

Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees

YUNYI ZHANG*,

Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA

*yuz334@ucsd.edu

AND

DIMITRIS N. POLITIS

Department of Mathematics and Halicioglu Data Science Institute, University of California, San Diego, La Jolla, CA 92093-0112, USA

dpolitis@ucsd.edu

[Received on 20 March 2022]

It can be argued that optimal prediction should take into account all available data. Therefore, to evaluate a prediction interval's performance one should employ conditional coverage probability, conditioning on all available observations. Focusing on a linear model, we derive the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction interval and the conditional coverage probability of a prediction interval obtained via a residual-based bootstrap. Applying this result, we show that a prediction interval generated by the residual-based bootstrap has approximately 50% probability to yield conditional under-coverage. We then develop a new bootstrap algorithm that generates a prediction interval that asymptotically controls both the conditional coverage probability as well as the possibility of conditional under-coverage. We complement the asymptotic results with several finite-sample simulations.

Keywords: Prediction, regression, bootstrap, conditional validity

2000 Math Subject Classification: 62J05, 62F40

1. Introduction

Statistical inference comes in two flavors: explaining the world and predicting the future state of the world. To explain the world based on data, statisticians create models like linear regression and use data to fit the models. After doing that, they will gauge the goodness-of-fit, and assess the accuracy of estimation, e.g., via confidence intervals of the fitted model. Focusing on regression, the literature is huge; to pick 3–4 papers, see Shao [35] on model selection, Xie and Huang [46] or Liu and Yu [22] on model fitting, and Freedman [12] on statistical analysis.

Prediction is not a new topic in statistical inference; we refer to Geisser [13] for a comprehensive introduction, or Politis [28] for a more recent exposition. Notably, prediction has seen a resurgence in the 21st century with the advent of statistical learning; see Hastie et al. [15] for an introduction. Similarly to the aforementioned linear model procedure, statisticians use data to fit a model that can yield a predictor for future observations, and use prediction intervals to quantify uncertainty in the prediction; see e.g. Romano et al. [31] and Wang and Politis [45]. Under a regression setting, there are

several ways to construct a prediction interval. The classical prediction interval was typically obtained under a Gaussian assumption on the errors; see Section 2 in that follows. One of the earliest methods foregoing the restrictive normality assumption employed the residual-based bootstrap; see Stine [38] and the references therein. More recent methods include the *Model-free* (MF) bootstrap and the hybrid *Model-free/Model-based* (MF/MB) bootstrap of Politis [28].

For all bootstrap methods, the aim is to provide an asymptotically valid prediction interval. Suppose Γ is a prediction interval for the future observation \mathcal{Y}_f . If $\text{Prob}(\mathcal{Y}_f \in \Gamma) \approx 1 - \alpha$ (where \approx indicates an asymptotic approximation), then Γ is an asymptotically valid $1 - \alpha$ prediction interval for \mathcal{Y}_f . On the other hand, if we wish to ensure that $\text{Prob}(\mathcal{Y}_f \in \Gamma) \geq 1 - \alpha$, i.e., an *unconditional lower-bound guarantee*, then we may apply the conformal prediction idea of Shafer and Vovk [34] and Vovk et al. [43], which has been applied to several complex models, including non-parametric regression; see Lei and Wasserman [20], Lei et al. [19], Romano et al. [31], and Sesia and Candès [33].

In the paper at hand, we assume a linear model and discuss how to construct an asymptotically valid prediction interval in the context of conditional coverage that also possesses some unconditional guarantees as discussed above. To be more concrete, suppose we have an $n \times p$ design matrix \mathcal{X} , independent and identically distributed residuals $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, dependent variables $\mathcal{Y} = (\mathcal{Y}_1, \dots, \mathcal{Y}_n)^T$ where $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$ and a fixed new regressor (a vector) \mathcal{X}_f that is of interest. We would like to provide a $1 - \alpha$ prediction interval $\Gamma = \Gamma(\mathcal{X}, \mathcal{Y}, \mathcal{X}_f)$ for the future observation $\mathcal{Y}_f = \mathcal{X}_f^T \beta + \xi$; here ξ is independent of \mathcal{X}, \mathcal{Y} and has the same distribution as ε_1 . The aforementioned bootstrap methods will ensure that $\text{Prob}(\mathcal{Y}_f \in \Gamma) \approx 1 - \alpha$, but without a lower-bound guarantee. On the other hand, the conformal prediction (e.g., Chernozhukov et al. [9]) method yields an interval Γ such that $\text{Prob}(\mathcal{Y}_f \in \Gamma) \geq 1 - \alpha$, i.e., an *unconditional lower-bound guarantee*. However, we are more interested in quantifying the performance of a prediction interval in terms of its conditional coverage probability $\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y})$ (or $\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}, \mathcal{X}_f, \mathcal{X})$ under random design).

The reason for our interest comes from two aspects. On one hand, the conditional probability precisely describes how statisticians make prediction in practice. By using the unconditional probability

$$\text{Prob}(\mathcal{Y}_f \in \Gamma) = E(\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y})) \quad (1.1)$$

it is as if we assume that the statistician has not observed \mathcal{Y} before making the prediction.

Realistically, however, statisticians have observed \mathcal{Y} and have fitted the model before they make predictions. Therefore, it is informative to understand what happens to \mathcal{Y}_f given our knowledge of all data (including \mathcal{Y}) rather than “on average” among all possible \mathcal{Y} .

On the other hand, according to eq. (1.1), analysis of the conditional probability is a more fundamental topic than the unconditional one. For example, if for any given $\delta > 0$, $\text{Prob}(\{|\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}) - (1 - \alpha)| > \delta\}) \rightarrow 0$ as $n \rightarrow \infty$, then we can take the conditional expectation and have

$$\begin{aligned} |\text{Prob}(\mathcal{Y}_f \in \Gamma) - (1 - \alpha)| &\leq E(|\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}) - (1 - \alpha)|) \\ &\leq \delta + \text{Prob}(\{|\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}) - (1 - \alpha)| > \delta\}) \end{aligned} \quad (1.2)$$

which implies $\text{Prob}(\mathcal{Y}_f \in \Gamma) \rightarrow 1 - \alpha$.

Consequently, the aforementioned performance goals of asymptotic validity and lower bound guarantee should be recast in terms of conditional coverage. Note, however, that $\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y})$ is a random variable itself – see e.g. definition 1.3 in Çinlar [7]. Hence, the performance goals are now stochastic, i.e., $\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}) \rightarrow_p 1 - \alpha$ and $\text{Prob}(\mathcal{Y}_f \in \Gamma | \mathcal{Y}) \geq 1 - \alpha$ with a specific probability. Surprisingly, we can achieve these goals simultaneously through a careful re-design of our prediction

intervals. Definition 1.2 in what follows describes our new performance aim. Before stating it, however, we need to clarify our notation since our results hold true for both fixed and random design. In the latter case, however, all probabilities and expectations will be understood as being conditional on \mathcal{X} ; see Definition 1.1 below.

DEFINITION 1.1 Consider the two cases:

- (a) **Fixed design**, i.e., there is no randomness involved in the design matrix \mathcal{X} and the new regressor \mathcal{X}_f . In this case, we define $\mathbf{P}(\cdot) = \text{Prob}(\cdot)$, $\mathbf{P}^*(\cdot) = \text{Prob}(\cdot|\mathcal{Y})$, $\mathbf{E}\cdot = E\cdot$, and $\mathbf{E}^*\cdot = E(\cdot|\mathcal{Y})$.
- (b) **Random design**, i.e., there is randomness involved in the design matrix \mathcal{X} (and possibly in the new regressor \mathcal{X}_f as well). In this case, we define $\mathbf{P}(\cdot) = \text{Prob}(\cdot|\mathcal{X}, \mathcal{X}_f)$, $\mathbf{P}^*(\cdot) = \text{Prob}(\cdot|\mathcal{Y}, \mathcal{X}, \mathcal{X}_f)$, $\mathbf{E}\cdot = E(\cdot|\mathcal{X}, \mathcal{X}_f)$, and $\mathbf{E}^*\cdot = E(\cdot|\mathcal{Y}, \mathcal{X}, \mathcal{X}_f)$. Furthermore, convergences and probability statements will be understood to hold almost surely in \mathcal{X} and \mathcal{X}_f .

We can now state our new performance aims in general.

