Approximate Leave-one-out Cross Validation for Regression with ℓ_1 Regularizers

Arnab Auddy
University of Pennsylvania

Haolin Zou
Columbia University

Kamiar Rahnama Rad
Columbia University

Arian Maleki
Baruch College
City University of New York

Abstract

The out-of-sample error (OO) is the main quantity of interest in risk estimation and model selection. Leave-one-out cross validation (LO) offers a (nearly) distribution-free yet computationally demanding method to estimate OO. Recent theoretical work showed that approximate leave-one-out cross validation (ALO) is a computationally efficient and statistically reliable estimate of LO (and OO) for generalized linear models with twice differentiable regularizers. For problems involving non-differentiable regularizers, despite significant empirical evidence, the theoretical understanding of ALO's error remains unknown. In this paper, we present a novel theory for a wide class of problems in the generalized linear model family with the non-differentiable ℓ_1 regularizer. We bound the error $|\text{ALO}-\text{LO}|$ in terms of intuitive metrics such as the size of leave-i-out perturbations in active sets, sample size n, number of features p and signal-to-noise ratio (SNR). As a consequence, for the ℓ_1 regularized problems, we show that $|\text{ALO}-\text{LO}| \xrightarrow{p\to\infty} 0$ while n/p and SNR remain bounded.

1 INTRODUCTION

Suppose we observe a dataset $\mathcal{D} = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ denote the features and response of the i^{th} observation, respectively. We assume observations are independent and identically distributed draws from some unknown joint distribution $q(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}^*)p(\mathbf{x}_i)$. We estimate $\boldsymbol{\beta}^*$ using the optimization problem

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \right\} \tag{1}$$

where $\ell(y|z)$ is the loss function, and $r(\boldsymbol{\beta})$ is the regularizer. Consider the problem of estimating the out-of-sample prediction error (OO), which is defined as

$$\text{OO} := \mathbb{E}[\phi(y_{\text{new}}, \mathbf{x}_{\text{new}}^\top \widehat{\boldsymbol{\beta}})|\mathcal{D}]$$

where $\phi(y, z)$ is another loss function (possibly but not necessarily the same as $\ell(y|z)$), $(y_{\text{new}}, \mathbf{x}_{\text{new}})$ is a sample from the same joint distribution $q(y|\mathbf{x}^\top \boldsymbol{\beta}^*)p(\mathbf{x})$, but is independent of the training set $\mathcal{D}$. As demonstrated through empirical and theoretical studies, in high-dimensional settings ($n, p \to \infty$ while n/p is fixed), the leave-one-out cross validation (LO) estimator

$$\text{LO} := \frac{1}{n} \sum_{i=1}^{n} \phi(y_i; \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}) \tag{2}$$

where

$$\widehat{\boldsymbol{\beta}}_{/i} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \sum_{j\neq i} \ell(y_j|\mathbf{x}_j^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \right\}, \tag{3}$$

provides an accurate estimation of the risk: (Rahnama Rad et al., 2020; Patil et al., 2021).

A significant limitation of LO is its necessity to fit the model repeatedly n times, making it computationally impractical for many high-dimensional problems. As a result, several recent researches have considered the problem of approximating LO (Beirami et al., 2017; Stephenson and Broderick, 2020; Rahnama Rad and Maleki, 2020; Giordano et al., 2019b,a; Wang et al., 2018; Rahnama Rad et al., 2020; Patil et al., 2021, 2022). For instance in (Rahnama Rad and Maleki, 2020; Rahnama Rad et al., 2020) it was theoretically and empirically shown that for twice differentiable regularizers and loss functions, the approximate leave-one-out cross validation (ALO) is a statistically reliable and computationally efficient approach

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

for estimating LO and OO in high-dimensional settings where $n, p \to \infty$ while n/p and SNR remain fixed and bounded. Regarding non-differentiable regularizers, while the extensive simulations in (Wang et al., 2018; Rahnama Rad and Maleki, 2020) provided empirical evidence, the theoretical understanding of ALO's error remains unknown.

In this paper, we present a novel theory for non-differentiable regularizers applied to a wide class of problems, e.g., linear regression, as well as generalized linear models like Poisson and logistic regression. Using intuitive metrics such as the size of leave-i-out perturbations in active sets we bound the error $|\text{ALO} - \text{LO}|$ in terms of fundamental quantities such as sample size n, number of features p, and signal-to-noise ratio (SNR). For ℓ_1 regularized least-square problems, we place bounds on the size of leave-i-out perturbations in active sets, and as a consequence, show that $|\text{ALO} - \text{LO}| \xrightarrow{p\to\infty} 0$ when n/p and SNR remain bounded [1].

The remainder of this paper is organized as follows. In the first two subsections of Section 2 we briefly present and review the key idea of ALO and the challenge to use it for non-differentiable regularizers. We review related work in section 2.3. In section 2.4 we briefly describe the main theoretical contributions of this paper. Section 3 presents Theorem 1 which allows us to bound the error $|\text{ALO}-\text{LO}|$ in terms of the typical size of leave-i-out perturbations in active sets. In Section 4 we present Theorem 2 which bounds metrics related to the size of leave-i-out perturbations in active sets for ℓ_1 regularized problems. Next we explain how Theorem 1 and Theorem 2 together lead to $|\text{ALO}-\text{LO}| \xrightarrow{p\to\infty} 0$ when n/p and SNR remain bounded for ℓ_1 regularized least squares problems.

Concluding remarks are given in section 5. Detailed proofs can be found in the online supplementary material.

2 APPROXIMATE LEAVE-ONE-OUT CROSS VALIDATION

In this section, we briefly review the key idea of ALO and the challenges to use it for non-differentiable regularizers.

[1] In Section 1.3 of the online supplementary material we rigorously discuss what we mean by a bounded SNR. Roughly speaking, we mean that $\text{var}[\mathbf{x}_j^\top \boldsymbol{\beta}^*]$ and $\text{var}[y_i|\mathbf{x}_j^\top \boldsymbol{\beta}^*]$ are stochastically bounded away from 0 and ∞, regardless of problem dimensions.

2.1 ALO for twice-differentiable losses and regularizers

ALO replaces the computationally demanding procedure of repeatedly fitting the model with finding an approximate model that is easy to compute. Instead of exactly computing $\widehat{\boldsymbol{\beta}}_{/i}$ as in (3), ALO adjusts the estimate $\widehat{\boldsymbol{\beta}}$ based on the entire dataset $\mathcal{D}$, and uses one Newton step to compute the approximation $\tilde{\boldsymbol{\beta}}_{/i}$ as follows:

$$\tilde{\boldsymbol{\beta}}_{/i} := \widehat{\boldsymbol{\beta}} + \left(\sum_{j\neq i} \mathbf{x}_j \mathbf{x}_j^\top \ddot{\ell}(\mathbf{y}_j|\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}) + \lambda \nabla^2 r(\widehat{\boldsymbol{\beta}})\right)^{-1} \mathbf{x}_i \dot{\ell}(y_i|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}), \tag{4}$$

where $\dot{\ell}(y|z)$ and $\ddot{\ell}(y|z)$ denote the first and second derivatives of $\ell(y|z)$ with respect to its second argument. Furthermore,

$$[\nabla^2 r(\mathbf{w})]_{ij} := \left.\frac{\partial^2 r(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right|_{\boldsymbol{\beta}=\mathbf{w}}.$$

It might seem that the matrix inversion required in (4) is computationally (nearly) as demanding as refitting the model, but this can actually be bypassed by using the Woodbury lemma (see Lemma 10 in the online supplement), leading to the following approximation

$$\begin{aligned}\text{ALO} &:= \frac{1}{n}\sum_{i=1}^n \phi(y_i; \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}_i) \\ &= \frac{1}{n}\sum_{i=1}^n \phi\left(y_i; \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}(y_i|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})}{\ddot{\ell}(y_i|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right)\end{aligned} \tag{5}$$

where H_{ii} is the (i,i) element of the matrix $\mathbf{H}$ defined as

$$\mathbf{H} := \mathbf{X}\left(\mathbf{X}^\top[\text{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))]\mathbf{X} + \lambda\nabla^2 r(\widehat{\boldsymbol{\beta}})\right)^{-1}\mathbf{X}^\top \text{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})]$$

with $\ddot{\ell}(\mathbf{w}) := \left[\ddot{\ell}(y_1|\mathbf{x}_1^\top \mathbf{w}), \cdots, \ddot{\ell}(y_n|\mathbf{x}_n^\top \mathbf{w})\right]^\top$ and $\text{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})]$ being the diagonal matrix with $\ddot{\ell}(\mathbf{w})$ as its diagonal elements. Most of the theoretical work about the consistency of ALO in estimating LO (and OO) has focused on differentiable regularizers, such as ridge, smoothed LASSO and Huber loss when n/p remain fixed (Rahnama Rad et al., 2020; Rahnama Rad and Maleki, 2020; Patil et al., 2022; Xu et al., 2021).

2.2 ALO for non-differentiable regularizers

The ALO formula above and the corresponding theory supporting it require twice differentiability but some of the most important loss functions and regularizers, such as elastic net, nuclear norm, or hinge loss are not

twice-differentiable. For example, consider the following estimate:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\ h(\boldsymbol{\beta}), \tag{6}$$

where $h(\boldsymbol{\beta}) := \sum_{i=1}^n \ell(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}) + \lambda(1-\eta)\|\boldsymbol{\beta}\|_1 + \lambda\eta\|\boldsymbol{\beta}\|_2^2$, and suppose that as before our goal is to approximate

$$\text{LO} := \frac{1}{n}\sum_{i=1}^n \phi(y_i; \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}) \tag{7}$$

where

$$\widehat{\boldsymbol{\beta}}_{/i} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j\neq i} \ell(y_j|\mathbf{x}_j^\top\boldsymbol{\beta}) + \lambda(1-\eta)\|\boldsymbol{\beta}\|_1 + \lambda\eta\|\boldsymbol{\beta}\|_2^2 \right\}. \tag{8}$$

Due to the regularizer's non-differentiability we cannot use the one step Newton approximation as proposed in the previous section. However, the following heuristic argument serves as a motivation of our new method. Let $\mathcal{S}$ denote the active set of $\widehat{\boldsymbol{\beta}}$, i.e.,

$$\mathcal{S} = \{i \ :\ \widehat{\beta}_i \neq 0\}.$$

Suppose that the active set of $\widehat{\boldsymbol{\beta}}_{/i}$ remains the same as $\mathcal{S}$ for all i. Then, we can solve (3) on the set $\mathcal{S}$ only. The validity of this heuristic assumption, and our remedies when it is mildly violated, are discussed in later sections. For now, since the regularizer is twice differentiable on $\mathcal{S}$, we can use the Newton method to obtain the following approximation for LO of the elastic-net:

$$\text{ALO} = \frac{1}{n}\sum_{i=1}^n \phi\left(y_i, \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right) \tag{9}$$

where H_{ii} is the (i,i) element of

$$\mathbf{H} := \mathbf{X}_{\mathcal{S}}\left(2\lambda\eta\mathbb{I} + \mathbf{X}_{\mathcal{S}}^\top \operatorname{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})]\mathbf{X}_{\mathcal{S}}\right)^{-1}\mathbf{X}_{\mathcal{S}}^\top \operatorname{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})].$$

with $\mathbf{X}_{\mathcal{S}}$ contains the columns of X that are in $\mathcal{S}$.

Unfortunately, the assumption that led to (9), i.e., the assumption that the active set does not change when a data point is removed, is not correct. While it is true that some of leave-i-out estimated coefficients retain the active set, most estimated active sets do change. Figure 1 confirms this claim.

Despite the observation in Figure 1, extensive empirical results presented in (Beirami et al., 2017; Wang et al., 2018; Rahnama Rad and Maleki, 2020; Stephenson and Broderick, 2020) confirm that (9) offer an accurate estimation of LO (and OO).

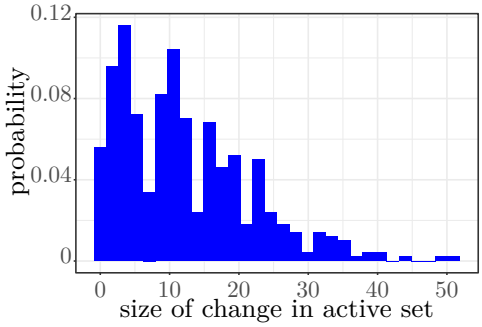


Figure 1: The histogram above shows the size of the difference in active sets when performing linear regression with the elastic net penalty on the entire dataset, vs leave-i-out (i.e., leaving out the i-th observation for $i = 1, \ldots, n$). The parameters are $n = 500$, $p = 1000$ with 20% of the true coefficients being non-zero. The design matrix $\mathbf{X}$ has iid $N(0, 1/n)$ rows, and the nonzero coefficients are iid $N(0, 1)$. The penalty strengths are $\lambda = 2, \eta = 0.5$. We use the `ElasticNet` function from the Python library `scikit-learn`.

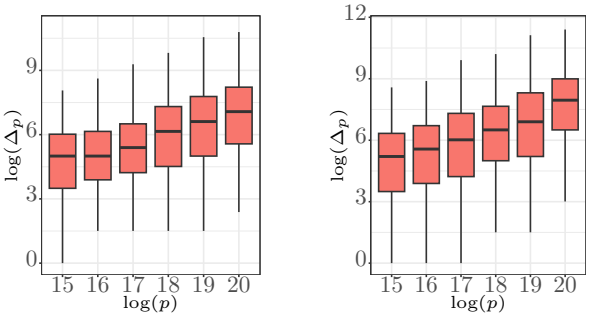
To understand these two contradictory observations, we ran another simulation that appears in Figure 2. In this figure we find that, while the active set does change, the number of changes in the active set (denoted by Δ_p) scales at a sub-linear rate with respect to p (and hence n) as p, n increases. Indeed, the comparable points in the boxplots (e.g., the median, maximum) of the logarithms, lie on a line that has slope smaller than one. A linear regression of the medians of $\log(\Delta_p)$ on $\log(p)$ showed a slope of 0.43 when $p = n$ and 0.52 when $p = 2n$.

This implies that $\Delta_p/p \to 0$ as p, n increases. As will be clarified later, this sub-linear rate of growth that will be proved in Theorem 2, is the main reason that the simulation results confirm the accuracy of (9).

2.3 Related work

Various approaches to estimating the out-of-sample error have been proposed. Examples include (but are not limited to) cross validation (Stone, 1974), predicted residual error sum of squares (Allen, 1974), and generalized cross validation (Craven and Wahba, 1979; Golub et al., 1979), just to name a few.

In the past, the use of n-fold cross validation (also known as LO) has been limited due to the high computational cost of repeatedly refitting the model n times, and due to concerns about the high variance, especially when n and p grow unboundedly. Recently, these concerns have been (mostly) alleviated by a large body of work that showed: 1) that the variance of LO in estimating OO goes to zero as n and p grow (Kumar et al., 2013; Bayle et al., 2020; Rahnama Rad et al., 2020; Patil et al., 2022; Luo et al., 2023), and 2) that



(a) $p = n$ (b) $p = 2n$

Figure 2: The figure shows boxplots of the change in sizes of leave-i-out active sets (denoted by Δ_p), plotted against the dimension, (on a logarithmic scale) when performing linear regression with the elastic net penalty. The upper and lower edges of the box in each boxplot represent the 1st and the 3rd quartiles respectively, and the black line represents the median. The whiskers extend up to the most extreme value in 1.5 times the interquartile range. The parameters are taken as $p = n$ (left) and $p = 2n$ (right), and in either figure p is then varied from 1000 to 10000 with six equal increments on the log scale. We take 20% of the true coefficients to be non-zero. The design matrix $\mathbf{X}$ has iid $N(0, 1/n)$ rows, and the nonzero coefficients are iid $N(0, 1)$. The penalty strengths are $\lambda = 2, \eta = 0.5$. We use the `ElasticNet` function from the Python library `scikit-learn`.

computationally efficient approximations to the leave-one-out cross validation (can) provide statistically reliable estimates of LO and OO (Beirami et al., 2017; Wang et al., 2018; Rahnama Rad and Maleki, 2020; Rahnama Rad et al., 2020; Patil et al., 2022; Stephenson and Broderick, 2020; Obuchi and Sakata, 2019; Opper and Winther, 2000; Cawley and Talbot, 2008; Meijer and Goeman, 2013; Vehtari et al., 2017, 2016; Obuchi and Kabashima, 2016, 2018; Xu et al., 2021). Most theoretical work about the consistency of ALO in estimating LO has focused on differentiable regularizers, such as ridge, smoothed LASSO and Huber loss (Rahnama Rad et al., 2020; Rahnama Rad and Maleki, 2020; Patil et al., 2022). Assuming that the SNR grows unboundedly, as n grows, (Stephenson and Broderick, 2020) considered ℓ_1 regularizers. In this regime the optimal value of the regularization parameter λ goes to zero, and tuning becomes (nearly) irrelevant because of the unbounded SNR.

Despite significant empirical evidence, to the best of our knowledge, there is no theoretical study of the consistency of ALO for non-differentiable regularizers in a regime where n and p grow to infinity while n/p and SNR remains bounded, a framework typical in high dimensional risk estimation problems (Donoho et al., 2011; Donoho and Montanari, 2016; Maleki, 2011; Mousavi et al., 2018; Wang et al., 2020, 2022; Xu et al., 2021; Guo et al., 2022; Rahnama Rad et al., 2020; Rahnama Rad and Maleki, 2020; Patil et al., 2022). Given the importance of risk estimation for non-differertiable estimation problems, this paper addresses the problem in the finite-SNR regime where tuning significantly impacts the selected model and estimated coefficients (as we discuss in Section 3.2).

2.4 Our technical contributions

In this paper, our primary focus is on risk estimation for ℓ_1 regularized problems within the generalized linear model family. Specifically, we aim to establish an upper bound for the error $|\text{ALO} - \text{LO}|$ under the conditions of large n and p, while maintaining fixed and bounded values for n/p and SNR. In the following, we outline some of the key theoretical innovations that have enabled us to undertake a comprehensive analysis of $|\text{ALO} - \text{LO}|$ in this context. For detailed proofs and further insights, please refer to the supplementary material available online.

Our initial step involves a smooth approximation $r_\alpha(\mathbf{z})$ for the ℓ_1-norm $\|\mathbf{z}\|_1$, where α is a parameter such that, as α approaches infinity, this approximation becomes increasingly accurate. While smoothing techniques have been extensively employed for deriving approximate minimizers of non-differentiable convex functions, their application as proof techniques in high-dimensional statistics has been unexplored. The primary challenge arises from the fact that as $\alpha \to \infty$, it becomes considerably difficult to bound the quantities that are of particular interest to statisticians, such as $|\text{ALO}^\alpha - \text{LO}^\alpha|$ in our specific problem. Here, ALO^α and LO^α represent the smoothed approximations of ALO and LO, respectively.

In our paper (in the proof of Theorem 1), we have devised an innovative method to bound such quantities, which we anticipate to have broader applications in studying non-differentiable losses and regularizers in high-dimensional settings.

Our smoothing technique, in essence, simplifies the task of bounding $|\text{ALO} - \text{LO}|$ by transforming it into the challenge of limiting the changes in the active set between the full-data estimate and the leave-one-out estimate. Hence, in this paper, we develop new techniques for understanding the relationships between two estimates that are using the same samples (proof of Theorem 2). As an example of our approach, we employ our technique to establish an upper bound for the disparity between $\mathcal{S}$ and $\mathcal{S}_{/i}$, where roughly speaking, $\mathcal{S}$ and $\mathcal{S}_{/i}$ denote the locations of non-zero coefficients in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{/i}$, respectively. We anticipate that this technique will prove valuable for other problems, such as the analysis of ensemble methods within the context of high-dimensional settings.

3 MAIN THEORETICAL RESULT

We begin by introducing our notations in Section 3.1 and assumptions in Section 3.2. We discuss the assumptions in Section 3.3 and show these assumption encompass a large class of typical problems. Then we present our main theorem in Section 3.4.

3.1 Notations

In this manuscript, vectors are denoted with boldfaced lowercase letter, such as $\mathbf{x}$. Matrices are represented with boldfaced capital letters, such as $\mathbf{X}$. Calligraphic letters, such as $\mathcal{F}$ are used for sets and events. For a matrix $\mathbf{X}$, $\sigma_{\min}(\mathbf{X})$, $\|\mathbf{X}\|$, $\|\mathbf{X}\|_{HS}$, $\mathrm{Tr}(\mathbf{X})$ denote the minimum singular value, the spectral norm (equal to the maximum singular value $\sigma_{\max}(\mathbf{X})$), the Hilbert-Schmidt norm, and the trace of the matrix $\mathbf{X}$ respectively. Suppose $\mathcal{F}$ represents a subset of indices corresponding to the columns of matrix $\mathbf{X}$. In such a case, the notation $\mathbf{X}_{\mathcal{F}}$ refers to a matrix formed by selecting only those columns of $\mathbf{X}$ whose indices belong to the set $\mathcal{F}$. The subscript "$/i$" refers a quantity related to the leave-i-out model, e.g. $\mathbf{X}_{/i}$ refers to matrix $\mathbf{X}$ after deleting the ith row, and $\widehat{\boldsymbol{\beta}}_{/i}$, $\widehat{\boldsymbol{\beta}}^{\alpha}_{/i}$ refer to the leave-i-out estimate and the smoothed leave-i-out estimate, respectively. For two probability measures μ, ν, $W_q(\mu,\nu)$ denote their Wasserstein-q distance. Moreover, we use the following definitions in this paper:

$$\dot{\ell}_i(\boldsymbol{\beta}) := \frac{\partial \ell(y_i|z)}{\partial z}\Big|_{z=\mathbf{x}_i^\top\boldsymbol{\beta}}, \ddot{\ell}_i(\boldsymbol{\beta}) := \frac{\partial^2 \ell(y_i|z)}{\partial z^2}\Big|_{z=\mathbf{x}_i^\top\boldsymbol{\beta}}$$

$$\dot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}) := \left[\dot{\ell}_1(\boldsymbol{\beta}), \cdots, \dot{\ell}_{i-1}(\boldsymbol{\beta}), \dot{\ell}_{i+1}(\boldsymbol{\beta}), \cdots, , \dot{\ell}_n(\boldsymbol{\beta})\right]^\top,$$

$$\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}) := \left[\ddot{\ell}_1(\boldsymbol{\beta}), \cdots, \ddot{\ell}_{i-1}(\boldsymbol{\beta}), \ddot{\ell}_{i+1}(\boldsymbol{\beta}), \cdots, , \ddot{\ell}_n(\boldsymbol{\beta})\right]^\top.$$

We also denote any polynomials of $\log(n)$ by $\mathrm{PolyLog}(n)$. For $x, y \in \mathbb{R}$, we write $x \wedge y$ and $x \vee y$ to denote $\min\{x,y\}$ and $\max\{x,y\}$ respectively.

3.2 Assumption group A

The following assumptions have been extensively used in the literature of high-dimensional statistics. We will explain the rationale for making these assumptions in the next section.

A1 $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ where $\mathbf{x}_i \in \mathbb{R}^p$ are iid $N(0, \boldsymbol{\Sigma})$. Moreover, there exist constants $0 < c_X \le C_X$ such that $p^{-1}c_X \le \sigma_{\min}(\boldsymbol{\Sigma}) \le \sigma_{\max}(\boldsymbol{\Sigma}) \le p^{-1}C_X$.

A2 $n/p = \gamma_0 \in (0, \infty)$.

A3 ϕ has continuous derivative $\dot{\phi}$, and $\ell(y|z)$ has continuous second derivative w.r.t. z.

A4 There exists $\overset{\star}{\epsilon} > 0$, and $q_n, \check{q}_n, \bar{q}_n \in [0,1)$, such that

$$\mathbb{P}\left(\sup_{\substack{1\le i\le n\\ \mathbf{v}\in\mathcal{D}}} \dot{\ell}_i(\mathbf{v}) \le \mathrm{PolyLog}(n)\right) \ge 1 - \check{q}_n, \quad (10)$$

$$\mathbb{P}\left(\sup_{\substack{1\le i\le n\\ \mathbf{v}\in\mathcal{D}}} \ddot{\ell}_i(\mathbf{v}) \le \mathrm{PolyLog}(n)\right) \ge 1 - q_n, \quad (11)$$

$$\mathbb{P}\left(\sup_{\substack{1\le i\le n,\\ \mathbf{v},\mathbf{v}'\in\mathcal{D}}} \frac{\|\ddot{\ell}_i(\mathbf{v}) - \ddot{\ell}_{/i}(\mathbf{v}')\|}{\|\mathbf{v}-\mathbf{v}'\|_2} \le \mathrm{PolyLog}(n)\right) \ge 1 - \bar{q}_n \quad (12)$$

where

$$\mathcal{D} := \bigcup_{1\le i\le n}\bigcup_{t\in[0,1]} \mathcal{B}(t\widehat{\boldsymbol{\beta}} + (1-t)\widehat{\boldsymbol{\beta}}_{/i}, \overset{\star}{\epsilon}),$$
$$\mathcal{B}(\mathbf{w}, r) := \{\mathbf{z} : \|\mathbf{z}-\mathbf{w}\|_2 \le r\}.$$

A5 $\eta \in (0,1)$, and $\lambda \in (0, \lambda_{\max}]$ for an arbitrary constant $\lambda_{\max} > 0$.

3.3 Discussion of the assumptions

In the previous section, we made five assumptions that will be used for our theoretical results. In this section, we clarify our rationale for making these assumptions.

3.3.1 About assumption A1 and A2

Assumptions A1 and A2 ensure that $\mathbf{x}_i^\top\boldsymbol{\beta}^*$ remains finite as n and p grow unboundedly as long as $\|\boldsymbol{\beta}^*\|^2/p$ is bounded:

$$\frac{c_X}{p}\|\boldsymbol{\beta}^*\|^2 \le \mathbb{E}(\mathbf{x}_i^\top\boldsymbol{\beta}^*)^2 \le \frac{C_X}{p}\|\boldsymbol{\beta}^*\|^2.$$

For instance, if each element of $\boldsymbol{\beta}^*$ remains bounded, then $\mathbf{x}_i^\top\boldsymbol{\beta}^*$ will be $O_p(1)$.

In addition to the aforementioned rationale for Assumptions A1 and A2, there is another compelling justification that we explain below.

Let $\lambda^*(n,p)$ denote the value of λ that minimizes the prediction error of $\widehat{\boldsymbol{\beta}}$. Suppose that we are interested in the asymptotic setting $n, p \to \infty$, such that $n/p = \gamma_0$ remains fixed. Then, if Assumption A1 holds, for many problems it has been shown that $\lambda^*(n,p) \to \bar{\lambda}$, in probability, where $\bar{\lambda}$ is a fixed number in the range $(0, \infty)$. See for instance (Mousavi et al., 2018; Wang

et al., 2020, 2022). Intuitively speaking, if under another scaling $\lambda^*(n,p)$ goes to zero as $n,p\to\infty$, then it indicates that the estimation problem is becoming easier as p grows and hence a regularizer is not required. Similarly, if under another scaling $\lambda^*(n,p)$ goes to infinity as $n,p\to\infty$, it indicates that the estimation problem is becoming so difficult that we end up choosing zero estimator as the best one. Hence, the scaling we have chosen here seems to be one of the most useful scalings for practice.

In summary, we believe that Assumptions A1 and A2 provide a good scaling regime for studying risk estimation (or the related problem of hyperparameter tuning) problem.

3.3.2 About assumption A4

Assumption A4 introduces a regularity condition for the data generating mechanism and the loss function ℓ. To clarify this point, we present a proposition below, which outlines a sufficient condition based on simpler regularity conditions for ℓ and the data generation mechanism to satisfy A4.

Proposition 1. *Suppose that $\mathbb{P}(|y_i|>\mathrm{PolyLog}(n))\le q_n^y$ for some $q_n^y=o(1/n)$. Furthermore, suppose that $\ell(y|z)$ is three times differentiable with respect to z and that $\ell(y|z),\dot\ell(y|z),\ddot\ell(y|z)$, and $\dddot\ell(y|z)$ grow polynomially in y,z, i.e., there exists a positive integer m and a constant $C>0$ such that*

$$\max\{|\ell(y|z)|,|\dot\ell(y|z)|,|\ddot\ell(y|z)|,|\dddot\ell(y|z)|\}\le C(1+|y|^m+|z|^m). \tag{13}$$

for all (y,z). Then, Assumption A4 holds.

The proof of the above proposition can be found in Section 2 of the online supplementary material.

The condition of polynomial growth for the loss function is not unduly restrictive, as it encompasses many commonly used loss functions.

To illustrate this point, we present a few examples of popular loss functions, viz., squared error, logistic, and Poisson. Therefore, Assumption A4 holds for the majority of loss functions encountered in applications.

For simplicity we assume $\mathbf{x}_i\sim N(0,\frac{1}{n}\mathbb{I}_p)$, but the examples are still valid for a general covarance matrix as in Assumption A1.

Example 3.1 (Linear regression)**.** *Suppose $y_i|\mathbf{x}_i\sim N(\mathbf{x}_i^\top\boldsymbol\beta^*,\sigma^2)$, then $y_i\sim N(0,\sigma^2+\frac{1}{n}\|\boldsymbol\beta^*\|^2)$. Denote $\nu^2:=\sigma^2+\frac{1}{n}\|\boldsymbol\beta^*\|^2$, we then have, for arbitrary $q>1$:*

$$\mathbb{P}(|y_i|>\nu\sqrt{2q\log(n)})\le 2e^{-\frac{2q\nu^2\log(n)}{2\nu^2}}=n^{-q}.$$

If we use negative log-likelihood as loss function, then $\ell(y|z)$ and its derivatives w.r.t. z are:

$$\ell(y|z)=\frac{1}{2\sigma^2}(y-z)^2,$$
$$\dot\ell(y|z)=\frac{1}{\sigma^2}(z-y)\,;\ \ddot\ell(y|z)=\frac{1}{\sigma^2}\,;\ \dddot\ell(y|z)=0,$$

and hence they are all dominated by $\frac{1}{\sigma^2}(1+y^2+z^2)$.

Example 3.2 (Logistic regression)**.** *Suppose $y_i|\mathbf{x}_i\sim$ Bernoulli$(1/(1+e^{-\mathbf{x}_i^\top\boldsymbol\beta^*}))$, then the boundedness of $|y_i|$ is natually satisfied since $y_i\in\{0,1\}$. The negative log-likelihood loss and its derivatives w.r.t. z are*

$$\begin{aligned}|\ell(y|z)|&=\left|y\log(1+e^{-z})+(1-y)\log(1+e^z)\right|\\&\le 2\log(2)+2|z|,\\|\dot\ell(y|z)|&=\left|\frac{e^z}{1+e^z}-y\right|\le 1+|y|,\\|\ddot\ell(y|z)|&=\left|\frac{e^{-z}}{(1+e^{-z})^2}\right|\le 1,\\|\dddot\ell(y|z)|&=\left|\frac{e^z-e^{-z}}{(e^{-z}+e^z+2)^2}\right|\le 4.\end{aligned}$$

The bound of $\ell(y|z)$ uses the fact that $y\in\{0,1\}$.

Example 3.3 (Poisson regression)**.** *Suppose $y_i\sim$ Poisson(λ) where $\lambda=\log(1+e^{\mathbf{x}_i^\top\boldsymbol\beta^*})$, then we have, by the Chernoff bound (see, e.g., Exercise 2.3.3 of (Vershynin, 2018)) that:*

$$\mathbb{P}(y_i>\log(n))\le\left(\frac{C}{\sqrt{\log(n)}}\right)^{\log(n)}=o(n^{-1}).$$

Since $\mathbf{x}_i^\top\boldsymbol\beta^*\sim N(0,\frac{1}{n}\|\boldsymbol\beta^*\|^2)$, we have

$$\mathbb{P}(|\mathbf{x}_i^\top\boldsymbol\beta^*|>2\nu\sqrt{\log(n)})\le n^{-2}$$

where $\nu^2=\frac{1}{n}\|\boldsymbol\beta^*\|^2$. So we have, with probability at least $1-n^{-2}$, that

$$\lambda=\log(1+e^{\mathbf{x}_i^\top\boldsymbol\beta^*})\log(2)+2n^{-1/2}\|\boldsymbol\beta^*\|\sqrt{\log(n)}.$$

It can be checked that the negative log-likelihood loss is

$$\begin{aligned}|\ell(y|z)|&=|\log(y!)+\log(1+e^z)-y\log\log(1+e^z)|\\&\le C(y^2+z^2+1)\end{aligned}$$

where $C=1+\log\log(2)+(2\log(2))^{-1}$.

The derivatives of the loss function w.r.t. z satisfy

$$\begin{aligned}|\dot\ell(y|z)|&=\left|\frac{1}{1+e^{-z}}-\frac{ye^z}{(1+e^z)\log(1+e^z)}\right|\le 1+|y|,\\|\ddot\ell(y|z)|&=\left|y\left(\frac{e^z}{(1+e^z)\log(1+e^z)}\right)^2\right.\\&\qquad\left.-\frac{ye^z}{(1+e^z)^2\log(1+e^z)}+\frac{e^z}{(1+e^z)^2}\right|\\&\le 1+2|y|,\\|\dddot\ell(y|z)|&\le 3+14|y|.\end{aligned}$$

We use the fact that $((1+e^{-z})\log(1+e^z))^{-1} \leq 1$ *for all* $z \in \mathbb{R}$. *We omit the exact expression for* $\dddot{\ell}(y|z)$ *for brevity.*

3.4 Main theorem

Based on Assumptions A1-A5, we would like to state our main theorem. However, our main theorem uses two important sets. We first require the notion of the subgradient vector $g(\boldsymbol{\beta})$ defined as

$$g(\boldsymbol{\beta}) := \frac{1}{\lambda(1-\eta)} \sum_{i=1}^{n} \dot{\ell}(\mathbf{y}_i|\mathbf{x}_i^\top \boldsymbol{\beta}) - \frac{2\eta}{1-\eta}\boldsymbol{\beta} \tag{14}$$

It can be directly verified that $g(\widehat{\boldsymbol{\beta}}) \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$, i.e. $g(\widehat{\boldsymbol{\beta}})$ belongs to the subgradient of ℓ_1-term in (6). In fact, the first order derivative of the elastic net problem (6) gives:

$$0 \in \sum_{i=1}^{n} \dot{\ell}(\mathbf{y}_i|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) + \lambda(1-\eta)\partial\|\widehat{\boldsymbol{\beta}}\|_1 + 2\lambda\eta\widehat{\boldsymbol{\beta}}$$

where $\partial\|\widehat{\boldsymbol{\beta}}\|_1$ denotes the subgradient of $\|\boldsymbol{\beta}\|_1$ evaluated at $\widehat{\boldsymbol{\beta}}$. We then obtain (14) by rearranging the terms. Hence we hereafter refer to $g(\widehat{\boldsymbol{\beta}})$ as "the" subgradient of $\|\widehat{\boldsymbol{\beta}}\|_1$.

The active set of $\widehat{\boldsymbol{\beta}}$ is $\mathcal{A} := \{k \in [p] : |\widehat{\beta}_k| > 0\}$, and similarly the non-active set is $\mathcal{A}^c = \{k \in [p] : |\widehat{\beta}_k| = 0\}$. It is well known that $|g(\widehat{\boldsymbol{\beta}})_k| = 1$ for $k \in \mathcal{A}$ and $|g(\widehat{\boldsymbol{\beta}})_k| < 1$ for $k \in \mathcal{A}^c$. We now define two sets which, heuristically speaking, stand for "strongly active" and "strongly non-active" sets. We gather the active *large coefficients* into $\mathcal{S}^{(1)} \subset \mathcal{A}$, and the non-active *small sub-differential* coefficients into $\mathcal{S}^{(0)} \subset \mathcal{A}^c$. That is,

$$\begin{aligned} \mathcal{S}^{(1)} &:= \{k \in [p] : |\widehat{\beta}_k| > \kappa_1(n)\}, \\ \mathcal{S}^{(0)} &:= \{k \in [p] : |g(\widehat{\boldsymbol{\beta}})_k| \leq 1 - \kappa_0(n)\}, \end{aligned} \tag{15}$$

where $\kappa_1(n)$ and $\kappa_0(n)$ both are $o(1)$ as $n \to \infty$.[2] We will clarify our choice of these parameters later. Likewise, we define $\mathcal{S}^{(1)}_{/i}, \mathcal{S}^{(0)}_{/i}$ for the leave-i-out problems.

Note that $\mathcal{S}^{(1)}$ is a subset of the active set of $\widehat{\boldsymbol{\beta}}$ by only including active elements that are not too close to zero. On the one hand, the condition $\kappa_1(n) \to 0$ implies that $\mathcal{S}^{(1)}$ is close to the active set. On the other hand, when the gap $\kappa_1(n)$ is selected to be sufficiently large (the choice will be clarified later), it is intuitively expected that only a very small fraction of the indices in $\mathcal{S}^{(1)}$ will move out of the active set (i.e. the corresponding coefficient becomes zero) in the leave-i-out problem. These two points makes $\mathcal{S}^{(1)}$ a good substitution of the active set discussed in Section 2.2.

To understand $\mathcal{S}^{(0)}$, consider the non-active set of $\widehat{\boldsymbol{\beta}}$, i.e. the indices of zero coefficients of $\widehat{\boldsymbol{\beta}}$. $\mathcal{S}^{(0)}$ is actually a subset of the non-active set, which only includes elements with sub-gradients bounded away from 1, as indicated by its definition $\{k \in [p] : |g(\widehat{\boldsymbol{\beta}})_k| \leq 1-\kappa_0(n)\}$. Again $\kappa_0(n) \to 0$ makes $\mathcal{S}^{(0)}$ close to the non-active set, and by setting the right rate of the convergence of $\kappa_0(n)$, it should be expected that only a very small number of regression coefficients corresponding to indices in $\mathcal{S}^{(0)}$ will become nonzero in the leave-i-out problem.

To sum up, our choice of $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(0)}$ serve as proxies of the active and non-active set of $\widehat{\boldsymbol{\beta}}$ that are more resilient to the leave-i-out procedure. In other words, we expect the size of $(\mathcal{S}^{(1)} \cup \mathcal{S}^{(0)})^c$ to be small compared to p. As is clear from the above discussion, the speed at which $\kappa_1(n)$ and $\kappa_0(n)$ go to zero must be carefully selected to satisfy the two contrasting objectives. On the one hand, we want them both to go to zero as slowly as possible so that $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(0)}$ only include elements from which we have strong evidence to be active and non-active, respectively. On the other hand, we want $\kappa_1(n)$ and $\kappa_0(n)$ to go to zero as fast as possible, so that $(\mathcal{S}^{(1)} \cup \mathcal{S}^{(0)})^c$ is as small as possible. The details are presented in Theorem 2.

To describe our theoretical result, consider the following sets:

$$\begin{aligned} &\mathcal{B}_{1,i} := \mathcal{S}^{(1)} \cap \mathcal{S}^{(1)}_{/i}, \mathcal{B}_{0,i} := \mathcal{S}^{(0)} \cap \mathcal{S}^{(0)}_{/i} \\ &\mathcal{B}_{1,i,+} := \mathcal{B}_{1,i} \cap \{k : \widehat{\beta}_k \cdot \widehat{\beta}_{/i,k} > 0\} \end{aligned} \tag{16}$$

Heuristically, $\mathcal{B}_{1,i}$ contains the indices of large coefficients that remain large after leaving the i^{th} observation out, and $\mathcal{B}_{0,i}$ contains the indices of zero coefficients with small subderivative (of LASSO component) that remain so after leaving the i^{th} observation out. Note that $\mathcal{B}_{1,i}$ and $\mathcal{B}_{0,i}$ are mutually exclusive, and $\mathcal{B}_{1,i,+}$ is the subset of $\mathcal{B}_{1,i}$ that rules out the coefficients with flipped signs. Ideally we wish leaving-i-out would not change the fact of each coefficient being zero or non-zero, i.e., $\mathcal{B}_{1,i} \cup \mathcal{B}_{0,i} = [p]$. This is clearly not true according to Figure 1, but the violation is actually controllable. Indeed, Figure 2 shows that the size of the leave-i-out perturbations, i.e., $|(\mathcal{B}_{0,i} \cup \mathcal{B}_{1,i,+})^c|$ scale at a rate which is slower than p. Our main theoretical result proves that the difference between LO and the ALO formula of (9) is proportional to $|(\mathcal{B}_{0,i} \cup \mathcal{B}_{1,i,+})^c|/n$ and thus disappears for large p, n.

Theorem 1. *Under Assumptions A1-A5, let* $1 \leq d_n \leq$

[2] eg. $\kappa_0(n) = (\log p)^{1/6} p^{-\delta}$ where $\delta \in (0, \frac{1}{6})$, and $\kappa_1(n) = p^{-1/12}$.

p/C for some sufficiently large C be such that

$$\max_{1\le i\le n} |(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c| \le d_n$$

with probability at least $1-\tilde{q}_n$. Then we have

$$|\text{ALO}-\text{LO}| \le \frac{\text{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{d_n\text{PolyLog}(n)}{n\lambda^2\eta^2} + \frac{\text{PolyLog}(n)}{\sqrt{n\lambda\eta}}$$

with probability at least $1-(n+1)\mathrm{e}^{-p}-(n+2)p^{-d_n}-2q_n-2\check{q}_n-2\bar{q}_n-2\tilde{q}_n$.

