# Theoretical Analysis of Leave-one-out Cross Validation for Non-differentiable Penalties under High-dimensional Settings

**Haolin Zou**
Columbia University

**Arnab Auddy**
The Ohio State University

**Kamiar Rahnama Rad**
Baruch College, CUNY

**Arian Maleki**
Columbia University

## Abstract

Despite a large and significant body of recent work focusing on the tuning of hyperparameters for regularized models in the high dimensional regime, a theoretical understanding of this problem for non-differentiable penalties such as generalized LASSO and nuclear norm is missing. In this paper we resolve this challenge. We study this problem in the proportional high dimensional regime where both the sample size $n$ and number of features $p$ are large, and $n/p$ and the signal-to-noise ratio (per observation) remain finite. To achieve this goal, we first provide finite-sample upper bounds on the expected squared error of leave-one-out cross-validation (LO) in estimating the out-of-sample risk. Building on this result, we establish the consistency of hyperparameter tuning based on minimizing LO's estimate. Our simulation results confirm the accuracy and sharpness of our theoretical results.

## 1 INTRODUCTION

### 1.1 Background

Setting model complexity based on hyperparameter tuning is a crucial task in statistical learning. Most hyperparamter tuning techniques first estimate the out-of-sample risk (OO), and then find the optimal choice of the hyperparameter by optimizing the risk estimate. A large body of recent work has studied the theoretical and empirical properties of cross-validation (or its variants) for risk estimation and hyperparameter tuning, for example: Rahnama Rad et al. (2020); Rahnama Rad and Maleki (2020); Patil et al. (2021, 2022a, 2024, 2022b); Xu et al. (2021); Bellec (2023); Beirami et al. (2017); Auddy et al. (2023); Stephenson and Broderick (2020); Stephenson et al. (2021); Luo et al. (2023); Nobel et al. (2024).

In spite of these recent advances, a complete quantitative description of the risk in using leave-one-out cross-validation (LO) is missing when it comes to problems that have the following characteristics 1) non-differentiable penalties such as generalized LASSO and nuclear norm, and 2) finite signal-to-noise ratio while the dimension of the feature space grows proportionally with the number of observations. In this paper, we focus on the aforementioned characteristics for regularized generalized linear models (including logistic and Poisson regression). We use minor assumptions about the data generating process to provide finite sample upper bounds on the expected squared error of leave-one-out cross-validation in estimating the out-of-sample error. We will then use this result to prove that the hyperparameter tuning scheme, based on minimizing LO, is consistent in the high-dimensional setting considered in this paper.

### 1.2 Problem Statement and Related Work

Consider the dataset $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ denote the features and response of the $i^{\text{th}}$ data point, respectively. Observations are independent and identically distributed (iid), drawn from some unknown joint distribution $q(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}^*)p(\mathbf{x}_i)$, where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ represents the true parameter.

The class of estimates, known as regularized empirical risk minimizers (R-ERM), are based on the following optimization:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}), \tag{1}$$

where $\ell(y, z)$ is a non-negative convex loss function between $y$ and $z$, e.g., square loss $\ell(y, z) = \frac{1}{2}(y - z)^2$, $r(\boldsymbol{\beta})$ is a non-negative convex regularizer, and $\boldsymbol{\Theta}$ is a convex set to which $\boldsymbol{\beta}^*$ belongs. By selecting particular forms for $\ell$ and $r$ we can cover a wide range of popular

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

estimators within the high-dimensional setting. For instance, one may use logistic or Poisson (negative log-likelihoods) for the loss function, and fused LASSO or nuclear norm for the regularizer.

One widely used criterion for model selection and hyperparameter tuning (e.g. tuning of $\lambda$ in (1)) is the out-of-sample prediction error (OO), defined as

$$\mathrm{OO} := \mathbb{E}[\phi(y_0, \mathbf{x}_0^\top \widehat{\boldsymbol{\beta}})|\mathcal{D}],$$

where $\phi(y, z)$ is a function measuring the difference between $y$ and $z$ (often, but not necessarily the same as $\ell(y, z)$), and $(y_0, \mathbf{x}_0)$ is a new sample from the same joint distribution $q(y|\mathbf{x}^\top\boldsymbol{\beta}^*)p(\mathbf{x})$, independent of the training set $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$.

Several methods are proposed to estimate OO, such as $K$-fold cross-validation, Generalized Cross-Validation, and Bootstrap, among others. In this paper, we employ the term "risk estimate" as a generic term for referring to these techniques. Despite the abundance of theoretical and empirical results in the literature, studies specifically examining the accuracy of risk estimates in high-dimensional settings, where the number of features is either larger than or comparable to the sample size $n$, are noticeably scarce. As a result, many fundamental questions regarding the performance of risk estimates and their applications in hyperparameter tuning have remained open.

The first objective of this paper is to prove the accuracy of the leave one out risk estimate LO for a large class of R-ERMs under the high-dimensional setting. To formalize the problem, for $1 \le i \le n$ define

$$\widehat{\boldsymbol{\beta}}_{/i} := \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{j\neq i} \ell(y_j, \mathbf{x}_j^\top\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}), \tag{2}$$

and define LO as:

$$\mathrm{LO} := \frac{1}{n}\sum_{i=1}^n \phi(y_i, \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i}). \tag{3}$$

The success of LO in practical scenarios has encouraged numerous researchers to investigate its accuracy and develop several computationally-efficient approximations for it, e.g. Rahnama Rad et al. (2020); Rahnama Rad and Maleki (2020); Patil et al. (2021, 2022a, 2024); Xu et al. (2021); Bellec (2023); Beirami et al. (2017); Auddy et al. (2023); Stephenson and Broderick (2020); Stephenson et al. (2021); Luo et al. (2023); Nobel et al. (2024). Notably, key references that are related to our current work are Burman (1989); Austern and Zhou (2020); Rahnama Rad et al. (2020). In their study, Burman (1989) studied the accuracy of LO in the classical low-dimensional setting where $p$ remains fixed and $n \to \infty$ and established the consistency of LO, i.e. LO $\xrightarrow{p}$ OO. Going a step further, Austern and Zhou (2020) has characterized the limiting distribution of the $k$-fold cross-validation under the same low-dimensional setting. As for the high dimensional setting, in Rahnama Rad et al. (2020), the authors studied the accuracy of the LO under the high-dimensional setting, similar to the setting we consider in this paper. However, it is worth noting that the conclusions drawn in Rahnama Rad et al. (2020) are contingent upon the following two assumptions that restrict the broader applicability of their outcomes:

- : The regularizer is twice differentiable.
- : The set $\boldsymbol{\Theta}$ is the same as $\mathbb{R}^p$.

Removing these restrictions significantly broadens the scope of LO in estimating OO, making it applicable across a diverse range of uses. Eliminating the first restriction allows for the inclusion of key regularizers frequently used in real-world settings, such as the $\ell_1$ norm, nuclear norm and group-LASSO, among others. The removal of the second restriction expands the scope of the theory to cases involving parameters that must conform to particular set and shape constraints. This encompasses issues that necessitate shape constraints, for instance, ensuring matrices are positive definite, as well as techniques like constrained lasso, ordered lasso, isotonic regression, convex regression, and regression with compositional covariates. It also includes scenarios where the elements of $\boldsymbol{\beta}^*$ need to be positive, monotonically increasing, or both, as highlighted in various studies including Gaines et al. (2018); Tibshirani and Suo (2016); Shi et al. (2016); Lin et al. (2014); Guntuboyina and Sen (2018). Applications that benefit from addressing these constraints include vaccine development Hu et al. (2015), document categorization El-Arini et al. (2013), and investment portfolio optimization Zhao et al. (2014).

The second objective of this paper is to establish the accuracy of the hyperparameter tuning scheme based on LO. As explained in Section 6.6, this problem is more challenging than proving the accuracy of LO for risk estimation. For example, proving this requires demonstrating that the risk estimate is uniformly accurate across all possible choices of hyperparameters.

Before we start the technical contributions of our paper, we would like to discuss the computational complexity of LO. While the heavy computational demand of LO has not been addressed in our paper, recent literature has explored approximations for leave-one-out cross-validation. Several papers have proposed various approximations, including those discussed in Rahnama Rad and Maleki (2020), Beirami et al. (2017), Giordano et al. (2019), Auddy et al. (2023), along with

references therein. In cases where the regularizer is non-differentiable, such as in our paper, Wang et al. (2018); Nobel et al. (2024) has introduced a range of heuristic methods to obtain computationally efficient approximations of *LO*. Additionally, recently Auddy et al. (2023) has made progress in proving the conjecture presented in Wang et al. (2018) when the regularizer is $\lambda(1-\eta)\|\cdot\|_1 + \lambda\eta\|\cdot\|_2^2$.

### 1.3 Notations

In this section, we summarize the symbols we use throughout the article. Vectors are denoted by boldfaced lowercase letters, such as $\mathbf{x}$. Matrices are represented by boldfaced capital letters, such as $\mathbf{X}$. For a matrix $\mathbf{X}$, $\sigma_{\min}(\mathbf{X})$, $\|\mathbf{X}\|$, $\mathrm{Tr}(\mathbf{X})$ denote the minimum singular value, the spectral norm (equal to the maximum singular value $\sigma_{\max}(\mathbf{X})$), and the trace of the matrix $\mathbf{X}$ respectively. We denote $\mathcal{D}$ as the full dataset $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, and $\mathcal{D}_{/i}$ as the dataset excluding the $i^{\text{th}}$ observation, i.e. $\mathcal{D}_{/i} = \{(y_j, \mathbf{x}_j) : j \neq i\}$. We also use $[n] := \{1, 2, ..., n\}$ for any positive integer $n$.

For brevity we denote $\ell_j(\boldsymbol{\beta}) := \ell(y_j, \mathbf{x}_j^\top \boldsymbol{\beta})$, and similarly $\phi_j(\boldsymbol{\beta}) := \phi(y_j, \mathbf{x}_j^\top \boldsymbol{\beta})$. Derivatives of these functions are by default taken with respect to the second variable, i.e.

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

Polynomials of $\log(n)$ are denoted by $\mathrm{PolyLog}(n)$ when we do not specify the polynomial itself, but a subscript will be added to distinguish between different polynomials. 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.

The usage of $O(1)$ and $o(1)$ are conventional, and $a_n = \Theta(1)$ iff both $a_n$ and $a_n^{-1}$ are $O(1)$. Similarly, for a sequence of random variables, $X_n = O_p(1)$ means $X_n$ being stochastically bounded, i.e. $\exists C > 0$ s.t. $\mathbb{P}(|X_n| > C) \to 0$. Similarly $X_n = o_p(1)$ means $X_n$ converge to zero in probability, and $X_n = \Theta_p(1)$ iff both $X_n$ and $X_n^{-1}$ are $O_p(1)$. For a closed subset $\boldsymbol{\Theta}$ of $\mathbb{R}^p$, let $\mathcal{C}^k(\boldsymbol{\Theta})$ denote the collection of all functions on $\boldsymbol{\Theta}$ with continuous $k^{\text{th}}$ Fréchet derivative on the interior of $\boldsymbol{\Theta}$. In particular, $\mathcal{C}^0(\boldsymbol{\Theta})$ consists of all continuous functions on $\boldsymbol{\Theta}$.

### 1.4 Organization of Our Paper

The rest of the paper is organized as follows. The assumptions are listed in Section 2.1, followed by discussions on them in Section 2.2. The main theoretical results are stated in Section 2.3. In Section 3, we present simulation experiments that confirm the sharpness and accuracy of our theoretical results. Section 4 discusses the main challenges and proof techniques. The concluding remarks are in Section 5. Detailed proofs of the lemmas are postponed to the Appendix.

## 2 THEORETICAL RESULTS

### 2.1 Assumptions

As we discussed in Section 1.2, our goal is to prove the asymptotic accuracy of LO under the high-dimensional setting for both risk estimation and hyperparameter tuning. We achieve this by first providing a finite-sample error bound for LO. Here, 'finite sample' means that the results are stated explicitly in $n, p$ and their ratio $\gamma_0 = n/p$. In contrast, 'asymptotic results' describe the behavior of a *sequence* of such problems with increasing $n$ and $p$, with $n/p \to \gamma_0 > 0$ or more generally when $n/p$ remain in a fixed interval. Such asymptotic regime has become one of the standard frameworks for studying high-dimensional problems Mousavi et al. (2018); Obuchi and Sakata (2019); Xu et al. (2021); Miolane and Montanari (2021); Donoho et al. (2009, 2011); Patil et al. (2021); Jalali and Maleki (2016); Bellec (2023); Celentano et al. (2023); Li and Wei (2022); Liang and Sur (2022), since it has been able to prove some of the peculiar features estimators exhibit in high-dimensional settings, such as phase transitions Weng et al. (2018).

Before we state our main theorems, we list the assumptions we make and discuss their validity.

A1 $\boldsymbol{\Theta} \subset \mathbb{R}^p$ is a closed convex set.

A2 $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ where $\mathbf{x}_i \in \mathbb{R}^p$ are i.i.d. $N(0, \boldsymbol{\Sigma})$ samples. Moreover, there exist constants $0 < c_X \leq C_X$ such that $p^{-1}c_X \leq \sigma_{\min}(\boldsymbol{\Sigma}) \leq \sigma_{\max}(\boldsymbol{\Sigma}) \leq p^{-1}C_X$.

A3 $r(\boldsymbol{\beta}) = (1-\eta)r_0(\boldsymbol{\beta}) + \eta\boldsymbol{\beta}^\top\boldsymbol{\beta}$ where $\eta \in (0,1)$ and $r_0$ is non-negative, convex and 1-Lipschitz continuous.

A4 Assume that $\ell(y,z)$ and $\phi(y,z)$ are non-negative, convex and continuously differentiable with respect to $z$ and that $\ell$, $\dot{\ell}$, $\phi$ and $\dot{\phi}$ (defined in Section 1.3) grow at most polynomially in $y, z$, i.e., $\exists s > 0$ such that $\forall y, z \in \mathbb{R}$,

$$\begin{aligned}&\max\{|\ell(y,z)|, |\dot{\ell}(y,z)|, |\phi(y,z)|, |\dot{\phi}(y,z)|\}\\ \leq& 1 + |y|^s + |z|^s.\end{aligned}$$

Furthermore, assume that for all integers $0 < m \leq 8s \wedge 4s^2$,

$$\mathbb{E}|y_i|^m \leq C_Y(m).$$

for a sequence $C_Y(m)$ depending only on $m$.

### 2.2 Discussion of the Assumptions

The goal of this section is to provide further intuitions on the assumptions we introduced in the previous section.

First of all, Assumption A1 is standard in convex optimization. In Assumption A2, the Gaussianity of $\mathbf{x}_i$ is prevalent in theoretical papers dealing with high-dimensional problems. As will be clear in the proof of our main theorem, this assumption can be relaxed to the following moment conditions:

$$\mathbb{E}|\mathbf{x}_1^\top \mathbf{a}|^{2t} \le C(t)p^{-t}, \forall \|\mathbf{a}\| \le 1, t = 2, 4, 4s, 8s$$
$$\mathbb{E}\|\mathbf{x}_1\|^8 \le C,$$

where $s$ is the constant defined in Assumption A4. That said, we present our initial result considering Gaussian covariates.

The scaling we have adopted in Assumption A2, i.e.,

$$p^{-1}c_X \le \sigma_{\min}(\mathbf{\Sigma}) \le \sigma_{\max}(\mathbf{\Sigma}) \le p^{-1}C_X,$$

is based on the following rationale. First notice that

$$\frac{c_X}{p}\|\boldsymbol{\beta}^*\|_2^2 \le \mathbb{E}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)^2 \le \frac{C_X}{p}\|\boldsymbol{\beta}^*\|_2^2.$$

Heuristically speaking, when $n, p \to \infty$ with $n/p \to \gamma_0$ and when the elements of $\boldsymbol{\beta}^*$ are $O(1)$, we have $\|\boldsymbol{\beta}^*\|_2 = O(\sqrt{p})$, and hence $\mathbb{E}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)^2 = O(1)$. Therefore, under the settings of the paper we can see that the signal-to-noise ratio (SNR) of each data point, defined as $\frac{\mathrm{var}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)}{\mathrm{var}(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}^*)}$, remains bounded. The reader may check Appendix 6.3 for details. The point of keeping SNR bounded is that, if the SNR grows unboundedly, the noises become negligible for large $n$, so the problem of risk estimation or hyperparameter tuning is not of particular interest; maximum likelihood estimators perform very well. On the other hand, if SNR per data point goes to zero, the accurate estimation of $\widehat{\boldsymbol{\beta}}$ will not be possible unless we impose conditions on $\boldsymbol{\beta}^*$, such as stringent sparsity assumption. However, since we want our results to be generic in terms of $\boldsymbol{\beta}^*$, we do not want to impose any stringent constraints on $\boldsymbol{\beta}^*$.

Another way we can justify the scaling in Assumption A2 is that, under this scaling, neither the loss $\sum_j \ell(y_j, \mathbf{x}_j^\top \boldsymbol{\beta}^*)$ nor the regularizer $\lambda r(\boldsymbol{\beta}^*)$ dominate the other when $n, p \to \infty$. In other words, for a large class of losses $\sum_j \ell(y_j, \mathbf{x}_j^\top \boldsymbol{\beta}^*) = O_p(n)$ and for a large class of regularizers $\lambda r(\boldsymbol{\beta}^*) = O_p(p)$. Given that $n/p \to \gamma_0$, the two terms have the same order. Effectively, this makes the optimal choice of $\lambda$ (that gives the minimum out-of-sample error), $O_p(1)$. For more precise arguments behind this, we refer to the readers Mousavi et al. (2018); Wang et al. (2020, 2022).

Assumption A3 guarantees the loss function to be $2\lambda\eta$-strongly convex. Researchers have noticed that in a wide range of applications, adding $\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}$ in addition to the non-differentiable regularizer often improves the prediction performance, e.g. Hastie et al. (2017); Mazumder et al. (2023); Wang et al. (2022); Guo et al. (2023).

Assumption A3 also posits the Lipschitz continuity of $r_0$, which is a sufficient condition for continuous approximation (see Lemma 4.1). Nearly all popular non-differentiable regularizers satisfy this condition. Appendix 6.4.1 presents several examples including LASSO, generalized LASSO and Schatten norms with the nuclear norm as a special case.

Assumption A4 also holds for a wide range of models that are used in practice. For instance, Appendix 6.4.2 considers standard data generation mechanisms that are used in linear regression, logistic regression and Poisson regression, and show that for all those models Assumption A4 holds.

### 2.3 Main Theorems

The following theorem is the first major contribution of this paper:

**Theorem 2.1.** *Under Assumption A1-A4, we have*

$$\mathbb{E}(\mathrm{LO} - \mathrm{OO})^2 \le \frac{C_v}{n}.$$

*where* $C_v = \frac{24C_\phi^2 C_X^2 C_\ell^{\frac{1}{4}}\gamma_0}{(2\lambda\eta\wedge 1)^2\sqrt{1+C_\beta(2s)}}$ *with*

- $C_\phi^2 = \max\{\sqrt{27(1+C_Y(4s))}, \sqrt{27(4s-1)!!}C_X^s\}$
- $C_\ell = 3^7(1 + C_Y(8s) + (8s-1)!!C_X^{4s}C_\beta(4s))$
- $C_\beta(t) = \left(\frac{\gamma_0}{\lambda\eta}\right)^t A(t)\max\{1+C_Y(st), (1+C_Y(s))^t\}$
- $A(t) = \max\{2, \mathbb{E}P^t\}$ *where* $P \sim \mathrm{Poisson}(1)$.

The main ideas of the proof are presented in Section 4. Below we would like to discuss this theorem and provide some intuition.

*Remark* 2.2. The finite sample result in Theorem 2.1 can be easily translated into an asymptotic one: under the asymptotic setting in which $n, p \to \infty$, $n/p \to \gamma_0$, and assume that all models share the same constants $\{c_X, C_X, \eta, \lambda, s, \{C_Y(m)\}_{m=1}^{8s\wedge 4s^2}\}$ in Assumptions A1-A4, LO offers a consistent estimate of the out-of-sample prediction error in the sense that LO $\to$ OO in probability. However, note that Theorem 2.1 offers more than the consistency, as it captures the convergence rate as well.

*Remark* 2.3. The rate $O(1/n)$ obtained in Theorem 2.1 is expected to be sharp. Note that the leave-one-out cross-validation takes an average of $n$ estimates of the OO, namely $\phi(y_i, \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i})$. The variance of each of these estimates is $O(1)$. Hence, if these terms were independent for all $i$, the variance of LO would be still proportional to $O(1/n)$, which is the same as the bound we have obtained. In reality, the terms tend to be positively correlated due to the overlap of $\widehat{\boldsymbol{\beta}}_{/i}$'s. This implies that we should not expect to obtain a rate better than $O(1/n)$. Our simulation results reported in Section 3 also demonstrate the correctness of this rate.

