E., Kropff, E., and Moser, M. (2008). Place cells, grid cells, and the brain's spatial representation system. *Ann. Rev. Neurosci.*, 31:69–89.

[Moser et al., 2014] Moser, E., Moser, M., and Roudi, Y. (2014). Network mechanisms of grid cells. *Phil. Trans. R. Soc. B*, 369(1635).

[Mousavi et al., 2018] Mousavi, A., Maleki, A., Baraniuk, R. G., et al. (2018). Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148.

[Negahban et al., 2012] Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2012). High-dimensional generalized linear models and the lasso. *Statistical Science*, 4(27):538–557.

[Nesterov, 2013] Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

[Nevo and Ritov, 2016] Nevo, D. and Ritov, Y. (2016). On Bayesian robust regression with diverging number of predictors. *Electron. J. Statist.*, 10(2):3045–3062.

[Obuchi and Kabashima, 2016] Obuchi, T. and Kabashima, Y. (2016). Cross validation in lasso and its acceleration. *J. Stat. Mech. Theor. Exp.*, 53(304):1–36.

[Opper and Winther, 2000] Opper, M. and Winther, O. (2000). Gaussian processes and SVM: Mean field results and leave-one-out. In Smola, A., Bartlett, P., Scholkopf, B., and Schuurmans, D., editors, *Advances Large Margin Classifiers*, pages 43–56. MIT Press, Cambridge, MA.

[O'Sullivan et al., 1986] O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *JASA*, 81(393):96–103.

[Paninski, 2004] Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15:243–262.

[Paninski et al., 2010] Paninski, L., Ahmadian, Y., Ferreira, D., Koyama, S., Rahnama Rad, K., Vidne, M., Vogelstein, J., and Wu, W. (2010). A new look at state-space models for neural data. *J. Comput. Neurosci.*, 29(0):107–126.

[Park et al., 2014] Park, M., Weller, J., Horowitz, G., and Pillow, J. (2014). Bayesian active learning of neural firing rate maps with transformed Gaussian process priors. *Neural Computation*, 26:1519–1541.

[Pillow, 2007] Pillow, J. (2007). Likelihood-based approaches to modeling the neural code. *Bayesian brain: Probabilistic approaches to neural coding*, pages 53–70.

[Pnevmatikakis et al., 2014] Pnevmatikakis, E., Rahnama Rad, K., Huggins, J., and Paninski, L. (2014). Fast Kalman filtering and forward-backward smoothing via low-rank perturbative approach. *J. Comp. Graph. Stat.*, 23(316-339).

[Qian et al., 2013] Qian, J., Hastie, T., Friedman, J., Tibshirani, R., and Simon, N. (2013). Glmnet for Matlab.

[Rahnama Rad and Paninski, 2010] Rahnama Rad, K. and Paninski, L. (2010). Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Computation in Neural Systems*, 21:142–168.

[Rowland et al., 2016] Rowland, D., Roudi, Y., Moser, M., and Moser, E. (2016). Ten years of grid cells. *Ann. Rev. Neurosci.*, 39:19–40.

[Schmidt et al., 2007] Schmidt, M., Fung, G., and Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *ECML*, pages 286–297. Springer.

[Stein, 1981] Stein, C. (1981). Estimation of the mean of a multivariate normal. *Ann. Stat.*, 9(6):1135–1151.

[Stensola et al., 2012] Stensola, H., Stensola, T., Solstad, T., Froland, K., Moser, M., and Moser, E. (2012). The entorhinal grid map is discretized. *Nature*, 492(7427):72–80.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assesment of statistical predictions. *J R Stat Soc Series B*, 36(2):111–147.

[Stone, 1977] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J R Stat Soc Series B*, pages 44–47.

[Su et al., 2017] Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the Lasso path. *Ann. Stat.*, 45(5):2133–2150.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288.

[Tibshirani and Taylor, 2012] Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232.

[Tibshirani et al., 2013] Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

[Van de Geer, 2008] Van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 2(36):614–645.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[Vehtari et al., 2017] Vehtari, A., Gelman, A., and Gabry, J. (2017). High-dimensional generalized linear models and the lasso. *Statistics and Computing*, 5(27):1413–1432.

[Vehtari et al., 2016] Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research*, 17(1):3581–3618.

[Wahba et al., 1995] Wahba, G., Johnson, D., Gao, F., and Gong, J. (1995). Adaptive tuning of numerical weather prediction models: Randomized GCV in three- and four-dimensionalassimilation. *Monthly Weather Review*, 123:3358–3369.

[Wainwright, 2009] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ell_1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202.

[Weng et al., 2018] Weng, H., Maleki, A., and Zheng, L. (2018). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Annals of Statistics*.

[Xiang and Wahba, 1996] Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, 6:675–692.

[Zolrowski and Pillow, 2018] Zolrowski, D. and Pillow, J. (2018). Scaling the Poisson GLM to massive neural datasets through polynomial approximations. *Advances in Neural Information Processing Systems*, pages 3517–3527.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.*, 67(2):301–320.

[Zou et al., 2007] Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Stat.*, 35(5):2173–2192.

A Proofs (FOR ON-LINE PUBLICATION ONLY)

A.1 Several concentration results for Gaussian random vectors and matrices

In this section, we mention a few concentration results that will be used multiple times in the proofs of our main results. We standard and well-known Gaussian tail bound:

Lemma 10. *Let $Z \sim N(0,1)$. Further assume that $t > 1$. Then,*

$$P(Z > t) \leq \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}.$$

Our next lemma obtains a tail bound for the magnitude of a Gaussian random vector and the maximum eigenvalue of a Gaussian matrix.

Lemma 11 (Due to [Boucheron et al., 2013]). *Let $\boldsymbol{x} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ with $\rho_{\max} \triangleq \sigma_{\max}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathrm{R}^{p \times p}$ then*

$$\Pr\left[\|\boldsymbol{x}\|_2^2 > 5p\rho_{\max}\right] \quad \leq \quad e^{-p}. \tag{30}$$

Furthermore, if $\boldsymbol{X} \in \mathrm{R}^{n \times p}$ is composed of independently distributed $\mathrm{N}(0, \frac{1}{n})$ entries, then

$$\Pr\left[\sqrt{\sigma_{\max}(\boldsymbol{X}^\top \boldsymbol{X})} \geq 1 + \sqrt{\frac{p}{n}} + t\right] \quad \leq \quad e^{-\frac{nt^2}{2}}. \tag{31}$$

The above lemma shows how we can find a tail bound for the maximum singular value of an iid Gaussian matrix. Below we extend the result to Gaussian matrices whose columns are dependent on each other. Note that Lemma 12 is the same as Lemma 5 with $n\rho_{\max} = c$ and $\sqrt{\frac{p}{n}} = \frac{1}{\sqrt{\delta_0}}$. Hence we present the proof Lemma 12 which can be easily used to prove Lemma 5.

Lemma 12. *$\boldsymbol{X} \in \mathrm{R}^{n \times p}$ is composed of independently distributed $\mathrm{N}(0, \boldsymbol{\Sigma})$ rows, with $\rho_{\max} \triangleq \sigma_{\max}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathrm{R}^{p \times p}$ then*

$$\Pr\left[\sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \geq \left(\sqrt{n} + 3\sqrt{p}\right)^2 \rho_{\max}\right] \leq e^{-p}. \tag{32}$$

Proof. Since $\boldsymbol{X} \in \mathrm{R}^{n \times p}$ is composed of independently distributed $\mathrm{N}(0, \boldsymbol{\Sigma})$ rows, then

$$\begin{aligned}
&\Pr\left[\sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \geq \sigma_0\right] = \Pr\left[\sigma_{\max}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) \geq \sigma_0\right] \\
&= \Pr\left[\max_{\|\boldsymbol{u}\|_2^2 \leq 1} \|\boldsymbol{X}\boldsymbol{u}\|_2^2 \geq \sigma_0\right] = \Pr\left[\max_{\|\boldsymbol{u}\|_2^2 \leq 1} \left\|\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\right\|_2^2 \geq \sigma_0\right] \\
&= \Pr\left[\max_{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}\|_2^2 \leq 1} \|\boldsymbol{Z}\boldsymbol{u}\|_2^2 \geq \sigma_0\right] \leq \Pr\left[\max_{\|\frac{\boldsymbol{u}}{\sqrt{\rho_{\max}}}\|_2^2 \leq 1} \|\boldsymbol{Z}\boldsymbol{u}\|_2^2 \geq \sigma_0\right] \\
&= \Pr\left[\max_{\|\boldsymbol{u}\|_2^2 \leq 1} \|\boldsymbol{Z}\boldsymbol{u}\|_2^2 \geq \frac{\sigma_0}{\rho_{\max}}\right] = \Pr\left[\sqrt{\sigma_{\max}\left(\frac{\boldsymbol{Z}^\top\boldsymbol{Z}}{n}\right)} \geq \sqrt{\frac{\sigma_0}{n\rho_{\max}}}\right],
\end{aligned} \tag{33}$$

where $\boldsymbol{Z} \in \mathrm{R}^{n \times p}$ is composed of independently distributed $\mathrm{N}(0,1)$ entries. As a consequence of Lemma 11, and letting

$\sigma_0 = n\rho_{\max}\left(1+\sqrt{\frac{p}{n}}+t\right)^2$, we get

$$\Pr\left[\sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \geq n\rho_{\max}\left(1+\sqrt{\frac{p}{n}}+t\right)^2\right] \leq e^{-nt^2/2}. \tag{34}$$

By substituting $t = \sqrt{\frac{2p}{n}}$ in (34), and noting that $3 > 1+\sqrt{2}$, we get

$$\Pr\left[\sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \geq n\rho_{\max}\left(1+3\sqrt{\frac{p}{n}}\right)^2\right] \leq e^{-p}.$$

□

A.2 Proof of Theorem 1

A.2.1 Roadmap of the proof

We first remind the reader that $r_\alpha(z) \triangleq \frac{1}{\alpha}(\log(1+e^{-\alpha z})+\log(1+e^{\alpha z}))$. Before we discuss the proof, let us mention the following definitions:

$$h_\alpha(\boldsymbol{\beta}) \triangleq \sum_{i=1}^n \ell(y_i|\boldsymbol{x}_i^\top\boldsymbol{\beta}) + \lambda\sum_{i=1}^p r_\alpha(\beta_i), \qquad h(\boldsymbol{\beta}) \triangleq \sum_{i=1}^n \ell(y_i|\boldsymbol{x}_i^\top\boldsymbol{\beta}) + \lambda\sum_{i=1}^p |\beta_i|, \tag{35}$$

$$\hat{\boldsymbol{\beta}}^\alpha \triangleq \arg\min_{\boldsymbol{\beta}} h_\alpha(\boldsymbol{\beta}), \qquad \hat{\boldsymbol{\beta}} \triangleq \arg\min_{\boldsymbol{\beta}} h(\boldsymbol{\beta}). \tag{36}$$

Note that according to Assumptions 1 and 2, $\hat{\boldsymbol{\beta}}^\alpha$ and $\hat{\boldsymbol{\beta}}$ are unique. We first mention a few structural properties of $r_\alpha(z)$ that will be used throughout our proof. Since the proofs of these results are straightforward, we skip them.

Lemma 13. *For any $\alpha > 0$ we have $r_\alpha(z) \geq |z|$, and*

$$\sup_z |r_\alpha(z) - |z|| \leq \frac{2\log 2}{\alpha}.$$

In particular, as $\alpha \to \infty$, $r_\alpha(z)$ uniformly converges to $|z|$.

Lemma 14. *$r_\alpha(z)$ is infinitely many times differentiable, and*

$$\begin{aligned}\dot{r}_\alpha(z) &= \frac{\mathrm{e}^{\alpha z}-\mathrm{e}^{-\alpha z}}{\mathrm{e}^{\alpha z}+\mathrm{e}^{-\alpha z}+2} \\ \ddot{r}_\alpha(z) &= \frac{2\alpha}{(\mathrm{e}^{\alpha z}+\mathrm{e}^{-\alpha z}+2)}.\end{aligned} \tag{37}$$

Furthermore, if $|z_\alpha| < \frac{\zeta_1}{\alpha}$ for a constant $\zeta_1 > 0$, then $\lim_{\alpha\to\infty}\ddot{r}_\alpha(z_\alpha) = +\infty$. Finally, if $|z_\alpha| > \zeta_2$ for a constant $\zeta_2 > 0$, then $\lim_{\alpha\to\infty}\ddot{r}_\alpha(z_\alpha) = 0$ and $\lim_{\alpha\to\infty}\dot{r}_\alpha(z_\alpha) = 1$.

Now, we show the main steps for finding the following limit

$$\lim_{\alpha\to\infty}\boldsymbol{H}^\alpha \triangleq \lim_{\alpha\to\infty}\boldsymbol{X}\left(\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha)] + \boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)].$$

In that vein, let

$$\begin{aligned} \boldsymbol{A} &\triangleq \boldsymbol{X}_{S^c}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_{S^c} + \mathrm{diag}[\ddot{\boldsymbol{r}}_{S^c}^\alpha(\hat{\boldsymbol{\beta}}^\alpha)], \qquad \boldsymbol{B} \triangleq \boldsymbol{X}_{S^c}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_S, \\ \boldsymbol{C} &\triangleq \boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_S + \mathrm{diag}[\ddot{\boldsymbol{r}}_S^\alpha(\hat{\boldsymbol{\beta}}^\alpha)], \qquad \boldsymbol{D} \triangleq (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{A}^{-1}\boldsymbol{B})^{-1}, \end{aligned} \tag{38}$$

where $S = \{i : |\hat{\beta}_i| \neq 0\}$. Based on Theorem 4 in Section A.2.3, for large enough α, there exist fixed numbers $\zeta_1, \zeta_2 > 0$ such that

$$\max_{i \in S^c} |\hat{\beta}_i^\alpha| < \frac{\zeta_1}{\alpha}, \text{ and } \min_{i \in S} |\hat{\beta}_i^\alpha| > \zeta_2,$$

which with Lemma 14 implies $\ddot{r}_\alpha(\hat{\beta}_i^\alpha) \to \infty$ for $i \in S^c$ and $\ddot{r}_\alpha(\hat{\beta}_i^\alpha) \to 0$ for $i \in S$, as $\alpha \to \infty$. Since the diagonal elements of $\mathrm{diag}[\ddot{\boldsymbol{r}}_{S^c}^\alpha(\hat{\boldsymbol{\beta}}^\alpha)]$ go off to infinity, $\boldsymbol{A}^{-1} \to 0$, as $\alpha \to \infty$. Furthermore, since the diagonal elements of $\mathrm{diag}[\ddot{\boldsymbol{r}}_S^\alpha(\hat{\boldsymbol{\beta}}^\alpha)]$ converge to zero, $\lim_{\alpha \to \infty} \boldsymbol{D} = (\boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_S)^{-1}$. Therefore, by using the following identity

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D} \\ -\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} & \boldsymbol{D} \end{bmatrix}, \tag{39}$$

and noting that $\lim_{\alpha \to \infty} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} = 0$, $\lim_{\alpha \to \infty} -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D} = 0$, we obtain

$$\begin{aligned} \lim_{\alpha \to \infty} \boldsymbol{H}^\alpha &= \lim_{\alpha \to \infty} \boldsymbol{X} \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^\alpha)] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)] \\ &= \boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]. \end{aligned}$$

Note that in Lemma 15 in Section A.2.2 we prove that $\|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2 \to 0$ as $\alpha \to \infty$. Hence, from the continuity of the second derivative of ℓ (Assumption 3) we have $\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha) \to \ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})$ as $\alpha \to \infty$.

A.2.2 Proof of $\|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2 \to 0$

Lemma 15. *If Assumptions 1 and 2 hold, i.e. uniqueness of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^\alpha$, then* $\lim_{\alpha \to \infty} \|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2 = 0$.