We present the proof of Theorem 1 in Section 3 of the online supplementary material.

Let us now clarify the statement of the theorem. First note that while the bound we have obtained for the difference between ALO and LO is a finite sample bound, one way to interpret and understand the result is through the asymptotic setting we described in Section 3.3.1, i.e. $n,p\to\infty$ while $n/p=\gamma_0$. Theorem 1 shows that ALO, while being computationally much more tractable, is a valid approximation to LOOCV in the regime we have considered. Indeed, according to Theorem 1, suppose λ,η remain fixed and $d_n=o(p^\zeta)$ with some $\zeta<1$, then the upper bound of Theorem 1 is $o(p^{\frac{1}{2}(\zeta-1)}\text{PolyLog}(n))$, which goes to zero as $n,p\to\infty$. Thus, ALO provides a computationally efficient, consistent estimate of the out-of-sample risk.

4 THE EXAMPLE OF LINEAR REGRESSION

Theorem 1 shows that if d_n grows slowly enough (e.g. $d_n=p^\zeta$ for $\zeta<1$), then the difference between ALO and LO will go to zero in probability. To see what the growth rate of d_n in terms of p is, in this section we focus on the concrete example of linear regression and obtain an upper bound for $|(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c|$. Even though the result of this section is given for the linear models, it is expected that a similar conclusion holds for more general models and under more general assumptions. However, given the length of the current paper, the complete investigation of the size of $(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c$ under the generalized linear model is left for a future research. Let us start with our modelling assumptions. In addition to Assumption group A we also assume the following:

Assumption group B

- B1 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}^*+\mathbf{w}$, where $\mathbf{w}\sim N(0,\sigma_w^2 I)$ is the noise or error in the observations.
- B2 The loss function $l(y|\mathbf{x}^\top\boldsymbol{\beta})=\frac{1}{2}(y-\mathbf{x}^\top\boldsymbol{\beta})^2$.
- B3 The true coefficients $\boldsymbol{\beta}^*$ satisfy $\frac{1}{p}\|\boldsymbol{\beta}^*\|_2^2\le\xi$ for some constant $\xi>0$.
- B4 $\mathbf{X}$ has iid entries $X_{ij}\sim N(0,\frac{1}{n})$.
- B5 $\lambda^2(1-\eta)^2=\omega(p^{-\frac{7}{12}})$.

Note that Assumptions B1-B3 are standard in the literature of linear regression. Assumption B4 is also frequently encountered in the high-dimensional asymptotic analysis of estimators (Miolane and Montanari, 2021; Bradic and Chen, 2015; Donoho et al., 2009; Bayati and Montanari, 2012; Weng et al., 2018; Dobriban and Wager, 2018; Wang et al., 2020; Maleki et al., 2013; Thrampoulidis et al., 2015; Rangan, 2011; Li and Wei, 2021). However, it is expected that this assumption can be relaxed as well given the more recent results in the literature (Celentano et al., 2020). Finally, Assumption B5 is a technical assumption on the rate of λ and η. We note here that it shows that for large values of p, one can choose λ to be quite small. The following theorem uses Assumptions B1-B5 to find an upper bound for $|(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c|$.

Theorem 2. *Under Assumptions A1-A5 and B1-B5, in (15) set $\kappa_0=(\frac{8\log p}{c_0 p})^{1/6}$ and $\kappa_1=p^{-1/12}(\log p)^{1/4}$ where c_0 is a positive constant.*[3] *Then there exist constants $C,C'>0$ such that*

$$\max_{1\le i\le n} |(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c| \le Cp^{11/12}(\log p)^{1/4}$$

with probability at least $1-C'p^{-6}-q_n-\mathrm{e}^{-c_0 p}$.

The proof of Theorem 2 can be found in Section 4 of the online supplementary material.

5 CONCLUSION

In this paper, we have introduced a novel theoretical framework that offers error bounds for the disparity between the computationally intensive leave-one-out risk estimate (LO) and its more computationally efficient approximation (ALO). We focus in a regime where n/p and SNR remain bounded regardless of how large n and p grow. For problems in the generalized linear model family such as linear Gaussian, Poisson and logistic, we bound the error between ALO and LO in terms of intuitive metrics such as perturbation size of leave-i-out active sets. Next, for least squares problems with elastic-net regularization, we show that these purturbations scales sub-linearly with n and p, and consequently, the difference $|\text{ALO}-\text{LO}|$ approaches zero as $n,p\to\infty$.

[3] The specific choice of c_0 can be found in Section 4 in the supplementary material.

Acknowledgements

Arian Maleki would like to thank NSF (National Science Foundation) for their generous support through grant number DMS-2210506. Kamiar Rahnama Rad would like to thank NSF (National Science Foundation) for their generous support through grant number DMS-1810888.

References

David M Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation Confidence Intervals for Test Error. *arXiv e-prints*, page arXiv:2007.12671, 2020.

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Jelena Bradic and Jiao Chen. Robustness in sparse linear models: relative efficiency based on robust approximate message passing. *arXiv preprint arXiv:1507.08726*, 2015.

Gavin C Cawley and Nicola LC Talbot. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71:243–264, 2008.

Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

Peter Craven and Grace Wahba. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation. *Numerische Mathematik*, 31:377–403, 1979.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

David Donoho and Andrea Montanari. High Dimensional Robust M-estimation: Asymptotic Variance via Approximate Message Passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

David Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

David Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

Ryan Giordano, Michael I Jordan, and Tamara Broderick. A Higher-Order Swiss Army Infinitesimal Jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A Swiss Army Infinitesimal Jackknife. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1139–1147, 2019b.

Gene Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2): 215–223, 1979.

Yilin Guo, Haolei Weng, and Arian Maleki. Signal-to-noise ratio aware minimaxity and higher-order asymptotics. *arXiv preprint arXiv:2211.05954*, 2022.

Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR, 2013.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

Yuetian Luo, Zhimei Ren, and Rina Barber. Iterative approximate cross-validation. In *International Conference on Machine Learning*, pages 23083–23102. PMLR, 2023.

Arian Maleki. *Approximate message passing algorithm for compressed sensing*. PhD thesis, Stanford University, 2011.

Arian Maleki, Laura Anitori, Zai Yang, and Richard G Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7): 4290–4308, 2013.

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

Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Ali Mousavi, Arian Maleki, and Richard G. Baraniuk. Consistent parameter estimation for LASSO and ap-

proximate message passing. *The Annals of Statistics*, 46(1):119 – 148, 2018.

Tomoyuki Obuchi and Yoshiyuki Kabashima. Cross validation in LASSO and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(5), 2016.

Tomoyuki Obuchi and Yoshiyuki Kabashima. Accelerating Cross-Validation in Multinomial Logistic Regression with ℓ_1-Regularization. *The Journal of Machine Learning Research*, 19(1):2030–2059, 2018.

Tomoyuki Obuchi and Ayaka Sakata. Cross validation in sparse linear regression with piecewise continuous nonconvex penalties and its acceleration. *Journal of Physics A: Mathematical and Theoretical*, 52(41): 414003, 2019.

Manfred Opper and Ole Winther. Gaussian processes and SVM: Mean field results and leave-one-out. 2000.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani. Uniform Consistency of Cross-Validation Estimators for High-Dimensional Ridge Regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3178–3186. PMLR, 2021.

Pratik Patil, Alessandro Rinaldo, and Ryan Tibshirani. Estimating Functionals of the Out-of-Sample Error Distribution in High-Dimensional Ridge Regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6087–6120. PMLR, 2022.

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 2020.

Kamiar Rahnama Rad, Wenda Zhou, and Arian Maleki. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 4067–4077. PMLR, 2020.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

William Stephenson and Tamara Broderick. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2424–2434. PMLR, 2020.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(2):111–147, 1974.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1683–1709, 2015.

Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1):3581–3618, 2016.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Shuaiwen Wang, Wenda Zhou, Arian Maleki, Haihao Lu, and Vahab Mirrokni. Approximate leave-one-out for high-dimensional non-differentiable learning problems. *arXiv preprint arXiv:1810.02716*, 2018.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Does SLOPE outperform bridge regression? *Information and Inference: A Journal of the IMA*, 11(1):1–54, 2022.

Haolei Weng, Arian Maleki, and Le Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics*, 46(6a):3099–3129, 2018.

Ji Xu, Arian Maleki, Kamiar Rahnama Rad, and Daniel Hsu. Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, all such descriptions can be found immediately preceding our main results, i.e., Theorems 1 and 2.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes, all such descriptions can be found immediately preceding our main results, i.e., Theorems 1 and 2.]

(b) Complete proofs of all theoretical results. [Yes, all detailed proofs are given in the online supplement.]

(c) Clear explanations of any assumptions. [Yes, for both groups of assumptions A and B, we explain the assumptions in detail. See Section 3.3 and the paragraph preceding Theorem 2.]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, all details are given in the caption of the figures 1 and 2.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, all details are given in the caption of the figures 1 and 2.]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, all details are given in the caption of the figures 1 and 2.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes, in the caption we mention the Python package used to create our figures.]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Approximate Leave-one-out Cross Validation for Regression with ℓ_1 Regularizers: Supplementary Materials

This supplement is organized as follows. In Section 1, we restate our main results for the reader's convenience. In Sections 2, 3 and 4, we present the proofs of Proposition 1, Theorem 1 and Theorem 2 respectively. Section 5 contains the proofs of lemmas and auxiliary theorems. Finally, the Appendix 6 prepares some results on the elastic net regularized least squares optimization problem which we use in the proof of Theorem 2.

1 MAIN RESULTS

Before providing the proofs, we restate the results in the main paper for completeness.

Approximate Leave One out (ALO) replaces the computationally demanding procedure of repeatedly fitting the model with finding an approximate model that is easy to compute. Instead of exactly computing the leave-one-out estimate $\widehat{\boldsymbol{\beta}}_{/i}$, ALO adjusts the estimate $\widehat{\boldsymbol{\beta}}$ based on the entire dataset $\mathcal{D}$, and uses one Newton step to compute the approximation $\tilde{\boldsymbol{\beta}}_{/i}$ as follows:

$$\tilde{\boldsymbol{\beta}}_{/i} = \widehat{\boldsymbol{\beta}} + \left(\sum_{j \neq i} \mathbf{x}_j \mathbf{x}_j^\top \ddot{\ell}(\mathbf{y}_j | \mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}) + \lambda \nabla^2 r(\widehat{\boldsymbol{\beta}}) \right)^{-1} \mathbf{x}_i \dot{\ell}(y_i | \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) \tag{1}$$

We use the following approximation for LO of the elastic net.

$$\text{ALO} = \frac{1}{n} \sum_{i=1}^{n} \phi \left(y_i, \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})} \right) \left(\frac{H_{ii}}{1 - H_{ii}} \right) \right) \tag{2}$$

where

$$\mathbf{H} := \mathbf{X}_{\mathcal{S}} \left(2\lambda\eta \mathbb{I} + \mathbf{X}_{\mathcal{S}}^\top \text{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})] \mathbf{X}_{\mathcal{S}} \right)^{-1} \mathbf{X}_{\mathcal{S}}^\top \text{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}})].$$

1.1 Assumption group A

The following assumptions have been extensively used in the literature of high-dimensional statistics. We will explain the rationale for making these assumptions in the next section.

A1 $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ where $\mathbf{x}_i \in \mathbb{R}^p$ are iid $N(0, \boldsymbol{\Sigma})$. Moreover, there exist constants $0 < c_X \leq C_X$ such that $p^{-1} c_X \leq \sigma_{\min}(\boldsymbol{\Sigma}) \leq \sigma_{\max}(\boldsymbol{\Sigma}) \leq p^{-1} C_X$.

A2 $n/p = \gamma_0 \in (0, \infty)$.

A3 ϕ has continuous derivative $\dot{\phi}$, and $\ell(y|z)$ has continuous second derivative w.r.t. z.

A4 There exists $\overset{\star}{\epsilon} > 0$, and $q_n, \breve{q}_n, \bar{q}_n \in [0, 1)$, such that

$$\mathbb{P} \left(\sup_{\substack{1 \leq i \leq n \\ \mathbf{v} \in \mathcal{D}}} \dot{\ell}_i(\mathbf{v}) \leq \text{PolyLog}(n) \right) \geq 1 - \breve{q}_n, \tag{3}$$

$$\mathbb{P}\left(\sup_{\substack{1\le i\le n\\ \mathbf{v}\in\mathcal{D}}} \ddot{\ell}_i(\mathbf{v}) \le \mathrm{PolyLog}(n)\right) \ge 1-q_n \tag{4}$$

$$\mathbb{P}\left(\sup_{\substack{1\le i\le n,\\ \mathbf{v},\mathbf{v}'\in\mathcal{D}}} \frac{\|\ddot{\ell}_i(\mathbf{v})-\ddot{\ell}_{/i}(\mathbf{v}')\|}{\|\mathbf{v}-\mathbf{v}'\|_2} \le \mathrm{PolyLog}(n)\right) \ge 1-\bar{q}_n \tag{5}$$

where

$$\begin{aligned}\mathcal{D} &:= \bigcup_{1\le i\le n}\bigcup_{t\in[0,1]} \mathcal{B}(t\widehat{\boldsymbol{\beta}}+(1-t)\widehat{\boldsymbol{\beta}}_{/i}, \overset{\star}{\epsilon}),\\ \mathcal{B}(\mathbf{w},r) &:= \{\mathbf{z}:\|\mathbf{z}-\mathbf{w}\|_2\le r\}.\end{aligned}$$

A5 $\eta\in(0,1)$, and $\lambda\in(0,\lambda_{\max}]$ for an arbitrary constant $\lambda_{\max}>0$.

where

$$\mathcal{D} := \bigcup_{1\le i\le n}\bigcup_{t\in[0,1]} \mathcal{B}(t\widehat{\boldsymbol{\beta}}+(1-t)\widehat{\boldsymbol{\beta}}_{/i}, \overset{\star}{\epsilon}), \text{ and } \mathcal{B}(\mathbf{w},r) := \{\mathbf{z}:\|\mathbf{z}-\mathbf{w}\|_2\le r\}.$$

To show that our assumptions are satisfied for a wide range of regression models, we state

Proposition 1. *Suppose that $\mathbb{P}(|y_i|>\mathrm{PolyLog}(n))\le q_n^y$ for some $q_n^y=o(1/n)$. Furthermore, suppose that $\ell(y|z)$ is three times differentiable with respect to z and that $\ell(y|z),\dot{\ell}(y|z),\ddot{\ell}(y|z)$, and $\dddot{\ell}(y|z)$ grow polynomially in y,z, i.e., there exists a positive integer m and a constant $C>0$ such that*

$$\max\{|\ell(y|z)|,|\dot{\ell}(y|z)|,|\ddot{\ell}(y|z)|,|\dddot{\ell}(y|z)|\}\le C(1+|y|^m+|z|^m)$$

for all (y,z). Then, Assumption A4 holds.

Next, let us define $g(\boldsymbol{\beta})$ as

$$g(\boldsymbol{\beta}) := \frac{1}{\lambda(1-\eta)}\sum_{i=1}^n \dot{\ell}(\mathbf{y}_i|\mathbf{x}_i^\top\boldsymbol{\beta}) - \frac{2\eta}{1-\eta}\boldsymbol{\beta} \tag{6}$$

Finally, we have the main theorem which proves that the difference of ALO and LO goes to zero in the asymptotic regime considered in the paper. Based on the subgradient, define the following subsets of $[p]=\{1,2,3,\ldots,p\}$:

$$\begin{aligned}\mathcal{S}^{(1)} &:= \{k\in[p]:|\widehat{\beta}_k|>\kappa_1(n)\},\\ \mathcal{S}^{(0)} &:= \{k\in[p]:|g(\widehat{\boldsymbol{\beta}})_k|\le 1-\kappa_0(n)\},\end{aligned} \tag{7}$$

We also consider the following sets:

$$\begin{aligned}&\mathcal{B}_{1,i} := \mathcal{S}^{(1)}\cap\mathcal{S}^{(1)}_{/i}, \mathcal{B}_{0,i} := \mathcal{S}^{(0)}\cap\mathcal{S}^{(0)}_{/i}\\ &\mathcal{B}_{1,i,+} := \mathcal{B}_{1,i}\cap\{k:\widehat{\beta}_k\cdot\widehat{\beta}_{/i,k}>0\}\end{aligned} \tag{8}$$

Theorem 1. *Under Assumptions A1-A5, for a sufficiently large constant $C>0$, let $1\le d_n\le p/C$ be such that*

$$\max_{1\le i\le n}|(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c|\le d_n$$

with probability at least $\tilde{q}_n$. Then we have

$$|\mathrm{ALO}-\mathrm{LO}|\le \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}}+\frac{d_n\mathrm{PolyLog}(n)}{n\lambda^2\eta^2}+\sqrt{\frac{\mathrm{PolyLog}(n)}{n\lambda\eta}}$$

with probability at least $1-(n+1)\mathrm{e}^{-p}-(n+2)p^{-d_n}-2q_n-2\check{q}_n-2\bar{q}_n-2\tilde{q}_n$.

Finally, we consider the example of multiple linear regression, with the elastic net penalty, and show that in this case, with high probability, $|(\mathcal{B}_{0,i}\cup\mathcal{B}_{1,i,+})^c|$ indeed grows at a sub-linear rate as $n,p\to\infty$. Let us start with our modelling assumptions:

1.2 Assumption group B

B1 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w}$, where $\mathbf{w} \sim N(0, \sigma_w^2 I)$ is the noise or error in the observations.

B2 The loss function $l(y|\mathbf{x}^\top\boldsymbol{\beta}) = \frac{1}{2}(y - \mathbf{x}^\top\boldsymbol{\beta})^2$.

B3 The true coefficients $\boldsymbol{\beta}^*$ satisfy:

$$\frac{1}{p}\|\boldsymbol{\beta}^*\|_2^2 \leq \xi$$

for some constant $\xi > 0$.

B4 $\mathbf{X}$ has iid entries $X_{ij} \sim N(0, \frac{1}{n})$.

B5 $\lambda^2(1-\eta)^2 = \omega(p^{-\frac{7}{12}})$.

Theorem 2. *Under Assumptions A1-A5 and B1-B5, in (7) set* $\kappa_0 = (\frac{8\log p}{cp})^{1/6}$ *and* $\kappa_1 = p^{-1/12}(\log p)^{1/4}$ *where* c *is a constant* $c > 0$. *Then there exist constants* $C, C' > 0$ *such that*

$$\max_{1\leq i\leq n} |(\mathcal{B}_{1,i,+} \cup \mathcal{B}_{0,i})^c| \leq Cp^{11/12}(\log p)^{1/4}$$

with probability at least $1 - C'p^{-6} - q_n - \mathrm{e}^{-cp}$.

1.3 Bounded Signal to Noise Ratio

We here briefly revisit and explain the issue of 'bounded signal to noise ratio (SNR)' mentioned in Section 1 of the main paper. Although our method and the conclusion of this paper do not rely explicitly on the SNR, the notion of keeping a delicate balance between the signal and noise is one of the most important foundations of our theory. We define the signal to noise ratio as

$$\mathrm{SNR} = \frac{\mathrm{var}(\mathbf{x}_i^\top\boldsymbol{\beta}^*)}{\mathrm{var}(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}^*)}.$$

It can be shown that the SNR is $O_p(1)$ in the three GLM examples in the previous section. In fact, under Assumption A1

$$\mathrm{var}(\mathbf{x}_i^\top\boldsymbol{\beta}^*) = (\boldsymbol{\beta}^*)^\top\boldsymbol{\Sigma}\boldsymbol{\beta}^* \leq \frac{C_X}{p}\|\boldsymbol{\beta}^*\|_2^2 = O(1)$$

and for the three generalized linear models (linear, logistic, and Poisson regression) in Section 3 of the main paper, we have

$$\frac{1}{\mathrm{var}(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}^*))} = \begin{cases} \sigma^{-2} & \text{for linear} \\ \left(\mathrm{e}^{-\frac{\mathbf{x}_i^\top\boldsymbol{\beta}^*}{2}} + \mathrm{e}^{\frac{\mathbf{x}_i^\top\boldsymbol{\beta}^*}{2}}\right)^2 & \text{for logistic} \\ (\log(1+\mathbf{x}_i^\top\boldsymbol{\beta}^*))^{-1} & \text{for Poisson} \end{cases}$$

all of which are $O_p(1)$ using the fact that

$$\mathbf{x}_i^\top\boldsymbol{\beta}^* \sim N(0, C_X\|\boldsymbol{\beta}^*\|_2^2/p) = O_p(1),$$
$$2 \leq \mathrm{e}^{-\mathbf{x}_i^\top\boldsymbol{\beta}^*} + \mathrm{e}^{\mathbf{x}_i^\top\boldsymbol{\beta}^*} + 2 \leq 2(\mathrm{e}^{|\mathbf{x}_i^\top\boldsymbol{\beta}^*|} + 1) = O_p(1).$$

Moreover if we fix $\frac{1}{p}\|\boldsymbol{\beta}^*\|_2^2 = \xi$, then SNR^{-1} is also $O_p(1)$.

2 Proof of Proposition 1

Without loss of generality we assume $C = 1$. The first step is to show that $\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i} = O_p(\text{PolyLog}(n))$. Throughout this proof we use the following notations:

$$
\begin{aligned}
h(\boldsymbol{\beta}) &= \sum_{j=1}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda(1-\eta) \sum_{i=1}^{p} |\beta_i| + \lambda\eta \boldsymbol{\beta}^\top \boldsymbol{\beta}, \\
h_{/i}(\boldsymbol{\beta}) &= \sum_{j\neq i}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda(1-\eta) \sum_{i=1}^{p} |\beta_i| + \lambda\eta \boldsymbol{\beta}^\top \boldsymbol{\beta}, \\
h_{\alpha}(\boldsymbol{\beta}) &= \sum_{j=1}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda r_\alpha(\boldsymbol{\beta}), \\
h_{\alpha,/i}(\boldsymbol{\beta}) &= \sum_{j\neq i}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda r_\alpha(\boldsymbol{\beta}).
\end{aligned} \tag{9}
$$

(a) First note that for all i,

$$
\lambda\eta \|\widehat{\boldsymbol{\beta}}_{/i}\|_2^2 \leq \sum_j \ell(y_j | \mathbf{x}_j \top \widehat{\boldsymbol{\beta}}_{/i}) + \lambda(1-\eta)\|\widehat{\boldsymbol{\beta}}_{/i}\|_1 + \lambda\eta \|\widehat{\boldsymbol{\beta}}_{/i}\|_2^2 \leq \sum_j \ell(y_j|0),
$$

where the last inequality is due to the fact that $h_{/i}(\widehat{\boldsymbol{\beta}}_{/i}) \leq h_{/i}(\mathbf{0})$. Under the event that $\forall i, |y_i| \leq \text{PolyLog}(n)$ which holds with probability at least $1 - nq_n^{(y)}$ according to the assumptions, we have

$$
\max_i \|\widehat{\boldsymbol{\beta}}_{/i}\|^2 \leq \frac{1}{\lambda\eta} \sum_j \ell(y_j|0) \leq \frac{1}{\lambda\eta} \sum_j (|y_j|^m + 1) \leq n(\text{PolyLog}(n))^m.
$$

Therefore

$$
\begin{aligned}
\mathbb{P}(\max_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}| > t) &\leq \sum_i \mathbb{EP}(|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}| > t | \mathbf{X}_{/i}, \mathbf{y}_{/i}) \\
&= \sum_i \mathbb{EP}(|N(0, \frac{\|\widehat{\boldsymbol{\beta}}_{/i}\|^2}{n})| > t | \mathbf{X}_{/i}, \mathbf{y}_{/i}) \\
&= \sum_i \mathbb{EP}(|N(0,1)| > \frac{t\sqrt{n}}{\|\widehat{\boldsymbol{\beta}}_{/i}\|} | \mathbf{X}_{/i}, \mathbf{y}_{/i}) \\
&\leq \mathbb{P}(\max_i \|\widehat{\boldsymbol{\beta}}_{/i}\|^2 > n(\text{PolyLog}(n))^m) \\
&+ n\mathbb{P}\left(|N(0,1)| > \frac{t}{(\text{PolyLog}(n))^{m/2}}\right).
\end{aligned} \tag{10}
$$

Let $t = (\text{PolyLog}(n))^{m/2} \cdot 2\sqrt{\log(n)} := \text{PolyLog}(n)$. Then, we can use (10) to obtain

$$
\mathbb{P}(\max_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}| > \text{PolyLog}(n)) \leq nq_n^{(y)} + 2ne^{-\frac{1}{2}(2\sqrt{\log(n)})^2} \leq nq_n^{(y)} + \frac{2}{n}. \tag{11}
$$

With a similar strategy we can also prove that for any α we have

$$
\mathbb{P}(\max_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}^{\alpha}| > \text{PolyLog}(n)) \leq nq_n^{(y)} + \frac{2}{n}. \tag{12}
$$

(b) Now we set $\alpha = 1$ and work within the event Ξ under which all the following hold:

(a) $\max_i |y_i| \leq \text{PolyLog}(n)$,
(b) $\max_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}| \leq \text{PolyLog}(n)$,
(c) $\max_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}^{1}| \leq \text{PolyLog}(n)$.

(d) $\max_i \|\mathbf{x}_i\| \leq 2\sqrt{C_X}$.

Note that by combining the assumption of theorem with (11), (12), and Lemma 17 we have

$$\mathbb{P}(\Xi) \geq 1 - 3nq_n^{(y)} - \frac{4}{n} - ne^{-p/2}.$$

Under Ξ we have

$$\begin{aligned}\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1_{/i}) &= \dot{\ell}_i(y_i|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^1_{/i}) \\ &\leq 1 + (\text{PolyLog}(n))^m + (\text{PolyLog}(n))^m \\ &= \text{PolyLog}(n).\end{aligned}$$

Next, consider the following first order optimality conditions for $\widehat{\boldsymbol{\beta}}^1$ and $\widehat{\boldsymbol{\beta}}^1_{/i}$:

$$\begin{aligned}\sum_j \mathbf{x}_j \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1) + \lambda \dot{r}_\alpha(\widehat{\boldsymbol{\beta}}^1) = 0, \\ \sum_{j\neq i} \mathbf{x}_j \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1_{/i}) + \lambda \dot{r}_\alpha(\widehat{\boldsymbol{\beta}}^1_{/i}) = 0\end{aligned}$$

By subtracting the two equations and using mean value theorem we have

$$\mathbf{X}^\top \operatorname{diag}(\ddot{\ell}_j(\boldsymbol{\xi}))\mathbf{X}(\widehat{\boldsymbol{\beta}}^1 - \widehat{\boldsymbol{\beta}}^1_{/i}) + \lambda \operatorname{diag}(\ddot{r}_\alpha(\boldsymbol{\xi}))(\widehat{\boldsymbol{\beta}}^1 - \widehat{\boldsymbol{\beta}}^1_{/i}) = -\mathbf{x}_i \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1_{/i}), \tag{13}$$

where $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{\xi}}$ are convex combinations of $\widehat{\boldsymbol{\beta}}^1$ and $\widehat{\boldsymbol{\beta}}^1_{/i}$. Therefore

$$\widehat{\boldsymbol{\beta}}^1_{/i} - \widehat{\boldsymbol{\beta}}^1 = \left[\mathbf{X}^\top \operatorname{diag}(\ddot{\ell}_j(\boldsymbol{\xi}))\mathbf{X} + \lambda \operatorname{diag}(\ddot{r}_\alpha(\tilde{\boldsymbol{\xi}}))\right]^{-1} \mathbf{x}_i \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1_{/i}).$$

Hence, it is straightforward to see that

$$\|\widehat{\boldsymbol{\beta}}^1 - \widehat{\boldsymbol{\beta}}^1_{/i}\| \leq \frac{1}{2\lambda\eta}\|\mathbf{x}_i\||\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^1_{/i})| \leq \frac{1}{2\lambda\eta} 2\sqrt{C_X}\text{PolyLog}_3(n) = \text{PolyLog}(n) \tag{14}$$

given that $\frac{1}{\lambda\eta} = O(\text{PolyLog}(n))$. We can now use (14) to obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\| &\leq \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^1\| + \|\widehat{\boldsymbol{\beta}}^1 - \widehat{\boldsymbol{\beta}}^1_{/i}\| + \|\widehat{\boldsymbol{\beta}}^1_{/i} - \widehat{\boldsymbol{\beta}}_{/i}\| \\ &\leq \text{PolyLog}(n) + 2\sqrt{\frac{4p\log(2)}{\eta}}.\end{aligned} \tag{15}$$

Recall that

$$\mathcal{D} := \cup_{1\leq i\leq n, t\in[0,1]} \mathcal{B}(t\widehat{\boldsymbol{\beta}} + (1-t)\widehat{\boldsymbol{\beta}}_{/i}, \overset{\star}{\epsilon}),$$

where $\mathcal{B}(x, r)$ is the ball with center x and radius r. Hence we can conclude that $\forall i, \forall \mathbf{v} \in \mathcal{D}$:

$$\|\mathbf{v} - \widehat{\boldsymbol{\beta}}_{/i}\| \leq \|\mathbf{v} - \widehat{\boldsymbol{\beta}}\| + \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\| \leq \overset{\star}{\epsilon} + \text{PolyLog}(n) + 8\sqrt{\frac{p\log(2)}{\eta}}, \tag{16}$$

and

$$\begin{aligned}|\mathbf{x}_i^\top \mathbf{v}| &\leq |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}| + |\mathbf{x}_i^\top (\mathbf{v} - \widehat{\boldsymbol{\beta}}_{/i})| \\ &\leq \text{PolyLog}(n) + \|\mathbf{x}_i\|\|\mathbf{v} - \widehat{\boldsymbol{\beta}}_{/i}\| \\ &\leq \text{PolyLog}(n) + 2\sqrt{C_X}(\overset{\star}{\epsilon} + 2\text{PolyLog}(n)) + 8\sqrt{\frac{4C_x p\log(2)}{\eta}} \\ &= \text{PolyLog}(n).\end{aligned}$$

Hence,

$$\begin{aligned}\dot{\ell}_i(\mathbf{v}) &\leq 1 + |y_i|^m + |\mathbf{x}_i^\top \mathbf{v}|^m \\ &\leq 1 + (\text{PolyLog}(n))^m + (\text{PolyLog}(n))^m \\ &= \text{PolyLog}(n)\end{aligned}$$

Using the same arguments one can show that

$$\ddot{\ell}_i(\mathbf{v}) \leq \text{PolyLog}(n)\,,\ \dddot{\ell}_i(\mathbf{v}) \leq \text{PolyLog}(n).$$

This completes the proof. □

3 Proof of Theorem 1

As we mentioned in the introduction, one of ingredients of the proof is the smoothing idea that we would like to describe first. Define

$$h(\boldsymbol{\beta}) := \left\{ \sum_{i=1}^{n} \ell(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda(1-\eta)\|\boldsymbol{\beta}\|_1 + \lambda\eta\|\boldsymbol{\beta}\|_2^2 \right\}. \tag{17}$$

If $h(\boldsymbol{\beta})$ were a differentiable function we could use one step of the Newton method to obtain an approximate leave-one-out estimate. However, the main issue here is the non-differentiability of $\|\cdot\|_1$. For that reason we start with approximating the $\|\cdot\|_1$ with a smooth function. Let

$$r_\alpha^{(1)}(z) = \frac{1}{\alpha}\left(\log(1+e^{\alpha z}) + \log(1+e^{-\alpha z})\right) \tag{18}$$

denote the α-smoothed l_1 regularizer. The following lemma proved in (Rahnama Rad and Maleki, 2020) shows the accuracy of this approximation:

Lemma 1 (Lemma 13 in (Rahnama Rad and Maleki, 2020)). *If $r_\alpha^{(1)}(z)$ denotes the α-smoothed l_1 regularizer. Then we have*

$$r_\alpha^{(1)}(z) \geq |z|,$$

and

$$\sup_z |r_\alpha^{(1)}(z) - |z|| \leq \frac{2\log 2}{\alpha}.$$

This lemma suggests that for large values of α, $r_\alpha^{(1)}(z)$ can be a good approximation of $|z|$. Hence, based on this approximation we now introduce the smoothed cost function

$$h_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda r_\alpha(\boldsymbol{\beta}), \tag{19}$$

where

$$r_\alpha(\boldsymbol{\beta}) := (1-\eta)\sum_{i=1}^{p} r_\alpha^{(1)}(\beta_i) + \eta\boldsymbol{\beta}^\top\boldsymbol{\beta} \tag{20}$$

denotes the smoothed regularizer. If we define

$$\widehat{\boldsymbol{\beta}}^\alpha = \arg\min_{\boldsymbol{\beta}} h_\alpha(\boldsymbol{\beta}), \tag{21}$$

and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ as its leave-one-out estimate, then we can use Theorem 3 of (Rahnama Rad and Maleki, 2020) to prove that

$$\max_{1\leq i\leq n}\left|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^\alpha_{/i} - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^\alpha - \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)}{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)}\right)\left(\frac{H_{ii}^\alpha}{1-H_{ii}^\alpha}\right)\right| \leq \frac{C_0(\alpha)\text{PolyLog}(n)}{\sqrt{p}}, \tag{22}$$

where $\mathbf{H}^\alpha$ is defined as

$$\mathbf{X}\left(\lambda\text{diag}[\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)] + \mathbf{X}^\top[\text{diag}(\ddot{\ell}(\boldsymbol{\beta}^\alpha))]\mathbf{X}\right)^{-1}\mathbf{X}^\top\text{diag}[\ddot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha)]. \tag{23}$$

For completeness we have mentioned Theorem 3 of (Rahnama Rad and Maleki, 2020) in Section 5 (Theorem 6). The main issue in the approximation of (22) is that $C_0(\alpha) \to \infty$ as $\alpha \to \infty$. This creates a dilemma. On one hand we would like α to be large to make $|r_\alpha^{(1)}(z) - |z||$ small. But on the other hand, the upper bound in Theorem 3 of (Rahnama Rad and Maleki, 2020) goes to infinity as $\alpha \to \infty$. In addition to these two problems, as will be discussed later, some of the elements of $\text{diag}[\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)]$ go to infinity as $\alpha \to \infty$ that may cause inaccuracies and instabilities if this procedure is used in practice. Despite the fact that smoothing idea is not useful for approximating the leave-one-out risk of the elastic net, as will be shown in this proof, it still serves as a good theoretical tool for proving the accuracy of (2). Hence, we pursue two goals here:

- Use a different strategy than the one pursued in (Rahnama Rad and Maleki, 2020) to find an upper bound on $\max_{1\le i\le n}\left|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^\alpha_{/i} - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^\alpha - \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)}{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)}\right)\left(\frac{H^\alpha_{ii}}{1-H^\alpha_{ii}}\right)\right|$. Our new bounds will not go off to infinity as $\alpha \to \infty$.
- We will then prove that for large values of α, $\mathbf{H}^\alpha_{ii}$ is close to $\mathbf{H}_{ii}$ used in (2).

To understand the challenge for achieving the above two goals, let us start with the following lemma:

Lemma 2 (Lemma 14 in (Rahnama Rad and Maleki, 2020)). *$r_\alpha^{(1)}(z)$ is infinitely many times differentiable, and*

$$\dot{r}_\alpha^{(1)}(z) = \frac{e^{\alpha z} - e^{-\alpha z}}{e^{\alpha z} + e^{-\alpha z} + 2},$$

$$\ddot{r}_\alpha^{(1)}(z) = \frac{2\alpha}{(e^{\alpha z} + e^{-\alpha z} + 2)^2}.$$

Suppose that $z_\alpha = O(\frac{1}{\alpha})$. Then, as $\alpha \to \infty$, $\ddot{r}_\alpha^{(1)}(z) \to \infty$. It may seem to the reader that $z_\alpha = O(\frac{1}{\alpha})$ is a condition that may not happen and hence it won't cause any issues. However, this is not the case. In fact, as will be shown in the next lemma, many of the regression coefficients satify the condition $|\widehat{\boldsymbol{\beta}}^\alpha_{/i,k}| = O(\frac{1}{\alpha}\log p)$. For these elements as $\alpha \to \infty$, $\ddot{r}(\widehat{\boldsymbol{\beta}}^\alpha_{/i,k}) \to \infty$.

Lemma 3. *Suppose Assumptions A1-A5 hold. Let $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(0)}$ be as defined in (7). Then we have:*

1. $\max_{1\le i\le n} \|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\| \le \sqrt{\frac{4\log(2)p}{\alpha\eta}}$.
2. $\|\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\| \le \frac{|\dot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha)|\|\mathbf{x}_i\|}{2\lambda\eta}$
3. $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\| \le \frac{|\dot{\ell}(\widehat{\boldsymbol{\beta}})|\|\mathbf{x}_i\|}{2\lambda\eta}$.
4. $\max_{1\le i\le n} \|g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i})\| \le \frac{\text{PolyLog}(n)}{\lambda^2\eta(1-\eta)}$ *with probability at least* $1 - q_n - \mathrm{e}^{-p} - \check{q}_n - ne^{-p/2}$.
5. *Suppose* $\alpha = \omega\left(\frac{p}{\kappa_1^2\eta}\right)$, *then for large enough* p, $\min_{0\le i\le n}\min_{k\in\mathcal{S}^{(1)}_{/i}} |\widehat{\beta}^\alpha_{/i,k}| \ge \frac{\kappa_1}{2}$.
6. *Suppose* $\alpha = \omega\left(\frac{n\text{PolyLog}(n)}{\kappa_0^2\lambda^2(1-\eta)\eta}\right)$, *then for large enough* p, *with probability at least* $1 - q_n - e^{-p}$:

$$\max_{0\le i\le n}\max_{k\in\mathcal{S}^{(0)}_{/i}} |\widehat{\beta}^\alpha_{/i,k}| \le \frac{1}{\alpha}\log\left(\frac{4}{\kappa_0}\right).$$

The proof of this lemma is presented in Section 5.2.