*Remark* 2.4. The rate $O(1/n)$ has also been seen in the previous work on the analysis of LO under the low-dimensional asymptotic, where $p$ is fixed, while $n \to \infty$. For instance, Burman (1989) provided a rate estimate of the variance of LO (Theorem 6.2(c) in Burman (1989), notations modified) and showed:

$$\text{var}(\text{LO} - \text{OO}) = O(1/n)$$

in the low-dimensional setting. Note that $\text{var}(\text{LO} - \text{OO}) \leq \mathbb{E}(\text{LO} - \text{OO})^2$. Our result shows that the rate is still $O(1/n)$ in high-dimensional settings. More recently, the limiting distribution of $\sqrt{n}(\text{LO}-\text{OO})$ (and more generally for k-fold CV) was studied in Austern and Zhou (2020) under the low-dimensional setting. This scaling translates to the same rate $O(1/n)$ as in Burman (1989).

Theorem 2.1 focuses on the problem of risk estimation. As previously discussed, one important application of risk estimation is its use in hyperparameter tuning. The key question here is whether the hyperparameter tuning technique based on minimizing the LO risk estimate is accurate in high-dimensional settings. Our next result addresses this question. Suppose we aim to find $\lambda_*$ and $\eta_*$, defined as follows:

$$(\lambda^*, \eta^*) = \text{argmin}_{\lambda \in [\lambda_{\min}, \infty), \eta \in [\eta_{\min}, 1)} \, \text{OO}(\lambda, \eta).$$

In order to estimate $(\lambda^*, \eta^*)$ we use LO and obtain

$$(\widehat{\lambda}, \widehat{\eta}) = \text{argmin}_{\lambda \in [\lambda_{\min}, \infty), \eta \in [\eta_{\min}, 1)} \, \text{LO}(\lambda, \eta).$$

The following theorem proves that $(\widehat{\lambda}, \widehat{\eta})$ offer accurate estimates of $(\lambda_*, \eta_*)$.

**Theorem 2.5.** *Suppose that $\lambda^* < \infty$ and $\eta^* < \infty$ and that for every $\epsilon > 0$, we have*

$$\inf_{(\lambda,\eta): |\lambda - \lambda^*| + |\eta - \eta^*| > \epsilon} \text{OO}(\lambda, \eta) > \text{OO}(\lambda^*, \eta^*). \quad (4)$$

*Then, we have that as $n, p \to \infty$ and $n/p \to \gamma_0$*

$$|\widehat{\lambda} - \lambda^*| = o_p(1), \quad |\widehat{\eta} - \eta^*| = o_p(1).$$

The proof of this result is presented in Section 6.6.

*Remark* 2.6. Note that the only additional condition required in Theorem 2.5, which was not needed in Theorem 2.1, is (4). This is a very mild condition, intended to ensure that OO has a unique global minimizer. Without this, the consistency of hyperparameter tuning would clearly be unattainable.

# 3 NUMERICAL EXPERIMENTS

To verify the $O(1/n)$ bound given in Theorem 2.1, we present two numerical experiments: linear and logistic regression, comparing OO and LO using synthetic data. In each experiment there are 4 models: LASSO vs elastic net penalty, with and without positive constraints over the regression coefficients.

## 3.1 Independent Design

In all the examples in this subsection, the rows of $\mathbf{X}$ are iid $N(\mathbf{0}, \mathbf{I}/n)$, and $\phi(y, z) = \ell(y, z)$. We use `scikit-learn` in Python and `glmnet` in `R` to implement linear regression, and our own `R` code to implement logistic regression with the Elastic Net penalty, all available at `https://github.com/RahnamaRad/LO-elnet`.

**Linear Regression:** In our first simulation experiment, we run linear regression with the LASSO and Elastic Net penalties, as well as without and with the positivity constraints on coefficients. We chose 6 values of $p$ from 200 to 2000, equally spaced on the log scale, and $n/p = \gamma_0 = 0.5$ in each case. The first $k = p/5$ coefficients of $\boldsymbol{\beta}^*$ are taken to be 1, and 0 for the rest. We set $\lambda = 1$ and set the ridge factor $\eta = 0$ for LASSO, and $\eta = 0.5$ for the Elastic Net penalty. Figure 1 plots the MSE of LO in estimating OO (based on 400 Monte Carlo samples) against $n$ on $\log - \log$ scale. It is evident that our theoretical rate of $1/n$ for the MSE is indeed correct, for both Elastic Net and LASSO: the regression lines of $\log(\text{MSE})$ vs $\log(p)$, for LASSO and Elastic Net respectively, show a slope of -1.03 and -1.04 for fits without the positivity constraint, and -0.96 and -1.02 for fits with the positivity constraint on coefficients. The 0.95 confidence intervals for the slopes are reported in the figure and contains the value of -1 in all cases. The $R^2$ values of the regression fits are above 0.99 in each case, showing a great fit. We repeated the experiment with $n = 3p/2$ and $p$ increasing from 120 to 1200, setting $\lambda = 0.25$ and set the ridge factor $\eta = 0$ for LASSO, and $\eta = 0.5$ for the Elastic Net penalty. As before, we similarly found that the 0.95 confidence intervals for the slope contain the value -1 in all cases. See Figure 2.

**Logistic Regression:** We also evaluated our results

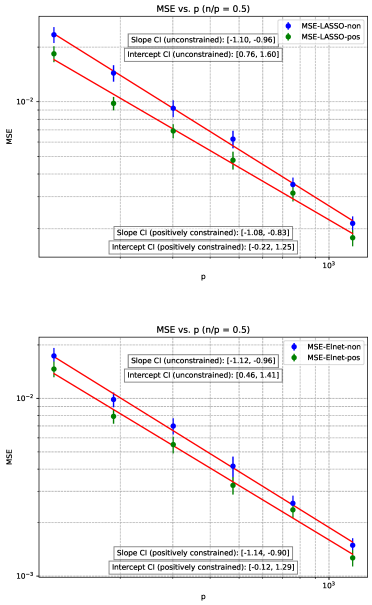


Figure 1: MSE of LO when estimating OO for linear regression with LASSO (top) and Elastic Net (bottom), plotted against $p$ on the log scale. Here $n = p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

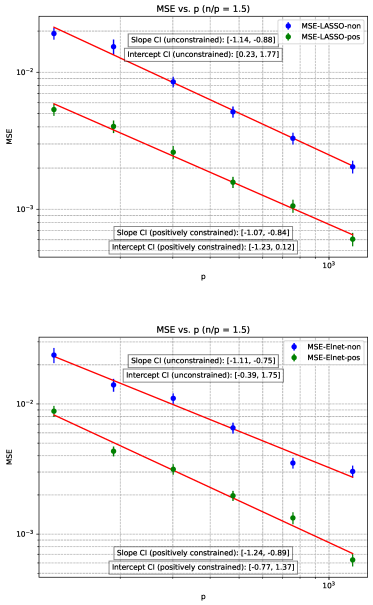


Figure 2: MSE of LO when estimating OO for linear regression with LASSO (top) and Elastic Net (bottom), plotted against $p$ on the log scale. Here $n = 3p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

on logistic regression. In all cases, we set $\lambda = 1$ and set the ridge factor $\eta = 0$ for LASSO, and $\eta = 0.5$ for the Elastic Net penalty. First, we chose $p$ to have 6 values from 360 to 3600, equally spaced on the log scale. Figure 3 plots the MSE of LO in estimating OO, and it again shows that our theoretical rate of $1/n$ for the MSE is indeed correct, for both Elastic Net and LASSO. The regression lines of log(MSE) vs log($p$), for LASSO and Elastic Net respectively, show a slope of -0.80 and -1.16 for fits without the positivity constraint on coefficients, and -0.98 and -0.91 for fits with the positivity constraint. The 0.95 confidence intervals for the slopes are reported in the figure and contain the value -1 in all cases. The $R^2$ values of the regression fits are above 0.97 in each case, showing a great fit for the linear regression of log(MSE) on log($p$). We repeated the experiment with $n = 3p/2$ and $p$ increasing from 120 to 1200. As above, we similarly found that the 0.95 confidence intervals for the slope contain the value -1 in all cases. See Figure 4.

### 3.2 Correlated Design

In this subsection, we explore the robustness of the LO evaluation to correlations among the columns of $\mathbf{X}$. We assume that the rows of $\mathbf{X}$ are iid $N(\mathbf{0}, \boldsymbol{\Sigma}/n)$ where $\boldsymbol{\Sigma}$ is a Toeplitz matrix, with elements $\sigma_{ij} = (1/3)^{|i-j|}$ for $1 \leq i, j \leq p$. We now repeat the experiments from the previous subsection and describe the results.

**Linear Regression:** In our first simulation experiment, we run linear regression with the LASSO and Elastic Net penalties, as well as without and with the positivity constraints on coefficients. We chose 6 values of $p$ from 200 to 2000, equally spaced on the log scale, and $n/p = \gamma_0 = 0.5$ in each case. The parameter settings are identical to the ones for independent design. Figure 5 plots the MSE of LO in estimating OO (based on 200 Monte Carlo samples) against $n$ on $\log - \log$ scale. Even for correlated designs, it is evident that our theoretical rate of $1/n$ for the MSE is indeed correct, for both Elastic Net and LASSO: the regression lines of log(MSE) vs log($p$), for LASSO and Elastic Net respectively, show a slope of -1.07 and -1.04 for fits without the positivity constraint, and -0.98 and -1.15 for fits with the positivity constraint on coefficients. The 0.95 confidence intervals for the slopes are reported in the figure and contains the value of -1 in all cases. The $R^2$ values of the regression fits are above 0.99 in each case, showing a great fit.

### 3.3 Real Data

We evaluate the performance of LO for estimating the out-of-sample error for the IMDb reviews dataset. More specifically, we look at the task of predicting the sentiment of each review using logistic regression over a bag-of-words model for the review text. We used 1039 reviews for training our model, and 48961 reviews for

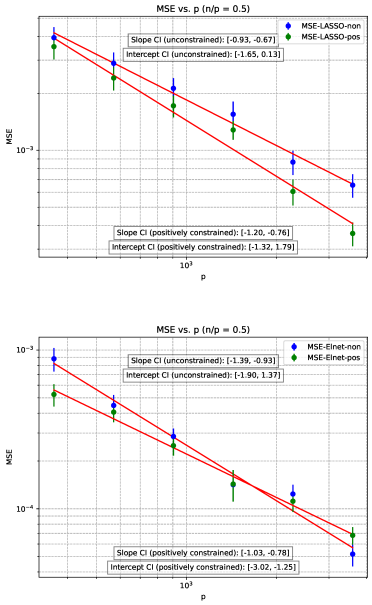


Figure 3: MSE of LO when estimating OO for logistic regression with LASSO (top) and Elastic Net (bottom), plotted against $p$ on the log scale. Here $n = p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

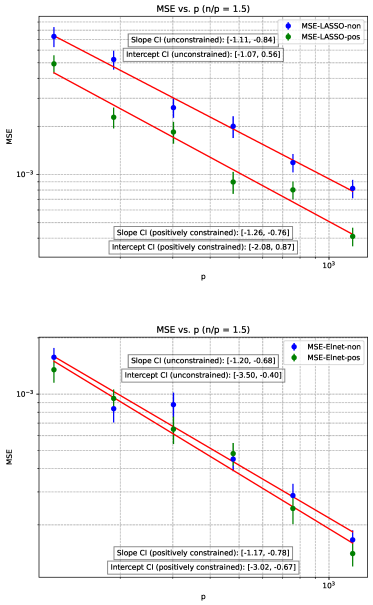


Figure 4: MSE of LO when estimating OO for logistic regression with LASSO (top) and Elastic Net (bottom), plotted against $p$ on the log scale. Here $n = 3p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

calculating the test, i.e., out-of-sample error. Figure 6 plots the negative log likelihood values resulting from a series of choices for the penalty parameter $\lambda$. We used an elastic net penalty with $\eta = 0.5$ as the ridge factor. It is clear that the leave-one-out estimator approximates the test error very well, and the two are virtually indistinguishable for large $\lambda$. We also compared the performance of 5-fold CV, which seems to overestimate the value of test error for all values of $\lambda$. The minimizers of the LO error and the test error are also very close, thus verifying Theorem 2.5 even for this real dataset, where assumptions for our theoretical development are very likely to be not satisfied.

# 4 MAIN THEORETICAL CHALLENGES

In this section, we aim to highlight the technical innovations of our proofs, with a focus on Theorem 2.1 due to space constraints. A summary of the proof for Theorem 2.5, along with the key challenges it addresses, is provided in Section 6.6.

## 4.1 Technical Novelties

We introduce two novel elements that allowed us to significantly broaden the scope of Theorem 2.1 well beyond what was offered by Rahnama Rad et al. (2020): smoothing and projection.

1. **Smoothing**. We start by approximating the non-smooth regularizer $r_0(\boldsymbol{\beta})$ with a smooth function:

$$r_0^{\alpha}(\boldsymbol{\beta}) = \int_{\boldsymbol{\Theta}} r_0(\boldsymbol{\beta} - \mathbf{z})\alpha\phi(\alpha\mathbf{z})d\mathbf{z} \tag{5}$$

where $\phi(\mathbf{z}) = (2\pi)^{-p/2}e^{-\frac{1}{2}\mathbf{z}^\top\mathbf{z}}$ is the density of a standard Gaussian vector.*

As defined in Section 1.3, we denote the collection of k-times differentiable functions on (the interior of) $\boldsymbol{\Theta}$ by $\mathcal{C}^k(\boldsymbol{\Theta})$ ($\mathcal{C}^0(\boldsymbol{\Theta})$ consists of continuous functions on $\boldsymbol{\Theta}$). The following lemma shows that the smooth approximation is accurate.

**Lemma 4.1.** *Suppose $\boldsymbol{\Theta}$ is closed and $r \in \mathcal{C}^0(\boldsymbol{\Theta})$.*

*(a) Suppose there exists a positive integer $k$ such that $r(\mathbf{z})\,\|\mathbf{z}\|_2^k\, e^{-\frac{1}{2}\|\boldsymbol{\beta}-\mathbf{z}\|_2^2}$ is integrable. Then $r^{\alpha}$ defined in (5) is in $\mathcal{C}^k(\boldsymbol{\Theta})$.*

*(b) Suppose there exist constants $L > 0$ and $k \in (0, 1]$ such that for all $\mathbf{x}, \mathbf{y} \in \boldsymbol{\Theta}$,*

$$|r(\mathbf{x}) - r(\mathbf{y})| \le L\,\|\mathbf{x} - \mathbf{y}\|_1^k.$$

*Then as $\alpha \to \infty$, we have*

$$\|r^{\alpha} - r\|_{\infty} \to 0.$$

*Note that when $\boldsymbol{\beta}$ is a $p_1$ by $p_2$ matrix, we concatenate its rows to form a vector of dimension $p = p_1 \times p_2$.

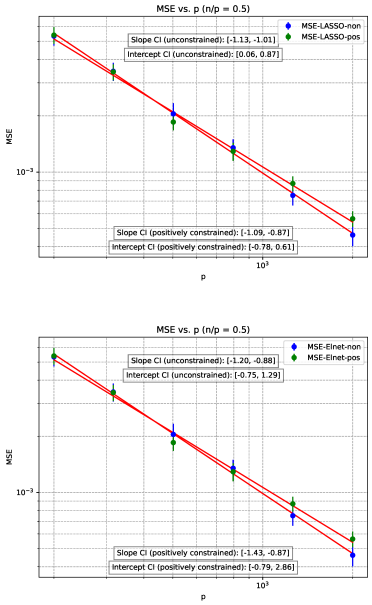


Figure 5: MSE of LO when estimating OO for linear regression with LASSO (top) and Elastic Net (bottom) for correlated design, plotted against $p$ on the log scale. Here $n = p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

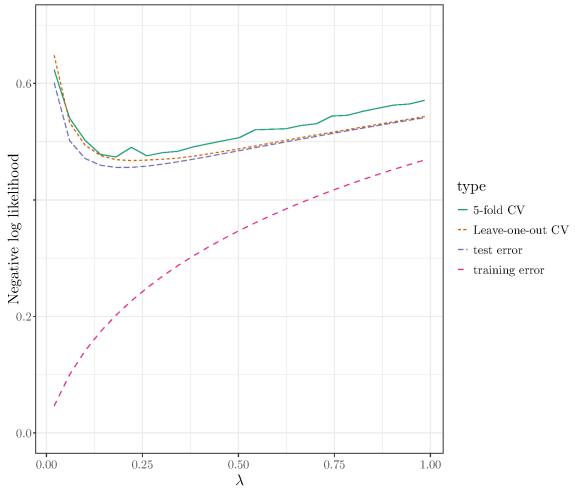


Figure 6: Performance of LO for estimating test error when elastic net penalized logistic regression was used to predict sentiment of reviews in the IMDb Reviews dataset.

The proof of this lemma is presented in Appendix 6.5.2. Note that all assumptions in Lemma 4.1 are satisfied when $r$ is Lipschitz continuous as in Assumption A3. This lemma implies that $r_0^\alpha(\boldsymbol{\beta})$ is an accurate approximation of $r_0(\boldsymbol{\beta})$ for all values of $\boldsymbol{\beta}$. Now if we replace $r_0$ by $r_0^\alpha$ and define

$$\widehat{\boldsymbol{\beta}}^\alpha := \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{j=1}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}$$
$$\widehat{\boldsymbol{\beta}}^\alpha_{/i} := \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{j\neq i} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},$$

we can analyze LO for these surrogate models and infer the properties of LO for the original model, provided that the surrogate estimators are close to the original ones (Lemma 6.13).

2. Projection operator: The second challenge we have to address is the existence of the constraint set $\boldsymbol{\Theta}$. Although one can write the optimization problem $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{i=1}^n \ell_i(\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})$ as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n \ell_i(\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) + \mathbf{I}_{\boldsymbol{\Theta}}(\boldsymbol{\beta})$$

where $\mathbf{I}_{\boldsymbol{\Theta}}(\boldsymbol{\beta})$ is the convex indicator function of the set $\boldsymbol{\Theta}$ [†] and can be absorbed into the regularizer, the problem is that $\mathbf{I}_{\boldsymbol{\Theta}}(\boldsymbol{\beta})$ is not Lipschitz continuous and, in fact, it cannot be uniformly approximated by smooth functions at all. Hence the smoothing argument discussed before will not work. We pick a different approach and represent $\widehat{\boldsymbol{\beta}}^\alpha$ in a different way.

Using a result on the proximal operator of the convex indicator function $\mathbf{I}_{\boldsymbol{\Theta}}$ (Lemma 6.1) , we have that $\widehat{\boldsymbol{\beta}}^\alpha$ should satisfy

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^\alpha &= \mathbf{prox}_{\mathbf{I}_{\boldsymbol{\Theta}}}(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha)) \\ &= \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha)), \end{aligned} \tag{6}$$

where $\boldsymbol{\Pi}_{\boldsymbol{\Theta}}$ is the metric projection operator: $\boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\mathbf{x}) := \underset{\mathbf{z}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \|\mathbf{x}-\mathbf{z}\|_2.$, and $h^\alpha(\boldsymbol{\beta}) := \sum_{j=1}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}$. Note that to obtain the last equality in (6), we have used the fact that $\mathbf{prox}_{\mathbf{I}_{\boldsymbol{\Theta}}} = \boldsymbol{\Pi}_{\boldsymbol{\Theta}}$. Again, this step would be impossible if we didn't have the smooth approximation $r_0^\alpha$. Similarly, we have $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ satisfies:

$$\widehat{\boldsymbol{\beta}}^\alpha_{/i} = \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})),$$

where $h^\alpha_{/i}(\boldsymbol{\beta}) = \sum_{j\neq i} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}$. These representations of $\widehat{\boldsymbol{\beta}}^\alpha$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ turn out to be very helpful in the subsequent analysis, because they allow us to write the difference between $\widehat{\boldsymbol{\beta}}^\alpha$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ as:

$$\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i} = \bar{\mathbf{J}}\left(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \widehat{\boldsymbol{\beta}}^\alpha_{/i} + \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right)$$

[†]The convex indicator function of set $\boldsymbol{\Theta}$ is defined as: $\mathbf{I}_{\boldsymbol{\Theta}}(\boldsymbol{\beta}) = 0$ if $\boldsymbol{\beta} \in \boldsymbol{\Theta}$ and $\mathbf{I}_{\boldsymbol{\Theta}}(\boldsymbol{\beta}) = +\infty$ otherwise.

for a matrix $\bar{\mathbf{J}} \in \mathbb{R}^{p\times p}$ that acts as a Jacobian of the projection operator and has eigenvalues between 0 and 1 (Lemma 6.2). This equation is crucial in bounding $\|\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i}\|_2$ and will be useful in our downstream analysis.