Proof. First note that according to Lemma 13, we have

$$|h(\boldsymbol{\beta}) - h_\alpha(\boldsymbol{\beta})| \leq \frac{2p \log 2}{\alpha}.$$

Hence, we have

$$h_\alpha(\hat{\boldsymbol{\beta}}^\alpha) \geq h(\hat{\boldsymbol{\beta}}^\alpha) - \frac{2p \log 2}{\alpha} \geq h(\hat{\boldsymbol{\beta}}) - \frac{2p \log 2}{\alpha}, \tag{40}$$

and

$$h_\alpha(\hat{\boldsymbol{\beta}}) \leq h(\hat{\boldsymbol{\beta}}) + \frac{2p \log 2}{\alpha}. \tag{41}$$

Suppose that $\|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2$ does not go to zero as $\alpha \to \infty$. Then, there exists an $\epsilon > 0$ for which we can find a sequence $\alpha_1, \alpha_2, \ldots$, such that

$$\|\hat{\boldsymbol{\beta}}^{\alpha_i} - \hat{\boldsymbol{\beta}}\|_2 > \epsilon. \tag{42}$$

According to Lemma 13, we have

$$\lambda\|\hat{\boldsymbol{\beta}}^{\alpha_i}\|_1 \overset{(a)}{\leq} \lambda\sum_{j=1}^{p} r_{\alpha_i}(\hat{\beta}_j^{\alpha_i}) \overset{(b)}{\leq} h_{\alpha_i}(\hat{\boldsymbol{\beta}}^{\alpha_i}) \overset{(c)}{\leq} h_{\alpha_i}(0)=\sum_{j=1}^{n}\ell(y_j|0) + \frac{2p\log 2}{\alpha_i}. \tag{43}$$

Note that Inequality (a) uses Lemma 13 which proves $|\hat{\beta}_j^{\alpha_i}| \leq r_{\alpha_i}(\hat{\beta}_j^{\alpha_i})$. Inequality (b) is due to the fact that $h_\alpha(\boldsymbol{\beta}) = \sum_{i=1}^n \ell(y_i|\boldsymbol{x_i}^\top\boldsymbol{\beta}) + \sum_{i=1}^p r_\alpha(\beta_i)$ and we assume that the loss function returns positive numbers. Inequality (c) is due to the fact that $\hat{\boldsymbol{\beta}}^{\alpha_i}$ is the minimizer of $h_{\alpha_i}(\boldsymbol{\beta})$.

According to (43) the sequence $\hat{\boldsymbol{\beta}}^{\alpha_1}, \hat{\boldsymbol{\beta}}^{\alpha_2}, \ldots$ belongs to a compact set, and hence has a converging subsequence, called $\hat{\boldsymbol{\beta}}^{\tilde{\alpha}_1}, \hat{\boldsymbol{\beta}}^{\tilde{\alpha}_2}, \ldots$. Suppose that $\hat{\boldsymbol{\beta}}^{\tilde{\alpha}_1}, \hat{\boldsymbol{\beta}}^{\tilde{\alpha}_2}, \ldots$ converges to $\tilde{\boldsymbol{\beta}}$. Therefore,

$$h(\hat{\boldsymbol{\beta}}^{\tilde{\alpha}_j}) \overset{(d)}{\leq} h_{\tilde{\alpha}_j}(\hat{\boldsymbol{\beta}}^{\tilde{\alpha}_j}) + \frac{2p\log 2}{\tilde{\alpha}_j} \overset{(e)}{\leq} h_{\tilde{\alpha}_j}(\hat{\boldsymbol{\beta}}) + \frac{2p\log 2}{\tilde{\alpha}_j} \overset{(f)}{\leq} h(\hat{\boldsymbol{\beta}}) + \frac{4p\log 2}{\tilde{\alpha}_j}. \tag{44}$$

Inequality (d) is due to (40). Inequality (e) is true because $\hat{\boldsymbol{\beta}}^{\tilde{\alpha}_j}$ is the minimizer of $h_{\tilde{\alpha}_j}(\boldsymbol{\beta})$, and finally Inequality (f) is due to (41). By taking the limit $j \to \infty$ from both sides of (44), we have

$$h(\tilde{\boldsymbol{\beta}}) \leq h(\hat{\boldsymbol{\beta}}).$$

But $\tilde{\boldsymbol{\beta}}$ is different from $\hat{\boldsymbol{\beta}}$, according to (42), contradicting the uniqueness of $\hat{\boldsymbol{\beta}}$ in Assumption 1. □.

A.2.3 Bounds for regression coefficients in smoothed LASSO

Theorem 4. *Let S denote the active set of $\hat{\boldsymbol{\beta}}$, i.e., the location of its non-zero coefficients. Under assumptions 1, 2, 3, and 4, there exists a fixed numbers $\zeta_1, \zeta_2 > 0$, such that for α large enough, we have*

$$\begin{aligned} \max_{i\in S^c} |\hat{\beta}_i^\alpha| &< \frac{\zeta_1}{\alpha} \\ \min_{i\in S} |\hat{\beta}_i^\alpha| &> \zeta_2. \end{aligned}$$

Proof. The optimality conditions

$$\sum_{i=1}^n \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}^\alpha) + \lambda\dot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha) = 0, \tag{45}$$

$$\sum_{i=1}^n \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}^\alpha) + \lambda\hat{\boldsymbol{g}} = 0, \tag{46}$$

lead to

$$\left\|\lambda\dot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha) - \lambda\hat{\boldsymbol{g}}\right\|_2 = \left\|-\sum_{i=1}^n \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}^\alpha) + \sum_{i=1}^n \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}})\right\|_2. \tag{47}$$

We know $\|\hat{\boldsymbol{\beta}}^\alpha - \hat{\boldsymbol{\beta}}\|_2 \to 0$ from Lemma 15. And since ℓ is twice differentiable (Assumption 3), we can argue that $\| - \sum \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}}^\alpha) + \sum \boldsymbol{x_i}\dot{\ell}(y_i|\boldsymbol{x_i}^\top\hat{\boldsymbol{\beta}})\|_2 \to 0$ as $\alpha \to \infty$. Hence,

$$\|\lambda\dot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha) - \lambda\hat{\boldsymbol{g}}\|_\infty \leq \|\lambda\dot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha) - \lambda\hat{\boldsymbol{g}}\|_2 \to 0, \tag{48}$$

as $\alpha \to \infty$. This shows that for every $i \in S^c$, $|\alpha\hat{\beta}_i^{\alpha}|$ should remain bounded as $\alpha \to \infty$. Suppose that this is not true. Then we find a subsequence that $\alpha_j \hat{\beta}_i^{\alpha_j} \to \infty$ as $j \to \infty$. Then

$$\lim_{j\to\infty} \dot{r}_{\alpha_j}(\hat{\beta}_i^{\alpha_j}) = \lim_{j\to\infty} \frac{e^{\alpha_j \hat{\beta}_i^{\alpha_j}} - e^{-\alpha_j \hat{\beta}_i^{\alpha_j}}}{e^{\alpha_j \hat{\beta}_i^{\alpha_j}} + e^{-\alpha_j \hat{\beta}_i^{\alpha_j}} + 2} = 1.$$

If we combine this with Assumption 4, we conclude that $\|\lambda \dot{\boldsymbol{r}}_\alpha(\hat{\boldsymbol{\beta}}^\alpha) - \lambda\hat{\boldsymbol{g}}\|_\infty$ will be a constant due to the assumption $\sup_{i\in S^c} |\hat{g}_i| < 1$. This is in contradiction with (48). Hence, we have proved that for every $i \in S^c$, $|\alpha\hat{\beta}_i^{\alpha}|$ remains bounded.

Next, we show that $\min_{i\in S} |\hat{\beta}_i^{\alpha}|$ is bounded away from zero in the limit $\alpha \to \infty$. Define $\min_{i\in S} |\hat{\beta}_i| = \gamma > 0$. Lemma 15 implies $\max_{i\in S} |\hat{\beta}_i^{\alpha} - \hat{\beta}_i| \to 0$, and therefore, for α large enough, we have

$$\max_{i\in S} |\hat{\beta}_i^{\alpha} - \hat{\beta}_i| < \gamma/2,$$

leading to

$$\min_{i\in S} |\hat{\beta}_i^{\alpha}| > \min_{i\in S} |\hat{\beta}_i| - \max_{i\in S} |\hat{\beta}_i^{\alpha} - \hat{\beta}_i| > \zeta_2 \triangleq \gamma/2.$$

□

A.3 Proof of Theorem 2

The following lemma plays a critical role in our proof of Theorem 2.

Lemma 16. *Consider a class of symmetric positive definite matrices of the form*

$$\boldsymbol{\Gamma}_\delta \triangleq \begin{bmatrix} a+\delta & \boldsymbol{b}^\top \\ \boldsymbol{b} & \boldsymbol{C} \end{bmatrix}, \tag{49}$$

where $a > 0$, $\delta \geq 0$ and $\boldsymbol{C} \in \mathbb{R}^{n-1\times n-1}$. Then, for any vector $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$\lim_{\delta\to\infty} \boldsymbol{v}^\top \boldsymbol{\Gamma}_\delta^{-1} \boldsymbol{v} \leq \boldsymbol{v}^\top \boldsymbol{\Gamma}_\delta^{-1} \boldsymbol{v} \leq \boldsymbol{v}^\top \boldsymbol{\Gamma}_0^{-1} \boldsymbol{v}.$$

Furthermore, if we define $\boldsymbol{v}_{/1} \triangleq (v_2, v_3, \ldots, v_n)^\top$, then $\lim_{\delta\to\infty} \boldsymbol{v}^\top \boldsymbol{\Gamma}_\delta^{-1} \boldsymbol{v} = \boldsymbol{v}_{/1}^\top \boldsymbol{C}^{-1} \boldsymbol{v}_{/1}$.

Proof: Define $\kappa \triangleq a + \delta - \boldsymbol{b}^\top \boldsymbol{C}^{-1} \boldsymbol{b}$. Note that since the matrix $\boldsymbol{\Gamma}_\delta$ is always positive definite, for any value of δ, $\kappa > 0$. By using the formulas for the inverse of block matrices we have

$$\boldsymbol{\Gamma}_\delta^{-1} = \begin{bmatrix} \frac{1}{\kappa} & -\frac{\boldsymbol{b}^\top \boldsymbol{C}^{-1}}{k} \\ -\frac{\boldsymbol{C}^{-1}\boldsymbol{b}}{\kappa} & \frac{\boldsymbol{C}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}^{-1}}{\kappa} + \boldsymbol{C}^{-1} \end{bmatrix}. \tag{50}$$

Define $\boldsymbol{v}_{/1} \triangleq (v_2, v_3, \ldots, v_n)^\top$.

$$\begin{aligned} \boldsymbol{v}^\top \boldsymbol{\Gamma}_\delta^{-1} \boldsymbol{v} &= \frac{v_1^2}{\kappa} + \boldsymbol{v}_{/1}^\top \boldsymbol{C}^{-1} \boldsymbol{v}_{/1} + \frac{\boldsymbol{v}_{/1}^\top \boldsymbol{C}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}^{-1}\boldsymbol{v}_{/1}}{\kappa} - 2\frac{\boldsymbol{v}_{/1}^\top \boldsymbol{C}^{-1}\boldsymbol{b} v_1}{\kappa} \\ &= \boldsymbol{v}_{/1}^\top \boldsymbol{C}^{-1} \boldsymbol{v}_{/1} + \frac{1}{\kappa}(v_1 - \boldsymbol{b}^\top \boldsymbol{C}^{-1} \boldsymbol{v}_{/1})^2. \end{aligned} \tag{51}$$

Lemma 16 follows from the monotonicity of $\boldsymbol{v}^\top \boldsymbol{\Gamma}_\delta^{-1}\boldsymbol{v}$ in terms of κ. $\square$

Proof of Theorem 2. Before we start the proof, let us emphasize on the following facts that will be used later in the proof.

1. Consider an index $i \in (S\cup T)^c$. We know that $\hat{\beta}_i = 0$ and the subgradient $|\hat{g}_i| < 1$. Hence, according to the proof of Theorem 4 we have $\alpha\hat{\beta}_i^\alpha < \zeta$. Therefore, according to Lemma 14 we have $\ddot{r}(\hat{\beta}_i^\alpha) \to \infty$ as $\alpha \to \infty$.
2. Consider $i \in S$. Then by definition $\hat{\beta}_i \neq 0$. Similar to the proof of Theorem 1, we have $\ddot{r}(\hat{\beta}_i^\alpha) \to 0$ as $\alpha \to \infty$.

Hence, we already know the limiting behavior of $\ddot{r}(\hat{\beta}_i^\alpha)$ for $i \in S$ and $i \in (S\cup T)^c$ as $\alpha \to \infty$. The only remaining index set is T. Unfortunately, for $i \in T$ we can not specify the limiting behavior of $\ddot{r}(\hat{\beta}_i^\alpha)$. Hence, our goal is to use Lemma 16 to get around this issue. Set $U \triangleq (S\cup T)^c$ and define the matrices

$$\begin{aligned}
\tilde{\boldsymbol{A}}^\alpha &\triangleq \boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_S + \mathrm{diag}[\ddot{\boldsymbol{r}}_S^\alpha(\hat{\boldsymbol{\beta}}^\alpha)], \qquad &\tilde{\boldsymbol{B}}^\alpha \triangleq \boldsymbol{X}_T^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_T,\\
\tilde{\boldsymbol{C}}^\alpha &\triangleq \boldsymbol{X}_U^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_U + \mathrm{diag}[\ddot{\boldsymbol{r}}_U^\alpha(\hat{\boldsymbol{\beta}}^\alpha)], \qquad &\tilde{\boldsymbol{D}}^\alpha \triangleq \boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_T,\\
\tilde{\boldsymbol{E}}^\alpha &\triangleq \boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_U, \qquad &\tilde{\boldsymbol{F}}^\alpha \triangleq \boldsymbol{X}_T^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\alpha)]\boldsymbol{X}_U.
\end{aligned} \tag{52}$$

Given this notation we have

$$H_{ii}^\alpha = \boldsymbol{x}_i^\top \begin{bmatrix} \tilde{\boldsymbol{A}}^\alpha & \tilde{\boldsymbol{D}}^\alpha & \tilde{\boldsymbol{E}}^\alpha \\ (\tilde{\boldsymbol{D}}^\alpha)^\top & \tilde{\boldsymbol{B}}^\alpha + \mathrm{diag}[\ddot{\boldsymbol{r}}_T^\alpha(\hat{\boldsymbol{\beta}}^\alpha)] & \tilde{\boldsymbol{F}}^\alpha \\ (\tilde{\boldsymbol{E}}^\alpha)^\top & (\tilde{\boldsymbol{F}}^\alpha)^\top & \tilde{\boldsymbol{C}}^\alpha \end{bmatrix}^{-1} \boldsymbol{x}_i \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha). \tag{53}$$

Here each element of $\mathrm{diag}[\ddot{\boldsymbol{r}}_T^\alpha(\hat{\boldsymbol{\beta}}^\alpha)]$ may converge to any number in the range $[0,\infty]$. Hence we use Lemma 16 to find upper and lower bounds for H_{ii}^α. According to Lemma 16 we have

$$H_{ii}^\alpha \leq \boldsymbol{x}_i^\top \begin{bmatrix} \tilde{\boldsymbol{A}}^\alpha & \tilde{\boldsymbol{D}}^\alpha & \tilde{\boldsymbol{E}}^\alpha \\ (\tilde{\boldsymbol{D}}^\alpha)^\top & \tilde{\boldsymbol{B}}^\alpha & \tilde{\boldsymbol{F}}^\alpha \\ (\tilde{\boldsymbol{E}}^\alpha)^\top & (\tilde{\boldsymbol{F}}^\alpha)^\top & \tilde{\boldsymbol{C}}^\alpha \end{bmatrix}^{-1} \boldsymbol{x}_i \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha), \tag{54}$$

and

$$\begin{aligned}
H_{ii}^\alpha &\geq \lim_{\delta_{|T|}\to\infty} \cdots \lim_{\delta_1\to\infty} \boldsymbol{x}_i^\top \begin{bmatrix} \tilde{\boldsymbol{A}}^\alpha & \tilde{\boldsymbol{D}}^\alpha & \tilde{\boldsymbol{E}}^\alpha \\ (\tilde{\boldsymbol{D}}^\alpha)^\top & \tilde{\boldsymbol{B}}^\alpha + \mathrm{diag}[\delta_1,\delta_2,\ldots,\delta_{|T|}] & \tilde{\boldsymbol{F}}^\alpha \\ (\tilde{\boldsymbol{E}}^\alpha)^\top & (\tilde{\boldsymbol{F}}^\alpha)^\top & \tilde{\boldsymbol{C}}^\alpha \end{bmatrix}^{-1} \boldsymbol{x}_i \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha)\\
&= \boldsymbol{x}_{i,,S\cup U}^\top \begin{bmatrix} \tilde{\boldsymbol{A}}^\alpha & \tilde{\boldsymbol{E}}^\alpha \\ (\tilde{\boldsymbol{E}}^\alpha)^\top & \tilde{\boldsymbol{C}}^\alpha \end{bmatrix}^{-1} \boldsymbol{x}_{i,S\cup U} \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\alpha).
\end{aligned} \tag{55}$$

The rest of the proof is similar to the proof of Theorem 1; we take the limit $\alpha \to \infty$ from both sides of (54) and (55), and then use the block matrix inversion formulas (similar to those used in the proof of Theorem 1) and the fact that $(\tilde{\boldsymbol{A}}^\alpha)^{-1} \to 0$ as $\alpha \to \infty$ to complete the proof. $\square$

A.4 Derivation of (18)

A.4.1 Roadmap of the derivations

The goal of this section is to derive the ALO formula, presented in (18), for the following class of bridge estimators:

$$\hat{\boldsymbol{\beta}} \triangleq \arg\min_{\boldsymbol{\beta}\in \mathrm{R}^p} \Big\{ \sum_{i=1}^{n} \ell(y_i|\boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q^q \Big\}, \tag{56}$$

where $q \in (1,2)$. Since $\|\boldsymbol{\beta}\|_q^q$ is not twice differentiable at zero, similar to what we did for LASSO, we first consider a smoothed version of the bridge regularizer:

$$r_\gamma^q(z) = \frac{1}{\gamma}\int |u|^q \psi((z-u)/\gamma)du, \tag{57}$$

where ψ satisfies the following conditions:

(i) ψ has a compact support, i.e., $\mathrm{supp}(\psi) = [-1,1]$. Also, $\psi(w) \geq 0$ for every w.