A source of difficulty in handling the smoothed regularizer is that we do not have much control over the curvature, $\ddot{r}_\alpha^{(1)}(\widehat{\beta}^\alpha_{/i,k})$ when $k \in (\mathcal{S}^{(1)}_{/i} \cup \mathcal{S}^{(0)}_{/i})^c$. Keeping this issue in mind, let us first simplify the error between the leave-one-out cross validation risk and the ALO for the smoothed problem. To simplify the calculations we introduce the following notations. Let $\dot{r}_\alpha(\boldsymbol{\theta})$ denote the vector $[\dot{r}_\alpha(\theta_1), \dot{r}_\alpha(\theta_2), \ldots, \dot{r}_\alpha(\theta_p)]^\top$. Similarly,

$$\dot{\ell}(\boldsymbol{\theta}) := [\dot{\ell}(y_1; \mathbf{x}_1^\top\boldsymbol{\theta}), \dot{\ell}(y_2; \mathbf{x}_2^\top\boldsymbol{\theta}), \ldots, \dot{\ell}(y_n; \mathbf{x}_n^\top\boldsymbol{\theta})]^T$$

and $\dot{\ell}_{/i}(\boldsymbol{\theta})$ is the same vector as $\dot{\ell}(\boldsymbol{\theta})$ except that its i^{th} element is removed. Furthermore, define $\mathbf{f}_{/i}(\boldsymbol{\theta})$ as the gradient of $h_\alpha(\cdot)$ at $\boldsymbol{\theta}$, i.e.,

$$\mathbf{f}_{/i}(\boldsymbol{\theta}) := \lambda \dot{r}_\alpha(\boldsymbol{\theta}) + \mathbf{X}_{/i}^\top \dot{\ell}_{/i}(\boldsymbol{\theta}). \tag{24}$$

Notice that $\mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) = 0$, where $\widehat{\boldsymbol{\beta}}_{/i}^\alpha$ is the true LO estimate. Similarly, define th Hessian of $h_\alpha(\cdot)$ and its leave-one-out coounter part as

$$\mathbf{J}(\boldsymbol{\theta}) = \lambda \text{diag}(\ddot{r}_\alpha(\boldsymbol{\theta})) + \mathbf{X}^\top \text{diag}[\ddot{\ell}(\boldsymbol{\theta})]\mathbf{X}, \tag{25}$$

and

$$\mathbf{J}_{/i}(\boldsymbol{\theta}) = \lambda \text{diag}(\ddot{r}_\alpha(\boldsymbol{\theta})) + \mathbf{X}_{/i}^\top \text{diag}[\ddot{\ell}_{/i}(\boldsymbol{\theta})]\mathbf{X}_{/i}. \tag{26}$$

By using the first order optimality conditions, we have

$$\begin{aligned} \mathbf{f}(\widehat{\boldsymbol{\beta}}^\alpha) &= 0, \\ \mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) &= 0. \end{aligned} \tag{27}$$

Define $\Delta_{/i}^\alpha := \widehat{\boldsymbol{\beta}}_{/i}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha$. Using the multivariate mean-value theorem we have

$$\begin{aligned} 0 &= \mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) \\ &= \mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha + \Delta^*{}_{/i}) \\ &= \mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha) + \left(\int_0^1 \mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha + t\Delta_{/i}^\alpha) dt \right) \Delta_{/i}^\alpha. \end{aligned} \tag{28}$$

Moreover,

$$0 = \lambda \dot{r}(\widehat{\boldsymbol{\beta}}^\alpha) + X^\top \dot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha) = \mathbf{f}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha) + \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)\mathbf{x}_i. \tag{29}$$

Combining (28) and (29) we have

$$\begin{aligned} \Delta_{/i}^\alpha &= \widehat{\boldsymbol{\beta}}_{/i}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha \\ &= -\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha) \left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}_{/i}^\alpha) dt \right)^{-1} \mathbf{x}_i. \end{aligned} \tag{30}$$

As is clear from (1), the ALO approximation of $\Delta_{/i}^\alpha$ is given by

$$\widehat{\Delta}_{/i}^\alpha = \dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha) \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha - \Delta_{/i}^\alpha) \right)^{-1} \mathbf{x}_i. \tag{31}$$

Also, it is straightforward to see that

$$\begin{aligned} &|\text{ALO}^\alpha - \text{LO}^\alpha| \\ &= \frac{1}{n} \sum_{i=1}^n (\phi(y_i, \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}^\alpha) - \phi(y_i, \mathbf{x}_i^\top (\widehat{\boldsymbol{\beta}}^\alpha + \widehat{\Delta}_{/i}^\alpha)) \\ &\le \max_{1 \le i \le n} \left| \dot{\phi}(y_i, \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}_i^\alpha) \right| \cdot \frac{1}{n} \cdot \sum_{i=1}^n \left| \mathbf{x}_i^\top \Delta_i^\alpha - \mathbf{x}_i^\top \widehat{\Delta}_{/i}^\alpha \right|, \end{aligned} \tag{32}$$

where $\tilde{\boldsymbol{\beta}}_i^\alpha$ is a point on the line that connects $\widehat{\boldsymbol{\beta}}^\alpha$ and $\widehat{\boldsymbol{\beta}}_{/i}^\alpha$. Similar to the proof of Proposition 1, we can see that for many reasonable models, $\max_{1 \le i \le n} \left| \dot{\phi}(y_i, \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}_i^\alpha) \right| = O_p(\text{PolyLog}(n))$. Hence, it is enough to obtain an upper bound for $\mathbf{x}_i^\top \Delta_i^\alpha - \mathbf{x}_i^\top \widehat{\Delta}_{/i}^\alpha$. From equations (30) and (31), we have

$$\begin{aligned} &\left| \mathbf{x}_i^\top \Delta_{/i}^\alpha - \mathbf{x}_i^\top \widehat{\Delta}_{/i}^\alpha \right| \\ &\le |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)| \, \mathbf{x}_i^\top \left[\left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}_{/i}^\alpha) dt \right)^{-1} - \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha - \Delta_{/i}^\alpha) \right)^{-1} \right] \mathbf{x}_i \\ &\le |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)| \mathbf{x}_i^\top \left[\left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}_{/i}^\alpha) dt \right)^{-1} - \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) \right)^{-1} \right] \mathbf{x}_i \\ &+ |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)| \mathbf{x}_i^\top \left[\left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) \right)^{-1} - \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha - \Delta_{/i}^\alpha) \right)^{-1} \right] \mathbf{x}_i. \end{aligned} \tag{33}$$

We again emphasize that we cannot let $\alpha \to \infty$ in these expressions, since some of the elements of $\mathbf{J}$ matrix go to infinity. Hence we have to find proper ways for obtaining an upper bound for (33) for large values of α. As is clear from this discussion, we have to be careful about $\ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k})$. The next lemma provides some information about these quantities:

Lemma 4. *Suppose the assumptions of Lemma 3 hold, and assume* $\alpha = \omega\left(\frac{n\text{PolyLog}(n)}{\kappa_0^2\lambda^2(1-\eta)\eta} \vee \frac{p}{\kappa_1^2\eta}\right)$. *Then the following statements hold with probability at least* $1 - q_n - e^{-p}$*:*

1. $\forall i, \forall k \in \mathcal{B}_{0,i}$*:*
$$\int_0^1 \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k} - t\Delta^\alpha_{/i,k})dt \geq 2\eta + \frac{1}{8}\alpha(1-\eta)\kappa_0$$
2. $\forall i, \forall k \in \mathcal{B}_{0,i}$*:*
$$\ddot{r}_\alpha(\widehat{\beta}^\alpha_k),\ \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k}) \geq 2\eta + \frac{1}{8}\alpha(1-\eta)\kappa_0.$$
3. $\forall i, \forall k \in \mathcal{B}_{1,i,+}$*:*
$$\left|\int_0^1 \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k} - t\Delta^\alpha_{/i,k})dt - 2\eta\right| \leq 2\alpha e^{-\frac{1}{2}\alpha\kappa_1}.$$
4. $\forall i, \forall k \in \mathcal{B}_{1,i,+}$*:*
$$\left|\ddot{r}_\alpha(\widehat{\beta}^\alpha_k) - 2\eta\right| \leq 2\alpha e^{-\frac{1}{2}\alpha\kappa_1}\ ;\ \left|\ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k}) - 2\eta\right| \leq 2\alpha e^{-\frac{1}{2}\alpha\kappa_1}.$$

Note that the rate assumption on α ensures that both rates hold in both Part 5 and Part 6 of Lemma 3. The proof of Lemma 4 is presented in Section 5.3. This lemma confirms that we have some control over the curvature of the regularizer on sets $\mathcal{B}_{0,i}$ and $\mathcal{B}_{1,i,+}$. The following lemma enables us to obtain an upper bound for the error between the leave-one-out estimate and ALO by finding an upper bound on the error over the set $\mathcal{B}_0^c$. In other words, the next lemma enables us to remove the indices k for which we are certain $\ddot{r}_\alpha(\widehat{\beta}^\alpha_k)$ converge to infinity from our analysis.

Theorem 3. *Suppose Assumptions A1-A5 hold, the conclusions of Lemma 3 are true, and* $\alpha \geq \dfrac{4p\log(2)}{(\overset{\star}{\epsilon})^2\eta(1-\eta)}$. *Then we have*

$$\begin{aligned}
&\left|\mathbf{x}_i^\top\left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i})dt\right)^{-1}\mathbf{x}_i - \mathbf{x}_{i,\mathcal{B}^c_{0,i}}^\top\left(\lambda\text{diag}(\underline{\ddot{r}}^{\alpha/i}_{\mathcal{B}^c_{0,i}}) + \mathbf{X}^\top_{/i,\mathcal{B}^c_{0,i}}\text{diag}(\underline{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{/i,\mathcal{B}^c_{0,i}}\right)^{-1}\mathbf{x}_{i,\mathcal{B}^c_{0,i}}\right| \\
&\leq \frac{16\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)\kappa_0}\left(\frac{\text{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|}{2\lambda\eta} + 1\right)^2, \qquad (34)
\end{aligned}$$

and similarly

$$\begin{aligned}
&\left|\mathbf{x}_i^\top\left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right)^{-1}\mathbf{x}_i - \mathbf{x}_{i,\mathcal{B}^c_{0,i}}^\top(\lambda\text{diag}(\underset{\sim}{\ddot{r}}^{\alpha/i}_{\mathcal{B}^c_{0,i}}) + \mathbf{X}^\top_{\mathcal{B}^c_{0,i}}\text{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{B}^c_{0,i}})^{-1}\mathbf{x}_{i,\mathcal{B}^c_{0,i}}\right| \\
&\leq \frac{16\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)\kappa_0}\left(\frac{\text{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|}{2\lambda\eta} + 1\right)^2 \qquad (35)
\end{aligned}$$

Here $\underline{\ddot{r}}^{\alpha/i}_k := \int_{-1}^0(\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i} + t\Delta^\alpha_{/i}))_k dt$, $\underline{\ddot{\ell}}^{\alpha/i}_k := \int_{-1}^0(\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i} + t\Delta^\alpha_{/i}))_k dt$, $\underset{\sim}{\ddot{r}}^{\alpha/i} := \ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \Delta^\alpha_{/i})$ *and* $\underset{\sim}{\ddot{\ell}}^{\alpha/i} := \ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \Delta^\alpha_{/i})$.

The complete proof of this theorem is presented in Section 5.4.

Applying (33) and Theorem 3, we have reduced the problem to the quadratic forms on the subset $\mathcal{B}_0^c$. The main remaining difficulty is that for the indices outside $\mathcal{B}^c_{0,i} \setminus \mathcal{B}_{1,i,+}$ we do not have much control over $\ddot{r}_\alpha(\widehat{\beta}^\alpha_k)$. So the question is whether those terms can cause any issue in our approximations or not. Our next theorem will show that these elements will not cause any issue if there are not too many of them. More specifically, in the asymptotic regime we are interested in, i.e. the asymptotic regime in which n, p grow at the same rate, if the size of the set $|\mathcal{B}_{0,i} \setminus \mathcal{B}_{1,i,+}|$ is sublinear in n, then the difference between ALO and LO converges to zero.

Theorem 4. *Suppose the assumptions of Lemma 3 hold, and assume* $\alpha = \omega\left(\frac{n\mathrm{PolyLog}(n)}{\kappa_0^2\kappa_1^2\lambda(1-\eta))\eta}\right)$. *Moreover, for a sufficiently large constant* $C > 0$, *let* $1 \le d_n \le p/C$ *be such that*

$$\max_{1\le i\le n} |\mathcal{B}_{0,i}^c \setminus \mathcal{B}_{1,i,+}| \le d_n$$

with probability at least $1-\tilde{q}_n$. *Let* $\mathcal{F}$ *denote a set such that* $\mathcal{B}_{1,i,+} \subset \mathcal{F} \subset \mathcal{B}_{0,i}^c$. *Then we have*

$$\begin{aligned}
&\left|\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\mathrm{diag}(\ddot{r}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\mathrm{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}} \;-\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\mathrm{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\mathrm{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}}\right| \\
&\le \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \sqrt{\frac{C\log p}{n\lambda\eta}}
\end{aligned}$$

with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}} - (n+2)p^{-d_n} - 2q_n - 2\check{q}_n - 2\bar{q}_n - 2\tilde{q}_n$, *for sufficiently large* p.

The proof of this claim is long and will be presented in Section 5.5. We will now use this theorem to complete the proof of Theorem 1.

First note that the leave-one-out cross validation risk of elastic net and smoothed elastic-net are

$$\begin{aligned}
\mathrm{LO}(\lambda) &= \frac{1}{n}\sum_{i=1}^n \phi(y_i, \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{/i}). \\
\mathrm{LO}^\alpha(\lambda) &= \frac{1}{n}\sum_{i=1}^n \phi(y_i, \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{/i}^\alpha).
\end{aligned} \tag{36}$$

Hence the difference of the two is

$$\begin{aligned}
|\mathrm{LO}(\lambda) - \mathrm{LO}^\alpha(\lambda)| &\le \max_{i,z_i}|\dot{\phi}(y_i,z_i)||\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}_{/i}^\alpha)| \\
&\le \max_{i,z_i}|\dot{\phi}(y_i,z_i)|\|\mathbf{x}_i\|\|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}_{/i}^\alpha)\| \\
&\le \max_{i,z_i}|\dot{\phi}(y_i,z_i)|\|\mathbf{x}_i\|\sqrt{\frac{4\log 2p}{\alpha\eta}}.
\end{aligned} \tag{37}$$

According to (32) we have

$$|\mathrm{ALO}^\alpha - \mathrm{LO}^\alpha| \le \max_{1\le i\le n}\left|\dot{\phi}(y_i, \mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_i^\alpha)\right| \cdot \frac{1}{n}\cdot\sum_{i=1}^n\left|\mathbf{x}_i^\top\Delta_i^\alpha - \mathbf{x}_i^\top\widehat{\Delta}_{/i}^\alpha\right|, \tag{38}$$

Combining (36), (37), and (38), we have

$$|\mathrm{ALO}^\alpha - \mathrm{LO}| \le \max_{i,z_i}|\dot{\phi}(y_i,z_i)|\|\mathbf{x}_i\|\sqrt{\frac{4\log 2p}{\alpha\eta}} + \max_{1\le i\le n}\left|\dot{\phi}(y_i, \mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_i^\alpha)\right| \cdot \frac{1}{n}\cdot\sum_{i=1}^n\left|\mathbf{x}_i^\top\Delta_i^\alpha - \mathbf{x}_i^\top\widehat{\Delta}_{/i}^\alpha\right|. \tag{39}$$

Furthermore, according to (33), Theorem 3, and Theorem 4 we have that with probability at least $1-(n+$

$1)\mathrm{e}^{-\frac{p}{2}} - (n+2)p^{-d_n} - 2q_n - 2\check{q}_n - 2\tilde{q}_n - 2\bar{q}_n$:

$$\begin{aligned}
&\left|\mathbf{x}_i^\top \Delta_{/i}^\alpha - \mathbf{x}_i^\top \widehat{\Delta}_{/i}^\alpha\right| \qquad (40)\\
\leq& |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\ \mathbf{x}_i^\top \left[\left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}_{/i}^\alpha)dt\right)^{-1} - \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right)^{-1}\right]\mathbf{x}_i\\
&+ |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\mathbf{x}_i^\top \left[\left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right)^{-1} - \left(\mathbf{J}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha - \Delta_{/i}^\alpha)\right)^{-1}\right]\mathbf{x}_i,\\
\leq& |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\Big|\mathbf{x}_{i,\mathcal{B}_{0,i}^c}^\top(\lambda\mathrm{diag}(\ddot{r}_{\mathcal{B}_{0,i}^c}^{\alpha/i}) + \mathbf{X}_{/i,\mathcal{B}_{0,i}^c}^\top \mathrm{diag}(\ddot{\underline{\ell}}^{\alpha/i})\mathbf{X}_{/i,\mathcal{B}_{0,i}^c})^{-1}\mathbf{x}_{i,\mathcal{B}_{0,i}^c}\\
&- \mathbf{x}_{i,\mathcal{B}_{0,i}^c}^\top(\lambda\mathrm{diag}(\ddot{r}_{\sim\mathcal{B}_{0,i}^c}^{\alpha/i}) + \mathbf{X}_{\mathcal{B}_{0,i}^c}^\top \mathrm{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{B}_{0,i}^c})^{-1}\mathbf{x}_{i,\mathcal{B}_{0,i}^c}\Big|\\
&+ \frac{32|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)\kappa_0}\Big(\frac{\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|}{2\lambda\eta} + 1\Big)^2\\
\leq& |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\left(\frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \sqrt{\frac{C\log p}{n\lambda\eta}}\right)\\
&+ \frac{32|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)\kappa_0}\Big(\frac{\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|}{2\lambda\eta} + 1\Big)^2. \qquad (41)
\end{aligned}$$

As is clear from this equation, as $\alpha\to\infty$, and for large values of n, p, if d_n grows slowly enough in n (or equivalently in p) the difference $|\mathrm{ALO}^\alpha - \mathrm{LO}|$ will be negligible. The last step of the proof, is to show that the difference between $|\mathrm{ALO}^\alpha - \mathrm{ALO}|$ is also negligible. Note that our approximate ALO formula for elastic net can be written as

$$\mathrm{ALO} = \frac{1}{n}\sum_{i=1}^n \phi(y_i, \mathbf{x}_{i,\mathcal{S}}^\top\widehat{\boldsymbol{\beta}}_{\mathcal{S}} + \mathbf{x}_{i,\mathcal{S}}^\top\widehat{\Delta}_{/i}), \qquad (42)$$

where

$$\widehat{\Delta}_{/i} = \dot{\ell}(\widehat{\boldsymbol{\beta}})(2\lambda\eta\mathbb{I} + \mathbf{X}_{\mathcal{S}}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{x}_{i,\mathcal{S}}.$$

Hence, we have

$$\begin{aligned}
&|\mathrm{ALO} - \mathrm{ALO}^\alpha|\\
=& \frac{1}{n}\sum_{i=1}^n |\phi(y_i, \mathbf{x}_{i,\mathcal{S}}^\top(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} + \widehat{\Delta}_{/i})) - \phi(y_i, (\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}^\alpha + \widehat{\Delta}_{/i}^\alpha))|\\
\leq& \max_{i,z_i}|\dot{\phi}(y_i,z_i)||(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}^\alpha - \mathbf{x}_{i,\mathcal{S}}^\top\widehat{\boldsymbol{\beta}}_{\mathcal{S}}| + \max_{i,z_i}|\dot{\phi}(y_i,z_i)||\mathbf{x}_{i,\mathcal{S}}^\top\widehat{\Delta}_{/i} - \mathbf{x}_i^\top\widehat{\Delta}_{/i}^\alpha|\\
\leq& \max_{i,z_i}|\dot{\phi}(y_i,z_i)|\|\mathbf{x}_i\|\sqrt{\frac{4\log 2p}{\alpha\eta}} + \max_{i,z_i}|\dot{\phi}(y_i,z_i)||\mathbf{x}_{i,\mathcal{S}}^\top\widehat{\Delta}_{/i} - \mathbf{x}_i^\top\widehat{\Delta}_{/i}^\alpha|, \qquad (43)
\end{aligned}$$

where to obtain the last inequality we have used the first part of Lemma 3.

Similar to the proof of Theorem 4 we can prove that

$$\begin{aligned}
&|\mathbf{x}_{i,\mathcal{S}}^\top\widehat{\Delta}_{/i} - \dot{\ell}(\widehat{\boldsymbol{\beta}})\mathbf{x}_{i,\mathcal{B}_{1,i,+}}^\top(2\lambda\eta\mathbb{I} + \mathbf{X}_{\mathcal{B}_{1,i,+}}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{B}_{1,i,+}})^{-1}\mathbf{x}_{i,\mathcal{B}_{1,i,+}}|\\
\leq& \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \sqrt{\frac{C\log p}{n\lambda\eta}} \qquad (44)
\end{aligned}$$

with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}} - (n+2)p^{-d_n} - 2q_n - 2\check{q}_n - 2\tilde{q}_n - 2\bar{q}_n$, for sufficiently large p. Similarly,

$$\begin{aligned}
&|\mathbf{x}_i^\top\widehat{\Delta}_{/i}^\alpha - \dot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha)\mathbf{x}_{i,\mathcal{B}_{1,i,+}}^\top(\lambda\mathrm{diag}(\ddot{r}^\alpha(\widehat{\boldsymbol{\beta}}^\alpha)) + \mathbf{X}_{\mathcal{B}_{1,i,+}}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha))\mathbf{X}_{\mathcal{B}_{1,i,+}})^{-1}\mathbf{x}_{i,\mathcal{B}_{1,i,+}}|\\
\leq& \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \sqrt{\frac{C\log p}{n\lambda\eta}} \qquad (45)
\end{aligned}$$

with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}}-(n+2)p^{-d_n}-2q_n-2\check{q}_n-2\tilde{q}_n-2\bar{q}_n$, for sufficiently large p. In the rest of the proof, for the notational simplicity, we use the notation $\mathcal{B}_+$ instead of $\mathcal{B}_{1,i,+}$.

$$
\begin{aligned}
&|\mathbf{x}_{i,\mathcal{S}}^\top\widehat{\Delta}_{/i}-\mathbf{x}_i^\top\widehat{\Delta}_{/i}^{\alpha}|\\
&\leq|\dot{\ell}(\widehat{\boldsymbol{\beta}})\mathbf{x}_{i,\mathcal{B}_+}^\top(2\lambda\eta\mathbb{I}+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\\
&-\dot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha})\mathbf{x}_{i,\mathcal{B}_+}^\top(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}|\\
&+\frac{2\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}}+\frac{Cd_n}{n\lambda^2\eta^2}+\sqrt{\frac{C\log p}{n\lambda\eta}}\\
&\leq|\dot{\ell}(\widehat{\boldsymbol{\beta}})|\times\\
&\times\big|\mathbf{x}_{i,\mathcal{B}_+}^\top(2\lambda\eta\mathbb{I}+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\\
&-\mathbf{x}_{i,\mathcal{B}_+}^\top(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\big|\\
&+|\dot{\ell}(\widehat{\boldsymbol{\beta}})-\dot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha})|\times\\
&\times\mathbf{x}_{i,\mathcal{B}_+}^\top(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\\
&+\frac{2\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}}+\frac{Cd_n}{n\lambda^2\eta^2}+\sqrt{\frac{C\log p}{n\lambda\eta}}
\end{aligned}
\tag{46}
$$

Since the minimum eigenvalue of $(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})$ is larger than $2\lambda\eta$, we can conclude that

$$
\begin{aligned}
&|\dot{\ell}(\widehat{\boldsymbol{\beta}})-\dot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha})|\mathbf{x}_{i,\mathcal{B}_+}^\top(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\\
\leq&\frac{\|\mathbf{x}_i\|_2^2}{2\lambda\eta}|\dot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha})-\dot{\ell}(\widehat{\boldsymbol{\beta}})|\overset{(a)}{=}\frac{\|\mathbf{x}_i\|_2^2}{2\lambda\eta}|\ddot{\ell}(\theta)(\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}))|\\
\leq&\frac{\|\mathbf{x}_i\|_2^3}{2\lambda\eta}|\ddot{\ell}(\boldsymbol{\theta})|\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}\|\\
\overset{(b)}{\leq}&\frac{\|\mathbf{x}_i\|_2^3}{2\lambda\eta}|\ddot{\ell}(\boldsymbol{\theta})|\sqrt{\frac{4\log(2)p}{\alpha\eta}}\\
\overset{(c)}{\leq}&\frac{\mathrm{PolyLog}(n)}{\lambda\eta}\sqrt{\frac{p}{\alpha\eta}},
\end{aligned}
\tag{47}
$$

with probability larger than $1-n\mathrm{e}^{-p}-q_n-\check{q}_n$. To obtain Equality (a) we have used the mean value theorem and $\boldsymbol{\theta}=t\widehat{\boldsymbol{\beta}}+(1-t)\widehat{\boldsymbol{\beta}}^{\alpha}$ for some $t\in[0,1]$. To obtain inequality (b) we have used Part 1 of Lemma 3. Inequality (c) is based on Assumption A.4 along with Lemma 17.

Also, using Lemma 13 we have that

$$
\begin{aligned}
&|\mathbf{x}_{i,\mathcal{B}_+}^\top(2\lambda\eta\mathbb{I}+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}\\
&-\mathbf{x}_{i,\mathcal{B}_+}^\top(\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+})^{-1}\mathbf{x}_{i,\mathcal{B}_+}|\\
&\leq\frac{\|\mathbf{x}_i\|^2\lambda_{\max}(\boldsymbol{\Gamma})}{(2\lambda\eta)^2}+\frac{\|\mathbf{x}_i\|^2\lambda_{\max}^2(\boldsymbol{\Gamma})}{(2\eta\lambda)^2(2\lambda\eta-\lambda_{\max}(\boldsymbol{\Gamma}))},
\end{aligned}
\tag{48}
$$

where

$$
\boldsymbol{\Gamma}:=\lambda\mathrm{diag}(\ddot{r}^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}))+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\mathbf{X}_{\mathcal{B}_+}-2\lambda\eta\mathbb{I}+\mathbf{X}_{\mathcal{B}_+}^\top\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}))\mathbf{X}_{\mathcal{B}_+}).
\tag{49}
$$

Therefore, by using Weyl's theorem

$$\begin{aligned}\lambda_{\max}(\mathbf{\Gamma}) &\leq \lambda_{\max}(\lambda \mathrm{diag}(\ddot{r}^{\alpha}_{\mathcal{B}_+}(\widehat{\boldsymbol{\beta}}^{\alpha})) - 2\lambda\eta\mathbb{I}) + \lambda_{\max}(\mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}})) - \mathrm{diag}(\ddot{\ell}(\widehat{\boldsymbol{\beta}}^{\alpha}))\|\mathbf{X}^\top\mathbf{X}\| \\ &\leq 2\lambda\alpha \mathrm{e}^{-\frac{1}{2}\alpha\kappa_1} + \|\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}\|\|\mathbf{X}^\top\mathbf{X}\|\mathrm{PolyLog}(n) \\ &\leq 2\lambda\alpha \mathrm{e}^{-\frac{1}{2}\alpha\kappa_1} + \sqrt{\frac{4\log 2p}{\alpha\eta}}\|\mathbf{X}^\top\mathbf{X}\|\mathrm{PolyLog}(n) \\ &\leq \frac{1}{\sqrt{p}}, \end{aligned} \tag{50}$$

with probability larger than $1 - \mathrm{e}^{-p} - \tilde{q}_n$. To obtain the two penultimate two inqualities we have used Lemma 4 and (3). The last inequality uses $\alpha = \omega\left(\frac{p\mathrm{PolyLog}(n)}{\eta} \wedge \frac{\mathrm{PolyLog}(n)}{\kappa_1}\right)$, where we use the fact Lemma 19 to conclude that $\|\mathbf{X}^\top\mathbf{X}\| \leq (\sqrt{\gamma_0}+3)^2 C_X$ with probability at least $1 - \mathrm{e}^{-p}$.

Plugging in these bounds into (46) we obtain

$$\begin{aligned}&|\mathbf{x}_{i,\mathcal{S}}^\top \widehat{\Delta}_{/i} - \mathbf{x}_i^\top \widehat{\Delta}^{\alpha}_{/i}| \\ &\leq |\dot{\ell}(\widehat{\beta})|\left(\frac{\|\mathbf{x}_i\|^2}{p(2\lambda\eta)^2} + \frac{\|\mathbf{x}_i\|^2}{(2p\eta\lambda)^2(\lambda\eta)}\right) \\ &\leq \frac{\mathrm{PolyLog}(n)}{\sqrt{p}(\lambda\eta)(1\vee\lambda\eta)}\left(1 + \frac{1}{p\lambda\eta}\right) + \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \sqrt{\frac{C\log p}{n\lambda\eta}}\end{aligned}$$

where we use $\alpha = \omega\left(\frac{p\mathrm{PolyLog}(n)}{\eta} \wedge \frac{\mathrm{PolyLog}(n)}{\kappa_1}\right)$.

Returning to (43) we thus write

$$|\mathrm{ALO} - \mathrm{ALO}^{\alpha}| \leq \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{d_n\mathrm{PolyLog}(n)}{n\lambda^2\eta^2} + \sqrt{\frac{\mathrm{PolyLog}(n)}{n\lambda\eta}},$$

by the assumption $\dot{\phi}(y_i, z_i) = O_p(\mathrm{PolyLog}(n))$. To conclude the proof of the theorem, note that

$$\begin{aligned}&|\mathrm{ALO} - \mathrm{LO}| \\ \leq\ &|\mathrm{ALO} - \mathrm{ALO}^{\alpha}| + |\mathrm{ALO}^{\alpha} - \mathrm{LO}| \\ \leq\ &\frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{d_n\mathrm{PolyLog}(n)}{n\lambda^2\eta^2} + \sqrt{\frac{\mathrm{PolyLog}(n)}{n\lambda\eta}}\end{aligned}$$

provided $\alpha = \omega\left(\frac{n\mathrm{PolyLog}(n)}{\kappa_0^2\kappa_1^2\lambda(1-\eta))\eta} \wedge \frac{\mathrm{PolyLog}(n)}{\kappa_1}\right)$, with probability at least $1 - (n+1)\mathrm{e}^{-p} - (n+2)p^{-d_n} - 2q_n - 2\check{q}_n - 2\tilde{q}_n - 2\bar{q}_n$. □

4 Proof of Theorem 2

One of the main components of this proof uses concentration results on the empirical distribution of $\widehat{\boldsymbol{\beta}}$, the subgradient of ℓ_1-regularizer (see below for precise definition), and the sparsity of $\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the minimizer of

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\eta\|\boldsymbol{\beta}\|_2^2 + \lambda(1-\eta)\|\boldsymbol{\beta}\|_1.$$

In order to state the results, we first introduce the following terms:

- Let $\widehat{\mu}$ denotes the empirical distribution of $\widehat{\boldsymbol{\beta}}$.
- Let Θ be a random variable with its value uniformly distributed among the elements of $\boldsymbol{\beta}^*$, and let $Z \sim N(0,1)$ and independent of Θ.

- Let $\text{soft}(x, r)$ denote the soft thresholding function

$$\text{soft}(x, r) = (|x| - r)_+ \text{sign}(x).$$

- For a couple (τ, b), define

$$\widehat{w}^f(\tau, b) = \frac{b}{b + 2\lambda\eta\tau} \text{soft}\Big(\tau Z + \Theta, \frac{\lambda(1-\eta)\tau}{b}\Big) - \Theta.$$

- Let the couple (τ_*, b_*) be the unique solution of the following equations:*

$$\tau^2 = \sigma^2 + \frac{1}{\gamma_0}\mathbb{E}[\widehat{w}^f(\beta, \tau)]^2, \tag{51}$$

$$\beta = \tau - \frac{1}{\gamma_0}\mathbb{E}Z \cdot \widehat{w}^f(\beta, \tau). \tag{52}$$

- Let μ^* denote the law of the random variable $\widehat{w}^f(\tau_*, b_*) + \Theta$.
- Let s_* be defined as

$$s_* = \mathbb{P}\left(|\Theta + \tau_* Z| \geq \frac{\lambda\tau_*}{b_*}\right).$$

Lemma 5 (stated and proved later as Theorem 8)**.** *Under Assumptions A1-A5 and B1-B5, there exist constants $C, c > 0$ such that for all $\varepsilon \in (0, 0.5]$:*

$$\mathbb{P}\left(W_2(\widehat{\mu}, \mu^*)^2 > \varepsilon\right) \leq C\varepsilon^{-2} e^{-cp\varepsilon^3 (\log \varepsilon)^{-2}},$$

where W_2 denotes the Wasserstein 2-distance.

We can now start the proof of the theorem. By the definition of $\mathcal{B}_{0,i}$ and $\mathcal{B}_{1,i}$ we have, for each fixed i that:

$$\begin{aligned}
&(\mathcal{B}_{1,i} \cup \mathcal{B}_{0,i})^c \\
&= \Big\{k : \Big(|\widehat{\beta}_k| \leq \kappa_1 \text{ or } |\widehat{\beta}_{/i,k}| \leq \kappa_1\Big) \text{ and } \\
&\quad \Big(|g(\widehat{\beta}_k)| > 1 - \kappa_0 \text{ or } |g(\widehat{\beta}_{/i,k})| > 1 - \kappa_0\Big)\Big\} \\
&\subset \Big\{k : 0 < |\widehat{\beta}_k| \leq \kappa_1\Big\} \\
&\quad \cup \Big\{k : 0 < |\widehat{\beta}_{/i,k}| \leq \kappa_1\Big\} \\
&\quad \cup \Big\{k : 1 - \kappa_0 \leq |g(\widehat{\beta}_k)| < 1\Big\} \\
&\quad \cup \Big\{k : 1 - \kappa_0 \leq |g(\widehat{\beta}_{/i,k})| < 1\Big\} \\
&\quad \cup \Big\{k : |\widehat{\beta}_k| > \kappa_1; |g(\widehat{\beta}_{/i,k})| \leq 1 - \kappa_0\Big\} \\
&\quad \cup \Big\{k : |\widehat{\beta}_{/i,k}| > \kappa_1; |g(\widehat{\beta}_k)| \leq 1 - \kappa_0\Big\} \\
&:= \mathcal{K}_1 \cup \mathcal{K}_1' \cup \mathcal{K}_2 \cup \mathcal{K}_2' \cup \mathcal{K}_3 \cup \mathcal{K}_3'.
\end{aligned} \tag{53}$$

We can now bound the sizes of each of the above sets. Since the full model and the leave-one-out models are the same in nature, we only bound the sets $(\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3)$ related to the full model, since the same proofs apply also to the leave-one-out models $(\mathcal{K}_1', \mathcal{K}_2', \mathcal{K}_3')$.

1. Bounding $|\mathcal{K}_1|$: The following lemma helps us bound the size of this set:

*The uniqueness of the solution is proved in Lemma 23.

Lemma 6. *Suppose* $\kappa_1 = o(p^{-\frac{1}{12}}(\log p)^{\frac{1}{4}})$. *Then, then there exists constants* C, C' *such that for all* $0 \le k \le n$,

$$\left|\{k : 0 < |\widehat{\beta}_{/i,k}| \le \kappa_1\}\right| \le Cp^{\frac{11}{12}}(\log p)^{\frac{1}{4}}$$

with probability at least $1 - C'p^{-7}$. *It then follows from a union bound over* i *that*

$$\max_{0\le i\le n}\left|\{k : 0 < |\widehat{\beta}_{/i,k}| \le \kappa_1\}\right| \le Cp^{\frac{11}{12}}(\log p)^{\frac{1}{4}}$$

with probability at least $1 - C'p^{-6}$.

Note that the size of the set $\left|\{k : 0 < |\widehat{\beta}_k| \le \kappa_1\}\right|$ can be calculated from the empirical distribution $\widehat{\mu}$. Also, Lemma 5 connects $\widehat{\mu}$ with μ^*. Hence, it is expected that we should be able to find a concentration result for $\left|\{k : 0 < |\widehat{\beta}_k| \le \kappa_1\}\right|$. However, there are several technical issues that need to be addressed in order to prove Lemma 6. Hence, the complete proof of this lemma will appear in Section 5.6.

Using Lemma 6, it is straightforward to confirm that

$$|\mathcal{K}_1| \le Cp^{11/12}(\log p)^{1/4}$$

with probability at least $1 - C'p^{-7}$ for some $C, C' > 0$.

2. Bounding $\mathcal{K}_2$: To find an upper bound for the size of the set $\mathcal{K}_2$, consider the following two sets:

$$\mathcal{T}_1 = \{k : \widehat{\beta}_k \ne 0\}$$

and

$$\mathcal{T}_2(\kappa_0) = \{k : |g(\widehat{\boldsymbol{\beta}})_k| \in [1 - \kappa_0, 1]\}.$$

First note that

$$\mathcal{T}_1 \subset \mathcal{T}_2 \text{ and } \mathcal{K}_2 = \mathcal{T}_2/\mathcal{T}_1.$$

Our first goal is to show that $\frac{1}{p}|\mathcal{T}_1|$ and $\frac{1}{p}|\mathcal{T}_2|$ are close to each other. The following two lemmas enable us to compare the two sets.

Lemma 7 (restated and proved later as Theorem 10)**.** *Under Assumptions A1-A5 and B1-B5, there exist constants* $C, C', c > 0$ *such that for all* $\varepsilon \in (0, 1]$,

$$\mathbb{P}\left(\frac{1}{p}\sum_{k=1}^{p}\mathbb{1}_{\{|g(\widehat{\boldsymbol{\beta}})_k| \ge 1-\varepsilon\}} \ge s_* + C\varepsilon\right) \le C'\varepsilon^{-3}e^{-cn\varepsilon^6}.$$

Lemma 8 (restated and proved later as Theorem 11)**.** *Under the same assumptions as the above theorem, there exist constants* $C, c > 0$ *such that for all* $\varepsilon \in (0, 1]$ *we have*

$$\mathbb{P}\left(\left|\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|_0 - s_*\right| \ge \varepsilon\right) \le C\varepsilon^{-6}e^{-cn\varepsilon^6}.$$

We should mention that the above two lemmas were originally proved in (Miolane and Montanari, 2021) for the LASSO problem. Theorems 10 and 11 are the extensions of the results of (Miolane and Montanari, 2021) and their proof strategies are the same.

From Lemma 8 and Lemma 7, it follows that, for some constant C, c_1, c_2,

$$\left|\frac{1}{p}|\mathcal{T}_1| - s_*\right| \le \kappa_0 \text{ and } \frac{1}{p}|\mathcal{T}_2(\kappa_0)| - s_* \le C\kappa_0$$

with probability at least $1 - \dfrac{2c_1}{\kappa_0^6}\exp(-c_2p\kappa_0^6)$. Here $s_* \in [0, 1]$ is the constant in Lemma 7. Setting $\kappa_0 = (\frac{8\log p}{c_2 p})^{1/6}$, using the above concentration, we obtain

$$|\mathcal{K}_2| \le |\mathcal{T}_1 \setminus \mathcal{T}_2| \le Cp^{5/6}(\log p)^{1/6}$$

with probability at least $1 - C'p^{-7}$.

3. To obtain an upper bound for $|\mathcal{K}_3|$, let $g(\widehat{\boldsymbol{\beta}}), g(\widehat{\boldsymbol{\beta}}_{/i})$ denote the sub-gradients of the LASSO penalty defined in (6). Note that $|g(\widehat{\boldsymbol{\beta}})_k| = 1$ and $|g(\widehat{\boldsymbol{\beta}}_{/i,k})| \leq 1 - \kappa_0$ for $k \in \mathcal{K}_3$, and hence

$$|g(\widehat{\boldsymbol{\beta}})_k - g(\widehat{\boldsymbol{\beta}_{/i}})_k| \geq \kappa_0 \quad \text{for } k \in \mathcal{K}_3.$$

Thus

$$\begin{aligned}
\sqrt{\kappa_0^2 |\mathcal{K}_3|} &\leq \sqrt{\sum_{k \in \mathcal{K}_3} |g(\widehat{\boldsymbol{\beta}})_k - g(\widehat{\boldsymbol{\beta}}_{/i})_k|^2} \\
&\leq \left\| g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i}) \right\| \\
&\leq \frac{2|\dot{\ell}_i(\widehat{\boldsymbol{\beta}})| \|\mathbf{x}_i\|}{\lambda(1-\eta)} \\
&\leq \frac{C\text{PolyLog}(p)}{\lambda(1-\eta)}
\end{aligned}$$

with probability at least $1 - q_n - \mathrm{e}^{-cp}$. The penultimate line follows from Part 4 of Lemma 3, and the last line uses Assumption A4 and Lemma 17. Therefore

$$|\mathcal{K}_3| \leq \frac{C\text{PolyLog}(p)}{\kappa_0^2 \lambda^2 (1-\eta)^2} = \frac{C\text{PolyLog}(p)}{\lambda^2(1-\eta)^2} p^{\frac{1}{3}} \leq C p^{\frac{11}{12}}$$

with probability at least $1 - q_n - \mathrm{e}^{-cp}$, provided $\lambda^2(1-\eta)^2 = \omega(p^{-\frac{7}{12}})$

Note that the same arguments hold for all $1 \leq i \leq n$. By cases 1-3 above, along with (53), we have proved that

$$\max_{1 \leq i \leq n} |(\mathcal{B}_{1,i} \cup \mathcal{B}_{0,i})^c| \leq C p^{11/12} (\log p)^{1/4} \tag{54}$$

with probability at least $1 - Cp^{-6} - q_n - \mathrm{e}^{-cp}$, for sufficiently large p.

The last step of the proof is to bound $|\mathcal{B}_{1,i}/\mathcal{B}_{1,i,+}| := |\mathcal{B}_{1,i,-}|$. The proof follows the arguments made for $\mathcal{K}_3$ above. More precisely, note that

$$|g(\widehat{\boldsymbol{\beta}})_k - g(\widehat{\boldsymbol{\beta}}_{/i})_k| = 2 \text{ for } k \in \mathcal{B}_{1,i,-}.$$

Thus

$$\begin{aligned}
&\max_{1 \leq i \leq n} \sqrt{4|\mathcal{B}_{1,i,-}|} \\
&= \max_{1 \leq i \leq n} \sqrt{\sum_{k \in \mathcal{B}_{1,i,-}} |g(\widehat{\boldsymbol{\beta}})_k - g(\widehat{\boldsymbol{\beta}}_{/i})_k|^2} \\
&\leq \max_{1 \leq i \leq n} \|g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i})\| \\
&\leq \frac{2\|\mathbf{x}_i\| |\dot{\ell}_i(\widehat{\boldsymbol{\beta}})|}{\lambda(1-\eta)} \\
&\leq \frac{C\text{PolyLog}(p)}{\lambda(1-\eta)}
\end{aligned}$$

with probability at least $1 - q_n - \mathrm{e}^{-cp}$, for sufficiently large p. The last inequality again follows from Part 4 of Lemma 3. Hence with probability at least $1 - q_n - \mathrm{e}^{-cp}$ we have

$$\max_{1 \leq i \leq n} |\mathcal{B}_{1,i,-}| \leq \frac{C\text{PolyLog}(p)}{\lambda^2(1-\eta)^2} \leq C p^{\frac{7}{12}} \text{PolyLog}(p)$$

provided $\lambda^2(1-\eta)^2 = \omega(p^{-\frac{7}{12}})$. Since $\mathcal{B}_{1,i,+} = \mathcal{B}_{1,i} \setminus \mathcal{B}_{1,i,-}$, we therefore have from (54) that

$$\max_{1 \leq i \leq n} |(\mathcal{B}_{1,i,+} \cup \mathcal{B}_{0,i})^c| \leq C p^{11/12} (\log p)^{1/4}$$

with probability at least $1 - Cp^{-6} - q_n - \mathrm{e}^{-cp}$, for sufficiently large p. This finishes the proof. □

5 PROOF OF LEMMAS AND OTHER THEOREMS

5.1 Preliminaries

5.1.1 Basic Linear Algebra Results

Lemma 9 (Weyl's Theorem)**.** *[Theorem 4.3.1 in (Horn and Johnson, 1994)] Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *be symmetric, and let the eigenvalues of* $\mathbf{A}, \mathbf{B}$ *and* $\mathbf{A}+\mathbf{B}$ *be* $\{\lambda_i(\mathbf{A})\}_{i=1}^n$, $\{\lambda_i(\mathbf{B})\}_{i=1}^n$ *and* $\{\lambda_i(\mathbf{A}+\mathbf{B})\}_{i=1}^n$, *in increasing order. Then*

$$|\lambda_i(\mathbf{A}+\mathbf{B}) - \lambda_i(\mathbf{A})| \leq \lambda_1(\mathbf{B})$$

for $i = 1, \ldots, n$.