### 4.2 Proof Summary of Theorem 2.1

In this subsection, we present a brief sketch of the proof. Similar to the proof of Theorem 1 in Rahnama Rad et al. (2020) we use the following two definitions:

$$V_1 := \mathrm{LO} - \frac{1}{n}\sum_{i\in[n]} \mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}],$$
$$V_2 := \frac{1}{n}\sum_{i\in[n]} \mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] - \mathrm{OO}, \tag{7}$$

and use the following upper bound:

$$\mathbb{E}\left[\mathrm{LO} - \mathrm{OO}\right]^2 = \mathbb{E}\left(V_1 + V_2\right)^2 \le 2\mathbb{E}V_1^2 + 2\mathbb{E}V_2^2. \tag{8}$$

Hence, the remaining steps are to obtain upper bounds for $\mathbb{E}V_1^2$ and $\mathbb{E}V_2^2$. To see how the ideas we introduced in Section 4.1 enable us to obtain the required upper bound, we outline the details of proving an upper bound for $\mathbb{E}V_2^2$ here. The detailed proof can be found in the Appendix 6.5.1.

To bound $\mathbb{E}V_2^2$, first notice that for all $i \in [n]$ we have

$$\mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] = \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] = \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}].$$

By the mean-value theorem, for each $i$ there exists a random variable $\boldsymbol{\xi}_i = t_i\widehat{\boldsymbol{\beta}}_{/i} + (1-t_i)\widehat{\boldsymbol{\beta}}$ with $t_i \in [0,1]$ such that

$$\phi_0(\widehat{\boldsymbol{\beta}}_{/i}) - \phi_0(\widehat{\boldsymbol{\beta}}) = \dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}).$$

Then we have

$$\begin{aligned}
\mathbb{E}(V_2^2) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}})|\mathcal{D}]\right)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}] - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}})|\mathcal{D}]\right)^2 \\
&\le \mathbb{E}\Big(\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_1)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\Big)^2
\end{aligned}$$

where we use Cauchy Schwarz inequality in the final step. Next we have

$$\begin{aligned}
&\left|\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\right| \\
\le&\sqrt{\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]}\sqrt{\frac{C_X}{p}\|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}\|^2}
\end{aligned}$$

where the last inequality uses the independence between $\mathbf{x}_0$ and $\mathcal{D}$, and Assumption A2. Hence, in order to bound $\mathbb{E}(V_2^2)$, we have to obtain bounds on $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\|$ and also $\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]$. The following lemma connects $\widehat{\boldsymbol{\beta}}_{/i}$ and $\widehat{\boldsymbol{\beta}}$:

**Lemma 4.2.** *Under assumptions A1-A3, we have for all $\alpha > 0$ that*

$$\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\right\| \le \frac{|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}_{/i})|\,\|\mathbf{x}_i\|}{2\lambda\eta \wedge 1}.$$

The proof of this lemma is presented in Section 6.5.3 and uses the ideas that we mentioned in Section 2, i.e. smoothing and projection operator. Using moment bounds and Assumption A4 one can bound $\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]$. Details can be found in Lemma 6.9. Combining all the above inequalities, we have:

$$\mathbb{E}V_2^2 \le \frac{1}{n}\frac{\sqrt{8}(24)^{\frac{1}{4}}C_\phi^2 C_X^2 C_\ell^{1/4}\gamma_0}{(2\lambda\eta\wedge 1)^2}\sqrt{1+C_\beta(2s)} := \frac{C_{v2}}{n}. \tag{9}$$

Obtaining an upper bound for $\mathbb{E}(V_1^2)$ uses similar techniques, although it involves more cumbersome calculations. In fact we can prove that

**Lemma 4.3.** *Under assumptions A1-A4, for large enough $n, p$ we have: $\mathbb{E}V_1^2 \le \frac{C_{v1}}{n}$ where*

$$C_{v1} = \frac{\sqrt{6}(24)^{\frac{1}{4}}C_\phi^2 C_X^2 C_\ell^{1/4}\gamma_0}{(2\lambda\eta\wedge 1)^2}\sqrt{1+C_\beta(2s)}$$

*depends only on $\{\gamma_0, \lambda, \eta, C_Y(\cdot), s, C_X\}$.*

The proof of Lemma 4.3 is given in Section 6.5.6. Now using (17) and (8) finishes the proof of Theorem 2.1.

## 5 CONCLUDING REMARKS

In this paper, our focus was on the class of regularized empirical risk minimization (R-ERM) techniques that incorporate non-differentiable regularizers. We studied the accuracy of leave-one-out cross-validation techniques within a high-dimensional setting, where both the number of observations $n$ and the number of features $p$ are large while the ratio $n/p$ is bounded. We derived a finite-sample upper bound for the difference between the out-of-sample prediction error, OO, and its leave-one-out cross-validation estimate LO. Our upper bound shows that $\mathbb{E}(\mathrm{LO} - \mathrm{OO})^2 = O(\frac{1}{n})$. Our simulations confirmed the sharpness of this theoretical result. Finally, we used the upper bound on the MSE of LO to establish the consistency of the hyperparameter tuning scheme based on minimizing LO in the high-dimensional asymptotic setting, where $n, p \to \infty$ and $n/p \to \gamma_0$.

## References


Arnab Auddy, Haolin Zou, Kamiar Rahnama Rad, and Arian Maleki. Approximate leave-one-out cross validation for regression with $\ell_1$ regularizers (extended version). *arXiv preprint arXiv:2310.17629*, 2023.

Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

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

Pierre C Bellec. Out-of-sample error estimation for m-estimators with convex penalty. *Information and Inference: A Journal of the IMA*, 12(4):2782–2817, 2023.

Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005. doi: 10.1137/050626090.

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.

Khalid El-Arini, Min Xu, Emily B Fox, and Carlos Guestrin. Representing documents through their readers. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining*, pages 14–22, 2013.

Brian R Gaines, Juhyun Kim, and Hua Zhou. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018.

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, 2019.

Adityanand Guntuboyina and Bodhisattva Sen. Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594, 2018.

Yilin Guo, Haolei Weng, and Arian Maleki. Signal-to-noise ratio aware minimaxity and higher-order asymptotics. *IEEE Transactions on Information Theory*, 2023.

Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

Zonghui Hu, Dean A Follmann, and Kazutoyo Miura. Vaccine design via nonnegative lasso-based variable selection. *Statistics in medicine*, 34(10):1791–1798, 2015.

R. Ibragimov and Sh. Sharakhmetov. The best constant in the rosenthal inequality for nonnegative random variables. *Statistics & Probability Letters*, 55(4):367–376, 2001. doi: https://doi.org/10.1016/S0167-7152(01)00134-1.

Shirin Jalali and Arian Maleki. New approach to bayesian high-dimensional linear regression. *Information and Inference: A Journal of the IMA*, 7, 07 2016.

Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-$\ell_1$-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669–1695, 2022.

Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

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

Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *Operations Research*, 71(1):129–147, 2023.

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

Parth T Nobel, Daniel LeJeune, and Emmanuel J Candès. Randalo: Out-of-sample risk estimation in no time flat. *arXiv preprint arXiv:2409.09781*, 2024.

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.

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, 2022a.

Pratik Patil, Yuchen Wu, and Ryan Tibshirani. Failures and successes of cross-validation for early-stopped gradient descent in high-dimensional least squares. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 02–04 May 2024.

Pratik V. Patil, Alessandro Rinaldo, and Ryan J. Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2022b. URL `https://api.semanticscholar.org/CorpusID:248923838`.

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.

H.P. Rosenthal. On the subspaces of $L_p(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8:273–303, 1970. doi: 10.1007/BF02771562.

Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. 2016.

Will Stephenson, Zachary Frangella, Madeleine Udell, and Tamara Broderick. Can we globally optimize cross-validation loss? quasiconvexity in ridge regression. *Advances in Neural Information Processing Systems*, 34:24352–24364, 2021.

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.

Robert Tibshirani and Xiaotong Suo. An ordered lasso and sparse time-lagged regression. *Technometrics*, 58(4):415–423, 2016.

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.

Ming Yuan and Yi Lin. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 12 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00532.x.

Hongke Zhao, Le Wu, Qi Liu, Yong Ge, and Enhong Chen. Investment recommendation in p2p lending: A portfolio perspective with risk management. In *2014 IEEE International Conference on Data Mining*, pages 1109–1114. IEEE, 2014.

## Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes,

No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

**In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.**

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] See Section 2.1 and Section 3.
   (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. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 2.1.
   (b) Complete proofs of all theoretical results. [Yes] See supplement.
   (c) Clear explanations of any assumptions. [Yes] See Section 2.2.

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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See details in experiments section.
   (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] See details of experiments in Section 3.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

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. [Not Applicable]
   (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]

# 6 SUPPLEMENTARY MATERIALS

## 6.1 Additional Numerical Experiments

**Logistic regression with correlated design:** We now present the results of our experiments on logistic regression with a correlated design. We allow the columns of the design matrix $\mathbf{X}$ to be correlated. We assume that the rows of $\mathbf{X}$ are iid $N(\mathbf{0}, \mathbf{\Sigma}/n)$ where $\mathbf{\Sigma}$ is a Toeplitz matrix, with elements $\sigma_{ij} = (1/3)^{|i-j|}$ for $1 \le i, j \le p$. We now repeat the experiments from the previous subsection and describe the results.

In all cases, we set $\lambda = 1.2$ and set the ridge factor $\eta = 0$ for LASSO, and $\eta = 0.5$ for the Elastic Net penalty. First, we chose $p$ to have 6 values from 360 to 3600, equally spaced on the log scale. Figure 7 plots the MSE of LO in estimating OO, and it again shows that our theoretical rate of $1/n$ for the MSE is indeed correct, for both Elastic Net and LASSO. The regression lines of log(MSE) vs log($p$), for LASSO and Elastic Net respectively, show a slope of -0.79 and -0.83 for fits without the positivity constraint on coefficients, and -0.86 and -0.88 for fits with the positivity constraint. The 0.95 confidence intervals for the slopes are reported in the figure and are very close to the value -1 in all cases. The $R^2$ values of the regression fits are above 0.97 in each case, showing a great fit for the linear regression of log(MSE) on log($p$).

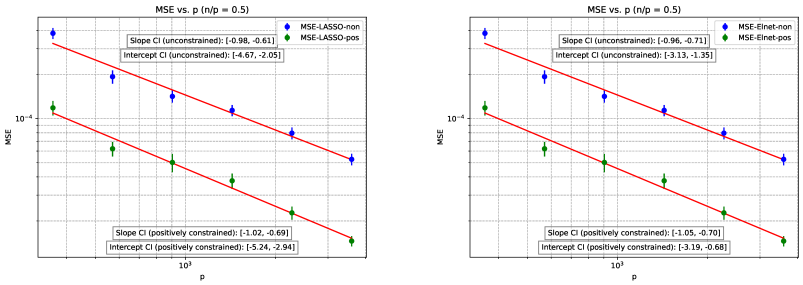


Figure 7: MSE of LO when estimating OO for logistic regression with LASSO (top) and Elastic Net (bottom) for correlated design, plotted against $p$ on the log scale. Here $n = p/2$ and $k/p = 0.2$ where $k$ is the no. of active elements of $\boldsymbol{\beta}^*$.

## 6.2 Technical Lemmas

To improve the readability of the rest of the manuscript, we include several standard technical results used in our detailed proofs which are presented in the rest of the Appendix.

**Lemma 6.1.** *Suppose*

- $R : \mathbb{R}^p \to \mathbb{R} \cup \{-\infty, \infty\}$ *is proper convex with* $dom(R)$ *being closed.*
- $L : \mathbb{R}^p \to \mathbb{R}$ *is differentiable on the relative interior of* $dom(R)$ *and proper convex on* $dom(R)$.

*Define the proximal operator of the function $R$ as*

$$\mathbf{prox}_R(\mathbf{u}) := \text{argmin}_{\mathbf{x}} \left\{ R(\mathbf{x}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}.$$

*Then a solution of the following equation:*

$$\mathbf{x} = \mathbf{prox}_R(\mathbf{x} - \nabla L(\mathbf{x})) \tag{10}$$

*is also a minimizer of the problem*

$$\min_{\mathbf{x}} \{L(\mathbf{x}) + R(\mathbf{x})\} \tag{11}$$

*and vice versa.*

*Proof.* See Propositon 3.1 (iii)(b) in Combettes and Wajs (2005). □

**Lemma 6.2.** *Suppose $\mathbf{\Theta}$ is a closed convex set in $\mathbb{R}^p$, and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Let $\mathbf{\Pi}_{\mathbf{\Theta}}$ be the metric projection onto $\mathbf{\Theta}$. Then there exists a matrix-valued function $\mathbf{J}(t)$ with*

$$0 \leq \lambda_{\min}(\mathbf{J}(t)) \leq \lambda_{\max}(\mathbf{J}(t)) \leq 1, \quad \forall t \in [0,1]$$

*such that*

$$\mathbf{\Pi}_{\mathbf{\Theta}}(\mathbf{y}) - \mathbf{\Pi}_{\mathbf{\Theta}}(\mathbf{x}) = \int_0^1 \mathbf{J}(t)dt(\mathbf{y}-\mathbf{x}).$$

*Proof of Lemma 6.2.* Fix $\mathbf{x}, \mathbf{y}$ and consider $f(t) := \mathbf{\Pi}_{\mathbf{\Theta}}((1-t)\mathbf{x}+t\mathbf{y})$. By the firm non-expansiveness of projection operators, we have $\forall t_1, t_2 \in [0,1]$,

$$\begin{aligned}
&0 \leq (f(t_1) - f(t_2))^2 \\
=&\|\mathbf{\Pi}_{\mathbf{\Theta}}((1-t_1)\mathbf{x}+t_1\mathbf{y}) - \mathbf{\Pi}_{\mathbf{\Theta}}((1-t_2)\mathbf{x}+t_2\mathbf{y})\|^2 \\
\leq& \langle \mathbf{\Pi}_{\mathbf{\Theta}}((1-t_1)\mathbf{x}+t_1\mathbf{y}) - \mathbf{\Pi}_{\mathbf{\Theta}}((1-t_2)\mathbf{x}+t_2\mathbf{y}), (t_1-t_2)(\mathbf{y}-\mathbf{x})\rangle \\
=&\langle f(t_1) - f(t_2), (t_1-t_2)(\mathbf{y}-\mathbf{x})\rangle \\
\leq&(t_1-t_2)^2\|\mathbf{y}-\mathbf{x}\|^2. \qquad (12)
\end{aligned}$$

By taking a square root, we thus have $f(t)$ is $\|\mathbf{x}-\mathbf{y}\|$ Lipschitz on $[0,1]$. Therefore it is absolutely continuous and differentiable almost everywhere. Let $t_1 = t+\epsilon$ and $t_2 = t$, and then divide (12) by $\epsilon^2$, then if $f'(t)$ exists we have:

$$0 \leq \|f'(t)\|^2 \leq \langle f'(t), \mathbf{y}-\mathbf{x}\rangle \leq \|\mathbf{y}-\mathbf{x}\|^2$$

For each $t$, consider solving the following equation for $\mathbf{J}(t)$:

$$f'(t) = \mathbf{J}(t)(\mathbf{y}-\mathbf{x}). \qquad (13)$$

The claim is that we can find $\mathbf{J}(t)$ with its eigenvalues between $[0,1]$ such that (13) holds. In fact, since the rotation in $\mathbb{R}^p$ is a unitary matrix with all eigenvalues being 1, we can choose $\mathbf{J}(t) = \frac{\|f'(t)\|}{\|\mathbf{y}-\mathbf{x}\|}\mathbf{R}$ where $\mathbf{R}$ is a rotation that rotates $\mathbf{y}-\mathbf{x}$ to the direction of $f'(t)$, and this choice of $\mathbf{J}(t)$ has its all eigenvalues being $\frac{\|f'(t)\|}{\|\mathbf{y}-\mathbf{x}\|} \in [0,1]$.

With a slight abuse of notation we assume $f'(t) = \mathbf{0}$ wherever the derivative does not exist. Then by the Newton-Leibniz formula (for Lebesgue integral and absolute continuous functions):

$$\begin{aligned}
\mathbf{\Pi}_{\mathbf{\Theta}}(\mathbf{y}) - \mathbf{\Pi}_{\mathbf{\Theta}}(\mathbf{x}) &= f(1) - f(0) \\
&= \int_0^1 f'(t)dt \\
&= \int_0^1 \mathbf{J}(t)(\mathbf{y}-\mathbf{x})dt
\end{aligned}$$

where the last line uses (13). □

*Remark* 6.3. In the proof of Lemma 6.2, the matrix $\mathbf{J}(t)$ is not explicitly derived. However by equation (13) one can easily see that, when $\mathbf{\Pi}_{\mathbf{\Theta}}((1-t)\mathbf{x}+t\mathbf{y})$ is smooth at $t = t_0$, we can use its Jacobian as $\mathbf{J}(t)$, i.e.

$$\mathbf{J}(t_0) = \frac{\partial}{\partial \mathbf{z}}\mathbf{\Pi}_{\mathbf{\Theta}}(\mathbf{z})\Big|_{\mathbf{z}=(1-t_0)\mathbf{x}+t_0\mathbf{y}}.$$

**Lemma 6.4** (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 6.5.** *Let $\gamma_0 = n/p$ as usual. For $p \geq 2$ we have*

*(a)*

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

*(b)*

$$\mathbb{E}\|\mathbf{x}_i\|^8 \le 24C_X^4$$

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

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

The last line uses Lemma 6.4. Finally a union bound over all $1 \le i \le n$ finishes the proof for this part.

(b) Let $\mathbf{z} = \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{x}_i$ then $\mathbf{z} \sim N(0, \mathbf{I}_p)$ and $\|\mathbf{z}\|^2 \sim \chi^2(p)$. By standard results on $\chi^2$ distribution we have

$$\mathbb{E}\|\mathbf{z}\|^8 = p(p+2)(p+4)(p+6)$$

. Then we have for $p \ge 2$:

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_i\|^8 &= \mathbb{E}(\mathbf{z}^\top \mathbf{\Sigma} \mathbf{z})^4 \\ &\le \mathbb{E}\left(\frac{C_X}{p}\|\mathbf{z}\|^2\right)^4 \\ &\le \frac{C_X^4}{p^4}\mathbb{E}\|\mathbf{z}\|^8 \\ &\le C_X^4\left(1+\frac{2}{p}\right)\left(1+\frac{4}{p}\right)\left(1+\frac{6}{p}\right) \\ &\le 24C_X^4.\end{aligned}$$

□

**Lemma 6.6.** *Suppose $(y_1, \cdots, y_n)$ are i.i.d. and $\mathbb{E}|y_1|^{8s} < \infty$. Then for $s > 0$ we have:*

*(a) denote $s_k = \mathbb{E}(|y_1|^s - \mu)^k < \infty$ for $k \le 4$, then*

$$\mathbb{E}\left[\frac{1}{n}\sum_i |y_i|^s - \mathbb{E}|y_1|^s\right]^8 \le \frac{c_1}{n^4}$$

*where $c_1 = \frac{1}{27}s_8 + \frac{28}{9}s_2s_6 + \frac{35}{9}s_4^2 + 70s_2^2s_4 + \frac{280}{3}s_3^2s_2 + 105s_2^4$.*

*(b) denote $s'_k = \mathbb{E}(|y_1|^{2s} - \mu)^k < \infty$ for $k \le 4$, then*

$$\mathbb{E}\left[\frac{1}{n}\sum_i |y_i|^{2s} - \mathbb{E}|y_1|^{2s}\right]^8 \le \frac{c_2}{n^4}$$

*where $c_2 = \frac{1}{27}s'_8 + \frac{28}{9}s'_2s'_6 + \frac{35}{9}(s'_4)^2 + 70(s'_2)^2s'_4 + \frac{280}{3}(s'_3)^2s'_2 + 105(s'_2)^4$.*

*Proof.* (a) For notational simplicity we denote $\bar{y} := \frac{1}{n}\sum_i |y_i|^s$, $\mu = \mathbb{E}|y_1|^s$. Then we have, for $n \ge 3$:

$$\begin{aligned}\mathbb{E}(\bar{y} - \mu)^8 &= \frac{1}{n^8}\mathbb{E}[\sum_i (|y_i^s| - \mu)]^8 \\ &= \frac{1}{n^8}\left[ns^8 + \binom{8}{2}n(n-1)s_2s_6 + \frac{1}{2}\binom{8}{4}n(n-1)s_4^2 + \frac{1}{2}\binom{8}{4}\binom{4}{2}n(n-1)(n-2)s_2^2s_4\right. \\ &\qquad \left. + \frac{1}{2}\binom{8}{3}\binom{5}{3}n(n-1)(n-2)s_3^2s_2 + \frac{1}{4!}\binom{8}{2}\binom{6}{2}\binom{4}{2}n(n-1)(n-2)(n-3)s_2^4\right] \\ &\le \frac{1}{n^4}\left(\frac{1}{27}s_8 + \frac{28}{9}s_2s_6 + \frac{35}{9}s_4^2 + 70s_2^2s_4 + \frac{280}{3}s_3^2s_2 + 105s_2^4\right) \\ &:= \frac{c_1}{n^4}.\end{aligned}$$

The last line uses the fact that $\frac{n}{n^4} \le \frac{1}{27}$, $\frac{n(n-1)}{n^4} \le \frac{1}{9}$ and so on, since we assume $n \ge 3$.