(ii) $\int \psi(w)dw = 1$ and $\psi(0) > 0$;

(iii) ψ is infinitely many times smooth and symmetric around 0 on $\mathbb{R}$;

The two important properties of $r_\gamma^q(z)$ are

1. $r_\gamma^q(z)$ is infinitely many times differentiable for any nonzero value of γ.
2. $|r_\gamma^q(z) - |z|^q| \to 0$ as $\gamma \to 0$. This claim will be proved in Lemma 17 below.

Hence, instead of finding the ALO formula directly for (56), we start with

$$\hat{\boldsymbol{\beta}}^\gamma \triangleq \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ell(y_i|\boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} r_\gamma^q(\beta_i). \tag{58}$$

Given that both the loss function and the regularizer are smooth in (58), we can use (6) to obtain the following formula as the estimate of the out-of-sample prediction error of $\hat{\boldsymbol{\beta}}^\gamma$:

$$\mathrm{ALO}^\gamma \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}^\gamma + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma)}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma)}\right)\left(\frac{H_{ii}^\gamma}{1-H_{ii}^\gamma}\right)\right), \tag{59}$$

where

$$\boldsymbol{H}^\gamma \triangleq \boldsymbol{X}\left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}_\gamma^q(\hat{\boldsymbol{\beta}}^\gamma)] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\gamma)]\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\gamma)]. \tag{60}$$

Note that we are interested in ALO^γ for large values of γ. Hence, as suggested for the LASSO problem in Section 2.2, we calculate $\lim_{\gamma\to 0} \mathrm{ALO}^\gamma$. In Section A.4.3 we prove the following theorem:

Theorem 5. *If the loss function is twice continuously differentiable with respect to its second argument, and the optimization problem in* (58) *has a unique solution for every* γ*, then*

$$\lim_{\gamma\to 0} \mathrm{ALO}^\gamma \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right), \tag{61}$$

where

$$\boldsymbol{H} \triangleq \boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})] \boldsymbol{X}_S + \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}_S^q(\hat{\boldsymbol{\beta}})] \right)^{-1} \boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})], \tag{62}$$

S is the active set of $\hat{\boldsymbol{\beta}}$, and $\ddot{r}^q(u) = q(q-1)|u|^{q-2}$.

Proof of this Theorem is presented in Section A.4.3. We will first prove in Lemma 19 that

$$\|\hat{\boldsymbol{\beta}}^\gamma - \hat{\boldsymbol{\beta}}\|_2 \to 0.$$

Since, $\|\hat{\boldsymbol{\beta}}^\gamma - \hat{\boldsymbol{\beta}}\|_2 \to 0$, and $\dot{\ell}$ and $\ddot{\ell}$ functions are continuous, it is straightforward to prove that as $\gamma \to 0$

$$\dot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma) \to \dot{\ell}_i(\hat{\boldsymbol{\beta}}), \qquad \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma) \to \ddot{\ell}_i(\hat{\boldsymbol{\beta}}).$$

Hence, the final remaining challenge in proving Theorem 5 is to calculate $\lim_{\gamma\to 0} H_{ii}^\gamma$. In Section A.4.3 we prove that

$$\lim_{\gamma \to 0} \boldsymbol{H}^\gamma = \boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})] \boldsymbol{X}_S + \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}_S^q(\hat{\boldsymbol{\beta}})] \right)^{-1} \boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})].$$

A.4.2 Basic properties of $r_\gamma^q(\cdot)$

Lemma 17. *The smoothed regularizer $r_\gamma^q(\cdot)$ satisfies*

$$\sup_{|z|<M} |r_\gamma^q(z) - |z|^q| \leq q(M+\gamma)^{q-1}\gamma.$$

Proof. According to the symmetry, we only consider $z \geq 0$. We have

$$\begin{aligned} |r_\gamma^q(z) - |z|^q| &= \frac{1}{\gamma} \left| \int_{-\gamma}^{\gamma} (|z-u|^q - |z|^q) \psi\left(\frac{u}{\gamma}\right) du \right| \\ &\overset{(a)}{\leq} (z+\gamma)^q - z^q \overset{(b)}{=} q\tilde{z}^{q-1}\gamma \leq q(M+\gamma)^{q-1}\gamma. \end{aligned} \tag{63}$$

Note that inequality (a) is due to the fact that since $z > 0$, the difference between $|z-u|^q - |z|^q$ is maximized when $u = -\gamma$. In other words,

$$||z-u|^q - |z|^q| \leq (z+\gamma)^q - z^q, \quad \forall u \in [-\gamma, \gamma].$$

Furthermore, equality (b) is a result of the mean value theorem and $\tilde{z} \in (z, z+\gamma)$. □

A.4.3 Proof of Theorem 5

Consider the following definitions:

$$h_\gamma^q(\boldsymbol{\beta}) \triangleq \sum_{i=1}^n \ell(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda \sum_{i=1}^p r_\gamma^q(\beta_i), \qquad h^q(\boldsymbol{\beta}) \triangleq \sum_{i=1}^n \ell(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda \sum_{i=1}^p |\beta_i|^q, \tag{64}$$

As discussed in Lemma 17, the difference $|r_\gamma^q(z) - |z|^q|$ is bounded by the maximum value that z takes. Our first lemma shows that $\sup_{\gamma\in(0,1]} \|\hat{\boldsymbol{\beta}}^\gamma\|_\infty < M$. Hence, according to Lemma 17 the discrepancy between $|r_\gamma^q(\hat{\beta}_i^\gamma) - |\hat{\beta}_i^\gamma|^q|$ goes to zero as $\gamma \to 0$.

Lemma 18. *There exists an $M < \infty$ such that $\sup_{\gamma \in [0,1]} \|\hat{\boldsymbol{\beta}}^\gamma\|_\infty < M$ and $\|\hat{\boldsymbol{\beta}}\|_\infty < M$.*

Here we only present the sketch of the proof, and skip the straightforward details. If $\|\hat{\boldsymbol{\beta}}^\gamma\|_\infty \to \infty$, then $h_\gamma^q(\hat{\boldsymbol{\beta}}^\gamma) \to \infty$. Hence, since $h_\gamma^q(\mathbf{0})$ is bounded, $\|\hat{\boldsymbol{\beta}}^\gamma\|_\infty$ cannot go off to infinity. We can now use this lemma to prove that as $\gamma \to 0$, $\|\hat{\boldsymbol{\beta}}^\gamma - \hat{\boldsymbol{\beta}}\|_2 \to 0$.

Lemma 19. *If the optimization problems in (64) have unique solutions, then as $\gamma \to 0$*

$$\|\hat{\boldsymbol{\beta}}^\gamma - \hat{\boldsymbol{\beta}}\|_2 \to 0.$$

The proof of this lemma is similar to the proof of Lemma 15, and is hence skipped here. As mentioned in Section A.4.1, the main step in proving Theorem 5 is to find the limit of $\lim_{\gamma \to 0} \boldsymbol{H_\gamma}$. The main step in this calculation is to calculate $\lim_{\gamma \to 0} \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^\gamma)$. The following lemma shows how this limit can be calculated.

Lemma 20. *Let z_γ denote a function of γ. If $z_\gamma \to 0$ as $\gamma \to 0$, then*

$$\lim_{\gamma \to 0} \ddot{r}_\gamma^q(z_\gamma) = \infty.$$

Proof. Without loss of generality we consider the case $z_\gamma \geq 0$. We consider three different cases. Each case has a slightly different proof strategy.

1. Case I: $\frac{z_\gamma}{\gamma} \to \infty$ or $\frac{z_\gamma}{\gamma} \to c \geq 1$.
2. Case II: $\frac{z_\gamma}{\gamma} \to c$, where $c \in (0,1)$.
3. Case III: $\frac{z_\gamma}{\gamma} \to 0$.

It is straightforward to show that

$$\ddot{r}_\gamma^q(z_\gamma) = \int_{-\infty}^{\infty} q|z_\gamma - u|^{q-1} \text{sign}(z_\gamma - u) \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) du. \tag{65}$$

Note that $\dot{\psi}\left(\frac{u}{\gamma}\right) = 0$ for u outside the interval $[-\gamma, \gamma]$. Now we consider the three cases we described above.

Case I: We assume that for large enough values of γ, $z_\gamma > \gamma$. Clearly, this holds when $z_\gamma/\gamma \to c > 1$. However, it may be violated when $\frac{z_\gamma}{\gamma} \to 1$. But, this special case can be handled with a similar approach and is hence skipped. We have

$$\begin{aligned}
|\ddot{r}_\gamma^q(z_\gamma)| &= |\int_{-\gamma}^{0} q(z_\gamma - u)^{q-1} \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) du + \int_0^\gamma q(z_\gamma - u)^{q-1} \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) du| \\
&= |\int_0^\gamma q[(z_\gamma - u)^{q-1} - (z_\gamma + u)^{q-1}] \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) dt| \\
&\overset{(a)}{=} 2q(q-1)|\int_0^\gamma \tilde{z}_u^{q-2} u \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) du| \\
&\overset{(b)}{\geq} 2q(q-1)(z_\gamma + \gamma)^{q-2} |\int_0^\gamma u \frac{1}{\gamma^2} \dot{\psi}\left(\frac{u}{\gamma}\right) du| \\
&\overset{(c)}{=} q(q-1)(z_\gamma + \gamma)^{q-2}.
\end{aligned} \tag{66}$$

Equality (a) is due to the mean-value theorem. To obtain (b) we used the fact that $\tilde{z}_u \in [z_\gamma - \gamma, z_\gamma + \gamma]$, and that $z_\gamma - \gamma > 0$ (hence $\tilde{z}_u > 0$). The last equality is the result of integration by parts. Note that since $z_\gamma \to 0$ as $\gamma \to 0$,

$q(q-1)(z_\gamma+\gamma)^{q-2} \to \infty$.

Case II: $\frac{z_\gamma}{\gamma} \to c$, where $c \in (0,1)$. For large enough values of γ, we know that $z_\gamma < \gamma$. Hence, according to (65) we have

$$\begin{aligned}
\ddot{r}^q_\gamma(z_\gamma) &= -\int_0^\gamma q(z_\gamma+u)^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du + \int_0^{z_\gamma} q(z_\gamma-u)^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\quad -\int_{z_\gamma}^\gamma q|z_\gamma-u|^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\geq -\int_{z_\gamma}^\gamma q(z_\gamma+u)^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\geq -q(2z_\gamma)^{q-1}\int_{z_\gamma}^\gamma \frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du = \frac{q(2z_\gamma)^{q-1}}{\gamma}\psi\left(\frac{z_\gamma}{\gamma}\right). \qquad (67)
\end{aligned}$$

It is straightforward to confirm that $\frac{q(2z_\gamma)^{q-1}}{\gamma}\psi(\frac{z_\gamma}{\gamma}) \to \infty$.

Case III: First note that since $z_\gamma/\gamma \to 0$, for large enough γ, $z_\gamma < \gamma/2$. Similar to the derivation in (67), we have

$$\begin{aligned}
\ddot{r}^q_\gamma(z_\gamma) &\geq -\int_{z_\gamma}^\gamma q|z_\gamma-u|^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\geq -\int_{z_\gamma}^\gamma q(z_\gamma+u)^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\geq -\int_{\gamma/2}^\gamma q(z_\gamma+u)^{q-1}\frac{1}{\gamma^2}\dot{\psi}\left(\frac{u}{\gamma}\right)du \\
&\geq q(z_\gamma+\frac{\gamma}{2})^{q-1}\frac{r(0.5)}{\gamma}. \qquad (68)
\end{aligned}$$

Again, it is straightforward to see that the last expression goes to ∞ as $\gamma \to 0$. □

We remind the reader that our goal is to show that

$$\lim_{\gamma\to 0}\boldsymbol{H}^\gamma = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + \lambda\operatorname{diag}[\ddot{\boldsymbol{r}}^q_S(\hat{\boldsymbol{\beta}})]\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})],$$

where S denotes the active set of $\hat{\boldsymbol{\beta}}$. Since, $\|\hat{\boldsymbol{\beta}}^\gamma - \hat{\boldsymbol{\beta}}\|_2 \to 0$, and $\dot{\ell}$ and $\ddot{\ell}$ functions are continuous, it is straightforward to prove that

$$\dot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma) \to \dot{\ell}_i(\hat{\boldsymbol{\beta}}), \qquad \ddot{\ell}_i(\hat{\boldsymbol{\beta}}^\gamma) \to \ddot{\ell}_i(\hat{\boldsymbol{\beta}}).$$

Let S denote the set of indices of the non-zero elements of $\hat{\boldsymbol{\beta}}$, and define

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{X}_{S^c}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\gamma)]\boldsymbol{X}_{S^c} + \operatorname{diag}[\ddot{\boldsymbol{r}}^q_{\gamma,S^c}(\hat{\boldsymbol{\beta}}^\gamma)], \qquad \boldsymbol{B} = \boldsymbol{X}_{S^c}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\gamma)]\boldsymbol{X}_S, \\
\boldsymbol{C} &= \boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}^\gamma)]\boldsymbol{X}_S + \operatorname{diag}[\ddot{\boldsymbol{r}}^q_{\gamma,S}(\hat{\boldsymbol{\beta}}^\gamma)]. \qquad (69)
\end{aligned}$$

Also define $\boldsymbol{D} \triangleq (\boldsymbol{C} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}$. According to Lemmas 19 and 20, the diagonal elements of $\operatorname{diag}[\ddot{\boldsymbol{r}}^q_{\gamma,S^c}(\hat{\boldsymbol{\beta}}^\gamma)]$ go off to infinity. Hence, it is straightforward to show that $\boldsymbol{A}^{-1} \to 0$, as $\gamma \to 0$. By using the following identity

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D} \\ -\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} & \boldsymbol{D} \end{bmatrix}, \tag{70}$$

and noting that $\lim_{\gamma\to 0} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top \boldsymbol{A}^{-1} = 0$, $\lim_{\gamma\to 0} -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{D} = 0$, and $\lim_{\gamma\to 0} \boldsymbol{D} = (\boldsymbol{X}_S^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S)^{-1}$ we obtain

$$\lim_{\gamma\to 0} \boldsymbol{H}^\gamma \;=\; \boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \,\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S + \lambda \,\mathrm{diag}[\ddot{\boldsymbol{r}}_S^q(\hat{\boldsymbol{\beta}}_\lambda)\right)^{-1} \boldsymbol{X}_S^\top \,\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})].$$

A.5 Proofs of the Lemmas of Section 4

A.5.1 Proof of Lemma 3

Since

$$\dot{\ell}_i(\boldsymbol{\beta}) \;=\; -y_i + \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}, \quad \ddot{\ell}_i(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})^2}, \quad \dddot{\ell}_i(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}(1-e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})}{(1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})^3}$$

using simple algebra it is straightforward to show that for any $\boldsymbol{\beta}$, we have

$$\|\dot{\ell}(\boldsymbol{\beta})\|_\infty \leq 1, \quad \|\ddot{\ell}(\boldsymbol{\beta})\|_\infty \leq 1/4, \quad \|\dddot{\ell}(\boldsymbol{\beta})\|_\infty \leq 1/10$$

Therefore,

$$\begin{aligned}
\|\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta})\|_2 \;&\leq\; \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_2 = \sqrt{\sum_i \left(\ddot{\ell}(\beta_i+\delta_i) - \ddot{\ell}(\beta_i)\right)^2} \\
&=\; \sqrt{\sum_i \dddot{\ell}(\beta_i+\epsilon_i)^2(\boldsymbol{x}_i^T\boldsymbol{\delta})^2} \quad \text{using the mean-value Theorem where} \quad \epsilon_i \in [0,\delta_i] \\
&\leq\; \sqrt{\boldsymbol{\delta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\delta}} \leq \sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}\|\boldsymbol{\delta}\|_2.
\end{aligned}$$

Finally, based on the inequality above, we have

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \;\leq\; \sup_{t\in[0,1]} \frac{(1-t)\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2\sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq \sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}.$$

The last statement of the Theorem is a direct result of Lemma 12.

