Accelerating Conformal Prediction via Approximate Leave-One-Out

Jiachen Cong[1] and Jingbo Liu[1]

[1] Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA

December 27, 2025

Abstract

While conformal prediction provides a general framework for uncertainty quantification in predictive inference, its application is often limited by computational cost. Recent methods, including Jackknife+ and Jackknife-minmax, achieve faster computation by trading a slight loss of efficiency relative to full conformal prediction, but still requires computing leave-one-out refits for all observations. In this paper, we further accelerate conformal prediction by incorporating approximate leave-one-out (ALO) estimators, and establish asymptotic coverage and efficiency. While our proof draws on methods developed for analyzing the consistency of ALO cross-validation risk estimators in high-dimensional statistics, it requires adaptations to handle conformal prediction, where leave-i-out residuals are needed for predictions at x_{n+1} rather than just at the training covariate x_i. Simulation results validate our theoretical findings, showing that the ALO-based methods achieve coverage and efficiency comparable to the exact methods, while significantly reducing the runtime.

1 Introduction

Quantifying uncertainty is a central problem in statistics. In regression settings, this entails constructing prediction intervals for the new response y_{n+1} given a feature vector $\boldsymbol{x}_{n+1}$ and i.i.d. training data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. Among various approaches, Full Conformal Prediction (FCP) [Vovk et al., 2005] requires substantial computation cost to construct prediction intervals, which limits its practical use. In contrast, Jackknife+ and Jackknife-minmax [Barber et al., 2021] are more commonly adopted and relatively more efficient. However, their computational speed still remains slow. In fact, the complexities of Jackknife+ and Jackknife-minmax are on the order of n times that of the original regression problem, underscoring the need for more scalable alternatives.

Recently, [Clarté and Zdeborová, 2024] proposed to use approximate message passing (AMP) to speed up full conformal prediction, leveraging an ALO formula suggested by a heuristic derivation of AMP. The ALO formula in [Clarté and Zdeborová, 2024] is asymptotically equivalent to the one in [Rad and Maleki, 2020] with appropriate denoiser for the AMP, since AMP also asymptotically solves an optimization problem. However, a rigorous bound on ALO error for AMP, in the application to full conformal prediction, does not immediately follow from existing proofs based on state-evolution, since the proposed y in (2.2) does not follow the data generating distribution. While [Clarté and Zdeborová, 2024] did not prove an ALO error bound, coverage probability is guaranteed since only symmetry is required. However, Figure 2 and Table 4 in [Clarté and Zdeborová, 2024] indeed show that the efficiency (interval length) can be poor, likely stemming from undercontrolled ALO approximation error. Also, in [Clarté and Zdeborová, 2024], they assumed that the covariate are i.i.d. generated and the covariance matrix of each $\boldsymbol{x}_i$ is a diagonal matrix. By comparison, in our paper, we proposed a more general assumption that the covariate vector follows a normal distribution $\mathcal{N}(0, \Sigma)$, where Σ is not necessarily a diagonal matrix. This is a more realistic assumption that is easier to satisfy when analyzing real-world data.

In this paper, to address this challenge in the context of generalized linear regression, we incorporate ALO estimators, derived via the *Newton step* and *Woodbury identity*, into Jackknife+ and Jackknife-minmax, yielding

their accelerated versions that produce prediction intervals that align with the original methods and remain the coverage while substantially reducing computation.

Contribution.

- We establish uniform error bounds for ALO-based predictions with a new covariate x_{n+1}, showing that the approximation error

$$\left| \boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i} \right|$$

  remains uniformly controlled, whereas prior work [Wang et al., 2018, Rad and Maleki, 2020, Auddy et al., 2023] focused only on approximation for $\boldsymbol{x}_i, i \in \{1, 2, \dots, n\}$. Our proof leverages some decomposition techniques to transform (4.29) into (4.33), thereby reducing the new-covariate case to the setting studied in [Rad and Maleki, 2020]. Compared with [Clarté and Zdeborová, 2024], we weaken the assumption on the covariance structure of the distribution of $\boldsymbol{x}_i$; in particular, we no longer require the covariance matrix to be diagonal.

- We propose the accelerated Jackknife+ and Jackknife-minmax methods for constructing prediction intervals based on ALO estimator $\tilde{\theta}_{/i}$, and prove they retain asymptotically the same coverage probability and efficiency (interval length) as their original counterparts.

- In a special high-dimensional linear model, we establish that *Full Conformal, Split Conformal, Jackknife+*, and *Jackknife-minmax* prediction intervals are asymptotically equivalent.

Related work. A substantial body of work has examined the construction and error bounds of ALO estimators. [Beirami et al., 2017, Wang et al., 2018, Rad et al., 2020, Rad and Maleki, 2020, Xu et al., 2021, Auddy et al., 2023, Zou et al., 2024] develop explicit forms of ALO estimators using the Woodbury identity and analyze their theoretical and empirical performance in estimating out-of-sample risk. In particular, [Rad and Maleki, 2020, Wang et al., 2018, Zou et al., 2024] investigate ALO estimators under smoothed regularizers, covering ℓ_1 penalties, non-smooth but piecewise twice-differentiable regularizers, and more general non-smooth convex Lipschitz regularizers. Another line of work [Karoui et al., 2013, Karoui, 2018] prove ALO error bounds and use them to show central limit theorems of the estimator, but the latter requires stronger assumptions.

Notations. We first review the notations that will be used in the rest of the paper. Let $\boldsymbol{x}_i^\top \in \mathbb{R}^{1 \times p}$ stand for the i-th row of $X := (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$; and we define $\mathbf{y} := (y_1, y_2, ..., y_n)^\top$. $\mathbf{y}_{/i} \in \mathbb{R}^{(n-1) \times 1}$ and $X_{/i} \in \mathbb{R}^{(n-1) \times p}$ stand for $\mathbf{y}$ and X, excluding the ith entry y_i and the ith row $x_i^\top$, respectively. For brevity, we denote $\ell_i(\theta) := \ell(y_i, \boldsymbol{x}_i^\top \theta)$. The following definitions are also used:

$$\dot{\ell}_i(\theta) := \left. \frac{\partial \ell(y_i, z)}{\partial z} \right|_{z = \boldsymbol{x}_i^\top \theta}, \quad \ddot{\ell}_i(\theta) := \left. \frac{\partial^2 \ell(y_i, z)}{\partial z^2} \right|_{z = \boldsymbol{x}_i^\top \theta}.$$

Similarly, we define $\ddot{r}(\theta) = \nabla_\theta^2 r(\theta)$. Polynomials of $\log(n)$ are denoted by $\text{polylog}(n)$. In addition, we define $\hat{q}_{n,\alpha}^+\{d_i\}$ as the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value of $d_1, \dots, d_n$, and $\hat{q}_{n,\alpha}^-\{d_i\}$ as the $\lfloor \alpha(n+1) \rfloor$-th smallest value of $d_1, \dots, d_n$. The notation $\text{polylog}(n)$ denotes polynomial of $\log n$ with a finite degree. Let $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ stand for the largest and smallest singular values of $\mathbf{A}$, respectively. We use $o(1)$ to denote a term that vanishes as $n \to \infty$, i.e., $o(1) \to 0$ when $n \to \infty$.

2 Preliminaries

We define $\hat{y} = \hat{\mu}(\boldsymbol{x}) = f(\boldsymbol{x}^\top \hat{\theta})$ as the model trained on the full data, with $f : \mathbb{R} \to \mathbb{R}$. For each i, $\hat{\mu}_{/i}(\boldsymbol{x}) = f(\boldsymbol{x}^\top \hat{\theta}_{/i})$ and $\tilde{\mu}_{/i}(\boldsymbol{x}) = f(\boldsymbol{x}^\top \tilde{\theta}_{/i})$, where $\hat{\theta}_{/i}$ and $\tilde{\theta}_{/i}$ are defined in (2.12) and (2.13) separately.

2.1 Conformal prediction

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n+1}$ be a set of covariate–ground truth pairs where $\boldsymbol{x}_{n+1}$ denotes a test value for which we would like to output a response. Then, full conformal prediction (FCP) outputs a prediction set $\hat{C}(\boldsymbol{x}_{n+1})$ such that

$$\mathbb{P}\{y_{n+1} \in \hat{C}(\boldsymbol{x}_{n+1})\} \geq 1 - \alpha, \tag{2.1}$$

for any user-specified value $\alpha \in (0, 1)$. In the standard expression of FCP,

$$\hat{C}(\boldsymbol{x}_{n+1}) = \left\{ y : \sigma_{n+1}(y) \leq \hat{q}_{\lceil(1-\kappa)(n+1)\rceil/n}(\sigma(y)) \right\}, \tag{2.2}$$

where $\sigma_i(y) := \left| \hat{\theta}(y)^\top \boldsymbol{x}_i - y_i \right|$ and $\hat{\theta}(y)$ denotes the estimator for the augmented dataset $\mathcal{D} \cup \{(\boldsymbol{x}_{n+1}, y)\} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{n+1}, y)\}$; see [Vovk et al., 2005, Angelopoulos and Bates, 2022].

Specifically, in our paper, we use two variants of FCP, called Jackknife+ and Jackknife-minmax. The construction of prediction intervals is introduced in (2.3) and (2.4), separately. Let $R_i^{LOO} := |y_i - f(\boldsymbol{x}_i^\top \hat{\theta}_{/i})|$, and we define

$$\hat{C}_{n,\alpha}^{jackknife+} := [L_{jk+}, U_{jk+}], \tag{2.3}$$

$$\hat{C}_{n,\alpha}^{jack-mm} := [L_{jkm}, U_{jkm}], \tag{2.4}$$

where the corresponding endpoints are defined as

$$U_{jk+} := \hat{q}_{n,\alpha}^+ \{\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) + R_i^{LOO}\}, \tag{2.5}$$

$$L_{jk+} := \hat{q}_{n,\alpha}^- \{\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - R_i^{LOO}\}, \tag{2.6}$$

$$U_{jkm} := \max_{i=1,\ldots,n} \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) + \hat{q}_{n,\alpha}^+ \{R_i^{LOO}\}, \tag{2.7}$$

$$L_{jkm} := \min_{i=1,\ldots,n} \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \hat{q}_{n,\alpha}^+ \{R_i^{LOO}\}. \tag{2.8}$$

In [Barber et al., 2021], the authors established following results:

$$\mathbb{P}\{y_{n+1} \in \hat{C}_{n,\alpha}^{jackknife+}\} \geq 1 - 2\alpha \tag{2.9}$$

$$\mathbb{P}\{y_{n+1} \in \hat{C}_{n,\alpha}^{jack-mm}\} \geq 1 - \alpha. \tag{2.10}$$

2.2 Leave-one-out approximation

Let $\ell(y, \boldsymbol{x}^\top \theta)$ and $r(\theta)$ be the loss function and regularizer, and λ be a regularization parameter. We define $r(\theta) = (1 - \eta)r_0(\theta) + \eta\theta^\top\theta$ where $\eta \in (0, 1)$ and r_0 is a smooth non-negative convex regularizer. Then, $\hat{\theta} \in \mathbb{R}^{p \times 1}$ can be expressed as

$$\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{j=1}^n \ell(y_j, \boldsymbol{x}_j^\top \theta) + \lambda \cdot r(\theta) \right\}. \tag{2.11}$$

We define the leave-i-out coefficient estimator $\hat{\theta}_{/i} \in \mathbb{R}^{p \times 1}$ as follows

$$\hat{\theta}_{/i} := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{j \neq i}^n \ell(y_j, \boldsymbol{x}_j^\top \theta) + \lambda \cdot r(\theta) \right\}. \tag{2.12}$$

According to [Rad and Maleki, 2020], based on the Newton step and the Woodbury identity, they define the approximate leave-i-out estimator $\tilde{\theta}_{/i} \in \mathbb{R}^{p \times 1}$ as follows

$$\tilde{\theta}_{/i} := \hat{\theta} + \frac{\mathbf{J}^{-1}\boldsymbol{x}_i \, \dot{\ell}_i(\hat{\theta})}{1 - \boldsymbol{x}_i^\top \mathbf{J}^{-1}\boldsymbol{x}_i \, \ddot{\ell}_i(\hat{\theta})}, \tag{2.13}$$

where $\mathbf{J} := \left(\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j^\top \ddot{\ell}_j(\hat{\theta}) + \lambda \operatorname{diag}\left[\ddot{r}(\hat{\theta}) \right] \right)$.

2.3 Assumptions in [Rad and Maleki, 2020] and [Zou et al., 2024]

Assumption 1. *(a) r_0 is a non-negative, convex and twice differentiable function.*
(b) Loss function $\ell(y, z)$ is non-negative, convex and continuously differentiable with respect to z.

Remark 1. *In our main analysis, we assume that r_0 is twice differentiable so that the Hessian is well-defined. For certain non-smooth regularizers, such as the ℓ_1 penalty, smoothing approximations can be employed as in [Zou et al., 2024, Rad and Maleki, 2020] to construct ALO estimators with smoothed regularizers. While extending our results to this setting is possible, it lies beyond the main scope of the present paper.*

Assumption 2. *(a) $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ are i.i.d. $N(0, \Sigma)$ samples. Let ρ_{max} denote the largest eigenvalue of Σ. Moreover, there exist constants $0 < c_X \le C_X$ such that*

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

(b) We assume observations are independent and identically distributed draws from some unknown joint distribution $p(y_i | \boldsymbol{x}_i^\top \theta^*) q(\boldsymbol{x}_i)$, where $\theta^* \in \mathbb{R}^p$ represents the true parameter.