DEFINITION 1.2 (Prediction interval with unconditional guarantee) Assume an $n \times p$ design matrix \mathcal{X} , independent and identically distributed (i.i.d.) residuals $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T \in \mathbf{R}^n$, and that the dependent variables \mathcal{Y} satisfy a linear model $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$. For a new regressor $\mathcal{X}_f \in \mathbf{R}^p$ and a potential future observation \mathcal{Y}_f , we say that $\Gamma = \Gamma(\mathcal{X}, \mathcal{Y}, \mathcal{X}_f)$ is the $1 - \alpha$ prediction interval with $1 - \gamma$ unconditional guarantee if the following conditions hold true:

1. For any given $\delta > 0$,

$$\mathbf{P}(\{|\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) - (1 - \alpha)| > \delta\}) \rightarrow 0 \quad (1.3)$$

- 2.

$$\mathbf{P}(\{\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) \geq 1 - \alpha\}) \rightarrow 1 - \gamma \quad (1.4)$$

as $n \rightarrow \infty$; here, α, γ are constants in $(0, 1)$. We call $1 - \alpha$ the nominal (conditional) coverage probability and $1 - \gamma$ the *guarantee level*.

Intuitively, Definition 1.2 requires the prediction interval Γ to have an asymptotically correct conditional coverage probability $1 - \alpha$. Meanwhile, the hope is that Γ 's conditional coverage probability is greater than $1 - \alpha$ with a specific (unconditional) probability.

REMARK 1.1 In Definition 1.2, the validity condition (eq.(1.3)) is ubiquitous and easily understood, but the second condition (eq.(1.4)) needs some clarifications. This remark aims to stress that the second condition is not redundant.

Suppose a prediction interval Γ satisfies (1.3) with $1 - \alpha = 95\%$. If the sample size n is very large, then Γ 's conditional coverage probability is close to 95%. In this situation, whether or not the conditional coverage probability is greater than 95% is not important. However, if the sample size is merely moderate, then Γ 's conditional coverage probability can be significantly smaller than 95%. Indeed, in table 2 and 3 (in section 6), a nominal 95% prediction interval may have a conditional coverage probability less than 91%.

In addition suppose Γ satisfies (1.4) with $1 - \gamma = 85\%$. When the sample size is moderate, the guarantee level may also be smaller than 85%. However, this condition still gives us an extra assurance that Γ is 'not likely' to have an under-coverage issue. Moreover, it is even unlikely for Γ 's conditional coverage probability to be far less than 95%.

Notably, statisticians have already noticed a gap between theoretical validity and finite sample performance. That is, an asymptotic valid prediction interval (e.g., Stine [38]) will often manifest under-coverage in practice; see Politis [27] for a discussion. In order to fill this gap, Politis [28] proposed

the definition of a ‘pertinent prediction interval’, which is a notion stronger than (1.3). Definition 1.2 provides a new perspective on this problem.

REMARK 1.2 (Further discussion on eq. (1.4)) A drawback of eq. (1.3) is that it takes place asymptotically (as the sample size $n \rightarrow \infty$). Hence, a prediction interval may satisfy (1.3) with a given $\delta > 0$, but for a given sample size n , the probability of the event $\{|\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) - (1 - \alpha)| > \delta\}$ may not be negligible. If the event $\{|\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) - (1 - \alpha)| > \delta\}$ is to happen, we may prefer $\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) > 1 - \alpha + \delta$ (i.e., overcoverage) to $\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) < 1 - \alpha - \delta$ (i.e., undercoverage). Eq.(1.4) reflects the intensity of this preference, i.e., overcoverage is more likely to happen if we choose large $1 - \gamma$. Notably, we require (1.3) and (1.4) to happen simultaneously. Therefore, (1.4) calibrates the usual prediction interval—e.g., the prediction interval generated by the residual-based bootstrap [38]—instead of creating a new one.

REMARK 1.3 This remark compares definition 1.2 with classical bootstrap methods and conformal predictions. Recall bootstrap methods always require $\mathbf{P}(\mathcal{Y}_f \in \Gamma) \rightarrow 1 - \alpha$ like Stine [38], or $\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) \rightarrow_p 1 - \alpha$ like Politis [28]. On the other hand, conformal prediction is considered a model-free, non-asymptotic method to generate a prediction interval. But its guarantee is **on average over the observations and over the future random regressor \mathcal{X}_f** . In table 1, it appears that the guarantee level of a conformal prediction is only 10.2% even when the sample size is 1600, implying that in 89.8% of the samples we have conditional coverage probability less than $1 - \alpha$. The new regressor \mathcal{X}_f is fixed (or conditioned upon) in our paper, so a complete model-free procedure (i.e., a procedure that constructs a consistent prediction interval for any models) is impossible; see Barber et al. [2].

In order to increase the guarantee level, Vovk [42] introduced the idea of a tolerance region; Vovk’s tolerance region is constructed as follows. First, perform the split-conformal prediction introduced in Lei et al. [19] to make the $1 - \alpha$ prediction interval $C_{1-\alpha}(\mathcal{X}_f)$ for \mathcal{Y}_f . Denote n_{calib} the size of the calibration set (i.e., I_2 in algorithm 2 of Lei et al. [19]). Then choose α' such that

$$\gamma \geq \text{binom}_{n_{calib}, \alpha}(\lfloor \alpha'(n_{calib} + 1) - 1 \rfloor) \quad (1.5)$$

where $\text{binom}_{n, \alpha}$ denotes the cumulative distribution function of a binominal distribution with size n and probability α , and $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to x . Then Vovk’s tolerance region is defined as $C_{1-\alpha'}(\mathcal{X}_f)$. According to proposition 2b in Vovk [42], this prediction interval satisfies

$$\mathbf{P}(\mathbf{P}^*(\mathcal{Y}_f \in C_{1-\alpha'}(\mathcal{X}_f)) \geq 1 - \alpha) \geq 1 - \gamma \quad (1.6)$$

which is similar to condition (1.4). However, Vovk’s tolerance region might not satisfy (1.3); that is why (1.5) is an inequality rather than an equality. In section 6, we compare several prediction methods via finite-sample simulations; it looks like Vovk’s tolerance region is typically wider than other prediction intervals.

Table 1 shows that this tolerance region has high guarantee levels among various linear models.

Definition 1.2 still follows a bootstrap framework but additionally requires $\mathbf{P}^*(\mathcal{Y}_f \in \Gamma) \geq 1 - \alpha$ for a specific proportion of observations. This definition is useful for understanding an existing bootstrap algorithm, like corollary 4.1. It also maintains the balance between Γ ’s length and its possibility of under-coverage.

Definition 1.2 is not easy to achieve; to see why, we present a simulation in table 1. The guarantee level (i.e., proportion of observations having conditional coverage probability $\geq 1 - \alpha$) of the aforementioned methods are not very high.

Our paper has two main contributions. On the one hand, it derives the Gaussian approximation for the difference between the conditional probability of a nominal prediction interval and the conditional

Table 1. Quantiles of conditional coverage probabilities and guarantee levels of prediction intervals on the **Experiment model** (see section 6). The errors are generated by i.i.d. normal random variables with mean 0 and variance 1. The nominal coverage probability is 95%. We use the R-package maintained by Tibshirani et al. [40] to perform conformal predictions. **For Vovk's tolerance region, we chose $\gamma = 15\%$ in (1.5).**

Sample size	Algorithm	Quantiles of coverage probabilities			Guarantee level
		15%	30%	50%	
100	Residual bootstrap	91.0%	92.5%	93.9%	31.3%
	MF/MB bootstrap	93.5%	94.7%	95.8%	66.7%
	Conformal prediction	90.0%	91.8%	93.5%	27.9%
	Split conformal prediction	95.2%	96.7%	97.8%	87.0%
	Jackknife conformal prediction	92.7%	95.3%	97.2%	56.5%
	Vovk's tolerance region	97.3%	98.4%	99.2%	95.5%
400	Residual bootstrap	93.3%	94.0%	94.7%	40.8%
	MF/MB bootstrap	93.8%	94.6%	95.2%	56.3%
	Conformal prediction	91.9%	92.7%	93.6%	15.5%
	Split conformal prediction	93.9%	94.8%	95.6%	66.4%
	Jackknife conformal prediction	93.8%	95.0%	96.2%	52.6%
	Vovk's tolerance region	96.1%	96.8%	97.5%	95.2%
1600	Residual bootstrap	94.0%	94.5%	95.0%	48.0%
	MF/MB bootstrap	94.2%	94.6%	95.0%	52.8%
	Conformal prediction	92.0%	92.7%	93.4%	10.2%
	Split conformal prediction	94.0%	94.6%	95.2%	57.7%
	Jackknife conformal prediction	93.0%	93.6%	94.3%	25.5%
	Vovk's tolerance region	94.8%	95.3%	95.9%	81.3%

probability of a prediction interval based on residual-based bootstrap. In practice, bootstrap approximates the former by the latter, and the non-zero difference will make the former deviate from $1 - \alpha$. This leads to the fact that the *residual-based bootstrap algorithm asymptotically has guarantee level of 50%*. On the other hand, we develop a new method to construct a prediction interval satisfying definition 1.2 with arbitrarily chosen α, γ .