A.5.2 Proof of Lemma 8

Since

$$\begin{aligned}
\dot{\ell}_i(\boldsymbol{\beta}) \;&=\; f'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - y_i f'(\boldsymbol{x}_i^\top \boldsymbol{\beta})/f(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \\
\ddot{\ell}_i(\boldsymbol{\beta}) \;&=\; f''(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - y_i (f'/f)'(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \\
\dddot{\ell}_i(\boldsymbol{\beta}) \;&=\; f'''(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - y_i (f'/f)''(\boldsymbol{x}_i^\top \boldsymbol{\beta}),
\end{aligned}$$

where

$$f'(z) = \frac{e^z}{1+e^z}, \quad f''(z) = \frac{e^z}{(1+e^z)^2}, \quad f'''(z) = \frac{e^z(1-e^z)}{(1+e^z)^3}. \tag{71}$$

Concerning equations (71), the following inequalities hold

$$f'(z) \le 1, \quad f''(z) \le 1/4, \quad f'''(z) \le 1/10.$$

For any $x > 0$, consider the function

$$h(x) := \frac{x}{(1+x)\log(1+x)}.$$

It is straightforward to check that $h(x)$ is a decreasing function of $x > 0$ and that $\lim_{x\to 0} h(x) = 1$. Hence, by simply using $x = e^z$, for any z, we have

$$f'(z)/f(z) \le 1 \quad \text{leading to} \quad \|\dot{\ell}(\boldsymbol{\beta})\|_\infty \le 1 + \|\boldsymbol{y}\|_\infty.$$

Moreover,

$$\begin{aligned} f''/f &= f'/f \times 1/(1+e^z) \le 1 \\ f'''/f &= f''/f \times (1-e^z)/(1+e^z) \le 1. \end{aligned}$$

Since

$$(f'/f)'' = f'''/f + 2f'^3/f^3 - 3f'f''/f^2 \quad \text{leading to} \quad |(f'/f)''| \le 6.$$

Therefore, $\|\dddot{\ell}(\boldsymbol{\beta})\|_\infty \le 1 + 6\|\boldsymbol{y}\|_\infty$, leading to

$$\begin{aligned} \|\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta})\|_2 &\le \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_2 = \sqrt{\sum_i \left(\ddot{\ell}(\beta_i+\delta_i) - \ddot{\ell}(\beta_i)\right)^2} \\ &= \sqrt{\sum_i \dddot{\ell}(\beta_i+\epsilon_i)^2\delta_i^2} \quad \text{using the mean-value Theorem where} \quad \epsilon_i \in [0,\delta_i] \\ &\le (1+6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}\|\boldsymbol{\delta}\|_2. \end{aligned}$$

Finally, based on the inequality above, we have

$$\begin{aligned} \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} &\le (1+6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X})} \sup_{t\in[0,1]} \frac{(1-t)\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \\ &\le (1+6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}. \end{aligned}$$

A.5.3 Proof of Lemma 9

We prove Lemma 9 using the following inequality (for large enough n such that $\sqrt{\tilde{c}}\log^{3/2} n > 1$)

$$\begin{aligned}\Pr\left((1+6\|\boldsymbol{y}\|_\infty)\sqrt{\sigma_{max}(\boldsymbol{X}^\top\boldsymbol{X})} \geq \zeta_1 \log^{3/2} n\right) \;&\leq\; \Pr\left(\|\boldsymbol{y}\|_\infty \geq (2\log n)\sqrt{9\tilde{c}\log n}\right)\\ &+\; \Pr\left(\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X}) \geq c(1+\frac{3}{\sqrt{\delta_0}})^2\right),\end{aligned}$$

where $\zeta_1 = 37\sqrt{c\tilde{c}}\left(1+\frac{3}{\sqrt{\delta_0}}\right)$. Note that according to Lemma 12, we have

$$\Pr\left[\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X}) \geq c(1+\frac{3}{\sqrt{\delta_0}})^2\right] \leq \mathrm{e}^{-p}.$$

In the next step, we bound $\|\boldsymbol{y}\|_\infty$. We have

$$\Pr(\|\boldsymbol{y}\|_\infty \geq t \mid \boldsymbol{X}) \leq \sum_{i=1}^n \Pr(y_i \geq t \mid \boldsymbol{X}). \tag{72}$$

For $t > \|\boldsymbol{\lambda}\|_\infty$, set $\gamma_i = \log\left(\frac{t}{\lambda_i}\right)$. We have

$$\Pr(y_i \geq t \mid \boldsymbol{X}) = \Pr(\mathrm{e}^{\gamma_i y_i} \geq \mathrm{e}^{\gamma_i t} \mid \boldsymbol{X}) \overset{(a)}{\leq} \mathrm{e}^{-\gamma_i t}\mathbb{E}(\mathrm{e}^{\gamma_i y_i}) \overset{(b)}{=} \mathrm{e}^{-\gamma_i t}\mathrm{e}^{\lambda_i(\mathrm{e}_i^{\gamma}-1)} = \mathrm{e}^{-t\log\frac{t}{\lambda_i}+t-\lambda_i},$$

where (a) is a result of Markov's inequality, and (b) uses the formula for the moment generating function of a Poisson random variable. It is straightforward to see that $\mathrm{e}^{-t\log\frac{t}{\lambda_i}+t-\lambda_i} \leq \mathrm{e}^{-t\log\frac{t}{\|\boldsymbol{\lambda}\|_\infty}+t-\|\boldsymbol{\lambda}\|_\infty}$. Hence,

$$\Pr(\|\boldsymbol{y}\|_\infty \geq t \mid \boldsymbol{X}) \leq n\mathrm{e}^{-t\log\frac{t}{\|\boldsymbol{\lambda}\|_\infty}+t-\|\boldsymbol{\lambda}\|_\infty}.$$

If we set $t = 2\|\boldsymbol{\lambda}\|_\infty \log n$, then (for large enough n such that $\log\log n > 2$) we will have

$$\Pr(\|\boldsymbol{y}\|_\infty \geq 2\|\boldsymbol{\lambda}\|_\infty \log n \mid \boldsymbol{X}) \leq n^{1-\|\boldsymbol{\lambda}\|_\infty \log\log n}.$$

Define the event

$$\mathcal{E}_p = \{1 \leq \|\boldsymbol{\lambda}\|_\infty \leq \sqrt{9\tilde{c}\log n}\},$$

and the set

$$\mathcal{X}_p = \{\boldsymbol{X} \mid 1 \leq \|\boldsymbol{\lambda}\|_\infty \leq \sqrt{9\tilde{c}\log n}\}.$$

First note that

$$\lambda_i = \log(1+\mathrm{e}^{\boldsymbol{x}_i^\top\boldsymbol{\beta}^*}) \leq \log(1+\mathrm{e}^{|\boldsymbol{x}_i^\top\boldsymbol{\beta}^*|}) \leq \log(\mathrm{e}^{|\boldsymbol{x}_i^\top\boldsymbol{\beta}^*|}+\mathrm{e}^{|\boldsymbol{x}_i^\top\boldsymbol{\beta}^*|}) \leq \log 2 + |\boldsymbol{x}_i^\top\boldsymbol{\beta}^*|.$$

If we define the event

$$\tilde{\mathcal{E}}_p = \{\log 2 + \max_i |\boldsymbol{x}_i^\top\boldsymbol{\beta}^*| \leq \sqrt{9\tilde{c}\log n}\} \cap \{\max_i |\boldsymbol{x}_i^\top\boldsymbol{\beta}^*| \geq 1\},$$

then $\tilde{\mathcal{E}}_p \subset \mathcal{E}_p$, leading to

$$\Pr\left(\mathcal{E}_p^c\right) \;\; \leq \;\; \Pr\left(\tilde{\mathcal{E}}_p^c\right) \leq \Pr\left(\{\log 2 + \max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \leq \sqrt{9\tilde{c}\log n}\}^c\right) + \Pr(\{\max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \geq 1\}^c).$$

Note that (for large enough n)

$$\begin{aligned} &\Pr\left(\{\log 2 + \max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \leq \sqrt{9\tilde{c}\log n}\}^c\right) \leq n\Pr\left(|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \geq \sqrt{9\tilde{c}\log n} - 1\right) \\ \leq \;\; & n\Pr\left(|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \geq 2\sqrt{\tilde{c}\log n}\right) \overset{(c)}{\leq} 2n\mathrm{e}^{-2\log n} \leq \frac{2}{n}, \end{aligned} \tag{73}$$

where (c) is a direct consequence of the Gaussian tail bound in Lemma 10 and the fact that $\boldsymbol{x}_i^\top \boldsymbol{\beta}^* \sim \mathrm{N}(0, \tilde{c})$. Furthermore, if $Z \sim N(0, \tilde{c})$, then

$$\Pr(\{\max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \geq 1\}^c) = \left(\Pr(|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| < 1)\right)^n = (P(Z < 1))^n. \tag{74}$$

We now have

$$\begin{aligned} &\Pr(\|\boldsymbol{y}\|_\infty \geq (2\log n)\sqrt{9\tilde{c}\log n}) \\ = \;\; & \int_{\boldsymbol{X}\in\mathcal{X}_p} \Pr(\|\boldsymbol{y}\|_\infty \geq (2\log n)\sqrt{9\tilde{c}\log n} \mid \boldsymbol{X}) dp_{\boldsymbol{X}} + \int_{\boldsymbol{X}\in\mathcal{X}_p^c} \Pr(\|\boldsymbol{y}\|_\infty \geq (2\log n)\sqrt{9\tilde{c}\log n} \mid \boldsymbol{X}) dp_{\boldsymbol{X}} \\ \leq \;\; & \int_{\boldsymbol{X}\in\mathcal{X}_p} \Pr(\|\boldsymbol{y}\|_\infty \geq 2\log n \|\boldsymbol{\lambda}\|_\infty \mid \boldsymbol{X}) dp_{\boldsymbol{X}} + \Pr(\mathcal{E}_p^c) \\ \leq \;\; & \int_{\boldsymbol{X}\in\mathcal{X}_p} n^{1-\|\boldsymbol{\lambda}\|_\infty \log\log n} dp_{\boldsymbol{X}} + \Pr\left(\tilde{\mathcal{E}}_p^c\right) \\ \leq \;\; & \int_{\boldsymbol{X}\in\mathcal{X}_p} n^{1-\log\log n} dp_{\boldsymbol{X}} + \Pr\left(\{\log 2 + \max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \leq \sqrt{9\tilde{c}\log n}\}^c\right) + \Pr(\{\max_i |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \geq 1\}^c) \\ \leq \;\; & n^{1-\log\log n} + \frac{2}{n} + \mathrm{e}^{-n\log\left(\frac{1}{P(Z\leq 1)}\right)}, \end{aligned} \tag{75}$$

where $Z \sim N(0, \tilde{c})$.

A.5.4 Proof of Lemma 4

Note that

$$\begin{aligned} f'(a) &= \frac{a}{\sqrt{1 + \left(\frac{a}{\gamma}\right)^2}} \leq \gamma, \\ f''(a) &= \left(1 + \left(\frac{a}{\gamma}\right)^2\right)^{\frac{3}{2}}, \\ f'''(a) &= \frac{-3a\gamma^3}{(a^2+\gamma^2)^{\frac{5}{2}}}. \end{aligned} \tag{76}$$

Note that $|f'''(a)| \leq \frac{3}{\gamma}$. To see this consider the following two cases:

- Case I, $|a| \leq \gamma$:

$$|f'''(a)| \leq \frac{3\gamma^4}{\gamma^5} \leq \frac{3}{\gamma}.$$

- Case II, $|a| > \gamma$:

$$|f'''(a)| \leq \frac{3|a|\gamma^3}{|a|^5} \leq \frac{3\gamma^3}{a^4|} \leq \frac{3}{\gamma}.$$

Therefore,

$$\begin{aligned}
\|\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta})\|_2 &\leq \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_2 = \sqrt{\sum_i \Big(\ddot{\ell}(\beta_i+\delta_i) - \ddot{\ell}(\beta_i)\Big)^2} \\
&= \sqrt{\sum_i \dddot{\ell}(\beta_i+\epsilon_i)^2(\boldsymbol{x}_i^\top \boldsymbol{\delta})^2} \quad \text{using the mean-value Theorem where} \quad \epsilon_i \in [0, \delta_i] \\
&\leq \frac{3}{\gamma}\sqrt{\boldsymbol{\delta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\delta}} \leq \frac{3}{\gamma}\sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}\|\boldsymbol{\delta}\|_2.
\end{aligned}$$

Finally, based on the inequality above, we have

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq \sup_{t\in[0,1]} \frac{(1-t)\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2\sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq \sqrt{\sigma_{max}(\boldsymbol{X}^\top \boldsymbol{X})}.$$

A.5.5 Proof of Lemma 7

Define

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^n \frac{(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2}{2} + \lambda \sum_{j=1}^p r(\beta_i), \\
\hat{\boldsymbol{\beta}}_{/i} &= \arg\min_{\boldsymbol{\beta}} f_{/i}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{j=1, j\neq i}^n \frac{(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2}{2} + \lambda \sum_{j=1}^p r(\beta_i)
\end{aligned} \tag{77}$$

Furthermore, define $r_{0.5}(\beta) = \frac{\gamma}{2}\beta^2 + (1-\gamma)r^\alpha(\beta)$. Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{I}\sigma_\epsilon^2)$, the optimality conditions yield

$$\begin{aligned}
\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{y} + \lambda \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta}^* + (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon} \\
&\quad + \lambda \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}).
\end{aligned} \tag{78}$$

We bound $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_\infty$ and complete the proof of Lemma 7, by separately bounding the infinity norm of each of the three terms in 78 using Lemma 21, 22 and 23, and defining

$$\tilde{\zeta} = 2\sqrt{c\tilde{c}} + 2\sigma_\epsilon + \lambda\bar{\zeta}, \tag{79}$$

where $\bar{\zeta}$ is introduced in Lemma 23.