(b) The proof is essentially the same as that of part (a) (by replacing $y_i$ with $y_i^2$).

□

**Lemma 6.7** (Rosenthal (1970), Ibragimov and Sharakhmetov (2001)). *Let $\{X_i\}_{i\in[n]}$ be a sequence of independent non-negative random variables, and $t \geq 1$. Then we have*

$$\mathbb{E}\left(\sum_{i\in[n]} X_i\right)^t \leq A(t)\max\left\{\sum_{i\in[n]}\mathbb{E}X_i^t, \left(\sum_{i\in[n]}\mathbb{E}X_i\right)^t\right\},$$

*where the optimal choice of $A(t)$ is $A(t) = 2, 1 \leq t \leq 2$ and $A(t) = e^{-1}\sum_{k=0}^{\infty}\frac{k^t}{k!} = \mathbb{E}P^t, t > 2$ where $P \sim Poisson(1)$.*

*Proof.* See Theorem 3 in Rosenthal (1970). The optimal choice of $A(t)$ was given in Ibragimov and Sharakhmetov (2001), Theorem 1. [‡] □

### 6.3 Bounded Signal-to-noise Ratio

First we explain what we mean by 'bounded SNR' in Section 2.1, using the three examples of linear, logistic and Poisson regression (with log exponential link):

- Linear: $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i^\top\boldsymbol{\beta}^*, \sigma^2)$
- Logistic: $y_i|\mathbf{x}_i \sim \text{Binomial}((1+e^{-\mathbf{x}_i^\top\boldsymbol{\beta}^*})^{-1})$
- Poisson: $y_i|\mathbf{x}_i \sim \text{Poisson}(\log(1+e^{\mathbf{x}_i^\top\boldsymbol{\beta}^*}))$

Define the signal-to-noise ratio as

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

When $\|\boldsymbol{\beta}^*\|_2 = O(\sqrt{p})$ or each elements of of $\boldsymbol{\beta}^*$ is $O(1)$, by Assumption A2 we have

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

On the other hand,

$$[\text{var}(y_i|\mathbf{x}_i)]^{-1} = \begin{cases}\sigma^{-2}, & \text{Linear}\\ (e^{-\frac{1}{2}\mathbf{x}_i^\top\boldsymbol{\beta}^*} + e^{\frac{1}{2}\mathbf{x}_i^\top\boldsymbol{\beta}^*})^2, & \text{Logistic}\\ [\log(1+e^{\mathbf{x}_i^\top\boldsymbol{\beta}^*})]^{-1}, & \text{Poisson}\end{cases}$$

They are all $O_p(1)$ when $n, p$ increase. To see this, notice that

$$\mathbf{x}_i^\top\boldsymbol{\beta}^* \sim N(0, \boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}\boldsymbol{\beta}^*)$$

and

$$4 \leq (e^{-\frac{1}{2}z} + e^{\frac{1}{2}z})^2 \leq 2(e^{|z|}+1),$$
$$(\log(2) + z_+)^{-1} \leq [\log(1+e^z)]^{-1} \leq z_+^{-1}$$

where $z_+ = z$ when $z \geq 0$ and 0 otherwise. Suppose $C_1 \leq p^{-1/2}\|\boldsymbol{\beta}^*\| \leq C_2$ for some constants $C_1, C_2$, then $\mathbf{x}_i^\top\boldsymbol{\beta}^* = \Theta_p(1)$ and so is the ratio $\frac{\text{var}(\mathbf{x}_i^\top\boldsymbol{\beta}^*)}{\text{var}(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}^*)}$.

[‡] When $t \geq 1$ is an integer, $A(t)$ is called the Bell number.

### 6.4 Discussion of the Assumptions

#### 6.4.1 On Assumption A3

In this subsubsection we present several commonly used regularizers in machine learning, and show that they are all Lipshitz continuous.

*Example* 6.1 (LASSO). The classic LASSO penalty is $r_0(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and is clearly non-negative, convex and Lipschitz continuous.

*Example* 6.2 (Group LASSO). The group LASSO was introduced in Yuan and Lin (2005) to achieve joint variable selection among different data groups. Assume the features are partitioned in $G$ groups $J_1, J_2, \ldots, J_G$, and the penalty takes the form $r_0(\boldsymbol{\beta}) = \sum_{j=1}^{G} (\boldsymbol{\beta}_{J_j}^\top \mathbf{K}_j \boldsymbol{\beta}_{J_j})^{1/2}$ where $\boldsymbol{\beta}_{J_j} \in \mathbb{R}^{p_j}$ is the coefficient vector for group $j$, $\mathbf{K}_j \in \mathbb{R}^{p_j \times p_j}$ is a positive definite matrix, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{J_1}^\top, \cdots, \boldsymbol{\beta}_{J_G}^\top)^\top \in \mathbb{R}^p$ is the concatenated coefficient with $p = \sum_{j=1}^{G} p_j$.

Clearly $r_0$ is non-negative. Moreover it is also convex since it is a sum of convex functions $r_{0j}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_{J_j}^\top \mathbf{K}_j \boldsymbol{\beta}_{J_j})^{1/2}$. Finally, $r_0$ is also $\sqrt{\sum_{k=1}^{J} \sigma_{\max}(K_j)}$-Lipschitz with respect to $\boldsymbol{\beta}$.

*Example* 6.3 (Generalized LASSO). The generalized LASSO penalty takes the form of $r_0(\boldsymbol{\beta}) = \|\mathbf{D}\boldsymbol{\beta}\|_1$ and encompasses many LASSO type penalty such as LASSO ($\mathbf{D} = \mathbf{I}_p$) and fused LASSO ($\mathbf{D} = (d_{ij})_{(p-1)\times p}$ where $d_{ij} = 1$ if $i = j$, $d_{ij} = -1$ if $j = i + 1$ and $d_{ij} = 0$ otherwise). The nonnegativity and convexity are immediate since $r_0$ is the $\ell_1$ norm of $\mathbf{D}\boldsymbol{\beta}$. Moreover, it follows that $\|\mathbf{D}\boldsymbol{\beta}\|_1$ is $\sigma_{\max}(\mathbf{D})$-Lipschitz in $\boldsymbol{\beta}$.

*Example* 6.4 (Nuclear norm and Schatten norms). When the estimand is a matrix $\mathbf{B}$ with rank $d$, the nuclear norm is a popular regularizer:

$$r_0(\mathbf{B}) = \sum_{i=1}^{d} \sigma_i(\mathbf{B})$$

where $\sigma_i(\mathbf{B})$ is the $i$-th largest singular value of $\mathbf{B}$. More generally, the Schatten norm of $\mathbf{B}$ is defined as

$$r_0(\mathbf{B}) = \left( \sum_{i=1}^{d} \sigma_i^p(\mathbf{B}) \right)^{\frac{1}{p}}$$

and takes the nuclear norm as a special case when $p = 1$. Since $r_0(\mathbf{B})$ is a norm on the singular values of $\mathbf{B}$, we obtain the nonnegativity and convexity of $r_0$. To show the Lipschitz continuity we argue as follows. Since $r_0(\mathbf{B})$ is the p-norm of the vector $(\sigma_1(\mathbf{B}), \cdots, \sigma_d(\mathbf{B}))$ and using the triangular inequality we have

$$\begin{aligned} |r_0(\mathbf{x}) - r_0(\mathbf{y})| &\le r_0(\mathbf{x} - \mathbf{y}) \\ &\le \begin{cases} K^{\frac{1}{p}-\frac{1}{2}} \|\mathbf{x} - \mathbf{y}\|_2 & \text{if } 1 \le p \le 2 \\ \|\mathbf{x} - \mathbf{y}\|_2 & \text{if } p \ge 2 \end{cases} \\ &\le \sqrt{K} \|\mathbf{x} - \mathbf{y}\|_2 , \end{aligned}$$

where $\|\mathbf{B}\|_2$ denotes the Frobenius norm of $\mathbf{B}$. The second line uses the relationship between p-norms: for $\mathbf{u} \in \mathbb{R}^p$,

$$\begin{aligned} \|\mathbf{u}\|_p &= \left( \sum_{k=1}^{K} u_k^p \right)^{\frac{1}{p}} \\ &\le \begin{cases} K^{\frac{1}{p}-\frac{1}{2}} \|\mathbf{u}\|_2 & \text{if } 1 \le p \le 2 \\ \|\mathbf{u}\|_2 & \text{if } p \ge 2. \end{cases} \end{aligned}$$

Then if we let $\mathbf{u} = (\sigma_1(\mathbf{B}), \cdots, \sigma_K(\mathbf{B}))^\top$ we then have $\|\mathbf{u}\|_p = r_0(\mathbf{B})$ and

$$\|\mathbf{u}\|_2 = \left( \sum_{k=1}^{K} \sigma_k^p(\mathbf{B}) \right)^{\frac{1}{2}} = \sqrt{\mathrm{tr}(\mathbf{B}^\top \mathbf{B})} = \|\mathbf{B}\|_2.$$

### 6.4.2 On Assumption A4

In this subsection, we show that the moment bound of $y_i$ and polynomial growth of $\ell, \dot{\ell}, \phi, \dot{\phi}$ in Assumption A4 are justified for many popular data generating mechanisms. We show this for three examples: linear, logistic and Poisson regression (with log exponential link). In this subsubsection we assume $n/p = \gamma_0 > 0$ and Assumption A2 hold, i.e. $\mathbf{x}_i$ are i.i.d. $N(0, \mathbf{\Sigma})$ with $\sigma_{\max}(\mathbf{\Sigma}) \leq C_X/p$. In addition we assume $p^{-1/2}\|\boldsymbol{\beta}^*\| \leq \xi$ for some $\xi > 0$. Finally, both $\ell$ and $\phi$ are set to the negative log-likelihood.

*Example* 6.5 (Linear). Suppose $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}^*, \sigma^2)$. Then $y_i \sim N(0, \tilde{\sigma}^2)$ where $\tilde{\sigma}^2 = \sigma^2 + \boldsymbol{\beta}^{*\top}\mathbf{\Sigma}\boldsymbol{\beta}^*$. Using standard results on Gaussian moments, we have

$$\mathbb{E}|y|^m = \tilde{\sigma}^m \times \frac{2^{m/2}\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi}} \leq \left(\sigma^2 + C_X\xi^2\right)^{m/2} \frac{2^{m/2}\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi}} \text{ for any } m \geq 0.$$

For Gaussian linear model, the negative log-likelihood is known to be the $\ell_2$ loss:

$$\ell(y, z) = \frac{1}{2}(y - z)^2 \leq y^2 + z^2,$$

and $|\dot{\ell}(y, z)| = |z - y| \leq |z| + |y|$.

*Example* 6.6 (Logistic). Suppose $y_i|\mathbf{x}_i \sim \text{Bernoulli}(p)$ where $p = (1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}^*})^{-1}$. Then $y_i \in \{0, 1\}$ and obviously

$$\mathbb{E}|y|^m \leq 1 \text{ for all } m \geq 0.$$

The negative log-likelihood of this model is

$$\ell(y, z) = y\log(1 + e^{-z}) + (1 - y)\log(1 + e^z), \quad y \in \{0, 1\},$$

and

$$|\ell(y, z)| \leq 2\log(2) + 2|z|.$$

For its derivative,

$$|\dot{\ell}(y, z)| = \left|\frac{e^z}{1 + e^z} - y\right| \leq 1 + |y|.$$

*Example* 6.7 (Poisson). Suppose $y_i|\mathbf{x}_i \sim \text{Poisson}(\mu)$ with $\mu = \log(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}^*})$. Then $y_i$ are non-negative. Using the equivalence between Poisson and exponential distributions, it can be shown that $y \sim \text{Poisson}(\mu)$ implies $y \sim \text{Sub} - \exp(\mu)$. It then follows by results for sub-exponential random variables (see, e.g., Proposition 2.7.1 of Vershynin (2018)) that

$$\mathbb{E}\left(|y|^m|\mathbf{x}_i\right) \leq (C_*\mu)^m m^m$$

whence taking expectation over $\mathbf{x}$ it follows that for any $m \geq 0$ we have

$$\begin{aligned}
\mathbb{E}y^m &\leq \mathbb{E}(\log(1 + e^{\mathbf{x}^\top \boldsymbol{\beta}^*}))^m \times (C_* m)^m \\
&\leq (C_* m)^m \mathbb{E}e^{m\mathbf{x}^\top \boldsymbol{\beta}^*} \\
&\leq (C_* m)^m e^{\boldsymbol{\beta}^{*\top}\mathbf{\Sigma}\boldsymbol{\beta}^* m^2/2} \\
&\leq (C_* m)^m e^{m^2 C_X \xi^2/2}.
\end{aligned}$$

for a numerical constant $C_* > 0$. In the last series of inequalities we have used first the fact that $\log(1 + x) \leq x$ for $x > 0$. Next we have used the moment generating function of the normal distribution $\mathbf{x}^\top \boldsymbol{\beta}^* \sim N(0, \boldsymbol{\beta}^{*\top}\mathbf{\Sigma}\boldsymbol{\beta}^*)$ and finally the inequality $\boldsymbol{\beta}^{*\top}\mathbf{\Sigma}\boldsymbol{\beta} \leq C_X\xi^2$ in the last step.

The negative log-likelihood is

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

and the derivative satisfies

$$|\dot{\ell}(y,z)| = \left| \frac{1}{1+e^{-z}} - \frac{ye^z}{(1+e^z)\log(1+e^z)} \right| \le 1 + |y|.$$

### 6.5 Detailed Proofs

#### 6.5.1 Proof of Theorem 2.1

We decompose the proof into bounding $\mathbb{E}V_1^2$ and $\mathbb{E}V_2^2$ following (7) and (8). To bound $\mathbb{E}V_2^2$, first notice that for all $i \in [n]$ we have

$$\mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] = \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] = \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}].$$

By the mean-value theorem, for each $i$ there exists a random variable $\boldsymbol{\xi}_i = t_i\widehat{\boldsymbol{\beta}}_{/i} + (1-t_i)\widehat{\boldsymbol{\beta}}$ with $t_i \in [0,1]$ such that

$$\phi_0(\widehat{\boldsymbol{\beta}}_{/i}) - \phi_0(\widehat{\boldsymbol{\beta}}) = \dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}).$$

Then we have

$$\begin{aligned}
\mathbb{E}(V_2^2) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}] - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}})|\mathcal{D}]\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}] - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}})|\mathcal{D}]\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n \mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\right)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\Big(\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}] \cdot \mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_j)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/j} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\Big) \\
&\le \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \sqrt{\mathbb{E}\Big(\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\Big)^2} \cdot \sqrt{\mathbb{E}\Big(\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_j)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/j} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\Big)^2} \\
&= \mathbb{E}\Big(\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_1)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\Big)^2, \qquad (14)
\end{aligned}$$

where we use Cauchy Schwarz inequality in the penultimate step. Next we have

$$\begin{aligned}
&\left|\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_i)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]\right| \le \sqrt{\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]\mathbb{E}[(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})^\top\mathbf{x}_0\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})|\mathcal{D}]} \\
\le &\sqrt{\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]}\sqrt{\frac{C_X}{p}\|\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}\|^2} \qquad (15)
\end{aligned}$$

where the last inequality uses the independence between $\mathbf{x}_0$ and $\mathcal{D}$ and also the fact that $\sigma_{\max}(\boldsymbol{\Sigma}) \le \frac{C_X}{p}$ by Assumption A2.

Hence, in order to bound $\mathbb{E}(V_2^2)$, we have to obtain bounds on $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\|$ and also $\mathbb{E}[\dot{\phi}_0^2(\boldsymbol{\xi}_i)|\mathcal{D}]$. That is what the next two lemmas aim to do. The first lemma connects $\widehat{\boldsymbol{\beta}}_{/i}$ and $\widehat{\boldsymbol{\beta}}$:

**Lemma 6.8.** *Under assumptions A1-A3, we have for all $\alpha > 0$ that*

$$\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\right\| \le \frac{|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}_{/i})|\,\|\mathbf{x}_i\|}{2\lambda\eta \wedge 1}.$$

The proof of this lemma is presented in Section 6.5.3 and uses the ideas that we mentioned in Section 2, i.e. smoothing and projection operator.

The next lemma bounds the moments of $\phi_0$ and $\dot{\phi}_0$:

**Lemma 6.9.** *Suppose assumptions A1-A4 hold. Then there exists a constant $C_\phi$ depending only on $s$ (from Assumption A4) and $C_X$ (from Assumption A2) such that $\forall \boldsymbol{\beta} \in \mathbb{R}^p$:*

$$\sqrt{\mathbb{E}[\phi_0^2(\boldsymbol{\beta})]} \le C_\phi + C_\phi\left(\frac{1}{p}\|\boldsymbol{\beta}\|^2\right)^{\frac{s}{2}}, \text{ and } \sqrt{\mathbb{E}[\dot{\phi}_0^{2k}(\boldsymbol{\beta})]} \le C_\phi^k + C_\phi^k\left(\frac{1}{p}\|\boldsymbol{\beta}\|^2\right)^{\frac{sk}{2}} \text{ for } k = 1, 2.$$

The proof of this lemma can be found in Section 6.5.4.

Inserting Lemma 6.8 and Lemma 6.9 back into (15) we have

$$\left|\mathbb{E}[\dot{\phi}_0(\boldsymbol{\xi}_1)\mathbf{x}_0^\top(\widehat{\boldsymbol{\beta}}_{/1}-\widehat{\boldsymbol{\beta}})|\mathcal{D}]\right| \leq \left(C_\phi + C_\phi\left(\frac{1}{p}\|\boldsymbol{\xi}_1\|^2\right)^{\frac{s}{2}}\right)\sqrt{\frac{C_X}{p}}\frac{|\dot{\ell}_1(\widehat{\boldsymbol{\beta}}_{/1})|\|\mathbf{x}_1\|}{2\lambda\eta\wedge 1}.$$

Hence, if we use (14), then we will obtain

$$\begin{aligned}
\mathbb{E}V_2^2 &\leq \mathbb{E}\left(C_\phi^2\left(1+\frac{\|\boldsymbol{\xi}_1\|^s}{p^{s/2}}\right)^2\frac{C_X}{p(2\lambda\eta\wedge 1)^2}\dot{\ell}_1^2(\widehat{\boldsymbol{\beta}}_{/1})\|\mathbf{x}_1\|^2\right)\\
&\overset{(a)}{\leq} \frac{C_\phi^2 C_X}{p(2\lambda\eta\wedge 1)^2}\sqrt{\mathbb{E}\left(1+\frac{1}{p^{s/2}}\|\boldsymbol{\xi}_1\|^s\right)^4}\times\sqrt{\mathbb{E}\left(\dot{\ell}_1^4(\widehat{\boldsymbol{\beta}}_{/1})\|\mathbf{x}_1\|^4\right)}\\
&\overset{(b)}{\leq} \frac{1}{n}\frac{C_\phi^2 C_X\gamma_0}{(2\lambda\eta\wedge 1)^2}\sqrt{8\left(1+\mathbb{E}\left(\frac{1}{p}\|\boldsymbol{\xi}_1\|^2\right)^{2s}\right)}\times\left(\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1})\right)^{\frac{1}{4}}\cdot\left(\mathbb{E}\|\mathbf{x}_1\|^8\right)^{\frac{1}{4}}. \qquad (16)
\end{aligned}$$

where steps (a) and (b) both used Cauchy Schwarz Inequality. The following lemma bounds $\mathbb{E}\left(\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2\right)^t$ and $\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1})$:

**Lemma 6.10.** *Under assumptions A1-A4, there exist constants $C_\beta(t)$ depending on $\{\gamma_0,\lambda,\eta,C_Y(\cdot),s,t\}$ and $C_\ell$ depending on $\{\gamma_0,\lambda,\eta,C_Y(\cdot),s,C_X\}$ such that*

*(a) For $t\geq 1$, $\mathbb{E}[p^{-1}\|\widehat{\boldsymbol{\beta}}\|^2]^t \leq C_\beta(t), \mathbb{E}[p^{-1}\|\widehat{\boldsymbol{\beta}}_{/1}\|^2]^t \leq C_\beta(t)$,*

*(b) $\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1}) \leq C_\ell$.*

The proof of this lemma can be found in Section 6.5.5. Observe that $\boldsymbol{\xi}_1$ is a convex combination of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{/i}$, then using part (a) of Lemma 6.10 we have that the same bound applies to $p^{-1}\|\boldsymbol{\xi}_1\|^2$. Next, by Lemma 6.5 we have $\mathbb{E}\|\mathbf{x}_1\|^8 \leq 24C_X^4$. Inserting the above bounds into (16) we have

$$\begin{aligned}
\mathbb{E}V_2^2 &\leq \frac{1}{n}\frac{C_\phi^2 C_X\gamma_0}{(2\lambda\eta\wedge 1)^2}\sqrt{8\left(1+C_\beta(2s)\right)}\cdot C_\ell^{\frac{1}{4}}\cdot(24C_X^4)^{\frac{1}{4}}\\
&= \frac{1}{n}\frac{\sqrt{8}(24)^{\frac{1}{4}}C_\phi^2 C_X^2 C_\ell^{1/4}\gamma_0}{(2\lambda\eta\wedge 1)^2}\sqrt{1+C_\beta(2s)} := \frac{C_{v2}}{n}. \qquad (17)
\end{aligned}$$

Now using Lemma 4.3 along with (8) finishes the proof.