We employ a simple example to illustrate why a classical prediction interval becomes problematic under the conditional coverage context in section 2. After that, we introduce the frequently used notations and assumptions in section 3. In section 4, we derive the Gaussian approximation result. In section 5, we develop the algorithm to construct the newly proposed prediction interval. We perform some simulations to illustrate the proposed algorithm's finite sample performance in section 6, and provide some conclusions in section 7. The proofs of the theoretical results will be deferred to the online supplement [49].

2. An intuitive illustration in the Gaussian case

For the sake of illustration, in this section only we suppose the residual ε_1 has a normal distribution with mean 0 and known variance σ^2 . Assume $\mathcal{X}^T \mathcal{X}$ is invertible. Denote $\Phi(x)$ as the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\alpha), \alpha \in (0, 1)$ as its α -quantile, i.e., $\Phi(\Phi^{-1}(\alpha)) = \alpha$. Adopt the notations \mathbf{P}, \mathbf{P}^* in definition 1.1. If we do not care about the conditional coverage, we can define $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}$ and use the normal distribution $1 - \alpha$ prediction interval $\mathcal{P}_1 = [\mathcal{X}_f^T \hat{\beta} + \sigma \Phi^{-1}(\frac{\alpha}{2}) \sqrt{1 + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}, \mathcal{X}_f^T \hat{\beta} + \sigma \Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{1 + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}]$ for the future response \mathcal{Y}_f . Since the random variable $\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}$ has normal distribution with mean 0

and variance $\sigma^2(1 + \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f)$, it follows that

$$\mathbf{P}(\mathcal{Y}_f \in \mathcal{P}_1) = \mathbf{P}\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}}{\sigma \sqrt{1 + \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha. \quad (2.1)$$

In other words, \mathcal{P}_1 has precise unconditional coverage probability. However, if we take the conditional coverage into consideration, the random variable $\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta} | \mathcal{Y}$ (or $\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta} | \mathcal{Y}, \mathcal{X}_f, \mathcal{X}$ under random design) has normal distribution with mean $\mathcal{X}_f^T \beta - \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}$ and variance σ^2 . According to Taylor's theorem,

$$\begin{aligned} \mathbf{P}^*(\mathcal{Y}_f \in \mathcal{P}_1) &= \mathbf{P}^*\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}}{\sigma \sqrt{1 + \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \\ &= \Phi\left(\sqrt{1 + \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f} \times \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \frac{\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right) \\ &\quad - \Phi\left(\sqrt{1 + \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f} \times \Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right) \\ &\approx 1 - \alpha + \Phi'(\Phi^{-1}(1 - \frac{\alpha}{2})) \times \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f \times \Phi^{-1}(1 - \frac{\alpha}{2}) \\ &\quad + \Phi''(\Phi^{-1}(1 - \frac{\alpha}{2})) \times \left(\frac{\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right)^2 \end{aligned} \quad (2.2)$$

The last line of (2.2) is derived by expanding the second line on $\Phi^{-1}(1 - \frac{\alpha}{2})$, and expanding the third line on $\Phi^{-1}(\frac{\alpha}{2})$. Since $\Phi''(\Phi^{-1}(1 - \frac{\alpha}{2})) < 0$, $\frac{(\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon)^2}{\sigma^2(\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f)}$ has χ_1^2 distribution and $\Phi''(x) = -x\Phi'(x)$ for any x ,

$$\mathbf{P}(\{\mathbf{P}^*(\mathcal{Y}_f \in \mathcal{P}_1) \geq 1 - \alpha\}) \approx \mathbf{P}\left(\frac{(\mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon)^2}{\sigma^2 \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f} \leq \frac{\Phi'(\Phi^{-1}(1 - \frac{\alpha}{2})) \times \Phi^{-1}(1 - \frac{\alpha}{2})}{-\Phi''(\Phi^{-1}(1 - \frac{\alpha}{2}))}\right) \quad (2.3)$$

which approximately equals 0.683. Therefore, the prediction interval \mathcal{P}_1 has about 68% guarantee level.

However, it is possible to find a prediction interval with a desired guarantee level, say $1 - \gamma$. Wallis [44], Lieberman and Miller [21] and De Gryze et al. [10] considered this problem and defined the ‘tolerance interval’ that controlled the guarantee level. However, their work assumed that the residuals ε_1 had normal distribution. Moreover, an $1 - \gamma$ tolerance interval does not ensure having asymptotic coverage probability $1 - \alpha$. We define $C_{1-\gamma}$ as the $1 - \gamma$ quantile of a χ_1^2 distribution, and let $c_{1-\gamma} = -\Phi''(\Phi^{-1}(1 - \frac{\alpha}{2})) \mathcal{X}_f^T(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f \times C_{1-\gamma} / (2\Phi'(\Phi^{-1}(1 - \frac{\alpha}{2}))) > 0$. We construct the prediction

interval $\mathcal{P}_2 = [\mathcal{X}_f^T \hat{\beta} + \sigma \times (\Phi^{-1}(\frac{\alpha}{2}) - c_{1-\gamma}), \mathcal{X}_f^T \hat{\beta} + \sigma \times (\Phi^{-1}(1 - \frac{\alpha}{2}) + c_{1-\gamma})]$. We can now compute

$$\begin{aligned} \mathbf{P}^*(\mathcal{Y}_f \in \mathcal{P}_2) &= \mathbf{P}^*\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - c_{1-\gamma} \leq \frac{\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}}{\sigma} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + c_{1-\gamma}\right) \\ &= \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + c_{1-\gamma} + \frac{\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right) - \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - c_{1-\gamma} + \frac{\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right) \\ &\approx 1 - \alpha + 2\Phi'\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \times c_{1-\gamma} + \Phi''\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \times \left(\frac{\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon}{\sigma}\right)^2 \\ &\quad \text{which implies that } \mathbf{P}(\{\mathbf{P}^*(\mathcal{Y}_f \in \mathcal{P}_2) \geq 1 - \alpha\}) \\ &\approx \mathbf{P}\left(-\Phi''\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \frac{(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon)^2}{\sigma^2 \times \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f} \leq -\Phi''\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \times C_{1-\gamma}\right) = 1 - \gamma \end{aligned} \quad (2.4)$$

Hence, prediction interval \mathcal{P}_2 has guarantee level about $1 - \gamma$. Note that since $c_{1-\gamma}$ has order $O(1/n)$, this correction does not significantly enlarge the width of the prediction interval. In other words, if the dimension of the parameter vector is fixed, then **the uncorrected and the corrected prediction intervals coincide with each other asymptotically**.

In the end of this section, we would like to briefly discuss the prediction problem under the high dimensional setting, i.e., $p/n \rightarrow s \in (0, 1)$. Bates et al. [3] and Dobriban and Wager [11] also considered this problem but they focused on estimating the prediction error. Steinberger and Leeb [37] and Zhang and Politis [48] constructed asymptotically valid prediction intervals for a (sparse) high dimensional linear model. Suppose $\exists 0 < c \leq C < \infty$ such that all eigenvalues of $\frac{1}{n} \mathcal{X}^T \mathcal{X}$ is greater than c and smaller than C . This assumption is achievable according to Bai and Yin [1]. If p is large and the new regressor \mathcal{X}_f is not sparse, then the term $\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f \geq \frac{\mathcal{X}_f^T \mathcal{X}_f}{Cn}$, which does not tend to 0 as $n \rightarrow \infty$. Therefore, despite that $\frac{\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}}{\sigma \sqrt{1 + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}}$ has normal distribution (so (2.1) is satisfied), we cannot use Taylor expansion in (2.2) and (2.4). So we need a new method to construct a prediction interval in order to satisfy Definition 1.2. Moreover, $c_{1-\gamma}$ will not converge to 0 as $n \rightarrow \infty$, and

$$\begin{aligned} &\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + c_{1-\gamma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{1 + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f} \\ &= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f \times \left(\frac{C_{1-\gamma}}{2} - \frac{1}{1 + \sqrt{1 + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_f}}\right) \end{aligned} \quad (2.5)$$

which does not converge to 0 as the sample size $n \rightarrow \infty$. So modification (2.4) will not be negligible asymptotically, and the prediction intervals (2.2) and (2.4) will not be close to each other even when n is large. In other words, constructing a ‘good’ prediction interval (e.g., a prediction interval satisfying definition 1.2) can be a challenging problem if the dimension of parameters is large. This paper will focus on the finite dimensional situation. However, our work should lay a good foundation for the high dimensional prediction problem.