Remark 2. *We consider the regime where $n, p \to \infty$, $n/p = \delta_0 \in (0, \infty)$, and elements of θ^* are $O(1)$; we have $\|\theta^*\| = O(\sqrt{p})$, and hence $\mathbb{E}[(\boldsymbol{x}^\top \theta^*)^2] = O(1)$ [Zou et al., 2024].*

Assumption 3. *(a) We define*

$$\dot{\boldsymbol{\ell}}(\cdot) := (\dot{\ell}_1(\cdot), \ldots, \dot{\ell}_n(\cdot))^\top;$$
$$\dot{\boldsymbol{\ell}}_{/i}(\cdot) := (\dot{\ell}_1(\cdot), \ldots, \dot{\ell}_{i-1}(\cdot), \dot{\ell}_{i+1}(\cdot), \ldots, \dot{\ell}_n(\cdot))^\top;$$
$$\ddot{\boldsymbol{\ell}}_{/i}(\cdot) := (\ddot{\ell}_1(\cdot), \ldots, \ddot{\ell}_{i-1}(\cdot), \ddot{\ell}_{i+1}(\cdot), \ldots, \ddot{\ell}_n(\cdot))^\top.$$

We assume that $c_1(n) = O(polylog(n))$ and $c_2(n) = O(polylog(n))$, and $q_n \to 0$ are all functions of n, such that with probability at least $1 - q_n$ for all $i = 1, \ldots, n$

$$c_1(n) > \|\dot{\boldsymbol{\ell}}(\hat{\theta})\|_\infty;$$
$$c_2(n) > \sup_{t \in [0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}\{(1-t)\hat{\theta}_{/i} + t\hat{\theta}\} - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta})\|_2}{\|\hat{\theta}_{/i} - \hat{\theta}\|_2};$$
$$c_2(n) > \sup_{t \in [0,1]} \frac{\|\ddot{r}\{(1-t)\hat{\theta}_{/i} + t\hat{\theta}\} - \ddot{r}(\hat{\theta})\|_2}{\|\hat{\theta}_{/i} - \hat{\theta}\|_2}.$$

(b) There exists a constant $\nu > 0$ and a sequence $\tilde{q}_n \to 0$ such that for all $i = 1, ..., n$

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

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

Remark 3. *For part (a) of the assumption, [Rad and Maleki, 2020, Section 4] provided its justification for the ridge, elastic net, logistic regression, robust regression, and Poisson regression by choosing $c_1(n)$ and $c_2(n)$ to be polynomials of n. For part (b) of the assumption, since we chose $r(\theta) = (1 - \eta)r_0(\theta) + \eta\theta^\top\theta$, we have $\nu = 2\lambda\eta$.*

2.4 Approximation error bound in [Rad and Maleki, 2020]

Under Assumption 1 to Assumption 3, [Rad and Maleki, 2020] proved that

$$\max_{1 \le i \le n} \left| x_i^\top \hat{\theta}_{/i} - x_i^\top \hat{\theta} - \left(\frac{\dot{\ell}_i(\hat{\theta})}{\ddot{\ell}_i(\hat{\theta})} \right) \left(\frac{H_{ii}}{1 - H_{ii}} \right) \right| \le \frac{C_0}{\sqrt{p}}$$

holds with probability at least $1 - 4ne^{-p} - \frac{8n}{p^3} - \frac{8n}{(n-1)^3} - q_n - \tilde{q}_n$, where C_0 is defined as same as (3.25). This result however is not sufficient for our application to conformal prediction, and we will adapt the proof technique to bound the approximation error of $x_{n+1}^\top \hat{\theta}_{/i}$.

2.5 Lemma 25 in [Rad and Maleki, 2020]

Lemma 1 (Lemma 25 in [Rad and Maleki, 2020]). *Assume that $X^\top(D + \Gamma)X$ and $X^\top DX$ are positive definite, and define:*

$$\Gamma := \mathrm{diag}(\gamma), \quad \bar{\omega}_{\max} := \sigma_{\max}(XX^\top), \quad \nu_{\min} := \sigma_{\min}(X^\top(D + \Gamma)X), \quad A := X^\top DX.$$

Then,

$$\left| z^\top \left(X^\top(D + \Gamma)X\right)^{-1} z - z^\top\left(X^\top DX\right)^{-1} z \right| \leq \left(\|\gamma\|_2 + \frac{\bar{\omega}_{\max}}{\nu_{\min}} \|\gamma\|_4^2 \right) \left\| XA^{-1} z\right\|_4^2. \tag{2.14}$$

3 Main results

Assumption 4. *Under assumptions 1 to 3, there exists a large enough constant $L < \infty$ such that $\mathbb{P}\{y \in [a, b] | \boldsymbol{x}^\top \theta\} \leq L \cdot |b - a|$ and $|f(b) - f(a)| \leq L \cdot |b - a|$ for $\forall a, b \in \mathbb{R}$.*

Compared with the Assumptions 1-3, the newly added Assumption 4 is used to control the effect of the deviation of approximate intervals from the exact ones on the resulting coverage guarantee. For simplicity, under Assumption 3, we define

$$p_1(n) := \left(\frac{16n}{(n-1)^3} + \frac{16n}{p^3} + 8ne^{-p} \right) + q_n + \tilde{q}_n, \tag{3.15}$$

$$p_2(n) := \left(\frac{8n}{(n-1)^3} + \frac{8n}{p^3} + 4ne^{-p} \right) + q_n + \tilde{q}_n, \tag{3.16}$$

and when $n/p = \delta_0 \in (0, \infty)$ it is clear that

$$p_1(n) \xrightarrow{n \to \infty} 0, \quad p_2(n) \xrightarrow{n \to \infty} 0. \tag{3.17}$$

3.1 Algorithm

Algorithm 1 Accelerating Jackknife+ with ALO Estimators

Input: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n), \boldsymbol{x}_{n+1}, \alpha$
Output: Prediction interval for y_{n+1}
1. Calculate the full-sample estimator $\hat{\theta}$ (see Eq. (2.11)).
2. Calculate the approximate leave-i-out estimator $\tilde{\theta}_{/i}$ (see Eq. (2.13)) for $i \in \{1, 2, ..., n\}$.
3. Obtain the approximate i-th LO residual $\tilde{R}_i^{LOO} := |y_i - f(\boldsymbol{x}_i^\top \tilde{\theta}_{/i})|$ and the approximate leave-i-out estimation $\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) = f(\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i})$ with $\boldsymbol{x}_{n+1}$ for $i \in \{1, 2, ..., n\}$.
4. Obtain the prediction interval for y_{n+1}, defined as $[\tilde{L}_{jk+}, \tilde{U}_{jk+}]$, where $\tilde{U}_{jk+}$ is defined in Eq. (3.18) and $\tilde{L}_{jk+}$ is defined in Eq. (3.19).

Algorithm 2 Accelerating Jackknife-minmax with ALO Estimators

---

Input: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n), \boldsymbol{x}_{n+1}, \alpha$
Output: Prediction interval for y_{n+1}
1. Calculate the full-sample estimator $\hat{\theta}$ (see Eq. (2.11)).
2. Calculate the approximate leave-i-out estimator $\tilde{\theta}_{/i}$ (see Eq. (2.13)) for $i \in \{1, 2, ..., n\}$.
3. Obtain the approximate i-th LO residual $\tilde{R}_i^{LOO} := |y_i - f(\boldsymbol{x}_i^\top \tilde{\theta}_{/i})|$ and the approximate leave-i-out estimation $\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) = f(\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i})$ with $\boldsymbol{x}_{n+1}$ for $i \in \{1, 2, ..., n\}$.
4. Obtain the prediction interval for y_{n+1}, defined as $[\tilde{L}_{jkm}, \tilde{U}_{jkm}]$, where $\tilde{U}_{jkm}$ is defined in Eq. (3.20) and $\tilde{L}_{jkm}$ is defined in Eq. (3.21).

With ALO estimator introduced in 2.13, let $\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) = f(\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i})$ and $\tilde{R}_i^{LOO} := |y_i - f(\boldsymbol{x}_i^\top \tilde{\theta}_{/i})|$. For simplicity, we define

$$\tilde{U}_{jk+} := \hat{q}_{n,\alpha}^+\{\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) + \tilde{R}_i^{LOO}\}, \tag{3.18}$$

$$\tilde{L}_{jk+} := \hat{q}_{n,\alpha}^-\{\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{R}_i^{LOO}\}, \tag{3.19}$$

$$\tilde{U}_{jkm} := \max_{i=1,\ldots,n} \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) + \hat{q}_{n,\alpha}^+\{\tilde{R}_i^{LOO}\}, \tag{3.20}$$

$$\tilde{L}_{jkm} := \min_{i=1,\ldots,n} \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \hat{q}_{n,\alpha}^+\{\tilde{R}_i^{LOO}\}. \tag{3.21}$$

Then the jackknife+ prediction interval based on estimators in (2.13) is constructed as in (3.22),

$$\tilde{C}_{n,\alpha}^{jackknife+} := [\tilde{L}_{jk+}, \tilde{U}_{jk+}], \tag{3.22}$$

and the jackknife-minmax prediction interval based on estimators in (2.13) is constructed as in (3.23)

$$\tilde{C}_{n,\alpha}^{jack-mm} := [\tilde{L}_{jkm}, \tilde{U}_{jkm}]. \tag{3.23}$$

3.2 Approximation guarantees

Theorem 3.1. *Let Assumption 1 to Assumption 3 hold with $\rho_{max} = c/p$ and $n/p = \delta_0 \in (0, \infty)$. Moreover, suppose that n is large enough such that $q_n + \tilde{q}_n < 0.5$. Then with probability at least $1 - p_1(n)$, where $p_1(n)$ is defined in (3.15), the following bound is valid:*

$$\max_{1 \le i \le n} |\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i}| \le \frac{C_0}{\sqrt{p}}, \tag{3.24}$$

where

$$C_0 := \left(\frac{216 c^{3/2}}{\nu^3}\right) \left(1 + \sqrt{\delta_0}(\sqrt{\delta_0} + 3)^2 \frac{c \log n}{\log p}\right) \left(c_1^2(n)c_2(n) + c_1^3(n)c_2^2(n) \frac{5\big(c^{1/2} + c^{3/2}(\sqrt{\delta_0} + 3)^2\big)}{\nu^2}\right). \tag{3.25}$$

Theorem 3.2. *Under Assumption 1 to Assumption 4, when $n, p \to \infty$, $n/p = \delta_0 \in (0, \infty)$, with probability at least $(1 - p_1(n))(1 - p_2(n))$,*

$$\tilde{U}_{jk+} = U_{jk+} + o(1), \ \tilde{L}_{jk+} = L_{jk+} + o(1);$$

$$\tilde{U}_{jkm} = U_{jkm} + o(1), \ \tilde{L}_{jkm} = L_{jkm} + o(1).$$

Theorem 3.3. *Under Assumption 1 to Assumption 4, when $n, p \to \infty$, $n/p = \delta_0 \in (0, \infty)$, we have*

$$\mathbb{P}\{y_{n+1} \in \tilde{C}_{n,\alpha}^{jackknife+}\} \ge 1 - 2\alpha - o(1). \tag{3.26}$$

Theorem 3.4. *Under Assumption 1 to Assumption 4, when $n, p \to \infty$, $n/p = \delta_0 \in (0, \infty)$, we have*

$$\mathbb{P}\{y_{n+1} \in \tilde{C}_{n,\alpha}^{jack-mm}\} \ge 1 - \alpha - o(1). \tag{3.27}$$

4 Proof sketch

4.1 Sketch of proof for Theorem 3.1

Let $\Delta_{/i}^* := \hat{\theta}_{/i} - \hat{\theta}$, $\bar{X}_{/i} := \begin{bmatrix} X_{/i} \\ I \end{bmatrix}$, $\boldsymbol{J}_{/i}(\theta) := \bar{X}_{/i}^\top \mathbf{D}_{/i}(\theta) \bar{X}_{/i}$ with $\mathbf{D}_{/i}(\theta) := \text{diag}\big[\ddot{\boldsymbol{\ell}}_{/i}(\theta), \lambda \ddot{r}(\theta)\big]$. In addition, for simplicity, we define

$$\gamma_{\xi/i}(\theta) := \begin{bmatrix} \ddot{\boldsymbol{\ell}}_{/i}(\theta + \xi) - \ddot{\boldsymbol{\ell}}_{/i}(\theta) \\ \lambda\big(\ddot{r}(\theta + \xi) - \ddot{r}(\theta)\big) \end{bmatrix},$$

6

$$\Phi_1 := \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\Big[\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\, dt\Big] \bar{X}_{/i}\right)^{-1},$$

$$\Phi_2 := \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\Big[\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\Big] \bar{X}_{/i}\right)^{-1}.$$

We want to show that

$$|\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i}| \le \frac{C_0}{\sqrt{p}} \tag{4.28}$$

with probability $1 - p_1(n)$ for $i = 1, 2, ..., n$.