Lemma 10 (Woodbury Inversion Formula)**.** *Suppose* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is nonsingular, and* $\mathbf{M} = \mathbf{A} + \mathbf{UBV}$, *then*

$$\mathbf{M}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

provided that all relevant inverse matrices exist.

Lemma 11. *Let* $\mathbf{M} \in \mathbb{R}^{n\times n}$ *be non-singular, and partitioned as a 2-by-2 block matrix*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{n_1\times n_1}$, $\mathbf{C} \in \mathbb{R}^{n_2\times n_2}$ *with* $n_1 + n_2 = n$. *Then*

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BDB}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{BD} \\ -\mathbf{DB}^\top\mathbf{A}^{-1} & \mathbf{D} \end{pmatrix}$$

where $\mathbf{D} = (\mathbf{C} - \mathbf{B}^\top\mathbf{AB})^{-1}$, *provided that all relevant inverse matrices exist.*

Lemma 12. *Suppose that* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *is an invertible matrix. Furthermore, assume that* $\mathbf{C} \in \mathbb{R}^{n\times n}$ *is a diagonal matrix. Finally,* $\mathbf{B} \in \mathbb{R}^{p\times n}$. *If* $\mathbf{A} + \mathbf{BCB}^\top$ *is invertible, then its inverse is:*

$$\begin{aligned}(\mathbf{A} + \mathbf{BCB}^\top)^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}^\top\mathbf{CBA}^{-1} \\ &+ \mathbf{A}^{-1}\mathbf{BCB}^\top(\mathbf{A} + \mathbf{BCB}^\top)^{-1}\mathbf{BCB}^\top\mathbf{A}^{-1}\end{aligned}$$

Proof. First assume $\mathbf{C}$ is invertible. Applying Woodbury formula twice yields

$$\begin{aligned}&(\mathbf{A} + \mathbf{BCB}^\top)^{-1} \\ =&\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{A}^{-1} \\ =&\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BCB}^\top\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BCB}^\top(\mathbf{A} + \mathbf{BCB}^\top)^{-1}\mathbf{BCB}^\top\mathbf{A}^{-1}\end{aligned}$$

If $\mathbf{C}$ is not invertible, WLOG assume it has non-zero diagonal elements, i.e. by rearranging its rows and columns there is a diagonal matrix $\mathbf{C}_1 \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{C} = \text{diag}(\mathbf{C}_1, 0, .., 0)$$

Split $\mathbf{B}$ in the same way:

$$\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$$

where $\mathbf{B}_1 \in \mathbb{R}^{p\times k}$ and $\mathbf{B}_2 \in \mathbb{R}^{p\times(n-k)}$, and we have

$$\mathbf{A} + \mathbf{BCB}^\top = \mathbf{A} + \mathbf{B}_1\mathbf{C}_1\mathbf{B}_1^\top$$

so we can still use the above formula by replacing $\mathbf{B}, \mathbf{C}$ by $\mathbf{B}_1, \mathbf{C}_1$ respectively. □

Lemma 13. *Suppose that* $\mathbf{A}, \mathbf{\Gamma} \in \mathbb{R}^{n\times n}$ *and that both* $\mathbf{A}+\mathbf{\Gamma}$ *and* $\mathbf{\Gamma}$ *are invertible. Then for any* $\mathbf{v} \in \mathbb{R}^n$ *we have*

$$|\mathbf{v}^\top(\mathbf{A}+\mathbf{\Gamma})^{-1}\mathbf{v} - \mathbf{v}^\top(\mathbf{A})^{-1}\mathbf{v}| \leq \frac{\lambda_{\max}(\Gamma)\mathbf{v}^\top\mathbf{v}}{\lambda^2_{\min}(A)} + \frac{\lambda^2_{\max}(\Gamma)\mathbf{v}^\top\mathbf{v}}{\lambda^2_{\min}(A)(\lambda_{\min}(A) - \lambda_{\max}(\Gamma))}$$

Proof. Using Lemma 11 we obtain

$$\begin{aligned}
&|\mathbf{v}^\top(\mathbf{A}+\mathbf{\Gamma})^{-1}\mathbf{v} - \mathbf{v}^\top(\mathbf{A})^{-1}\mathbf{v}| \\
&\leq |\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{\Gamma}\mathbf{A}^{-1}\mathbf{v}| + |\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{\Gamma}(\mathbf{A}+\mathbf{\Gamma})^{-1}\mathbf{\Gamma}\mathbf{A}^{-1}\mathbf{v}| \\
&\leq \frac{\lambda_{\max}(\Gamma)\mathbf{v}^\top\mathbf{v}}{\lambda^2_{\min}(A)} + \frac{\lambda^2_{\max}(\Gamma)\mathbf{v}^\top\mathbf{v}}{\lambda^2_{\min}(A)(\lambda_{\min}(A)-\lambda_{\max}(\Gamma))}.
\end{aligned}$$

In the last inequality, we have used Weyl's theorem for bounding the maximum eigenvalue of $(\mathbf{A}+\mathbf{\Gamma})^{-1}$. □

5.1.2 Basic Probability and Statistics

Lemma 14 (Stirling's approximation). *For $1 < s \leq p \in \mathbb{Z}$ and $p > 2$, we have*

$$\binom{p}{s} \leq e^{s\log\frac{ep}{s}}$$

Proof. By (Robbins, 1955),

$$n! = \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{r_n}$$

with

$$\frac{1}{12n+1} \leq r_n \leq \frac{1}{12n}$$

So

$$\begin{aligned}
\binom{p}{s} &= \frac{\sqrt{2\pi}p^{p+\frac{1}{2}}e^{-p}e^{r_p}}{\sqrt{2\pi}s^{s+\frac{1}{2}}e^{-s}e^{r_s}\sqrt{2\pi}(p-s)^{(p-s)+\frac{1}{2}}e^{-(p-s)}e^{r_(p-s)}} \\
&= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{p}{s(p-s)}}\left(\frac{p}{s}\right)^s\left(\frac{p}{p-s}\right)^{p-s}e^{r_n-r_s-r_{n-s}} \qquad (55)
\end{aligned}$$

The last term satisfies

$$e^{r_n-r_s-r_{n-s}} = e^{\frac{1}{12n}-\frac{1}{12s+1}-\frac{1}{12(n-s)+1}} \leq 1 \qquad (56)$$

Furthermore, if we asuume that $1 < s < p$, and $p > 2$ then

$$\frac{p}{p(p-s)} < 2.$$

Hence,

$$\frac{1}{\sqrt{2\pi}}\sqrt{\frac{p}{p-s}} \leq \sqrt{\frac{1}{\pi}} \leq 1. \qquad (57)$$

Finally,

$$\left(\frac{p-s}{s}\right)^{p-s} = \mathrm{e}^{(p-s)\log\frac{p}{p-s}} = \mathrm{e}^{(p-s)\log\left(1+\frac{s}{p-s}\right)} \leq \mathrm{e}^s. \qquad (58)$$

Combining (55), (56), (57), and (58) we conclude the result:

$$\binom{p}{s} \leq \mathrm{e}^{s\log\frac{ep}{s}}.$$

□

Lemma 15. *(Theorem 1.1 in (Rudelson and Vershynin, 2013)) Let $\mathbf{x} \in \mathbb{R}^p$ be a random vector with independent sub-Gaussian entries that satisfy $\mathbb{E}x_i = 0$ and $\|x_i\|_{\psi_2} \leq K$. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$. Then for every $t > 0$,*

$$\mathbb{P}(|\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^\top\mathbf{A}\mathbf{x}| > t) \leq 2e^{-c\min\{\frac{t}{K^2\|\mathbf{A}\|}, \frac{t^2}{K^4\|\mathbf{A}\|^2_{HS}}\}},$$

where $\|\mathbf{A}\|$, and $\|\mathbf{A}\|_{HS}$ denote, respectively, the spectral norm and the Hilbert-Schmidt norm of matrix $\mathbf{A}$.

Lemma 16 (Lemma 6 of (Jalali and Maleki, 2016)). *Let* $\mathbf{x} \sim N(0, \mathbf{I}_p)$, *then*

$$\mathbb{P}(\mathbf{x}^\top \mathbf{x} \geq p + pt) \leq \mathrm{e}^{-\frac{p}{2}(t - \log(1+t))}$$

Lemma 17. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N(0, \mathbf{\Sigma}) \in \mathbb{R}^{p \times p}$ *and suppose* $\rho_{\max}(\mathbf{\Sigma}) \leq p^{-1} C_X$ *for some constant* $C_X > 0$, *then*

$$\mathbb{P}(\max_{1 \leq i \leq n} \|\mathbf{x}_i\| \geq 2\sqrt{C_X}) \leq n e^{-p/2}$$

Proof. Let $\mathbf{z} = \Sigma^{-\frac{1}{2}} \mathbf{x}$, then $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_p)$ and

$$\begin{aligned} \mathbb{P}(\|\mathbf{x}\| \geq 2\sqrt{C_X}) &= \mathbb{P}(\mathbf{z}^\top \mathbf{\Sigma} \mathbf{z} \geq 4C_X) \\ &\leq \mathbb{P}(\mathbf{z}^\top \mathbf{z} \geq 4p) \\ &\leq e^{-\frac{p}{2}(3 - \log(4))} \leq e^{-p/2} \end{aligned}$$

The last line uses Lemma 16. A union bound over all $1 \leq i \leq n$ finishes the proof. □

Lemma 18 (Lemma 12 in (Rahnama Rad and Maleki, 2020)). $\mathbf{X} \in \mathbb{R}^{p \times p}$ *is composed of independently distributed* $N(0, \mathbf{\Sigma})$ *rows, with* $\rho_{\max} = \sigma_{\max}(\Sigma)$, *where* $\Sigma \in \mathbb{R}^{p \times p}$. *Then*

$$\mathbb{P}(\|\mathbf{X}^\top \mathbf{X}\| \geq (\sqrt{n} + 3\sqrt{p})^2 \rho_{\max}) \leq \mathrm{e}^{-p}.$$

Lemma 19. *If* $\rho_{\max} \leq p^{-1} C_X$ *and* $n/p = \gamma_0$, *then*

$$\mathbb{P}(\|\mathbf{X}^\top \mathbf{X}\| \geq (\sqrt{\gamma_0} + 3)^2 C_X) \leq \mathrm{e}^{-p}.$$

This is a straightforward application of Lemma 18.

Lemma 20 (Lemma 4.10 of (Chatterjee, 2014)). *Let* $V_1, V_2, \ldots, V_p$ *denote dependent zero mean Gaussian random variables with mean zero and variances* $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. *We then have*

$$\mathbb{E}(\max_i V_i) \leq \sqrt{2 \log 2p} \left(\max_i \sigma_i \right).$$

Lemma 21 (Borell-TIS inequality). *Let* $V_1, V_2, \ldots, V_p$ *denote dependent zero mean Gaussian random variables with mean zero and variances* $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. *Then,*

$$\mathbb{P}(|\max_i V_i - \mathbb{E}(\max_i V_i)| > t) \leq 2\mathrm{e}^{-\frac{t^2}{2\sigma^2}},$$

where $\sigma = \max(\sigma_1, \sigma_2, \ldots, \sigma_p)$.

Proof. It is a special case of the original Borell-TIS inequality, see Theorem 2.1.1 of (Adler and Taylor, 2007). □

Corollary 5. *Let* $\mathbf{v} \in \mathbb{R}^p$ *be an* $N(0, \mathbf{\Sigma})$ *random vector. Let* $\rho_{\max}$ *denotet the maximum eigenvalue of* $\mathbf{\Sigma}$. *We then have*

$$\mathbb{P}(\|\mathbf{v}\|_\infty > \sqrt{2\rho_{\max} \log 2p} + t) \leq 2\mathrm{e}^{-\frac{t^2}{2\rho_{\max}}}.$$

Proof. This corollary is a direct application of the previous two lemmas and using the fact that $\max_i \sigma_i^2 < \rho_{\max}$. □

5.1.3 Accuracy of ALO for Smooth Loss Functions and Regularizers

Theorem 6 (Theorem 3 in (Rahnama Rad and Maleki, 2020)). *Under assumptions A1-A5 and B1-B5, with probability at least* $1 - 4n\mathrm{e}^{-p} - \frac{8n}{p^3} - \frac{8n}{(n-1)^3} - q_n$ *the following bound is valid:*

$$\max_{1 \leq i \leq n} \left| \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i} - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} - \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})}{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})} \right) \left(\frac{H_{ii}}{1 - H_{ii}} \right) \right| \leq \frac{C_0 \mathrm{PolyLog}(n)}{\sqrt{p}}$$

for some constant $C_0 > 0$.

5.2 Proof of Lemma 3

The elastic net objective function is

$$h(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda(1-\eta) \sum_{i=1}^{p} |\beta_i| + \lambda\eta \boldsymbol{\beta}^\top \boldsymbol{\beta}$$

and its surrogate smoothed objective function is

$$h_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda(1-\eta) \sum_{i=1}^{p} r_\alpha^{(1)}(\beta_i) + \lambda\eta \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

where $r_\alpha^{(1)}$ is defined in (18).

1. According to Lemma 1,

$$\sup_{\boldsymbol{\beta}} |h(\boldsymbol{\beta}) - h_\alpha(\boldsymbol{\beta})| \leq \frac{2\lambda(1-\eta)p(\log 2)}{\alpha}$$

Hence,

$$\begin{aligned} 0 &\leq h_\alpha(\widehat{\boldsymbol{\beta}}) - h_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) \\ &= h_\alpha(\widehat{\boldsymbol{\beta}}) - h(\widehat{\boldsymbol{\beta}}^\alpha) + h(\widehat{\boldsymbol{\beta}}^\alpha) - h_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) \\ &\leq h_\alpha(\widehat{\boldsymbol{\beta}}) - h(\widehat{\boldsymbol{\beta}}) + \frac{2\lambda(1-\eta)p(\log 2)}{\alpha} \\ &\leq \frac{4p\lambda(1-\eta)(\log 2)}{\alpha}. \end{aligned} \tag{59}$$

In the first and second inequalities above, we have used the facts that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^\alpha$ are the optimizers of $h(\boldsymbol{\beta})$ and $h_\alpha(\boldsymbol{\beta})$ respectively. By the Taylor series expansion at $\mathbf{z} = \widehat{\boldsymbol{\beta}}_\alpha$, we obtain

$$\begin{aligned} & h_\alpha(\widehat{\boldsymbol{\beta}}) - h_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) \\ &= \nabla h_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha) + \frac{1}{2}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha)^\top \nabla^2 h_\alpha(\boldsymbol{\xi})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha) \\ &= (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha)^\top \nabla^2 h_\alpha(\boldsymbol{\xi})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha)/2 \\ &\geq \lambda\eta \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha\|^2. \end{aligned} \tag{60}$$

Here $\boldsymbol{\xi} = t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}$ for some $t \in [0,1]$. Note that to obtain the second equality, we have used the fact that $\nabla h_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) = 0$ due to the optimality of $\widehat{\boldsymbol{\beta}}^\alpha$. The last line of (60) is due to the existence of the ridge penalty term $\lambda\eta\|\boldsymbol{\beta}\|^2$ in h_α. Comparing (60) and (59), one has that

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha\| \leq \sqrt{\frac{4p(1-\eta)(\log 2)}{\alpha\eta}} \leq \sqrt{\frac{4p(\log 2)}{\alpha\eta}} \tag{61}$$

Using a similar approach we can also prove that

$$\|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}_{/i}^\alpha\| \leq \sqrt{\frac{4p(\log 2)}{\alpha\eta}}.$$

This finishes the proof of Part 1.

2. Consider the first-order optimality equations of $\widehat{\boldsymbol{\beta}}^\alpha$ and $\widehat{\boldsymbol{\beta}}_{/i}^\alpha$:

$$\begin{aligned} \sum_j \mathbf{x}_j \dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha) + \lambda(1-\eta)\dot{r}_\alpha^{(1)}(\widehat{\boldsymbol{\beta}}^\alpha) + \lambda\eta\widehat{\boldsymbol{\beta}}^\alpha = 0 \\ \sum_{j\neq i} \mathbf{x}_j \dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) + \lambda(1-\eta)\dot{r}_\alpha^{(1)}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha) + \lambda\eta\widehat{\boldsymbol{\beta}}_{/i}^\alpha = 0. \end{aligned}$$

By subtracting one from the other we have

$$0 = \sum_{j\neq i} \mathbf{x}_j[\dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha) - \dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha_{/i})] + \mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha) \ + \lambda(1-\eta)[\dot{r}^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \dot{r}^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})] + 2\lambda\eta\left(\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\right).$$

It is straightforward to simplify this expression by using the mean value theorem for $\dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha) - \dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha_{/i})$ and $\dot{r}^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \dot{r}^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})$:

$$\left[\mathbf{X}_{/i}^\top \operatorname{diag}(\underline{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{/i} + \lambda(1-\eta)\operatorname{diag}\underline{\ddot{r}}^{(1)}_{\alpha/i}) + 2\lambda\eta\mathbb{I}_p\right]\left(\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\right) = -\mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha). \tag{62}$$

In this equation, $\mathbf{X}_{/i}$ is identical to $\mathbf{X}$, except for the removal of the i^{th} row. As expected from the mean value theorem, in each diagonal element of the two terms $\operatorname{diag}(\underline{\ddot{\ell}}^{\alpha/i})$ and $\operatorname{diag}(\underline{\ddot{r}}^{(1)}_{\alpha/i})$ the second derivative of ℓ and r are calculated at a point $\xi = t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ for some $t \in [0,1]$. The choice of t can be different for different diagonal elements and is dictated by the mean value theorem. Defining $\operatorname{diag}(\underline{\ddot{r}}^\alpha) := (1-\eta)\operatorname{diag}(\underline{\ddot{r}}^{(1)}_{\alpha/i}) + 2\eta\mathbb{I}_p$, it is straightforward to use (62) and obtain

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\| \\
&\leq \left\|\left[\mathbf{X}_{/i}^\top \operatorname{diag}(\underline{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{/i} + \lambda\operatorname{diag}(\underline{\ddot{r}}^\alpha)\right]^{-1}\right\| \|\mathbf{x}_i\| |\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)| \\
&\leq \frac{|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^\alpha)|\|\mathbf{x}_i\|}{2\lambda\eta},
\end{aligned}$$

where the last inequality is because of the existence of $2\eta\mathbb{I}_p$ in $\operatorname{diag}(\underline{\ddot{r}}^\alpha)$.

3. According to Part 1, $\lim_{\alpha\to\infty}\widehat{\boldsymbol{\beta}}^\alpha_{/i} = \widehat{\boldsymbol{\beta}}_{/i}$, $\forall 0 \leq i \leq n$. The result then follows by letting $\alpha \to \infty$ in Part 2, and using the fact that $\dot{\ell}$ is continuous.

4. By Part 3,

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\| \leq \frac{|\dot{\ell}(\widehat{\boldsymbol{\beta}})|\|\mathbf{x}_i\|}{2\lambda\eta} \tag{63}$$

Note that the first order optimaility equations of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{/i}$ are:

$$\begin{aligned}
&\sum_j \mathbf{x}_j\dot{\ell}_j(\widehat{\boldsymbol{\beta}}) + \lambda(1-\eta)g(\widehat{\boldsymbol{\beta}}) + 2\lambda\eta\widehat{\boldsymbol{\beta}} = 0 \\
&\sum_{j\neq i} \mathbf{x}_j\dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i}) + \lambda(1-\eta)g(\widehat{\boldsymbol{\beta}}_{/i}) + 2\lambda\eta\widehat{\boldsymbol{\beta}}_{/i} = 0.
\end{aligned}$$

By subtracting one from another and applying the mean value theorem we have

$$\begin{aligned}
&\lambda(1-\eta)[g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i})] \\
=& -\sum_{j\neq i}\mathbf{x}_j\left(\dot{\ell}_j(\widehat{\boldsymbol{\beta}}) - \dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i})\right) - \mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}) - 2\lambda\eta(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}) \\
=& -\left[\mathbf{X}_{/i}^\top \operatorname{diag}[\underline{\ddot{\ell}}]\mathbf{X}_{/i} + 2\lambda\eta\mathbb{I}_p\right](\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}) - \mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}).
\end{aligned}$$

Here, we have defined $\underline{\ddot{\ell}}$ as it was in the proof of Part 2. Under the event that

$$\begin{aligned}
\{\sup_i \sup_{j\neq i} \sup_{t\in[0,1]} \ddot{\ell}_i(t\widehat{\boldsymbol{\beta}} + (1-t)\widehat{\boldsymbol{\beta}}_{/i}) &\leq \operatorname{PolyLog}(n), \\
\|\mathbf{X}^\top\mathbf{X}\| &\leq (\sqrt{\gamma_0}+3)^2 C_X, \\
\sup_i \|\mathbf{x}_i\| &\leq 2\sqrt{C_X}, \\
\sup_i |\dot{\ell}_i(\widehat{\boldsymbol{\beta}})| &\leq \operatorname{PolyLog}(n)\}
\end{aligned}$$

with probability at least $1 - q_n - \mathrm{e}^{-p} - \check{q}_n - ne^{-p/2}$ according to Assumption A4, Lemma 17 and Lemma 19, we have

$$\begin{aligned}
&\lambda(1-\eta)\|g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i})\| \\
\leq\ & (\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\| + 2\lambda\eta)\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\| + \|\mathbf{x}_i\||\dot{\ell}_i(\widehat{\boldsymbol{\beta}})| \\
\leq\ & \left(\frac{(\sqrt{\gamma_0}+3)^2 C_X \mathrm{PolyLog}(n)}{2\lambda\eta} + 2\right)\|\mathbf{x}_i\||\dot{\ell}_i(\widehat{\boldsymbol{\beta}})| \\
\leq\ & \frac{\mathrm{PolyLog}(n)}{\lambda\eta}
\end{aligned}$$

where in the second inequality we also used (63). Combining the above results we have

$$\max_{1\leq i\leq n}\|g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\beta}}_{/i})\| \leq \frac{\mathrm{PolyLog}(n)}{\lambda^2\eta(1-\eta)}$$

with probability at least $1 - q_n - \mathrm{e}^{-p} - \check{q}_n - ne^{-p/2}$.

5. ($k \in \mathcal{S}^{(1)}$) We only provide a proof for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^\alpha$, since the arguments are exactly the same for the leave-one-out estimators $\widehat{\boldsymbol{\beta}}_{/i}$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$.

   For $k \in \mathcal{S}^{(1)}$, we have $|\widehat{\beta}_k| > \kappa_1$ so

$$|\widehat{\beta}^\alpha_k| \geq |\widehat{\beta}_k| - \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha\| \geq \kappa_1 - \sqrt{\frac{4p(1-\eta)\log 2}{\alpha\eta}} \geq \frac{\kappa_1}{2}$$

   provided that $\alpha\eta\kappa_1^2 \geq 16(1-\eta)(\log 2)p$. The second inequality uses (61) to bound $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha\|$.

6. ($k \in \mathcal{S}^{(0)}$) From the first order optimality conditions on $\widehat{\boldsymbol{\beta}}_{/i}$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$, we have

$$\begin{aligned}
&\sum_{j\neq i}\mathbf{x}_j\dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i}) + \lambda(1-\eta)g(\widehat{\boldsymbol{\beta}}_{/i}) + 2\lambda\eta\widehat{\boldsymbol{\beta}}_{/i} = 0 \\
&\sum_{j\neq i}\mathbf{x}_j\dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) + \lambda(1-\eta)\nabla r^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) + 2\lambda\eta\widehat{\boldsymbol{\beta}}^\alpha_{/i} = 0.
\end{aligned}$$

   By subtracting the two equalities we obtain

$$\begin{aligned}
\nabla r^{(1)}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) - g(\widehat{\boldsymbol{\beta}}_{/i}) =& \frac{1}{\lambda(1-\eta)}\left(\sum_{j\neq i}\mathbf{x}_j(\dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i}) - \dot{\ell}_j(\widehat{\boldsymbol{\beta}}^\alpha_{/i})) + 2\lambda\eta(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}^\alpha_{/i})\right) \\
=& \frac{1}{\lambda(1-\eta)}\left(\mathbf{X}^\top_{/i}\,\mathrm{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_i)]_{j\neq i}\mathbf{X}_{/i} + 2\lambda\eta\mathbb{I}_p\right)(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}^\alpha_{/i}). \qquad (64)
\end{aligned}$$

   The last line follows from the mean value theorem applied to $\dot{\ell}(\cdot)$, and $\boldsymbol{\xi}_i^\alpha = t\widehat{\boldsymbol{\beta}}_{/i} + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ for some $t \in [0,1]$, where t can be different for different i, j. By Parts 1-3, we have $\forall i$, $\boldsymbol{\xi}_i$ lies in set $\mathcal{D}$ in Assumption A4 for large enough p. To see this, let $\boldsymbol{\xi}_i := t\widehat{\boldsymbol{\beta}}_{/i} + (1-t)\widehat{\boldsymbol{\beta}}_{/i}$. By definition of $\mathcal{D}$, $\boldsymbol{\xi}_i \in \mathcal{D}$. Now since

$$\|\boldsymbol{\xi}_i^\alpha - \boldsymbol{\xi}_i\|_2 = (1-t)\|\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \widehat{\boldsymbol{\beta}}_{/i}\|_2 \leq \sqrt{\frac{4p(\log 2)}{\alpha\eta}},$$

   the difference can be arbitrarily small for large p, when we assume $\alpha\eta = \omega(p)$. So there exists p_0 such that $\forall p \geq p_0$, $\|\boldsymbol{\xi}_i^\alpha - \boldsymbol{\xi}_i\|_2 \overset{\star}{\leq} \epsilon$, i.e. $\boldsymbol{\xi}_i^\alpha \in \mathcal{D}$.

So $\max_{0\le i\le n, j\neq i} \ddot{\ell}_j(\boldsymbol{\xi}_i) \le \text{PolyLog}(n)$ with probabilty at least $1-q_n$.[†] Then we have

$$\begin{aligned}
&\max_{0\le i\le n} \|\nabla r_\alpha^{(1)}(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) - g(\widehat{\boldsymbol{\beta}}_{/i})\| \\
\le& \frac{1}{\lambda(1-\eta)}\Big(\max_{0\le i\le n, j\neq i} |\ddot{\ell}_j(\boldsymbol{\xi}_i)|\|\mathbf{X}^\top\mathbf{X}\| + 2\lambda\eta\Big)\|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\| \\
\le& \frac{1}{\lambda(1-\eta)}(\text{PolyLog}(n)(\sqrt{\gamma_0}+3)^2 C_X + 2\lambda\eta)\sqrt{\frac{4p(1-\eta)(\log 2)}{\alpha\eta}} \\
\le& \frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{p(1-\eta)}{\alpha\eta}}
\end{aligned} \tag{65}$$

with probability at least $1-q_n-e^{-p}$. The second inequality uses Lemma 19 to bound $\|\mathbf{X}^\top\mathbf{X}\|$ and Part 1 to bound $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\alpha\|$. The last line uses boundedness of λ and η (Assumption A5) to absorb $2\lambda\eta$ into the constant C.

Without loss of generality we assume that $0 < g(\widehat{\boldsymbol{\beta}})_k \le 1-\kappa_0$ (negative subgradients can be handled similarly). We first obtain

$$\nabla r_\alpha^{(1)}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})_k - 1 = \frac{e^{\alpha\widehat{\beta}^\alpha_{/i,k}} - e^{-\alpha\widehat{\beta}^\alpha_{/i,k}}}{e^{\alpha\widehat{\beta}^\alpha_{/i,k}} + e^{-\alpha\widehat{\beta}^\alpha_{/i,k}} + 2} - 1 = -\frac{2}{1+e^{\alpha\widehat{\beta}^\alpha_{/i,k}}}.$$

It follows from (65) that, with probability at least $1-q_n-\mathrm{e}^{-p}$, $\forall i\ge 0$, $k\in[p]$:

$$-\frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{p(1-\eta)}{\alpha\eta}} \le 1 - g(\widehat{\boldsymbol{\beta}}_{/i})_k - \frac{2}{1+e^{\alpha\widehat{\beta}^\alpha_{/i,k}}} \le \frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{p(1-\eta)}{\alpha\eta}} \tag{66}$$

By rearranging the terms in the second inequality of (66) and using the fact that $1-g(\widehat{\boldsymbol{\beta}}_{/i})_k \ge \kappa_0$, we obtain

$$\kappa_0 - \frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{(1-\eta)}{\alpha\eta}} \le \frac{2}{1+e^{\alpha\widehat{\beta}^\alpha_{/i,k}}} \le 2e^{-\alpha\widehat{\beta}^\alpha_{/i,k}}$$

Therefore

$$\widehat{\beta}^\alpha_{/i,k} \le \frac{1}{\alpha}\left(\log 2 - \log\left[\kappa_0 - \frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{p(1-\eta)}{\alpha\eta}}\right]\right)$$

By our assumption that $\alpha = \omega\left(\frac{n\text{PolyLog}(n)}{\kappa_0^2\lambda^2(1-\eta)\eta}\right)$, we have

$$\kappa_0 - \frac{\text{PolyLog}(n)}{\lambda(1-\eta)}\sqrt{\frac{p(1-\eta)}{\alpha\eta}} \ge \frac{1}{2}\kappa_0,$$

therefore with probability at least $1-q_n-\mathrm{e}^{-p}$, $\forall i\ge 0$, $k\in\mathcal{S}^{(0)}_{/i}$:

$$\widehat{\beta}^\alpha_{/i,k} \le \frac{1}{\alpha}\left(\log 2 - \log\left(\frac{1}{2}\kappa_0\right)\right) = \frac{1}{\alpha}\log\left(\frac{4}{\kappa_0}\right).$$

Using a symmetric argument on the case of negative subgradients, we conclude that

$$\max_{0\le i\le n}\max_{k\in\mathcal{S}^{(0)}_{/i}} |\widehat{\beta}^\alpha_{/i,k}| \le \frac{1}{\alpha}\log\left(\frac{4}{\kappa_0}\right)$$

with probability at least $1-q_n-\mathrm{e}^{-p}$. □

[†]Please note that for notational simplicity we use the same notation for all the terms that are polynomial functions of $\log(n)$, and have dropped the subscript 2 of the term $\text{PolyLog}(n)$ that appeared in Assumption A4.

5.3 Proof of Lemma 4

Recall that the penalty function is

$$r_\alpha(z) = \eta z^2 + \frac{1-\eta}{\alpha} \cdot \left(\log(1+e^{\alpha z}) + \log(1+e^{-\alpha z})\right),$$

It can be verified that

$$\ddot{r}_\alpha(z) = 2\eta + (1-\eta) \cdot \frac{2\alpha}{e^{\alpha z} + e^{-\alpha z} + 2}. \tag{67}$$

We have

$$\int_0^1 \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k} - t\Delta^\alpha_{/i,k})dt = \frac{\dot{r}_\alpha(\widehat{\beta}^\alpha_k) - \dot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k})}{\widehat{\beta}^\alpha_k - \widehat{\beta}^\alpha_{/i,k}}.$$

Note that $\dot{r}_\alpha(z) = 2\eta z + (1-\eta)\left(1 - \frac{2}{1+e^{\alpha z}}\right)$ is an increasing odd function, concave on $[0, +\infty)$ and convex on $(-\infty, 0]$. Furthermore, using the fact that $\frac{1}{4}e^{-x} \le \frac{e^x}{(1+e^x)^2} \le e^{-x}$ for $x \ge 0$, we have, for $z \ge 0$:

$$\ddot{r}_\alpha(z) \ge 2\eta + \frac{1}{2}\alpha(1-\eta)e^{-\alpha z}, \tag{68}$$

and

$$\ddot{r}_\alpha(z) \le 2\eta + 2\alpha(1-\eta)e^{-\alpha z}. \tag{69}$$

With this background, we can now state the proof of each part.

1. $(k \in \mathcal{B}_{0,i})$**:** By Part 6 of Lemma 3, for large enough p, with probability at least $1 - q_n - e^{-p}$ we have $\forall i, \forall k \in \mathcal{B}_{0,i}$:

   $$\max\left\{|\widehat{\beta}^\alpha_k|,\ |\widehat{\beta}^\alpha_{/i,k}|\right\} \le \frac{1}{\alpha}\log\left(\frac{4}{\kappa_0}\right).$$

   With the same probability we then have

   $$\begin{aligned}\int_0^1 \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k} - t\Delta^\alpha_{/i,k})dt &\overset{(a)}{\ge} \ddot{r}_\alpha(\frac{1}{\alpha}\log(\frac{4}{\kappa_0})) \\ &\overset{(b)}{\ge} 2\eta + \frac{1}{2}\alpha(1-\eta)\frac{\kappa_0}{4} \\ &= 2\eta + \frac{1}{8}\alpha(1-\eta)\kappa_0.\end{aligned}$$

   Inequality (a) is because $\ddot{r}_\alpha(z)$ is decreasing in $|z|$. Inequality (b) uses (69).

2. $(k \in \mathcal{B}_{0,i})$**:** An argument similar to the one presented for part (1) proves

   $$\ddot{r}_\alpha(\widehat{\beta}^\alpha_k) \ge 2\eta + \frac{1}{8}\alpha(1-\eta)\kappa_0.$$

   with probability at least $1 - q_n - e^{-p}$.

3. $(k \in \mathcal{B}_{1,i,+})$**:**

   $$\begin{aligned}2\eta &\le \int_0^1 \ddot{r}_\alpha(\widehat{\beta}^\alpha_{/i,k} - t\Delta^\alpha_{/i,k})dt \\ &\le \ddot{r}_\alpha\left(\frac{\kappa_1}{2}\right) \le 2\eta + 2\alpha e^{-\frac{1}{2}\alpha\kappa_1}.\end{aligned}$$

4. $(k \in \mathcal{B}_{1,i,+})$**:** Similarly we have

   $$2\eta \le \ddot{r}_\alpha(\widehat{\beta}^\alpha_k) \le 2\eta + 2\alpha e^{-\frac{1}{2}\alpha\kappa_1}.$$

5. $(k \in \mathcal{B}_{1,i,+})$**:** The proof is identical to 4) by substituting $\widehat{\boldsymbol{\beta}}^\alpha$ with $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$.

□

5.4 Proof of Theorem 3

We begin with the proof of (34). To simplify notation we will use the compact notation $\ddot{\mathbf{L}}_{/i}$ and $\ddot{\mathbf{R}}_{/i}$ to denote the diagonal matrices $\mathrm{diag}[\int_0^1 \ddot{\ell}_{/i}(\boldsymbol{\theta}(t))dt]$ and $\mathrm{diag}[\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))dt]$ respectively, where $\boldsymbol{\theta} = t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}$. We also fix an index i and write $\mathcal{B}_{1,+}$ and $\mathcal{B}_0$ to denote $\mathcal{B}_{1,i,+}$ and $\mathcal{B}_{0,i}$ respectively.

Plugging in $\boldsymbol{\theta} = t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}$, with a possible permutation of the rows and columns of $\mathbf{J}_{/i}(\boldsymbol{\theta})$, by (25), we have that

$$\begin{aligned}
&\int_0^1 \mathbf{J}_{/i}(\boldsymbol{\theta}(t))dt =: \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \\
&= \begin{pmatrix} \lambda\ddot{\mathbf{R}}_{/i,\mathcal{B}_0^c} + \mathbf{X}^\top_{/i,\mathcal{B}_0^c}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0^c} & \mathbf{X}^\top_{/i,\mathcal{B}_0^c}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0} \\ \mathbf{X}^\top_{/i,\mathcal{B}_0}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0^c} & \lambda\ddot{\mathbf{R}}_{/i,\mathcal{B}_0} + \mathbf{X}^\top_{/i,\mathcal{B}_0}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0} \end{pmatrix}.
\end{aligned}$$

By the block matrix inversion lemma, i.e. Lemma 11, we have

$$\begin{aligned}
&\left(\int_0^1 \mathbf{J}_{/i}(\boldsymbol{\theta}(t))dt\right)^{-1} \\
&= \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{D}\mathbf{B}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{D} \\ -\mathbf{D}\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{D} \end{pmatrix}
\end{aligned}$$

where $\mathbf{D} := (\mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}$. We will now estimate the norms of each of these terms separately, using Lemma 4.

Bounding $\|\mathbf{A}^{-1}\|$: Note that for $\boldsymbol{\theta}(t) = t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}$, we have

$$\begin{aligned}
\sigma_{\min}(\mathbf{A}) &= \sigma_{\min}\left(\lambda\ddot{\mathbf{R}}_{/i,\mathcal{B}_0^c} + \mathbf{X}^\top_{/i,\mathcal{B}_0^c}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0^c}\right) \\
&\overset{(a)}{\geq} \sigma_{\min}\left(\lambda\ddot{\mathbf{R}}_{/i,\mathcal{B}_0^c}\right) \geq 2\lambda\eta,
\end{aligned}$$

where for inequality (a) we have used the fact that the matrix $\mathbf{X}^\top_{/i,\mathcal{B}_0^c}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0^c}$ is positive semidefinite because of the convexity of the loss function. Hence,

$$\|\mathbf{A}^{-1}\| = \frac{1}{\sigma_{\min}(\mathbf{A})} \leq (2\lambda\eta)^{-1}. \tag{70}$$

Bounding $\|\mathbf{B}\|$: If $\mathbb{S}^{p-1}$ denotes the unit sphere in $\mathbb{R}^p$, then we can write the cross term

$$\begin{aligned}
\|\mathbf{B}\| &= \sup_{\mathbf{u}\in\mathbb{S}^{|\mathcal{B}_0^c|},\mathbf{v}\in\mathbb{S}^{|\mathcal{B}_0|}} \mathbf{u}^\top\mathbf{B}\mathbf{v} \\
&\leq \sup_{\mathbf{u}\in\mathbb{S}^{|\mathcal{B}_0^c|},\mathbf{v}\in\mathbb{S}^{|\mathcal{B}_0|}} \mathbf{u}^\top\left(\mathbf{X}^\top_{/i,\mathcal{B}_0^c}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i,\mathcal{B}_0}\right)\mathbf{v} \\
&\leq \sup_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}\in\mathbb{S}^p} \tilde{\mathbf{u}}^\top\left(\mathbf{X}^\top_{/i}\ddot{\mathbf{L}}_{/i}\mathbf{X}_{/i}\right)\tilde{\mathbf{v}} \\
&\leq \max_{j\neq i}\int_0^1 \ddot{\ell}_j(\boldsymbol{\theta}(t))dt\|\mathbf{X}^\top_{/i}\mathbf{X}_{/i}\| \\
&\leq \mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|
\end{aligned} \tag{71}$$

In the third line above we define $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ based on $\mathbf{u}$ and $\mathbf{v}$ as follows. Let $\mathcal{B}_0^c = \{j_1, \ldots, j_{|\mathcal{B}_0^c|}\}$. Then we define $\tilde{\mathbf{u}}_{j_k} = \mathbf{u}_k$ for $k \in [|\mathcal{B}_0^c|]$, and $\tilde{\mathbf{u}}_j = 0$ for all $j \in \mathcal{B}_0$. $\tilde{\mathbf{v}}$ is defined similarly. Note that $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}} \in \mathbb{S}^{p-1}$ so that the

supremum in the next line is justified. In the last line we use the second part of (3), in conjunction with part 1) of Lemma 3.[‡]

Bounding $\|\mathbf{D}\|$: By the repeated use of Weyl's theorem, i.e., Lemma 9, we have

$$\begin{aligned}
&\sigma_{\min}(\mathbf{C}-\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})\\
&\geq\ \lambda\min\left\{\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))_k dt : k\in\mathcal{B}_0\right\} - \|\mathbf{B}\|^2\|\mathbf{A}^{-1}\|\\
&\geq\ \lambda\min\left\{\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))_k dt : k\in\mathcal{B}_0\right\}\\
&\quad - \frac{(\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|)^2}{2\lambda\eta},
\end{aligned} \tag{72}$$

where in the last line we use the upper bounds on $\|\mathbf{B}\|$ and $\|\mathbf{A}^{-1}\|$ we obtained above. From Lemma 4 we have

$$\min_{k\in\mathcal{B}_0}\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))_k dt \geq 2\eta + \frac{1}{8}\alpha(1-\eta)\kappa_0, \tag{73}$$

By combining (72) and (73) we obtain

$$\sigma_{\min}(\mathbf{C}-\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}) \geq \frac{\lambda\alpha}{16}(1-\eta)\kappa_0. \tag{74}$$

provided

$$2\lambda\eta + \frac{\lambda\alpha}{16}(1-\eta)\kappa_0 \geq \frac{(\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|)^2}{2\lambda\eta}$$

Note that according to Lemma 19, $\|\mathbf{X}^\top\mathbf{X}\|$ is bounded by a constant with probability at least $1-\mathrm{e}^{-p}$. By the assumptions of this theorem, we know that α grows fast enough, and thus the above event holds with probability at least $1-\mathrm{e}^{-p}$.