Lemma 21. *Under the assumptions of Lemma 7 we have*

$$\Pr(\|(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta}^*\|_\infty > 2\sqrt{c\tilde{c}\log n}) \leq \frac{2}{n}.$$

Proof. First note that

$$(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta}^* = \frac{\lambda\gamma}{2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{\beta}^*. \tag{80}$$

Define $\boldsymbol{D}_i = (\boldsymbol{X}_{/i}^\top \boldsymbol{X}_{/i} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}$. According to the matrix inversion lemma we have

$$\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{\beta}^* = \boldsymbol{x}_i^\top \boldsymbol{D}_i\boldsymbol{\beta}^* - \frac{\boldsymbol{x}_i^\top \boldsymbol{D}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{D}_i\boldsymbol{\beta}^*}{1+\boldsymbol{x}_i^\top \boldsymbol{D}_i \boldsymbol{x}_i} = \frac{\boldsymbol{x}_i^\top \boldsymbol{D}_i\boldsymbol{\beta}^*}{1+\boldsymbol{x}_i^\top \boldsymbol{D}_i \boldsymbol{x}_i}. \tag{81}$$

Note that conditioned on $\boldsymbol{X}_{/i}$ the distribution of $\boldsymbol{x}_i^\top \boldsymbol{D}_i\boldsymbol{\beta}^*$ is a zero mean Gaussian random variable with variance $v_i = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{D}_i\boldsymbol{\beta}^*\|_2^2 \leq \frac{4\rho_{\max}}{\lambda^2\gamma^2}\|\boldsymbol{\beta}^*\|_2^2$. Hence, (81) and the Gaussian tail bound, i.e. Lemma 10, lead to

$$\Pr(|\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{\beta}^*| > t \mid \boldsymbol{X}_{/i}) \leq \Pr(|\boldsymbol{x}_i^\top \boldsymbol{D}_i\boldsymbol{\beta}^*| > t \mid \boldsymbol{X}_{/i}) \leq 2\mathrm{e}^{-\frac{t^2}{2\|\Sigma^{1/2}\boldsymbol{D}_i\boldsymbol{\beta}^*\|_2^2}} \leq 2\mathrm{e}^{-\frac{\lambda^2\gamma^2t^2}{8\rho_{\max}\|\boldsymbol{\beta}^*\|_2^2}}. \tag{82}$$

Hence, by marginalizing $\boldsymbol{X}_{/i}$, we get

$$\Pr(|\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{\beta}^*| > t) \leq 2\mathrm{e}^{-\frac{\lambda^2\gamma^2t^2}{8\rho_{\max}\|\beta\|_2^2}} = 2\mathrm{e}^{-\frac{\lambda^2\gamma^2t^2}{8c\tilde{c}}}.$$

By setting $t = \frac{4\sqrt{c\tilde{c}\log n}}{\lambda\gamma}$ we have

$$\Pr(|\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{\beta}^*| > \frac{4\sqrt{c\tilde{c}\log n}}{\lambda\gamma}) \leq \frac{2}{n^2}.$$

This combined with a union bound and (80) proves that

$$\Pr\left(\|(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty > 2\sqrt{c\tilde{c}\log n}\right) \leq \frac{2}{n}.$$

□

Lemma 22. *If* $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{I}\sigma_\epsilon^2)$*, then*

$$\Pr\left[\|(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon}\|_\infty \geq 2\sigma_\epsilon\sqrt{\log n}\right] \leq \frac{2}{n}.$$

Proof. Note that conditioned on $\boldsymbol{X}$, the distribution of $\boldsymbol{v} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon}$ is multivariate Gaussian with mean zero and covariance matrix $\sigma_\epsilon^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)^2$. We have

$$(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top)^2 = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top - \frac{\lambda\gamma}{2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-2}\boldsymbol{X}^\top. \tag{83}$$

We define $\sigma_i^2(\boldsymbol{X}) = \left(1 - \boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\boldsymbol{x}_i - \frac{\lambda\gamma}{2}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-2}\boldsymbol{x}_i\right)\sigma_\epsilon^2$. Clearly $\sigma_i^2(\boldsymbol{X}) \leq \sigma_\epsilon^2$, hence,

$$\Pr(\|\boldsymbol{v}\|_\infty > t \mid \boldsymbol{X}) \leq \sum_{i=1}^n \Pr(|v_i| > t \mid \boldsymbol{X}) \leq \sum_{i=1}^n 2\mathrm{e}^{-\frac{t^2}{2\sigma_i^2(X)}} = 2n\mathrm{e}^{-\frac{t^2}{2\sigma_\epsilon^2}}. \tag{84}$$

Hence, by setting $t = 2\sigma_\epsilon\sqrt{\log n}$, we have

$$\Pr(\|\boldsymbol{v}\|_\infty > t \mid \boldsymbol{X}) \leq \frac{2}{n}.$$

□

Lemma 23. *Under the assumptions of Lemma 7 we have*

$$\Pr\left[\|\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})\|_\infty > \bar{\zeta}\sqrt{\log n}\right] \leq \frac{6}{n}+2ne^{-n+1}+ne^{-p}, \tag{85}$$

where

$$\begin{aligned}
\bar{\zeta} &= \frac{5c}{\lambda^2\gamma\delta_0}\Big(1+\frac{\alpha(1-\gamma)}{\gamma}\Big)\left(2\sqrt{(c\tilde{c}+\sigma_\epsilon^2)}+\sqrt{\frac{10c(c\tilde{c}+\sigma_\epsilon^2)}{\lambda\gamma}}\right)+\sqrt{20\zeta(c\tilde{c}+\sigma_\epsilon^2)},\\
\zeta &= \frac{2c}{\lambda^3\gamma^2}\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right).
\end{aligned} \tag{86}$$

Proof. Since $f_{/i}(\hat{\boldsymbol{\beta}}_{/i}) \leq f_{/i}(\boldsymbol{0})$, we have

$$2\lambda\gamma\|\hat{\boldsymbol{\beta}}_{/i}\|_2^2 \leq \|\boldsymbol{y}_{/i}\|_2^2. \tag{87}$$

Furthermore, due to $\ddot{r}_{0.5}(\beta) \leq \gamma + \frac{\alpha(1-\gamma)}{2}$, $\dot{r}_{0.5}(0)=0$ (see Lemma 14), and (87), we have

$$\|\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})\|_2^2 \leq \left(\gamma+\frac{\alpha(1-\gamma)}{2}\right)\|\hat{\boldsymbol{\beta}}_{/i}\|_2^2 \leq \left(\frac{1}{2\lambda}+\frac{\alpha(1-\gamma)}{4\lambda\gamma}\right)\|\boldsymbol{y}_{/i}\|_2^2. \tag{88}$$

The first order optimality condition yields

$$\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{/i}-\hat{\boldsymbol{\beta}})+\lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i})-\lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}) = -\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}).$$

Since the minimum eigenvalue of the Hessian of $\boldsymbol{r}(\boldsymbol{\beta})$ is 2γ, therefore the minimum eigenvalue of $\boldsymbol{X}^\top\boldsymbol{X}+\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\boldsymbol{\beta})]$ (for all $\boldsymbol{\beta}$) is greater than $2\lambda\gamma$, leading to

$$\|\hat{\boldsymbol{\beta}}_{/i}-\hat{\boldsymbol{\beta}}\|_2 \leq \frac{|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}|}{2\lambda\gamma}\|\boldsymbol{x}_i\|_2.$$

This together with $\ddot{r}_{0.5}(\beta) \leq \gamma+\frac{\alpha(1-\gamma)}{2}$ yields

$$\|\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})-\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})\|_2 \leq \left(\gamma+\frac{\alpha(1-\gamma)}{2}\right)\|\hat{\boldsymbol{\beta}}_{/i}-\hat{\boldsymbol{\beta}}\|_2 \leq \left(\frac{1}{2\lambda}+\frac{\alpha(1-\gamma)}{4\lambda\gamma}\right)|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}|\|\boldsymbol{x}_i\|_2.$$

Define $\boldsymbol{D}_i = (\boldsymbol{X}_{/i}^\top\boldsymbol{X}_{/i}+\frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}$. According to the matrix inversion lemma we have

$$\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X}+\frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i^\top\boldsymbol{D}_i\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) - \frac{\boldsymbol{x}_i^\top\boldsymbol{D}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{D}_i\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})}{1+\boldsymbol{x}_i^\top\boldsymbol{D}_i\boldsymbol{x}_i} = \frac{\boldsymbol{x}_i^\top\boldsymbol{D}_i\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})}{1+\boldsymbol{x}_i^\top\boldsymbol{D}_i\boldsymbol{x}_i}. \tag{89}$$

Furthermore, we have

$$|\boldsymbol{x}_i^\top\boldsymbol{D}_i\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})| \leq |\boldsymbol{x}_i^\top\boldsymbol{D}_i\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})| + |\boldsymbol{x}_i^\top\boldsymbol{D}_i(\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})-\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i}))|. \tag{90}$$

First note that, since the maximum eigenvalue of $\boldsymbol{D}_i$ is $\frac{\lambda\gamma}{2}$ we have

$$\begin{aligned}
&|\boldsymbol{x}_i^\top \boldsymbol{D}_i(\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) - \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i}))| \\
&\leq \frac{2}{\lambda\gamma}\|\boldsymbol{x}_i\|_2\|\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) - \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})\|_2 \leq \frac{1}{\lambda^2\gamma}\|\boldsymbol{x}_i\|_2^2\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right)|y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}| \\
&\leq \frac{1}{\lambda^2\gamma}\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right)\|\boldsymbol{x}_i\|_2^2(|y_i| + |\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}|). \qquad (91)
\end{aligned}$$

Furthermore, we have

1. Due to Lemma 11, $\Pr(\|\boldsymbol{x}_i\|_2^2 > 5p\rho_{\max}) \leq \mathrm{e}^{-p}$, leading to

$$\Pr(\|\boldsymbol{x}_i\|_2^2 > \frac{5c}{\delta_0}) \leq \mathrm{e}^{-p}. \qquad (92)$$

2. Note that $y_i \sim N(0, \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma_\epsilon^2)$. Furthermore, $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma_\epsilon^2 \leq \rho_{\max}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \sigma_\epsilon^2 \leq c\tilde{c} + \sigma_\epsilon^2$. Hence, using the Gaussian tail bound, i.e. Lemma 10, we have

$$\Pr(|y_i| > t) \leq 2\mathrm{e}^{-\frac{t^2}{2(c\tilde{c}+\sigma_\epsilon^2)}}. \qquad (93)$$

Hence,

$$\Pr(|y_i| > 2\sqrt{(c\tilde{c}+\sigma_\epsilon^2)\log n}) \leq \frac{2}{n^2}. \qquad (94)$$

3. Given $\boldsymbol{X}_{/i}, \boldsymbol{y}_{/i}$, the distribution of $\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}$ is $N(0, \hat{\boldsymbol{\beta}}_{/i}^\top\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_{/i})$. Furthermore, $\hat{\boldsymbol{\beta}}_{/i}^\top\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_{/i} \leq \frac{c\hat{\boldsymbol{\beta}}_{/i}^\top\hat{\boldsymbol{\beta}}_{/i}}{n} \leq \frac{c\|\boldsymbol{y}_{/i}\|_2^2}{2n\lambda\gamma}$, where the last inequality is due to (87). Hence, we have

$$\Pr(|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}| > t|\boldsymbol{X}_{/i}, \boldsymbol{y}_{/i}) \leq 2\mathrm{e}^{-\frac{n\lambda\gamma t^2}{c\|\boldsymbol{y}_{/i}\|_2^2}}. \qquad (95)$$

According to Lemma 11 since $y_i \overset{i.i.d.}{\sim} N(0, \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma_\epsilon^2)$, and $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma_\epsilon^2 \leq c\tilde{c} + \sigma_\epsilon^2$, we have

$$\Pr(\|\boldsymbol{y}_{/i}\|_2^2 > 5(n-1)(c\tilde{c}+\sigma_\epsilon^2)) \leq \mathrm{e}^{-n+1}. \qquad (96)$$

Let B denote the event that $\|\boldsymbol{y}_{/i}\|_2^2 \leq 5(n-1)(c\tilde{c}+\sigma_\epsilon^2)$. Then, combining (95) and (96), we have

$$\Pr(|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}| > t) \leq \Pr(|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}| > t \mid B) + \Pr(B^c) \leq 2\mathrm{e}^{-\frac{\lambda\gamma t^2}{5c(c\tilde{c}+\sigma_\epsilon^2)}} + \mathrm{e}^{-n+1}.$$

Hence,

$$\Pr\left[|\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{/i}| > \sqrt{\frac{10c(c\tilde{c}+\sigma_\epsilon^2)}{\lambda\gamma}\log n}\right] \leq \frac{2}{n^2} + \mathrm{e}^{-n+1}. \qquad (97)$$

By combining (92), (94), (97), and (91) we conclude that

$$\begin{aligned}
&\Pr\left[|\boldsymbol{x}_i^\top D_i(\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}) - \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i}))| > \frac{5c}{\lambda^2\gamma\delta_0}\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right)\left(2\sqrt{(c\tilde{c}+\sigma_\epsilon^2)} + \sqrt{\frac{10c(c\tilde{c}+\sigma_\epsilon^2)}{\lambda\gamma}}\right)\sqrt{\log n}\right] \\
&\leq \frac{4}{n^2} + \mathrm{e}^{-n+1} + \mathrm{e}^{-p}. \qquad (98)
\end{aligned}$$

Next, we compute an upper bound on $|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})|$. Since $\boldsymbol{x}_i$ is independent of $\boldsymbol{y}_{/i}$ and $\boldsymbol{X}_{/i}$, we conclude that given $\boldsymbol{X}_{/i}$ and $\boldsymbol{y}_{/i}$, $\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})$ is a Gaussian random variable with mean zero and variance

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})\|_2^2 \leq \frac{4\rho_{\max}}{\lambda^2\gamma^2}\|\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})\|_2^2 \leq \frac{2\rho_{\max}}{\lambda^3\gamma^2}\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right)\|\boldsymbol{y}_{/i}\|_2^2 = \frac{\zeta\|\boldsymbol{y}_{/i}\|_2^2}{n},$$

where $\zeta = \frac{2c}{\lambda^3\gamma^2}\left(1+\frac{\alpha(1-\gamma)}{2\gamma}\right)$, and the second inequality is due to (88). Hence,

$$\mathbb{P}(|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})| > t \mid \boldsymbol{X}_{/i}, \boldsymbol{y}_{/i}) \leq 2\mathrm{e}^{-\frac{nt^2}{2\zeta\|\boldsymbol{y}_{/i}\|_2^2}}.$$

Considering the event B of $\|\boldsymbol{y}_{/i}\|_2^2 \leq 5(n-1)(c\tilde{c}+\sigma_\epsilon^2)$, we have

$$\Pr\left(|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})| > t\right) \leq \Pr\left(|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})| > t \big| B\right) + \Pr(B^c) \leq 2\mathrm{e}^{-\frac{t^2}{10\zeta(c\tilde{c}+\sigma_\epsilon^2)}} + \mathrm{e}^{-n+1}. \tag{99}$$

Hence,

$$\Pr\left(|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}_{/i})| > \sqrt{20\zeta(c\tilde{c}+\sigma_\epsilon^2)\log n}\right) \leq \frac{2}{n^2} + \mathrm{e}^{-n+1}. \tag{100}$$