Following the derivations in Eqs. (58) and (59) of [Rad and Maleki, 2020], we obtain

$$\Delta_{/i}^* = \dot{\ell}_i(\hat{\theta}) \cdot \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta_{/i}^*)dt\right)^{-1} \boldsymbol{x}_i,$$

$$\frac{\boldsymbol{J}^{-1}\boldsymbol{x}_i\,\dot{\ell}_i(\hat{\theta})}{1 - \boldsymbol{x}_i^\top \boldsymbol{J}^{-1}\boldsymbol{x}_i\,\ddot{\ell}_i(\hat{\theta})} = \dot{\ell}_i(\hat{\theta}) \cdot \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i} - \Delta_{/i}^*)\boldsymbol{x}_i,$$

then we obtain the following decomposition

$$\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i} \tag{4.29}$$
$$= \dot{\ell}_i(\hat{\theta})\boldsymbol{x}_{n+1}^\top M \boldsymbol{x}_i \tag{4.30}$$
$$= \frac{1}{2}\dot{\ell}_i(\hat{\theta})(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top M(\boldsymbol{x}_i + \boldsymbol{x}_{n+1}) - \frac{1}{2}\dot{\ell}_i(\hat{\theta})\boldsymbol{x}_{n+1}^\top M \boldsymbol{x}_{n+1} - \frac{1}{2}\dot{\ell}_i(\hat{\theta})\boldsymbol{x}_i^\top M \boldsymbol{x}_i, \tag{4.31}$$

where we define

$$M := \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta_{/i}^*)dt\right)^{-1} - \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i} - \Delta_{/i}^*). \tag{4.32}$$

Let A_1, A_2, A_3 denote the three terms in (4.31). Then clearly

$$|\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i}| \le |A_1| + |A_2| + |A_3|. \tag{4.33}$$

Terms A_1, A_2 and A_3 can be controlled in probability by using similar approach in [Rad and Maleki, 2020]: For A_1, we have For the new A_1 term, we have

$$\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top M(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$
$$\overset{1}{\le} \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\left[\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})dt\right] \bar{X}_{/i}\right)^{-1}\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$
$$+ \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\left[\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right] \bar{X}_{/i}\right)^{-1}\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$
$$= \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right| + \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|. \tag{4.34}$$

Following conclusions can be derived when Assumption 1-3 hold:

$$\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 \le 2c_2(n)\|\Delta_{/i}^*\|_2,$$

$$\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\, dt\right\|_2 \le 2c_2(n)\|\Delta_{/i}^*\|_2,$$

$$\|\Delta_{/i}^*\|_2 \le \left(\frac{|\dot{\ell}_i(\hat{\theta})|}{\nu}\right)\|\boldsymbol{x}_i\|_2 \le \frac{c_1(n)}{\nu}\|\boldsymbol{x}_i\|_2.$$

Now we focus on the $\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2 (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$ part in A_1. Using result in Section 2.5 and the Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
&\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2 (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right| \\
&\leq 2\left(\left\|\gamma_{-\Delta_{/i}^* /i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^* /i}(\hat{\theta}_{/i})\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \\
&\quad + 2\left(\left\|\gamma_{-\Delta_{/i}^* /i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^* /i}(\hat{\theta}_{/i})\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2, \qquad (4.35)
\end{aligned}
$$

where $\bar{\omega}_{\max,i} = \sigma_{\max}(\bar{X}_{/i}\bar{X}_{/i}^\top)$, ν has been defined in Assumption 3.

Similarly, for $\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1 (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$, we obtain

$$
\begin{aligned}
&\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1 (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right| \\
&\leq 2\left(\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^* /i}(\hat{\theta}_{/i})\,dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^* /i}(\hat{\theta}_{/i})\,dt\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \\
&\quad + 2\left(\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^* /i}(\hat{\theta}_{/i})\,dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^* /i}(\hat{\theta}_{/i})\,dt\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2.
\end{aligned}
$$
$$(4.36)$$

For $|A_2|$ and $|A_3|$, we can use a similar method to show that

$$
|A_2| \leq \frac{1}{2}\bar{C}_i\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2\right] \qquad (4.37)
$$

and

$$
|A_3| \leq \frac{1}{2}\bar{C}_i\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2\right]. \qquad (4.38)
$$

Combining results in 4.35, 4.36, 4.37 and 4.38, we obtain

$$
\begin{aligned}
|\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i}| &\leq \frac{3}{2}\bar{C}_i\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2\right] \\
&\quad + \frac{3}{2}\bar{C}_i\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2\right] \\
&\leq 3\bar{C}_i \cdot \max_{j\in\{i,n+1\}}\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2\right], \qquad (4.39)
\end{aligned}
$$

where $\bar{C}_i := 4\|\boldsymbol{x}_i\|_2\left(\frac{c_1^2(n)c_2(n)}{\nu}\right)\left(1 + \frac{2c_1(n)c_2(n)}{\nu^2}(1+\omega_{\max})\|\boldsymbol{x}_i\|_2\right)$. For simplicity, we define $\Xi_j := \left(\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2\right)$ and event G as follows

$$
G := \left\{\max_{1\leq i\leq n}|\boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i}| > C\frac{\log(p)}{\sqrt{p}}\right\}, \qquad (4.40)
$$

where
$$C := 96\sqrt{5}\left(\frac{c_1^2(n)c_2(n)\,(p\rho_{\max})^{3/2}}{\nu^3}\right)$$
$$\times \left(1 + \left(\sqrt{\frac{n}{p}} + 3\right)^2 p\rho_{\max}\sqrt{\frac{n-1}{p}\frac{\log(n-1)}{\log p}}\right)\left(1 + \frac{2c_1(n)c_2(n)\sqrt{5}\left(1 + \left(\sqrt{\frac{n}{p}}+3\right)^2 p\rho_{\max}\right)\sqrt{p\rho_{\max}}}{\nu^2}\right).$$

Let S denote the event that Assumption 3 holds, we can trivially obtain that $\mathbb{P}[S^c] \leq q_n + \tilde{q}_n$. By Eq. (122) to (124) in [Rad and Maleki, 2020], we have

$$
\begin{aligned}
\mathbb{P}[G] &\leq \mathbb{P}[G|S] + \mathbb{P}[S^c] \\
&\leq \frac{\mathbb{P}\left[\max_{1\leq i\leq n} 3\bar{C}_i\Xi_i \geq C\frac{\log(p)}{\sqrt{p}}\right]}{1 - q_n - \tilde{q}_n} + \frac{\mathbb{P}\left[\max_{1\leq i\leq n} 3\bar{C}_i\Xi_{n+1} \geq C\frac{\log(p)}{\sqrt{p}}\right]}{1 - q_n - \tilde{q}_n} + q_n + \tilde{q}_n. \qquad (4.41)
\end{aligned}
$$

First, we obtain

$$\mathbb{P}\left[\max_{1\leq i\leq n} 3\bar{C}_i \Xi_i \geq C\frac{\log(p)}{\sqrt{p}}\right] \leq \sum_{i=1}^{n} \mathbb{P}[\tilde{E}_i] \tag{4.42}$$

and

$$\mathbb{P}\left[\max_{1\leq i\leq n} 3\bar{C}_i \Xi_{n+1} \geq C\frac{\log(p)}{\sqrt{p}}\right] \leq \sum_{i=1}^{n} \mathbb{P}[\tilde{E}_{n+1,i}], \tag{4.43}$$

where events $\tilde{E}_i$ and $\tilde{E}_{n+1,i}$ are defined as

$$\tilde{E}_i := \left\{ 3\bar{C}_i \left(\left\| X_{/i} \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) \boldsymbol{x}_i \right\|_4^2 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) \boldsymbol{x}_i \right\|_4^2 \right) > C\frac{\log p}{\sqrt{p}} \right\}$$

and

$$\tilde{E}_{n+1,i} := \left\{ 3\bar{C}_i \left(\left\| X_{/i} \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) \boldsymbol{x}_{n+1} \right\|_4^2 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) \boldsymbol{x}_{n+1} \right\|_4^2 \right) > C\frac{\log p}{\sqrt{p}} \right\}.$$

For each $\mathbb{P}[\tilde{E}_i]$ and $\mathbb{P}[\tilde{E}_{n+1,i}]$, we can obtain the following results after some derivations

$$\mathbb{P}[\tilde{E}_i] \leq \frac{4}{(n-1)^3} + \frac{4}{p^3} + 2e^{-p} \text{ and } \mathbb{P}[\tilde{E}_{n+1,i}] \leq \frac{4}{(n-1)^3} + \frac{4}{p^3} + 2e^{-p}. \tag{4.44}$$

Here we used the techniques in [Rad and Maleki, 2020], but with a key modification. Specifically, [Rad and Maleki, 2020] upper bounded $\mathbb{P}[\tilde{E}_i]$, and we showed that the same argument can be used to control $\mathbb{P}[\tilde{E}_{n+1,i}]$, because $\boldsymbol{x}_i$ and $\boldsymbol{x}_{n+1}$ are i.i.d.

Then taking the results in (4.44) into (4.41), replacing $96\sqrt{5}$, $2\sqrt{3}$ and $\sqrt{\frac{n-1}{p}\frac{\log(n-1)}{\log p}}$ with the upper-bounds 216, 5 and $\sqrt{\frac{n}{p}\frac{\log n}{\log p}}$ respectively, we obtain the result described in Theorem 3.1.

4.2 Sketch of proof for Theorem 3.2

Firstly, we introduce the following lemma.

Lemma 2. *Let $\{a_i\}_{i=1}^n$ be a sorted sequence such that $a_1 \leq a_2 \leq \cdots \leq a_n$, and suppose there exist sequences $\{b_i\}_{i=1}^n$ satisfying $|b_i - a_i| \leq \varepsilon$, for some $\varepsilon > 0$ and all $i = 1, \ldots, n$. Then for any $q \in \{1, 2, ..., n\}$ we have*

$$|b_{(q)} - a_q| \leq \epsilon,$$

where $b_{(q)}$ is the q-th smallest value of $\{b_i\}_{i=1}^n$.

Proof of the lemma 2 is deferred to Section S.5 in the supplementary materials. Based on Lemma 2, we can take the part $\tilde{U}_{jk+} = U_{jk+} + o(1)$ in Theorem 3.2 as an example, and the proofs of the other three follow analogously.

For all $i \in \{1, 2, ..., n\}$, using Theorem 3.1 and Assumption 4, we can show that

$$\begin{aligned} &\left| \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) \right| \\ \leq &L \cdot \left| \boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i} \right| = o(1) \end{aligned} \tag{4.45}$$

with probability at least $1 - p_1(n)$.

Using the triangle inequality, Assumption 4, and existing results introduced in Section 2.4 we obtain that

$$\begin{aligned} &\left| R_i^{LOO} - \tilde{R}_i^{LOO} \right| \\ = &\left| (|y_i - \Phi(\boldsymbol{x}_i^\top \hat{\theta}_{/i})| - |y_i - \Phi(\boldsymbol{x}_i^\top \tilde{\theta}_{/i})|) \right| \\ \leq &L \cdot \max_{1\leq i\leq n} |\boldsymbol{x}_i^\top \hat{\theta}_{/i} - \boldsymbol{x}_i^\top \tilde{\theta}_{/i}| = o(1) \end{aligned} \tag{4.46}$$

with probability at least $1 - p_2(n)$.

As a result, with probability at least $(1 - p_1(n))(1 - p_2(n))$, the following bounds hold simultaneously for all $i = 1, 2, ..., n$

$$\left| (\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{R}_i^{LOO}) - (\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - R_i^{LOO}) \right| = o(1). \tag{4.47}$$

Using Lemma 2, we can show that

$$\tilde{U}_{jk+} = U_{jk+} + o(1) \tag{4.48}$$

holds with probability at least $(1 - p_1(n))(1 - p_2(n))$.

4.3   Sketch of proof for Theorem 3.3-3.4

Based on Theorem 3.1 and 3.2, the proofs of Theorem 3.3 and 3.4 are similar and straightforward. We outline the proof of Theorem 3.3 here, while the argument for Theorem 3.4 follows analogously.

In Theorem 3.2, using Assumption 4, we can show that

$$\tilde{U}_{jk+} = U_{jk+} + o(1), \ \tilde{L}_{jk+} = L_{jk+} + o(1)$$

and

$$\begin{aligned} &\mathbb{P}[y_{n+1} \in \tilde{C}_{n,\alpha}^{jackknife+}] \\ =&\mathbb{P}[y_{n+1} \in \hat{C}_{n,\alpha}^{jackknife+}] + L \cdot o(1) \end{aligned} \tag{4.49}$$

hold with probability at least $(1 - p_1(n))(1 - p_2(n))$.