Hence, under the event $\|\mathbf{X}^\top\mathbf{X}\| \leq C$, which occurs with probability at least $1-\mathrm{e}^{-p}$, we expect $\|\mathbf{D}\|$ to go to zero as $\alpha\to\infty$. Therefore,

$$\|\mathbf{D}\| = \ \|(\mathbf{C}-\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}\| = (\sigma_{\min}(\mathbf{C}-\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}))^{-1} \leq \frac{16}{\lambda\alpha(1-\eta)\kappa_0}. \tag{75}$$

By the block matrix inversion lemma,

$$\begin{aligned}
&\Big|\mathbf{x}_i^\top\left(\int_0^1 \mathbf{J}_{/i}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i})dt\right)^{-1}\mathbf{x}_i \ - \mathbf{x}_{i,\mathcal{B}_0^c}^\top(\lambda\mathrm{diag}(\ddot{\underline{r}}^{\alpha/i}_{\mathcal{B}_0^c})) + \mathbf{X}^\top_{/i,\mathcal{B}_0^c}\mathrm{diag}(\ddot{\underline{\ell}}^{\alpha/i})\mathbf{X}_{/i,\mathcal{B}_0^c})^{-1}\mathbf{x}_{i,\mathcal{B}_0^c}\Big|\\
&= \Big|\mathbf{x}_i^\top\begin{pmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\mathbf{D}\mathbf{B}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{D}\\ -\mathbf{D}\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{D}\end{pmatrix}\mathbf{x}_i - \mathbf{x}_{i,\mathcal{B}_0^c}^\top\mathbf{A}^{-1}\mathbf{x}_{i,\mathcal{B}_0^c}\Big|\\
&\leq \mathbf{x}_{i,\mathcal{B}_0^c}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{D}\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{x}_{i,\mathcal{B}_0^c} + 2|\mathbf{x}_{i,\mathcal{B}_0^c}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{D}\mathbf{x}_{i,\mathcal{B}_0}| + \mathbf{x}_{i,\mathcal{B}_0}^\top\mathbf{D}\mathbf{x}_{i,\mathcal{B}_0}\\
&\leq \|\mathbf{x}_{i,\mathcal{B}_0^c}\|^2\|\mathbf{A}^{-1}\|^2\|\mathbf{B}\|^2\|\mathbf{D}\| + 2\|\mathbf{x}_{i,\mathcal{B}_0^c}\|\|\mathbf{A}^{-1}\mathbf{B}\mathbf{D}\mathbf{x}_{i,\mathcal{B}_0}\| + \ \|\mathbf{x}_{i,\mathcal{B}_0}\|^2\|\mathbf{D}\|\\
&\leq \|\mathbf{x}_i\|^2\|\mathbf{A}^{-1}\|^2\|\mathbf{B}\|^2\|\mathbf{D}\| + 2\|\mathbf{x}_{i,\mathcal{B}_0^c}\|\|\mathbf{A}^{-1}\|\|\mathbf{B}\|\|\mathbf{D}\|\|\mathbf{x}_{i,\mathcal{B}_0}\| + \ \|\mathbf{x}_i\|^2\|\mathbf{D}\|\\
&\leq \|\mathbf{x}_i\|^2\|\mathbf{D}\|(\|\mathbf{A}^{-1}\|^2\|\mathbf{B}\|^2 + 2\|\mathbf{A}^{-1}\|\|\mathbf{B}\| + 1)\\
&= \|\mathbf{x}_i\|^2\|\mathbf{D}\|(\|\mathbf{A}^{-1}\|\|\mathbf{B}\| + 1)^2\\
&\leq \frac{16\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)\kappa_0}\left(\frac{\mathrm{PolyLog}(n)\|\mathbf{X}^\top\mathbf{X}\|}{2\lambda\eta}+1\right)^2.
\end{aligned}$$

[‡]Please note that as mentioned before, for notational simplicity, we have dropped the subscript 24 of the term $\mathrm{PolyLog}(n)$ that appeared in the second part of Assumption A4, use the same notation for all the terms that are polynomial functions of $\log(n)$.

In the last step we use the previously derived bounds on the operator norms of $\mathbf{A}^{-1}$, $\mathbf{B}$ and $\mathbf{D}$, along with the Cauchy-Schwarz inequality. As discussed before, this error will be small for large values of α.

To show (35), we define $\boldsymbol{\theta} := \widehat{\boldsymbol{\beta}}^{\alpha}_{/i} - \Delta^{\alpha}_{/i}$. The bounds then follow by the same steps with very minor modifications. □

5.5 Proof of Theorem 4

We will use the notation

$$\omega_s := \|\mathbf{X}^\top \mathbf{X}\| \sup_{t\in[0,1]} \max_{1\le i\le n} \ddot{\ell}_i(t\widehat{\boldsymbol{\beta}}^{\alpha} + (1-t)\widehat{\boldsymbol{\beta}}^{\alpha}_{/i}); \; \rho(\alpha) := \lambda\left(2\eta + \frac{\alpha(1-\eta)\kappa_0}{8}\right). \tag{76}$$

By the assumptions of the Theorem we define the following two events:

1. $\mathcal{A}_L := \{\text{Assumption A4 holds}\}$
2. $\mathcal{A}_s := \{\max_{1\le i\le n} |\mathcal{B}^c_{0,i} \setminus \mathcal{B}_{1,i,+}| \le d_n\}$

which hold with probabilities at least $1-q_n$ and $1-\tilde{q}_n$ respectively. We will prove the theorem assuming that the two events above both hold, which by the union bound, happens with probability at least $1-q_n-\tilde{q}_n$.

Our strategy will be to bound the following quadratic forms:

$$\left|\mathbf{x}^\top_{i,\mathcal{F}}(\lambda \mathrm{diag}(\underset{\sim}{\ddot{r}}^{\alpha/i}_{\mathcal{F}}) + \mathbf{X}^\top_{\mathcal{F}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}} - \mathbf{x}^\top_{i,\mathcal{B}_{1,i,+}}(\lambda \mathrm{diag}(\underset{\sim\mathcal{B}_{1,i,+}}{\ddot{r}^{\alpha/i}}) + \mathbf{X}^\top_{\mathcal{B}_{1,i,+}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{B}_{1,i,+}})^{-1}\mathbf{x}_{i,\mathcal{B}_{1,i,+}}\right| \le f(d_n), \tag{77}$$

and similarly

$$\left|\mathbf{x}^\top_{i,\mathcal{F}}(\lambda \mathrm{diag}(\underset{\sim\mathcal{F}}{\ddot{r}^{\alpha/i}}) + \mathbf{X}^\top_{\mathcal{F}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}} - \mathbf{x}^\top_{i,\mathcal{B}_{1,i,+}}\; (\lambda \mathrm{diag}(\underset{\sim\mathcal{B}_{1,i,+}}{\ddot{r}^{\alpha/i}}) + \mathbf{X}^\top_{\mathcal{B}_{1,i,+}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{B}_{1,i,+}})^{-1}\mathbf{x}_{i,\mathcal{B}_{1,i,+}}\right| \le f(d_n) \tag{78}$$

for a suitable function $f(\cdot)$ of d_n, i.e., the cardinality of set differences.

In the rest of the proof, we fix an index i and write $\mathcal{B}_{1,+}$ and $\mathcal{B}_0$ to denote $\mathcal{B}_{1,i,+}$ and $\mathcal{B}_{0,i}$ respectively. Consider the following decomposition

$$\mathbf{H}_1 := \lambda\,\mathrm{diag}(\underset{\sim}{\ddot{r}}^{\alpha/i}_{\mathcal{B}^c_0}) + \mathbf{X}^\top_{/i,\mathcal{F}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{/i,\mathcal{F}} = \begin{pmatrix}\mathbf{A}_1 & \mathbf{B}_1\\ \mathbf{B}_1^\top & \mathbf{C}_1\end{pmatrix} \tag{79}$$

that is obtained by a permutation of rows and columns such that the rows and columns of $\mathbf{A}_1$ belong to $\mathcal{B}_{1,+}$, and rows and columns of $\mathbf{C}_1$ belong to $\mathcal{F}\setminus\mathcal{B}_{1,+}$. Define $\boldsymbol{\theta}(t) = t\widehat{\boldsymbol{\beta}}^{\alpha} + (1-t)\widehat{\boldsymbol{\beta}}^{\alpha}_{/i}$. Then, we have

$$\begin{aligned}
\mathbf{A}_1 &= \mathrm{diag}\left[\lambda\left\{\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))dt\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}^\top_{/i,\mathcal{B}_{1,+}}\mathrm{diag}\left[\int_0^1 \ddot{\ell}_{/i}(\boldsymbol{\theta}(t))dt\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}}\\
\mathbf{B}_1 &= \mathbf{X}^\top_{/i,\mathcal{B}_{1,+}}\mathrm{diag}\left[\int_0^1 \ddot{\ell}_{/i}(\boldsymbol{\theta}(t))dt\right]\mathbf{X}_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}\\
\mathbf{C}_1 &= \mathrm{diag}\left[\lambda\int_0^1 \ddot{r}_\alpha(\boldsymbol{\theta}(t))_{\mathcal{F}\setminus\mathcal{B}_{1,+}})dt\right] + \mathbf{X}^\top_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}\mathrm{diag}\left[\int_0^1 \ddot{\ell}_{/i}(\boldsymbol{\theta}(t))dt\right]\mathbf{X}_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}.
\end{aligned}$$

Similarly, we define $\mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2$ in the following way:

$$\mathbf{H}_2 := \lambda\,\mathrm{diag}(\underset{\sim}{\ddot{r}}^{\alpha/i}_{\mathcal{F}}) + \mathbf{X}^\top_{\mathcal{F}}\mathrm{diag}(\underset{\sim}{\ddot{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}} = \begin{pmatrix}\mathbf{A}_2 & \mathbf{B}_2\\ \mathbf{B}_2^\top & \mathbf{C}_2\end{pmatrix} \tag{80}$$

where

$$\begin{aligned}
\mathbf{A}_2 &= \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}_{/i,\mathcal{B}_{1,+}}^\top \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}}\\
\mathbf{B}_2 &= \mathbf{X}_{/i,\mathcal{B}_{1,+}}^\top \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}\\
\mathbf{C}_2 &= \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)\right\}_{\mathcal{F}\setminus\mathcal{B}_{1,+}}\right] + \mathbf{X}_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}^\top \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{F}}.
\end{aligned}$$

We then obtain

$$\begin{aligned}
&\left|\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\operatorname{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}} - \mathbf{x}_{i,\mathcal{F}}^\top(\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\operatorname{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}}\right|\\
&= \mathbf{x}_{i,\mathcal{F}}^\top\left[\begin{pmatrix}\mathbf{A}_1 & \mathbf{B}_1\\ \mathbf{B}_1^\top & \mathbf{C}_1\end{pmatrix}^{-1} - \begin{pmatrix}\mathbf{A}_2 & \mathbf{B}_2\\ \mathbf{B}_2^\top & \mathbf{C}_2\end{pmatrix}^{-1}\right]\mathbf{x}_{i,\mathcal{F}}.
\end{aligned}$$

By matrix inversion of block diagonal matrices, Lemma 11, we have

$$\mathbf{H}_k^{-1} = \begin{pmatrix}\mathbf{A}_k^{-1} + \mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{D}\mathbf{B}_k^\top\mathbf{A}_k^{-1} & -\mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{D}_k\\ -\mathbf{D}_k\mathbf{B}_k^\top\mathbf{A}_k^{-1} & \mathbf{D}_k\end{pmatrix} \tag{81}$$

where for $k = 1, 2$ we define

$$\mathbf{D}_k := (\mathbf{C}_k - \mathbf{B}_k^\top\mathbf{A}_k^{-1}\mathbf{B}_k)^{-1}.$$

From (81) we have

$$\begin{aligned}
&\left|\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\operatorname{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}} - \mathbf{x}_{i,\mathcal{F}}^\top(\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i}) + \mathbf{X}_{\mathcal{F}}^\top\operatorname{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}}\right|\\
&= \mathbf{x}_{i,\mathcal{F}}^\top\left[\begin{pmatrix}\mathbf{A}_1 & \mathbf{B}_1\\ \mathbf{B}_1^\top & \mathbf{C}_1\end{pmatrix}^{-1} - \begin{pmatrix}\mathbf{A}_2 & \mathbf{B}_2\\ \mathbf{B}_2^\top & \mathbf{C}_2\end{pmatrix}^{-1}\right]\mathbf{x}_{i,\mathcal{F}}\\
&\le |\psi_{01} - \psi_{02}| + |\psi_{11} - \psi_{12}| + |\psi_{21} - \psi_{22}| + 2|\psi_{31} - \psi_{32}|,
\end{aligned} \tag{82}$$

where

$$\begin{aligned}
\psi_{0k} &:= \mathbf{x}_{i,\mathcal{B}_{1,+}}^\top\mathbf{A}_k^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}},\\
\psi_{1k} &:= \mathbf{x}_{i,\mathcal{B}_{1,+}}^\top\mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{D}_k\mathbf{B}_k^\top\mathbf{A}_k^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}},\\
\psi_{2k} &:= \mathbf{x}_{i,\mathcal{F}\setminus\mathcal{B}_{1,+}}^\top\mathbf{D}_k\mathbf{x}_{i,\mathcal{F}\setminus\mathcal{B}_{1,+}},\\
\psi_{3k} &:= \mathbf{x}_{i,\mathcal{B}_{1,+}}^\top\mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{D}_k\mathbf{x}_{i,\mathcal{F}\setminus\mathcal{B}_{1,+}}.
\end{aligned}$$

The last inequality of (82) is the result of the triangle inequality. Our goal is to prove that all the terms in (82) are small with high probability for large values of n, p. Towards this goal, we will first prove that ψ_{11}, ψ_{12}, ψ_{21}, ψ_{22}, ψ_{31}, and ψ_{32} are all individually small. Then, we finally show that the difference $\psi_{01} - \psi_{02}$ is small too.

- **Finding upper bounds for ψ_{21} and ψ_{22}:**
  Note that

$$\sigma_{\max}(\mathbf{D}_k) \le \sigma_{\max}(\mathbf{H}_k^{-1}) = \frac{1}{\sigma_{\min}(\mathbf{H}_k)}. \tag{83}$$

  It is then straightforward to see that since $\mathbf{H}_1$ and $\mathbf{H}_2$ are summations of $\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{B}_0^c}^{\alpha/i})$ and $\lambda\operatorname{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{B}_0^c}^{\alpha/i})$ with a pair of positive semidefinite matrices, and that r has a ridge component in it, we have

$$\sigma_{\min}(\mathbf{H}) \ge 2\lambda\eta.$$

  Using this fact and (83) we obtain

$$\|\mathbf{D}_k\| \le (2\lambda\eta)^{-1} \text{ for } k = 1, 2. \tag{84}$$

It then follows that

$$\psi_{2k} \leq \|\mathbf{D}_k\| \times \|\mathbf{x}_{i,\mathcal{F}\setminus\mathcal{B}_{1,+}}\|^2 \overset{(a)}{\leq} \frac{\max_i \sup_{\mathcal{T}:|\mathcal{T}|=d_n} \|\mathbf{x}_{i,\mathcal{T}}\|^2}{2\lambda\eta}. \tag{85}$$

Note that the reason we have used the maximum on the set $\mathcal{T}$ in bounding $\|\mathbf{x}_{i,\mathcal{F}\setminus\mathcal{B}_{1,+}}\|^2$ is that $\mathcal{F}/\mathcal{B}_{1,i}$ depends on $\mathbf{x}_i$ and hence we cannot use standard concentration results for χ^2 random variables (e.g., Lemma 16). Furthermore, when taking the supremum over sets $\mathcal{T}$, we have to consider all the sets $\mathcal{T}$ whose sizes are smaller than d_n. But as is clear, in (85) we have only considered $\mathcal{T}$ with sizes equal to d_n. This is because the norm of $\|\mathbf{x}_{i,\mathcal{T}'}\|^2 \leq \|\mathbf{x}_{i,\mathcal{T}}\|^2$ if $\mathcal{T}' \subset \mathcal{T}$. We have

$$\begin{aligned}
&\mathbb{P}(\max_i \sup_{\mathcal{T}:|\mathcal{T}|=d_n} \|\mathbf{x}_{i,\mathcal{T}}\|^2 > d_n\rho_{\max}(1+t)) \\
&\leq \sum_i \sum_{\mathcal{T}:|\mathcal{T}|=d_n} \mathbb{P}(\|\mathbf{x}_{i,\mathcal{T}}\|^2 > d_n\rho_{\max}(1+t)) \\
&\leq n\binom{p}{d_n} \mathrm{e}^{-\frac{d_n}{2}(t-\log(1+t))} \\
&\leq n\mathrm{e}^{d_n \log\left(\frac{ep}{d_n}\right)} \mathrm{e}^{-\frac{d_n}{2}(t-\log(1+t))},
\end{aligned} \tag{86}$$

where to obtain the last two inequalities we have used Lemma 16 and Lemma 14. Setting $t = 8\log p$ in this equation, we conclude that

$$\mathbb{P}\left(\max_i \sup_{\mathcal{T}:|\mathcal{T}|=d_n} \|\mathbf{x}_{i,\mathcal{T}}\|^2 > d_n\rho_{max}(1+8\log p)\right) \leq n\mathrm{e}^{d_n \log\left(\frac{ep}{d_n}\right)} \mathrm{e}^{-\frac{d_n}{2}(8\log p - \log(1+8\log p))}. \tag{87}$$

Combining (85) and (87) we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\psi_{2k} > \frac{\rho_{\max} d_n(1+8\log p)}{2\eta\lambda}\right) \\
&\leq n\mathrm{e}^{d_n \log\left(\frac{ep}{d_n}\right)} \mathrm{e}^{-\frac{d_n}{2}(8\log p - \log(1+8\log p))} \\
&\leq \mathrm{e}^{\log(n) + d_n \log\left(\frac{ep}{d_n}\right) - \frac{d_n}{2}(8\log p - \log(1+8\log p))} \\
&\leq np^{-d_n},
\end{aligned} \tag{88}$$

provided $d_n \geq \mathrm{e}\sqrt{1+8\log p}$.

- **Finding upper bounds for ψ_{11} and ψ_{12}:** For better readability, we defer the proof of this bound to Lemma 22:

  Lemma 22. *For $k = 1, 2$ we have a sufficienty large constant $C > 0$ depending only on γ_0 and C_X such that*

$$\begin{aligned}
&\psi_{1k} := \mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}_2^{-1} \mathbf{B}_k \mathbf{B}_k^\top \mathbf{A}_k^{-1} \mathbf{x}_{i,\mathcal{B}_{1,+}} \\
&\leq \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \frac{d_n\log^2 p}{p-d_n} + \frac{C}{p\lambda\eta}
\end{aligned}$$

  with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}} - \mathrm{e}^{-d_n\log p} - q_n - \check{q}_n - \bar{q}_n - \tilde{q}_n$, for sufficiently large p.

- **Finding upper bounds for ψ_{31} and ψ_{32}:**
  We have

$$\begin{aligned}
\psi_{3k} &\leq \sqrt{\psi_{2k} \times \mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}_k^{-1} \mathbf{B}_k \mathbf{D}_k \mathbf{B}_k^\top \mathbf{A}_k^{-1} \mathbf{x}_{i,\mathcal{B}_{1,+}}} \\
&\leq \sqrt{\psi_{2k}\psi_{1k}} \leq (\psi_{1k}+\psi_{2k})/2.
\end{aligned} \tag{89}$$

- **Finding an upper bound for $|\psi_{01}-\psi_{02}|$:**

$$\mathbf{x}_{i,\mathcal{B}_{1,+}}^{\top}(\mathbf{A}_1^{-1}-\mathbf{A}_2^{-1})\mathbf{x}_{i,\mathcal{B}_{1,+}}.$$

The proof of this part is similar to our proof technique for bounding ψ_{11} and ψ_{12}, with a few important differences. As before, there are two sources of dependency between $\mathbf{A}_1, \mathbf{A}_2$ and $\mathbf{x}_i$: (i) The dependency of the input arguments of $\ddot{\ell}$ and $\ddot{r}$ on $\mathbf{x}_i$. (ii) the dependency between the set $\mathcal{B}_{1,+}$ and $\mathbf{x}_i$. The goal is to remove these dependencies. We start with removing the dependency of the input argument of r on $\mathbf{x}_i$. Define

$$\overset{\star}{\mathbf{A}}_2 := \ \mathrm{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}_{/i,\mathcal{B}_{1,+}}^{\top}\mathrm{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^{\alpha})\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}} \tag{90}$$

Define

$$\begin{aligned}\overset{\star}{\boldsymbol{\Delta}}_1 &= \overset{\star}{\mathbf{A}}_1^{-1} - \mathbf{A}_1^{-1},\\ \overset{\star}{\boldsymbol{\Delta}}_2 &= \overset{\star}{\mathbf{A}}_2^{-1} - \mathbf{A}_2^{-1}.\end{aligned} \tag{91}$$

The goal is to obtain a bound on $\|\overset{\star}{\boldsymbol{\Delta}}_1\|$ and $\|\overset{\star}{\boldsymbol{\Delta}}_2\|$. Since the proofs are similar and for notational simplicity we show our claim for $\overset{\star}{\boldsymbol{\Delta}}_2$. Since $\overset{\star}{\boldsymbol{\Delta}}_2 = \mathbf{A}_2^{-1}(\mathbf{A}_2 - \overset{\star}{\mathbf{A}}_2)\ \overset{\star}{\mathbf{A}}_2^{-1}$, and $\mathbf{A}_2 - \overset{\star}{\mathbf{A}}_2 = \mathrm{diag}\left[\lambda\left\{\ddot{r}(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\right\}_{\mathcal{B}_{1,+}}\right] - \mathrm{diag}\left[\lambda\left\{\ddot{r}(\widehat{\boldsymbol{\beta}}^{\alpha})\right\}_{\mathcal{B}_{1,+}}\right]$, we have

$$\begin{aligned}\|\overset{\star}{\boldsymbol{\Delta}}_2\| &\leq \frac{\|\mathbf{A}_2 - \overset{\star}{\mathbf{A}}_2\|}{\sigma_{\min}(\overset{\star}{\mathbf{A}}_2)\sigma_{\min}(\mathbf{A}_2)} \leq \frac{2\lambda\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{(2\lambda\eta)^2}\\ &= \frac{\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{2\lambda\eta^2}.\end{aligned} \tag{92}$$

To obtain (92), we used Part (4) and (5) of Lemma 4 to find a bound on $\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i,k}) - \ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^{\alpha}_{k})$. Furthermore, given the ridge part of the regularizer, the minimum eigenvalues of $\mathbf{A}_2$ and $\overset{\star}{\mathbf{A}}_2$ are $2\lambda\eta$. It is straight forward to see that

$$\left|\mathbf{x}_{i,\mathcal{B}_{1,+}}^{\top}\left(\mathbf{A}_2^{-1} - \overset{\star}{\mathbf{A}}_2^{-1}\right)\mathbf{x}_{i,\mathcal{B}_{1,+}}\right| \leq \|\mathbf{x}_i\|^2\|\overset{\star}{\boldsymbol{\Delta}}_2\| \leq \frac{\|\mathbf{x}_i\|_2^2\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{2\lambda\eta^2}. \tag{93}$$

It is clear that as $\alpha \to \infty$, the upper bound in (93) goes to zero. By replacing $\mathbf{A}_2^{-1}, \mathbf{A}_1^{-1} - \overset{\star}{\mathbf{A}}_2^{-1}$ with $\overset{\star}{\mathbf{A}}_2^{-1}, \overset{\star}{\mathbf{A}}_1^{-1}$ we have removed the dependency between the input argument of $\ddot{r}$ and $\mathbf{x}_i$. As before to remove the dependency of the set $\mathcal{B}_{1,+}$ we lift the problem to a higher dimensional space. Towards this goal, we define

$$\tilde{\mathbf{A}}_2 = \ \mathrm{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\right\}_{\mathcal{B}^+}\right] + \mathbf{X}_{/i,\mathcal{B}^+}^{\top}\mathrm{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^{\alpha})\right]\mathbf{X}_{/i,\mathcal{B}^+},$$

where $\mathcal{B}^+ := \mathcal{B}_{1,+} \cup \mathcal{B}_0$. We write $\tilde{\mathbf{A}}_2$ as

$$\tilde{\mathbf{A}}_2 = \begin{pmatrix}\overset{\star}{\mathbf{A}}_2 & \mathbf{B}\\ \mathbf{B}^{\top} & \mathbf{C}\end{pmatrix}$$

where

$$\begin{aligned}\mathbf{B} &= \mathbf{X}_{\mathcal{B}_{1,+}}^{\top}\,\mathrm{diag}\{\ddot{\underline{\ell}}^{\alpha/i}\}\mathbf{X}_{\mathcal{B}_{1,+}},\\ \mathbf{C} &= \lambda\,\mathrm{diag}\left[\{\ddot{\underline{r}}_\alpha\}_{\mathcal{B}_0}\right] + \mathbf{X}_{\mathcal{B}_0}^{\top}\,\mathrm{diag}\{\ddot{\underline{\ell}}^{\alpha/i}\}\mathbf{X}_{\mathcal{B}_0}.\end{aligned}$$

Using the block matrix inversion lemma, i.e. Lemma 11, we have

$$\tilde{\mathbf{A}}_2^{-1} = \begin{pmatrix} \overset{\star}{\mathbf{A}}_2^{-1} + \overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B}\mathbf{H}\mathbf{B}^\top \overset{\star}{\mathbf{A}}_2^{-1} & -\overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B}\mathbf{H} \\ -\mathbf{H}\mathbf{B}^\top \overset{\star}{\mathbf{A}}^{-1} & \mathbf{H} \end{pmatrix}$$

where $\mathbf{H} = (\mathbf{C} - \mathbf{B}^\top \overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B})^{-1}$. Then we have

$$\begin{aligned} & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}_{1,+}} - \mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}^+}| \\ \leq & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B}\mathbf{H}\mathbf{B}^\top \overset{\star}{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}_{1,+}}| + 2|\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B}\mathbf{H}\mathbf{x}_{i,\mathcal{B}_0}| + |\mathbf{x}_{i,\mathcal{B}_0}^\top \mathbf{H}\mathbf{x}_{i,\mathcal{B}_0}| \\ \leq & \|\mathbf{x}_i\|^2 \|\overset{\star}{\mathbf{A}}_2^{-1}\|^2 \|\mathbf{B}\|^2 \|\mathbf{H}\| + 2\|\mathbf{x}_i\|^2 \|\overset{\star}{\mathbf{A}}_2^{-1}\| \|\mathbf{B}\| \|\mathbf{H}\| + \|\mathbf{x}_i\|^2 \|\mathbf{H}\| \\ \leq & \|\mathbf{x}_i\|^2 \|\mathbf{H}\| \left(\|\overset{\star}{\mathbf{A}}_2^{-1}\| \|\mathbf{B}\| + 1 \right)^2 \end{aligned}$$

To bound the matrix norms in the above equation, we have

- $\|\overset{\star}{\mathbf{A}}_2^{-1}\|$:

$$\sigma_{\min}(\overset{\star}{\mathbf{A}}) \geq 2\lambda\eta \quad \text{and hence} \quad \|\overset{\star}{\mathbf{A}}_2^{-1}\| \leq \frac{1}{2\lambda\eta}.$$

- $\|\mathbf{B}\|$: It follows by definition of ω^s in (76) that

$$\|\mathbf{B}\| \leq \omega^s.$$

- $\|\mathbf{H}\|$: Using a derivation similar to (74), we have

$$\sigma_{\min}(\mathbf{C} - \mathbf{B}^\top \overset{\star}{\mathbf{A}}_2^{-1}\mathbf{B}) \geq \frac{\lambda\alpha}{16}(1-\eta)\kappa_0.$$

Inserting the above bounds for matrix norms, we have

$$\begin{aligned} & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}_{1,+}} - \mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}^+}| \\ & \leq \|\mathbf{x}_i\|^2 \|\mathbf{H}\| \left(\|\overset{\star}{\mathbf{A}}_2^{-1}\| \|\mathbf{B}\| + 1 \right)^2 \\ & \leq \frac{16\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)} \left(1 + \frac{\omega^s}{2\lambda\eta} \right)^2, \end{aligned} \tag{94}$$

with probability larger than $1 - q_n$. Again it is straightforward to check that as $\alpha \to \infty$, the upper bound in (94) goes to zero. Hence, we can now focus on the term $\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}_2^{-1} \mathbf{x}_{i,\mathcal{B}^+}$. Note that in matrix $\tilde{\mathbf{A}}_2^{-1}$, there are $\ddot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha)$. In the next step, we would like to show that the difference between this term and $\ddot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})$ is negligible for large values of n, p and any α. Note that once $\ddot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})$ the dependence between $\tilde{\mathbf{A}}_2^{-1}$ and $\mathbf{x}_i$ reduces to only the dependence of the two terms on $\mathcal{B}_{1,+}$. Define

$$\check{\mathbf{A}}_2 := \operatorname{diag}\left[\lambda \left\{ \ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) \right\} \right] + \mathbf{X}_{/i,\mathcal{B}^+}^\top \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) \right] \mathbf{X}_{/i,\mathcal{B}^+}, \tag{95}$$

The goal is to show that

$$\left| \mathbf{x}_{i,\mathcal{B}^+}^\top \left(\check{\mathbf{A}}_2^{-1} - \tilde{\mathbf{A}}_2^{-1} \right) \mathbf{x}_{i,\mathcal{B}^+} \right|$$

is small. Towards this goal, define

$$\boldsymbol{\Delta}_\ell := \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha) \right] - \operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i}) \right]$$

We have

$$\tilde{\mathbf{A}}_2 = \check{\mathbf{A}}_2 + \mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}$$

According to Lemma 12

$$\begin{aligned}
&\tilde{\mathbf{A}}_2^{-1} \\
=\ &\check{\mathbf{A}}_2^{-1} - \check{\mathbf{A}}_2^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}_2^{-1} \\
&+ \check{\mathbf{A}}_2^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\tilde{\mathbf{A}}_2^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}_2^{-1}.
\end{aligned}$$

Hence, if we define

$$\begin{aligned}
\check{\mathbf{v}} &= \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}_2^{-1}\mathbf{x}_{i,\mathcal{B}^+} \\
\check{\mathbf{u}}^\top &= \mathbf{x}_{i,\mathcal{B}^+}\check{\mathbf{A}}_2^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\tilde{\mathbf{A}}_2^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top,
\end{aligned} \tag{96}$$

then we have

$$\begin{aligned}
&|\mathbf{x}_{i,\mathcal{B}^+}^\top(\check{\mathbf{A}}_2^{-1} - \tilde{\mathbf{A}}_2^{-1})\mathbf{x}_{i,\mathcal{B}^+}| \le |\check{\mathbf{v}}^\top \boldsymbol{\Delta}_\ell \check{\mathbf{v}}| + |\check{\mathbf{u}}^\top \boldsymbol{\Delta}_\ell \check{\mathbf{v}}| \\
\le\ &\sqrt{\sum_i \boldsymbol{\Delta}_{\ell,ii}^2}\sqrt{\sum_i \check{\mathbf{u}}_i^2\check{\mathbf{v}}_i^2} + \sqrt{\sum_i \Delta_{\ell,ii}^2}\sqrt{\sum_i \check{\mathbf{v}}_i^4} \\
\overset{(c)}{\le}\ &\sqrt{\sum_i \boldsymbol{\Delta}_{\ell,ii}^2}(\sum_i \check{\mathbf{u}}_i^4)^{\frac{1}{4}}(\sum_i \check{\mathbf{v}}_i^4)^{\frac{1}{4}} + \sqrt{\sum_i \Delta_{\ell,ii}^2}(\sum_i \check{\mathbf{v}}_i^4)^{\frac{1}{2}} \\
\le\ &\sqrt{\sum_i \boldsymbol{\Delta}_{\ell,ii}^2}(\sum_i \check{\mathbf{u}}_i^2)^{\frac{1}{2}}(\max_i |\check{\mathbf{v}}_i|)^{\frac{1}{2}}(\sum_i \check{\mathbf{v}}_i^2)^{\frac{1}{4}} + \sqrt{\sum_i \boldsymbol{\Delta}_{\ell,ii}^2}(\max_i |\check{\mathbf{v}}_i|)(\sum_i \check{\mathbf{v}}_i^2)^{\frac{1}{2}}.
\end{aligned} \tag{97}$$

By our assumptions (see (146) for a more detailed calculation), we have

$$\sqrt{\sum_j \Delta_{\ell,ii}^2} \le \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|_2}{2\lambda\eta}, \tag{98}$$

with probability larger than $1 - \tilde{q}_n - \check{q}_n$. Furthermore,

$$\begin{aligned}
\|\check{\mathbf{v}}\|_2 &= \|\mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}\| \\
&\le \frac{\|\mathbf{X}\|\|\mathbf{x}_i\|_2}{\sigma_{\min}(\check{\mathbf{A}})} \le \frac{\|\mathbf{X}\|\|\mathbf{x}_i\|_2}{2\lambda\eta}
\end{aligned} \tag{99}$$

and

$$\|\check{\mathbf{u}}\| \le \frac{\|\mathbf{x}_i\|_2^2\|\mathbf{X}\|^2\|\boldsymbol{\Delta}_\ell\|}{\sigma_{\min}^2(\check{\mathbf{A}})} \le \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|_2^3\|\mathbf{X}\|^2}{8\lambda^3\eta^3}, \tag{100}$$

with probability larger than $1 - \check{q}_n$. Finally,

$$\begin{aligned}
&\mathbb{P}\left(\max_i \check{\mathbf{v}}_i > \gamma_5(n) + t\right) \\
\le\ &\mathrm{e}^{2d_n \log\frac{ep}{d_n}}\left(\mathrm{e}^{-p} + 2\mathrm{e}^{-\frac{n\lambda\eta t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right).
\end{aligned} \tag{101}$$

For the definition of $\gamma_5(n)$ and the detailed derivation of (101), please check (156) and (161). Hence, if we set $t = \sqrt{\frac{d_n \log^2 p}{\lambda\eta n}}$, and define

$$\gamma_7(n) = \gamma_5(n) + \sqrt{\frac{d_n \log^2 p}{\lambda\eta n}}$$

by combining (97), (98), (99), (100), (101), we obtain

$$\begin{aligned}
&\mathbb{P}(|\mathbf{x}_{i,\mathcal{B}^+}^\top(\check{\mathbf{A}}_2^{-1}-\tilde{\mathbf{A}}_2^{-1})\mathbf{x}_{i,\mathcal{B}^+}|>\gamma_7(n))\\
&\le \mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{d_n\log^2 p}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right)+q_n+\check{q}_n\\
&\le \mathrm{e}^{-d_n\log p}+q_n+\check{q}_n
\end{aligned}\tag{102}$$

provided $d_n\le\frac{p}{C}$ for a sufficiently large constant $C>0$, where we use the fact that $\rho_{\max}=O(p^{-1})$.

Note that we were originally interested in bounding $|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top(\mathbf{A}_1^{-1}-\mathbf{A}_2^{-1})\mathbf{x}_{1,\mathcal{B}_{1,+}}|$. We have

$$\begin{aligned}
&|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top(\mathbf{A}_1^{-1}-\mathbf{A}_2^{-1})\mathbf{x}_{1,\mathcal{B}_{1,+}}|\\
\le\ &|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top(\mathbf{A}_1^{-1}-\overset{\star}{\mathbf{A}}_1^{-1})\mathbf{x}_{1,\mathcal{B}_{1,+}}|\\
&+|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top(\mathbf{A}_2^{-1}-\overset{\star}{\mathbf{A}}_2^{-1})\mathbf{x}_{1,\mathcal{B}_{1,+}}|\\
&+|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top\overset{\star}{\mathbf{A}}_1^{-1}\mathbf{x}_{1,\mathcal{B}_{1,+}}-\mathbf{x}_{1,\mathcal{B}^+}^\top\tilde{\mathbf{A}}_1^{-1}\mathbf{x}_{1,\mathcal{B}_{1,+}}|\\
&+|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top\overset{\star}{\mathbf{A}}_2^{-1}\mathbf{x}_{1,\mathcal{B}_{1,+}}-\mathbf{x}_{1,\mathcal{B}^+}^\top\tilde{\mathbf{A}}_2^{-1}\mathbf{x}_{1,\mathcal{B}_{1,+}}|\\
&+|\mathbf{x}_{1,\mathcal{B}^+}^\top\left(\tilde{\mathbf{A}}_2^{-1}-\check{\mathbf{A}}_2^{-1}\right)\mathbf{x}_{1,\mathcal{B}^+}|\\
&+|\mathbf{x}_{1,\mathcal{B}^+}^\top\left(\tilde{\mathbf{A}}_1^{-1}-\check{\mathbf{A}}_1^{-1}\right)\mathbf{x}_{1,\mathcal{B}^+}|.
\end{aligned}\tag{103}$$

Hence, by combining (93), (94), (102) we obtain that if

$$\gamma_8(n,\alpha):=2\gamma_7(n)+2\frac{\|\mathbf{x}_i\|_2^2\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{2\lambda\eta^2}+2\frac{\|\mathbf{x}_i\|^2}{\gamma_1^s(\alpha)}\left(1+\frac{\omega^s}{2\lambda\eta}\right)^2,$$

then

$$\begin{aligned}
&\mathbb{P}(|\mathbf{x}_{1,\mathcal{B}_{1,+}}^\top(\mathbf{A}_1^{-1}-\mathbf{A}_2^{-1})\mathbf{x}_{1,\mathcal{B}_{1,+}}|>\gamma_8(n,\alpha))\\
&\le \mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{d_n\log^2 p}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right)+q_n+\check{q}_n+\tilde{q}_n\\
&\le \mathrm{e}^{-d_n\log p}+q_n+\check{q}_n+\tilde{q}_n,
\end{aligned}\tag{104}$$

provided $d_n\le\frac{p}{C}$ for a sufficiently large constant $C>0$, where we use the fact that $\rho_{\max}=O(p^{-1})$.

Equipped with the bounds on $|\psi_{01}-\psi_{02}|$, ψ_{1k}, ψ_{2k} and ψ_{3k}, we now return to equation (82) to obtain

$$\begin{aligned}
&\left|\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\mathrm{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i})+\mathbf{X}_{\mathcal{F}}^\top\mathrm{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}}-\mathbf{x}_{i,\mathcal{F}}^\top(\lambda\mathrm{diag}(\ddot{\underset{\sim}{r}}_{\mathcal{F}}^{\alpha/i})+\mathbf{X}_{\mathcal{F}}^\top\mathrm{diag}(\ddot{\underset{\sim}{\ell}}^{\alpha/i})\mathbf{X}_{\mathcal{F}})^{-1}\mathbf{x}_{i,\mathcal{F}}\right|\\
&=\mathbf{x}_{i,\mathcal{F}}^\top\left[\begin{pmatrix}\mathbf{A}_1&\mathbf{B}_1\\\mathbf{B}_1^\top&\mathbf{C}_1\end{pmatrix}^{-1}-\begin{pmatrix}\mathbf{A}_2&\mathbf{B}_2\\\mathbf{B}_2^\top&\mathbf{C}_2\end{pmatrix}^{-1}\right]\mathbf{x}_{i,\mathcal{F}}\\
&\le|\psi_{01}-\psi_{02}|+|\psi_{11}-\psi_{12}|+|\psi_{21}-\psi_{22}|+2|\psi_{31}-\psi_{32}|\\
&\le\gamma_8(n,\alpha)+\frac{2\rho_{\max}d_n(1+8\log p)}{\eta\lambda}+2(\psi_{11}+\psi_{12})\\
&\le\frac{2\rho_{\max}d_n(1+8\log p)}{\eta\lambda}+2(\psi_{11}+\psi_{12})\ +2\sqrt{\frac{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}{n\lambda\eta}\log 2n}\\
&\quad+2\sqrt{\frac{d_n\log^2 p}{n\lambda\eta}}+\frac{16\|\mathbf{x}_i\|^2}{\lambda\alpha(1-\eta)}\left(1+\frac{\omega^s}{2\lambda\eta}\right)^2+\frac{\alpha\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{\lambda\eta^2}\\
&\le\frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}}+\frac{Cd_n\log^2 p}{p\lambda\eta}+\frac{Cd_n}{n\lambda^2\eta^2}+\sqrt{\frac{C\log p}{p\lambda\eta}}
\end{aligned}$$

with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}}-(n+2)p^{-d_n}-2q_n-2\check{q}_n-2\tilde{q}_n$. This finishes the proof of Theorem 4. □

5.6 Proof of Lemma 6

We prove the lemma only for $\widehat{\boldsymbol{\beta}}$, because the same proof and constants apply to the leave-one-out estimators $\widehat{\boldsymbol{\beta}}_{/i}$.