Another limitation in this section is that the marginal distribution of the errors is assumed to be normal with known variance σ^2 , which is always not true. In the general situation, the marginal distribution of the errors is not normal and is unknown. As a consequence, the correction can be significantly larger

than $1/n$. Besides, we need to use resampling to find a satisfactory correction; this will be the subject of the following sections.

3. Preliminary notions

For the remainder of the paper, we revert to the general setup: an $n \times p$ design matrix \mathcal{X} (assumed to have full-rank), the dependent variable \mathcal{Y} satisfying the linear model $\mathcal{Y} = (\mathcal{Y}_1, \dots, \mathcal{Y}_n)^T = \mathcal{X}\beta + \varepsilon$ with respect to the i.i.d. errors $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$; here, ε_1 has mean zero, unknown variance σ^2 , and cumulative distribution function denoted by F . We denote $\mathcal{X}^T = (\mathcal{X}_1, \dots, \mathcal{X}_n)$, $\mathcal{X}_i = (\mathcal{X}_{i1}, \dots, \mathcal{X}_{ip})^T \in \mathbf{R}^p$, $i = 1, 2, \dots, n$, the new regressor $\mathcal{X}_f \in \mathbf{R}^p$ and the new dependent variable \mathcal{Y}_f (the subscript ‘ f ’ will only be used for future observations). Define

$$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y} \quad (3.1)$$

as the least squares estimator of the parameter vector β . Then, define the centered estimated residual $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T$ and the residual empirical process $\hat{F}(x)$ for any $x \in \mathbf{R}$ respectively as

$$\begin{aligned} \tilde{\varepsilon}_i &= \mathcal{Y}_i - \mathcal{X}_i^T \hat{\beta} = \varepsilon_i - \mathcal{X}_i^T (\hat{\beta} - \beta) \\ \hat{\varepsilon}_i &= \tilde{\varepsilon}_i - \frac{1}{n} \sum_{j=1}^n \tilde{\varepsilon}_j \\ \hat{F}(x) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\hat{\varepsilon}_i \leq x}. \end{aligned} \quad (3.2)$$

We also define $\overline{\mathcal{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathcal{X}_i \in \mathbf{R}^p$. From (3.2),

$$\int x d\hat{F} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i = 0, \quad \hat{\sigma}^2 = \int x^2 d\hat{F} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2. \quad (3.3)$$

Here and in the rest of this paper, the lower case letters x, y, z will be used to represent a scalar. For a function $g : \mathbf{R} \rightarrow \mathbf{R}$, define g' as its derivative. Denote $\mathbf{D} = \mathbf{D}[0, 1]$ the space of càdlàg functions on $[0, 1]$ with Skorohod topology—see chapter 3 of Billingsley [6].

To derive our results, we need the following assumptions.

Assumptions:

1. ε_1 's distribution is absolutely continuous with respect to Lebesgue measure. F is second order continuous differentiable and $\sup_{x \in \mathbf{R}} |F''(x)| < \infty$, $\mathbf{E}\varepsilon_1 = 0$, $\mathbf{E}|\varepsilon_1|^4 < \infty$. The new regressor $\mathcal{X}_f \in \mathbf{R}^p$ and the new dependent variable \mathcal{Y}_f satisfy $\mathcal{Y}_f = \mathcal{X}_f^T \beta + \xi$. ξ is independent of ε and has the same distribution as ε_1 .
2. One of the two following conditions holds true:
 - 2.1. **Fixed design:** \mathcal{X} and \mathcal{X}_f are fixed, i.e., non-random.
 - 2.2. **Random design:** \mathcal{X} and \mathcal{X}_f are random. However, \mathcal{X}_f is independent of ε, ξ ; and \mathcal{X} is independent of $\varepsilon, \xi, \mathcal{X}_f$.
3. $\mathcal{X}^T \mathcal{X}$ is invertible for $\forall n \geq p$ and $\lim_{n \rightarrow \infty} \frac{\mathcal{X}^T \mathcal{X}}{n} = A$, $\lim_{n \rightarrow \infty} \overline{\mathcal{X}}_n = b$; here A is an invertible matrix and $b \in \mathbf{R}^p$. Besides, there exists a constant $M > 0$ such that $\|\mathcal{X}_i\|_2 \leq M$ for $i = 1, 2, \dots, n$ and $\|\mathcal{X}_f\|_2 \leq M$. $\|\cdot\|_2$ denotes the Euclidean norm in \mathbf{R}^p .

4. Define $H(x) = \mathbf{E}\varepsilon_1 \mathbf{1}_{\varepsilon_1 \leq x}$ and for $\forall x, z \in \mathbf{R}$.

$$\begin{aligned} \mathcal{V}(x, z) &= \sigma^2 F'(x) F'(z) (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2 \mathcal{X}_f^T A^{-1} b) \\ &\quad - (F'(x) H(z) + F'(z) H(x)) (\mathcal{X}_f^T A^{-1} b - 1) + F(\min(x, z)) - F(x) F(z) \end{aligned} \quad (3.4)$$

We also define $\mathcal{U}(x) = \mathcal{V}(x, x) + \mathcal{V}(-x, -x) - 2\mathcal{V}(x, -x)$

Assume $F'(x) > 0, \forall x \in \mathbf{R}$, and $\mathcal{U}(x) > 0$ for $\forall 0 < x < \infty$.

For a function $g : \mathbf{R} \rightarrow \mathbf{R}$ and a point $x \in \mathbf{R}$, we define the limit from the left as

$$g^-(x) = \lim_{y \rightarrow x, y < x} g(y) \quad (3.5)$$

if this limit exists. Note that $g \in \mathbf{D}$ implies that $g^-(x)$ exists for $\forall x \in (0, 1)$. As in section 1.1.4 of Politis et al. [29], for any $0 < \alpha < 1$, we define the α quantile of a cumulative distribution function g as

$$c_\alpha = \inf\{x \in \mathbf{R} : g(x) \geq \alpha\}. \quad (3.6)$$

The meaning of notations $\mathbf{P}, \mathbf{P}^*, \mathbf{E}, \mathbf{E}^*$ is presented in definition 1.1. The symbol \rightarrow represents convergence in \mathbf{R} , and $\rightarrow_{\mathcal{L}}$ represents convergence in distribution. Without being specified, the convergence assumes the sample size $n \rightarrow \infty$. $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ respectively represents the cumulative distribution function and the quantile of the standard normal distribution. In the case of random design, the convergence results hold true for almost sure \mathcal{X} and \mathcal{X}_f .

REMARK 3.1 (a) We centered $\tilde{\varepsilon}_i$ in eq. (3.2), but if the design matrix \mathcal{X} has a column of ones, then summation of the estimated residuals will be 0 exactly, and re-centering is superfluous.

(b) In the case of random design, we assume assumption 3 and 4 happen for almost sure \mathcal{X} and \mathcal{X}_f .

(c) There are various linear model settings, e.g., presence of outliers, errors being dependent, errors being heteroskedastic, etc. This paper cannot discuss all situations simultaneously. So we focus on the classical setting, i.e., without outliers and errors are i.i.d., to present our work.

4. Gaussian approximation in bootstrap prediction

Residual-based bootstrap has been widely used in interval prediction for various models, such as Thombs and Schucany [39], and Li and Politis [26]. Stine [38] introduced a residual-based bootstrap algorithm for prediction, but this algorithm is typically characterized by finite sample undercoverage; see Li and Politis [25]. To alleviate the finite-sample undercoverage, Politis [28] proposed the *Model-free/Model-Based (MF/MB) bootstrap*, that resamples the *predictive* residuals $\hat{r} = (\hat{r}_1, \dots, \hat{r}_n)^T$ instead of the usual fitted residuals. The predictive residuals are sometimes called the ‘leave-one-out’ residuals, and are defined as:

$$\tilde{r}_i = \mathcal{Y}_i - \mathcal{X}_i^T (\mathcal{X}_{-i}^T \mathcal{X}_{-i})^{-1} \mathcal{X}_{-i}^T \mathcal{Y}_{-i}, \quad \hat{r}_i = \tilde{r}_i - \frac{1}{n} \sum_{j=1}^n \tilde{r}_j, \quad i = 1, 2, \dots, n \quad (4.1)$$

here \mathcal{X}_{-i} and \mathcal{Y}_{-i} are the design matrix \mathcal{X} and the dependent variable vector \mathcal{Y} respectively, having left out the i th row. For a least squares estimator, the predictive residuals can be efficiently computed using the hat matrix; see theorem 10.1 in Seber and Lee [32].