Then we have

$$\begin{aligned} &\mathbb{P}[y_{n+1} \in \tilde{C}_{n,\alpha}^{Jackknife+}] \\ \geq&(1 - 2\alpha - o(1)) \cdot (1 - p_1(n))(1 - p_2(n)) \\ =&1 - 2\alpha - o(1). \end{aligned} \tag{4.50}$$

By similar method, we can show that

$$\begin{aligned} &\mathbb{P}[y_{n+1} \in \tilde{C}_{n,\alpha}^{jack-mm}] \\ \geq&(1 - \alpha - o(1)) \cdot (1 - p_1(n))(1 - p_2(n)) \\ =&1 - \alpha - o(1). \end{aligned} \tag{4.51}$$

5   Numerical experiments

5.1   Synthetic data

In this section, we conduct simulations to support the results in Theorem 3.3-3.4. We report the mean of coverage, mean of operation time, mean of interval length and mean of Jaccard index of original methods (labeled as "JK+" and "JK-minmax", where "JK" stands for "Jackknife") and accelerated methods (labeled as "Fast JK+" and "Fast JK-minmax"). Recall that the Jaccard index between two sets $\mathcal{S}_1, \mathcal{S}_2$ is defined as

$$\mathcal{J}(\mathcal{S}_1, \mathcal{S}_2) = \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \in [0, 1].$$

Values closer to 1 indicate more precise approximations. In Table 3 we report the Jaccard index between Jackknife+ and Fast Jackknife+, Jackknife-minmax and Fast Jackknife-minmax.

In our simulation, we set $\alpha = 0.1$, which means our target coverage level is $1 - \alpha = 0.9$ (in the jackknife+ case, as shown in [Barber et al., 2021], the coverage level is $1 - 2\alpha = 0.8$). We use training sample size

$n = 100$, test sample size $n_{test} = 100$, and repeat the experiment at each dimension $p = 50, 100, 200$, with i.i.d. data points $(\boldsymbol{x}_i, y_i)$ generated as $\boldsymbol{x}_i \sim \mathcal{N}(0, I_p/\sqrt{p})$ and $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \theta, 1)$. The true coefficient vector θ is randomly generated from a standard normal distribution . For the Ridge regression model, we define the loss function as $\ell(y, \boldsymbol{x}^\top \theta) = (y - \boldsymbol{x}^\top \theta)^2/2$ and the regularization term as $r(\theta) = \frac{1}{2}r_0(\beta) + \frac{1}{2}\theta^\top \theta$, where Pseudo-Huber regularizer (this setting allows our simulation results to be compared with those reported in [Clarté and Zdeborová, 2024]) $r_0(\beta) = \sum_{j=1}^{p} 4(\sqrt{1 + \frac{\beta_j^2}{4}} - 1)$, with the ridge parameter $\lambda = 0.1, 1$, separately. To obtain stable results, we repeat the procedures above for 50 iterations and report the averaged outcomes.

Table 1: Comparison of JK+ and Fast JK+.

Parameters	Model	Coverage	Time(s)	Length
(n=100, p=50, λ=1)	JK+	0.880	0.075	3.808
(n=100, p=100, λ=1)	JK+	0.892	0.109	4.189
(n=100, p=200, λ=1)	JK+	0.897	0.185	4.409
(n=100, p=50, λ=0.1)	JK+	0.872	0.091	4.135
(n=100, p=100, λ=0.1)	JK+	0.888	0.172	4.729
(n=100, p=200, λ=0.1)	JK+	0.889	0.354	4.662
(n=100, p=50, λ=1)	Fast JK+	0.880	0.004	3.809
(n=100, p=100, λ=1)	Fast JK+	0.893	0.008	4.190
(n=100, p=200, λ=1)	Fast JK+	0.897	0.043	4.410
(n=100, p=50, λ=0.1)	Fast JK+	0.872	0.005	4.135
(n=100, p=100, λ=0.1)	Fast JK+	0.888	0.009	4.734
(n=100, p=200, λ=0.1)	Fast JK+	0.890	0.044	4.672

Table 2: Comparison of JK-minmax and Fast JK-minmax.

Parameters	Model	Coverage	Time(s)	Length
(n=100, p=50, λ=1)	JK-minmax	0.911	0.086	4.187
(n=100, p=100, λ=1)	JK-minmax	0.920	0.103	4.549
(n=100, p=200, λ=1)	JK-minmax	0.918	0.189	4.718
(n=100, p=50, λ=0.1)	JK-minmax	0.926	0.094	4.925
(n=100, p=100, λ=0.1)	JK-minmax	0.942	0.164	5.574
(n=100, p=200, λ=0.1)	JK-minmax	0.937	0.388	5.387
(n=100, p=50, λ=1)	Fast JK-minmax	0.911	0.004	4.188
(n=100, p=100, λ=1)	Fast JK-minmax	0.920	0.008	4.550
(n=100, p=200, λ=1)	Fast JK-minmax	0.918	0.042	4.719
(n=100, p=50, λ=0.1)	Fast JK-minmax	0.926	0.004	4.925
(n=100, p=100, λ=0.1)	Fast JK-minmax	0.943	0.008	5.750
(n=100, p=200, λ=0.1)	Fast JK-minmax	0.937	0.047	5.400

Table 3: Prediction-interval Overlap (Jaccard Index) of Fast JK+ & JK+ and Fast JK-minmax & JK-minmax.

Parameters	Fast JK+ & JK+	Fast JK-minmax & JK-minmax
(n=100, p=50, λ=1)	0.9997	0.9997
(n=100, p=100, λ=1)	0.9997	0.9996
(n=100, p=200, λ=1)	0.9998	0.9997
(n=100, p=50, λ=0.1)	0.9996	0.9996
(n=100, p=100, λ=0.1)	0.9983	0.9982
(n=100, p=200, λ=0.1)	0.9978	0.9976

Table 1 and Table 2 present the mean of coverage, mean of operation time and mean of interval length of Jackknife+, Fast Jackknife+, Jackknife-minmax and Fast Jackknife-minmax, respectively. Table 3 shows the similarity, measured by Jaccard Index, between the prediction intervals constructed by the accelerated methods and their corresponding original methods.

We find that our accelerated methods substantially reduce the average computational time while maintaining coverage, and without significantly altering the length (efficiency) of the prediction intervals. The prediction intervals constructed by the accelerated methods exhibit a high degree of similarity to those from the original methods. We can observe that, in most cases, a smaller value of λ tends to result in a wider interval, and the interval length also increases as p grows. Moreover, our methods exhibit more significant acceleration when the dimensionality of the covariates is higher.

Compared with the synthetic-data simulation results reported in Table 1 of Clarté and Zdeborová [2024], both of our acceleration methods achieve higher coverage than Taylor-AMP, while also providing more efficient prediction (in terms of shorter average interval length) compared with Taylor-AMP, SCP [Vovk et al., 2005] and CQP [Romano et al., 2019]. With respect to Table 2 of Clarté and Zdeborová [2024], fast Jackknife+ and fast Jackknife-minmax exhibit higher Jaccard similarity to exact LOO than Taylor-AMP and SCP. Under Gaussian settings (Table 3 of Clarté and Zdeborová [2024]), our methods deliver more efficient prediction than Bayes posterior and FCP combined with Taylor-AMP. In comparison with Table 4 of Clarté and Zdeborová [2024], both of our accelerated procedures achieve higher prediction efficiency and higher coverage than Taylor-AMP and approximate homotopy [Ndiaye and Takeuchi, 2019], while exhibiting comparable computation time. Furthermore, our framework systematically explores multiple dimensions ($p = 50, 100, 200$), corresponding to different n/p ratios, and two regularization strengths ($\lambda = 0.1, 1$), demonstrating consistent performance across regimes.

These improvements can be attributed to two main factors. First, unlike Clarté and Zdeborová [2024], which assumes i.i.d. features with a diagonal covariance matrix, our framework accommodates a general covariance structure Σ. This relaxation is more realistic in practice and ensures that the resulting ALO characterization remains accurate even when the features are correlated, thereby preventing the deterioration observed for AMP and Taylor-AMP under non-isotropic designs. Second, while AMP-based methods rely on an ALO heuristic whose approximation error is not rigorously controlled, our approach benefits from a explicit and tighter ALO error bound derived via the Newton step and Woodbury identity. This leads to more accurate leave-one-out predictions, which in turn improves the quality of the prediction intervals and enhances efficiency.

5.2 Application to Real Data

In this section, we compare the performance of the fast Jackknife+ with the original Jackknife+, and the fast Jackknife-minmax with the original Jackknife-minmax on real data. We use for this two datasets-the Concrete Compressive Strength Dataset [Yeh, 1998] and the Energy Efficiency Dataset [Tsanas and Xifara, 2012].We validate that our methods provide the correct coverage and equally efficiency with faster speed.

In particular, from Table 4 and Table 5, we observe that our proposed methods, Fast Jackknife+ and Fast Jackknife-minmax, substantially accelerate their corresponding baseline counterparts while maintaining the predictive accuracy and efficiency.

Table 4: Performance the Concrete Compressive Strength Dataset

Method	Coverage Rate	Operation Time (s)	Average Interval Length
Jackknife+	0.9660	1.7275	37.0917
Fast Jackknife+	0.9660	0.0651	37.0851
Jackknife-minmax	0.9709	1.5735	37.8552
Fast Jackknife-minmax	0.9709	0.0414	37.8671

Table 5: Performance on the Energy Efficiency Dataset

Method	Coverage Rate	Operation Time (s)	Average Interval Length
Jackknife+	0.9351	1.6567	12.0404
Fast Jackknife+	0.9351	0.0509	12.0403
Jackknife-minmax	0.9481	1.0640	12.2182
Fast Jackknife-minmax	0.9481	0.0371	12.2181

5.3 Code

The code used to produce the results in Table 1, Table 2, Table 3, Table 4 and Table 5 can be found in the following github repository: `https://github.com/JiachenCong/Accelerating_Conformal_Prediction_via_Approximate_Leave-One-Out.git`. All experiments were run on a Apple M4 laptop with 16 Go of memory.

6 Discussion

A few questions naturally arise concerning the proposed algorithm:

Q1: We applied the leave-one-out approximation formula to accelerate jackknife+ and jackknife-mm, rather than full conformal prediction, so that rigorous results in [Rad and Maleki, 2020, Auddy et al., 2023, Zou et al., 2024] can be used to control the approximation error. Does this lead to worse efficiency (longer intervals)?

Q2: [Karoui et al., 2013] used leave-one-out to prove the central limit theorem M-estimators. Is it possible to further accelerate our algorithm by directly using the central limit theorem result?

In this section, we address these questions by studying the asymptotic performances of full conformal prediction, split conformal prediction, jackknife+, and jackknife-mm in a high-dimensional setting rigorously studied in [Karoui, 2018], although some of the conclusions are expected to hold in broader settings. Following [Karoui, 2018], we consider the model $y_i = X_i^\top \beta_0 + \epsilon_i$, and

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho_i \Big(y_i - X_i^\top \beta \Big) + \frac{\tau}{2} \|\beta\|^2. \tag{6.52}$$

Note that, following [Karoui, 2018], we assume the scaling $\|X_i\|_2 = \Theta(\sqrt{n})$ and $\|\beta\|_2 = \Theta(1)$, which is consistent with the scaling in the preceding sections if we take $X_i := \sqrt{n} x_i$ and $\beta := \frac{1}{\sqrt{n}} \theta$. Within the general framework in the preceding sections, (6.52) can be understood as the special case where the regularizer is ridge, the log-likelihood $\ell(y, z) = \rho(y - z)$ for some ρ, and $y_i = X_i^\top \beta + \epsilon_i$ under $p(y_i | X_i^\top \beta)$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. noise. The main contribution of [Karoui, 2018] was to use leave-one-out (both observation and predictor) to prove the asymptotic distribution of $\hat{\beta}$. Though more general sufficient conditions can be found in [Karoui, 2018], for simplicity let us assume the following condition (see [Karoui, 2018, Section 2.1]):

C1 p/n has a finite nonzero limit.

C2 ρ is convex, whose first and second derivatives are uniformly bounded: $\|\rho'\|_\infty < \infty$, $\|\rho''\|_\infty < \infty$. Furthermore, $\text{sign}(\rho'(t)) = \text{sign}(t)$ and $\rho(t) \geq \rho(0) = 0$ for all $t \in \mathbb{R}$.

C3 $X_1, \ldots, X_n$ are i.i.d. following $\mathcal{N}(0, I_p)$, or with i.i.d. entries with bounded support, symmetric density, and unit variance.

C4 $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. following a differentiable, symmetric, unimodal distribution with variance σ_ϵ^2, whose density f satisfies $\lim_{t \to \infty} t f(t) = 0$.

13

C5 $\|\beta_0\|_2$ remains bounded as $n \to \infty$, and $\|\beta_0\|_\infty = O(n^{-\epsilon})$ for some $\epsilon > 1/4$.

We denote by prox_ρ the proximal map:

$$\text{prox}_\rho(x) := \underset{y \in \mathbb{R}}{\arg\min}\{\rho(y) + \frac{1}{2}(x-y)^2\}. \tag{6.53}$$

Under the conditions above, the general result in Theorem 2.1 in [Karoui, 2018] implies the following:

Theorem 6.5. *[Karoui et al., 2013]* *Let $\delta := \lim_{n \to \infty} n/p \in (0, \infty)$, and $b_0 := \lim_{n \to \infty} \|\beta_0\|_2^2$. Let $c, r > 0$ be the solution to the following equations:*

$$\mathbb{E}\big[(\text{prox}'_{c\rho}(Z)\big] = 1 - \delta^{-1} + \tau, \tag{6.54}$$

$$\delta^{-1}\mathbb{E}\big[(Z - \text{prox}_{c\rho}(Z))^2\big] + \tau^2 c^2 b_0 = \delta^{-2} r^2 \tag{6.55}$$

where we defined the scalar $\epsilon \sim \mathcal{N}(0, 1)$ and $Z \sim \mathcal{N}(0, r^2) + \epsilon$. Then we have $\lim_{p \to \infty} \|\hat{\beta} - \beta_0\| = r$ and $\lim_{p \to \infty} \text{var}(\|\hat{\beta} - \beta_0\|^2) = 0$ in probability.