By combining (89), (90), (98), and 100 we conclude that if

$$\bar{\zeta} = \frac{5c}{\lambda^2\gamma\delta_0}\left(1+\frac{\alpha(1-\gamma)}{\gamma}\right)\left(2\sqrt{(c\tilde{c}+\sigma_\epsilon^2)} + \sqrt{\frac{10c(c\tilde{c}+\sigma_\epsilon^2)}{\lambda\gamma}}\right) + \sqrt{20\zeta(c\tilde{c}+\sigma_\epsilon^2)},$$

then

$$\begin{aligned}
&\Pr\left[|\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})| > \bar{\zeta}\sqrt{\log n}\right] \\
\leq\ & \Pr\left[|\boldsymbol{x}_i^\top \boldsymbol{D}_i \dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}})| \geq \bar{\zeta}\sqrt{\log n}\right] \leq \frac{6}{n^2} + 2\mathrm{e}^{-n+1} + \mathrm{e}^{-p}.
\end{aligned} \tag{101}$$

Hence,

$$\Pr\left[\|\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda\gamma}{2}\boldsymbol{I})^{-1}\dot{\boldsymbol{r}}_{0.5}(\hat{\boldsymbol{\beta}}\|_\infty| > \bar{\zeta}\sqrt{\log n}\right] \leq \frac{6}{n} + 2n\mathrm{e}^{-n+1} + n\mathrm{e}^{-p}.$$

□

A.5.6 Proof of Lemma 2

Since $r^\alpha(z) = \alpha^{-1}\log\left(e^{\alpha z} + e^{-\alpha z} + 2\right)$, we have $e^{\alpha r^\alpha(z)} = e^{\alpha z} + e^{-\alpha z} + 2$, and because of Lemma 13, $e^{\alpha z} + e^{-\alpha z} + 2 \geq e^{\alpha|z|}$. Moreover, $\dddot{r}^\alpha(z) = 2\alpha^2(e^{-\alpha z} - e^{\alpha z})/(e^{\alpha z} + e^{-\alpha z} + 2)^2 \leq 2\alpha^2(e^{-\alpha z} - e^{\alpha z})/e^{2\alpha|z|} \leq 4\alpha^2 e^{-\alpha|z|}$. The next step is

$$\begin{aligned}
\frac{\|\ddot{\boldsymbol{r}}^\alpha(\boldsymbol{\beta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{r}}^\alpha(\boldsymbol{\beta})\|_2}{\|\boldsymbol{\delta}\|_2} &= \frac{\sqrt{\sum_i\left(\ddot{r}^\alpha(\beta_i+\delta_i) - \ddot{r}^\alpha(\beta_i)\right)^2}}{\|\boldsymbol{\delta}\|_2} \\
&= \frac{\sqrt{\sum_i \dddot{r}^\alpha(\beta_i+\epsilon_i)^2\delta_i^2}}{\|\boldsymbol{\delta}\|_2} \quad \text{using the mean-value Theorem where} \quad \epsilon_i \in [0,\delta_i] \\
&= \frac{4\alpha^2\sqrt{\sum_i \delta_i^2 e^{-2\alpha|\beta_i+\epsilon_i|}}}{\|\boldsymbol{\delta}\|_2} \\
&\leq 4\alpha^2.
\end{aligned}$$

A.6 Proof of Theorem 3

We first present lemmas necessary for the proof of Theorem 3. Lemmas are proved in section A.7.

Lemma 24. *Let $\boldsymbol{X} \in \mathrm{R}^{m\times p}$ be a matrix with $m > p = rank(\boldsymbol{X})$. Moreover, let $\boldsymbol{D} \in \mathrm{R}^{m\times m}$ and $\boldsymbol{D}+\boldsymbol{\Gamma} \in \mathrm{R}^{m\times m}$ be diagonal matrices with positive elements, then*

$$\begin{aligned}
&\left(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1} \\
&\quad = \boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma}\boldsymbol{X}\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1},
\end{aligned}$$

where $\boldsymbol{A} \triangleq \boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}$.

Lemma 25. *Assume that $\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}$ are positive definite, and define:*

$$\boldsymbol{\Gamma} \triangleq \mathrm{diag}(\boldsymbol{\gamma}), \tag{102}$$

$$\bar{\omega}_{\max} \triangleq \sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right), \tag{103}$$

$$\nu_{\min} \triangleq \sigma_{\min}\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right), \tag{104}$$

$$\boldsymbol{A} \triangleq \boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}. \tag{105}$$

Then,

$$\left|\boldsymbol{z}^\top\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{z} - \boldsymbol{z}^\top\left(\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}\right)^{-1}\boldsymbol{z}\right| \leq \left(\|\boldsymbol{\gamma}\|_2 + \left(\frac{\bar{\omega}_{\max}}{\nu_{\min}}\right)\|\boldsymbol{\gamma}\|_4^2\right)\left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right\|_4^2. \tag{106}$$

Lemma 26. *Let S denote the event that (22), (23), (24), and (25) hold. If S holds, then*

$$\left|\boldsymbol{x}_{\boldsymbol{i}}^\top\boldsymbol{\Delta}_{/i}^* - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}}\right| \leq \bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right),$$

where

$$\begin{aligned}
\boldsymbol{\Delta}_{/i}^* &\triangleq \hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}, \\
\boldsymbol{H} &\triangleq \boldsymbol{X}\left(\lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})], \\
\boldsymbol{J}_{/i} &\triangleq \lambda\,\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i})] + \boldsymbol{X}_{/i}^\top\mathrm{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i})]\boldsymbol{X}_{/i}, \\
\bar{C}_i &\triangleq 4\left\|\boldsymbol{x}_{\boldsymbol{i}}\right\|_2\left(\frac{c_1^2(n)c_2(n)}{\nu}\right)\left(1 + \frac{2c_1(n)c_2(n)(1+\omega_{\max,i})}{\nu^2}\left\|\boldsymbol{x}_{\boldsymbol{i}}\right\|_2\right),
\end{aligned}$$

and $c_1(n)$ and $c_2(n)$ are defined in Assumption 6, ν is defined in Assumption 7, and $\omega_{\max,i} \triangleq \sigma_{\max}\left(\boldsymbol{X}_{/i}\boldsymbol{X}_{/i}^\top\right)$.

Lemma 27. *Let $\boldsymbol{x} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ with $\rho_{\max} \triangleq \sigma_{\max}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathrm{R}^{p\times p}$ then*

$$\Pr\left[\|\boldsymbol{x}\|_4^2 > 2(1+c)\rho_{\max}\sqrt{p}\log p\right] \leq \frac{2}{p^c}. \tag{107}$$

Moreover, if

$$\omega_{\max} \triangleq \sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right), \tag{108}$$

$$\nu_{\min} \triangleq \sigma_{\min}\left(\boldsymbol{J}\right), \tag{109}$$

where $\boldsymbol{x}$ *is independent of the symmetric matrix* $\boldsymbol{J} \in \mathrm{R}^{p\times p}$ *and* $\boldsymbol{X} \in \mathrm{R}^{m\times p}$*, then*

$$\Pr\left[\left\|\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu_{\min}^2}\right)\sqrt{p}\log p\right] < \frac{2}{p^c}, \tag{110}$$

$$\Pr\left[\left\|\boldsymbol{X}\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu_{\min}^2}\omega_{\max}\right)\sqrt{m}\log m\right] < \frac{2}{m^c}. \tag{111}$$

Proof of Theorem 3. Let S denote the event that (22), (23), (24), and (25) hold. Furthermore, define the following events:

$$G \triangleq \left\{\max_{1\le i\le n}\left|\boldsymbol{x}_{\boldsymbol{i}}^\top\boldsymbol{\Delta}^*_{/i} - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}}\right| > C\frac{\log p}{\sqrt{p}}\right\}, \tag{112}$$

$$E_i \triangleq \left\{\left|\boldsymbol{x}_{\boldsymbol{i}}^\top\boldsymbol{\Delta}^*_{/i} - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}}\right| > C\frac{\log p}{\sqrt{p}}\right\}, \tag{113}$$

$$\tilde{E}_i \triangleq \left\{\bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right) > C\frac{\log p}{\sqrt{p}}\right\}, \tag{114}$$

$$F_i \triangleq \left\{\bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right) > \bar{C}_iC_i\sqrt{p}\log p\right\}, \tag{115}$$

$$K_i \triangleq \left\{\frac{C}{\sqrt{p}} \ge \bar{C}_iC_i\sqrt{p}\right\}, \tag{116}$$

$$W_i \triangleq \left\{\|\boldsymbol{x}_{\boldsymbol{i}}\|_2^2 > 5p\rho_{\max}\right\} \cup \left\{\omega_{\max} > \left(\sqrt{n}+3\sqrt{p}\right)^2\rho_{\max}\right\}, \tag{117}$$

where C in (113) is a positive constant (defined later in (123)), and

$$C_i \triangleq 2(1+c)\left(\frac{\rho_{\max}}{\nu^2}\right)\left(1+\omega_{\max}\sqrt{\frac{n-1}{p}}\frac{\log(n-1)}{\log p}\right), \tag{118}$$

$$\bar{C}_i \triangleq 4\|\boldsymbol{x}_{\boldsymbol{i}}\|_2\left(\frac{c_1^2(n)c_2(n)}{\nu}\right)\left(1+\frac{2c_1(n)c_2(n)(1+\omega_{\max})}{\nu^2}\|\boldsymbol{x}_{\boldsymbol{i}}\|_2\right), \tag{119}$$

$$\omega_{\max} \triangleq \sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right). \tag{120}$$

The variable c in (118) is later set to 3, but for now all we need to know is that it is a positive constant. Due to Lemma 26, if the event S holds, then for every i we have

$$\bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right) \ge \left|\boldsymbol{x}_{\boldsymbol{i}}^\top\boldsymbol{\Delta}^*_{/i} - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}}\right|.$$

Since $\Pr[S^c] \le q_n + \tilde{q}_n$, we have

$$\begin{aligned}\Pr[G] &\le \Pr[G|S] + \Pr[S^c] \le \Pr\left[\max_{1\le i\le n}\bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right) > C\frac{\log p}{\sqrt{p}} \mid S\right] + q_n + \tilde{q}_n \\ &\le \frac{1}{1-q_n-\tilde{q}_n}\Pr\left[\max_{1\le i\le n}\bar{C}_i\left(\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{\boldsymbol{i}}\right\|_4^2\right) > C\frac{\log p}{\sqrt{p}}\right] + q_n + \tilde{q}_n \\ &\le \frac{1}{1-q_n-\tilde{q}_n}\sum_{i=1}^n \Pr[\tilde{E}_i] + q_n + \tilde{q}_n. \end{aligned} \tag{121}$$

Hence, we now obtain an upper bound for $\Pr[\tilde{E}_i]$;

$$
\begin{aligned}
\Pr[\tilde{E}_i] &\leq \Pr[\tilde{E}_i|K_i] + \Pr[K_i^c], \\
&\leq \Pr[F_i|K_i] + \Pr[K_i^c] \leq \frac{\Pr(F_i)}{\Pr(K_i)} + \Pr[K_i^c] \\
&\leq \frac{\Pr\left[\left\|\boldsymbol{X}_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x_i}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu^2}\omega_{\max}\right)\sqrt{n-1}\log(n-1)\right]}{\Pr(K_i)} \\
&+ \frac{\Pr\left[\left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x_i}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu^2}\right)\sqrt{p}\log p\right]}{\Pr(K_i)} + \Pr[K_i^c] \\
&\overset{1}{\leq} \left(\frac{2}{(n-1)^c} + \frac{2}{p^c}\right)\frac{1}{\Pr(K_i)} + \Pr[K_i^c],
\end{aligned} \tag{122}
$$

where $\overset{1}{\leq}$ is due to Inequality (111) from Lemma 27. To bound $\Pr[K_i^c]$ we define

$$
\begin{aligned}
C &\triangleq 32\sqrt{5}\left(\frac{c_1^2(n)c_2(n)(p\rho_{\max})^{3/2}}{\nu^3}\right)\left(1 + \left(\sqrt{\frac{n}{p}} + 3\right)^2 p\rho_{\max}\sqrt{\frac{n-1}{p}\frac{\log(n-1)}{\log p}}\right) \\
&\times \left(1 + \frac{2c_1(n)c_2(n)\sqrt{5}\left(1 + \left(\sqrt{\frac{n}{p}}+3\right)^2 p\rho_{\max}\right)\sqrt{p\rho_{\max}}}{\nu^2}\right)
\end{aligned} \tag{123}
$$

obtained by setting $c = 3$, and computing $p\bar{C}_iC_i$ after putting $\sqrt{5p\rho_{\max}}$ and $\left(\sqrt{n}+3\sqrt{p}\right)^2\rho_{\max}$, bounds in event W_i, into $\|\boldsymbol{x_i}\|_2$ and $\omega_{\max}$, respectively. Next,

$$
\begin{aligned}
\Pr[K_i^c] &= \Pr\left[\frac{C}{p} < \bar{C}_iC_i\right] \leq \Pr\left[C < p\bar{C}_iC_i \,\middle|\, W_i^c\right] + \Pr[W_i] \\
&= \Pr\left[C < C\right] + \Pr[W_i] = \Pr[W_i].
\end{aligned}
$$

The term $\Pr[W_i]$ is exponentially small because $\boldsymbol{x_i}$ is $\mathrm{N}(0, \boldsymbol{\Sigma})$ with $\rho_{\max} = \sigma_{\max}(\boldsymbol{\Sigma})$, leading to

$$
\Pr[W_i] \leq \Pr\left[\|\boldsymbol{x_i}\|_2^2 > 5p\rho_{\max}\right] + \Pr\left[\sigma_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) > \left(\sqrt{n}+3\sqrt{p}\right)^2\rho_{\max}\right] \leq 2e^{-p}, \tag{124}
$$

due to Lemma 11 and Lemma 12. In summary, since for $p \geq 1$ we have $\frac{1}{1-\mathrm{e}^{-p}} < 2$, for $c = 3$ we obtain

$$
\Pr[\tilde{E}_i] \leq \frac{4}{(n-1)^3} + \frac{4}{p^3} + 2e^{-p}.
$$

This combined with (121) leas to

$$
\begin{aligned}
\Pr\left[\max_{1\leq i\leq n}\left|\boldsymbol{x_i}^\top\boldsymbol{\Delta}_{/i}^* - \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}}\right| > C\frac{\log p}{\sqrt{p}}\right] &\leq \left(\frac{4n}{(n-1)^3} + \frac{4n}{p^3} + 2ne^{-p}\right)\frac{1}{1-q_n-\tilde{q}_n} + q_n + \tilde{q}_n \\
&\leq \frac{8n}{(n-1)^3} + \frac{8n}{p^3} + 4ne^{-p} + q_n + \tilde{q}_n,
\end{aligned} \tag{125}
$$

where the last inequality is due to the assumption that $q_n + \tilde{q}_n \leq 0.5$. Hence, Inequality (26) in Theorem 3 follows. Note that in the presentation of Theorem 3, we replaced respectively $32\sqrt{5}$ and $2\sqrt{3}$ with the upper-bounds 72 and 5, and we replaced $\sqrt{\frac{n-1}{p}\frac{\log(n-1)}{\log p}}$ with the upper bound $\sqrt{\frac{n}{p}\frac{\log n}{\log p}}$. We also used δ_0 to denote n/p.