For concreteness, the algorithms are as follows:

Algorithm 4.1 (Residual-based bootstrap) **Input:** Design matrix \mathcal{X} and dependent variable data vector \mathcal{Y} satisfying $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$, the new regression vector \mathcal{X}_f of interest, number of bootstrap replicates B , nominal coverage probability $1 - \alpha$

1. Calculate statistics $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}$ and $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T$ as in eq. (3.2).
2. Generate i.i.d. residuals $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)^T$ and ξ^* by drawing from $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ with replacement. Then calculate $\mathcal{Y}^* = \mathcal{X}\hat{\beta} + \varepsilon^*$ and $\mathcal{Y}_f^* = \mathcal{X}_f^T \hat{\beta} + \xi^*$. Re-estimate $\hat{\beta}^* = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}^*$ and calculate the prediction root $\delta_b^* = \mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*$.
3. Repeat step 2 for $b = 1, 2, \dots, B$, and calculate the $1 - \alpha$ (unadjusted) sample quantile $\hat{c}_{1-\alpha}^*$ of $|\delta_b^*|$, $b = 1, 2, \dots, B$.
4. The prediction interval of \mathcal{Y}_f is given by $\left\{ \mathcal{Y}_f : |\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq \hat{c}_{1-\alpha}^* \right\}$

REMARK 4.1 If we replace $\hat{\varepsilon}$ by \hat{r} in algorithm 4.1, we then obtain the *MF/MB bootstrap* algorithm.

The Glivenko - Cantelli theorem ensures that the empirical process of the bootstrapped prediction root $\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*$ converges to $\mathbf{P}^* \left(\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^* \leq x \right)$ for any $x \in \mathbf{R}$ \mathbf{P}^* almost surely as $B \rightarrow \infty$. Therefore, the residual-based bootstrap approximates the unobservable conditional cumulative distribution function $\mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq x)$ by $\mathbf{P}^* \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \right)$, and estimates the latter distribution by the bootstrapped prediction root's empirical process; see Politis et al. [29].

Notably, the notation \mathbf{P}^* and \mathbf{E}^* are used for the conditional probability and expectation conditioning on all observed data in this paper. Note that this definition coincides with 'the probability and expectation in the bootstrap world' which is typical in the bootstrap literature; see e.g., Cheng and Huang [8]. The bootstrap approximation inevitably introduces errors. This section focuses on understanding the asymptotic behavior of the error process.

$$\mathcal{S}(x) = \sqrt{n} \left(\mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq x) - \mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x) \right) \quad (4.2)$$

here \mathcal{Y}_f^* and $\hat{\beta}^*$ are defined in algorithm 4.1. We refer to Bickel and Freedman [5] and Politis et al.[29] for the related researches.

The asymptotic behavior of \mathcal{S} is summarized in theorem 4.2.

THEOREM 4.2 Suppose assumption 1 to 4 hold true. Then for any given real numbers $0 < r < s < \infty$,

$$\sup_{x \in [r, s]} \sup_{y \in \mathbf{R}} \left| \mathbf{P}(\mathcal{S}(x) \leq y) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right) \right| \rightarrow 0 \quad (4.3)$$

here \mathcal{U} is defined in (3.4).

Hence, if a prediction interval Γ has the form $\left\{ y \in \mathbf{R} : |y - \mathcal{X}_f^T \hat{\beta}| \leq x \right\}$ (where x is a given positive number), then the conditional probability $\mathbf{P}^*(\mathcal{Y}_f \in \Gamma)$ and Γ 's coverage probability estimated by the residual-based bootstrap algorithm (i.e., $\mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x)$, where \mathcal{Y}_f^* and $\hat{\beta}^*$ are defined in algorithm 4.1) has an error. Moreover, $\sqrt{n} \times$ this error has an asymptotic normal distribution with mean 0 and a specific variance $\mathcal{U}(x)$ (depending on x).

In the conditional coverage context, an application of theorem 4.2 is to calculate a prediction interval's guarantee level. For example, by choosing $y = 0$, and $x = c_{1-\alpha}^*$ which denotes the $1 - \alpha$ quantile of the distribution $\mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x)$, we have the following corollary

COROLLARY 4.1 Under assumptions 1 to 4, the prediction interval generated by residual-based bootstrap has an asymptotically 50% guarantee level.

Alternatively, for a given $\gamma \in (0, 1)$, we could choose $y = \Phi^{-1}(\gamma)$, the γ quantile of the standard normal distribution, and $x = c_{1-\alpha-\Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)}/\sqrt{n}}^*$. Since \mathcal{U} is continuous, theorem 4.2 implies the event $\{\mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c_{1-\alpha-\Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)}/\sqrt{n}}^* - (1 - \alpha) \geq 0\}$, which is equivalent to the event

$$\sqrt{n} \left(\mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c_{1-\alpha-\Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)}/\sqrt{n}}^*) - (1 - \alpha - \frac{\Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)}}{\sqrt{n}}) \right) \geq \Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)} \quad (4.4)$$

asymptotically has unconditional probability $1 - \gamma$. In other words, the prediction interval $\{y \in \mathbf{R} : |y - \mathcal{X}_f^T \hat{\beta}| \leq c_{1-\alpha-\Phi^{-1}(\gamma) \times \sqrt{\mathcal{U}(c_{1-\alpha}^*)}/\sqrt{n}}^*\}$ has an asymptotic guarantee level $1 - \gamma$. Section 5 adopts this idea. However, in order to estimate \mathcal{U} , statisticians need to estimate $F(x) = \text{Prob}(\varepsilon_1 \leq x)$, the derivative $F'(x)$ and $H(x) = \mathbf{E}\varepsilon_1 \mathbf{1}_{\varepsilon_1 \leq x}$, which is complex. To make our work practical, section 5 presents a resampling algorithm that automatically generates the desired prediction interval without estimating \mathcal{U} .

5. Bootstrap prediction interval with unconditional guarantee

For a fixed dimensional linear model, bootstrap algorithms like the residual-based bootstrap and the *MF/MB bootstrap* generate asymptotically valid prediction intervals. Besides, Steinberger and Leeb [37] and Zhang and Politis [48] constructed asymptotically valid prediction intervals for high dimensional linear models. However, the statistician cannot adjust those prediction intervals' guarantee level; for example, corollary 4.1 says that the residual-based bootstrap has asymptotic guarantee level 50%. Therefore, in practice, the statistician cannot expect the possibility for a prediction interval to have a conditional coverage probability less than the nominal coverage probability. Ideally, we would wish for an algorithm that can generate an asymptotic valid prediction interval with a suitable guarantee level which is useful for both fixed and high dimensional regression. However, if the dimension is large, eq.(2.5) shows that the prediction intervals satisfying different purposes may not coincide with each other asymptotically. Therefore, finding a 'good' prediction interval can be a subtle problem for a high dimensional regression.

Focus on the fixed dimensional linear regression, this section proposes two new variations on these bootstrap methods, namely the *Residual bootstrap with unconditional guarantee (RBUG)* and the *Predictive residual bootstrap with unconditional guarantee (PRBUG)*, that maintain the asymptotic validity but also allows us to choose the prediction interval's guarantee level. These algorithms involve two steps: generating a valid prediction interval by residual-based bootstrap or *MF/MB bootstrap*; then calibrating the length of the prediction interval. Calibration of a confidence/prediction interval is not a new idea; see Loh [23, 24], Politis et al. [29] and Beran [4]. Their work calibrated a confidence interval based on the Edgeworth expansion. Our method does not use Edgeworth expansion. Instead, our method calibrates the prediction interval based on theorem 4.2 and the idea of eq. (4.4).

In order to use eq.(4.4), we need to estimate \mathcal{U} . In section A.2 of the supplement [49], we show that the error process $\mathcal{S}(x)$ (defined in (4.2)) can be approximated by a special stochastic process

$\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$, here

$$\tilde{M}_m(x) = \sqrt{n}F'(x_m) \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon - \frac{1}{n} \sum_{j=1}^n \varepsilon_j \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n (\mathbf{1}_{\varepsilon_j \leq x_m} - F(x_m)) \quad (5.1)$$

m is a sufficiently large positive integer and $x_m = 2mx - m$. As long as m is large, changing m does not affect the value of $\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$. Fortunately, simulating \tilde{M}_m in the bootstrap world is not difficult. So we can implicitly estimate \mathcal{W} by simulating \tilde{M}_m . Algorithm 5.1 adopts this idea, i.e., first estimate $c_{1-\alpha}^*$, the $1 - \alpha$ (unadjusted) quantile of the conditional distribution $\mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x)$.