Denote by U_{J} and U_{mm} the upper limits of the conformal prediction intervals for the $(n+1)$-th observation produced by the jackknife+ and jackknife-minmax algorithms. For any given $M > 0$, let $U_{\text{f}}(M)$ be the upper limit of the discretized interval $\hat{C}(X_{n+1}) \cap [-M, M] \cap \{0, \pm\frac{1}{M}, \frac{2}{M}, \dots\}$, where $\hat{C}(X_{n+1})$ is the full conformal interval defined in (2) in the the main text. This discretization procedure is standard in practice and will also be used in the proof. Similarly, define $L_{\text{f}}(M)$, L_{J}, L_{mm} as the lower limits of the prediction intervals. Under the conditions we assumed, the lengths and centers of these intervals are all order of a constant, whereas the differences of these intervals vanish as $n \to \infty$ and $M \to \infty$, as the following result shows:

Theorem 6.6. *Define $U := X_{n+1}^\top \hat{\beta} + \sqrt{\sigma_\epsilon^2 + r^2} z_{\alpha/2}$, where σ_ϵ is defined in C4, r in Theorem 5 in the main text, and $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$ quantile of a standard normal random variable. Let $a_n := \max\{|U - U_{\text{J}}|, |U - U_{\text{mm}}|\}$ and $b_{n,M} := \max_{i=1,\dots,n} |U_{\text{f}} - U|$. Then we have*

$$\lim_{n \to \infty} a_n = 0 \tag{6.56}$$

$$\lim_{M \to \infty} \lim_{n \to \infty} b_{n,M} = 0 \tag{6.57}$$

in the sense of weak convergence. Moreover, the same holds for the lower bounds, with $L := X_{n+1}^\top \hat{\beta} - \sqrt{\sigma_\epsilon^2 + r^2} z_{\alpha/2}$.

The proof of Theorem 6.6 can be found in the supplement. This theorem shows that under the idealized assumptions in [Karoui, 2018], all the conformal prediction algorithms considered are asymptotically equivalent, partially answering Q1. On the other hand, the validity of $[L(i), U(i)]$ relies on the central limit result in [Karoui, 2018], which is much stronger than leave-observation-out approximation. In particular, the central limit result also relies on leave-predictor-out approximation in [Karoui, 2018], which requires independence of $X_1, \dots, X_n$, and exchangeability is not sufficient. [1] On the other hand, jackknife+ (as well as the proposed ALO version) remains valid in the general exchangeable case. This partially answers Q2.

---

[1] It uses the lemma that $v^\top A v \approx \text{Tr}(A)$ when v has i.i.d. coordinates independent of A and with zero mean and unit variance).

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Arnab Auddy, Haolin Zou, Kamiar Rahnama Rad, and Arian Maleki. Approximate leave-one-out cross validation for regression with ℓ_1 regularizers (extended version), 2023. URL https://arxiv.org/abs/2310.17629.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. https://doi.org/10.1214/20-AOS1965.

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3458–3468, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Lucas Clarté and Lenka Zdeborová. Building conformal prediction intervals with approximate message passing, 2024. URL https://arxiv.org/abs/2410.16493.

Noureddine El Karoui. On the impact of predictor geometry on the performance of high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018. doi:10.1007/s00440-016-0732-0.

Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013. doi:10.1073/pnas.1307842110.

Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy, 2019. URL https://arxiv.org/abs/1909.09365.

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 Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4067–4077. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/rad20a.html.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C51307.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. 01 2005. doi:10.1007/b106715.

Shuaiwen Wang, Wenda Zhou, Arian Maleki, Haihao Lu, and Vahab Mirrokni. Approximate leave-one-out for high-dimensional non-differentiable learning problems, 2018. URL https://arxiv.org/abs/1810.02716.

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. doi:10.1109/TIT.2021.3095375.

I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5PK67.

Haolin Zou, Arnab Auddy, Kamiar Rahnama Rad, and Arian Maleki. Theoretical analysis of leave-one-out cross validation for non-differentiable penalties under high-dimensional settings, 2024. https://arxiv.org/abs/2402.08543.

Supplementary Materials

In this supplementary material, we provide detailed proofs of the theorems presented in the main paper.

Organization

Table 6 presents the organization of the appendix.

Table 6: Organization of Supplementary Materials

Section	Purpose
Appendix S.1	Presents existing results adapted from [Rad and Maleki, 2020] for our proof.
Appendix S.2	Provides the complete proof of Theorem 1.
Appendix S.3	Provides the complete proof of Theorem 2.
Appendix S.4	Provides the complete proof of Theorem 3 and Theorem 4.
Appendix S.5	Provides the complete proof of Lemma 2.
Appendix S.6	Provides the complete proof of Theorem 6 and relative lemmas.

S.1 Existing Results

In this section, we present existing results used in our proof, which are adapted from [Rad and Maleki, 2020].

S.1.1 [Rad and Maleki, 2020, Eq.136]

Define $g_{/i}(\theta) = \lambda \dot{r}(\theta) + X_{/i}^{\top} \dot{\boldsymbol{\ell}}_{/i}(\theta)$. The leave-one-out estimate, $\hat{\beta}_{/i} = \hat{\beta} + \Delta_{/i}^{*}$, satisfies $g_{/i}(\Delta_{/i}^{*} + \hat{\beta}) = 0$. The multivariate mean-value Theorem yields

$$0 = g_{/i}(\hat{\beta} + \Delta_{/i}^{*}) = g_{/i}(\hat{\beta}) + \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\beta} + t\Delta_{/i}^{*}) dt \right) \Delta_{/i}^{*}, \tag{S.1.58}$$

where the Jacobian is

$$\boldsymbol{J}_{/i}(\theta) = \lambda \operatorname{diag}[\ddot{r}(\theta)] + X_{/i}^{\top} \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\theta)] X_{/i}. \tag{S.1.59}$$

Moreover, $\hat{\beta}$ satisfies

$$0 = \lambda \dot{r}(\hat{\beta}) + X^{\top} \dot{\boldsymbol{\ell}}(\hat{\beta}) = g_{/i}(\hat{\beta}) + \dot{\boldsymbol{\ell}}_i(\hat{\beta}) x_i. \tag{S.1.60}$$

We obtain

$$\dot{\boldsymbol{\ell}}_i(\hat{\beta}) x_i = - \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\beta} + t\Delta_{/i}^{*}) dt \right) \Delta_{/i}^{*}, \tag{S.1.61}$$

so that

$$\Delta_{/i}^{*} = -\dot{\boldsymbol{\ell}}_i(\hat{\beta}) \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\beta} + t\Delta_{/i}^{*}) dt \right)^{-1} x_i, \tag{S.1.62}$$

leading to the following inequality

$$\|\Delta_{/i}^{*}\|_2 \leq \left(\frac{|\dot{\boldsymbol{\ell}}_i(\hat{\beta})|}{\nu} \right) \|\boldsymbol{x}_i\|_2. \tag{S.1.63}$$

By Assumption 3.(b), we obtain the Eq.136 in [Rad and Maleki, 2020]

$$\|\Delta_{/i}^{*}\|_2 \leq \frac{c_1(n)}{\nu} \|\boldsymbol{x}_i\|_2 \tag{S.1.64}$$

S.1.2 [Rad and Maleki, 2020, Lemma 27]

Lemma 3 (Lemma 27 in [Rad and Maleki, 2020]). *Let $x \sim \mathcal{N}(0, \Sigma)$ with $\rho_{\max} := \sigma_{\max}(\Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$, then*

$$\mathbb{P}\left[\|x\|_4^2 > 2(1+c)\rho_{\max}\sqrt{p \log p}\right] \leq \frac{2}{p^c}. \tag{S.1.65}$$

Moreover, if

$$\omega_{\max} := \sigma_{\max}(XX^\top),$$
$$\nu_{\min} := \sigma_{\min}(J), \tag{S.1.66}$$

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

$$\mathbb{P}\left[\|J^{-1}x\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu_{\min}^2}\right)\sqrt{p \log p}\right] < \frac{2}{p^c}, \tag{S.1.67}$$

$$\mathbb{P}\left[\|XJ^{-1}x\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}\omega_{\max}}{\nu_{\min}^2}\right)\sqrt{m \log m}\right] < \frac{2}{m^c}. \tag{S.1.68}$$

The proof of this lemma can be found on pages 60 and 61 of [Rad and Maleki, 2020].

S.1.3 [Rad and Maleki, 2020, Lemma 11]

Lemma 4 (Lemma 11 in [Rad and Maleki, 2020]). *Let $x \sim \mathcal{N}(0, \Sigma)$ with $\rho_{\max} \triangleq \sigma_{\max}(\Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$, then*

$$\mathbb{P}\left[\|x\|_2^2 > 5p\rho_{\max}\right] \leq e^{-p}. \tag{S.1.69}$$

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

$$\mathbb{P}\left[\sqrt{\sigma_{\max}(X^\top X)} \geq 1 + \sqrt{\frac{p}{n}} + t\right] \leq e^{-\frac{nt^2}{2}}. \tag{S.1.70}$$

S.1.4 [Rad and Maleki, 2020, Lemma 12]

Lemma 5 (Lemma 12 in [Rad and Maleki, 2020]). *If $X \in \mathbb{R}^{n \times p}$ is composed of independently distributed $\mathcal{N}(0, \Sigma)$ rows, with $\rho_{\max} := \sigma_{\max}(\Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$, then*

$$\mathbb{P}\left[\sigma_{\max}(XX^\top) \geq (\sqrt{n} + 3\sqrt{p})^2 \rho_{\max}\right] \leq e^{-p}.$$

The proof of this lemma can be found on pages 34 and 35 of [Rad and Maleki, 2020].

S.1.5 [Rad and Maleki, 2020, Theorem 3]

Theorem S.1.7 (Theorem 3 in [Rad and Maleki, 2020]). *Let $n/p = \delta_0$ and [Rad and Maleki, 2020, Assumption 5] hold with $\rho_{\max} = c/p$. Moreover, suppose that [Rad and Maleki, 2020, Assumption 6 and 7] are satisfied, and that n is large enough such that $q_n + \tilde{q}_n < 0.5$. Then with probability at least*

$$1 - 4ne^{-p} - \frac{8n}{p^3} - \frac{8n}{(n-1)^3} - q_n - \tilde{q}_n, \tag{S.1.71}$$

the following bound is valid:

$$\max_{1 \leq i \leq n}\left|x_i^\top \hat{\beta}_{/i} - x_i^\top \hat{\beta} - \left(\frac{\dot{\ell}_i(\hat{\beta})}{\ddot{\ell}_i(\hat{\beta})}\right)\left(\frac{H_{ii}}{1 - H_{ii}}\right)\right| \leq \frac{Q_0}{\sqrt{p}}, \tag{S.1.72}$$

where

$$Q_0 := \left(\frac{72c^{3/2}}{\nu^3}\right)\left(1 + \sqrt{\delta_0}(\sqrt{\delta_0} + 3)^2 \frac{c \log n}{\log p}\right)\left(c_1^2(n)c_2(n) + c_1^3(n)c_2^2(n)\frac{5(c^{1/2} + c^{3/2}(\sqrt{\delta_0} + 3)^2)}{\nu^2}\right). \tag{S.1.73}$$

S.2 Proof of Theorem 1

We introduce the following definitions:

- (1) $\Delta^*_{/i} := \hat{\theta}_{/i} - \hat{\theta}$, and $\boldsymbol{J}_{/i} := \lambda \operatorname{diag}[\ddot{r}_s(\hat{\theta}_{/i})] + X^{\top}_{/i} \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta}_{/i})]\, X_{/i}$, where $X_{/i} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_n)^{\top}$.

- (2) Events $\tilde{E}_i := \left\{ 3\bar{C}_i \left(\left\| X_{/i} \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) \boldsymbol{x}_i \right\|^2_4 + \left\| \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) \boldsymbol{x}_i \right\|^2_4 \right) > C \frac{\log p}{\sqrt{p}} \right\}$,

  and $\tilde{E}_{n+1,i} := \left\{ 3\bar{C}_i \left(\left\| X_{/i} \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) \boldsymbol{x}_{n+1} \right\|^2_4 + \left\| \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) \boldsymbol{x}_{n+1} \right\|^2_4 \right) > C \frac{\log p}{\sqrt{p}} \right\}$.

- (3) $\bar{C}_i := 4\|\boldsymbol{x}_i\|_2 \left(\frac{c_1^2(n) c_2(n)}{\nu} \right) \left(1 + \frac{2 c_1(n) c_2(n)}{\nu^2} (1 + \omega_{\max}) \|\boldsymbol{x}_i\|_2 \right)$.