□

A.7 Proofs of lemmas 24, 25, 26, 27 and 12

Proof of Lemma 24. Let $Q \triangleq \{i : \Gamma_{ii} \neq 0\}$. Moreover, let $\boldsymbol{X}_{Q,:} \in \mathrm{R}^{|Q|\times p}$ stand for the sub-matrix of $\boldsymbol{X}$ restricted to the rows indexed by Q, and let $\tilde{\boldsymbol{\Gamma}} \in \mathrm{R}^{|Q|\times|Q|}$ be the diagonal matrix with the diagonal elements of $\boldsymbol{\Gamma}$ indexed by Q. Then, $\boldsymbol{X}^\top \boldsymbol{\Gamma} \boldsymbol{X} = \boldsymbol{X}_{Q,:}^\top \tilde{\boldsymbol{\Gamma}} \boldsymbol{X}_{Q,:}$, and in turn, the Woodbury inversion lemma yields

$$
\begin{aligned}
(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X})^{-1} &= (\boldsymbol{A} + \boldsymbol{X}_{Q,:}^\top \tilde{\boldsymbol{\Gamma}} \boldsymbol{X}_{Q,:})^{-1} \\
&= \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}_{Q,:}^\top (\tilde{\boldsymbol{\Gamma}}^{-1} + \boldsymbol{X}_{Q,:}\boldsymbol{A}^{-1}\boldsymbol{X}_{Q,:}^\top)^{-1} \boldsymbol{X}_{Q,:}\boldsymbol{A}^{-1}. \qquad (126)
\end{aligned}
$$

Using the Woodbury lemma again we obtain

$$
\begin{aligned}
(\tilde{\boldsymbol{\Gamma}}^{-1} + \boldsymbol{X}_{Q,:}\boldsymbol{A}^{-1}\boldsymbol{X}_{Q,:}^\top)^{-1} &= \tilde{\boldsymbol{\Gamma}} - \tilde{\boldsymbol{\Gamma}}\boldsymbol{X}_{Q,:}(\boldsymbol{A} + \boldsymbol{X}_{Q,:}^\top \tilde{\boldsymbol{\Gamma}}\boldsymbol{X}_{Q,:})^{-1}\boldsymbol{X}_{Q,:}^\top \tilde{\boldsymbol{\Gamma}} \\
&= \tilde{\boldsymbol{\Gamma}} - \tilde{\boldsymbol{\Gamma}}\boldsymbol{X}_{Q,:}(\boldsymbol{X}^\top(\boldsymbol{\Gamma} + \boldsymbol{D})\boldsymbol{X})^{-1}\boldsymbol{X}_{Q,:}^\top \tilde{\boldsymbol{\Gamma}}. \qquad (127)
\end{aligned}
$$

Hence, by using (126) and (127) we have

$$
(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X})^{-1} = \boldsymbol{A}^{-1}\boldsymbol{X}^\top \left(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{X}(\boldsymbol{X}^\top(\boldsymbol{\Gamma} + \boldsymbol{D})\boldsymbol{X})^{-1}\boldsymbol{X}^\top \Gamma\right)\boldsymbol{X}\boldsymbol{A}^{-1}.
$$

□

Proof of Lemma 25. Let $\boldsymbol{A} \triangleq \boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}$, then

$$
\begin{aligned}
&\left|\boldsymbol{z}^\top \left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{z} - \boldsymbol{z}^\top \left(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}\right)^{-1}\boldsymbol{z}\right| \\
&\overset{1}{=} \left|\boldsymbol{z}^\top \left(\boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1}\right)\boldsymbol{z}\right| \\
&\overset{2}{\leq} \left|\boldsymbol{z}^\top \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right| + \boldsymbol{z}^\top \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z} \\
&\overset{3}{\leq} \|\gamma\|_2 \left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right\|_4^2 + \boldsymbol{z}^\top \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\left(\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z} \\
&\overset{4}{\leq} \|\gamma\|_2 \left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right\|_4^2 + \left(\frac{\bar{\omega}_{\max}}{\nu_{\min}}\right)\boldsymbol{z}^\top \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{\Gamma}^2\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z} \\
&\overset{5}{\leq} \|\gamma\|_2 \left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right\|_4^2 + \left(\frac{\bar{\omega}_{\max}}{\nu_{\min}}\right)\|\gamma\|_4^2 \left\|\boldsymbol{X}\boldsymbol{A}^{-1}z\right\|_4^2 \\
&= \left(\|\gamma\|_2 + \left(\frac{\bar{\omega}_{\max}}{\nu_{\min}}\right)\|\gamma\|_4^2\right)\left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{z}\right\|_4^2, \qquad (128)
\end{aligned}
$$

where $\overset{1}{=}$ is due to Lemma 24, $\overset{2}{\leq}$ is due to the triangle inequality, and the fact that $\boldsymbol{X}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{X}$ is positive definite, and $\overset{3}{\leq}$ and $\overset{5}{\leq}$ are due to Cauchy-Schwartz inequality:

$$
\boldsymbol{x}^\top \mathrm{diag}[\gamma]\boldsymbol{x} = \sum_{i=1}^{n} x_i^2\gamma_i \leq \sqrt{\|\boldsymbol{x}\|_4^4\|\gamma\|_2^2} = \|\boldsymbol{x}\|_4^2\|\gamma\|_2.
$$

Finally, $\overset{4}{\leq}$ is due to (108) and (109). □

Proof of Lemma 26. Define the approximate leave-i-out perturbation vector as

$$
\hat{\boldsymbol{\Delta}}_{/i} \triangleq \dot{\ell}_i(\hat{\boldsymbol{\beta}})[\boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{\Delta}^*_{/i})]^{-1}\boldsymbol{x}_{\boldsymbol{i}}, \qquad (129)
$$

where the exact leave-i-out perturbation vector is given by

$$\mathbf{\Delta}^*_{/i} \quad \triangleq \quad \hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}. \tag{130}$$

Woodbury lemma yields:

$$\begin{aligned}
\boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\mathbf{\Delta}}_{/i} &= \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}^\top \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i} - \mathbf{\Delta}^*_{/i})] + \boldsymbol{X}_{/i}^\top \mathrm{diag}[\ddot{\ell}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - \mathbf{\Delta}^*_{/i})]\boldsymbol{X}_{/i}\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}} \\
&= \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}^\top \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}_{/i}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_{/i}\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}} \\
&= \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}^\top \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X} - \boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{x}_{\boldsymbol{i}}^\top \ddot{\ell}_i(\hat{\boldsymbol{\beta}})\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}} \\
&= \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right) \frac{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}^\top \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}}}{1 - \ddot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}^\top \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}}} \\
&= \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right) \frac{H_{ii}}{1 - H_{ii}},
\end{aligned} \tag{131}$$

where $\boldsymbol{H} \triangleq \boldsymbol{X} \left(\lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})] + \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]$. Define

$$\boldsymbol{f}_{/i}(\boldsymbol{\theta}) \quad \triangleq \quad \lambda \dot{\boldsymbol{r}}(\boldsymbol{\theta}) + \boldsymbol{X}_{/i}^\top \dot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\theta}). \tag{132}$$

The leave-one-out estimate, $\hat{\boldsymbol{\beta}}_{/i} = \hat{\boldsymbol{\beta}} + \mathbf{\Delta}^*_{/i}$, satisfies $\boldsymbol{f}_{/i}(\mathbf{\Delta}^*_{/i}) = 0$. The multivariate mean-value Theorem yields

$$0 \quad = \quad \boldsymbol{f}_{/i}(\hat{\boldsymbol{\beta}} + \mathbf{\Delta}^*_{/i}) = \boldsymbol{f}_{/i}(\hat{\boldsymbol{\beta}}) + \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}} + t\mathbf{\Delta}^*_{/i})dt\right) \mathbf{\Delta}^*_{/i} \tag{133}$$

where the Jacobean is

$$\boldsymbol{J}_{/i}(\boldsymbol{\theta}) \quad = \quad \lambda\, \mathrm{diag}[\ddot{\boldsymbol{r}}(\boldsymbol{\theta})] + \boldsymbol{X}_{/i}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\theta})]\boldsymbol{X}_{/i}. \tag{134}$$

Moreover, $\hat{\boldsymbol{\beta}}$ satisfies

$$0 \quad = \quad \lambda \dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}) + \boldsymbol{X}^\top \dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{f}_{/i}(\hat{\boldsymbol{\beta}}) + \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}}.$$

We get

$$\dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\boldsymbol{i}} \quad = \quad -\left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}} + t\mathbf{\Delta}^*_{/i})dt\right) \mathbf{\Delta}^*_{/i},$$

so that

$$\mathbf{\Delta}^*_{/i} \quad = \quad -\dot{\ell}_i(\hat{\boldsymbol{\beta}}) \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}} + t\mathbf{\Delta}^*_{/i})dt\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}}, \tag{135}$$

leading to the following inequality

$$\left\|\mathbf{\Delta}^*_{/i}\right\|_2 \leq \left(\frac{|\dot{\ell}_i(\hat{\boldsymbol{\beta}})|}{\nu}\right) \|\boldsymbol{x}_{\boldsymbol{i}}\|_2\,, \tag{136}$$

as a consequence of Assumption 7. Next, we look at the part of $\boldsymbol{\Delta}^*_{i,\lambda}$ dependent on $\boldsymbol{x_i}$, so we rewrite (135) as

$$\boldsymbol{\Delta}^*_{/i} \quad = \quad -\dot{\ell}_i(\hat{\boldsymbol{\beta}}) \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - (1-t)\boldsymbol{\Delta}^*_{/i}) dt \right)^{-1} \boldsymbol{x_i}. \tag{137}$$

Let us rewrite the Jacobean in a more compact form:

$$\boldsymbol{J}_{/i}(\boldsymbol{\theta}) = \bar{\boldsymbol{X}}_{/i}^\top \boldsymbol{D}_{/i}(\boldsymbol{\theta}) \bar{\boldsymbol{X}}_{/i}, \tag{138}$$

where

$$\bar{\boldsymbol{X}}_{/i} \triangleq \begin{bmatrix} \boldsymbol{X}_{/i} \\ \boldsymbol{I} \end{bmatrix} \in \mathrm{R}^{(n-1+p)\times p}, \qquad \boldsymbol{D}_{/i}(\boldsymbol{\theta}) \triangleq \mathrm{diag} \begin{bmatrix} \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\theta}) \\ \lambda \ddot{\boldsymbol{r}}(\boldsymbol{\theta}) \end{bmatrix} \in \mathrm{R}^{(n-1+p)\times(n-1+p)}. \tag{139}$$

Define

$$\boldsymbol{\gamma}_{\boldsymbol{\delta}/i}(\boldsymbol{\theta}) \quad \triangleq \quad \begin{bmatrix} \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\theta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\theta}) \\ \lambda(\ddot{\boldsymbol{r}}(\boldsymbol{\theta}+\boldsymbol{\delta}) - \ddot{\boldsymbol{r}}(\boldsymbol{\theta})) \end{bmatrix} \tag{140}$$

so that

$$\boldsymbol{J}_{/i}(\boldsymbol{\theta}+\boldsymbol{\delta}) \quad = \quad \boldsymbol{J}_{/i}(\boldsymbol{\theta}) + \bar{\boldsymbol{X}}_{/i}^\top \mathrm{diag}\left[\boldsymbol{\gamma}_{\boldsymbol{\delta}/i}(\boldsymbol{\theta})\right] \bar{\boldsymbol{X}}_{/i}, \tag{141}$$

Note that $\boldsymbol{J}_{/i}(\boldsymbol{\theta}+\boldsymbol{\delta})$ is positive definite for all $t \in [0,1]$, $\boldsymbol{\theta} = \hat{\boldsymbol{\beta}}_{/i}$ and $\boldsymbol{\delta} = -(1-t)\boldsymbol{\Delta}^*_{/i}$, due to Assumption 7. The last steps of the proof are as follows:

$$
\begin{aligned}
\left|\boldsymbol{x_i}^\top \boldsymbol{\Delta}^*_{/i} - \boldsymbol{x_i}^\top \hat{\boldsymbol{\Delta}}_{/i}\right| &= |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left|\boldsymbol{x_i}^\top \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - (1-t)\boldsymbol{\Delta}^*_{/i})dt\right)^{-1} \boldsymbol{x_i} - \boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{\Delta}^*_{/i})\boldsymbol{x_i}\right| \\
&\leq |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left|\boldsymbol{x_i}^\top \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - (1-t)\boldsymbol{\Delta}^*_{/i})dt\right)^{-1} \boldsymbol{x_i} - \boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right| \\
&+ |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left|\boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i} - \boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{\Delta}^*_{/i})\boldsymbol{x_i}\right| \\
&\overset{0}{\leq} |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left|\boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i} - \boldsymbol{x_i}^\top \left(\boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i}) + \bar{\boldsymbol{X}}^\top_{/i} \operatorname{diag}\left[\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right] \bar{\boldsymbol{X}}_{/i}\right)^{-1} \boldsymbol{x_i}\right| \\
&+ |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left|\boldsymbol{x_i}^\top \boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i} - \boldsymbol{x_i}^\top \left(\boldsymbol{J}_{/i}(\hat{\boldsymbol{\beta}}_{/i}) + \bar{\boldsymbol{X}}^\top_{/i} \operatorname{diag}\left[\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right] \bar{\boldsymbol{X}}_{/i}\right)^{-1} \boldsymbol{x_i}\right| \\
&\overset{1}{\leq} |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left(\left\|\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right\|_4^2\right) \left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&+ |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left(\left\|\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_4^2\right) \left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&\overset{2}{\leq} |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left(\left\|\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{i,\lambda}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right\|_2^2\right) \left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&+ |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \left(\left\|\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{i,\lambda}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2^2\right) \left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&\overset{3}{\leq} 4c_1(n)c_2(n)\left\|\boldsymbol{\Delta}^*_{/i}\right\|_2 \left(1 + 2c_2(n)\left\|\boldsymbol{\Delta}^*_{/i}\right\|_2\left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\right)\left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&\overset{4}{\leq} 4\|\boldsymbol{x_i}\|_2\left(\frac{c_1^2(n)c_2(n)}{\nu}\right)\left(1 + \frac{2c_1(n)c_2(n)\bar{\omega}_{\max,i}}{\nu^2}\|\boldsymbol{x_i}\|_2\right)\left\|\bar{\boldsymbol{X}}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^2 \\
&\overset{5}{\leq} \underbrace{4\|\boldsymbol{x_i}\|_2\left(\frac{c_1^2(n)c_2(n)}{\nu}\right)\left(1 + \frac{2c_1(n)c_2(n)(1+\omega_{\max,i})}{\nu^2}\|\boldsymbol{x_i}\|_2\right)}_{\triangleq \bar{C}_i}\sqrt{\left\|\boldsymbol{X}_{/i}\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^4 + \left\|\boldsymbol{J}^{-1}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x_i}\right\|_4^4},
\end{aligned}
$$

where

- $\overset{0}{\leq}$ is due (141).
- $\overset{1}{\leq}$ is due to Assumption 7, and Lemma 25, where $\bar{\omega}_{\max,i} \triangleq \sigma_{\max}\left(\bar{\boldsymbol{X}}_{/i}\bar{\boldsymbol{X}}^\top_{/i}\right)$.
- $\overset{2}{\leq}$ is due the fact that for any $\boldsymbol{\gamma}$ we have $\|\boldsymbol{\gamma}\|_4^2 \leq \|\boldsymbol{\gamma}\|_2^2$,
- $\overset{3}{\leq}$ is due to Assumption 6 as illustrated below

$$
\begin{aligned}
\left\|\boldsymbol{\gamma}_{-\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 &\leq \left\|\ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{\Delta}^*_{/i}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 + \left\|\lambda(\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i} - \boldsymbol{\Delta}^*_{/i}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i}))\right\|_2 \\
&\leq 2c_2(n)\left\|\boldsymbol{\Delta}^*_{/i}\right\|_2.
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
\left\|\int_0^1 \boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})dt\right\|_2 &\leq \int_0^1 \left\|\boldsymbol{\gamma}_{-(1-t)\boldsymbol{\Delta}^*_{/i}/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 dt \\
&\leq \int_0^1 \left\|\ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i} - (1-t)\boldsymbol{\Delta}^*_{/i}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i})\right\|_2 dt \\
&+ \int_0^1 \left\|\lambda(\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i} - (1-t)\boldsymbol{\Delta}^*_{/i}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i}))\right\|_2 dt \\
&\leq 2c_2(n)\left\|\boldsymbol{\Delta}^*_{/i}\right\|_2. \qquad (142)
\end{aligned}
$$

Here we should emphasize that this is the main place in which we have used the smoothness of second derivatives of the loss and regularizer in Assumption 6. [8]