Then estimate the coverage probability adjustment $-\frac{\Phi^{-1}(\gamma) \times \sqrt{\mathcal{W}(c_{1-\alpha}^*)}}{\sqrt{n}}$ in eq.(4.4) by simulating \tilde{M}_m . Finally, calibrate the prediction interval based on the adjustment.

Algorithm 5.1 (RBUG/PRBUG) Input: Design matrix \mathcal{X} and dependent variable data vector \mathcal{Y} satisfying $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$, the new regression vector \mathcal{X}_f of interest, and number of bootstrap replicates B , number of replicates to find quantile's adjustment \mathcal{B}_1 , nominal coverage probability $1 - \alpha$, and nominal guarantee level $1 - \gamma$

Note: For *RBUG*, we define $\hat{\tau} = (\hat{\tau}_1, \dots, \hat{\tau}_n)^T = \hat{\varepsilon}$ as in (3.2), while for *PRBUG*, we define $\hat{\tau} = \hat{r}$ as in (4.1).

Calculate an unadjusted sample quantile

1. Calculate the statistics $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}$ and $\hat{\tau}$.
2. Generate i.i.d. residuals $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)^T$ and ξ^* by drawing from $\hat{\tau}_1, \dots, \hat{\tau}_n$ with replacement; calculate $\mathcal{Y}^* = \mathcal{X} \hat{\beta} + \varepsilon^*$, $\mathcal{Y}_f^* = \mathcal{X}_f^T \hat{\beta} + \xi^*$ and $\hat{\beta}^* = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}^*$; derive the prediction root $\delta_b^* = \mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*$.
3. Repeat 2 for $b = 1, 2, \dots, B$, and calculate the $1 - \alpha$ unadjusted sample quantile (denoted as $\hat{c}_{1-\alpha}^*$) of $|\delta_b^*|$, $b = 1, 2, \dots, B$.

Find the quantile adjustment

4. Generate i.i.d. $e^* = (e_1^*, \dots, e_n^*)^T$ by drawing from $\hat{\tau}_1, \dots, \hat{\tau}_n$ with replacement, then derive $\mathcal{Y}^\dagger = \mathcal{X} \hat{\beta} + e^*$, $\hat{\beta}^\dagger = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y}^\dagger$. Then define $\hat{\zeta}_i^* = \mathcal{X}_f^T \hat{\beta} + \hat{\tau}_i - \mathcal{X}_f^T \hat{\beta}^\dagger + \frac{1}{n} \sum_{j=1}^n e_j^*$ for $i = 1, 2, \dots, n$. Calculate

$$p_{b_1}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{|\hat{\zeta}_i^*| \leq \hat{c}_{1-\alpha}^*} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{|e_i^*| \leq \hat{c}_{1-\alpha}^*} \quad (5.2)$$

5. Repeat step 4 for $b_1 = 1, 2, \dots, \mathcal{B}_1$, then calculate the $1 - \gamma$ sample quantile (denoted as $\hat{d}_{1-\gamma}^*$) of $p_{b_1}^*$, $b_1 = 1, 2, \dots, \mathcal{B}_1$.

Calibrate the prediction interval

6. Calculate $\hat{c}_{1-\alpha+\hat{d}_{1-\gamma}^*/\sqrt{n}}^*$, the $1 - \alpha + \hat{d}_{1-\gamma}^*/\sqrt{n}$ sample quantile of $|\delta_b^*|$, $b = 1, 2, \dots, B$
7. The prediction interval with $1 - \alpha$ coverage probability and $1 - \gamma$ guarantee level is given by the set

$$\left\{ x \in \mathbf{R} : |x - \mathcal{X}_f^T \hat{\beta}| \leq \hat{c}_{1-\alpha+\hat{d}_{1-\gamma}^*/\sqrt{n}}^* \right\}. \quad (5.3)$$

REMARK 5.1 This remark explains why step 4 and 5 in *RBUG/PRBUG* simulates \tilde{M}_m . Suppose we use

RBUG. Then

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{|\hat{\zeta}_i^*| \leq x} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{-x + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^* \leq \hat{\zeta}_i^* \leq x + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^*} \\ &= \sqrt{n} \left(\hat{F} \left(x + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^* \right) - \hat{F}^- \left(-x + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^* \right) \right) \end{aligned} \quad (5.4)$$

so $p_{b_1}^*$ equals

$$\begin{aligned} &\sqrt{n} \left(\hat{F} \left(\hat{c}_{1-\alpha}^* + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^* \right) - \hat{F}(\hat{c}_{1-\alpha}^*) \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\mathbf{1}_{e_j^* \leq \hat{c}_{1-\alpha}^*} - \hat{F}(\hat{c}_{1-\alpha}^*) \right) \\ &- \sqrt{n} \left(\hat{F}^- \left(-\hat{c}_{1-\alpha}^* + \mathcal{X}_f^T (\hat{\beta}^\dagger - \hat{\beta}) - \frac{1}{n} \sum_{j=1}^n e_j^* \right) - \hat{F}^-(-\hat{c}_{1-\alpha}^*) \right) + \frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\mathbf{1}_{e_j^* < -\hat{c}_{1-\alpha}^*} - \hat{F}^-(-\hat{c}_{1-\alpha}^*) \right) \end{aligned} \quad (5.5)$$

which simulates $\tilde{M}_m \left(\frac{x+m}{2m} \right) - \tilde{M}_m^- \left(\frac{-x+m}{2m} \right)$ in the bootstrap world. The same discussion applies to *PRBUG* as well.

We focus on proving *RBUG*'s validity, i.e., that prediction interval (5.3) satisfies definition 1.2. Define the simulated stochastic process

$$\begin{aligned} \widehat{\mathcal{M}}(x) &= \sqrt{n} \hat{F} \left(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{j=1}^n e_j^* \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathbf{1}_{e_j^* \leq x} \\ \text{and } \widehat{\mathcal{S}}(x) &= \widehat{\mathcal{M}}(x) - \widehat{\mathcal{M}}(-x) \end{aligned} \quad (5.6)$$

and the quantiles

$$\begin{aligned} c_{1-\alpha}^* &= \inf \left\{ x \in \mathbf{R} : \mathbf{P}^* \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \right) \geq 1 - \alpha \right\} \\ \text{and } d_{1-\gamma}^*(x) &= \inf \left\{ z \in \mathbf{R} : \mathbf{P}^* \left(\widehat{\mathcal{S}}(x) \leq z \right) \geq 1 - \gamma \right\} \end{aligned} \quad (5.7)$$

See algorithm 5.1 for the meaning of the notations. Denote

$$c^*(1 - \alpha, 1 - \gamma) = c_{1-\alpha+d_{1-\gamma}^*(c_{1-\alpha}^*)/\sqrt{n}}^* \quad (5.8)$$

From theorem 1.2.1 of Politis et al. [29], $\hat{c}_{1-\alpha+d_{1-\gamma}^*/\sqrt{n}}^*$ converges to $c^*(1 - \alpha, 1 - \gamma)$ almost surely as $B, \mathcal{B}_1 \rightarrow \infty$. Therefore, the theoretical justification only focuses on $c^*(1 - \alpha, 1 - \gamma)$.

THEOREM 5.2 Consider the *RBUG* algorithm, i.e., algorithm 5.1 with $\hat{\tau} = \hat{\varepsilon}$ as in (3.2). Suppose assumption 1 to 4 hold true. Then, for any given $0 < \alpha, \gamma < 1, \delta > 0$,

$$\begin{aligned} \mathbf{P} \left(\left| \mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c^*(1 - \alpha, 1 - \gamma) \right) - (1 - \alpha) \right| \leq \delta \right) &\rightarrow 1 \\ \mathbf{P} \left(\left\{ \mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c^*(1 - \alpha, 1 - \gamma) \right) \geq 1 - \alpha \right\} \right) &\rightarrow 1 - \gamma \end{aligned} \quad (5.9)$$

In other words, *RBUG* is able to generate a prediction interval with desired asymptotic coverage probability and guarantee level.