### 6.5.2 Proof of Lemma 4.1

We start with the proof of Part (a). First we extend $r_0$ to $\mathbb{R}^p$ by simply letting $r_0(\boldsymbol{\beta})=0$ for $\boldsymbol{\beta}\notin\boldsymbol{\Theta}$. Then we have

$$\begin{aligned}
r_0^\alpha(\boldsymbol{\beta}) &= \int_{\mathbb{R}^p} r_0(\boldsymbol{\beta}-\mathbf{z})\alpha\phi(\alpha\mathbf{z})d\mathbf{z}\\
&= \alpha\int_{\mathbb{R}^p} r_0(\mathbf{z})\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))d\mathbf{z}.
\end{aligned}$$

For a point $\boldsymbol{\beta}\in\boldsymbol{\Theta}_0$, consider its directional derivative over a direction $\mathbf{v}\in\mathbb{R}^p$ with $\|\mathbf{v}\|_2=1$:

$$\begin{aligned}
&\nabla_{\mathbf{v}} r_0^\alpha(\boldsymbol{\beta})\\
&:= \lim_{h\to 0}\frac{1}{h}[r_0^\alpha(\boldsymbol{\beta}+h\mathbf{v})-r_0^\alpha(\boldsymbol{\beta})]\\
&= \lim_{h\to 0}\alpha\int_{\mathbb{R}^p} r_0(\mathbf{z})\frac{1}{h}[\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}+h\mathbf{v}))-\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))]d\mathbf{z}
\end{aligned}$$

Since $\phi \in \mathcal{C}^\infty(\mathbb{R}^p)$, we have that for small enough $h$:

$$\begin{aligned}
&\frac{1}{h}[\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}+h\mathbf{v}))-\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))]\\
&\le 2\alpha \mathbf{v}^\top \nabla\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))\\
&\le 2\alpha \left\|\nabla\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))\right\|\\
&\le 2\alpha(2\pi)^{-p/2} e^{-\frac{1}{2}\alpha^2\|\boldsymbol{\beta}-\mathbf{z}\|_2^2} \left\|\boldsymbol{\beta}-\mathbf{z}\right\|_2.
\end{aligned}$$

By assumption $r_0(\mathbf{z}) e^{-\frac{1}{2}\alpha^2\|\boldsymbol{\beta}-\mathbf{z}\|_2^2} \left\|\boldsymbol{\beta}-\mathbf{z}\right\|_2$ is integrable, then using Dominated Convergence, the limit can be taken within the integral, so that $\nabla_{\mathbf{v}} r_0^\alpha(\boldsymbol{\beta})$ exists. Using similar arguments, in order that $\nabla_{\mathbf{v}}^k r_0^\alpha(\boldsymbol{\beta})$ exists, we only need $r_0(\mathbf{z})\nabla_{\mathbf{v}}^k\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))$ to be integrable. To see this, notice that

$$\frac{\partial^k}{\partial u_1^{k_1}\cdots\partial u_p^{k_p}}\phi(\mathbf{u}) = e^{-\frac{1}{2}\|\mathbf{u}\|^2} P_{k_1\cdots k_p}(\mathbf{u})$$

where $P_{k_1\cdots k_p}(\mathbf{u}) = u_1^{k_1}u_2^{k_2}\cdots u_p^{k_p} + o(u_1^{k_1}u_2^{k_2}\cdots u_p^{k_p})$ is a polynomial with its dominating term being $u_1^{k_1}u_2^{k_2}\cdots u_p^{k_p}$ with order $K$. Then we have

$$\begin{aligned}
&\left|\nabla_{\mathbf{v}}^k\phi(\mathbf{u})\right|\\
&= \left|\sum_{k_1+\cdots+k_p=k} \frac{\partial^k}{\partial u_1^{k_1}\cdots\partial u_p^{k_p}}\phi(\mathbf{u}) v_1^{k_1}\cdots v_p^{k_p}\right|\\
&\le \sum_{k_1+\cdots+k_p=k} \left|v_1^{k_1}\cdots v_p^{k_p}\right| e^{-\frac{1}{2}\|\mathbf{u}\|^2} P_{k_1\cdots k_p}(\mathbf{u})\\
&\le \left(\sum_{k_1+\cdots+k_p=k} \left|P_{k_1\cdots k_p}(\mathbf{u})\right|\right) e^{-\frac{1}{2}\|\mathbf{u}\|^2} \left(\sum|v_i|\right)^k\\
&= O(\|\mathbf{u}\|^k e^{-\frac{1}{2}\|\mathbf{u}\|^2} p^{k/2} \|\mathbf{v}\|_2^k)\\
&= O(\|\mathbf{u}\|^k e^{-\frac{1}{2}\|\mathbf{u}\|^2})
\end{aligned}$$

Inserting $\mathbf{u} = \alpha(\boldsymbol{\beta}-\mathbf{z})$ we have

$$\begin{aligned}
\left|\nabla_{\mathbf{v}}^k\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))\right| &= O(\alpha^k \|\boldsymbol{\beta}-\mathbf{z}\|_2^k e^{-\frac{1}{2}\alpha^2\|\boldsymbol{\beta}-\mathbf{z}\|_2^2})\\
&= O(\|\mathbf{z}\|_2^k e^{-\frac{1}{2}\alpha^2\|\boldsymbol{\beta}-\mathbf{z}\|_2^2})
\end{aligned}$$

So $r_0(\mathbf{z})\nabla_{\mathbf{v}}^k\phi(\alpha(\boldsymbol{\beta}-\mathbf{z}))$ is integrable if $r_0(\mathbf{z})\|\mathbf{z}\|_2^k e^{-\frac{1}{2}\|\boldsymbol{\beta}-\mathbf{z}\|_2^2}$ is integrable. The continuity of the derivatives follows then from exchanging the integration and differentiation.

We now turn to the proof of Part (b) of the Lemma. For all $\boldsymbol{\beta} \in \boldsymbol{\Theta}_0$,

$$\begin{aligned}
|r_0^\alpha(\boldsymbol{\beta}) - r_0(\boldsymbol{\beta})| &\le \int_{\mathbb{R}^p} |r_0(\boldsymbol{\beta}-\mathbf{z}) - r_0(\boldsymbol{\beta})|\alpha\phi(\alpha\mathbf{z})d\mathbf{z}\\
&= \int_{\mathbb{R}^p} |r_0(\boldsymbol{\beta}-\alpha^{-1}\mathbf{u}) - r_0(\boldsymbol{\beta})| d\Phi(\mathbf{u})\\
&\le \frac{L}{\alpha^k}\int_{\mathbb{R}^p} \|\mathbf{u}\|_1^k \, d\Phi(\mathbf{u})\\
&\to 0 \quad (\alpha\to\infty)
\end{aligned}$$

The second line uses change of variable $\mathbf{u} = \alpha\mathbf{z}$ and the third line uses our assumption on $r_0$. The last line is because the integral is finite. □

*Remark* 6.11. Although Lemma 4.1 adopts a weaker assumption than Lipschitz continuity, it can be shown that $k = 1$ is the only value that allows $r_0$ to be convex. In fact, for $k < 1$ we have

$$r_0(\mathbf{x}) \le L\|\mathbf{x}\|^k.$$

It is obvious that unless $r_0$ is constant, it would eventually increase at least linearly thus cannot be bounded by $L\|\mathbf{x}\|^k$ for some $k < 1$.

*Example* 6.8. Generalized LASSO: $r_0(\boldsymbol{\beta}) = \|\mathbf{D}\boldsymbol{\beta}\|_1$ for a fixed $\mathbf{D} \in \mathbb{R}^{m\times p}$. It is continuous, and $|r_0(\mathbf{x}) - r_0(\mathbf{y})| = |\,\|\mathbf{D}\mathbf{x}\|_1 - \|\mathbf{D}\mathbf{y}\|_1\,| \le \|\mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{y}\|_1 \le \sqrt{m}\,\|\mathbf{x}-\mathbf{y}\|_2$. Note that the classical LASSO is a special case with $\mathbf{D} = \mathbf{I}_p$.

*Remark* 6.12. In fact if $r_0$ is a norm such that $\mathbb{E}r_0(\boldsymbol{\beta}) < \infty$ where $\boldsymbol{\beta}$ has i.i.d. $N(0,1)$ entries, then the results of the above two lemmas hold.

### 6.5.3 Proof of Lemma 6.8

*Proof.* Begin we start the proof, let us introduce the following notations:

$$
\begin{aligned}
h(\boldsymbol{\beta}) &:= \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\
h^\alpha(\boldsymbol{\beta}) &:= \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\
h_{/i}(\boldsymbol{\beta}) &:= \sum_{j\neq i}^{n} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\
h^\alpha_{/i}(\boldsymbol{\beta}) &:= \sum_{j\neq i}^{n} \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta) r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}.
\end{aligned}
$$

As mentioned before, in order to obtain a bound on $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\|_2$ we use the smoothing trick. Hence, we first obtain an upper bound for $\|\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}\|$.

Similar to $h$ and $h^\alpha$, let $h_{/i}$ and $h^\alpha_{/i}$ denote the loss functions for $\widehat{\boldsymbol{\beta}}_{/i}$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ respectively. By Lemma 6.1, $\widehat{\boldsymbol{\beta}}^\alpha$ and $\widehat{\boldsymbol{\beta}}^\alpha_{/i}$ satisfy

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^\alpha &= \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha))\\
\widehat{\boldsymbol{\beta}}^\alpha_{/i} &= \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})).
\end{aligned}
$$

By subtracting one from the other we have

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i} =& \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha)) - \boldsymbol{\Pi}_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}^\alpha_{/i} - \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i}))\\
=& \bar{\mathbf{J}} \times \left(\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \widehat{\boldsymbol{\beta}}^\alpha_{/i} + \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right) \qquad (18)
\end{aligned}
$$

where the second line comes from Lemma 6.2 and

$$
\bar{\mathbf{J}} := \int_0^1 \mathbf{J}(t)dt.
$$

It is straightforward to use Lemma 6.2 to show that

$$
0 \le \lambda_{\min}(\bar{\mathbf{J}}) \le \lambda_{\max}(\bar{\mathbf{J}}) \le 1.
$$

On the other hand,

$$
\begin{aligned}
&\widehat{\boldsymbol{\beta}}^\alpha - \nabla h^\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \widehat{\boldsymbol{\beta}}^\alpha_{/i} + \nabla h^\alpha_{/i}(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\\
=& \widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i} - \sum_{j\in[n]} \left[\dot\ell_j(\widehat{\boldsymbol{\beta}}^\alpha) - \dot\ell_j(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right]\mathbf{x}_j - \lambda\left[\nabla r^\alpha(\widehat{\boldsymbol{\beta}}^\alpha) - \nabla r^\alpha(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\right] - \dot\ell_i(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\mathbf{x}_i\\
=& \left(\mathbf{I} - \mathbf{X}^\top \operatorname{diag}[\ddot\ell_j(\boldsymbol{\xi}_j)]_{j\in[n]}\mathbf{X} - \lambda\nabla^2 r^\alpha(\boldsymbol{\Xi})\right)(\widehat{\boldsymbol{\beta}}^\alpha - \widehat{\boldsymbol{\beta}}^\alpha_{/i}) - \dot\ell_i(\widehat{\boldsymbol{\beta}}^\alpha_{/i})\mathbf{x}_i \qquad (19)
\end{aligned}
$$

where the second line uses the definition of $h^\alpha$ and the third line uses mean-value-theorem on $\dot\ell_j$ and $\nabla r^\alpha$:

$$
\ddot\ell_j(\boldsymbol{\xi}_j) := \int_0^1 \ddot\ell_j(t\widehat{\boldsymbol{\beta}}^\alpha + (1-t)\widehat{\boldsymbol{\beta}}^\alpha_{/i})dt
$$

with

$$\dot{\ell}_j(\widehat{\boldsymbol{\beta}}^{\alpha}) - \dot{\ell}_j(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i}) = \ddot{\ell}_j(\boldsymbol{\xi}_j)\mathbf{x}_j^{\top}(\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i})$$

and likewise

$$\nabla^2 r^{\alpha}(\boldsymbol{\Xi}) := \int_0^1 \nabla^2 r^{\alpha}(t\widehat{\boldsymbol{\beta}}^{\alpha} + (1-t)\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})dt$$

with

$$\nabla r^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}) - \nabla r^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i}) = \nabla^2 r^{\alpha}(\boldsymbol{\Xi})(\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i}).$$

Inserting this back to (18) we have

$$\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i} = -\mathbf{G}^{-1}\bar{\mathbf{J}}\left(\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\mathbf{x}_i\right)$$

where

$$\begin{aligned}\mathbf{G} &:= \mathbf{I} + \bar{\mathbf{J}}\left(\mathbf{X}^{\top}\operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\in[n]}\mathbf{X} + \lambda\nabla^2 r^{\alpha}(\boldsymbol{\Xi}) - \mathbf{I}\right)\\ &= \mathbf{I} - \bar{\mathbf{J}} + \bar{\mathbf{J}}\left(\mathbf{X}_{/i}^{\top}\operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\neq i}\mathbf{X}_{/i} + \lambda\nabla^2 r^{\alpha}(\boldsymbol{\Xi})\right).\end{aligned}$$

Since $\mathbf{I} - \bar{\mathbf{J}}$ is positive semidefinite (note that all the eigenvalues of $\bar{\mathbf{J}}$ are in $[0,1]$ ) and $\mathbf{X}^{\top}\operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\in[n]}\mathbf{X} + \lambda\nabla^2 r^{\alpha}(\boldsymbol{\Xi})$ is positive definite (due to the ridge component), $\mathbf{G}$ is also positive definite. Hence, we have

$$\left\|\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i}\right\|_2 \leq \left[\sigma_{\min}(\mathbf{G})\right]^{-1}\sigma_{\max}(\bar{\mathbf{J}})\left\|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\mathbf{x}_i\right\|_2. \tag{20}$$

We have already established $\sigma_{\max}(\bar{\mathbf{J}}) \leq 1$. Now we bound $\sigma_{\min}(\mathbf{G})$:

$$\begin{aligned}\sigma_{\min}(\mathbf{G}) &= \lambda_{\min}(\mathbf{G})\\ &= 1 + \lambda_{\min}\left(\bar{\mathbf{J}}\left(\mathbf{X}_{/i}^{\top}\operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\notin I_i}\mathbf{X}_{/i} + \lambda\nabla^2 r^{\alpha}(\boldsymbol{\Xi}) - \mathbf{I}\right)\right)\\ &:= 1 + \lambda_{\min}(\bar{\mathbf{J}}\mathbf{M})\end{aligned}$$

where $\mathbf{M} := \mathbf{X}_{/i}^{\top}\operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\notin I_i}\mathbf{X}_{/i} + \lambda\nabla^2 r^{\alpha}(\boldsymbol{\Xi}) - \mathbf{I}$. Observe that due to the existence of the ridge component in $r$, we have

$$\lambda_{\min}(\mathbf{M}) \geq 2\lambda\eta - 1.$$

- If $\lambda_{\min}(\mathbf{M}) \geq 0$, then we have
$$\lambda_{\min}\left(\bar{\mathbf{J}}\mathbf{M}\right) \geq \lambda_{\min}(\bar{\mathbf{J}})\lambda_{\min}(\mathbf{M}) \geq 0.$$
- If $2\lambda\eta - 1 \leq \lambda_{\min}(\mathbf{M}) < 0$, then we have
$$\lambda_{\min}\left(\bar{\mathbf{J}}\mathbf{M}\right) \geq \lambda_{\max}(\bar{\mathbf{J}})\lambda_{\min}(\mathbf{M}) \geq 2\lambda\eta - 1.$$

Therefore we have

$$\sigma_{\min}(\mathbf{G}) \geq 1 + (2\lambda\eta - 1) \wedge 0 = 2\lambda\eta \wedge 1.$$

which implies

$$\sigma_{\max}(\mathbf{G}^{-1}) \leq \frac{1}{2\lambda\eta \wedge 1} \tag{21}$$

inserting this back to (20) we have

$$\left\|\widehat{\boldsymbol{\beta}}^{\alpha} - \widehat{\boldsymbol{\beta}}^{\alpha}_{/i}\right\| \leq \frac{\left\|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}^{\alpha}_{/i})\mathbf{x}_i\right\|}{2\lambda\eta \wedge 1}.$$

The next step of the proof is to use this upper bound on the smoothed estimates and obtain an upper bound for $\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{/i}\right\|$. Towards this goal, we first prove the following lemma:

**Lemma 6.13.** *Under assumptions A1-A3, we have that*

$$\left\|\widehat{\boldsymbol{\beta}}^{\alpha}-\widehat{\boldsymbol{\beta}}\right\| \leq \sqrt{\frac{2(1-\eta)}{\eta}\left\|r_0^{\alpha}-r_0\right\|_{\infty}},$$

*and similarly*

$$\left\|\widehat{\boldsymbol{\beta}}_{/i}^{\alpha}-\widehat{\boldsymbol{\beta}}_{/i}\right\| \leq \sqrt{\frac{2(1-\eta)}{\eta}\left\|r_0^{\alpha}-r_0\right\|_{\infty}}.$$

*Proof.* We have

$$\begin{aligned} &h^{\alpha}(\widehat{\boldsymbol{\beta}})-h^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha})=h^{\alpha}(\widehat{\boldsymbol{\beta}})-h(\widehat{\boldsymbol{\beta}}^{\alpha})+h(\widehat{\boldsymbol{\beta}}^{\alpha})-h^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}) \\ &\leq h^{\alpha}(\widehat{\boldsymbol{\beta}})-h(\widehat{\boldsymbol{\beta}})+h(\widehat{\boldsymbol{\beta}}^{\alpha})-h^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}) \leq 2\lambda(1-\eta)\left\|r_0^{\alpha}-r_0\right\|_{\infty} \end{aligned} \tag{22}$$

where the first inequality uses the fact that $\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}^{\alpha}$ are the minimizers of $h(\boldsymbol{\beta})$ and $h^{\alpha}(\boldsymbol{\beta})$ respectively, and the last inequality uses the definition of $h, h^{\alpha}$. On the other hand we also have from the second order mean value theorem (or Taylor expansion) that

$$\begin{aligned} &h^{\alpha}(\widehat{\boldsymbol{\beta}})-h^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha}) \\ =&\nabla h^{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha})^{\top}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha})+\frac{1}{2}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha})^{\top}\nabla^2 h^{\alpha}(\boldsymbol{\Xi})(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}) \\ \geq&\frac{1}{2}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha})^{\top}\nabla^2 h^{\alpha}(\boldsymbol{\Xi})(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}) \\ \geq&\lambda\eta\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}\|^2. \end{aligned} \tag{23}$$

In the first equality, by a slight abuse of the notation, we have written $\nabla^2 h^{\alpha}(\boldsymbol{\Xi})$ as the Hessian in the second order Taylor expansion.[§] To obtain the second line, observe that

$$\nabla h_{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha})^{\top}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha})=\left.\frac{\partial}{\partial t}h^{\alpha}((1-t)\widehat{\boldsymbol{\beta}}^{\alpha}+t\widehat{\boldsymbol{\beta}})\right|_{t=0}.$$

Since $\widehat{\boldsymbol{\beta}}^{\alpha}$ is the minimizer of $h^{\alpha}$, $t=0$ is the minimizer of $h^{\alpha}((1-t)\widehat{\boldsymbol{\beta}}^{\alpha}+t\widehat{\boldsymbol{\beta}})$ for $t\in[0,1]$. And since $h^{\alpha}$ is smooth on $(0,1)$ we must have $\nabla h_{\alpha}(\widehat{\boldsymbol{\beta}}^{\alpha})^{\top}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha})=\left.\frac{\partial}{\partial t}h^{\alpha}((1-t)\widehat{\boldsymbol{\beta}}^{\alpha}+t\widehat{\boldsymbol{\beta}})\right|_{t=0}\geq 0$. The last line of (23) is because $\nabla^2 h^{\alpha}(\boldsymbol{\Xi})$ is positive-definite and $\sigma_{\min}(\nabla^2 h^{\alpha}(\boldsymbol{\Xi}))\geq 2\lambda\eta$ due to the existence of the ridge component.