Recall that in Lemmas 5, 7 and 8, $\widehat{\mu}$ is the empirical distribution of $\widehat{\boldsymbol{\beta}}$, μ^* is the distribution of $\frac{b_*}{b_*+2\lambda\eta\tau_*}\mathrm{soft}(\tau_* Z+\Theta, \frac{\lambda(1-\eta)\tau_*}{b_*})$ with $(\Theta, Z)\sim\frac{1}{p}\sum_{k=1}^{p}\delta_{\beta_k^*}\otimes N(0,1)$, and (b_*,τ_*) is the unique solution of the equations (51) and (52). Note that by Lemma 24, there exists $0<b_{\min}<b_{\max}$ and $\sigma<\tau_{\max}$ depending only on model parameters such that $\sigma<\tau_*<\tau_{\max}$ and $b_{\min}<b_*<b_{\max}$.

Let us first define a smoothed indicator function as follows:

$$h_\zeta(x)=\begin{cases}0 & ,|x|\geq\kappa_1+\zeta\\ 1-\dfrac{|x|-\kappa_1}{\zeta} & ,\kappa_1\leq|x|<\kappa_1+\zeta\\ 1 & ,|x|<\kappa_1\end{cases} \tag{105}$$

Note that h_ζ is $\frac{1}{\zeta}$-Lipschitz, and $\mathbb{1}_{[-\kappa_1,\kappa_1]}\leq h_\zeta\leq\mathbb{1}_{[-\kappa_1-\zeta,\kappa_1+\zeta]}$. We then have

$$\begin{aligned}
&\frac{1}{p}\#\{k:0<|\widehat{\beta}_k|\leq\kappa_1\}\\
&=\widehat{\mu}([-\kappa_1,\kappa_1])-\widehat{\mu}(\{0\})\\
&=\int\mathbb{1}_{[-\kappa_1,\kappa_1]}(x)d\widehat{\mu}(x)-\widehat{\mu}(\{0\})\\
&\leq\int h_\zeta(x)d\widehat{\mu}(x)-\widehat{\mu}(\{0\})\\
&\leq\int h_\zeta(x)d\mu_*(x)+\left|\int h_\zeta(x)d\mu_*(x)-\int h_\zeta(x)d\widehat{\mu}_*(x)\right|\\
&\quad-\widehat{\mu}(\{0\})\\
&\leq\mu_*([-\kappa_1-\zeta,\kappa_1+\zeta])-\mu_*(\{0\})\\
&\quad+\left|\int h_\zeta(x)d\mu_*(x)-\int h_\zeta(x)d\widehat{\mu}_*(x)\right|\\
&\quad+|\mu_*(\{0\})-\widehat{\mu}(\{0\})|.
\end{aligned}\tag{106}$$

We now aim to obtain upper bounds for the final three terms in (106). First note that 0 is the unique discontinuity of μ_*, therefore

$$\mu_*([-\kappa_1-\zeta,\kappa_1+\zeta])-\mu_*(\{0\})\leq 2(\kappa_1+\zeta)\max_{x\neq0}f_*(x)$$

where f_* is the density of the absolutely continuous part of μ_* w.r.t. the Lebesgue measure. It can be verified directly via definition that

$$\begin{aligned}
&f_*(x)\\
&=\frac{1}{p}\left(\frac{1}{\tau_*}+\frac{2\lambda\eta}{b_*}\right)\sum_{k=1}^{p}\phi\left(x\left(\frac{1}{\tau_*}+\frac{2\lambda\eta}{b_*}\right)+\mathrm{sign}(x)\frac{\lambda(1-\eta)}{b_*}-\frac{\beta_k^*}{\tau_*}\right)
\end{aligned}$$

where $\phi(\cdot)$ is the standard Gaussian density, and therefore

$$f_*(x)\leq(\sqrt{2\pi})^{-1}\left(\frac{1}{\tau_*}+\frac{2\lambda\eta}{b_*}\right)\leq(\sqrt{2\pi})^{-1}\left(\frac{1}{\sigma}+\frac{2\lambda_{max}}{b_{min}}\right)=:C_f$$

So we have

$$\mu_*([-\kappa_1-\zeta,\kappa_1+\zeta])-\mu_*(\{0\})\leq 2C_f(\kappa_1+\zeta). \tag{107}$$

To bound the term $\left|\int h_\zeta(x)\mu_*(x) - \int h_\zeta(x)\widehat{\mu}_*(x)\right|$ in (106) we use the following approach. First, note that h_ζ is $1/\zeta$-Lipschitz. Thus,

$$\begin{aligned}
&\left|\int h_\zeta(x)\mu_*(x) - \int h_\zeta(x)\widehat{\mu}_*(x)\right| \\
&\leq \frac{1}{\zeta}\sup_{g\in L_1}\left|\int g(x)d\mu_*(x) - \int g(x)d\widehat{\mu}_*(x)\right| \\
&= W_1(\widehat{\mu},\mu_*)/\zeta \leq W_2(\widehat{\mu},\mu_*)/\zeta. \qquad (108)
\end{aligned}$$

In these equations, W_q denotes the Wasserstein-q distance between two measures for $q = 1, 2$. The last inequality uses monotonicity of W_q in q: for $0 < p < q$:

$$W_p(\mu,\nu) := \inf_{X\sim\mu, Y\sim\nu}\|X - Y\|_{\mathcal{L}_p} \leq \inf_{X\sim\mu, Y\sim\nu}\|X - Y\|_{\mathcal{L}_q} =: W_q(\mu,\nu)$$

Furthermore, by Lemma 5, for any $0 < \epsilon < \frac{1}{2}$, we have

$$W_2(\widehat{\mu},\mu_*)/\zeta \leq \sqrt{\varepsilon}/\zeta \qquad (109)$$

with probability at least $1 - C_1\varepsilon^{-2}\mathrm{e}^{-c_1 p\varepsilon^3(\log\varepsilon)^{-2}}$, for some constants C_1, c_1. Combining (108) and (110) we conclude that

$$\left|\int h_\zeta(x)\mu_*(x) - \int h_\zeta(x)\widehat{\mu}_*(x)\right| \leq \sqrt{\varepsilon}/\zeta \qquad (110)$$

with probability at least $1 - C_1\varepsilon^{-2}\mathrm{e}^{-c_1 p\varepsilon^3(\log\varepsilon)^{-2}}$.

For the last term of (106) we use Lemma 8 to get:

$$|\widehat{\mu}(0) - \mu_*(0)| = |\|\widehat{\boldsymbol{\beta}}\|_0 - s_*| < \varepsilon' \qquad (111)$$

with probability at least $1 - C_2\varepsilon'^{-6}\mathrm{e}^{-c_2 p\varepsilon'^6}$.

Using (106), (107), (110), and (111) together we have

$$\frac{1}{p}\left|\{k : 0 < |\widehat{\beta}_k| \leq \kappa_1\}\right| \leq 2C_f(\kappa_1 + \zeta) + \sqrt{\varepsilon}/\zeta + \varepsilon'$$

with probability at least $1 - C_1\varepsilon^{-2}\mathrm{e}^{-c_1 p\varepsilon^3(\log\varepsilon)^{-2}} - C_2\varepsilon'^{-6}\mathrm{e}^{-c_2 p\varepsilon'^6}$. We observe first that the bound is minimized at $\zeta = (2C_f)^{-\frac{1}{2}}\varepsilon^{\frac{1}{4}}$ without changing the probability.

Next, we set $\varepsilon = 2(c_1 p)^{-\frac{1}{3}}\log p$ and $\varepsilon' = 3^{\frac{1}{3}}(c_2 p)^{-\frac{1}{6}}(\log p)^{\frac{1}{6}}$ to get:

$$\begin{aligned}
&\left|\{k : 0 < |\widehat{\beta}_k| \leq \kappa_1\}\right| \\
&\leq 2C_f p\kappa_1 + C_3 p^{\frac{11}{12}}(\log p)^{\frac{1}{4}} + C_4 p^{\frac{5}{6}}(\log p)^{\frac{1}{6}} \\
&\leq Cp\kappa_1 + Cp^{\frac{11}{12}}(\log p)^{\frac{1}{4}} \\
&\leq Cp^{\frac{11}{12}}(\log p)^{\frac{1}{4}}
\end{aligned}$$

with probability at least $1 - C'p^{-7}$, for sufficiently large p and some constants C, C', provided $\kappa_1 = o(p^{-\frac{1}{12}}(\log p)^{\frac{1}{4}})$ By an identical argument, an analogous statement holds for $\widehat{\beta}_{/i,k}$. The proof is completed by a union bound over $1 \leq i \leq n$. □

5.7 Proof of Lemma 22

First note that by using (84) we obtain

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{D}_k\mathbf{B}_k^\top\mathbf{A}_k^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} \\
&\leq (2\lambda\eta)^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}_k^{-1}\mathbf{B}_k\mathbf{B}_k^\top\mathbf{A}_k^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}}.
\end{aligned}$$

We will use the notation

$$\omega_s := \|\mathbf{X}^\top \mathbf{X}\| \sup_{t\in[0,1]} \max_{1\le i\le n} \ddot{\ell}_i(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}\,;$$
$$\rho(\alpha) := \lambda\left(2\eta + \frac{\alpha(1-\eta)\kappa_0}{8}\right). \tag{112}$$

In the rest of the proof, we focus on $k=2$ and find an upper bound for $|\psi_{2k}|$. The proof for $k=1$ is similar and will hence be skipped. Furthermore, for notational simplicity, we drop the subscripts of k from matrices $\mathbf{A}_k$ and $\mathbf{B}_k$.

Intuitively speaking, the matrix $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top\mathbf{A}^{-1}$ has rank d_n. Hence, if $\mathbf{G}$ were independent of $\mathbf{x}_i$, we could have used concentration of χ^2 random variables to show that $\psi_{1k} = O_p(d_n\rho_{\max})$. However, there are multiple sources of dependency between $\mathbf{x}_i$ and $\mathbf{G}$: (1) The input arguments of $\ddot{r}$, (2) the input arguments of $\ddot{\ell}$, and (3) The dependence of $\mathcal{B}_{1,+}$ and $\mathcal{F}$ on $\mathbf{x}_i$. In the rest of the proof, we aim to handle these three dependencies and show that ψ_{1k} is still small. We remind the reader that

$$\mathbf{A} = \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}^\top_{/i,\mathcal{B}_{1,+}}\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}}$$
$$\mathbf{B} = \mathbf{X}^\top_{/i,\mathcal{B}_{1,+}}\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{F}\setminus\mathcal{B}_{1,+}}.$$

We start by defining

$$\overset{\star}{\mathbf{A}} = \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}^\top_{/i,\mathcal{B}_{1,+}}\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}}. \tag{113}$$

Note that in $\overset{\star}{\mathbf{A}}$ we have removed one source of dependence, i.e., the dependence of the input argument of $\ddot{r}$ on $\mathbf{x}_i$. As the first step we would like to show that the difference

$$\left|\mathbf{x}^\top_{i,\mathcal{B}_{1,+}}\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top\mathbf{A}^{-1} - \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top\overset{\star}{\mathbf{A}}^{-1}\right)\mathbf{x}_{i,\mathcal{B}_{1,+}}\right|$$

is negligible for large values of α. Define

$$\overset{\star}{\boldsymbol{\Delta}} = (\mathbf{A})^{-1} - (\overset{\star}{\mathbf{A}})^{-1} = \mathbf{A}^{-1}(\overset{\star}{\mathbf{A}} - \mathbf{A})\,\overset{\star}{\mathbf{A}}^{-1} \tag{114}$$

Since

$$\overset{\star}{\mathbf{A}} - \mathbf{A} = \operatorname{diag}\left[\lambda\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})_{\mathcal{B}_{1,+}}\right] - \operatorname{diag}\left[\lambda\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha)_{\mathcal{B}_{1,+}}\right],$$

according to Lemma 4, $\|\overset{\star}{\mathbf{A}} - \mathbf{A}\| \le 4\lambda\alpha(1-\eta)e^{-\frac{1}{2}\alpha\kappa_1}$. We conclude that

$$\|\overset{\star}{\boldsymbol{\Delta}}\| \le \frac{\|\mathbf{A} - \overset{\star}{\mathbf{A}}\|}{\sigma_{\min}(\mathbf{A})\sigma_{\min}(\overset{\star}{\mathbf{A}})} \le \frac{\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{\lambda\eta^2},$$

where we have used the lower bound $2\lambda\eta$ for $\sigma_{\min}(\mathbf{A}_2)$ and $\sigma_{\min}(\overset{\star}{\mathbf{A}})$. This lower bound is obtained from the ridge part of the regularizer. Define

$$\gamma^s_0(\alpha) := \frac{\alpha(1-\eta)\mathrm{e}^{-\frac{1}{2}\alpha\kappa_1}}{\lambda\eta^2}. \tag{115}$$

As we will discuss later, we will ensure that $\gamma^s_0(\alpha) \to 0$ as $\alpha \to \infty$. Hence, $\| \overset{\star}{\boldsymbol{\Delta}} \|$ will be small too. Using this

result and the triangle inequality we have

$$
\begin{aligned}
& |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \Big(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top\mathbf{A}^{-1} - \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top\overset{\star}{\mathbf{A}}^{-1}\Big)\mathbf{x}_{i,\mathcal{B}_{1,+}}| \\
\leq\ & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{\Delta}}\mathbf{B}\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}}| + |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{\Delta}}\mathbf{B}\mathbf{B}^\top\overset{\star}{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}}| \\
\leq\ & \|\mathbf{x}_i\|^2\|\mathbf{B}\|^2\|\overset{\star}{\mathbf{\Delta}}\|\left(\frac{1}{\sigma_{\min}(\mathbf{A})}+\frac{1}{\sigma_{\min}(\overset{\star}{\mathbf{A}})}\right) \\
\leq\ & \|\mathbf{x}_i\|^2\|\mathbf{B}\|^2\|\overset{\star}{\mathbf{\Delta}}\|\left(\frac{1}{\sigma_{\min}(\mathbf{A})}+\frac{1}{\sigma_{\min}(\overset{\star}{\mathbf{A}})}\right) \\
\leq\ & \|\mathbf{x}_i\|^2\|\mathbf{B}\|^2\gamma_0^s(\alpha)\left(\frac{1}{\lambda\eta}\right). \qquad (116)
\end{aligned}
$$

Similar to the proof of (71), it is not hard to see that

$$\|\mathbf{B}_2\| \leq \omega_s$$

under event $\mathcal{A}_L$. Furthermore, given the scaling we have considered in our paper, as will be clarified later, $\|\mathbf{X}^\top\mathbf{X}\| = O_p(1)$. Hence, the term $\|\mathbf{x}_i\|^2\|\mathbf{B}\|^2\gamma_0^s(\alpha)(\frac{1}{\lambda\eta})$ will go to zero with high probability as $\alpha\to\infty$.

In the rest of our proof, we will work with $\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top\overset{\star}{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}}$. Still there are two sources of dependency between the middle matrix and $\mathbf{x}_{i,\mathcal{B}_{1,+}}$: (1) the input argument of $\ddot{\ell}$, and (2) The dependence of $\mathcal{B}_{1,+}$ and $\mathcal{F}$ on $\mathbf{x}_i$. In order to remove the dependence of $\mathcal{F}$ and $\mathcal{B}_{1,+}$, we lift the problem to a higher dimensional problem for the reasons that will become clearer later. Define the two sets:

$$
\begin{aligned}
\mathcal{B}^+ &:= \mathcal{B}_{1,+}\cup\mathcal{B}_0, \\
\mathcal{B}^- &:= \mathcal{F}\setminus\mathcal{B}_{1,+}. \qquad (117)
\end{aligned}
$$

Use these two sets to define

$$
\begin{aligned}
\tilde{\mathbf{A}} &= \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right\}_{\mathcal{B}^+}\right] + \mathbf{X}_{/i,\mathcal{B}^+}^\top\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}^+} \\
\tilde{\mathbf{B}} &= \mathbf{X}_{/i,\mathcal{B}^+}^\top\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}^-},
\end{aligned}
$$

As will be clarified later, working with $\mathcal{B}^+$ and $\mathcal{B}^-$ will be helpful when we would like to remove the dependancies that are caused by $\mathcal{B}^+$ and $\mathcal{B}^-$. Hence, as the first step our aim is to obtain an upper bound on the difference

$$|\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top\overset{\star}{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} - \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}|.$$

We write $\tilde{\mathbf{A}}$ as

$$\tilde{\mathbf{A}} = \begin{pmatrix}\mathbf{E} & \mathbf{F}\\ \mathbf{F}^\top & \mathbf{G}\end{pmatrix}, \qquad (118)$$

where

$$
\begin{aligned}
\mathbf{E} &= \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right\}_{\mathcal{B}_{1,+}}\right] + \mathbf{X}_{/i,\mathcal{B}_{1,+}}^\top\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_{1,+}}, \\
\mathbf{F} &= \mathbf{X}_{/i,\mathcal{B}_{1,+}}^\top\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_0}, \\
\mathbf{G} &= \operatorname{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right\}_{\mathcal{B}_0}\right] + \mathbf{X}_{/i,\mathcal{B}_0}^\top\operatorname{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}_0}.
\end{aligned}
$$

Similar to the proof of (71) it can be checked that

$$\|\mathbf{F}\|_2 \leq \omega^s \qquad (119)$$

with probability at least $1-q_n$. Using the matrix inversion of block matrices (Lemma 11) we have

$$\tilde{\mathbf{A}}^{-1} = \begin{pmatrix} \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}\mathbf{H}\mathbf{F}^\top\mathbf{E}^{-1} & -\mathbf{E}^{-1}\mathbf{F}\mathbf{H} \\ -\mathbf{H}^\top\mathbf{F}^\top\mathbf{E}^{-1} & \mathbf{H} \end{pmatrix}, \tag{120}$$

where

$$\mathbf{H} = (\mathbf{G} - \mathbf{F}^\top\mathbf{E}^{-1}\mathbf{F})^{-1}.$$

Furthermore, similar to the derivation in (74) we have

$$\|\mathbf{H}\| = \ (\sigma_{\min}(\mathbf{G} - \mathbf{F}^\top\mathbf{E}^{-1}\mathbf{F}))^{-1} \leq \frac{1}{\rho(\alpha) - \frac{(\omega^s)^2}{2\lambda\eta}}, \tag{121}$$

with probability $1-q_n$. Both $\rho(\alpha)$ and ω^s are defined at the beginning of the proof. It follows from the definition of $\rho(\alpha)$ that $\frac{1}{\rho(\alpha)-\frac{(\omega^s)^2}{2\lambda\eta}}$ goes to zero for large values of α and hence the norm of $\mathbf{H}$ should be considered as a small number. If we define

$$\bar{\mathbf{A}}^\dagger := \begin{pmatrix} \overset{\star}{\mathbf{A}}^{-1} & \mathbf{0}_{|\mathcal{B}_1\times\mathcal{B}_0|} \\ \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_1|} & \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_0|} \end{pmatrix}, \tag{122}$$

and

$$\boldsymbol{\Delta} := \bar{\mathbf{A}}^\dagger - \tilde{\mathbf{A}}^{-1},$$

then it is straightforward to show that

$$\begin{aligned} & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} - \ \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}| \\ = & |\mathbf{x}_{i,\mathcal{B}^+}^\top\bar{\mathbf{A}}^\dagger\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\bar{\mathbf{A}}^\dagger\mathbf{x}_{i,\mathcal{B}^+} - \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}| \\ \leq & |\mathbf{x}_{i,\mathcal{B}^+}^\top\boldsymbol{\Delta}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\boldsymbol{\Delta}\mathbf{x}_{i,\mathcal{B}^+}| + 2|\mathbf{x}_{i,\mathcal{B}^+}^\top\boldsymbol{\Delta}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\bar{\mathbf{A}}^\dagger\mathbf{x}_{i,\mathcal{B}^+}| \\ \leq & \|\mathbf{x}_i\|^2\|\tilde{\mathbf{B}}\|^2\|\boldsymbol{\Delta}\|\left(\|\boldsymbol{\Delta}\| + 2\|\bar{\mathbf{A}}^\dagger\|\right). \end{aligned} \tag{123}$$

To find an upper bound for (123) we have to bound $\|\boldsymbol{\Delta}\|$, $\|\tilde{\mathbf{B}}_2\|$ and $\|\bar{\mathbf{A}}^\dagger\|$. Similar to our previous calculations, we have under event $\mathcal{A}_L$ that

$$\|\tilde{\mathbf{B}}_2\| \leq \omega^s\,, \ \|\bar{\mathbf{A}}_2^\dagger\| = \frac{1}{\sigma_{\min}(\overset{\star}{\mathbf{A}}_2)} \leq \frac{1}{2\lambda\eta}. \tag{124}$$

Finally, by using Weyl's theorem (Lemma 9), we have

$$\begin{aligned} \|\boldsymbol{\Delta}\| \leq & \left\|\begin{pmatrix} \mathbf{E}^{-1}\mathbf{F}\mathbf{H}\mathbf{F}^\top\mathbf{E}^{-1} & \mathbf{0}_{|\mathcal{B}_1\times\mathcal{B}_0|} \\ \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_1|} & \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_0|}) \end{pmatrix}\right\| + \left\|\begin{pmatrix} \mathbf{0}_{|\mathcal{B}_1\times\mathcal{B}_1|} & \mathbf{0}_{|\mathcal{B}_1\times\mathcal{B}_0|} \\ \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_1|} & \mathbf{H} \end{pmatrix}\right\| + \left\|\begin{pmatrix} \mathbf{0}_{|\mathcal{B}_1\times\mathcal{B}_1|} & \mathbf{E}^{-1}\mathbf{F}\mathbf{H} \\ \mathbf{H}^\top\mathbf{F}^\top\mathbf{E}^{-1} & \mathbf{0}_{|\mathcal{B}_0\times\mathcal{B}_0|}) \end{pmatrix}\right\| \\ \leq & \frac{\|\mathbf{H}\|\|\mathbf{F}\|^2}{\sigma_{\min}^2(\mathbf{E})} + \|\mathbf{H}\| + 2\frac{\|\mathbf{F}\|\|\mathbf{H}\|}{\sigma_{\min}(\mathbf{E})} \\ \leq & \frac{1}{\rho(\alpha) - \frac{(\omega^s)^2}{2\lambda\eta}}\left(1 + \frac{\omega^s}{2\lambda\eta} + \frac{(\omega^s)^2}{(2\lambda\eta)^2}\right), \end{aligned} \tag{125}$$

under event $\mathcal{A}_L$. To obtain the last inequality we have used (119), and (121). Define

$$\gamma_1^s(\alpha) := \frac{1}{\rho(\alpha) - \frac{(\omega^s)^2}{2\lambda\eta}}\left(1 + \frac{\omega^s}{2\lambda\eta} + \frac{(\omega^s)^2}{(2\lambda\eta)^2}\right). \tag{126}$$

By combining (123), (124) and (125) we conclude that

$$\begin{aligned} & |\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{B}\mathbf{B}^\top \overset{\star}{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} - \ \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}| \\ = & \ \|\mathbf{x}_i\|^2\|\tilde{\mathbf{B}}\|^2\|\boldsymbol{\Delta}\|\left(\|\boldsymbol{\Delta}\| + 2\|\bar{\mathbf{A}}^\dagger\|\right) \\ \leq & \ \|\mathbf{x}_i\|^2(\omega^s)^2\|\boldsymbol{\Delta}\|(\|\boldsymbol{\Delta}\| + \frac{1}{\lambda\eta}), \\ \leq & \ \|\mathbf{x}_i\|^2(\omega^s)^2\gamma_1^s(\alpha)\left(\gamma_1^s(\alpha) + \frac{1}{\lambda\eta}\right), \end{aligned} \tag{127}$$

under event $\mathcal{A}_L$. As discussed before, we eventually show that as $\alpha \to \infty$, $\gamma_1^s(\alpha) \to 0$ and hence the upper bound in (127) will go to zero. Hence, in the rest of the proof, we aim to show that $\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top \tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$ is small.

In the next step we would like to remove the dependence of the input argument of $\ddot{\ell}$ on $\mathbf{x}_i$. Define,

$$\check{\mathbf{A}} := \text{diag}\left[\lambda\left\{\ddot{r}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right\}\right] + \mathbf{X}_{/i,\mathcal{B}^+}^\top \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}^+},$$
$$\check{\mathbf{B}} := \mathbf{X}_{/i,\mathcal{B}^+}^\top \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right]\mathbf{X}_{/i,\mathcal{B}^-}.$$

Define

$$\boldsymbol{\Delta}_\ell := \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha)\right] - \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right]$$

We have

$$\tilde{\mathbf{A}} = \check{\mathbf{A}} + \mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}$$

According to Lemma 12,

$$\tilde{\mathbf{A}}^{-1} = \check{\mathbf{A}}^{-1} - \check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1} + \check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\tilde{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1} \tag{128}$$

Also, we have

$$\tilde{\mathbf{B}} - \check{\mathbf{B}} = \mathbf{X}_{/i,\mathcal{B}^+}^\top \boldsymbol{\Delta}_\ell \mathbf{X}_{/i,\mathcal{B}^-}. \tag{129}$$

Define $\boldsymbol{\delta}_B = \tilde{\mathbf{B}} - \check{\mathbf{B}}$ and $\boldsymbol{\delta}_{A^{-1}} = \tilde{\mathbf{A}}^{-1} - \check{\mathbf{A}}^{-1}$. Then, we have

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top \tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} \\
&= \mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top \boldsymbol{\delta}_{A^{-1}}\mathbf{x}_{i,\mathcal{B}^+} + \mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\boldsymbol{\delta}_B^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} \\
&\quad + \mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\boldsymbol{\delta}_B\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} + \mathbf{x}_{i,\mathcal{B}^+}^\top \boldsymbol{\delta}_{A^{-1}}\check{\mathbf{B}}\check{\mathbf{B}}_2^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} \\
&\quad + \mathbf{x}_{i,\mathcal{B}^+}^\top \check{\mathbf{A}}^{-1}\check{\mathbf{B}}\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}.
\end{aligned} \tag{130}$$

Our goal is to show that each of the terms in (130) are small for large values of n, p. The two terms $\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}_2^\top \boldsymbol{\delta}_{A^{-1}}\mathbf{x}_{i,\mathcal{B}^+}$ and $\mathbf{x}_{i,\mathcal{B}^+}^\top \boldsymbol{\delta}_{A^{-1}}\check{\mathbf{B}}\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$ can be bounded in very similar ways and will have similar upper bounds. Also, the two terms $\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\boldsymbol{\delta}_B^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$ and $\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\boldsymbol{\delta}_B\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$ can be handled in similar fashion and will have similar upper bounds. Hence, we only study the following three terms:

1. Finding an upper bound for $\mathbf{x}_{i,\mathcal{B}^+}^\top \check{\mathbf{A}}^{-1}\check{\mathbf{B}}\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$:
   Note that if it were not for the dependance of $\mathcal{B}^+$ on $\mathbf{x}_i$ we could claim that $\mathbf{x}_{i,\mathcal{B}^+}$ is independent of $\check{\mathbf{A}}^{-1}\check{\mathbf{B}}\check{\mathbf{B}}^\top \check{\mathbf{A}}^{-1}$, and we could use the Hanson-Wright inequality (Lemma 15). So, in order to remove the dependence to be able to use the Hanson-Wright inequality, we will use union bound on sets $\mathcal{B}^+$ and $\mathcal{B}^-$ in the following way. First note that $|\mathcal{B}^+| \geq p - d_n$ and $|\mathcal{B}^-| \leq d_n$. Let $\mathcal{T}^+$ and $\mathcal{T}^-$ denote two fixed sets of size larger than $p - d_n$, and smaller than d_n (not dependent on $\mathbf{x}_i$) respectively. We define

$$\begin{aligned}
\check{\check{\mathbf{A}}} &:= \text{diag}\left[\lambda\left\{\ddot{r}_\alpha(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right\}\right] + \mathbf{X}_{/i,\mathcal{T}^+}^\top \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right]\mathbf{X}_{/i,\mathcal{T}^+}, \\
\check{\check{\mathbf{B}}} &= \mathbf{X}_{/i,\mathcal{T}^+}^\top \text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}_{/i}^\alpha)\right]\mathbf{X}_{/i,\mathcal{T}^-}.
\end{aligned} \tag{131}$$

   For fixed $\mathcal{T}^+$ and $\mathcal{T}^-$, $\mathbf{x}_{i,\mathcal{T}}$ is independent of $\check{\check{\mathbf{A}}}$ and $\check{\check{\mathbf{B}}}$. Therefore if

$$\mathbf{G} = \check{\check{\mathbf{A}}}^{-1}\check{\check{\mathbf{B}}}\check{\check{\mathbf{B}}}^\top \check{\check{\mathbf{A}}}^{-1},$$

   then from the Hanson-Wright inequality, Lemma 15, we have:

$$\begin{aligned}
&\mathbb{P}\left(|\mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} - \mathbb{E}(\mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+}|\ \mathbf{X}_{/i}, \mathbf{y}_{/i})| > t \mid \mathbf{X}_{/i}, \mathbf{y}_{/i}\right) \\
&\leq 2\text{e}^{-c\left(\frac{(p-d_n)^2 t^2}{\|\mathbf{G}\|_{HS}^2} \wedge \frac{(p-d_n)t}{\|\mathbf{G}\|_2}\right)}.
\end{aligned} \tag{132}$$

   To use this bound we have to calculate the three terms: $\|\mathbf{G}\|_2$, $\|\mathbf{G}\|_{HS}$, and $\mathbb{E}(\mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+}|\ \mathbf{X}_{/i}, \mathbf{y}_{/i})$:

(a) $\|\mathbf{G}\|_2$:
It is straightforward to see that (see e.g. the derivation of (71))

$$\|\mathbf{G}\|_2 \leq \frac{\|\check{\check{\mathbf{B}}}\|^2}{\sigma^2_{\min}(\check{\check{\mathbf{A}}})} \leq \frac{\omega^s}{4\lambda^2\eta^2}, \tag{133}$$

under event $\mathcal{A}_L$.

(b) $\|\mathbf{G}\|^2_{HS}$:
The rank of matrix $\mathcal{G}$ is at most d_n (the maximum size of $\mathcal{T}^-$). Hence,

$$\|\mathbf{G}\|^2_{HS} \leq d_n\|\mathbf{G}\|^2_2 = \frac{d_n(\omega^s)^2}{16\lambda^4\eta^4} \tag{134}$$

(c) $\mathbb{E}(\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+})$:
Let $\mathbf{\Sigma}_\mathcal{T}$ the covariance matrix of $\mathbf{x}_{i,\mathcal{T}}$. We have

$$\begin{aligned}
|\mathbb{E}\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+}| &= \left|\mathrm{Tr}(\mathbf{\Sigma}^{1/2}_\mathcal{T}\mathbf{G}\mathbf{\Sigma}^{1/2}_\mathcal{T})\right| \\
&\overset{(a)}{\leq} d_n\|\mathbf{\Sigma}^{1/2}_\mathcal{T}\mathbf{G}\mathbf{\Sigma}^{1/2}_\mathcal{T}\| \leq d_n\|\mathbf{\Sigma}_\mathcal{T}\|\|\mathbf{G}\| \\
&\leq d_n\rho_{\max}\|\mathbf{G}\| \leq \frac{d_n\rho_{\max}\omega^s}{4\lambda^2\eta^2},
\end{aligned} \tag{135}$$

with probability $1-q_n$. Here, to obtain Inequality (a), we used the fact that the rank of $\mathbf{\Sigma}^{1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}$ is less than or equal to the rank of $\mathbf{G}$ which is less than or equal to d_n.

Furthermore, from Lemma 19 we have

$$\mathbb{P}(\sigma_{\max}(\mathbf{X}^\top\mathbf{X}) \geq (\sqrt{n}+3\sqrt{p})^2\rho_{\max}) \leq \mathrm{e}^{-p}, \tag{136}$$

where $\rho_{\max} = \sigma_{\max}(\Sigma)$. Let the event $\mathcal{E}$ denote the event $\sigma_{\max}(\mathbf{X}^\top\mathbf{X}) < (\sqrt{n}+3\sqrt{p})^2\rho_{\max}$. We know that

$$\mathbb{P}(\mathcal{E}) > 1-\mathrm{e}^{-p}.$$

We remind the reader that

$$\omega^s = \|\mathbf{X}^\top\mathbf{X}\| \sup_{t\in[0,1]}\max_{1\leq i\leq n}\ddot{\ell}_i(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}.$$

Combining (132), (133), (134), (135), (136) and using the following definitions:

$$\gamma_3 \quad := \quad \frac{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}{4\eta^2\lambda^2} \tag{137}$$

we have

$$\begin{aligned}
&\mathbb{P}(\mathbf{x}^\top_{i,\mathcal{T}}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + \mathbb{E}\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+}) \\
&\leq \mathbb{P}(\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + \mathbb{E}\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+}, \mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \\
&\leq 2\mathrm{e}^{-c\left(\frac{(p-d_n))^2t^2}{d_n\gamma_3^2}\wedge\frac{(p-d_n)t}{\gamma_3}\right)} + \mathrm{e}^{-p}.
\end{aligned} \tag{138}$$

Hence, we have

$$\mathbb{P}(\mathbf{x}^\top_{i,\mathcal{T}^+}\mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + d_n\gamma_3) \leq 2\mathrm{e}^{-c\left(\frac{(p-d_n)^2t^2}{d_n\gamma_3^2}\wedge\frac{(p-d_n)t}{\gamma_3}\right)} + \mathrm{e}^{-p}. \tag{139}$$

If we define the event $\tilde{\mathcal{E}}$ as the event that (3) holds, then we have

$$\begin{aligned}
&\mathbb{P}\Big(\max_{\substack{\mathcal{T}^+\\ |\mathcal{T}^+|\geq p-d_n}}\ \max_{\substack{\mathcal{T}^-\\ |\mathcal{T}^-|\leq d_n}} \mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + d_n\gamma_3\Big)\\
&\leq \mathbb{P}\Big(\max_{\substack{\mathcal{T}^+\\ |\mathcal{T}^+|\geq p-d_n}}\ \max_{\substack{\mathcal{T}^-\\ |\mathcal{T}^-|\leq d_n}} \mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + d_n\gamma_3|\mathcal{E},\tilde{\mathcal{E}}\Big) + \mathbb{P}(\mathcal{E}) + \mathbb{P}(\tilde{\mathcal{E}})\\
&\leq \sum_{\substack{\mathcal{T}^+\\ |\mathcal{T}^+|\geq p-d_n}}\ \sum_{\substack{\mathcal{T}^-\\ |\mathcal{T}^-|\leq d_n}} \mathbb{P}(\mathbf{x}_{i,\mathcal{T}^+}^\top \mathbf{G}\mathbf{x}_{i,\mathcal{T}^+} > t + d_n\gamma_3|\mathcal{E},\tilde{\mathcal{E}}) + \mathbb{P}(\mathcal{E}) + \mathbb{P}(\tilde{\mathcal{E}})\\
&\leq 2d_n\binom{p}{d_n}^2\left(2\mathrm{e}^{-c\left(\frac{(p-d_n)^2t^2}{d_n\gamma_3^2}\wedge\frac{(p-d_n)t}{\gamma_3}\right)}\right) + \mathrm{e}^{-p} + q_n\\
&\overset{(a)}{\leq} \mathrm{e}^{4d_n\log\frac{ep}{d_n}}\left(2\mathrm{e}^{-c\left(\frac{(p-d_n)^2t^2}{d_n\gamma_3^2}\wedge\frac{(p-d_n)t}{\gamma_3}\right)}\right) + \mathrm{e}^{-p} + q_n,
\end{aligned} \tag{140}$$

where to obtain inequality (a) we used Stirling approximation together with the assumption $d_n < \mathrm{e}^{d_n\log(ep/d_n)}$. By setting $t = \frac{d_n\log^2 d_n}{p-d_n}$ we can obtain

$$\mathbb{P}(\mathbf{x}_{i,\mathcal{B}^+}^\top \check{\mathbf{A}}^{-1}\check{\mathbf{B}}\check{\mathbf{B}}^\top\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} > \frac{d_n\log^2 p}{p-d_n} + d_n\gamma_3) \leq 2\mathrm{e}^{d_n\log\frac{ep}{d_n} - c\left(\frac{d_n\log^4 p}{\gamma_3^2}\wedge\frac{d_n\log^2 p}{\gamma_3}\right)} + q_n + \mathrm{e}^{-p}. \tag{141}$$

The probability in (141) is small for large values of d_n and p.