- $\overset{4}{\leq}$ is due to inequality (136), and Assumption 6.
- $\overset{5}{\leq}$ is due to (139), and

$$\begin{aligned}\bar{\omega}_{\max,i} &= \sigma_{\max}\left(\bar{\boldsymbol{X}}_{/i}\bar{\boldsymbol{X}}_{/i}^\top\right) = \sigma_{\max}\left(\bar{\boldsymbol{X}}_{/i}^\top\bar{\boldsymbol{X}}_{/i}\right) = \sigma_{\max}\left(\begin{bmatrix}\boldsymbol{X}_{/i}\\ \boldsymbol{I}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{X}_{/i}\\ \boldsymbol{I}\end{bmatrix}\right)\\ &= \sigma_{\max}\left(\boldsymbol{I}+\boldsymbol{X}_{/i}^\top\boldsymbol{X}_{/i}\right) \leq 1+\sigma_{\max}\left(\boldsymbol{X}_{/i}^\top\boldsymbol{X}_{/i}\right) = 1+\omega_{\max,i}.\end{aligned} \tag{145}$$

The final result follows the basic inequality: $\sqrt{a^2+b^2} \leq |a|+|b|$. □

Proof of Lemma 27. First, we prove

$$\Pr\left[\|\boldsymbol{x}\|_\infty > \rho_{\max}\sqrt{2(1+c)\log p}\right] \leq \frac{2}{p^c} \tag{146}$$

as follows

$$\Pr\left[\|\boldsymbol{x}\|_\infty > t\right] \leq \sum_{i=1}^p \Pr\left[|x_i| > t\right] \leq 2\sum_{i=1}^p e^{-\frac{t^2}{2\Sigma_{ii}}} \leq 2pe^{-\frac{t^2}{2\max_{i=1,\dots,p}\Sigma_{ii}}} \leq 2e^{\log p - \frac{t^2}{2\rho_{\max}}},$$

where $t = \rho_{\max}\sqrt{2(1+c)\log p}$ and $\max_{i=1,\dots,p}\Sigma_{ii} \leq \rho_{\max}$. Second, we prove

$$\Pr\left[\|\boldsymbol{x}\|_4^2 > 2(1+c)\rho_{\max}\sqrt{p}\log p\right] \leq \frac{2}{p^c} \tag{147}$$

in the following way:

$$\begin{aligned}\Pr\left[\sqrt{\sum_{i=1}^p {x_i}^4} > t\right] &= \Pr\left[\sum_{i=1}^p {x_i}^4 > t^2\right] \leq \Pr\left[p\max_{i=1,\dots,p} {x_i}^4 > t^2\right]\\ &\leq \Pr\left[\|\boldsymbol{x}\|_\infty > \left(\frac{t^2}{p}\right)^{1/4}\right] \leq 2e^{\log p - \frac{t}{2\rho_{\max}\sqrt{p}}},\end{aligned} \tag{148}$$

where $t = 2(1+c)\rho_{\max}\sqrt{p}\log p$ yields the desired result. Let $\boldsymbol{z} \triangleq \boldsymbol{J}^{-1}\boldsymbol{x}$ and $\boldsymbol{u} \triangleq \boldsymbol{X}\boldsymbol{J}^{-1}\boldsymbol{x}$, then $\boldsymbol{z}$ is zero mean Gaussian with covariance $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{J}^{-1}\boldsymbol{\Sigma}\boldsymbol{J}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{X}\boldsymbol{J}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{J}^{-1}\boldsymbol{X}^\top$, leading to

$$\sigma_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{z}}\right) = \sigma_{\max}\left(\boldsymbol{J}^{-1}\boldsymbol{\Sigma}\boldsymbol{J}^{-1}\right) \leq \frac{\rho_{\max}}{\nu_{\min}^2}, \tag{149}$$

$$\sigma_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}\right) = \sigma_{\max}\left(\boldsymbol{X}\boldsymbol{J}^{-1}\boldsymbol{\Sigma}\boldsymbol{J}^{-1}\boldsymbol{X}^\top\right) \leq \frac{\rho_{\max}}{\nu_{\min}^2}\omega_{\max}. \tag{150}$$

[8] Note that by checking the derivation, it is clear that we can replace Assumption 6 with the following weaker assumptions:

$$c_2(n) > \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i}+t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}})\|_2}{\left\|\hat{\boldsymbol{\beta}}_{/i}-\hat{\boldsymbol{\beta}}\right\|_2^\zeta} \tag{143}$$

$$c_2(n) > \sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{r}}((1-t)\hat{\boldsymbol{\beta}}_{/i}+t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})\|_2}{\left\|\hat{\boldsymbol{\beta}}_{/i}-\hat{\boldsymbol{\beta}}\right\|_2^\zeta} \tag{144}$$

for some $\zeta > 0$, and still find an (weaker) upper bound for $\left|\boldsymbol{x}_i^\top\boldsymbol{\Delta}_{/i}^* - \boldsymbol{x}_i^\top\hat{\boldsymbol{\Delta}}_{/i}\right|$ that converges to zero as $n,p\to\infty$.

Therefore, we have

$$\begin{aligned}
&\Pr\left[\left\|\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu_{\min}^2}\right)\sqrt{p}\log p\right] \\
&\leq \quad \Pr\left[\left\|\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_4^2 > 2(1+c)\sigma_{\max}\left(\boldsymbol{\Sigma_z}\right)\sqrt{p}\log p\right] \leq \frac{2}{p^c},
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left[\left\|\boldsymbol{X}\boldsymbol{J}^{-1}x\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu_{\min}^2}\omega_{\max}\right)\sqrt{m}\log m\right] \\
&\leq \quad \Pr\left[\left\|\boldsymbol{X}\boldsymbol{J}^{-1}x\right\|_4^2 > 2(1+c)\sigma_{\max}\left(\boldsymbol{\Sigma_u}\right)\sqrt{m}\log m\right] \leq \frac{2}{m^c}.
\end{aligned}$$

□

A.8 Proof of Corollary 1

To bound $|\text{LO} - \text{ALO}|$ we use the following variable

$$\kappa_i \triangleq \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\frac{H_{ii}}{1-H_{ii}} - \boldsymbol{x_i}^\top \boldsymbol{\Delta}_{/i}^*$$

as follows:

$$\begin{aligned}
|\text{LO} - \text{ALO}| &= \left|\frac{1}{n}\sum_{i=1}^n \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i}\right) - \frac{1}{n}\sum_{i=1}^n \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}} + \boldsymbol{x_i}^\top \boldsymbol{\Delta}_{/i}^*\right) - \frac{1}{n}\sum_{i=1}^n \phi\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \dot{\phi}\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i} + a_i\kappa_i\right)\kappa_i\right|. \qquad (151)
\end{aligned}$$

where $a_1, \ldots, a_n$ denote n numbers between 0 and 1. Note that we have $\kappa_i < \frac{C_o}{\sqrt{p}}$ with probability at least $1 - \left(\frac{8n}{(n-1)^3} + \frac{8n}{p^3} + 4ne^{-p}\right) - q_n - \tilde{q}_n$. Therefore, with at least the same probability we have

$$|\text{LO} - \text{ALO}| \quad \leq \quad \frac{C_o}{\sqrt{p}} \times \max_{i=1,\ldots,n} \sup_{|b_i| < \frac{C_o}{\sqrt{p}}} \left|\dot{\phi}\left(y_i, \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{/i} + b_i\right)\right|. \qquad (152)$$

A.9 ALO and LO in the p fixed and large n regime

As we discussed so far, our main concern in this paper is high-dimensional settings in which n is proportional to p. However, to present a complete picture about ALO, in this section, we study it in the classical asymptotic regime where n is large and p is fixed. The assumptions presented here will be used throughout Section A.9 only. Let

$$\hat{\boldsymbol{\beta}}_{\lambda_n} \triangleq \arg\min_{\boldsymbol{\beta} \in \mathrm{R}^p} \Big\{\sum_{i=1}^n \ell(y_i | \boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda_n r(\boldsymbol{\beta})\Big\}.$$

We also assume that the samples $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$ are independent and identically distributed, and that $\frac{\lambda_n}{n} \to \lambda^*$. Define,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\ell(y_i | \boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda^* r(\boldsymbol{\beta}).$$

Also, define

$$\begin{aligned}\boldsymbol{R} &\triangleq \mathrm{E}\left\{\dot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta}^*)\dot{\phi}(y,\boldsymbol{x}^\top\boldsymbol{\beta}^*)\boldsymbol{x}\boldsymbol{x}^\top\right\},\\ \boldsymbol{K} &\triangleq \mathrm{E}\left\{\ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta}^*)\boldsymbol{x}\boldsymbol{x}^\top + \lambda^*\mathrm{diag}[\ddot{\boldsymbol{r}}(\boldsymbol{\beta}^*)]\right\}. \end{aligned} \tag{153}$$

For the sake of brevity, we follow [Stone, 1977] and make the following assumptions that enable us avoid repeating standard asymptotic arguments that can be found elsewhere, e.g. in [Van der Vaart, 2000]:

(B.1) As $n \to \infty$ $\hat{\boldsymbol{\beta}}_{\lambda_n} \xrightarrow{p} \boldsymbol{\beta}^*$.

(B.2) $\sup_i \|\hat{\boldsymbol{\beta}}_{\lambda_n/i} - \boldsymbol{\beta}^*\|_2 \xrightarrow{p} 0$.

(B.3) Define $\boldsymbol{\Delta}_{/i} \triangleq \hat{\boldsymbol{\beta}}_{\lambda_n} - \hat{\boldsymbol{\beta}}_{\lambda_n/i}$. Let $b_1, b_2, \ldots, b_n$ and $c_1, c_2, \ldots, c_n$ denote $2n$ numbers between $[0,1]$ that may depend on the dataset $\mathcal{D}$. Then, we assume that

$$\frac{\boldsymbol{X}\,\mathrm{diag}\left[\ddot{\boldsymbol{\ell}}\left(\hat{\boldsymbol{\beta}}_{\lambda_n} + b_i\boldsymbol{\Delta}_{/i}\right)\right]\boldsymbol{X}^\top + \lambda_n\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{\lambda_n} + c_i\boldsymbol{\Delta}_{/i})]}{n} \xrightarrow{p} \boldsymbol{K}.$$

(B.4) Let $a_1, \ldots, a_n$ denote n number between $0, 1$ that may depend on dataset $\mathcal{D}$. Then, assume that

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\lambda_n/i})\dot{\phi}\left(y_i, \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{\lambda_n} + a_i\boldsymbol{x}_i^\top\boldsymbol{\Delta}_{/i}\right) \xrightarrow{p} \boldsymbol{R}.$$

(B.5) Note that

$$H_{ii} = \boldsymbol{x}_i^\top\left(\boldsymbol{X}\,\mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_{\lambda_n})]\boldsymbol{X}^\top + \lambda\mathrm{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{\lambda_n})]\right)^{-1}\boldsymbol{x}_i\ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\lambda_n}).$$

Hence, we also assume that $H_{ii} \xrightarrow{p} 0$.

(B.6) Let $d_1, d_2, \ldots, d_n$ denote n numbers between $[0,1]$. Note that we have already assumed that $\sup_i H_{ii} \xrightarrow{p} 0$. We further assume that

$$\sum_{i=1}^n\left(\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{n}\right)\left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\lambda_n})}{1-H_{ii}}\right)\dot{\phi}\left(y_i, \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} + d_i\left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\lambda_n})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\lambda_n})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right) \xrightarrow{p} \boldsymbol{R}.$$

It should be clear that all these assumptions can be proved under appropriate regularity conditions on the loss function and the regularizer. Note that according to this theorem the error between ALO and LO is $o_p(1/n)$.

Theorem 6. *Under assumptions (B.1), (B.2), ..., (B.6), we have* $n(\mathrm{ALO}_{\lambda_n} - \mathrm{LO}_{\lambda_n}) \xrightarrow{p} 0$, *as* $n \to \infty$.

Proof. For notational simplicity instead of using λ_n we use λ in our formulas. However, the reader should note that $\lambda/n \to \lambda^*$. First note that the gradient condition implies that $\boldsymbol{X}_{/i}\dot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i}) + \lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i}) = 0$. Hence,

$$\boldsymbol{X}\dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_{/i}) + \lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i}) = \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x}_i. \tag{154}$$

Furthermore, we can use the fact that $\boldsymbol{X}\dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}) + \lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}) = 0$ to obtain

$$\boldsymbol{X}\dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_{/i}) + \lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{/i}) - \boldsymbol{X}\dot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}) - \lambda\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}) = \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{/i})\boldsymbol{x}_i$$

Since both the loss function and the regularizer are assumed to be twice continuously differentiable, we can use the mean value theorem to simplify this expression to

$$\left(\boldsymbol{X} \operatorname{diag}\left[\ddot{\boldsymbol{\ell}}\left(\hat{\boldsymbol{\beta}}+b_i \boldsymbol{\Delta}_{/i}\right)\right] \boldsymbol{X}^{\top}+\lambda \operatorname{diag}\left[\ddot{\boldsymbol{r}}\left(\hat{\boldsymbol{\beta}}+c_i \boldsymbol{\Delta}_{/i}\right)\right]\right) \boldsymbol{\Delta}_{/i}=\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{/i}) \boldsymbol{x}_{\boldsymbol{i}}, \tag{155}$$

where all b_is and c_is are in $[0,1]$. Furthermore, if ϕ is continuously differentiable, then we can again use the mean value theorem to obtain

$$\begin{aligned}
\mathrm{LO} &= \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}}) + \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\Delta}_{/i} \dot{\phi}\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}} + a_i \boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\Delta}_{/i}\right) = \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}}) \\
&+ \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{\boldsymbol{i}}^\top \left(\boldsymbol{X} \operatorname{diag}\left[\ddot{\boldsymbol{\ell}}\left(\hat{\boldsymbol{\beta}}+b_i \boldsymbol{\Delta}_{/i}\right)\right] \boldsymbol{X}^{\top}+\lambda \operatorname{diag}\left[\ddot{\boldsymbol{r}}\left(\hat{\boldsymbol{\beta}}+c_i \boldsymbol{\Delta}_{/i}\right)\right]\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}} \\
&\times \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{/i}) \dot{\phi}\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}} + a_i \boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\Delta}_{/i}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathrm{LO} - \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}}) \\
&= \frac{1}{n}\sum_{i=1}^n \operatorname{trace}\left[\left(\boldsymbol{X} \operatorname{diag}\left[\ddot{\boldsymbol{\ell}}\left(\hat{\boldsymbol{\beta}}+b_i \boldsymbol{\Delta}_{/i}\right)\right] \boldsymbol{X}^{\top}+\lambda \operatorname{diag}[\ddot{\boldsymbol{r}}\left(\hat{\boldsymbol{\beta}}+c_i \boldsymbol{\Delta}_{/i}\right)]\right)^{-1} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{\boldsymbol{i}}^\top\right] \\
&\times \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{/i}) \dot{\phi}\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}} + a_i \boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\Delta}_{/i}\right).
\end{aligned}$$

It is then straightforward to use Assumptions (B.3) and (B.4) to claim that

$$n(\mathrm{LO} - \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}})) \xrightarrow{p} \operatorname{trace}(\boldsymbol{K}^{-1}\boldsymbol{R}).$$

Similarly, we can use the mean value theorem to argue that

$$\begin{aligned}
\mathrm{ALO} &= \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}}) \\
&+ \frac{1}{n}\sum_{i=1}^n \left(\frac{H_{ii}}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right) \times \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{1-H_{ii}}\right) \dot{\phi}\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}} + d_i \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}}) + \frac{1}{n}\operatorname{trace}\left[\left(\frac{\boldsymbol{X} \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}^\top + \lambda \operatorname{diag}[\ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}})]}{n}\right)^{-1}\right. \\
&\times \left.\sum_{i=1}^n \left(\frac{\boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{x}_{\boldsymbol{i}}^\top}{n}\right)\left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{1-H_{ii}}\right) \dot{\phi}\left(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}} + d_i \left(\frac{\dot{\ell}_i(\hat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\hat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)\right]
\end{aligned}$$

with $|d_i| \leq 1$, $i = 1, \cdots, n$. Again it is straightforward to use Assumptions (B.3) and (B.6) to show that

$$n(\mathrm{ALO} - \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_{\boldsymbol{i}}^\top \hat{\boldsymbol{\beta}})) \xrightarrow{p} \operatorname{trace}(\boldsymbol{K}^{-1}\boldsymbol{R}). \tag{156}$$

□