Combining (22) and (23) we have

$$\left\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{\alpha}\right\|^2 \leq \frac{2(1-\eta)}{\eta}\left\|r_0^{\alpha}-r_0\right\|_{\infty}.$$

Using a similar argument we can also prove that

$$\left\|\widehat{\boldsymbol{\beta}}_{/i}-\widehat{\boldsymbol{\beta}}_{/i}^{\alpha}\right\|^2 \leq \frac{2(1-\eta)}{\eta}\left\|r_0^{\alpha}-r_0\right\|_{\infty}.$$

□

From lemma 4.1 we can see that $\left\|r_0^{\alpha}-r_0\right\|_{\infty}\to 0$ as $\alpha\to\infty$.

Continuing with the proof of Lemma 6.8, by Lemma 6.13, $\widehat{\boldsymbol{\beta}}^{\alpha}\to\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{/i}^{\alpha}\to\widehat{\boldsymbol{\beta}}_{/i}$ when $\alpha\to\infty$. Hence, we can let $\alpha\to\infty$ and have

$$\left\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{/i}\right\| \leq \frac{\left\|\dot{\ell}_i(\widehat{\boldsymbol{\beta}}_{/i})\mathbf{x}_i\right\|}{2\lambda\eta\wedge 1}.$$

□

[§]Since the mean-value theorem doesn't exist for vector-valued functions, the matrix $\nabla^2 h^{\alpha}(\boldsymbol{\Xi})$ is actually the Hessian of $h^{\alpha}$, with each row evaluated at a different convex combination of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^{\alpha}$, as the matrix $\boldsymbol{\Xi}$ indicates.

### 6.5.4 Proof of Lemma 6.9

We first obtain an upper bound for the $4^{\text{th}}$ moment. Using Assumption A4, we have

$$\begin{aligned}\sqrt{\mathbb{E}\dot{\phi}_0^4(\boldsymbol{\beta})} &\le \sqrt{\mathbb{E}\left(1+|y_0|^s+|\mathbf{x}_0^\top\boldsymbol{\beta}|^s\right)^4}\\ &\le \sqrt{27(1+\mathbb{E}|y_0|^{4s}+\mathbb{E}|\mathbf{x}_0^\top\boldsymbol{\beta}|^{4s})}\\ &\le \sqrt{27\left(1+C_Y(4s)+(4s-1)!!\left(\frac{C_X}{p}\|\boldsymbol{\beta}\|^2\right)^{2s}\right)}\\ &\le \sqrt{27(1+C_Y(4s))}+\sqrt{27(4s-1)!!}C_X^s\left(\frac{C_X}{p}\|\boldsymbol{\beta}\|^2\right)^s.\end{aligned}$$

where the second line uses the simple equation $(a+b+c)^4 \le 27a^4+27b^4+27c^4$, the third line uses $\mathbf{x}_0^\top\boldsymbol{\beta} \sim N(0,\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})$ with $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} \le \frac{C_X}{p}\|\boldsymbol{\beta}\|^2$. It also uses the moment formula for standard Gaussian variable. The last line uses the equation $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ for $a,b \ge 0$. Define $C_\phi^2 = \max\{\sqrt{27(1+C_Y(4s))}, \sqrt{27(4s-1)!!}C_X^s\}$, then

$$\sqrt{\mathbb{E}\dot{\phi}_0^4(\boldsymbol{\beta})} \le C_\phi^2\left(1+\left(\frac{C_X}{p}\|\boldsymbol{\beta}\|^2\right)^s\right)$$

For the second moment, notice that

$$\begin{aligned}\sqrt{\mathbb{E}\dot{\phi}_0^2(\boldsymbol{\beta})} &\le [\mathbb{E}\dot{\phi}_0^4(\boldsymbol{\beta})]^{\frac{1}{4}}\\ &\le C_\phi\sqrt{1+\left(\frac{C_X}{p}\|\boldsymbol{\beta}\|^2\right)^s}\\ &\le C_\phi\left(1+\left(\frac{C_X}{p}\|\boldsymbol{\beta}\|^2\right)^{s/2}\right).\end{aligned}$$

where the last step uses $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ again. The same arguments lead to the same bound for $\sqrt{\mathbb{E}\phi_0^2(\boldsymbol{\beta})}$.

### 6.5.5 Proof of Lemma 6.10

*Proof.* (a) Without loss of generality we assume $C = 1$ in Assumption A4. Throughout this proof we use $h(\boldsymbol{\beta}), h_{/i}(\boldsymbol{\beta}), h^\alpha(\boldsymbol{\beta})$ and $h^\alpha_{/i}(\boldsymbol{\beta})$ for the loss functions of the corresponding models, defined as:

$$\begin{aligned}h(\boldsymbol{\beta}) &:= \sum_{j=1}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta)r_0(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\ h^\alpha(\boldsymbol{\beta}) &:= \sum_{j=1}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta)r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\ h_{/i}(\boldsymbol{\beta}) &:= \sum_{j\neq i}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta)r_0(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta},\\ h^\alpha_{/i}(\boldsymbol{\beta}) &:= \sum_{j\neq i}^n \ell_j(\boldsymbol{\beta}) + \lambda(1-\eta)r_0^\alpha(\boldsymbol{\beta}) + \lambda\eta\boldsymbol{\beta}^\top\boldsymbol{\beta}.\end{aligned}$$

It is straightforward to show that

$$\lambda\eta\|\widehat{\boldsymbol{\beta}}\|_2^2 \le \sum_{j\in[n]}\ell(y_j,\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}}) + \lambda(1-\eta)r_0(\widehat{\boldsymbol{\beta}}) + \lambda\eta\|\widehat{\boldsymbol{\beta}}\|_2^2 \le \sum_{j=1}^n \ell(y_j,0),$$

where the last inequality is due to the fact that $h(\widehat{\boldsymbol{\beta}}) \leq h(\mathbf{0})$ and that $\ell(y,z) \geq 0$. Similarly we have $\lambda\eta\|\widehat{\boldsymbol{\beta}}_{/i}\|_2^2 \leq \sum_{j=1}^n \ell(y_j, 0)$.

We can then use Assumption A4 to obtain

$$\mathbb{E}\left(\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2\right)^t \leq (p\lambda\eta)^{-t}\mathbb{E}[n + |y_1|^s + \cdots + |y_n|^s]^t.$$

By Rosenthal inequality (Lemma 6.7), we have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2\right)^t &\leq (p\lambda\eta)^{-t}A(t)\max\left\{n^t + n\mathbb{E}|y_1|^{st}, (n + n\mathbb{E}|y_1|^s)^t\right\} \\
&\leq \left(\frac{\gamma_0}{\lambda\eta}\right)^t A(t)\max\left\{1 + n^{1-t}C_Y(st), (1 + C_Y(s))^t\right\} \\
&\leq \left(\frac{\gamma_0}{\lambda\eta}\right)^t A(t)\max\left\{1 + C_Y(st), (1 + C_Y(s))^t\right\} := C_\beta(t), \qquad (24)
\end{aligned}$$

where $C_Y(\cdot)$ is the constant in Assumption A4, and $A(\cdot)$ is the Rosenthal constant in Lemma 6.7.

(b) Our next goal is to obtain an upper bound for $\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1})$. By using Assumption A4 we have

$$\begin{aligned}
\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1}) &\leq \mathbb{E}[1 + |y_1|^s + |\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|^s]^8 \\
&\leq 3^7\left(1 + \mathbb{E}|y_1|^{8s} + \mathbb{E}|\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|^{8s}\right). \qquad (25)
\end{aligned}$$

Next we bound $\mathbb{E}|\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|^{8s}$:

$$\begin{aligned}
\mathbb{E}|\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|^{8s} &= \mathbb{E}\left[\mathbb{E}[|\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|^{8s}|\widehat{\boldsymbol{\beta}}_{/1}]\right] \\
&\leq \mathbb{E}[(8s-1)!!(\widehat{\boldsymbol{\beta}}_{/1}^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}_{/1})^{4s}] \\
&\leq (8s-1)!!\mathbb{E}\left(\frac{C_X}{p}\|\widehat{\boldsymbol{\beta}}_{/1}\|^2\right)^{4s} \\
&\leq (8s-1)!!C_X^{4s}C_\beta(4s)
\end{aligned}$$

where the second line uses the fact that $\mathbf{x}_1^\top\widehat{\boldsymbol{\beta}}_{/1}|\widehat{\boldsymbol{\beta}}_{/1} \sim N(0, \widehat{\boldsymbol{\beta}}_{/1}^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}_{/1})$ and that the $t^{\text{th}}$ moment of a standard Gaussian variable is $(t-1)!!$ whenever $t$ is a positive even number. The last line uses (24). Inserting this back to (25) we have:

$$\mathbb{E}\dot{\ell}_1^8(\widehat{\boldsymbol{\beta}}_{/1}) \leq 3^7(1 + C_Y(8s) + (8s-1)!!C_X^{4s}C_\beta(4s)) := C_\ell.$$

□

### 6.5.6 Proof of Lemma 4.3

Recall that

$$\text{LO} = \frac{1}{n}\sum_{i=1}^n \phi_i(\widehat{\boldsymbol{\beta}}_{/i}).$$

Then we have

$$\begin{aligned}
\mathbb{E}V_1^2 =& \mathbb{E}\left(\text{LO} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_i(\widehat{\boldsymbol{\beta}}_{/i})|\mathcal{D}_{/i}]\right)^2 \\
=& \frac{1}{n}\mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) - \mathbb{E}[\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1}]\right)^2 \\
&+ \frac{n-1}{n}\mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) - \mathbb{E}[\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1}]\right)\cdot\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/2}) - \mathbb{E}[\phi_2(\widehat{\boldsymbol{\beta}}_{/2})|\mathcal{D}_{/2}]\right). \qquad (26)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) - \mathbb{E}(\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1})\right)\right]^2 \\
&= \mathbb{E}\,\mathrm{var}(\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1}) \\
&\le \mathbb{E}\left(\mathbb{E}[\phi_1^2(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1}]\right) \\
&\le \mathbb{E}\left(C_\phi + C_\phi\left(\frac{1}{p}\|\widehat{\boldsymbol{\beta}}_{/1}\|^2\right)^{\frac{s}{2}}\right)^2 \\
&\le 2C_\phi^2 + 2C_\phi^2\mathbb{E}\left(\frac{1}{p}\|\widehat{\boldsymbol{\beta}}_{/1}\|^2\right)^s \\
&\le 2C_\phi^2\,(1 + C_\beta(s)) := C_{v1,1}
\end{aligned} \tag{27}$$

where the fourth line uses Lemma 6.9, and the last line uses part (a) of Lemma 6.10. Next, we study

$$\mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) - \mathbb{E}(\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1})\right)\cdot\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/2}) - \mathbb{E}(\phi_2(\widehat{\boldsymbol{\beta}}_{/2})|\mathcal{D}_{/2})\right). \tag{28}$$

Define $\widehat{\boldsymbol{\beta}}_{/12} := \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\mathrm{argmin}}\Big\{\sum_{j\ge 3}\ell(y_j, \mathbf{x}_j^\top\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})\Big\}$. By the mean-value theorem, there exists a random variable $t\in[0,1]$ such that

$$\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) = \phi_1(\widehat{\boldsymbol{\beta}}_{/12}) + \dot{\phi}_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}).$$

Hence,

$$\begin{aligned}
&\mathbb{E}[\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1}] \\
&= \mathbb{E}[\phi_1(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}] + \mathbb{E}[\dot{\phi}_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}] \\
&= \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}] + \mathbb{E}[\dot{\phi}_0(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_0^\top|\mathcal{D}_{/1}](\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}).
\end{aligned}$$

Similarly, we have

$$\phi_2(\widehat{\boldsymbol{\beta}}_{/2}) = \phi_2(\widehat{\boldsymbol{\beta}}_{/12}) + \dot{\phi}_2(t\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}).$$

and

$$\mathbb{E}[\phi_2(\widehat{\boldsymbol{\beta}}_{/2})|\mathcal{D}_{/2}] = \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}] + \mathbb{E}[\dot{\phi}_0(t\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_0^\top|\mathcal{D}_{/2}](\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})$$

Then (28) can be decomposed into four terms:

$$\begin{aligned}
&\mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/1}) - \mathbb{E}(\phi_1(\widehat{\boldsymbol{\beta}}_{/1})|\mathcal{D}_{/1})\right)\cdot\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/2}) - \mathbb{E}(\phi_2(\widehat{\boldsymbol{\beta}}_{/2})|\mathcal{D}_{/2})\right) \\
=&A_1 + B_1 + C_1 + D_1,
\end{aligned} \tag{29}$$

where $A_1, B_1, C_1$, and $D_1$ are defined as

$$\begin{aligned}
A_1 :=&\ \mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right)\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right) \\
B_1 :=&\ \mathbb{E}\Big[\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}\big]\right) \\
&\times\left(\dot{\phi}_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\dot{\phi}_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/2}]\right)\Big] \\
C_1 :=&\ \mathbb{E}\Big[\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right) \\
&\times\left(\dot{\phi}_1(\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(t\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\left[\dot{\phi}_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}\right]\right)\Big] \\
D_1 :=&\ \mathbb{E}\Bigg\{\left(\dot{\phi}_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\dot{\phi}_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}\big]\right) \\
&\times\left(\dot{\phi}_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\dot{\phi}_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/2}\big]\right)\Bigg\}.
\end{aligned}$$

We bound each of these four terms below. For $A_1$ we have

$$\begin{aligned} A_1 &:= \mathbb{E}\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right)\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right)\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right)\Big|\,\mathcal{D}_{/12}\right]\right] \\ &= 0. \end{aligned} \tag{30}$$

where the last equality is correct, because given $\mathcal{D}_{/12}$, $\phi_1(\widehat{\boldsymbol{\beta}}_{/12})$ and $\phi_2(\widehat{\boldsymbol{\beta}}_{/12})$ are conditionally independent, and hence the inner expectation can be taken to each term separately. Similarly,

$$\begin{aligned} B_1 :=& \mathbb{E}\Big[\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}\big]\right) \\ &\times \left(\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/2}]\right)\Big] \\ =& \mathbb{E}\Big\{\mathbb{E}\Big[\left(\phi_1(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}\big]\right) \\ &\times \left(\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/2}]\right)\Big|\,\mathcal{D}_{/2}\Big]\Big\} \\ =& 0. \end{aligned} \tag{31}$$

Again, the last equality is correct, because given $\mathcal{D}_{/2}$, $\phi_1(\widehat{\boldsymbol{\beta}}_{/12})$ and $\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2$ are conditionally independent. Similarly,

$$\begin{aligned} C_1 :=& \mathbb{E}\Big[\left(\phi_2(\widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}[\phi_0(\widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/12}]\right) \\ &\times \left(\dot\phi_1(\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(t\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\left[\dot\phi_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}\right]\right)\Big] \\ =& 0. \end{aligned} \tag{32}$$

Finally,

$$\begin{aligned} D_1 :=& \mathbb{E}\Bigg\{\left(\dot\phi_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\dot\phi_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}\big]\right) \\ &\times \left(\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12}) - \mathbb{E}\big[\dot\phi_2(\widehat{\boldsymbol{\beta}}_{/2} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_2^\top(\widehat{\boldsymbol{\beta}}_{/2} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/2}\big]\right)\Bigg\} \\ \overset{(a)}{\leq}& \mathbb{E}\,\mathrm{var}\left[\dot\phi_1(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})|\mathcal{D}_{/1}\right] \\ \leq& \mathbb{E}\left[\mathbb{E}\left[\dot\phi_1^2(t\widehat{\boldsymbol{\beta}}_{/1} + (1-t)\widehat{\boldsymbol{\beta}}_{/12})[\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})]^2|\mathcal{D}_{/1}\right]\right] \\ \overset{(b)}{\leq}& \mathbb{E}\left[\sqrt{\mathbb{E}[\dot\phi_1^4(\boldsymbol{\xi}_1)|\mathcal{D}_{/1}]}\sqrt{\mathbb{E}\left[[\mathbf{x}_1^\top(\boldsymbol{\beta}_{/1} - \boldsymbol{\beta}_{/12})]^4|\mathcal{D}_{/1}\right]}\right] \\ \overset{(c)}{\leq}& \mathbb{E}\left[C_\phi^2\left(1 + \left(\frac{1}{p}\|\boldsymbol{\xi}\|^2\right)^s\right)\cdot\frac{\sqrt{3}C_X}{p}\|\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}\|^2\right] \\ \overset{(d)}{\leq}& \frac{\sqrt{3}C_\phi^2C_X}{p(2\lambda\eta\wedge 1)^2}\mathbb{E}\left[\left(1 + \left(p^{-1}\|\boldsymbol{\xi}\|^2\right)^s\right)|\dot\ell_2(\widehat{\boldsymbol{\beta}}_{/12})|^2\|\mathbf{x}_2\|^2\right] \\ \overset{(e)}{\leq}& \frac{\sqrt{3}C_\phi^2C_X}{p(2\lambda\eta\wedge 1)^2}\sqrt{\mathbb{E}\left(1 + (p^{-1}\|\boldsymbol{\xi}\|^2)^s\right)^2}\cdot\left(\mathbb{E}|\dot\ell_2(\widehat{\boldsymbol{\beta}}_{/12})|^8\right)^{\frac{1}{4}}\left(\mathbb{E}\|\mathbf{x}_2\|^8\right)^{\frac{1}{4}} \\ \overset{(f)}{\leq}& \frac{\sqrt{3}\gamma_0C_\phi^2C_X}{n(2\lambda\eta\wedge 1)^2}\sqrt{2 + 2C_\beta(2s)}C_\ell^{\frac{1}{4}}(24C_X^4)^{\frac{1}{4}} \\ :=& \frac{C_{v1,2}}{n}. \end{aligned} \tag{33}$$

where inequality (a) uses Cauchy Schwarz inequality and symmetry between $\dot{\phi}_1$ and $\dot{\phi}_2$, (b) uses Cauchy Schwarz again, (c) uses Lemma 6.9 and the fact that, conditioned on $\mathcal{D}$, $\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}) \sim N(0, \widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}}_{/1} - \widehat{\boldsymbol{\beta}}_{/12})$. Inequality (d) uses Lemma 6.8. Inequality (e) uses Cauchy Schwarz twice. Finally inequality (f) uses Lemma 6.10 and Lemma 6.5. Plugging in the results of equations (30)-(33) into (29), and then using (27), the bound in (26) boils down to

$$\mathbb{E}V_1^2 \leq \frac{C_{v1,1} + C_{v1,2}}{n} := \frac{C_{v1}}{n}$$

where, after some simplification:

$$C_{v1} = 2C_\phi^2 (1 + C_\beta(s)) + \frac{\sqrt{6}(24)^{\frac{1}{4}} C_\phi^2 C_X^2 C_\ell^{\frac{1}{4}} \gamma_0}{(2\lambda\eta \wedge 1)^2} \sqrt{1 + C_\beta(2s)}. \tag{34}$$

This finishes the proof of Lemma 4.3.