2. Finding an upper bound for $\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\boldsymbol{\delta}_{A^{-1}}\mathbf{x}_{i,\mathcal{B}^+}$:

Define

$$\check{\mathbf{v}} := \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}, \tag{142}$$

and

$$\begin{aligned}
\mathbf{u}^\top &:= -\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top\\
\tilde{\mathbf{u}}^\top &:= \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top\boldsymbol{\Delta}_\ell\mathbf{X}_{/i,\mathcal{B}^+}\tilde{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top
\end{aligned} \tag{143}$$

Using Lemma 12 we have

$$\begin{aligned}
&|\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}_2^\top\boldsymbol{\delta}_{A^{-1}}\mathbf{x}_{i,\mathcal{B}^+}|\\
&\overset{(a)}{\leq} |\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top\boldsymbol{\Delta}_\ell\mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}|\\
&\quad + |\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\check{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top\boldsymbol{\Delta}_\ell\mathbf{X}_{/i,\mathcal{B}^+}\tilde{\mathbf{A}}^{-1}\mathbf{X}_{/i,\mathcal{B}^+}^\top\boldsymbol{\Delta}_\ell\mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}|\\
&= |\mathbf{u}^\top\boldsymbol{\Delta}_\ell\check{\mathbf{v}}| + |\tilde{\mathbf{u}}^\top\boldsymbol{\Delta}_\ell\check{\mathbf{v}}|\\
&\overset{(b)}{\leq} \sqrt{\sum_i\boldsymbol{\Delta}_{\ell,ii}^2}\sqrt{\sum_i\mathbf{u}_i^2\check{\mathbf{v}}_i^2} + \sqrt{\sum_i\Delta_{\ell,ii}^2}\sqrt{\sum_i\tilde{\mathbf{u}}_i^2\check{\mathbf{v}}_i^2}\\
&\overset{(c)}{\leq} \sqrt{\sum_i\boldsymbol{\Delta}_{\ell,ii}^2}(\sum_i\mathbf{u}_i^4)^{\frac{1}{4}}(\sum_i\check{\mathbf{v}}_i^4)^{\frac{1}{4}} + \sqrt{\sum_i\Delta_{\ell,ii}^2}(\sum_i\tilde{\mathbf{u}}_i^4)^{\frac{1}{4}}(\sum_i\check{\mathbf{v}}_i^4)^{\frac{1}{4}}\\
&\leq \sqrt{\sum_i\boldsymbol{\Delta}_{\ell,ii}^2}(\sum_i\mathbf{u}_i^2)^{\frac{1}{2}}(\max_i|\check{\mathbf{v}}_i|)^{\frac{1}{2}}(\sum_i\check{\mathbf{v}}_i^2)^{\frac{1}{4}} + \sqrt{\sum_i\boldsymbol{\Delta}_{\ell,ii}^2}(\sum_i\tilde{\mathbf{u}}_i^2)^{\frac{1}{2}}(\max_i|\check{\mathbf{v}}_i|)^{\frac{1}{2}}(\sum_i\check{\mathbf{v}}_i^2)^{\frac{1}{4}}.
\end{aligned} \tag{144}$$

Note that since $\sigma_{\min}(\tilde{\mathbf{A}}) > \frac{1}{2\lambda\eta}$ and $\sigma_{\min}(\check{\mathbf{A}}) > \frac{1}{2\lambda\eta}$ we have

$$\begin{aligned}
\|\mathbf{u}\|_2 &\leq \|\mathbf{x}_i\|_2\frac{\|\tilde{\mathbf{B}}\|^2\|\mathbf{X}\|}{(2\lambda\eta)^2},\\
\|\tilde{\mathbf{u}}\|_2 &\leq \|\mathbf{x}_i\|_2\frac{\|\tilde{\mathbf{B}}\|^2\|\mathbf{X}\|^2\|\boldsymbol{\Delta}_\ell\|}{2(\lambda\eta)^3}.
\end{aligned} \tag{145}$$

Furthermore,

$$\|\boldsymbol{\Delta}_\ell\| = \sqrt{\sum_j \Delta_{\ell,ii}^2} \overset{(a)}{\leq} \text{PolyLog}(n)\|\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\|_2$$

$$\overset{(b)}{\leq} \text{PolyLog}(n)\frac{|\dot{\ell}(\widehat{\boldsymbol{\beta}}^\alpha)|\|\mathbf{x}_i\|_2}{2\lambda\eta} \tag{146}$$

$$\leq \frac{\text{PolyLog}(n)\|\mathbf{x}_i\|}{2\eta\lambda} \tag{147}$$

under the event $\mathcal{A}_L \cap \mathcal{A}_s$. Here Inequality (a) is a result of (3) in Assumption A4. Note that we are using the same notation for all the terms that are polynomial functions of $\log(n)$. To obtain Inequality (b) we have used Part 2 of Lemma 3. Finally,

$$\|\check{\mathbf{v}}\|_2 = \|\mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}\| \leq \frac{\|\mathbf{X}\|\|\mathbf{x}_i\|_2}{\sigma_{\min}(\check{\mathbf{A}})} \leq \frac{\|\mathbf{X}\|\|\mathbf{x}_i\|_2}{2\lambda\eta}. \tag{148}$$

To finish our bound for (144) we have to bound $\|\check{\mathbf{v}}\|_\infty$. Note that

$$\check{\mathbf{v}} = \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}.$$

The main difficulty in bounding this term is that due to the dependence of $\mathcal{B}^+$ on $\mathbf{x}_i$ it is hard to characterize the distribution of the elements of $\check{\mathbf{v}}$. To remove this dependency we again want to use the union bound on the different choices of $\mathcal{B}^+$. Towards this goal, suppose that for a fixed set $\mathcal{T}$ of size larger than $p - d_n$ we define:

$$\check{\check{\mathbf{v}}} = \mathbf{X}_{/i,\mathcal{T}}\check{\check{\mathbf{A}}}^{-1}\mathbf{x}_{i,\mathcal{T}},$$

where

$$\check{\check{\mathbf{A}}} = \text{diag}\left[\lambda\left\{\ddot{r}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right\}\right] + \mathbf{X}^\top_{/i,\mathcal{T}}\text{diag}\left[\ddot{\ell}_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right]\mathbf{X}_{/i,\mathcal{T}}. \tag{149}$$

Note that the distribution of $\check{\check{\mathbf{v}}}$ given $\mathbf{X}_{/i}, \mathbf{y}_{/i}$ is $N\left(0, \frac{1}{n}\mathbf{X}_{/i,\mathcal{T}}\check{\check{\mathbf{A}}}^{-2}\mathbf{X}^\top_{/i,\mathcal{T}}\right)$. Furthermore, we have

$$\|\mathbf{X}_{/i,\mathcal{T}}\check{\check{\mathbf{A}}}^{-2}\mathbf{X}^\top_{/i,\mathcal{T}}\| \leq \frac{\|\mathbf{X}_{/i}\mathbf{X}^\top_{/i}\|}{4\eta^2\lambda^2}.$$

Hence, using Corollary 5 we have

$$\mathbb{P}\left(\max_i \check{\check{\mathbf{v}}}_i > \sqrt{\frac{\|\mathbf{X}_{/i}\mathbf{X}^\top_{/i}\|}{n\lambda\eta}\log 2n} + t \mid \mathbf{X}_{/i}, \mathbf{y}_{/i}\right) \leq 2\text{e}^{-\frac{4n\lambda^2\eta^2t^2}{\|\mathbf{X}_{/i}\mathbf{X}^\top_{/i}\|}}. \tag{150}$$

It is straightforward to check that

$$\max_i \|\mathbf{X}_{/i}\mathbf{X}^\top_{/i}\| \leq \|\mathbf{X}\mathbf{X}^\top\|.$$

Furthermore, according to Lemma 19 with probability larger than e^{-p} we have

$$\mathbb{P}(\|\mathbf{X}^\top\mathbf{X}\| \geq (\sqrt{n} + 3\sqrt{p})^2\rho_{\max}) \leq \text{e}^{-p}. \tag{151}$$

Define

$$\gamma_4(n) := \sqrt{\frac{(\sqrt{n} + 3\sqrt{p})^2\rho_{\max}}{n\lambda\eta}\log 2n}, \tag{152}$$

and let the event $\mathcal{E}$ denote the event that $\|\mathbf{X}^\top\mathbf{X}\| \leq (\sqrt{n} + 3\sqrt{p})^2\rho_{\max}$. Then, from (151) and (150) we obtain

$$\mathbb{P}\left(\max_i \check{\check{\mathbf{v}}}_i > \gamma_4(n) + t\right)$$

$$\leq \mathbb{P}\left(\max_i \check{\check{\mathbf{v}}}_i > \gamma_4(n) + t, \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)$$

$$\leq \text{e}^{-p} + 2\text{e}^{-\frac{4n\lambda^2\eta^2t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}. \tag{153}$$

Note that we expect $\rho_{\max} = O(\frac{1}{p})$. Hence, $\gamma_4(n)$ is expected to be $O\left(\sqrt{\frac{\log n}{n}}\right)$. So far, we have only considered a bound for $\check{\check{\mathbf{v}}}$ in which a fixed set $\mathcal{T}$ is only considered, despite the fact that the quantity we are interested in is $\|\check{\mathbf{v}}\|_\infty$ in which $\mathcal{T}$ is replaced with set $\mathcal{B}^+$. To resolve the issue we use the union bound. We have

$$\begin{aligned}
&\mathbb{P}\left(\max_{\mathcal{T}:|\mathcal{T}|>p-d_n} \max_i \check{\check{\mathbf{v}}}_i > \gamma_4(n)+t\right)\\
&\le \sum_{\mathcal{T}:|\mathcal{T}|>p-d_n} \mathbb{P}\left(\max_i \check{\check{\mathbf{v}}}_i > \gamma_4(n)+t\right)\\
&\le d_n \binom{p}{p-d_n}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{n\lambda^2\eta^2t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right)\\
&\le \mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{4n\lambda^2\eta^2t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right).
\end{aligned} \tag{154}$$

Hence, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\max_i \check{\mathbf{v}}_i > \gamma_4(n)+t\right)\\
&\le \mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{n\lambda^2\eta^2t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right).
\end{aligned} \tag{155}$$

Setting $t = \frac{1}{\lambda\eta}\sqrt{\frac{d_n\log^2 p}{n}}$ and combining this equation with (144), (145), (146), and (148), we conclude that if we define

$$\gamma_5(n) := \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|_2}{2\lambda\eta}\left(\gamma_4(n)+\frac{1}{\lambda\eta}\sqrt{\frac{d_n\log^2 p}{n}}\right)\frac{\|\mathbf{X}\|\|\mathbf{x}_i\|}{2\lambda\eta}\frac{\|\mathbf{x}_i\|\|\tilde{\mathbf{B}}\|^2\|\mathbf{X}\|}{4\lambda^2\eta^2}\left(1+\frac{\|\mathbf{X}\|c_3(n)c_4(n)}{4\lambda^2\eta^2}\right), \tag{156}$$

then

$$\mathbb{P}(|\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\boldsymbol{\delta}_{A^{-1}}\mathbf{x}_{i,\mathcal{B}^+}| > \gamma_6(n)) \le q_n+\check{q}_n+\mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p}+2\mathrm{e}^{-\frac{d_n\log^2 p}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right). \tag{157}$$

As discussed before, we have $\gamma_4(n) = O\left(\sqrt{\frac{\log n}{n}}\right)$. Also, all the norms $\|\mathbf{X}\|, \|\mathbf{x}_i\|, \|\tilde{\mathbf{B}}\|$ are $O_P(1)$, hence we expect $\gamma_5(n)$ to go to zero at the rate $\sqrt{d_n\mathrm{PolyLog}(n)/n}$. This heuristic argument will be be made rigorous later.

3. $\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\boldsymbol{\delta}_B^\top\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}$:
Using a similar argument as the one presented in (144) we obtain:

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\boldsymbol{\delta}_B^\top\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}\\
&= \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\mathbf{X}_{/i,\mathcal{B}^-}\boldsymbol{\Delta}_\ell\mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}\\
&= \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\mathbf{X}_{/i,\mathcal{B}^-}\boldsymbol{\Delta}_\ell\check{\mathbf{v}} = \check{\mathbf{u}}\boldsymbol{\Delta}_\ell\check{\mathbf{v}}\\
&\le \sqrt{\sum_i \boldsymbol{\Delta}_{\ell,ii}^2}\|\check{\mathbf{u}}\|(\max_i|\check{\mathbf{v}}_i|)^{\frac{1}{2}}\|\check{\mathbf{v}}\|^{\frac{1}{2}}.
\end{aligned} \tag{158}$$

In the above equations we have used

$$\check{\mathbf{u}}^\top := \mathbf{x}_{i,\mathcal{B}^+}^\top\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\mathbf{X}_{/i,\mathcal{B}^-},$$

and

$$\check{\mathbf{v}} = \mathbf{X}_{/i,\mathcal{B}^+}\check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}.$$

According to (146) and (155) we have

$$\sqrt{\sum_j \Delta_{\ell,ii}^2} \leq \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|_2}{2\lambda\eta}, \tag{159}$$

with probability larger than $1 - q_n - \check{q}_n$, and

$$\|\check{\mathbf{v}}\|_2 \leq \frac{\|\mathbf{X}\|\|\mathbf{x}_i\|_2}{2\lambda\eta}, \tag{160}$$

and

$$\mathbb{P}\left(\max_i \check{\mathbf{v}}_i > \gamma_5(n) + t\right) \leq \mathrm{e}^{2d_n \log \frac{ep}{d_n}} \left(\mathrm{e}^{-p} + 2\mathrm{e}^{-\frac{4n\lambda^2\eta^2 t^2}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right). \tag{161}$$

Hence, the only remaining term to bound is $\|\check{\mathbf{u}}\|$. It is straightforward to see that

$$\|\check{\mathbf{u}}\|_2 \leq \frac{\|\mathbf{x}_i\|\|\tilde{\mathbf{B}}\|\|\mathbf{X}\|}{\sigma_{\min}(\tilde{\mathbf{A}})} \leq \frac{\|\mathbf{x}_i\|\|\tilde{\mathbf{B}}\|\|\mathbf{X}\|}{2\lambda\eta}. \tag{162}$$

Hence, combining (158), (159), (160), (161), and (162) we will have that if we set $t = \frac{1}{\lambda\eta}\sqrt{\frac{d_n \log^2 p}{n}}$, and define

$$\gamma_6(n) := \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|^3\|\tilde{\mathbf{B}}\|\|\mathbf{X}\|^2}{(2\lambda\eta)^3}\left(\gamma_5(n) + \frac{1}{\lambda\eta}\sqrt{\frac{d_n \log^2 p}{n}}\right), \tag{163}$$

then

$$\mathbb{P}(|\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\boldsymbol{\delta}_B^\top \check{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+}| > \gamma_6(n)) \leq q_n + \check{q}_n + \mathrm{e}^{2s\log\frac{ep}{s}}\left(\mathrm{e}^{-p} + 2\mathrm{e}^{-\frac{4d_n\log^2 p}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right). \tag{164}$$

Plugging in the bounds from equations (141),(164) and similarly bounding the remaining terms in (130), we obtain

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}^+}^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{B}}_2^\top \tilde{\mathbf{A}}^{-1}\mathbf{x}_{i,\mathcal{B}^+} \\
&\leq 2\gamma_5(n) + 2\gamma_6(n) + \frac{d_n\log^2 p}{p-d_n} + d_n\gamma_3 \\
&\leq \frac{d_n\log^2 p}{p-d_n} + \frac{(\sqrt{n}+3\sqrt{p})^2 d_n\rho_{\max}^2}{4\eta^2\lambda^2} \\
&\quad + \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|^3\|\mathbf{X}\|^2\omega^s}{4\lambda^3\eta^3}\left(1 + \frac{\omega^s}{2\lambda\eta} + \frac{\omega^s\|\mathbf{X}\|\mathrm{PolyLog}(n)}{8\lambda^3\eta^3}\right)\left(\sqrt{\frac{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}{n\lambda\eta}\log 2n} + \frac{1}{\lambda\eta}\sqrt{\frac{d_n\log^2 p}{n}}\right).
\end{aligned}$$

with probability at least

$$\begin{aligned}
&1 - \mathrm{e}^{d_n\log\frac{ep}{d_n}}\left(2\mathrm{e}^{-c\left(\frac{d_n\log^4 p}{\gamma_3^2}\wedge\frac{d_n\log^2 p}{\gamma_3}\right)}\right) - q_n - \check{q}_n - \mathrm{e}^{-p} - 2\mathrm{e}^{2d_n\log\frac{ep}{d_n}}\left(\mathrm{e}^{-p} + 4\mathrm{e}^{-\frac{d_n\log^2 p}{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}}\right) \\
&\geq 1 - \mathrm{e}^{-d_n\log p} - q_n - \check{q}_n
\end{aligned}$$

provided $\mathrm{e} \leq d_n \leq \frac{p}{C}$ for a sufficiently large constant $C > 0$, where we use the definition of γ_3, γ_4, γ_5, γ_6 along with the fact that $\rho_{\max} = O(p^{-1})$.

We now return to equations (116) and (127), along with the bound derived immediately above, to obtain

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top \mathbf{A}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} \\
&\leq \|\mathbf{x}_i\|^2(\omega^s)^2\left(\frac{\gamma_0^s(\alpha)}{\lambda\eta} + (\gamma_0^s(\alpha))^2 + (\gamma_1^s(\alpha))^2\right) \\
&\quad + \frac{d_n \log^2 p}{p-d_n} + \frac{(\sqrt{n}+3\sqrt{p})^2 d_n \rho_{\max}^2}{4\eta^2\lambda^2} \\
&\quad + \frac{\mathrm{PolyLog}(n)\|\mathbf{x}_i\|^3\|\mathbf{X}\|^2\omega^s}{4\lambda^3\eta^3}\left(1+\frac{\omega^s}{2\lambda\eta}+\frac{\omega^s\|\mathbf{X}\|\mathrm{PolyLog}(n)}{8\lambda^3\eta^3}\right)\left(\sqrt{\frac{(\sqrt{n}+3\sqrt{p})^2\rho_{\max}}{n\lambda\eta}\log 2n} + \frac{1}{\lambda\eta}\sqrt{\frac{d_n\log^2 p}{n}}\right)
\end{aligned} \tag{165}$$

with probability at least $1 - p^{-d_n} - q_n - \check{q}_n$. We now simplify the above bound by plugging in high probability bounds on the respective parameters. To this end, we first state:

- By Assumption A1, $\rho_{\max} = \sigma_{\max}(\Sigma) \leq C_X/p$.
- By Lemmas 17 and 19,
$$\max_{1\leq i\leq n}\|\mathbf{x}_i\| \leq 2\sqrt{C_X} \text{ and } \|\mathbf{X}^\top\mathbf{X}\| \leq (\sqrt{\gamma_0}+3)^2 C_X$$
with probability at least $1-(n+1)\mathrm{e}^{-p/2}$.
- Next,
$$\begin{aligned}
\omega_s &= \|\mathbf{X}^\top\mathbf{X}\| \sup_{t\in[0,1]}\max_{1\leq i\leq n}\ddot{\ell}(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i}) \\
&\leq (\sqrt{n}+3\sqrt{p})^2\rho_{\max}\mathrm{PolyLog}(n) \\
&\leq C_X(\sqrt{\gamma_0}+3)^2\mathrm{PolyLog}(n)
\end{aligned}$$
with probability at least $1-\mathrm{e}^{-p}$, by Lemma 19 and Assumption A4.
- By equation (115), we have
$$\gamma_0^s(\alpha) \leq \frac{1}{p}$$
provided $\alpha \geq Cp/\lambda^2\eta(1-\eta)\kappa_0$ for sufficiently large p and a sufficiently large constant $C>0$.
- By equation (126) and the above inequalities, we obtain
$$\gamma_1^s(\alpha) \leq \frac{16}{\lambda\alpha(1-\eta)\kappa_0}\left(1+\frac{1}{\lambda\eta}(C_X)^2(\sqrt{\gamma_0}+3)^4\mathrm{PolyLog}(n)\right) \leq \frac{1}{p}$$
provided $\alpha \geq Cp\mathrm{PolyLog}(n)/\lambda^2\eta(1-\eta)\kappa_0$ for sufficiently large p and a sufficiently large constant $C>0$.

Plugging in all these bounds into (165) we conclude that there is a sufficiently large numerical constant $C>0$, depending only on γ_0, C_X and such that

$$\begin{aligned}
&\mathbf{x}_{i,\mathcal{B}_{1,+}}^\top \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top \mathbf{A}^{-1}\mathbf{x}_{i,\mathcal{B}_{1,+}} \\
&\leq \frac{\mathrm{PolyLog}(n)}{\lambda^3\eta^3(1\wedge\lambda\eta)^3}\sqrt{\frac{d_n}{n\lambda\eta}} + \frac{Cd_n}{n\lambda^2\eta^2} + \frac{d_n\log^2 p}{p-d_n} + \frac{C}{p\lambda\eta}
\end{aligned}$$

with probability at least $1-(n+1)\mathrm{e}^{-\frac{p}{2}} - p^{-d_n} - q_n - \check{q}_n$. □

6 STUDY OF THE ELASTIC NET ESTIMATOR

6.1 Objective

As mentioned in Section 3 of the main part of the paper, our proofs use concentration of measure results for the empirical distribution of the regression coefficeints and the subgradient vector. Some of the results we use in our paper are due to (Miolane and Montanari, 2021). However, the results of (Miolane and Montanari, 2021) are stated for the LASSO estimator and not the elastic-net. Hence, the results we require are different from those presented in (Miolane and Montanari, 2021). However, the changes do not constitute significant advancements that would warrant the derivation of elastic-net results as a major contribution. As a result, we have included these findings in a dedicated section, which will serve as an online appendix to our paper. This section will not be part of the formal submission to a journal but is included for the sake of completeness.

Throughout this appendix we will mainly focus on Theorem 3.1, Theorem E.5 and Theorem F.1 of (Miolane and Montanari, 2021). For the sake of brevity we do not present the proof with every details, since the proof technique is very similar to that of (Miolane and Montanari, 2021). We only focus on the differences. Consider the elastic net problem:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{n}\left(\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^*\|_1\right) + \frac{\eta}{n}\left(\|\boldsymbol{\beta}\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2\right).$$

Note that to simplify the proof we have subtracted $\|\boldsymbol{\beta}^*\|_1$ and $\|\boldsymbol{\beta}^*\|_2^2$ and used a different scaling than the one presented in the paper. However, as is obvious, these changes do not have any effect on $\widehat{\boldsymbol{\beta}}$. Furthermore, the definition here is slightly different from the one in the main paper, where the LASSO penalty is $\lambda(1-\eta)$, and the ridge penalty is $\lambda\eta$. The difference is not substantial and is only for notational brevity. Let $\widehat{\mathbf{w}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ denote the estimation error. Using the assumption $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\mathbf{z}$, the problem can be written as:

$$\begin{aligned}\widehat{\mathbf{w}} &= \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathcal{C}(\mathbf{w}) \\ &= \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\|\sigma\mathbf{z}-\mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^*\|_1\right) + \frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2\right). \end{aligned} \tag{166}$$

We assume there exist $0 < \lambda_{\min} < \lambda_{\max}$, $\eta_{\max} > 0$ such that $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and $\eta \in (0, \eta_{\max}]$.

Suppose that we are interested in the asymptotic distribution of the elements of $\widehat{\mathbf{w}}$. The main approach that can help us in characterizing the distribution is the convex Gaussian minimax theorem that we would like to introduce briefly next.

1.2 Convex Gaussian Minimax Theorem

In this appendix, we require a few notations that we aim to introduce in this section. Consider the mean square error of the elastic net, i.e. $\frac{1}{p}\|\widehat{\mathbf{w}}\|_2^2$. It has been shown in (Maleki, 2010; Donoho et al., 2011, 2009; Thrampoulidis et al., 2015, 2018) that under the asymptotic settings $n/p \to \delta$ and under Assumptions B1-B5 of our paper, the mean square error converges to $\delta(\tau_*^2 - \sigma^2)$, where τ_*^2 is a saddle point of the following function:

$$\begin{aligned}\psi(\tau, b) = &\left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{b^2}{2} + \frac{1}{n}\mathbb{E}\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{b}{2\tau}\|\mathbf{w}\|_2^2 - b\mathbf{g}^\top\mathbf{w} + \lambda\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^*\|_1\right)\right\} \\ &+ \eta\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2\right). \end{aligned} \tag{167}$$

Define

$$\widehat{\mathbf{w}}^f(\tau, b) := \frac{b}{b+2\eta\tau}\operatorname{soft}\left(\tau\mathbf{g}+\boldsymbol{\beta}^*, \frac{\lambda\tau}{b}\right) - \boldsymbol{\beta}^*$$

where $\mathbf{g} \sim N(0, I_p)$ and

$$[\operatorname{soft}(\mathbf{x}, r)]_i = ((|x_i| - r)_+ \operatorname{sgn}(x_i))_{i=1}^p$$

is the element-wise soft thresholding function. Let (τ_*, b_*) denote the unique saddle point of $\psi(\tau, b)$. Define

$$\widehat{\mathbf{w}}^f := \widehat{\mathbf{w}}^f(\tau^*, b^*)$$

The following theorems state that the saddle point exists, is unique, and is bounded.

Lemma 23. $\max_{b\geq 0}\min_{\tau\geq\sigma}\psi(\tau,b)$ *is achieved at a unique couple* (τ_*,b_*) *which is also the unique solution of the following system:*

$$\begin{cases}\tau^2=\sigma^2+\frac{1}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f\|_2^2,\\ b=\tau-\frac{1}{n}\mathbb{E}g^\top\widehat{\mathbf{w}}^f.\end{cases}\tag{168}$$

Lemma 24. *There exist* $b_{min}>0$, $\tau_{max}>0$, $b_{max}>0$ *that depend only on* δ,σ,ξ *such that* $b_{min}\leq b_*\leq b_{max}$ *and* $\sigma<\tau_*\leq\tau_{max}$

The proof of these theorems are postponed to Section 1.4.

The convex Gaussian minimax framework makes the connection between $\frac{1}{p}\|\widehat{\mathbf{w}}\|_2^2$ and τ_*^2 through a few steps that we clarify below. Again consider the optimization problem:

$$\begin{aligned}\widehat{\mathbf{w}}&=\underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\,\mathcal{C}(\mathbf{w})\\&=\underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\frac{1}{2n}\|\sigma\mathbf{z}-\mathbf{X}\mathbf{w}\|_2^2+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right).\end{aligned}$$

Note that we can rewite this optimization problem as the following saddle point problem using dual representation of the l_2 norm:

$$\widehat{\mathbf{w}}:=\arg\min_{\mathbf{w}\in\mathbb{R}^p}\max_{\mathbf{u}}\frac{1}{n}\mathbf{u}^\top\mathbf{X}\mathbf{w}-\frac{1}{2n}\mathbf{u}^\top\mathbf{u}-\frac{1}{n}\mathbf{u}^\top\sigma\mathbf{z}+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right).$$

According to the Convex Gaussian minimax framework we can construct a simpler auxilary saddle point problem that can provide useful information about $\widehat{\mathbf{w}}$. To clarify this point, define

$$\begin{aligned}\Phi(X)\quad:=\quad&\min_{\mathbf{w}\in\mathcal{S}_w}\max_{\mathbf{u}\in\mathcal{S}_u}\frac{1}{n}\mathbf{u}^\top\mathbf{X}\mathbf{w}-\frac{1}{2n}\mathbf{u}^\top\mathbf{u}-\frac{1}{n}\mathbf{u}^\top\sigma\mathbf{z}+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)\\&+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right)&(169)\\=\quad&\min_{\mathbf{w}\in\mathcal{S}_w}\max_{\mathbf{u}\in\mathcal{S}_u}\frac{1}{n^{3/2}}\mathbf{u}^\top(\tilde{\mathbf{X}},-\mathbf{z})\begin{pmatrix}\mathbf{w}\\-\sqrt{n}\sigma\end{pmatrix}-\frac{1}{2n}\mathbf{u}^\top\mathbf{u}+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)\\&+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right)&(170)\end{aligned}$$

where $\mathcal{S}_w$ and $\mathcal{S}_u$ are two convex, compact sets. Note that $(\tilde{\mathbf{X}},-\mathbf{z})$ is a matrix with i.i.d. $N(0,1)$ entries. Define the following auxillary optimization problem:

$$\begin{aligned}&\phi(\mathbf{g},\mathbf{h})\\:=\quad&\min_{\mathbf{w}\in\mathcal{S}_w}\max_{\mathbf{u}\in\mathcal{S}_u}\frac{1}{n}\sqrt{\frac{\|\mathbf{w}\|_2^2}{n}+\sigma^2}\mathbf{h}^\top\mathbf{u}-\frac{1}{n^{3/2}}\|\mathbf{u}\|\mathbf{g}^\top\mathbf{w}+\frac{1}{n}\|\mathbf{u}\|g'\sigma-\frac{1}{2n}\mathbf{u}^\top\mathbf{u}\\&+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right)&(171)\end{aligned}$$

where $\mathbf{h}\sim N(0,\mathbb{I}_n)$, $\mathbf{g}\sim N(0,\mathbb{I}_p)$, $g'\sim N(0,1)$, and all of them are independent and also independent of $\mathbf{z}$.

According to the Gaussian minimax theorem, i.e. Theorem 3 of (Thrampoulidis et al., 2015) , when $\mathcal{S}_w$ and $\mathcal{S}_u$ are convex compact sets we have

$$\mathbb{P}(\Phi(X)<t)\leq 2\mathbb{P}(\phi(\mathbf{g},\mathbf{h})<t)\tag{172}$$
$$\mathbb{P}(\Phi(X)>t)\leq 2\mathbb{P}(\phi(\mathbf{g},\mathbf{h})>t)\tag{173}$$

Using this theorem, and by using a proper choice for $\mathcal{S}_w$ we can analyze certain properties of $\widehat{\mathbf{w}}$ through the minimizer of the easier auxillary function:

$$\begin{aligned}L(\mathbf{w})\quad:=\quad&\frac{1}{2}\left(\sqrt{\frac{\|\mathbf{w}\|_2^2}{n}+\sigma^2}\frac{\|\mathbf{h}\|_2}{\sqrt{n}}-\frac{1}{n}\mathbf{g}^\top\mathbf{w}+\frac{g'\sigma}{\sqrt{n}}\right)_+^2+\frac{\lambda}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}^*\|_1\right)\\&+\frac{\eta}{n}\left(\|\mathbf{w}+\boldsymbol{\beta}^*\|_2^2-\|\boldsymbol{\beta}^*\|_2^2\right).&(174)\end{aligned}$$

Note that $L(\mathbf{w})$ is obtained from $\phi(\mathbf{g}, \mathbf{h})$, when the supremum with respect to $\mathbf{u} \in \mathbb{R}^p$ is calculated.

Let us explain how $L(\mathbf{w})$ can be connected with the scaler saddle point problem presented in (167). Actually $\min_{\mathbf{w}} L(\mathbf{w})$ concentrates around $\max_{b\geq 0} \min_{\tau \geq \sigma} \psi(\tau, b)$. Below is a heuristic argument. For simplicity we denote $r(\mathbf{w}) = \lambda(\|\mathbf{w} + \boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^*\|_1) + \eta(\|\mathbf{w} + \boldsymbol{\beta}^*\|_2 - \|\boldsymbol{\beta}^*\|_2)$

First notice that $\frac{\|\mathbf{h}\|_2}{\sqrt{n}}$ concentrates around 1, and $\frac{g'\sigma}{\sqrt{n}}$ concentrates around 0, so heuristically, they can be removed from the expression which lead to:

$$\min_{\mathbf{w}} L(\mathbf{w}) \approx \min_{\mathbf{w}} \frac{1}{2}\left(\sqrt{\frac{\|\mathbf{w}\|_2^2}{n} + \sigma^2} - \frac{1}{n}\mathbf{g}^\top \mathbf{w}\right)_+^2 + \frac{1}{n} r(\mathbf{w})$$

Next we use the fact that $a_+^2 = \max_{b\geq 0} ab - \frac{1}{2}b^2$ and $\sqrt{\frac{\|\mathbf{w}\|_2^2}{n} + \sigma^2} = \min_{\tau \geq \sigma} \frac{\frac{\|\mathbf{w}\|_2^2}{n} + \sigma^2}{2\tau} + \frac{\tau}{2}$ to obtain

$$\begin{aligned}
\min_{\mathbf{w}} L(\mathbf{w}) &\approx \min_{\mathbf{w}} \max_{b\geq 0} \min_{\tau \geq \sigma} \left(\frac{\frac{\|\mathbf{w}\|_2^2}{n} + \sigma^2}{2\tau} + \frac{\tau}{2} - \frac{1}{n}\mathbf{g}^\top \mathbf{w}\right) b - \frac{1}{2}b^2 + \frac{1}{n} r(\mathbf{w}) \\
&\overset{(a)}{=} \max_{b\geq 0} \min_{\tau \geq \sigma} \left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{1}{2}b^2 + \frac{1}{n} \min_{\mathbf{w}} \left\{\frac{b}{2\tau}\|\mathbf{w}\|_2^2 - \mathbf{g}^\top \mathbf{w} + r(\mathbf{w})\right\} \\
&:= \max_{b\geq 0} \min_{\tau \geq \sigma} F(\tau, b, \mathbf{w})
\end{aligned}$$

Step (a) requires some delicate arguments which are omitted here. They intend to prove that the minimum ad maximum operation are interchangeable. Finally, one can show that $\max_{b\geq 0} \min_{\tau \geq \sigma} F(\tau, b, \mathbf{w})$ concentrates around

$$\max_{b\geq 0} \min_{\tau \geq \sigma} \mathbb{E}F(\tau, b, \mathbf{w}) = \max_{b\geq 0} \min_{\tau \geq \sigma} \psi(\tau, b) = \psi(\tau^*, b^*).$$

Therefore, $\min_{\mathbf{w}} L(\mathbf{w})$ concentrates around $\psi(b_*, \tau_*)$.

Now we will use the above arguments for a specific choice of $\mathcal{S}_w$ that allows us to obtain the mean square error of $\widehat{\mathbf{w}}$. For this purpose we define $\mathcal{S}_w = \mathcal{D}(\varepsilon) := \{\mathbf{w} \in \mathbb{R}^p : \frac{1}{p}\|\mathbf{w} - \widehat{\mathbf{w}}^f\|_2^2 > \varepsilon\}$. We first mention the roadmap of the proof to help the readers navigate through the following theorems:

- We first show that the minimizer of $L(\mathbf{w})$ is with high probability in a ball of radius ϵ^2 around $\widehat{\mathbf{w}}^f$, and hence $\min_{\mathbf{w}} L(\mathbf{w})$ is with high probability very close to $\psi(b_*, \tau_*)$. This is done in Lemma 25, Lemma 26, and Corollary 7.
- In the next step we use the minimax theorem to show that the minimizer of $\mathcal{C}(\mathbf{w})$ defined in (166) is with high probability in the complement of $\mathcal{D}(\epsilon)$. This is proved in Lemmas 29 and 28. One can then use Lemma 28 to obtain information about $\frac{1}{p}\|\widehat{\mathbf{w}}\|_2^2$, which is the same as the mean square error of $\widehat{\boldsymbol{\beta}}$.

Our first result aims to connect the minimizer of $L(\mathbf{w})$ with $\widehat{\mathbf{w}}^f$.

Lemma 25. *There exist $\gamma, C, c > 0$ depending only on Ω, such that $\forall \varepsilon \in [0, 1]$,*

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{D}(\varepsilon)} L(\mathbf{w}) < \min_{\mathbf{w}\in\mathbb{R}^p} L(\mathbf{w}) + \gamma\varepsilon\right) \leq \frac{C}{\varepsilon} e^{-cn\varepsilon^2}$$

where $\mathcal{D}(\varepsilon) := \{\mathbf{w} \in \mathbb{R}^p : \frac{1}{p}\|\mathbf{w} - \widehat{\mathbf{w}}^f\|_2^2 > \varepsilon\}$. [§]

The proof is essentially the same as the proof of Theorem B.1 of (Miolane and Montanari, 2021), up to minor modifications to some constants. Hence, we do not repeat the proof here..

[§]Throughout the discussion, the term 'constants' refers to quantities that rely only on the following set of model parameters $\Omega := (\delta, \sigma, \xi, \lambda_{\min}, \lambda_{\max}, \eta_{\max})$. Recall that $\delta = n/p$, and $\xi = \frac{1}{\sqrt{p}}\|\boldsymbol{\beta}^*\|_2$; $\lambda_{\min}, \lambda_{\max}$ control the scale of λ, and $\eta_{\max}$ controls the scale of η. We do not assume η to be bounded away from 0, to be consistent with our main text.

Lemma 26. *There exist constants $C, c > 0$ such that for all $\epsilon \in [0,1]$,*

$$\mathbb{P}\left(|L(\widehat{\mathbf{w}}^f) - \psi(\tau_*, b_*)| > \varepsilon\right) \le Ce^{-cn\varepsilon^2}$$

Proof. We only need to prove it for $\varepsilon \le \varepsilon_0$ for some small constant ε_0 depending only on model parmeters Ω . The reason is that the probability is non-increasing in ε so we have a naive bound $Ce^{-cn\varepsilon_0^2}$ for $\varepsilon_0 \le \varepsilon \le 1$. This flat bound, combined with the sub-Gaussian bound for small ε, is further bounded by $Ce^{-cn\varepsilon_0^2\varepsilon^2}$ for all $\varepsilon \in [0,1]$ and this is the bound we desire.

Using the fix point equations (168) we have the following simplification for $\psi(\tau_*, b_*)$:

$$\psi(\tau_*, b_*) = \frac{1}{2}b_*^2 + \frac{\lambda}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1 + \frac{\eta}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2 - \frac{\eta}{n}\|\boldsymbol{\beta}^*\|_2^2$$

Define

$$\widehat{b}^f := \left(\sqrt{\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n} + \sigma^2}\frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \frac{1}{n}\mathbf{g}^\top\widehat{\mathbf{w}}^f + \frac{g'\sigma}{\sqrt{n}}\right)$$

And denote $\widehat{b}_+^f = \widehat{b}^f 1_{\widehat{b}^f > 0}$, we have

$$\begin{aligned}|L(\widehat{\mathbf{w}}^f) - \psi(\tau_*, b_*)| \le \frac{1}{2}&\left|(\widehat{b}_+^f)^2 - b_*^2\right| \\ &+ \frac{\lambda}{n}\left|\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1\right| \\ &+ \frac{\eta}{n}\left|\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2 - \mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2\right| \end{aligned} \tag{175}$$

The last two terms are relatively easy to bound. Notice that $\mathbf{g} \to \widehat{\mathbf{w}}^f = \frac{b_*}{b_* + 2\eta\tau_*}\text{soft}(\tau_*\mathbf{g} + \boldsymbol{\beta}^*, \frac{\lambda\tau_*}{b_*}) - \boldsymbol{\beta}^*$ is $\tau_{\max}$ Lipschitz, so

1. $\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1$ is Cn^{-1} sub-Gaussian, because $\mathbf{g} \mapsto \frac{1}{n}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1$ is $Cn^{-1/2}$ Lipschitz.

2. $\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2$ is Cn^{-1} sub-exponential because $\mathbf{g} \to n^{-1/2}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2$ is $Cn^{-1/2}$-Lipschitz. Therefore $n^{-1/2}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2$ is Cn^{-1}-sub-Gaussian, and hence $\frac{1}{n}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2$ is Cn^{-1} sub-exponential.

Therefore for the second and third terms of (175), we have the following concentrations. For $\varepsilon \in [0,1]$, there exist constants $C, c > 0$ such that

$$\mathbb{P}\left(\frac{\lambda}{n}\left|\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1\right| > \varepsilon\right) \le Ce^{-cn\varepsilon^2} \tag{176}$$

$$\mathbb{P}\left(\frac{\eta}{n}\left|\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2 - \mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_2^2\right| > \varepsilon\right) \le Ce^{-cn\varepsilon^2} \tag{177}$$

Now for the first term $(\widehat{b}_+^f)^2 - b_*^2$, first we have

$$\begin{aligned}&|\widehat{b}^f - b_*| \\ &= \left|\left(\sqrt{\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n} + \sigma^2} - \sqrt{\mathbb{E}\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n} + \sigma^2}\right)\frac{\|\mathbf{h}\|_2}{\sqrt{n}} + \tau_*\left(\frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1\right) - \frac{1}{n}(\mathbf{g}^\top\widehat{\mathbf{w}}^f - \mathbb{E}\mathbf{g}^\top\widehat{\mathbf{w}}^f) + \frac{\sigma g'}{\sqrt{n}}\right|.\end{aligned} \tag{178}$$

We aim to establish concentration results for each of the four terms that appear in (178). For the first term, we

have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n}+\sigma^2}-\sqrt{\mathbb{E}\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n}+\sigma^2}\right|\frac{\|\mathbf{h}\|_2}{\sqrt{n}}>\frac{\varepsilon}{4}\right)\\
&\leq \mathbb{P}(\|\mathbf{h}\|_2>2\sqrt{n})+\mathbb{P}\left(\left|\sqrt{\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n}+\sigma^2}-\sqrt{\mathbb{E}\frac{\|\widehat{\mathbf{w}}^f\|_2^2}{n}+\sigma^2}\right|>\frac{\varepsilon}{8}\right)\\
&\leq Ce^{-cn}+\mathbb{P}\left(\frac{1}{n}\left|\|\widehat{\mathbf{w}}^f\|_2^2-\mathbb{E}\|\widehat{\mathbf{w}}^f\|_2^2\right|>\frac{\varepsilon}{b}\cdot 2\sigma\right)\\
&\leq Ce^{-cn}+Ce^{-cn\varepsilon^2}
\end{aligned}$$

for small ε. The second inequality above uses sub-Gaussian concentration of $\|\mathbf{h}\|_2$, and the fact that $\sqrt{x+\sigma^2}$ is $\frac{1}{2\sigma}$ Lipschitz in x. To obtain the last inequality first note that $\mathbf{g}\to\|\widehat{\mathbf{w}}^f\|$ is τ_{max}-Lipschitz. Hence, $\frac{\|\widehat{\mathbf{w}}^f\|_2}{\sqrt{n}}$ is C/n sub-Gaussian, and $\frac{1}{n}\|\widehat{\mathbf{w}}^f\|_2^2$ is C/n sub-exponential. We can then use Bernstein inequality of sub-exponential random variables (e.g. Theorem 2.8.1 of (Vershynin, 2018)) to establish the last inequaltiy.

This implies that

$$\mathbb{P}(\widehat{b}^f<0)\leq\mathbb{P}(|\widehat{b}^f-b_*|>b_*)\leq Ce^{-cn}$$

and similarly

$$\mathbb{P}(\widehat{b}^f>b_{max}+1)\leq Ce^{-cn}$$

so

$$\mathbb{P}(|\widehat{b}_+^f-b_*|>\varepsilon)\leq\mathbb{P}(\widehat{b}^f<0)+\mathbb{P}(|\widehat{b}_-^f b_*|>\varepsilon)\leq Ce^{-cn\varepsilon^2}$$

then we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{2}|(\widehat{b}_+^f)^2-b_*^2|>\frac{\varepsilon}{4}\right)&=\mathbb{P}\left(|\widehat{b}_+^f-b_*|\cdot|\widehat{b}_+^f+b_*|>\frac{\varepsilon}{2}\right)\\
&\leq\mathbb{P}(\widehat{b}^f>b_{max}+1)+\mathbb{P}\left(|\widehat{b}_+^f-b_*|>\frac{\varepsilon}{2(2b_{max}+1)}\right)\\
&\leq Ce^{-cn}+Ce^{-c\varepsilon^2}\\
&\leq Ce^{-cn\varepsilon^2}
\end{aligned}\tag{179}$$

for small ε. Now inserting (176), (177), and (179) back into (175) we have our final result

$$\mathbb{P}(|L(\widehat{\mathbf{w}}^f)-\psi(\tau_*,b_*)|>\varepsilon)\leq Ce^{-cn\varepsilon^2}$$

for small ε. □

Lemma 27. *For all $R>0$ there exists constants $C,c>0$ that only depend on (Ω,R) such that for all $\varepsilon\in(0,1]$,*

$$\mathbb{P}(L(\widehat{\mathbf{w}}^f)>\min_{\|\mathbf{w}\|_2\leq\sqrt{n}R}L(\mathbf{w})+\varepsilon)\leq\frac{C}{\varepsilon}e^{-cn\varepsilon^2}$$

Proof. The proof is essentially the same as that of Proposition B.2 in (Miolane and Montanari, 2021) and thus omitted. □

We would now like to combine Lemma 25, Lemma 26 and Lemma 27 to prove the following result:

Corollary 7. *There exist $C,c>0$ depending only on Ω such that for all $\varepsilon\in[0,1]$,*

$$\mathbb{P}(|\min_{\mathbf{w}}L(\mathbf{w})-\psi(b_*,\tau_*)|>\varepsilon)\leq\frac{C}{\varepsilon}e^{-cn\varepsilon^2}$$

Proof. $L(\mathbf{w})$ is $\frac{2\eta}{n}$-strictly convex and $L(\mathbf{w}) \to +\infty$ when $\|\mathbf{w}\|_2 \to +\infty$. Therefore $L(\mathbf{w})$ possesses a unique global minimizer $\mathbf{w}^*$.