Corollary 5.1 proves the validity of *PRBUG*. In corollary 5.1, we choose $\hat{\tau} = \hat{r}$ in algorithm 5.1 and define $C_{1-\alpha}^* = \inf \left\{ x \in \mathbf{R} : \mathbf{P}^* (|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x) \geq 1 - \alpha \right\}$; $D_{1-\gamma}^*(x) = \inf \left\{ z \in \mathbf{R} : \mathbf{P}^* \left(\widehat{\mathcal{F}}(x) \leq z \right) \geq 1 - \gamma \right\}$. We define $C^*(1 - \alpha, 1 - \gamma) = C_{1-\alpha+D_{1-\gamma}^*(C_{1-\alpha}^*)/\sqrt{n}}^*$. That is, $C_{1-\alpha}^*$, $D_{1-\gamma}^*(x)$ and $C^*(1 - \alpha, 1 - \gamma)$ play the same roles as $c_{1-\alpha}^*$, $d_{1-\gamma}^*$ and $c^*(1 - \alpha, 1 - \gamma)$. The only reason for using another set of notations is that we change the sampling mechanism (i.e., replace $\hat{\varepsilon}$ in algorithm 5.1 by \hat{r}).

COROLLARY 5.1 Consider the *PRBUG* algorithm, i.e, algorithm 5.1 with $\hat{\tau} = \hat{r}$. Suppose assumptions 1 to 4 hold true. Then, for any given $0 < \alpha, \gamma < 1, \delta > 0$,

$$\begin{aligned} \mathbf{P} \left(\left| \mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq C^*(1 - \alpha, 1 - \gamma) \right) - (1 - \alpha) \right| \leq \delta \right) &\rightarrow 1 \\ \mathbf{P} \left(\left\{ \mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq C^*(1 - \alpha, 1 - \gamma) \right) \geq 1 - \alpha \right\} \right) &\rightarrow 1 - \gamma. \end{aligned} \quad (5.10)$$

REMARK 5.2 Similar to residual-based bootstrap and *MF / MB* bootstrap, section 6 shows that *PRBUG* tends to generate a wider, and of higher guarantee level, prediction interval than *RBUG*.

6. Numerical justification

This section applies numerical simulations to demonstrate the finite sample performance of *RBUG/PRBUG*. The alternatives are the residual-based bootstrap(RB), the *MF/MB* bootstrap(*MF/MB*), the split conformal prediction defined in Lei et al. [19] and Vovk's tolerance region [42]. The classical conformal prediction algorithm (e.g., Vovk et al. [43]) assumed \mathcal{X}_f is random, which is unsuitable for our setting. Vovk's tolerance region yields a prediction interval satisfying eq. (1.4) but not condition (1.3). Lei et al. [19] showed that the split conformal prediction could generate an asymptotic valid prediction interval when \mathcal{X}_f is fixed, which coincides with our setting.

Figure 1 plots point-wise prediction intervals for the linear model $\mathcal{Y}_i = 0.8 + 0.5\mathcal{X}_i + \varepsilon_i$. I.i.d. residuals are generated by normal distribution with mean 0 and variance 1. When the sample size is small, the prediction intervals generated by *RBUG / PRBUG* is significantly wider than the prediction intervals generated by classical bootstrap methods. On the other hand, when the sample size is large, the prediction intervals generated by different algorithms coincide with each other.

Our linear model of choice is denoted as the **Experiment model** and defined as follows: $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$, and β 's dimension is 8. $\beta = (\beta_0, \beta_1, \dots, \beta_7)^T$ with $\beta_0 = 1.0$, $\beta_1 = 0.5$, $\beta_2 = -1.0$, $\beta_3 = -0.5$, $\beta_4 = 1.5$, $\beta_5 = -1.5$ and $\beta_i = 0$ for $i \geq 6$. The design matrix \mathcal{X} is generated by i.i.d. standard normal random variables, and is fixed in each experiment. The new regressor $\mathcal{X}_f = (\mathcal{X}_{f,0}, \dots, \mathcal{X}_{f,7})^T$ is given by $\mathcal{X}_{f,i} = 0.1 \times i, i = 0, 1, \dots, 7$. The i.i.d. error vector ε is generated by various distributions. We choose the sample size $n = 50, 100, 200, 400, 1200$. The result is demonstrated in table 2, table 3, figure 2 and figure 3.

When the sample size is small (e.g. 50 or 100 in the example), the *MF/MB bootstrap* alleviates the residual-based bootstrap's under-coverage nature. Therefore, it has a higher guarantee level than the residual-based bootstrap. Yet this modification **does not** change the asymptotic guarantee level (in other words, the *MF/MB* bootstrap still has 50% asymptotic guarantee level). The split conformal prediction also has a high guarantee level when the sample size is small and a low guarantee level when the sample

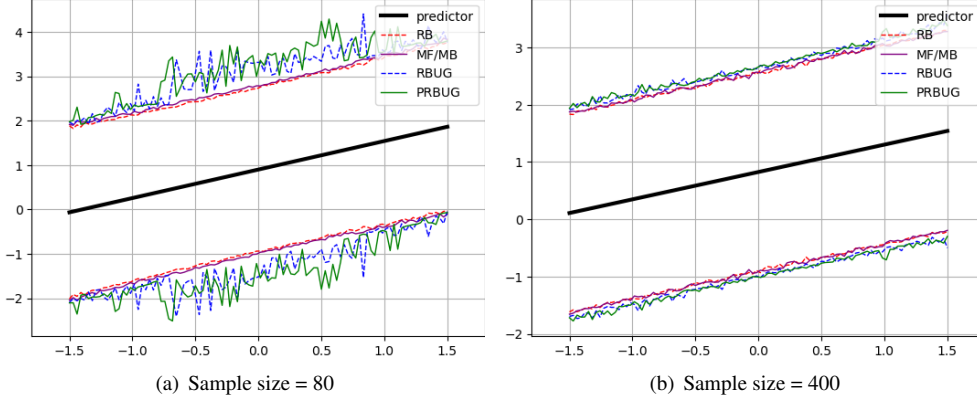


FIG. 1. Predictors and point-wise prediction intervals for the linear model $\mathcal{Y}_i = 0.8 + 0.5\mathcal{X}_i + \varepsilon_i$, $i = 1, 2, \dots$. The prediction intervals are generated by the following methods: RB for the residual-based bootstrap(Stein [38]); MF/MB for the model-free / model-based bootstrap(Politis [28]); RBUG and PRBUG for algorithm 5.1. We choose the nominal coverage probability $1 - \alpha = 95\%$ and the nominal guarantee level(in RBUG / PRBUG) $1 - \gamma = 90\%$.

size is moderate or large. **Vovk's tolerance region has the desired guarantee level when the sample size is large. However, when the sample size is small (e.g., 50 or 100), the tolerance region is always too wide.** On the other hand, the *RBUG* and the *PRBUG* algorithms improve the residual-based bootstrap's performance by controlling the asymptotic guarantee level. *PRBUG* reaches the desired guarantee levels when the sample size is moderate, while *RBUG* needs a large number of data in order to achieve the desired guarantee level. So we recommend using *PRBUG* in practice. When the sample size is large, the bootstrap algorithms' conditional coverage probabilities are close to 95%, and the adjustments made by *RBUG* / *PRBUG* are not significant.

In practice, our work can be particularly useful when the sample size n is not very large. Suppose we use the residual-based bootstrap. In table 2 we see that the 15% quantile of conditional coverage probabilities is 91.0% when the sample size is 100, which means 15% of the nominal 95% prediction intervals' conditional coverage probabilities are less than 91%. On the other hand, the *RBUG*'s 15% quantile is 93.5% and the *PRBUG*'s 15% quantile is 95.5%, which is significantly larger than the residual-based bootstrap's quantile.

7. Conclusion

Focusing on the fixed dimensional linear model, in this paper we derive the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction interval $\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq x \right)$ and the conditional coverage probability of a prediction interval for residual-based bootstrapped observations $\mathbf{P}^* \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \right)$. According to this result, the prediction interval generated by residual-based bootstrap has approximately 50% probability to yield conditional under-coverage.

We then develop a new bootstrap algorithm that generates prediction intervals with arbitrarily assigned conditional coverage probability and guarantee level, and prove its asymptotic validity. Our theoretical results are corroborated by several finite-sample simulations.

Table 2. Performance of different algorithms on the **Experiment model**. The nominal coverage probability is 95% and the nominal guarantee level is 85%(we also choose $\gamma = 15\%$ in (1.5)). The residuals are generated by normal random variables with mean 0 and variance 1. In the ‘Algorithm’ column, ‘RB’ means residual-based bootstrap; ‘*MF/MB*’ means *MF/MB* bootstrap; ‘split-conformal’ means the split conformal prediction(defined in Lei et al. [19]), **Vovk’s tolerance region was defined in remark 1.3**, and *RBUG* / *PRBUG* mean algorithm 5.1. We use the R package maintained by Tibshirani et al. [40] to perform the split conformal prediction. ‘Length’ represents the average length of the prediction interval. The number of bootstrap replicates is $B = 3000$, the number of replicates to find quantile’s adjustment is $\mathcal{B}_1 = 3000$. The result is generated by 1500 simulations.