### 6.6 Proof of Theorem 2.5

We first remind the readers of the following two definitions:

$$(\lambda^*, \eta^*) = \text{argmin}_{\lambda \in [\lambda_{\min}, 1), \eta \in [\eta_{\min}, \infty)} \text{OO}(\lambda, \eta).$$

and

$$(\widehat{\lambda}, \widehat{\eta}) = \text{argmin}_{\lambda \in [\lambda_{\min}, \infty), \eta \in [\eta_{\min}, 1)} \text{LO}(\lambda, \eta).$$

For simplicity we denote $\mathcal{K} = \{(\lambda, \eta) : \lambda_{\min} \leq \lambda \leq \infty, \eta_{\min} \leq \eta \leq 1\}$ to be the parameter space of interest. We also define $\widehat{\boldsymbol{\beta}}_{/0} := \widehat{\boldsymbol{\beta}}$, so by stating "$\forall 0 \leq i \leq n$, $\widehat{\boldsymbol{\beta}}_{/i}$ satisfy property A", we mean property A is true for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{/i}$ for all $i$.

Our first goal is to establish uniform consistency of $\text{LO}_{\lambda,\eta}$ on $\mathcal{K}$, i.e. $\forall \epsilon > 0$,

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} |\text{LO}(\lambda, \eta) - \text{OO}(\lambda, \eta)| > \epsilon) \to 0.$$

The following is a heuristic summary of our proof strategy. Suppose that the event $G_n(\epsilon) := \{\omega : \text{LO and OO are } C_L\text{PolyLog}(n) - \text{Lipschitz in } (\lambda, \eta) \in \mathcal{K}\}$ has probability $1 - q_L$. Then we divide the region $\mathcal{K}$ by squares with edge length $a_H = \frac{\epsilon}{6C_L \text{PolyLog}(n)}$. Let $\mathcal{H}$ be the collection of the centers of such squares. We have $|\mathcal{H}| \leq 2(\lambda_{\max} - \lambda_{\min} + 1)a_H^{-2} := C_H \epsilon^{-2}$. For $(\lambda, \eta) \in \mathcal{K}$, let $(\lambda_h, \eta_h)$ be the closest element in $\mathcal{K}$. Notice that for $\epsilon \in (0, 1]$, under the event $G_n(\epsilon)$:

- $\sup_{(\lambda,\eta)\in\mathcal{K}} |\text{LO}_{\lambda,\eta} - \text{LO}_{\lambda_h,\eta_h}| \leq 2a_H C_L \text{PolyLog}(n) \leq \epsilon/3$ and similarly $\sup_{(\lambda,\eta)\in\mathcal{K}} |\text{OO}_{\lambda,\eta} - \text{OO}_{\lambda_h,\eta_h}| \leq \epsilon/3$.
- $\sup_{(\lambda,\eta)\in\mathcal{K}} |\text{LO}_{\lambda,\eta} - \text{OO}_{\lambda,\eta}| \leq \sup_{(\lambda,\eta)\in\mathcal{K}} |\text{LO}_{\lambda,\eta} - \text{LO}_{\lambda_h,\eta_h}| + \sup_{(\lambda,\eta)\in\mathcal{K}} |\text{OO}_{\lambda,\eta} - \text{OO}_{\lambda_h,\eta_h}| + \max_{(\lambda_h,\eta_h)\in\mathcal{H}} |\text{LO}_{\lambda_h,\eta_h} - \text{OO}_{\lambda_h,\eta_h}| \leq 2/3\epsilon + \max_{(\lambda_h,\eta_h)\in\mathcal{H}} |\text{LO}_{\lambda_h,\eta_h} - \text{OO}_{\lambda_h,\eta_h}|$

It then follows that

$$\begin{aligned}
\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} |\text{LO} - \text{OO}| > \epsilon) &\leq q_L + \mathbb{P}(\max_{(\lambda,\eta)\in\mathcal{H}} |\text{LO} - \text{OO}| > \epsilon/3) \\
&\leq q_L + \sum_{j \leq |\mathcal{H}|} \mathbb{P}(|\text{LO}(\lambda_j, \eta_j) - \text{OO}(\lambda_j, \eta_j)| \geq 3\epsilon) \\
&\leq q_L + \sum_{j \leq |\mathcal{H}|} \frac{\mathbb{E}[\text{LO}(\lambda_j, \eta_j) - \text{OO}(\lambda_j, \eta_j)]^2}{\epsilon^2/9} \\
&\leq q_L + \frac{C_H}{\epsilon^2} \frac{9C_v}{n\epsilon^2} \\
&= q_L + 9C_H \epsilon^{-4} \frac{\text{PolyLog}(n)}{n} \to 0.
\end{aligned}$$

The second line uses a union bound, the third uses Chebyshev inequality, and the 4th line uses 2.1 and the formula of $|\mathcal{H}|$.

In the actual proof however, the Lipschitzness of LO does not always hold, so we prove in Lemma 6.19 that it is 'nearly' Lipschitz. To prove this fact we will require the results of Lemma 6.15, Lemma 6.17 and Lemma 6.18. The goal of this subsection is the following theorem that can be used for proving Theorem 2.5:

**Theorem 6.14.** *Suppose Assumptions A1-A4 hold. Then* $\forall t \in (0,1], \exists C_7, C_8 > 0$ *such that for large enough* $n$,

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{OO}_{\lambda,\eta}| > t) \le 1 - 2ne^{-p/2} - \frac{2c_1+c_2}{n^2} - \frac{C_7\mathrm{PolyLog}(n)}{t^4 n} - \frac{C_8}{n^2\log(n)}.$$

*where* $c_1, c_2$ *are the two constants in Lemma 6.6.*

As mentioned in the heuristic discussion of the proof, proving Theorem 6.14 requires several auxiliary lemmas, one of which is presented below.

**Lemma 6.15.** *Let* $C_r$ *be the Lipschitz constant of* $r_0$ *in Assumption A3. Define the event* $E_n := \{\omega : \sup_{\mathcal{K}} \max_{0\le i\le n} \frac{1}{p}\|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}\|_2^2 > C_1\}$, *where* $C_1 = \frac{\gamma_0}{\lambda_{\min}\eta_{\min}}(2 + C_Y(s))$. *Then, under Assumptions A1-A4, we have:*

1. $\mathbb{P}(E_n) \le p_1 = \frac{c_1}{n^4}$ *where* $c_1$ *is defined in Lemma 6.6.*
2. *Under the event* $E_n^c$ , $\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}$ *is* $C_2\sqrt{n}$*-Lipschitz in* $(\lambda,\eta) \in \mathcal{K}$ *for all* $0 \le i \le n$, *where* $C_2 = \max\{\frac{1}{2\lambda_{\min}\eta_{\min}}, 1, \frac{1}{2\eta_{\min}}, \lambda_{\max}\} \cdot \frac{2(C_r+C_1)}{\sqrt{\gamma_0}}$.

*Proof.* 1. Suppose that $0 \notin \boldsymbol{\Theta}$. First, since $\widehat{\boldsymbol{\beta}}_{\lambda,\eta}$ is the minimizer of $h(\boldsymbol{\beta})$,

$$\lambda\eta\|\widehat{\boldsymbol{\beta}}_{\lambda,\eta}\|^2 \le \sum_i \ell_i(\widehat{\boldsymbol{\beta}}_{\lambda,\eta}) + \lambda(1-\eta)r_0(\widehat{\boldsymbol{\beta}}_{\lambda,\eta}) + \lambda\eta\|\widehat{\boldsymbol{\beta}}_{\lambda,\eta}\|^2 \le \sum_i \ell_i(0).$$

and similarly $\lambda\eta\|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}\|^2 \le \sum_{j\neq i}\ell_j(0) \le \sum_i \ell_i(0)$. Then $\forall t > 0$,

$$\begin{aligned}\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \max_{0\le i\le n} \frac{1}{p}\|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}\|_2^2 > t) &\le \mathbb{P}(\frac{1}{p\lambda\eta}\sum_i \ell_i(0) > t)\\ &\le \mathbb{P}(\frac{1}{p\lambda_{\min}\eta_{\min}}\sum_i (1+|y_i|^s)) > t)\\ &= \mathbb{P}(\frac{1}{n}\sum_i |y_i|^s > \frac{\lambda_{\min}\eta_{\min}t}{\gamma_0} - 1),\end{aligned}$$

where the second line uses Assumption A4 (with the constant $C$ assumed to be 1 without loss of generality). By Lemma 6.6 and Markov inequality we have

$$\mathbb{P}(\frac{1}{n}\sum_i |y_i|^s > \mathbb{E}|y_i|^s + 1) \le \mathbb{E}(\frac{1}{n}\sum_i |y_i|^s - \mathbb{E}|y_1|^s)^8 \le \frac{c_1}{n^4}$$

Let $\frac{\lambda_{\min}\eta_{\min}t}{\gamma_0} - 1 = C_Y(s) + 1$, or equivalently $t = \frac{\gamma_0}{\lambda_{\min}\eta_{\min}}(2 + C_Y(s)) := C_1$, then we have

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \max_{0\le i\le n} \frac{1}{p}\|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}\|_2^2 > C_1) \le \frac{c_1}{n^4}.$$

2. The proof is similar to that of Lemma 6.8 in Section 6.5.3. In the following proof, we drop the irrelevant subscript(s) of $\widehat{\boldsymbol{\beta}}$, e.g. we use $\widehat{\boldsymbol{\beta}}_\lambda$ instead of $\widehat{\boldsymbol{\beta}}_{\lambda,\eta}$ if $\eta$ is held fixed and not of the current interest.

   For now we fix $\eta$ and, for any $\lambda, \tilde{\lambda} \in [\lambda_{\min}, \lambda_{\max}]$, consider the smoothed estimators $\widehat{\boldsymbol{\beta}}_\lambda^\alpha, \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha$ defined in Section 4.1, where $\alpha$ is the smoothing parameter. By Lemma 6.1, they satisfy

$$\widehat{\boldsymbol{\beta}}_\lambda^\alpha = \Pi_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \nabla h_\lambda^\alpha(\widehat{\boldsymbol{\beta}}_\lambda^\alpha)) \tag{35}$$

$$\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha = \Pi_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha - \nabla h_{\tilde{\lambda}}^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha)) \tag{36}$$

where

$$\nabla h_\lambda^\alpha(\widehat{\boldsymbol{\beta}}_\lambda^\alpha) = \sum_i \ell_i(\widehat{\boldsymbol{\beta}}_\lambda^\alpha)\mathbf{x}_i + \lambda(1-\eta)\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_\lambda^\alpha) + 2\lambda\eta\widehat{\boldsymbol{\beta}}_\lambda^\alpha$$
$$\nabla h_{\tilde{\lambda}}^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) = \sum_i \ell_i(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha)\mathbf{x}_i + \tilde{\lambda}(1-\eta)\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) + 2\tilde{\lambda}\eta\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha$$

. By subtracting (36) from (35), we have

$$\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha = \bar{\mathbf{J}} \times (\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha - \nabla h_\lambda^\alpha(\widehat{\boldsymbol{\beta}}_\lambda^\alpha) + \nabla h_{\tilde{\lambda}}^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha))$$

where $\bar{\mathbf{J}} = \int_0^1 \mathbf{J}(t)dt$ with $\mathbf{J}(t)$ being the matrix in Lemma 6.2.
By mean value theorem as was used in Section 6.5.3,

$$\begin{aligned}
&\nabla h_\lambda^\alpha(\widehat{\boldsymbol{\beta}}_\lambda^\alpha) - \nabla h_{\tilde{\lambda}}^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) \\
=&(\mathbf{X}^\top \operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\in[n]}\mathbf{X} + \lambda(1-\eta)\nabla^2 r_0^\alpha(\boldsymbol{\Xi}) + 2\lambda\eta\mathbb{I})(\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) \\
&+ (\tilde{\lambda}-\lambda)[(1-\eta)\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) + 2\eta\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha].
\end{aligned}$$

so

$$\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha = \mathbf{G}^{-1}\bar{\mathbf{J}}(\lambda-\tilde{\lambda})[(1-\eta)\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) + 2\eta\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha],$$

where $\mathbf{G} = \mathbb{I} - \bar{\mathbf{J}} + \bar{\mathbf{J}}\left(\mathbf{X}^\top \operatorname{diag}[\ddot{\ell}_j(\boldsymbol{\xi}_j)]_{j\in[n]}\mathbf{X} + \lambda(1-\eta)\nabla^2 r_0^\alpha(\boldsymbol{\Xi}) + 2\lambda\eta\mathbb{I}\right)$., and by (21) we have $\|\mathbf{G}^{-1}\| \le \frac{1}{2\lambda\eta\wedge 1} \le \frac{1}{2\lambda_{\min}\eta_{\min}\wedge 1}$. Therefore we have

$$\|\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha\| \le \frac{1}{2\lambda_{\min}\eta_{\min}\wedge 1}|\lambda-\tilde{\lambda}|[(1-\eta)\|\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha)\|_2 + 2\eta\|\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha\|].$$

Under event $E_n^c$ (and the fact that $\eta \le 1$), $2\eta\|\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha\| \le 2C_1\sqrt{p}$. Moreover, since $r_0$ is $C_r-$Lipschitz, we have $\|\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha)\|_2 \le 2C_r\sqrt{p}$ (for large enough $\alpha$). Putting these together we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_\lambda^\alpha - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha\| &\le \frac{1}{2\lambda_{\min}\eta_{\min}\wedge 1}|\lambda-\tilde{\lambda}|[2C_r\sqrt{p} + 2C_1\sqrt{p}] \\
&\le C_2\sqrt{n}|\lambda-\tilde{\lambda}|.
\end{aligned}$$

Finally, by letting $\alpha \to \infty$ and using Lemma 6.13, we have

$$\|\widehat{\boldsymbol{\beta}}_\lambda - \widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}\| \le C_2\sqrt{n}|\lambda-\tilde{\lambda}|.$$

The proof for the Lipschitzness in $\eta$ is similar:

$$\widehat{\boldsymbol{\beta}}_\eta - \widehat{\boldsymbol{\beta}}_{\tilde{\eta}} = \mathbf{G}^{-1}\bar{\mathbf{J}}\lambda(\eta-\tilde{\eta})[\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha) - 2\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha]$$

so under event $E_n^c$,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_\eta - \widehat{\boldsymbol{\beta}}_{\tilde{\eta}}\| &\le \frac{\lambda}{2\lambda\eta\wedge 1}|\eta-\tilde{\eta}|[\|\nabla r_0^\alpha(\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha)\|_2 + 2\|\widehat{\boldsymbol{\beta}}_{\tilde{\lambda}}^\alpha\|] \\
&\le \frac{\lambda_{\max}}{2\lambda_{\min}\eta_{\min}\wedge 1}|\eta-\tilde{\eta}|[2C_r\sqrt{n} + 2C_1\sqrt{n}] \\
&\le C_2\sqrt{n}|\eta-\tilde{\eta}|
\end{aligned} \tag{37}$$

□

*Remark* 6.16. The only place $\lambda_{\max}$ appears is in Equation (37), which arises from the existence of the constraint $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. For an unconstrained problem, *i.e.*$\boldsymbol{\beta} \in \mathbb{R}^p$, all $\lambda\eta\wedge 1$ are replaced by $\lambda\eta$, thus the constant in front of (37) becomes $\frac{\lambda}{2\lambda\eta} \le \frac{1}{2\eta_{\min}}$ so that $\lambda_{\max}$ wouldn't occur.

The following two lemmata are two key steps in establishing the (nearly) Lipschitzness of OO and LO.

**Lemma 6.17.** *Under Assumptions A1-A4, $\forall \epsilon \in (0,1], \exists C_3 > 0$ such that*

$$\begin{aligned}&\mathbb{P}(\sup_{(\lambda,\eta),(\tilde{\lambda},\tilde{\eta})\in\mathcal{K}} \max_{0\le i\le n} |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda},\tilde{\eta}})| \le C_3\sqrt{\log(n)}(4\sqrt{n}\epsilon + |\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}|))\\ &\ge 1 - ne^{-p/2} - \frac{c_1}{n^4} - \frac{2C_H^2(c_1+1)}{\epsilon^4 n^3}\end{aligned}$$

*Proof.* We use the so-called '$\epsilon-$net argument', as we have introduced in the beginning of this subsection, to obtain the uniform concentration. Again, we divide $\mathcal{K}$ by squares with edge length $2\epsilon$ and let $\mathcal{H}$ be the collections of square centers. For $0 < \epsilon \le 1$, we know $|\mathcal{H}| \le \frac{\lambda_{\max}-\lambda_{\min}+1}{2\epsilon}\frac{1+1}{2\epsilon} = C_H\epsilon^{-2}$. For all $(\lambda,\eta)\in\mathcal{K}$, let $(\lambda_h,\eta_h)$ be the closest element in $\mathcal{H}$. By triangular inequality we have

$$|\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda},\tilde{\eta}})| \le |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h})| + |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda},\tilde{\eta}} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda}_h,\tilde{\eta}_h})| + |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda}_h,\tilde{\eta}_h})|.$$

For $t_1, t_2 > 0$, define the following two events:

$$\begin{aligned}F_1(t_1) &:= \{\omega : \exists(\lambda,\eta)\in\mathcal{K}, \exists 0\le i\le n : |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h})| > t_1\}\\ F_2(t_2) &:= \{\omega : \exists(\lambda_h,\eta_h),(\tilde{\lambda}_h,\tilde{\eta}_h)\in\mathcal{K}, \exists 0\le i\le n : |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda}_h,\tilde{\eta}_h})| > t_2\}.\end{aligned}$$

Then the probability we want to calculate is just

$$\mathbb{P}(\sup_{(\lambda,\eta),(\tilde{\lambda},\tilde{\eta})\in\mathcal{K}} \max_{0\le i\le n} |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda},\tilde{\eta}})| > 2t_1 + t_2) \le \mathbb{P}(F_1(t_1)) + \mathbb{P}(F_2(t_2)),$$

for some $t_1, t_2$ that will be determined later. For $\mathbb{P}(F_1(t_1))$, let $t_1 = 4\sqrt{C_X}C_2\sqrt{n}\epsilon$, we have

$$\begin{aligned}\mathbb{P}(F_1(t_1) &\le \mathbb{P}(\exists(\lambda,\eta)\in\mathcal{K}, \exists 0\le i\le n : \|\mathbf{x}_i\|\|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h}\| > 4\sqrt{C_X}C_2\sqrt{n}\epsilon)\\ &\le \mathbb{P}(\exists 0\le i\le n : \|\mathbf{x}_i\| > 2\sqrt{C_X}) + \mathbb{P}(\exists\lambda,\eta,i : \|\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h}\| > 2C_2\sqrt{n}\epsilon)\\ &\le ne^{-p/2} + \frac{c_1}{n^4},\end{aligned} \tag{38}$$

where the first line uses Cauchy-Schwarts inequality, the last line uses part 1 of Lemma 6.5, and also part 1 of Lemma 6.15.

For $\mathbb{P}(F_2(t_2))$, by a union bound on the probability we have

$$\begin{aligned}\mathbb{P}(F_2(t_2)) &\le |\mathcal{H}|^2 n\mathbb{P}(|\mathbf{x}_1^\top(\widehat{\boldsymbol{\beta}}_{/1,\lambda_h,\eta_h} - \widehat{\boldsymbol{\beta}}_{/1,\tilde{\lambda}_h,\tilde{\eta}_h})| > t_2)\\ &\le \frac{C_H^2 n}{\epsilon^4}\mathbb{E}\, 2\exp\left\{-\frac{t_2^2}{2C_X p^{-1}\|\widehat{\boldsymbol{\beta}}_{/1,\lambda_h,\eta_h} - \widehat{\boldsymbol{\beta}}_{/1,\tilde{\lambda}_h,\tilde{\eta}_h}\|^2}\right\}\end{aligned}$$

where the second line uses the standard Gaussian concentration $\mathbb{P}(|Z| > t) \le 2\exp -t^2/2$ for $Z\sim N(0,1)$, conditioned on $\mathcal{D}_{/1}$.