- (4) Define $\Phi_1 := \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}^{\top}_{/i} \operatorname{diag}\left[\int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i}) dt \right] \bar{X}_{/i} \right)^{-1}$ and

  $\Phi_2 := \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}^{\top}_{/i} \operatorname{diag}\left[\gamma_{-\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right] \bar{X}_{/i} \right)^{-1}$, where for $\delta, \theta \in \mathbb{R}^p$

  $$\gamma_{\delta/i}(\theta) := \begin{bmatrix} \ddot{\boldsymbol{\ell}}_{/i}(\theta + \delta) - \ddot{\boldsymbol{\ell}}_{/i}(\theta) \\ \lambda\big(\ddot{r}_s(\theta + \delta) - \ddot{r}_s(\theta)\big) \end{bmatrix}, \quad \bar{X}_{/i} := \begin{bmatrix} X_{/i} \\ I \end{bmatrix} \in \mathbb{R}^{(n-1+p) \times p},$$

  $$\mathbf{D}_{/i}(\theta) := \operatorname{diag}\begin{bmatrix} \ddot{\boldsymbol{\ell}}_{/i}(\theta) \\ \lambda \ddot{r}_s(\theta) \end{bmatrix} \in \mathbb{R}^{(n-1+p) \times (n-1+p)}.$$

  Define $\boldsymbol{J}_{/i}(\theta) := \bar{X}^{\top}_{/i} \mathbf{D}_{/i}(\theta) \bar{X}_{/i}$.

  Consider

  $$\begin{aligned} & \boldsymbol{x}^{\top}_{n+1} \tilde{\theta}_{/i} - \boldsymbol{x}^{\top}_{n+1} \hat{\theta}_{/i} \\ &= \dot{\ell}_i(\hat{\theta}) \boldsymbol{x}^{\top}_{n+1} M \boldsymbol{x}_i \end{aligned} \tag{S.2.74}$$

  $$= \frac{1}{2}\dot{\ell}_i(\hat{\theta})(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^{\top} M (\boldsymbol{x}_i + \boldsymbol{x}_{n+1}) - \frac{1}{2}\dot{\ell}_i(\hat{\theta}) \boldsymbol{x}^{\top}_{n+1} M \boldsymbol{x}_{n+1} - \frac{1}{2}\dot{\ell}_i(\hat{\theta}) \boldsymbol{x}^{\top}_i M \boldsymbol{x}_i, \tag{S.2.75}$$

where we defined

$$M := \left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta^*_{/i}) dt \right)^{-1} - \boldsymbol{J}^{-1}_{/i}(\hat{\theta}_{/i} - \Delta^*_{/i}). \tag{S.2.76}$$

Let A_1, A_2, A_3 denote the three terms in (S.2.75). Then clearly

$$|\boldsymbol{x}^{\top}_{n+1} \tilde{\theta}_{/i} - \boldsymbol{x}^{\top}_{n+1} \hat{\theta}_{/i}| \le |A_1| + |A_2| + |A_3|. \tag{S.2.77}$$

By Assumption 3.(a), we have

$$\dot{\ell}_i(\hat{\theta}) = O(\text{polylog}(n)). \tag{S.2.78}$$

By the definition of $\gamma_{\delta/i}(\theta)$, we have

$$\boldsymbol{J}_{/i}(\theta + \delta) = \boldsymbol{J}_{/i}(\theta) + \bar{X}^{\top}_{/i} \operatorname{diag}\left[\gamma_{\delta/i}(\theta) \right] \bar{X}_{/i}. \tag{S.2.79}$$

For the new A_1 term, we have

$$\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top M(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$= \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta_{/i}^*)dt\right)^{-1} - \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i} - \Delta_{/i}^*)\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$= \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\left(\int_0^1 \boldsymbol{J}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta_{/i}^*)dt\right)^{-1} - \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$+ \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i} - \Delta_{/i}^*)\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$\leq^1 \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\left[\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})dt\right]\bar{X}_{/i}\right)^{-1}\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$+ \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \left[\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i}) - \left(\boldsymbol{J}_{/i}(\hat{\theta}_{/i}) + \bar{X}_{/i}^\top \operatorname{diag}\left[\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right]\bar{X}_{/i}\right)^{-1}\right](\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$= \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right| + \left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|, \tag{S.2.80}$$

where $\leq^1$ is due to (S.2.79).

For the $\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$ term, using [Rad and Maleki, 2020, Lemma 25] we have

$$\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$\leq \left(\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right\|_4^2$$

$$\leq \left(\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2^2\right)\left(\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4 + \left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4\right)^2$$

$$\leq 2\left(\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2$$

$$+ 2\left(\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2. \tag{S.2.81}$$

Similarly, for the term $\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$, we can prove the following bound

$$\left|(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1(\boldsymbol{x}_i + \boldsymbol{x}_{n+1})\right|$$

$$\leq 2\left(\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\,dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\,dt\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2$$

$$+ 2\left(\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\,dt\right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu}\right)\left\|\int_0^1 \gamma_{-(1-t)\Delta_{/i}^*/i}(\hat{\theta}_{/i})\,dt\right\|_2^2\right)\left\|\bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2. \tag{S.2.82}$$

Under Assumption 4, we have

$$\left\|\gamma_{-\Delta_{/i}^*/i}(\hat{\theta}_{/i})\right\|_2 \leq \left\|\ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta}_{/i} - \Delta_{/i}^*) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta}_{/i})\right\|_2 + \left\|\lambda\left(\ddot{r}_s(\hat{\theta}_{/i} - \Delta_{/i}^*) - \ddot{r}_s(\hat{\theta}_{/i})\right)\right\|_2$$

$$\leq 2c_2(n)\left\|\Delta_{/i}^*\right\|_2. \tag{S.2.83}$$

Likewise,

$$\left\| \int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i})\,dt \right\|_2 \leq \int_0^1 \left\| \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right\|_2 dt$$

$$\leq \int_0^1 \left\| \ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta}_{/i} - (1-t)\Delta^*_{/i}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\theta}_{/i}) \right\|_2 dt$$

$$+ \int_0^1 \left\| \lambda\big(\ddot{r}_s(\hat{\theta}_{/i} - (1-t)\Delta^*_{/i}) - \ddot{r}_s(\hat{\theta}_{/i})\big) \right\|_2 dt$$

$$\leq 2c_2(n)\left\| \Delta^*_{/i} \right\|_2. \tag{S.2.84}$$

By the result in Section S.1.1, we have

$$\left\| \Delta^*_{/i} \right\|_2 \ \leq\ \left(\frac{|\dot{\ell}_i(\hat{\theta})|}{\nu} \right) \left\| \boldsymbol{x}_i \right\|_2 \leq \frac{c_1(n)}{\nu} \|\boldsymbol{x}_i\|_2. \tag{S.2.85}$$

Using results in (S.2.83), (S.2.84) and (S.2.85), we obtain

$$|A_1| \leq \frac{1}{2}|\dot{\ell}_i(\hat{\theta})| \cdot \left| (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_1(\boldsymbol{x}_i + \boldsymbol{x}_{n+1}) \right| + \frac{1}{2}|\dot{\ell}_i(\hat{\theta})| \cdot \left| (\boldsymbol{x}_i + \boldsymbol{x}_{n+1})^\top \Phi_2(\boldsymbol{x}_i + \boldsymbol{x}_{n+1}) \right|$$

$$\leq |\dot{\ell}_i(\hat{\theta})| \cdot \left(\left\| \gamma_{-\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu} \right) \left\| \gamma_{-\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right\|_2^2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^2$$

$$+ |\dot{\ell}_i(\hat{\theta})| \cdot \left(\left\| \gamma_{-\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu} \right) \left\| \gamma_{-\Delta^*_{/i}/i}(\hat{\theta}_{/i}) \right\|_2^2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2$$

$$+ |\dot{\ell}_i(\hat{\theta})| \cdot \left(\left\| \int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i})\,dt \right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu} \right) \left\| \int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i})\,dt \right\|_2^2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^2$$

$$+ |\dot{\ell}_i(\hat{\theta})| \cdot \left(\left\| \int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i})\,dt \right\|_2 + \left(\frac{\bar{\omega}_{\max,i}}{\nu} \right) \left\| \int_0^1 \gamma_{-(1-t)\Delta^*_{/i}/i}(\hat{\theta}_{/i})\,dt \right\|_2^2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2$$

$$\leq 4\|\boldsymbol{x}_i\|_2 \left(\frac{c_1^2(n)c_2(n)}{\nu} \right) \left(1 + \frac{2c_1(n)c_2(n)}{\nu^2}(1+\omega_{\max,i})\|\boldsymbol{x}_i\|_2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2$$

$$+ 4\|\boldsymbol{x}_i\|_2 \left(\frac{c_1^2(n)c_2(n)}{\nu} \right) \left(1 + \frac{2c_1(n)c_2(n)}{\nu^2}(1+\omega_{\max,i})\|\boldsymbol{x}_i\|_2 \right) \left\| \bar{X}_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^2$$

$$\leq 4\|\boldsymbol{x}_i\|_2 \left(\frac{c_1^2(n)c_2(n)}{\nu} \right) \left(1 + \frac{2c_1(n)c_2(n)}{\nu^2}(1+\omega_{\max,i})\|\boldsymbol{x}_i\|_2 \right) \sqrt{\left\| X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^4 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^4}$$

$$+ 4\|\boldsymbol{x}_i\|_2 \left(\frac{c_1^2(n)c_2(n)}{\nu} \right) \left(1 + \frac{2c_1(n)c_2(n)}{\nu^2}(1+\omega_{\max,i})\|\boldsymbol{x}_i\|_2 \right) \sqrt{\left\| X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^4 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^4}$$

$$\leq \bar{C}_i \left[\left\| X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2 \right] + \bar{C}_i \left[\left\| X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^2 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i \right\|_4^2 \right],$$

$$\tag{S.2.86}$$

where $\omega_{\max,i} := \sigma_{\max}(X_{/i}X_{/i}^\top)$, $\omega_{\max} := \sigma_{\max}(XX^\top)$ and $\bar{\omega}_{\max,i} := \sigma_{\max}\big(\bar{X}_{/i}\bar{X}_{/i}^\top\big)$. By this definition, we can derive that

$$\bar{\omega}_{\max,i} = \sigma_{\max}\big(\bar{X}_{/i}\bar{X}_{/i}^\top\big) = \sigma_{\max}\big(\bar{X}_{/i}^\top\bar{X}_{/i}\big) = \sigma_{\max}\left(\begin{bmatrix} X_{/i} \\ I \end{bmatrix}^\top \begin{bmatrix} X_{/i} \\ I \end{bmatrix} \right)$$

$$= \sigma_{\max}\big(I + X_{/i}^\top X_{/i}\big) \leq 1 + \sigma_{\max}\big(X_{/i}^\top X_{/i}\big) = 1 + \omega_{\max,i}, \tag{S.2.87}$$

For $|A_2|$ and $|A_3|$, we can use a similar method to show that

$$|A_2| \leq \frac{1}{2}\bar{C}_i \left[\left\| X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2 + \left\| \boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1} \right\|_4^2 \right] \tag{S.2.88}$$

and

$$|A_3| \leq \frac{1}{2}\bar{C}_i \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \right]. \tag{S.2.89}$$

Combining the results in (S.2.86), (S.2.88) and (S.2.89), we obtain

$$
\begin{aligned}
|\boldsymbol{x}_{n+1}^{\top}\tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^{\top}\hat{\theta}_{/i}| \leq &\frac{3}{2}\bar{C}_i \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \right] \\
&+ \frac{3}{2}\bar{C}_i \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 \right] \\
\leq &3\bar{C}_i \cdot \max_{j \in \{i, n+1\}} \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_j\right\|_4^2 \right]. \tag{S.2.90}
\end{aligned}
$$

Next, we define the event

$$G := \left\{ \max_{1 \leq i \leq n} |\boldsymbol{x}_{n+1}^{\top}\hat{\theta}_{/i} - \boldsymbol{x}_{n+1}^{\top}\tilde{\theta}_{/i}| \geq C\frac{\log(p)}{\sqrt{p}} \right\}.$$

Adopting the reasoning on pages 54–55 of [Rad and Maleki, 2020], where S denotes the event that Assumption 3 in our paper holds (trivially we have $\mathbb{P}(S^c) \leq q_n + \tilde{q}_n$), and q_n and $\tilde{q}_n$ follow the definitions in Assumption 3, we have

$$
\begin{aligned}
\mathbb{P}[G] \leq & \mathbb{P}[G|S] + \mathbb{P}[S^c] \\
\leq & \mathbb{P}[\max_{1 \leq i \leq n}(\max_{j \in \{i, n+1\}} |\mathbf{x}_j^{\top}\hat{\theta}_{/i} - \mathbf{x}_j^{\top}\tilde{\theta}_{/i}| > C\frac{\log(p)}{\sqrt{p}}|S] + q_n + \tilde{q}_n \\
\leq & \frac{\mathbb{P}[\max_{1 \leq i \leq n}|\boldsymbol{x}_i^{\top}\hat{\theta}_{/i} - \boldsymbol{x}_i^{\top}\tilde{\theta}_{/i}| > C\frac{\log(p)}{\sqrt{p}}] + \mathbb{P}[\max_{1 \leq i \leq n}|\boldsymbol{x}_{n+1}^{\top}\hat{\theta}_{/i} - \boldsymbol{x}_{n+1}^{\top}\tilde{\theta}_{/i}| > C\frac{\log(p)}{\sqrt{p}}]}{1 - q_n - \tilde{q}_n} + q_n + \tilde{q}_n \\
\leq & \frac{\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}]}{1 - q_n - \tilde{q}_n} \\
& + \frac{\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}]}{1 - q_n - \tilde{q}_n} + q_n + \tilde{q}_n. \tag{S.2.91}
\end{aligned}
$$