By Lemma 25, the event $\{\frac{1}{p}\|\mathbf{w}^* - \widehat{\mathbf{w}}^f\|_2^2 \le 1\}$ has probability at least $1 - Ce^{-cn}$. On this event we have

$$\|\mathbf{w}^*\|_2 \le \|\widehat{\mathbf{w}}^f\|_2 + \|\mathbf{w}^* - \widehat{\mathbf{w}}^f\|_2 \le \|\widehat{\mathbf{w}}^f\|_2 + \sqrt{p}$$

Recall that $\frac{\|\widehat{\mathbf{w}}^f\|}{\sqrt{n}}$ is C/n sub-Gaussian so $\mathbb{P}(\frac{1}{\sqrt{p}}\|\widehat{\mathbf{w}}^f\|_2 \le \frac{1}{\sqrt{p}}\mathbb{E}\|\widehat{\mathbf{w}}^f\|_2 + 1) \ge 1 - Ce^{-cn}$. Therefore

$$\begin{aligned}
\|\mathbf{w}^*\|_2 &\le \mathbb{E}\|\widehat{\mathbf{w}}^f\|_2 + C\sqrt{n} \\
&\le \sqrt{\mathbb{E}\|\widehat{\mathbf{w}}^f\|_2^2} + C\sqrt{n} \\
&\le \sqrt{n(\tau_{max}^2 - \sigma^2)} + C\sqrt{n} \\
&= C\sqrt{n}
\end{aligned} \tag{180}$$

where the second inequality uses Jensen's Inequality, and the third uses Lemma 23.

Let event A be the intersection of above events, i.e. $A := \{\frac{1}{p}\|\mathbf{w}^* - \widehat{\mathbf{w}}^f\|_2^2 \le 1,\ \frac{1}{\sqrt{p}}\|\widehat{\mathbf{w}}^f\|_2 \le \frac{1}{\sqrt{p}}\mathbb{E}\|\widehat{\mathbf{w}}^f\|_2 + 1\}$ then A has probability at least $1 - Ce^{-cn}$. On event A we have:

$$\min_{\mathbf{w}} L(\mathbf{w}) = \min_{\|\mathbf{w}\|_2 \le C\sqrt{n}} L(\mathbf{w}).$$

This means

$$\mathbb{P}\left(L(\widehat{\mathbf{w}}^f) > \min_{\mathbf{w}} L(\mathbf{w}) + \frac{\varepsilon}{2}\right) \le \mathbb{P}\left(L(\widehat{\mathbf{w}}^f) > \min_{\|\mathbf{w}\|_2 \le C\sqrt{n}} L(\mathbf{w}) + \frac{\varepsilon}{2}\right) + Ce^{-cn} \tag{181}$$

So we have

$$\begin{aligned}
\mathbb{P}(|\min_{\mathbf{w}} L(\mathbf{w}) - \psi(b_*, \tau_*)| > \varepsilon) &\le \mathbb{P}\left(|\min_{\mathbf{w}} L(\mathbf{w}) - L(\widehat{\mathbf{w}}^f)| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|L(\widehat{\mathbf{w}}^f) - \psi(b_*, \tau_*)| > \frac{\varepsilon}{2}\right) \\
&\le \mathbb{P}\left(L(\widehat{\mathbf{w}}^f) > \min_{\mathbf{w}} L(\mathbf{w}) + \frac{\varepsilon}{2}\right) + Ce^{-cn\varepsilon^2} \\
&\le Ce^{-cn\varepsilon^2} + Ce^{-cn} + \mathbb{P}\left(L(\widehat{\mathbf{w}}^f) > \min_{\|\mathbf{w}\|_2 \le C\sqrt{n}} L(\mathbf{w}) + \frac{\varepsilon}{2}\right) \\
&\le Ce^{-cn\varepsilon^2} + Ce^{-cn} + Ce^{-cn\varepsilon^2} \\
&\le Ce^{-cn\varepsilon^2}
\end{aligned}$$

for small ε. The second inequality uses Lemma 26 and the fact that $L(\widehat{\mathbf{w}}^f) > \min_{\mathbf{w}} L(\mathbf{w})$. The third inequality use (181). The penultimate inequality uses Lemma 27 with R chosen to be the constant C in (180). Note that all above constants C, c may vary line by line but depend only on model parameters in Ω. □

Lemma 28. *There exists $C, c > 0$ depending only on Ω such that for all $\varepsilon \in [0, 1]$,*

$$\mathbb{P}\left(\left|\min_{\mathbf{w}} \mathcal{C}(\mathbf{w}) - \psi(\tau_*, b_*)\right| \ge \varepsilon\right) \le \frac{C}{\varepsilon} e^{-cn\varepsilon^2}$$

Proof. Using the convex Gaussian minimax theorem stated in (172) and Corollary 7, we have

$$\mathbb{P}\left(\left|\min_{\mathbf{w}} \mathcal{C}(\mathbf{w}) - \psi(\tau_*, b_*)\right| \ge \varepsilon\right) \le 2\mathbb{P}\left(\left|\min_{\mathbf{w}} \mathrm{L}(\mathbf{w}) - \psi(\tau_*, b_*)\right| \ge \varepsilon\right) \le \frac{C}{\varepsilon} e^{-cn\varepsilon^2}.$$

□

Lemma 29. *There exists constants C, c depending only on Ω such that for all closed set $D \in \mathbb{R}^p$, $\forall \varepsilon \in (0, 1]$:*

$$\mathbb{P}(\min_{\mathbf{w} \in \mathcal{D}_\epsilon} \mathcal{C}(\mathbf{w}) \le \min_{\mathbf{w}} \mathcal{C}(\mathbf{w}) + \varepsilon) \le 2\mathbb{P}(\min_{\mathbf{w} \in \mathcal{D}_\epsilon} L(\mathbf{w}) \le \min_{\mathbf{w}} L(\mathbf{w}) + 3\varepsilon) + \frac{C}{\varepsilon} e^{-cn\varepsilon^2}$$

Proof. The proof is essentially the same as the proof of Proposition C.1 in (Miolane and Montanari, 2021) and thus omitted here. □

1.3 The Asymptotic Distribution and Sparsity

As we mentioned in our paper we need to evaluate a few quantities such as the number of non-zero elements $\widehat{\boldsymbol{\beta}}$. Clearly, the empirical disribution of $\widehat{\boldsymbol{\beta}}$ can be used for this purpose. Hence, if we can evaluate what the empirical distribution of $\widehat{\beta}$ converges to, then we can hopefully obtain accurate bounds on e.g., $\|\widehat{\boldsymbol{\beta}}\|_0$.

Let $\widehat{\mu}$ be the empirical distribution of the couple $(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$. According to the results that have appeared in the approximate message framework and CGMT framework (Maleki, 2010; Donoho et al., 2011, 2009; Thrampoulidis et al., 2015, 2018; Wang et al., 2022), we expect $\widehat{\mu}$ to converge to μ^* that is the distribution of the couple

$$\Big(\frac{b_*}{b_*+2\eta\tau_*}\text{soft}\Big(\tau_* Z+\Theta, \frac{\lambda\tau_*}{b_*}\Big), \Theta\Big)$$

where $(Z,\Theta) \sim \mathcal{N}(0,1)\bigotimes \frac{1}{p}\sum \delta_{\beta_k^*}$, and (τ_*, b_*) is the unique saddle point of $\psi(\tau, b)$ defined in (167). The following theorem is a finite sample size confirmation of this claim:

Theorem 8. *There exists constants $C, c > 0$ depending only on Ω such that for all $\varepsilon \in (0, \frac{1}{2}]$,*

$$\mathbb{P}\left(W_2(\widehat{\mu}, \mu^*)^2 \geq \varepsilon\right) \leq C\varepsilon^{-2} e^{cp\varepsilon^3 (\log \varepsilon)^{-2}}$$

Proof. The proof of this Theorem is very similar to the proof of Theorem 3.1 of (Miolane and Montanari, 2021), and hence will be skipped. Please note that there will be some minor changes due to the fact that our regulaizer is elastic net, i.e. $\lambda\|\boldsymbol{\beta}\|_1 + \eta\|\boldsymbol{\beta}\|_2^2$ compared to LASSO in (Miolane and Montanari, 2021), i.e. $\lambda\|\boldsymbol{\beta}\|_1$. For that reason as we described in Section 1.4, our scaler optimization function $\psi(\tau, b)$ is slightly different from the corresponding function in (Miolane and Montanari, 2021). □

As we described before, one of our goals is to use $\widehat{\mu}$ to evaluate the properties of $\widehat{\boldsymbol{\beta}}$. Hence, in most of our results we are more interested in the empirical law of $\widehat{\boldsymbol{\beta}}$, denotes as $\widehat{\mu}_1$, rather than $\widehat{\mu}$. However, it turns out that we can simply obtain a bound for $W_2(\widehat{\mu}_1, \mu_1^*)^2$, where μ_1^* is the law of $\frac{b_*}{b_*+2\eta\tau_*} soft(\tau_* Z + \Theta, \frac{\lambda\tau_*}{b_*})$ using Theorem 8.

Corollary 9. *There exists constants $C, c > 0$ depending only on Ω such that for all $\varepsilon \in (0, \frac{1}{2}]$,*

$$\mathbb{P}\left(W_2(\widehat{\mu}_1, \mu_1^*)^2 \geq \varepsilon\right) \leq C\varepsilon^{-2} e^{cp\varepsilon^3 (\log \varepsilon)^{-2}}$$

Proof. We have

$$\begin{aligned}
W_2^2(\widehat{\mu}, \mu^*) = & \inf_{\substack{(X_1,X_2)\sim\widehat{\mu} \\ (Y_1,Y_2)\sim\mu^*}} \mathbb{E}[(X_1-Y_1)^2 + (X_2-Y_2)^2] \\
\geq & \inf_{\substack{(X_1,X_2)\sim\widehat{\mu} \\ (Y_1,Y_2)\sim\mu^*}} \mathbb{E}(X_1-Y_1)^2 \\
= & \inf_{\substack{X_1\sim\widehat{\mu}_1 \\ Y_1\sim\mu_1^*}} \mathbb{E}(X_1-Y_1)^2 \\
= & W_2^2(\widehat{\mu}_1, \mu_1^*)
\end{aligned} \tag{182}$$

Hence, (182) combined with Theorem 8 completes the proof. □

As we discussed in our main paper, we also need to bound the size of sets for which we have bounds on the magnitude of the subgradients of the ℓ_1-norm. The rest of this section is dedicated to explaining how the sizes of such sets can be bounded. Define

$$\widehat{\mathbf{v}} := \frac{1}{\lambda}\left[\mathbf{X}^\top \mathbf{y} - (\mathbf{X}^\top\mathbf{X} + \eta\mathbf{I}_p)\widehat{\boldsymbol{\beta}}\right]$$

where $\mathbf{I}_p$ is the identity matrix on $\mathbb{R}^{p\times p}$. It can be shown that $\widehat{\mathbf{v}}$ is a subgradient of $\|\widehat{\boldsymbol{\beta}}\|_1$. In fact, the first order condition of $\widehat{\boldsymbol{\beta}}$ gives

$$0 \in \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \lambda\partial\|\widehat{\boldsymbol{\beta}}\|_1 + 2\eta\widehat{\boldsymbol{\beta}}.$$

By simple algebra this is equivalent to $\widehat{\mathbf{v}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$. Note that if $|\widehat{\mathbf{v}}_i| < 1$, then $\widehat{\boldsymbol{\beta}}_i$ has to be zero. Hence, analyzing $\widehat{\mathbf{v}}$ provides an upper bound on $\|\widehat{\boldsymbol{\beta}}\|_0$. Let μ_1^* denote the distribution of

$$\frac{b_*}{b_* + 2\eta\tau_*}\text{soft}(\tau_* Z + \Theta, \frac{\lambda\tau_*}{b_*})$$

as defined previously. Define

$$s_* = \mu_*(\{0\}) = \frac{1}{p}\sum_{k=1}^{p}\left[\Phi\left(\frac{\lambda}{b_*} - \frac{\beta_k^*}{\tau_*}\right) - \Phi\left(-\frac{\lambda}{b_*} - \frac{\beta_k^*}{\tau_*}\right)\right]$$

Note that if $\widehat{\boldsymbol{\beta}}_i \neq 0$ then $|\widehat{\mathbf{v}}_i| = 1$. If there are not many zero coefficients with subgradients whose magnitude is close to 1, we should expect, $\frac{1}{p}\sum_i \mathbb{1}_{\{|\widehat{\mathbf{v}}_i|=1\}}$ to be close to s^*. The following theorem confirms this:

Theorem 10. *There exist constants $C, c > 0$ depending only on Ω such that, for all $\varepsilon \in [0, 1]$,*

$$\mathbb{P}\left(\frac{1}{p}\sum_i \mathbb{1}_{\{|\widehat{\mathbf{v}}_i|\geq 1-\varepsilon\}} \geq s_* + 2(1 + \frac{\lambda}{b_{\min}})\varepsilon\right) \leq \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6},$$

where $b_{\min} > 0$ is the lower bound of b_* in Lemma 24.

Sketch of proof. The proof is similar to that of Lemma 25 using the convex Gaussian minimax theorem (CGMT) that was stated in (172), and is essentially the same as the proof of Theorem E.5 in (Miolane and Montanari, 2021). Therefore we provide a sketch of proof here while omitting the details. First we conctruct the primary optimization (PO) with $\widehat{\mathbf{v}}$ being its unique optimizer. Then we identify the auxillary optimization (AO) of CGMT and study the local stability of AO, similar to Lemma 25. Finally we use CGMT to connect the local stability of AO to that of PO. Define the primary optimizatio (PO) as the following:

$$\mathcal{V}(\mathbf{v}) = \min_{\mathbf{w}} \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \sigma\mathbf{z}\|_2^2 + \frac{\lambda}{n}\mathbf{v}^\top(\boldsymbol{\beta}^* + \mathbf{w}) - \frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1 + \frac{\eta}{n}\|\mathbf{w} + \boldsymbol{\beta}^*\|_2^2 - \frac{\eta}{n}\|\boldsymbol{\beta}^*\|_2^2.$$

It can be verified using dual norm and interchangeability of min-max that

$$\widehat{\mathbf{v}} = \text{argmax}_{\|\mathbf{v}\|_\infty \leq 1}\, \mathcal{V}(\mathbf{v}).$$

Hence, the goal would be to use this optimization and CGMT to provide useful information about $\widehat{\mathbf{v}}$. As we described before, we expect $\frac{1}{p}\sum_k 1_{\{|\mathrm{v_k}|\geq 1-\varepsilon\}}$ to be close to the number of nonzero coefficients and that should be close to s_*. Hence, we set

$$D_\varepsilon := \left\{\mathbf{v} : \|\mathbf{v}\|_\infty \leq 1, \frac{1}{p}\sum_k 1_{\{|\mathrm{v_k}|\geq 1-\varepsilon\}} \geq \mathrm{s_*} + 2(1 + \frac{\lambda}{\mathrm{b_{min}}})\varepsilon\right\}.$$

We have

$$\mathbb{P}\left(\frac{1}{p}\sum_i \mathbb{1}_{\{|\widehat{\mathbf{v}}_i|\geq 1-\varepsilon\}} \geq s_* + 2(1 + \frac{\lambda}{b_{\min}})\varepsilon\right) = \mathbb{P}(\widehat{\mathbf{v}} \in D_\varepsilon) \leq \mathbb{P}\left(\max_{D_\varepsilon}\mathcal{V}(\mathbf{v}) \geq \max_{\|\mathbf{v}\|_\infty\leq 1}\mathcal{V}(\mathbf{v}) - \varepsilon'\right). \tag{183}$$

for any $\varepsilon' > 0$. Note that ε' will be decided after the analysis of PO is finished. Using the same arguments as the ones presented in Section 1.2, we can obtain the following auxillary optimization for this problem:

$$\begin{aligned} V(\mathbf{v}) = \min_{\mathbf{w}} &\frac{1}{2}\left(\sqrt{\frac{\|\mathbf{w}\|_2^2}{n} + \sigma^2}\frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \frac{1}{n}\mathbf{g}^\top\mathbf{w} + \frac{g'\sigma}{\sqrt{n}}\right)_+^2 + \frac{\lambda}{n}\left(\mathbf{v}^\top(\mathbf{w} + \boldsymbol{\beta}^*) - \|\boldsymbol{\beta}^*\|_1\right) \\ &+ \frac{\eta}{n}\left(\|\mathbf{w} + \boldsymbol{\beta}^*\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2\right) \end{aligned} \tag{184}$$

where $\mathbf{g} \sim N(0, \mathbb{I}_p)$, $\mathbf{h} \sim N(0, \mathbb{I}_n)$ and $g' \sim N(0, 1)$, independent with each other. Directly working with D_ε is quite difficult, but recall in Lemma 25 we have defined a set $\{\mathbf{w} : \frac{1}{p}\|\mathbf{w} - \widehat{\mathbf{w}}^f\|_2^2 > \varepsilon\}$ and it was easier to work with. In fact, we can define a similar set

$$\tilde{D}_\varepsilon := \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_\infty \leq 1, \frac{1}{p}\|\mathbf{v} - \widehat{\mathbf{v}}^f\|_2^2 \geq \varepsilon\},$$

with

$$\widehat{\mathbf{v}}^f := -\frac{b_*}{\lambda\tau_*}(\widehat{\mathbf{w}}^f - \tau_*\mathbf{g}) - \frac{2\eta}{\lambda}(\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*).$$

And one can then show that, for some constants $C, c, \gamma > 0$:

$$\mathbb{P}(\max_{\mathbf{v}\in\tilde{D}_\varepsilon} V(\mathbf{v}) \geq \max_{\mathbf{v}} V(\mathbf{v}) - c\varepsilon) \leq \frac{C}{\varepsilon}e^{-cn\varepsilon^2} \tag{185}$$

The proof is essentially the same as that of Theorem E.7 of (Miolane and Montanari, 2021) and thus omitted here.

The next step is to substitute the $\tilde{D}_\varepsilon$ in (185) back to D_ε. The goal is to prove that, for some constants $C, c, \gamma > 0$, for all $\varepsilon \in (0, 1]$:

$$\mathbb{P}(\max_{\mathbf{v}\in D_\varepsilon} V(\mathbf{v}) \geq \max_{\mathbf{v}} V(\mathbf{v}) - 3\gamma\varepsilon^3) \leq \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6} \tag{186}$$

The proof is essentially the same as that of Lemma E.9 of (Miolane and Montanari, 2021) and thus omitted.

Then we connect (186) with its $\mathcal{V}(\mathbf{v})$ version via CGMT. Notice that by interchanging min-max and using dual norm expressions, we have

$$\max_{\|\mathbf{v}\|_\infty\leq 1} \mathcal{V}(\mathbf{v}) = \min_{\mathbf{w}} \mathcal{C}(\mathbf{w}), \qquad \max_{\|\mathbf{v}\|_\infty\leq 1} V(\mathbf{v}) = \min_{\mathbf{w}} L(\mathbf{w}).$$

Hence, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{D_\varepsilon} \mathcal{V}(\mathbf{v}) \geq \max_{\|\mathbf{v}\|_\infty\leq 1} \mathcal{V}(\mathbf{v}) - 3\gamma\varepsilon^3\right) \\
&\leq \mathbb{P}\left(\max_{\|\mathbf{v}\|_\infty\leq 1} \mathcal{V}(\mathbf{v}) < \psi(\tau_*, b_*) - \gamma\varepsilon^3\right) + \mathbb{P}\left(\max_{D_\varepsilon} \mathcal{V}(\mathbf{v}) \geq \psi(\tau_*, b_*) - 2\gamma\varepsilon^3\right) \\
&\leq \mathbb{P}\left(\min_{\mathbf{w}} \mathcal{C}(\mathbf{w}) < \psi(\tau_*, b_*) - \gamma\varepsilon^3\right) + 2\mathbb{P}\left(\max_{D_\varepsilon} V(\mathbf{v}) \geq \psi(\tau_*, b_*) - 2\gamma\varepsilon^3\right).
\end{aligned} \tag{187}$$

Now by Lemma 28, the first term is bounded by $\frac{C}{\varepsilon^3}e^{-cn\varepsilon^6}$. For the second term, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{D_\varepsilon} V(\mathbf{v}) \geq \psi(\tau_*, b_*) - 2\gamma\varepsilon^3\right) \\
&\leq \quad \mathbb{P}\left(\max_{D_\varepsilon} V(\mathbf{v}) \geq \max_{\|\mathbf{v}\|_\infty\leq 1} V(\mathbf{v}) - 3\gamma\varepsilon^3\right) + \mathbb{P}(\max_{\|\mathbf{v}\|_\infty\leq 1} V(\mathbf{v}) > \psi(\tau_*, b_*) + \gamma\varepsilon^3) \\
&\leq \quad \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6} + \mathbb{P}(\min_{\mathbf{w}} L(\mathbf{w}) > \psi(\tau_*, b_*) + \gamma\varepsilon^3) \\
&\leq \quad \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6} + \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6}
\end{aligned} \tag{188}$$

The penultimate inequality uses (186) and the last inequality uses Corollary 7. Putting (183), (187) and (188) together we have

$$\mathbb{P}\left(\frac{1}{p}\sum_i \mathbb{1}_{\{|\widehat{\mathbf{v}}_i|\geq 1-\varepsilon\}} \geq s_* + 2(1 + \frac{\lambda}{b_{\min}})\varepsilon\right) \leq \frac{C}{\varepsilon^3}e^{-cn\varepsilon^6}.$$

□

The final theorem that we would like to mention in the appendix is a concentration result on the number of nonzero elements of $\widehat{\boldsymbol{\beta}}$.

Theorem 11. *There exist constants C, c depending only on Ω such that for all $0 < \varepsilon < 1$,*

$$\mathbb{P}\left(\left|\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|_0 - s_*\right| \geq \varepsilon\right) \leq C\varepsilon^{-6}e^{-cp\varepsilon^6}$$

Proof. Note that if an elements of $\widehat{\boldsymbol{\beta}}_i \neq 0$, then its corresponding subgradient has to be either 1 or -1. Hence, the upper bound of $\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|_0$ is a direct result of Lemma 10. For the lower bound, the proof follows a similar PO-AO-local stability path and is essentially the same as the proof of Theorem F.1 in (Miolane and Montanari, 2021). □

1.4 Study of the Scalar Optimization

Let $h_r(x)$ be the Huber loss

$$h_r(x) := \frac{1}{2r}\|\text{soft}(\mathbf{x}, r) - \mathbf{x}\|_2^2 + \|\text{soft}(\mathbf{x}, r)\|_1.$$

We remind the reader that

$$\begin{aligned}\psi(\tau, b) = &\left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{b^2}{2} \\ &+ \frac{1}{n}\mathbb{E}\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{b}{2\tau}\|\mathbf{w}\|_2^2 - b\mathbf{g}^\top\mathbf{w} + \lambda\left(\|\mathbf{w} + \boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^*\|_1\right) + \eta\left(\|\mathbf{w} + \boldsymbol{\beta}^*\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2\right)\right\},\end{aligned} \tag{189}$$

and that

$$\widehat{\mathbf{w}}^f(\tau, b) := \frac{b}{b + 2\eta\tau}\text{soft}\left(\tau\mathbf{g} + \boldsymbol{\beta}^*, \frac{\lambda\tau}{b}\right) - \boldsymbol{\beta}^*$$

Inserting back $\mathbf{w} = \widehat{\mathbf{w}}^f(\tau, b)$, it can be shown that

$$\psi(\tau, b) = \left(\frac{\sigma^2}{\tau'} + \frac{b\tau'}{b - 2\eta\tau'}\right)\frac{b}{2} - \frac{b^2}{2} - \eta\sigma^2 + \frac{\lambda}{n}\mathbb{E}h_{\frac{\lambda\tau'}{b}}(\tau'\tilde{\mathbf{g}} + \boldsymbol{\beta}^*) - \frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1 - \frac{b\tau'}{2n}\mathbb{E}\|\tilde{\mathbf{g}}\|_2^2$$

where $\tau' = \frac{b}{b+2\eta\tau}\tau$ and $\tilde{\mathbf{g}} = \mathbf{g} - \frac{2\eta}{b}\boldsymbol{\beta}^*$. The variable $\widehat{\mathbf{w}}^f$ and function $\psi(\tau, b)$ play an important role in our analysis later. In this section, we study the saddle point of $\psi(\tau, b)$:

$$(\tau_*, b_*) := \operatorname*{argmax}_{b\geq 0}\operatorname*{argmin}_{\tau\geq\sigma}\psi(\tau, b)$$

In the rest of this section we prove Lemma 23 and Lemma 24.

Proof of Lemma 23. Note that ψ is convex-concave and differentiable with respect to (τ, b), and the differentiation can be taken inside the expectation. Using the formulae

$$\begin{aligned}\frac{\partial}{\partial r}h_r(x) &= -\frac{1}{2r^2}[\text{soft}(x, r) - x]^2 \\ \frac{\partial}{\partial x}h_r(x) &= \frac{1}{r}\left(x - \text{soft}(x, r)\right)\end{aligned}$$

one can obtain the following:

$$\begin{aligned}\frac{\partial}{\partial\tau}\psi(\tau, b) &= \frac{b}{2\tau^2}\left(\tau^2 - \sigma^2 - \frac{1}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f(\tau, b)\|_2^2\right), \\ \frac{\partial}{\partial b}\psi(\tau, b) &= \tau - b - \frac{1}{n}\mathbb{E}\mathbf{g}^\top\widehat{\mathbf{w}}^f(\tau, b).\end{aligned}$$

First, for each $b \geq 0$ consider $\min_{\tau\geq\sigma}\psi(\tau, b)$. Let $f_b(\tau) = \frac{\partial}{\partial\tau}\psi(\tau, b)$. We have

$$f_b(\tau) = \frac{b}{2}\left(1 - \frac{\sigma^2}{\tau^2} - \frac{1}{n}\sum_{k=1}^{p}\mathbb{E}\left[\frac{b}{b + 2\eta\tau}\text{soft}(g_k + \frac{\beta_k^*}{\tau}, \frac{\lambda}{b}) - \frac{\beta_k^*}{\tau}\right]^2\right).$$

Hence, we have

- $f_b(\tau)$ is differentiable since the integrand is almost surely differentiable, and the distribution of $\mathbf{g}$ is continuous

- $f_b'(\tau) > 0$ because $\psi(\cdot, b)$ is strictly convex
- $f_b(\sigma) = -\frac{b}{2n}\mathbb{E}\|\widehat{\mathbf{w}}^f(\tau, b)\|^2 < 0$ since $\widehat{\mathbf{w}}^f(\tau, b)$ is non-degenerate
- $f_b(+\infty) = 1$ using the Dominated Convergence Theorem.

Therefore $\forall b \geq 0$, $\exists$ unique $\tau_0(b) > \sigma$ such that $f_b(\tau_0(b)) = 0$. Moreover $f_b(\tau) < 0$ on the left and > 0 on the right. Hence, $\tau_0(b)$ is the unique minimizer of $\psi(\tau, b)$ at b. Moreover, by the Implicit Function Theorem, $\tau_0(b)$ is differentiable.

Next we define $G(b) = \psi(\tau_0(b), b) = \min_{\tau \geq \sigma} \psi(\tau, b)$ so that

$$\max_{b \geq 0} \min_{\tau \geq \sigma} \psi(\tau, b) = \max_{b \geq 0} G(b)$$

Its derivative is given by

$$\begin{aligned} g(b) :&= G'(b) \\ &= \frac{\partial}{\partial b}\psi(\tau, b)\Big|_{\tau=\tau_0(b)} + \frac{\partial}{\partial \tau}\psi(\tau, b)\Big|_{\tau=\tau_0(b)} \tau_0'(b) \\ &= \tau_0(b) - b - \frac{1}{n}\mathbb{E}\mathbf{g}^\top \widehat{\mathbf{w}}^f(\tau_0(b), b) \end{aligned}$$

The last line is because $\frac{\partial}{\partial \tau}\psi(\tau, b)\big|_{\tau=\tau_0(b)} = 0$. Next we show that $G(b)$ has a unique maximizer $b_* > 0$ and it is also the unique solution of $g(b) = 0$.

- $g(b)$ is a decreasing function, because $G(b)$ is the pointwise minimum of a collection of strictly-concave functions, and is hence strictly-concave itself
- $\liminf_{b \to 0_+} g(b) \geq \sigma$. To see this, first notice that $\tau_0(b)$ is the zero of $\frac{2}{b} f_b(\tau)$ and

  $$\lim_{b \to 0_+} \frac{2}{b} f_b(\tau) = 1 - \tau^{-2}(\sigma^2 + \frac{1}{n}\|\boldsymbol{\beta}^*\|^2).$$

  Therefore, we have

  $$\lim_{b \to 0_+} \tau_0(b) = \sigma^2 + \frac{1}{n}\|\boldsymbol{\beta}^*\|^2,$$

  By using Fatou's lemma we obtain

  $$\begin{aligned} \lim \inf_{b \to 0_+} g(b) &= \lim_{b \to 0_+} \tau_0(b) - \lim \sup_{b \to 0_+} \frac{1}{n}\mathbb{E}\mathbf{g}^\top \widehat{\mathbf{w}}^f(\tau_0(b), b) \\ &\geq \sigma^2 + \frac{1}{n}\|\boldsymbol{\beta}^*\|^2 - \frac{1}{n}\sum_{k=1}^{p} \mathbb{E}\left[g_k\left(0 - \beta_k^*\right)\right] \\ &= \sigma^2 + \frac{1}{n}\|\boldsymbol{\beta}^*\|^2 \\ &\geq \sigma^2 \end{aligned}$$

- As mentioned before $G(b)$ is the pointwise minimum of a collection of strictly concave functions so it is itself strictly concave and therefore admits a unique maximizer $b_* \geq 0$. By the last point $\liminf_{b \to 0_+} g(b) \geq \sigma^2 > 0$. Therefore the maximizer cannot be 0, and hence $b_* > 0$. Since $G(b)$ is differentiable b_* must be a zero of $g(b)$, and the zero must be unique as the maximizer is unique.

We conclude that the unique saddle point is $(\tau_0(b_*), b_*)$ and it satisfies (168). □

Proof of Lemma 24. We remind the reader of the following notation:

$$\widehat{\mathbf{w}}^f := \widehat{\mathbf{w}}^f(\tau_*, b_*).$$

We divide the proof into the following steps:

Step 1 (Lower bound for b_*): Using the same notations as the ones used in the proof of Lemma 23, we have $G(b) = \min_{\tau \geq \sigma} \psi(\tau, b)$ and $g(b) = G'(b)$. Using the fact that $\mathbb{E}Z \cdot \text{soft}(Z + a, r) = \Phi(a - r) + \Phi(-a - r)$ (which can be verified directly via integration), we have

$$\begin{aligned}
g(b) = G'(b) &= \tau_0(b) - \frac{1}{n}\mathbb{E}\mathbf{g}^\top \widehat{\mathbf{w}}^f(\tau_0(b), b) - b \\
&= \tau_0(b)\left(1 - \frac{1}{n}\frac{b}{b + 2\eta\tau_0(b)}\sum_{k=1}^{p}\mathbb{E}g_k \cdot \text{soft}(g_k + \frac{\theta_k^*}{\tau_0(b)}, \frac{\lambda}{b})\right) - b \\
&= \tau_0(b)\left(1 - \frac{1}{n}\frac{b}{b + 2\eta\tau_0(b)}\sum_{k=1}^{p}\left[\Phi(\frac{\theta_k^*}{\tau_0(b)} - \frac{\lambda}{b}) + \Phi(-\frac{\theta_k^*}{\tau_0(b)} - \frac{\lambda}{b})\right]\right) - b \\
&\geq \tau_0(b)\left(1 - \frac{1}{\gamma_0}\mathbb{E}\left[\Phi(\frac{\Theta}{\tau_0(b)} - \frac{\lambda}{b}) + \Phi(-\frac{\Theta}{\tau_0(b)} - \frac{\lambda}{b})\right]\right) - b \\
&:= \tau_0(b)\left(1 - \frac{1}{\gamma_0}\mathbb{E}h_b(\frac{\Theta}{\tau_0(b)})\right) - b,
\end{aligned}$$

where Θ has uniform distribution over the elements of β^*, and $h_b(x) = \Phi(x - \alpha) + \Phi(-x - \alpha)$ is an even funcion, decreasing for $x < 0$ and increasing for $x \geq 0$. Let $K = \frac{2\xi}{\sigma\sqrt{\gamma_0}}$. By Markov inequality we have

$$\begin{aligned}
\mathbb{E}h_b\left(\frac{\Theta}{\tau_0(b)}\right) &\leq h_b(K) + \mathbb{P}\left(\left|\frac{\Theta}{\tau_0(b)}\right| \geq K\right) \\
&\leq h_b(K) + \frac{\mathbb{E}\Theta^2}{\tau_0^2(b)K^2} \\
&\leq h_b(K) + \frac{\gamma_0}{4}.
\end{aligned}$$

Finally it can be verified directly that $\lim_{b \to 0_+} h_b(K) = 0$. Henec we can find $b_0 > 0$ depending only on $\xi, \lambda, \sigma, \gamma_0$ such that $\forall b \leq b_0, h_b(K) \leq \frac{\gamma_0}{4}$. Hence $\forall b \leq b_0$ we have

$$\begin{aligned}
g(b) &\geq \tau_0(b)(1 - \frac{1}{\gamma_0}(\frac{\gamma_0}{4} + \frac{\gamma_0}{4})) - b \\
&= \frac{1}{2}\tau_0(b) - b \\
&\geq \frac{\sigma}{2} - b
\end{aligned}$$

If we let $b_{min} = \min\{b_0, \frac{\sigma}{4}\}$, then $\forall b \leq b_{min}$, $g(b) \geq \frac{\sigma}{4} > 0$. Since b_* is the zero of $g(b)$ and $g(b)$ is decreasing, we conclude that $b_* > b_{min}$. This finishes the proof of the first step.

Step 2 ($\tau_* < \tau_{max}$): First note that if we insert $\mathbf{w} = 0$ in the definition of $\psi(\tau, b)$, then we will have $\psi(\tau, b) \leq \left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{b^2}{2}$. Hence,

$$\psi(\tau_*, b_*) \leq \max_{b \geq 0}\min_{\tau \geq \sigma}\left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{b^2}{2} = \frac{\sigma^2}{2} \tag{190}$$

Next, we show that $\psi(\tau, b)$ has an increasing lower bound $g(\tau)$ for all $b \geq b_{min}$. In fact,

$$\begin{aligned}
\psi(\tau, b) &= \left(\frac{\sigma^2}{\tau} + \tau\right)\frac{b}{2} - \frac{b^2}{2} + \frac{1}{n}\mathbb{E}\left\{\|\widehat{\mathbf{w}}^f\|_2^2(\frac{b}{2\tau} + \eta) - (b\mathbf{g} - 2\eta\beta^*)^\top\widehat{\mathbf{w}}^f + \lambda\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \lambda\|\boldsymbol{\beta}^*\|_1\right\} \\
&= \frac{b}{2\tau}(\sigma^2 + \tau^2) - \frac{b^2}{2} + (\frac{b}{2\tau} + \eta)(\tau^2 - \sigma^2) - b(\tau - b) + \frac{2\eta}{n}\mathbb{E}\boldsymbol{\beta}^{*\top}\widehat{\mathbf{w}}^f + \frac{\lambda}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1 \\
&= \eta(\tau^2 - \sigma^2) + \frac{b^2}{2} + \frac{2\eta}{n}\mathbb{E}\boldsymbol{\beta}^{*\top}\widehat{\mathbf{w}}^f + \frac{\lambda}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 - \frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1.
\end{aligned} \tag{191}$$

The first two terms in the last line of the above equation are nonnegative when $\tau = \tau_*$. For the third term,

$$\begin{aligned}
\frac{2\eta}{n}\mathbb{E}\boldsymbol{\beta}^{*\top}\widehat{\mathbf{w}}^f &= \frac{2\eta}{n}\frac{b_*\tau_*}{b_* + 2\eta\tau_*}\sum_i \beta_i^* \cdot \mathbb{E}\text{soft}\left(Z + \frac{\beta_i^*}{\tau_*}, \frac{\lambda}{b_*}\right) \\
&\geq -\frac{2\eta}{n}\frac{b_*\tau_*}{b_* + 2\eta\tau_*}\sum |\beta_i^*| \cdot \left(\frac{|\beta_i^*|}{\tau_*} + \frac{\lambda}{b_*}\right) \\
&\geq -\frac{2\eta}{n}\frac{b_*}{b_* + 2\eta\tau_*}\|\boldsymbol{\beta}^*\|_2^2 - -\frac{2\eta}{n}\frac{\lambda\tau_*}{b_* + 2\eta\tau_*}\|\boldsymbol{\beta}^*\|_1 \\
&\geq -\frac{2\eta_{\max}}{\delta}\xi^2 - \frac{\lambda_{\max}}{\delta}\xi \\
&:= -C_1
\end{aligned} \tag{192}$$

In the above equations, the first inequality uses $|\mathbb{E}\text{soft}(Z + x, r)| \leq |x| + r$, the third uses Cauchy-Schwarz inequality, and the last uses $\frac{b_*}{b_* + 2\eta\tau_*} \leq 1$, $\frac{\eta\tau_*}{b_* + 2\eta\tau_*} \leq \frac{1}{2}$.

For the fourth term of (191), using the notation Θ taking values uniformly from the elements of $\boldsymbol{\beta}^*$ and $Z \sim N(0, 1)$ and independent from Θ, we have

$$\begin{aligned}
\frac{\lambda}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 &= \frac{\lambda}{\gamma_0}\frac{b_*\tau_*}{b_* + 2\eta\tau_*}\mathbb{E}\left|\text{soft}(Z + \frac{\Theta}{\tau_*}, \frac{\lambda}{b_*})\right| \\
&\geq \frac{\lambda}{\gamma_0}\frac{b_{\min}\tau_*}{b_{\min} + 2\eta\tau_*}\mathbb{E}\left|\text{soft}(Z + \frac{\Theta}{\tau_*}, \frac{\lambda}{b_{\min}})\right|.
\end{aligned}$$

Now consider the function $h(a) = \mathbb{E}|\text{soft}(Z + a, r)|$ with $r = \frac{\lambda}{b_{min}}$. We have $h(0) = \mathbb{E}|\text{soft}(Z, r)| := C > 0$ and by Monotone Convergence Theorem we have $\lim_{a\to\infty} h(a) = +\infty$. Therefore there exists a constant $C_2 > 0$ that depends only on $r = \frac{\lambda}{b_{\min}}$ such that $h(a) \geq C_2$. Hence, we have

$$\begin{aligned}
\frac{\lambda}{n}\mathbb{E}\|\widehat{\mathbf{w}}^f + \boldsymbol{\beta}^*\|_1 &\geq \frac{\lambda}{\gamma_0}\frac{b_{min}\tau_*}{b_{min} + 2\eta\tau_*}\mathbb{E}h(\frac{\Theta}{\tau_*}) \\
&\geq \frac{\lambda}{\gamma_0}\frac{b_{min}\tau_*}{b_{min} + 2\eta\tau_*}C_2
\end{aligned} \tag{193}$$

Finally, for the last term of (191) we have

$$\frac{\lambda}{n}\|\boldsymbol{\beta}^*\|_1 \leq \frac{\lambda\xi}{\gamma_0} \leq \frac{\lambda_{\max}\xi}{\gamma_0}. \tag{194}$$

Combining (190), (192), (193) and (194) together, we have

$$\eta(\tau_*^2 - \sigma^2) - C_1 + \frac{\lambda}{\gamma_0}\frac{b_{\min}\tau_*}{b_{\min} + 2\eta\tau_*}C_2 - \frac{\lambda_{\max}\xi}{\gamma_0} \leq \psi(\tau_*, b_*) \leq \frac{\sigma^2}{2}$$

If we define $C_1' = \frac{sigma^2}{2} + \eta_{\max}\sigma^2 + C_1 + \frac{\lambda_{\max}\xi}{\gamma_0}$, $C_2' = \frac{\lambda_{\max}}{\gamma_0}b_{\min}C_2$, we can rephrase the above inequality into

$$\eta\tau_*^2 + \frac{C_2'\tau_*}{b_{\min} + 2\eta\tau_*} \leq C_1'$$

It can be verified directly that, the minimum of the left hand side over $\eta \geq 0$ is $\max\{(\sqrt{2C_2'} - \frac{b_{\min}}{2}), \frac{C_2'}{b_{\min}}\}\tau_* := C_3\tau_*$. Hence, we conclude that $\tau_* < \tau_{\max} = \frac{C_1'}{C_3}$ depending on model parameters.

Step 3 (Upper bound for b_*): First note that

$$\frac{1}{n}\mathbb{E}\mathbf{g}^\top\widehat{\mathbf{w}}^f = \frac{1}{n}\frac{b_*\tau_*}{b_* + 2\eta\tau_*}\sum_i \mathbb{P}\left(\left|g_i + \frac{\beta_i^*}{\tau_*}\right| > \frac{\lambda}{b_*}\right) \geq 0.$$

Therefore

$$b_* = \tau_* - \frac{1}{n}\mathbb{E}\mathbf{g}^\top\widehat{\mathbf{w}}^f \leq \tau_{max}.$$

Step 4 (Lower bound for τ_*): It is trivial that $\tau_* > \sigma$ and is hence skipped. □

References

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer New York, NY, 2007.

Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6): 2340–2381, 2014.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis, second edition*. Cambridge University Press, 1994.

Shirin Jalali and Arian Maleki. New approach to bayesian high-dimensional linear regression. *Information and Inference: A Journal of the IMA*, 7, 07 2016.

Mohammad Ali Maleki. *Approximate message passing algorithms for compressed sensing*. Stanford University, 2010.

Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 2020.

Herbert Robbins. A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1683–1709, 2015.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Does SLOPE outperform bridge regression? *Information and Inference: A Journal of the IMA*, 11(1):1–54, 2022.