Notice that by Lemma 6.15, under the event $E_n^c$ with probability $1 - c_1 n^{-4}$,

$$\|\widehat{\boldsymbol{\beta}}_{/1,\lambda_h,\eta_h} - \widehat{\boldsymbol{\beta}}_{/1,\tilde{\lambda}_h,\tilde{\eta}_h}\| \le C_2\sqrt{n}(|\lambda_h - \tilde{\lambda}_h| + |\eta_h - \tilde{\eta}_h|) \le C_2\sqrt{n}(|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}| + 4\epsilon).$$

Now let $t_2^2 = 8C_X p^{-1}\log(n)[C_2\sqrt{n}(|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}| + 4\epsilon)]^2$, i.e. $t_2 = \sqrt{8\gamma_0 C_X\log(n)}C_2(|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}| + 4\epsilon)$, then we have

$$\begin{aligned}\mathbb{P}(F_2(t_2)) &\le \frac{2C_H^2 n}{\epsilon^4}[\mathbb{P}(E_n) + \exp\{-4\log(n)\}]\\ &\le \frac{2C_H^2(c_1+1)}{\epsilon^4 n^3}.\end{aligned} \tag{39}$$

Combining (38) and (39), with probability at least $1 - ne^{-p/2} - \frac{c_1}{n^4} - \frac{2C_H^2(c_1+1)}{\epsilon^4 n^3}$, we have that $\forall \epsilon \in (0,1], \forall (\lambda,\eta),(\tilde{\lambda},\tilde{\eta}) \in \mathcal{K}, \forall 0 \le i \le n$:

$$\begin{aligned}
|\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\tilde{\lambda},\tilde{\eta}})| &\le 8\sqrt{C_X}C_2\sqrt{n}\epsilon + 2\sqrt{\gamma_0 C_X \log(n)}C_2(|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}| + 4\epsilon) \\
&= 8C_2\sqrt{C_X}(\sqrt{n} + \sqrt{\gamma_0\log(n)})\epsilon + 2\sqrt{\gamma_0 C_X\log(n)}C_2(|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}|) \\
&\le 4C_3\sqrt{n\log(n)}\epsilon + C_3\sqrt{\log(n)}|\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}|) \\
&= C_3\sqrt{\log(n)}(4\sqrt{n}\epsilon + |\lambda-\tilde{\lambda}| + |\eta-\tilde{\eta}|).
\end{aligned}$$

where the penultimate line holds for large enough $n$ such that $\sqrt{n} + \sqrt{\gamma_0\log(n)} \le \sqrt{n\gamma_0\log(n)}$, and $C_3 = 2C_2\sqrt{\gamma_0 C_X}$. This finishes the proof. □

**Lemma 6.18.** *Under Assumptions A1-A4, the following statements hold:*

1. $\forall \epsilon \in (0,1] : \mathbb{P}(\exists \lambda,\eta,i : |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}| > 3\sqrt{C_X C_1 \log(n)}) \le ne^{-p/2} + \frac{c_1}{n^4} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}$
2. $\mathbb{P}(\exists\lambda,\eta : \sqrt{\frac{1}{n}\sum_i \phi_i^2(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta})} > C_5\text{PolyLog}(n)) \le \frac{c_1+c_2}{n^4} + ne^{-p/2} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}$, *where* $C_5 = \sqrt{3C_Y(2s)+6} + 3^{s+0.5}(C_X C_1)^{s/2}$, *where* $c_2$ *is the constant in Lemma 6.6.*

*Proof.* 1. Again let $\mathcal{H}$ be the collection of square centers that divide $\mathcal{K}$ with edge length $2\epsilon$. $\forall t_1, t_2 > 0$, define the following events:

$$\begin{aligned}
F_1(t_1) &= \{\omega : \exists\lambda,\eta,i : |\mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta} - \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h})| > t_1\} \\
F_3(t_3) &= \{\omega : \exists(\lambda_h,\eta_h) \in \mathcal{H}, \exists 0 \le i \le n : |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h}| > t_3\}.
\end{aligned}$$

Then

$$\mathbb{P}(\exists\lambda,\eta,i : |\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}| > t_1 + t_3) \le \mathbb{P}(F_1(t_1)) + \mathbb{P}(F_3(t_3)).$$

Using (38) we have for $t_1 = 4\sqrt{C_X}C_2\sqrt{n}\epsilon$,

$$\mathbb{P}(F_1(t_1)) \le ne^{-p/2} + \frac{c_1}{n^4}.$$

For $F_3(t_3)$, by union bound and standard univariate Gaussian concentration,

$$\mathbb{P}(F_3(t_3)) \le |\mathcal{H}|n\mathbb{E}\, 2\exp\left\{-\frac{t_3^2}{2C_X p^{-1}\|\widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h}\|^2}\right\}$$

Under the event $E_n$ defined in Lemma 6.15, $\|\widehat{\boldsymbol{\beta}}_{/i,\lambda_h,\eta_h}\| \le C_1\sqrt{p}$. Let $t_2^2 = 4C_X C_1\log(n)$, we have

$$\mathbb{P}((F_3(t_3)) \le \frac{2nC_H}{\epsilon^2}(\mathbb{P}(E_n) + \frac{1}{n^4}) = \frac{2nC_H(c_1+1)}{\epsilon^2 n^4}$$

Finally let $t_1 = \frac{1}{2}t_2$, i.e. $4\sqrt{C_X}C_2\sqrt{n}\epsilon = \sqrt{C_X C_1\log(n)} \Leftrightarrow \epsilon = \frac{1}{4}\sqrt{\frac{C_1\log(n)}{C_2^2 n}}$, we have

$$t_1 + t_2 = 3\sqrt{C_X C_1\log(n)}$$

and

$$\begin{aligned}
\mathbb{P}(\exists\lambda,\eta,i : |\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}| > 3\sqrt{C_X C_1\log(n)}) &\le ne^{-p/2} + \frac{c_1}{n^4} + \frac{2C_H(c_1+1)}{\frac{1}{16}\frac{C_1\log(n)}{C_2^2 n}n^3} \\
&= ne^{-p/2} + \frac{c_1}{n^4} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}.
\end{aligned}$$

This finishes the proof of Part 1.

2. Notice that

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_i \dot{\phi}_i^2(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta})} &\le \sqrt{\frac{1}{n}\sum_i (1+|y_i|^s+|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^s)^2} \\ &\le \sqrt{\frac{3}{n}\sum_i (1+|y_i|^{2s}+|\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^{2s})} \\ &\le \sqrt{\frac{3}{n}\sum_i (1+|y_i|^{2s})}+\sqrt{\frac{3}{n}\sum_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^{2s}},\end{aligned}$$

where the first line uses Assumption A4, the second line uses $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, and the last line uses $\sqrt{a+b} \le \sqrt{a}+\sqrt{b},\ \forall a,b \ge 0$.

It then follows that

$$\begin{aligned}&\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \sqrt{\frac{1}{n}\sum_i \dot{\phi}_i^2(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta})} > t_1+t_2) \\ \le&\mathbb{P}(\sqrt{\frac{3}{n}\sum_i (1+|y_i|^{2s})} > t_1) + \mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \sqrt{\frac{3}{n}\sum_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^{2s}} > t_2). \end{aligned} \tag{40}$$

The first term is just

$$\mathbb{P}(\sqrt{\frac{3}{n}\sum_i (1+|y_i|^{2s})} > t_1) = \mathbb{P}(\frac{1}{n}\sum_i |y_i|^{2s} > \frac{t_1^2}{3}-1).$$

Let $\frac{t_1^2}{3}-1 = \mathbb{E}|y_1|^{2s}+1 = C_Y(2s)+1$, i.e. $t_1 = \sqrt{3(C_Y(2s)+2)}$, and by Markov inequality we have

$$\mathbb{P}(\sqrt{\frac{3}{n}\sum_i (1+|y_i|^{2s})} > t_1) \le \mathbb{E}[\frac{1}{n}\sum_i |y_i|^{2s} - \mathbb{E}|y_1|^{2s}]^8 \le \frac{c_2}{n^4}. \tag{41}$$

where $c_2$ is the constant in Lemma 6.6

For the second term,

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \sqrt{\frac{3}{n}\sum_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^{2s}} > t_2) \le \mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \max_{0\le i\le n} |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}| > \left(\frac{t_2}{\sqrt{3}}\right)^{1/s}).$$

Let $\left(\frac{t_2}{\sqrt{3}}\right)^{1/s} = 3\sqrt{C_X C_1 \log(n)}$ i.e. $t_2 = 3^{s+0.5}(C_X C_1 \log(n))^{s/2}$, by Part 1 we have

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \sqrt{\frac{3}{n}\sum_i |\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta}|^{2s}} > 3^{s+0.5}(C_X C_1 \log(n))^{s/2}) \le ne^{-p/2} + \frac{c_1}{n^4} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}. \tag{42}$$

For large enough $n$:

$$t_1+t_2 = \sqrt{3(C_Y(2s)+2)} + 3^{s+0.5}(C_X C_1\log(n))^{s/2} \le C_5 \mathrm{PolyLog}(n)$$

where $C_5 = \sqrt{3(C_Y(2s)+2)} + 3^{s+0.5}(C_X C_1)^{s/2}$, provided that $n \ge 3$.

Finally combining (40), (41) and (42) we have

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} \sqrt{\frac{1}{n}\sum_i \dot{\phi}_i^2(\widehat{\boldsymbol{\beta}}_{/i,\lambda,\eta})} > C_5\mathrm{PolyLog}(n)) \le ne^{-p/2} + \frac{c_1+c_2}{n^4} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}.$$

This finishes the proof of Part 2.

□

The next lemma show that with high probability OO is Lipschitz in $(\lambda, \eta)$, and LO is nearly Lipschitz.

**Lemma 6.19.** *Under Assumptions A1-A4, for $\epsilon \in (0,1]$, Let $G(\epsilon)$ denote the event that either of the following statements are false:*

- $\mathrm{OO}_{\lambda,\eta}$ *is $C_6$-Lipschitz in $\lambda$ and $\eta$, where $C_6 = C_\phi(1 + C_1^{s/2})\sqrt{C_X\gamma_0}C_2$.*
- $\forall(\lambda,\eta),(\tilde\lambda,\tilde\eta) \in \mathcal{K} : |\mathrm{LO}_{\lambda,\eta} - \mathrm{LO}_{\tilde\lambda,\tilde\eta}| \le C_3C_5\mathrm{PolyLog}(n)(4\sqrt{n}\epsilon + |\lambda-\tilde\lambda| + |\eta-\tilde\eta|)$.

*Then*

$$\mathbb{P}(G(\epsilon)) \le 2ne^{-p/2} + \frac{c_1+c_2}{n^4} + \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2}$$

*Proof.* On the event $E_n^c$ defined in Lemma 6.15 that has probability at least $1 - \frac{c_1}{n^4}$, fix $\eta \in [\eta_{\min}, 1]$ (thus we suppress the subscript $\eta$) and consider $\lambda, \tilde\lambda \in [\lambda_{\min}, \lambda_{\max}]$:

$$\begin{aligned}
|\mathrm{OO}_\lambda - \mathrm{OO}_{\tilde\lambda}| &= |\mathbb{E}\left[\phi_0(\widehat{\boldsymbol\beta}_\lambda) - \phi_0(\widehat{\boldsymbol\beta}_{\tilde\lambda})|\mathcal{D}\right]| \\
&= |\mathbb{E}\left[\dot\phi_0(\boldsymbol\xi)\mathbf{x}_0^\top(\widehat{\boldsymbol\beta}_\lambda - \widehat{\boldsymbol\beta}_{\tilde\lambda})|\mathcal{D}\right]| \\
&\le \sqrt{\mathbb{E}[\dot\phi_0^2(\boldsymbol\xi)|\mathcal{D}]}\cdot\sqrt{\mathbb{E}[(\widehat{\boldsymbol\beta}_\lambda - \widehat{\boldsymbol\beta}_{\tilde\lambda})^\top\mathbf{x}_0\mathbf{x}_0^\top(\widehat{\boldsymbol\beta}_\lambda - \widehat{\boldsymbol\beta}_{\tilde\lambda})|\mathcal{D}]} \\
&\le \left[C_\phi + C_\phi(p^{-1}\|\boldsymbol\xi\|^2)^{s/2}\right]\sqrt{C_Xp^{-1}\|\widehat{\boldsymbol\beta}_\lambda - \widehat{\boldsymbol\beta}_{\tilde\lambda}\|^2} \\
&\le C_\phi(1 + C_1^{s/2})\sqrt{C_X/p}C_2\sqrt{n}|\lambda-\tilde\lambda| \\
&\le C_\phi(1 + C_1^{s/2})\sqrt{C_X\gamma_0}C_2|\lambda-\tilde\lambda| \\
&:= C_6|\lambda-\tilde\lambda|,
\end{aligned}$$

where $C_6 = C_\phi(1 + C_1^{s/2})C_2\sqrt{C_X\gamma_0}$. The second line uses mean-value theorem where $\boldsymbol\xi$ is a convex combination of $\widehat{\boldsymbol\beta}_\lambda$ and $\widehat{\boldsymbol\beta}_{\tilde\lambda}$. The third line uses Cauchy-Schwarts inequality. The 4th line uses Lemma 6.9 and also the fact that $\mathbb{E}\mathbf{x}_0\mathbf{x}_0^\top = \boldsymbol\Sigma \le C_Xp^{-1}\mathbb{I}$. The 5th line uses the fact that under $E_n^c$, $\|\widehat{\boldsymbol\beta}_\lambda\|$ and $\|\widehat{\boldsymbol\beta}_{\tilde\lambda}\|$ are bounded by $C_1\sqrt{p}$, and so is $\|\boldsymbol\xi\|$. It also uses Lemma 6.17 to bound $\|\widehat{\boldsymbol\beta}_\lambda - \widehat{\boldsymbol\beta}_{\tilde\lambda}\|$.

Using the same arguments one can also show that, fix $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, for $\eta, \tilde\eta \in [\eta_{\min}, 1]$:

$$|\mathrm{OO}_\eta - \mathrm{OO}_{\tilde\eta}| \le C_6|\eta - \tilde\eta|.$$

For LO, under the joint event of Lemma 6.17 and Lemma 6.18, $\forall\epsilon \in (0,1]$, with probability at least

$$\begin{aligned}
&1 - ne^{-p/2} - \frac{c_1}{n^4} - \frac{2C_H^2(c_1+1)}{\epsilon^4n^3} - ne^{-p/2} - \frac{c_1+c_2}{n^4} - \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2} \\
=&1 - 2ne^{-p/2} - \frac{2c_1+c_2}{n^4} - \frac{2C_H^2(c_1+1)}{\epsilon^4n^3} - \frac{32C_H(c_1+1)C_2^2}{C_1\log(n)n^2},
\end{aligned}$$

the following hold:

$$\begin{aligned}
|\mathrm{LO}_{\lambda,\eta} - \mathrm{LO}_{\tilde\lambda,\tilde\eta}| &= \frac{1}{n}\sum_i|\phi_i(\widehat{\boldsymbol\beta}_{\lambda,\eta}) - \phi_i(\widehat{\boldsymbol\beta}_{\tilde\lambda,\tilde\eta})| \\
&= \frac{1}{n}\sum_i\dot\phi_i(\boldsymbol\xi_i)\mathbf{x}_i^\top(\widehat{\boldsymbol\beta}_{\lambda,\eta} - \widehat{\boldsymbol\beta}_{\tilde\lambda,\tilde\eta}) \\
&\le \sqrt{\frac{1}{n}\sum_i\dot\phi_i^2(\boldsymbol\xi_i)}\cdot\sqrt{\frac{1}{n}\sum_i\mathbf{x}_i^\top(\widehat{\boldsymbol\beta}_{\lambda,\eta} - \widehat{\boldsymbol\beta}_{\tilde\lambda,\tilde\eta})} \\
&\le C_5\mathrm{PolyLog}(n)C_3\sqrt{\log(n)}(4\sqrt{n}\epsilon + |\lambda-\tilde\lambda| + |\eta-\tilde\eta|) \\
&\le C_3C_5\mathrm{PolyLog}(n)(4\sqrt{n}\epsilon + |\lambda-\tilde\lambda| + |\eta-\tilde\eta|).
\end{aligned}$$

The second line uses mean-value theorem where $\boldsymbol{\xi}_i$ is a convex combination of $\widehat{\boldsymbol{\beta}}_{\lambda,\eta}$ and $\widehat{\boldsymbol{\beta}}_{\tilde{\lambda},\tilde{\eta}}$, the third line uses Cauchy-Schwartz inequality and the fourth line uses Lemma 6.17 and Lemma 6.18. □

Using the above lemmata, we are ready to prove the main theorem in this subsection:

**Theorem 6.20.** *Under Assumptions A1-A4,* $\forall t \in (0,1], \exists C_7, C_8 > 0$ *such that*

$$\mathbb{P}(\sup_{(\lambda,\eta)\in\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{OO}_{\lambda,\eta}| > t) \leq 1 - 2ne^{-p/2} - \frac{2c_1 + c_2}{n^2} - \frac{C_7 \mathrm{PolyLog}(n)}{t^4 n} - \frac{C_8}{n^2 \log(n)}.$$

*Proof.* The proof, without any surprise, starts with an $\epsilon-$net argument. Let $\tilde{\mathcal{H}}$ be the collections of the centers of squares with edge length $2\tilde{\epsilon}$ that divide $\mathcal{K}$, and for $(\lambda,\eta) \in \mathcal{K}$ let $(\lambda_h, \eta_h)$ be the closest element to it in $\tilde{\mathcal{H}}$. Define the event

$$F_4(t_4) = \{\omega : \max_{(\lambda_h,\eta_h)\in\tilde{\mathcal{H}}} |\mathrm{LO}_{\lambda_h,\eta_h} - \mathrm{OO}_{\lambda_h,\eta_h}| > t_4\},$$

then

$$\begin{aligned}
\mathbb{P}(F_4(t_4)) &\leq |\tilde{\mathcal{H}}| \mathbb{P}(|\mathrm{LO}_{\lambda_1,\eta_1} - \mathrm{OO}_{\lambda_1,\eta_1}| > t_4) \\
&\leq \frac{C_H}{\tilde{\epsilon}^2} \frac{\mathbb{E}[\mathrm{LO}_{\lambda_1,\eta_1} - \mathrm{OO}_{\lambda_1,\eta_1}]^2}{t_4^2} \\
&\leq \frac{C_H C_v}{\tilde{\epsilon}^2 t_4^2 n},
\end{aligned}$$

where the first line is a union bound, the second line uses Markov inequality, and the last line uses Theorem 2.1.

Let $G(\epsilon)$ denote the event in Lemma 6.19. Then under $F_4^c(t_4) \cup G^c(\epsilon)$, we have

$$\begin{aligned}
\sup_{\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{OO}_{\lambda,\eta}| &\leq \sup_{\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{LO}_{\lambda_h,\eta_h}| + \sup_{\mathcal{K}} |\mathrm{OO}_{\lambda,\eta} - \mathrm{OO}_{\lambda_h,\eta_h}| + \max_{\mathcal{H}} |\mathrm{LO}_{\lambda_h,\eta_h} - \mathrm{OO}_{\lambda_h,\eta_h}| \\
&\leq C_3 C_5 \mathrm{PolyLog}(n)(4\sqrt{n}\epsilon + 2\tilde{\epsilon}) + 2C_6\tilde{\epsilon} + t_4
\end{aligned}$$

Now let $\tilde{\epsilon} = 4\sqrt{n}\epsilon = \frac{t/2}{3C_3C_5\mathrm{PolyLog}(n)+2C_6}$ and $t_4 = t/2$, we have

$$\sup_{\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{OO}_{\lambda,\eta}| \leq C_3 C_5 \mathrm{PolyLog}(n) \cdot 3\tilde{\epsilon} + 2C_6\tilde{\epsilon} + t/2 = t/2 + t/2 = t,$$

so that

$$\begin{aligned}
\mathbb{P}(\sup_{\mathcal{K}} |\mathrm{LO}_{\lambda,\eta} - \mathrm{OO}_{\lambda,\eta}| \leq t) &\geq \mathbb{P}(F_4^c(t_4) \cap G^c(\epsilon)) \\
&\geq 1 - \mathbb{P}(F_4(t_4)) - \mathbb{P}(G(\epsilon)) \\
&\geq 1 - \frac{C_H C_v}{\tilde{\epsilon}^2 t_4^2 n} - 2ne^{-p/2} - \frac{2c_1 + c_2}{n^4} - \frac{2C_H^2(c_1+1)}{\epsilon^4 n^3} - \frac{32C_H(c_1+1)C_2^2}{C_1 \log(n) n^2} \\
&\geq 1 - 2ne^{-p/2} - \frac{2c_1 + c_2}{n^4} - \frac{32C_H(c_1+1)C_2^2}{C_1 \log(n) n^2} \\
&\quad - \frac{16C_H C_v(3C_3C_5\mathrm{PolyLog}(n) + 2C_6)^2 + 2C_H^2(c_1+1)8^4(3C_3C_5\mathrm{PolyLog}(n) + 2C_6)^4}{t^4 n} \\
&:= 1 - 2ne^{-p/2} - \frac{2c_1 + c_2}{n^2} - \frac{C_7 \mathrm{PolyLog}(n)}{t^4 n} - \frac{C_8}{n^2 \log(n)}
\end{aligned}$$

where[¶],

- $C_7 = 16C_H C_v(3C_3C_5 + 2C_6)^2 + 2C_H^2(c_1+1)8^4(3C_3C_5 + 2C_6)^4$
- $C_8 = \frac{32C_H(c_1+1)C_2^2}{C_1}$.

□

[¶]Here for notational simplicity, we require large enough $n$ such that both of the PolyLog$(n)$ terms are greater than one