For $\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}]$ and
$\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}]$, we have

$$\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_i\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}] \leq \sum_{i=1}^{n} \mathbb{P}[\tilde{E}_i]; \tag{S.2.92}$$

$$\mathbb{P}[3\bar{C}_i \cdot \left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}(\hat{\theta}_{/i})\boldsymbol{x}_{n+1}\right\|_4^2 \right] > C\frac{\log(p)}{\sqrt{p}}] \leq \sum_{i=1}^{n} \mathbb{P}[\tilde{E}_{n+1,i}]. \tag{S.2.93}$$

For simplicity, we define

$$
\begin{aligned}
C_i &:= 6(1+c)\left(\frac{\rho_{\max}}{\nu^2}\right)\left(1 + \omega_{\max}\sqrt{\frac{n-1}{p}\frac{\log(n-1)}{\log p}}\right), \\
F_{n+1,i} &:= \left\{ 3\bar{C}_i\left(\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{n+1}\right\|_4^2 + \left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{n+1}\right\|_4^2 \right) > \bar{C}_i C_i\sqrt{p}\log p \right\}, \\
K_i &:= \left\{ \frac{C}{\sqrt{p}} \geq \bar{C}_i C_i\sqrt{p} \right\}, \\
W_{n+1} &:= \left\{ \|\boldsymbol{x}_{n+1}\|_2^2 > 5p\rho_{\max} \right\} \cup \left\{ \omega_{\max} > (\sqrt{n} + 3\sqrt{p})^2\rho_{\max} \right\}.
\end{aligned}
$$

Hence, we now obtain an upper bound for $\mathbb{P}[\tilde{E}_{n+1,i}]$

$$\mathbb{P}[\tilde{E}_{n+1,i}] \leq \mathbb{P}[\tilde{E}_{n+1,i} \mid K_i] + \mathbb{P}[K_i^c]$$

$$\leq \mathbb{P}[F_{n+1,i} \mid K_i] + \mathbb{P}[K_i^c] \;\leq\; \frac{\mathbb{P}(F_{n+1,i})}{\mathbb{P}(K_i)} + \mathbb{P}[K_i^c]$$

$$\leq \frac{\mathbb{P}\left[\left\|X_{/i}\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{n+1}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu^2}\omega_{\max}\right)\sqrt{n-1}\,\log(n-1)\right]}{\mathbb{P}(K_i)}$$

$$+ \frac{\mathbb{P}\left[\left\|\boldsymbol{J}_{/i}^{-1}\boldsymbol{x}_{n+1}\right\|_4^2 > 2(1+c)\left(\frac{\rho_{\max}}{\nu^2}\right)\sqrt{p}\,\log p\right]}{\mathbb{P}(K_i)} + \mathbb{P}[K_i^c]$$

$$\leq^1 \left(\frac{2}{(n-1)^c} + \frac{2}{p^c}\right)\frac{1}{\mathbb{P}(K_i)} + \mathbb{P}[K_i^c], \tag{S.2.94}$$

where $\leq^1$ is due to (S.1.68). To bound $\mathbb{P}[K_i^c]$ we define

$$C := 96\sqrt{5}\left(\frac{c_1^2(n)c_2(n)\,(p\rho_{\max})^{3/2}}{\nu^3}\right)\left(1 + \left(\sqrt{\tfrac{n}{p}} + 3\right)^2 p\rho_{\max}\sqrt{\tfrac{n-1}{p}\,\tfrac{\log(n-1)}{\log p}}\right)$$

$$\times \left(1 + \frac{2c_1(n)c_2(n)\sqrt{5}\left(1 + \left(\sqrt{\tfrac{n}{p}} + 3\right)^2 p\rho_{\max}\right)\sqrt{p\rho_{\max}}}{\nu^2}\right).$$

obtained by setting $c = 3$, and computing $p\bar{C}_iC_i$ after putting $\sqrt{5p\rho_{\max}}$ and $(\sqrt{n/p} + 3\sqrt{p})^2\rho_{\max}$ bounds in event W_{n+1}, into $\|\boldsymbol{x}_{n+1}\|_2$ and $\omega_{\max}$, respectively. Next,

$$\mathbb{P}[K_i^c] = \mathbb{P}\left[\frac{C}{p} < \bar{C}_iC_i\right] \;\leq\; \mathbb{P}\left[C < p\bar{C}_iC_i \mid W_{n+1}^c\right] + \mathbb{P}[W_{n+1}]$$

$$= \mathbb{P}[C < C] + \mathbb{P}[W_{n+1}] \;=\; \mathbb{P}[W_{n+1}]. \tag{S.2.95}$$

The term $\mathbb{P}[W_{n+1}]$ is exponentially small because $\boldsymbol{x}_{n+1} \sim N(0, \Sigma)$ with $\rho_{\max} = \sigma_{\max}(\Sigma)$, leading to

$$\mathbb{P}[W_{n+1}] \;\leq\; \mathbb{P}\left[\|\boldsymbol{x}_{n+1}\|_2^2 > 5p\rho_{\max}\right] + \mathbb{P}\left[\sigma_{\max}(XX^\top) > (\sqrt{n} + 3\sqrt{p})^2\rho_{\max}\right] \;\leq\; 2e^{-p},$$

due to result in Section S.1.3 and Section S.1.4. In summary, since for $p \geq 1$ we have $\frac{1}{1-e^{-p}} < 2$, for $c = 3$ we obtain

$$\mathbb{P}[\tilde{E}_{n+1,i}] \;\leq\; \frac{4}{(n-1)^3} + \frac{4}{p^3} + 2e^{-p}.$$

Since $\boldsymbol{x}_i$ and $\boldsymbol{x}_{n+1}$ follow same distribution, that is $\boldsymbol{x}_i \sim N(0, \Sigma)$, the same argument implies that

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

The detailed proof for the bound of $\mathbb{P}[\tilde{E}_i]$ can also be found in [Rad and Maleki, 2020].

As a result, we have

$$\mathbb{P}[G] \leq \frac{1}{1 - q_n - \tilde{q}_n}\left(\frac{8n}{(n-1)^3} + \frac{8n}{p^3} + 4ne^{-p}\right) + q_n + \tilde{q}_n$$

$$\leq \left(\frac{16n}{(n-1)^3} + \frac{16n}{p^3} + 8ne^{-p}\right) + q_n + \tilde{q}_n. \tag{S.2.96}$$

It is clear that $\left[\left(\frac{16n}{(n-1)^3} + \frac{16n}{p^3} + 8ne^{-p}\right) + q_n + \tilde{q}_n\right] \to 0$ when $n \to \infty$.

Thus, we obtain a result similar to the result in Section S.1.5. That is, with probability at least $1 - \left(\frac{16n}{(n-1)^3} + \frac{16n}{p^3} + 8ne^{-p} \right) - q_n - \tilde{q}_n$, the following bound is valid

$$\max_{1 \leq i \leq n} |\boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i}| \leq \frac{C_0}{\sqrt{p}},$$

where C_0 is defined as follows

$$C_0 := \left(\frac{216 c^{3/2}}{\nu^3} \right) \left(1 + \sqrt{\delta_0}(\sqrt{\delta_0} + 3)^2 \frac{c \log n}{\log p} \right) \left(c_1^2(n) c_2(n) + c_1^3(n) c_2^2(n) \frac{5 \left(c^{1/2} + c^{3/2}(\sqrt{\delta_0} + 3)^2 \right)}{\nu^2} \right).$$

$$(S.2.97)$$

Note that in the presentation of Theorem 1, we replaced respectively $96\sqrt{5}$ and $2\sqrt{3}$ with the upper-bounds 216 and 5, and we replaced $\sqrt{\frac{n-1}{p} \frac{\log(n-1)}{\log p}}$ with the upper bound $\sqrt{\frac{n}{p} \frac{\log n}{\log p}}$. We also used δ_0 to denote n/p.

S.3 Proof of Theorem 2

Using result Section S.1.5, with probability at least $1 - p_2(n)$, the following holds

$$\max_{1 \leq i \leq n} |\boldsymbol{x}_i^\top \tilde{\theta}_{/i} - \boldsymbol{x}_i^\top \hat{\theta}_{/i}| \leq \frac{Q_0}{\sqrt{p}}. \tag{S.3.98}$$

Then with probability at least $1 - p_2(n)$, following holds

$$\max_{1 \leq i \leq n} |\boldsymbol{x}_i^\top \tilde{\theta}_{/i} - \boldsymbol{x}_i^\top \hat{\theta}_{/i}| = o(1). \tag{S.3.99}$$

Plugging (S.3.99) into $\left| R_i^{LOO} - \tilde{R}_i^{LOO} \right|$, using Assumption 4, with probability at least $1 - p_2(n)$ the following result holds

$$\begin{aligned}
\left| R_i^{LOO} - \tilde{R}_i^{LOO} \right| &= \left| (|y_i - f(\boldsymbol{x}_i^\top \hat{\theta}_{/i})| - |y_i - f(\boldsymbol{x}_i^\top \tilde{\theta}_{/i})|) \right| \\
&\leq \left| f(\boldsymbol{x}_i^\top \hat{\theta}_{/i}) - f(\boldsymbol{x}_i^\top \tilde{\theta}_{/i}) \right| \\
&\leq L \cdot \max_{1 \leq i \leq n} |\boldsymbol{x}_i^\top \hat{\theta}_{/i} - \boldsymbol{x}_i^\top \tilde{\theta}_{/i}| \\
&= o(1).
\end{aligned} \tag{S.3.100}$$

Using Theorem 1 and Assumption 4, with probability at least $1 - p_1(n)$ the following result holds

$$\left| \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) \right| \leq L \cdot \left| \boldsymbol{x}_{n+1}^\top \tilde{\theta}_{/i} - \boldsymbol{x}_{n+1}^\top \hat{\theta}_{/i} \right| = o(1). \tag{S.3.101}$$

Then with probability at least $(1 - p_1(n))(1 - p_2(n))$, the following results hold for $i = 1, 2, ..., n$:

$$\left| (\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{R}_i^{LOO}) - (\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - R_i^{LOO}) \right| = o(1); \tag{S.3.102}$$

$$\left| (\tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) + \tilde{R}_i^{LOO}) - (\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) + R_i^{LOO}) \right| = o(1). \tag{S.3.103}$$

With (S.3.102) and (S.3.103), using Lemma 2, we can easily derive that

$$\hat{q}_{n,\alpha}^- \{ \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) - \tilde{R}_i^{LOO} \} = \hat{q}_{n,\alpha}^- \{ \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - R_i^{LOO} \} + o(1); \tag{S.3.104}$$

$$\hat{q}_{n,\alpha}^+ \{ \tilde{\mu}_{/i}(\boldsymbol{x}_{n+1}) + \tilde{R}_i^{LOO} \} = \hat{q}_{n,\alpha}^+ \{ \hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) + R_i^{LOO} \} + o(1). \tag{S.3.105}$$

That is, with probability at least $(1 - p_1(n))(1 - p_2(n))$, the following results hold

$$\tilde{U}_{jk+} = U_{jk+} + o(1), \; \tilde{L}_{jk+} = L_{jk+} + o(1). \tag{S.3.106}$$

We can utilize similar method to conclude that, with probability at least $(1 - p_1(n))(1 - p_2(n))$, the following results hold

$$\tilde{U}_{jkm} = U_{jkm} + o(1), \; \tilde{L}_{jkm} = L_{jkm} + o(1). \tag{S.3.107}$$

Theorem 2 has been proved.

S.4 Proof of Theorem 3 & 4

Recall that in Assumption 4 we defined that

$$p_1(n) = \left(\frac{16n}{(n-1)^3} + \frac{16n}{p^3} + 8ne^{-p} \right) + q_n + \tilde{q}_n$$

and

$$p_2(n) = \left(\frac{8n}{(n-1)^3} + \frac{8n}{p^3} + 4ne^{-p} \right) + q_n + \tilde{q}_n.$$

Under Assumption 1 to Assumption 4, and the condition that $n, p \to \infty$, $n/p = \delta_0 \in (0, \infty)$, it is clear that

$$p_1(n) \xrightarrow{n \to \infty} 0, \quad p_2(n) \xrightarrow{n \to \infty} 0.$$

Using (S.3.104), (S.3.105) and Assumption 4, we obtain that, with probability at least $(1 - p_1(n))(1 - p_2(n))$, the following holds

$$\mathbb{P}\{y_{n+1} \in \tilde{C}_{n,\alpha}^{Jackknife+}\} = \mathbb{P}\{y_{n+1} \in \hat{C}_{n,\alpha}^{Jackknife+}\} + L \cdot o(1) = \mathbb{P}\{y_{n+1} \in \hat{C}_{n,\alpha}^{Jackknife+}\} + o(1). \tag{S.4.108}$$

According to [Barber et al., 2021], we have

$$\hat{C}_{n,\alpha}^{Jackknife+} = [\hat{q}_{n,\alpha}^- \{\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) - R_i^{LOO}\}, \hat{q}_{n,\alpha}^+ \{\hat{\mu}_{/i}(\boldsymbol{x}_{n+1}) + R_i^{LOO}\}] \tag{S.4.109}$$

and

$$\mathbb{P}\{y_{n+1} \in \hat{C}_{n,\alpha}^{Jackknife+}\} \geq 1 - 2\alpha. \tag{S.4.110}$$

As s result, using (S.4.110), we have

$$\mathbb{P}\{y_{n+1} \in \tilde{C}_{n,\alpha}^{Jackknife+}\} \geq (1 - 2\alpha - o(1)) \cdot (1 - p_1(n))(1 - p_2(n)) = 1 - 2\alpha - o(1). \tag{S.4.111}$$

Theorem 3 has been proved.
Proof of Theorem 4 is similar to Theorem 3.

S.5 Proof of lemmas

S.5.1 Proof of Lemma 2

Lemma 2 follows from the following more precise version:

Lemma 6. *Let $\{a_i\}_{i=1}^n$ be a sorted sequence such that $a_1 \leq a_2 \leq \cdots \leq a_n$, and suppose there exist sequences $\{b_i\}_{i=1}^n$ satisfying $|b_i - a_i| \leq \varepsilon$, for some $\varepsilon > 0$ and all $i = 1, \ldots, n$. Then for any $q \in \{1, 2, ..., n\}$ we have*

$$|b_{(q)} - a_q| \leq \epsilon,$$

where $b_{(q)}$ is the q-th smallest value of $\{b_i\}_{i=1}^n$.

Proof. Under the constraint of $b_i - a_i \leq \epsilon$, the maximum of $b_{(q)}$ is achieved when $b_i = a_i + \epsilon$ for each i, in which case $b_{(q)} = a_q + \epsilon$. Hence $b_{(q)} \leq a_q + \epsilon$. Similarly, $b_{(q)} \geq a_q - \epsilon$, and the proof is completed. $\square$

S.6 Proof of Theorem 6

In this section, we prove Theorem 6 in the main text. The statement of Theorem 6 and its preceding paragraph in the main text contain some typos (boundaries of the intervals, such as $U_{\mathrm{J}}(i)$, should be U_{J}, independent of i). Therefore, we reproduce Theorem 6 and its setting below for clarity:

The proof of Theorem 6.6 relies on some auxiliary results:

Lemma 7. *Under assumptions C1-C5, we have the following (in the sense of weak convergence)*

$$\max_{1 \le i \le n} |\hat{\mu}_{/i}(X_{n+1}) - \hat{\mu}(X_{n+1})| = o(1). \tag{S.6.112}$$

Proof. Recall that [Karoui, 2018, Theorem 2.2] showed that

$$\max_{1 \le i \le n} \|\hat{\beta} - \hat{\beta}_{/i} - \eta_i\|_2 = O_{L_2}\left(\frac{\mathrm{polylog}(n)}{n}\right), \tag{S.6.113}$$

where the notations are explained as follows: $O_{L_2}(\cdot)$ denotes a bound on a random variable in the square root of the second moment, and $\hat{\beta}_{/i} + \eta_i$ can be thought of as an approximate formula for $\hat{\beta}$ using the leave-one-out estimator $\hat{\beta}_{/i}$. The correction term

$$\eta_i := \frac{1}{n}(S_i + \tau I)^{-1} X_i \rho'\left(\mathrm{prox}_{c_i \rho_i}(r_{i,(i)})\right), \tag{S.6.114}$$

where we defined

$$c_i := \frac{1}{n} X_i^\top (S_i + \tau I)^{-1} X_i; \tag{S.6.115}$$

$$S_i := \frac{1}{n} \sum_{j \ne i} \rho''(r_{j,(i)}) X_j X_j^\top, \tag{S.6.116}$$

where

$$r_{j,(i)} := y_j - X_j^\top \hat{\beta}_{/j} \tag{S.6.117}$$

denotes the leave-one-out residual (in particular, $r_{i,(i)} = R_i^{LOO}$).

Noting that S_i, X_i, and X_{n+1} are independent, and X_i and X_{n+1} are i.i.d. with zero mean and identity covariance, we see that $\frac{1}{n} X_{n+1}^\top (S_i + \tau I)^{-1} X_i$ has zero mean and $O(\frac{1}{n})$ second moment. Under the assumption of $\|\rho'\|_\infty < \infty$, this in turn implies, by (S.6.114), that $X_{n+1}^\top \eta_i = o_{L_2}(1)$ (vanishing in the sense of second moment). Then using the fact that $\|X_{n+1}\|_2 = O_{L_2}(\sqrt{n})$ and (S.6.113) we obtain $\max_{1 \le i \le n} \|X_{n+1}^\top (\hat{\beta} - \hat{\beta}_{/i})\|_2 = o_{L_2}(1)$, which is equivalent to (S.6.112). $\square$

Next, recall that $\hat{\mu}^y(X_i) := X_i^\top \hat{\beta}^y$, $i = 1, \ldots, n+1$, where $\hat{\beta}^y$ denotes the output of the regression algorithm trained on the augmented dataset $\{(X_1, y_1), \ldots, (X_n, y_n), (X_{n+1}, y)\}$. Also, adopt the notations

$$R_i := y_i - X_i^\top \hat{\beta}, \quad i = 1, \ldots, n, \tag{S.6.118}$$

$$\psi(t) := \rho'(t), \quad t \in \mathbb{R}. \tag{S.6.119}$$

We then have:

Lemma 8. *For any given $M > 0$, we have the following (in the sense of convergence in probability):*

$$\max_{y \in [-M, M] \cap \{0, \pm \frac{1}{M}, \pm \frac{2}{M}, \ldots\}} \max_{1 \le i \le n} |\hat{\mu}^y(X_i) - \hat{\mu}(X_i)| = o(1); \tag{S.6.120}$$

$$\max_{y \in [-M, M] \cap \{0, \pm \frac{1}{M}, \pm \frac{2}{M}, \ldots\}} \max_{1 \le i \le n} \left|\hat{\mu}_{/i}(X_i) + c_i \psi(\mathrm{prox}_{c_i \rho_i}(r_{i,(i)})) - \hat{\mu}(X_i)\right| = o(1). \tag{S.6.121}$$

Proof. The proof of (S.6.120) is almost identical to Lemma 7 (by viewing $\hat{\mu}$ as the estimator leaving out the $(n+1)$-th observation, and then taking the union bound over all y in the set $[-M, M] \cap \{0, \pm\frac{1}{M}, \pm\frac{2}{M}, \dots\}$), which relies on the error bound on the leave-one-out approximation formula (S.6.113). It still remains, however, to check whether the proof depends on how $y_{n+1} = y$ is selected. Below, we provide details of justification that (S.6.112) continues to hold if y_i is a fixed value in $\mathbb{R}$ (rather than randomly generated from the stochastic model), which is equivalent to (S.6.120) (by switching the role of $n+1$ and i). This, in turn, relies on two properties:

a) (S.6.113) continues to hold when y_i is an arbitrary deterministic quantity.

b) $X_{n+1}^\top \eta_i = o_{L_2}(1)$ continues to hold when y_i is an arbitrary deterministic quantity.

It is easy to see that b) is true, under our assumption of $\|\rho'\|_\infty < \infty$, since it is still true that $\frac{1}{n} X_{n+1}^\top (S_i + \tau I)^{-1} X_i$ has zero mean and $O(\frac{1}{n})$ second moment. For a), we revisit the original proof [Karoui, 2018, p122-p125], which is mostly based on several deterministic non-asymptotic bounds therein, except [Karoui, 2018, Lemma 3.7], which relies on convergence in high probability. However, [Karoui, 2018, Lemma 3.7] uses nothing but a high probability bound that $\sup_{j \neq i} \frac{\|(S_i + \tau I)^{-1} X_j\|}{\tau \sqrt{n}} = O(1)$, the independence of $\frac{\|(S_i + \tau I)^{-1} X_j\|}{\tau \sqrt{n}}$ and X_i, and sub-Gaussianity of X_i, which in turn implies by union bound that $\sup_{j \neq i} \frac{X_j^\top (S_i + \tau I)^{-1} X_i}{n} = \frac{1}{\sqrt{n}} \operatorname{polylog}(n)$ with high probability. Thus a) also holds, and we have verified (S.6.120).

Next, we observe that (S.6.121) because of (S.6.113) and (S.6.114), where again we used the fact that (S.6.113) continues to hold for arbitrary deterministic y_i. $\qquad\square$

Lemma 9. *For any given $M > 0$, we have the following (in the sense of convergence in probability):*

$$\max_{y \in [-M,M] \cap \{0, \pm\frac{1}{M}, \pm\frac{2}{M}, \dots\}} |\hat{\mu}(X_{n+1}) + c\psi(\operatorname{prox}_{c\rho}(y - \hat{\mu}(X_{n+1}))) - \hat{\mu}^y(X_{n+1})| = o(1), \qquad \text{(S.6.122)}$$

where we defined

$$c := \frac{1}{n+1} X_{n+1}^\top (S + \tau I)^{-1} X_{n+1}; \qquad \text{(S.6.123)}$$

$$S := \frac{1}{n+1} \sum_{j=1}^n \rho''(y - \hat{\mu}(X_{n+1})) X_j X_j^\top. \qquad \text{(S.6.124)}$$

Proof. The proof is identical to the proof of (S.6.121), again noting that the approximation error bound ((S.6.113), but leaving out $n+1$ instead of i) continues to hold with fixed deterministic $y_{n+1} = y$. $\qquad\square$

Proof of Theorem 6.6. Lemma 7 and Lemma 6 imply that $|U_J - U_{\text{mm}}| = o(1)$, and

$$|\hat{\mu}(X_{n+1}) + \hat{q}_{n,\alpha}^+ \{R_i^{LOO}\} - U_J| = o(1). \qquad \text{(S.6.125)}$$

Next, recall that [Karoui, 2018, Appendix 5] shows that $r_{i,(i)} = R_i^{LOO}$ behaves like (in the sense of moment convergence)

$$\epsilon_i + \sqrt{\mathbb{E}\left(\|\hat{\beta} - \beta_0\|^2\right)} Z_i,$$

where $Z_i \sim \mathcal{N}(0, 1)$ independent of ϵ_i; furthermore, if $i \neq j$, $r_{i,(i)}$ and $r_{j,(j)}$ are asymptotically (pairwise) independent. Since $\lim_{p \to \infty} \|\hat{\beta} - \beta_0\| = r$, this implies that the empirical distribution $(r_{i,(i)})_{i=1}^n$ converges to $\mathcal{N}(0, \sigma_\epsilon^2 + r^2)$. Thus

$$|\hat{q}_{n,\alpha}^+ \{R_i^{LOO}\} - \sqrt{\sigma_\epsilon^2 + r^2} z_{\alpha/2}| = o(1) \qquad \text{(S.6.126)}$$

and we have proved (6.56).

It remains to prove (6.57). Note that by definition, $y \in [-M, M] \cap \{0, \pm\frac{1}{M}, \pm\frac{2}{M}, \dots\}$ is included in the full conformal interval if

$$|y - \hat{\mu}^y(X_{n+1})| \le \hat{q}^+\{R_i\}. \tag{S.6.127}$$

Suppose that $y_{\max}$ and $y_{\min}$ denote the maximum and minimum y under this criterion. By Lemma 9, we see that $y_{\max}$ satisfies

$$y_{\max} - \hat{\mu}(X_{n+1}) - c\psi(\mathrm{prox}_{c\rho}(y_{\max} - \hat{\mu}(X_{n+1}))) = \hat{q}^+\{R_i\} + e_{n,M}, \tag{S.6.128}$$

where $e_{n,M}$ is an error term satisfying $\lim_{M\to+\infty} \lim_{n\to\infty} e_{n,M} = 0$. By (S.6.113), we have

$$R_i - R_i^{LOO} = -c_i\psi(R_i) + o(1), \tag{S.6.129}$$

implying $R_i = \mathrm{prox}_{c_i\rho}(R_i^{LOO} + o(1))$. Furthermore, using the Sherman–Morrison formula we can show that $c_i = c + o(1)$, so that $R_i = \mathrm{prox}_{c\rho}(R_i^{LOO} + o(1))$. Since $\mathrm{prox}_{c\rho}$ is an increasing function, we have

$$\hat{q}^+\{R_i\} = \hat{q}^+\{\mathrm{prox}_{c\rho}(R_i^{LOO} + o(1))\} \tag{S.6.130}$$

$$= \mathrm{prox}_{c\rho}(\hat{q}^+\{R_i^{LOO} + o(1)\}). \tag{S.6.131}$$

Then (S.6.128) shows that

$$\mathrm{prox}_{c\rho}(y_{\max} - \hat{\mu}(X_{n+1})) = \mathrm{prox}_{c\rho}(\hat{q}^+\{R_i^{LOO} + o(1)\}) + e_{n,M}. \tag{S.6.132}$$

Since we assumed $\|\rho''\|_\infty < \infty$, and $c = O(1)$ with high probability, $\mathrm{prox}_{c\rho}^{-1}$ must be Lipschitz with high probability. Hence

$$y_{\max} - \hat{\mu}(X_{n+1}) = \hat{q}^+\{R_i^{LOO} + o(1)\} + \tilde{e}_{n,M} \tag{S.6.133}$$

for some $\tilde{e}_{n,M}$ satisfying $\lim_{M\to\infty} \lim_{n\to\infty} \tilde{e}_{n,M} = 0$. Then (6.57) follows by (S.6.126). $\qquad\square$