Sample size	Algorithm	Quantiles of coverage probabilities			Guarantee level	Length
		15%	30%	50%		
50	RB	87.8%	90.2%	92.4%	21.1%	3.63
	<i>MF / MB</i>	93.8%	95.5%	96.9%	75.6%	4.40
	split-conformal	95.9%	97.8%	99.0%	89.2%	5.65
	Vovk’s tolerance region	95.9%	98.0%	99.1%	89.2%	5.69
	<i>RBUG</i>	91.2%	93.7%	95.7%	57.9%	4.19
	<i>PRBUG</i>	95.8%	97.3%	98.5%	90.3%	5.02
100	RB	91.0%	92.6%	93.9%	29.0%	3.78
	<i>MF / MB</i>	93.7%	94.9%	95.9%	69.3%	4.14
	split-conformal	95.1%	96.7%	98.0%	86.0%	4.78
	Vovk’s tolerance region	97.3%	98.5%	99.2%	96.5%	5.55
	<i>RBUG</i>	93.5%	94.9%	96.1%	68.1%	4.22
	<i>PRBUG</i>	95.5%	96.6%	97.6%	89.1%	4.58
200	RB	92.5%	93.4%	94.3%	34.7%	3.83
	<i>MF / MB</i>	93.7%	94.5%	95.3%	58.9%	4.00
	split-conformal	93.6%	94.9%	96.0%	69.8%	4.19
	Vovk’s tolerance region	95.9%	97.0%	97.9%	92.7%	4.69
	<i>RBUG</i>	94.2%	95.0%	95.8%	71.0%	4.12
	<i>PRBUG</i>	95.1%	95.9%	96.7%	87.5%	4.29
400	RB	93.5%	94.1%	94.7%	41.3%	3.88
	<i>MF / MB</i>	94.0%	94.6%	95.2%	58.3%	3.96
	split-conformal	93.7%	94.7%	95.5%	65.2%	4.05
	Vovk’s tolerance region	96.1%	96.8%	97.5%	95.5%	4.50
	<i>RBUG</i>	94.6%	95.3%	95.9%	75.5%	4.08
	<i>PRBUG</i>	95.1%	95.7%	96.2%	87.5%	4.16
1200	RB	94.0%	94.5%	94.9%	47.9%	3.91
	<i>MF / MB</i>	94.2%	94.7%	95.1%	55.8%	3.94
	split-conformal	94.1%	94.6%	95.2%	60.5%	3.97
	Vovk’s tolerance region	95.1%	95.6%	96.2%	88.1%	4.15
	<i>RBUG</i>	94.7%	95.1%	95.6%	76.7%	4.03
	<i>PRBUG</i>	94.9%	95.3%	95.7%	81.0%	4.05

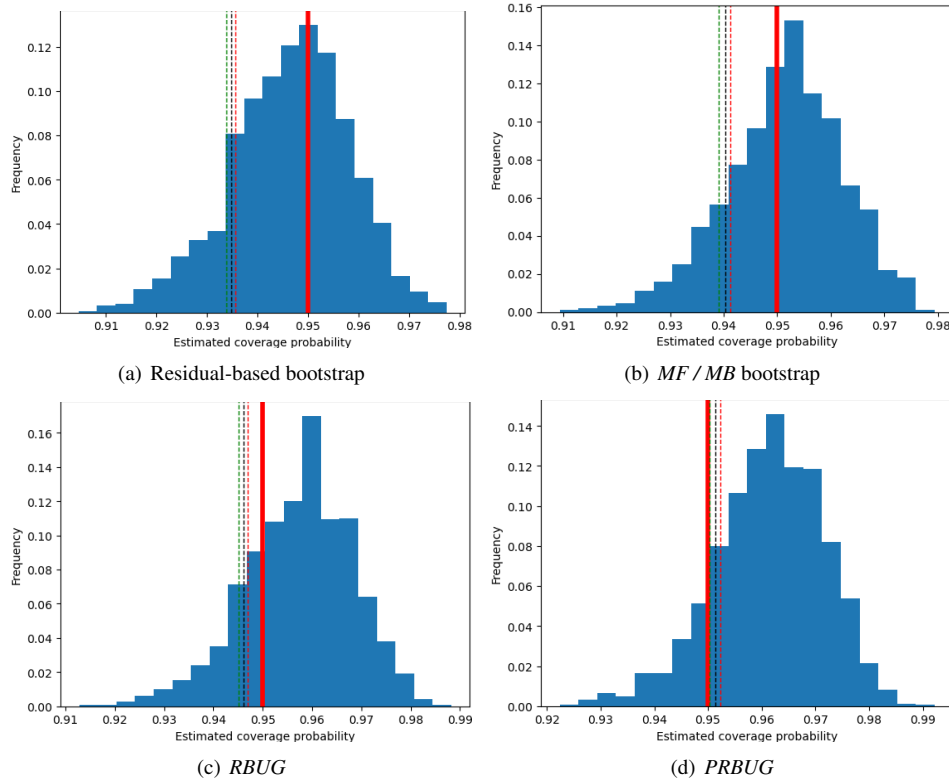


FIG. 2. Histograms for the conditional coverage probabilities. Here we use the **Experiment model** with sample size 400. The residuals are generated by i.i.d. normal random variables with mean 0 and variance 1. The solid red line is the nominal coverage probability (95%); the green, black and red dashed lines respectively represent the 13%, 15%, 17% quantile of conditional coverage probabilities. In order to have a $1 - \gamma = 85\%$ guarantee level, the solid red line should be close to the black dashed line.

Table 3. Performance of different algorithms on the **Experiment model**. The nominal coverage probability is 95%, and the nominal guarantee level is 85%. The residuals are generated by the Laplace distribution with mean 0 and scale $1/\sqrt{2}$; which makes the residuals' variance 1.

Sample size	Algorithm	Quantiles of coverage probabilities			Guarantee level	Length
		15%	30%	50%		
50	RB	87.7%	90.4%	92.4%	23.3%	3.83
	<i>MF / MB</i>	92.2%	94.1%	95.5%	58.7%	4.60
	split-conformal	94.4%	96.6%	98.1%	82.1%	6.14
	Vovk's tolerance region	94.6%	96.6%	98.2%	82.7%	6.15
	<i>RBUG</i>	91.1%	93.6%	95.7%	57.3%	4.75
	<i>PRBUG</i>	94.5%	96.1%	97.6%	81.4%	5.67
100	RB	91.0%	92.6%	93.9%	33.9%	4.03
	<i>MF / MB</i>	92.9%	94.2%	95.3%	57.1%	4.41
	split-conformal	94.5%	96.1%	97.3%	81.5%	5.30
	Vovk's tolerance region	96.8%	98.0%	98.9%	94.6%	6.68
	<i>RBUG</i>	93.6%	95.1%	96.4%	72.1%	4.80
	<i>PRBUG</i>	94.9%	96.2%	97.3%	84.7%	5.21
200	RB	92.6%	93.6%	94.5%	37.7%	4.13
	<i>MF / MB</i>	93.4%	94.4%	95.2%	54.7%	4.32
	split-conformal	93.3%	94.6%	95.6%	63.9%	4.54
	Vovk's tolerance region	95.6%	96.7%	97.6%	91.0%	5.37
	<i>RBUG</i>	94.5%	95.3%	96.1%	76.7%	4.64
	<i>PRBUG</i>	95.1%	95.8%	96.6%	86.0%	4.83
400	RB	93.4%	94.1%	94.8%	42.6%	4.18
	<i>MF / MB</i>	93.8%	94.5%	95.1%	53.7%	4.28
	split-conformal	93.5%	94.5%	95.4%	60.4%	4.40
	Vovk's tolerance region	95.9%	96.6%	97.3%	94.5%	5.16
	<i>RBUG</i>	94.7%	95.3%	95.9%	78.6%	4.53
	<i>PRBUG</i>	95.0%	95.6%	96.2%	84.5%	4.63
1200	RB	94.1%	94.5%	94.9%	47.9%	4.23
	<i>MF / MB</i>	94.2%	94.6%	95.0%	52.5%	4.26
	split-conformal	93.9%	94.5%	95.2%	57.1%	4.28
	Vovk's tolerance region	95.1%	95.6%	96.1%	86.4%	4.61
	<i>RBUG</i>	94.8%	95.2%	95.6%	77.6%	4.42
	<i>PRBUG</i>	94.9%	95.3%	95.7%	81.8%	4.45

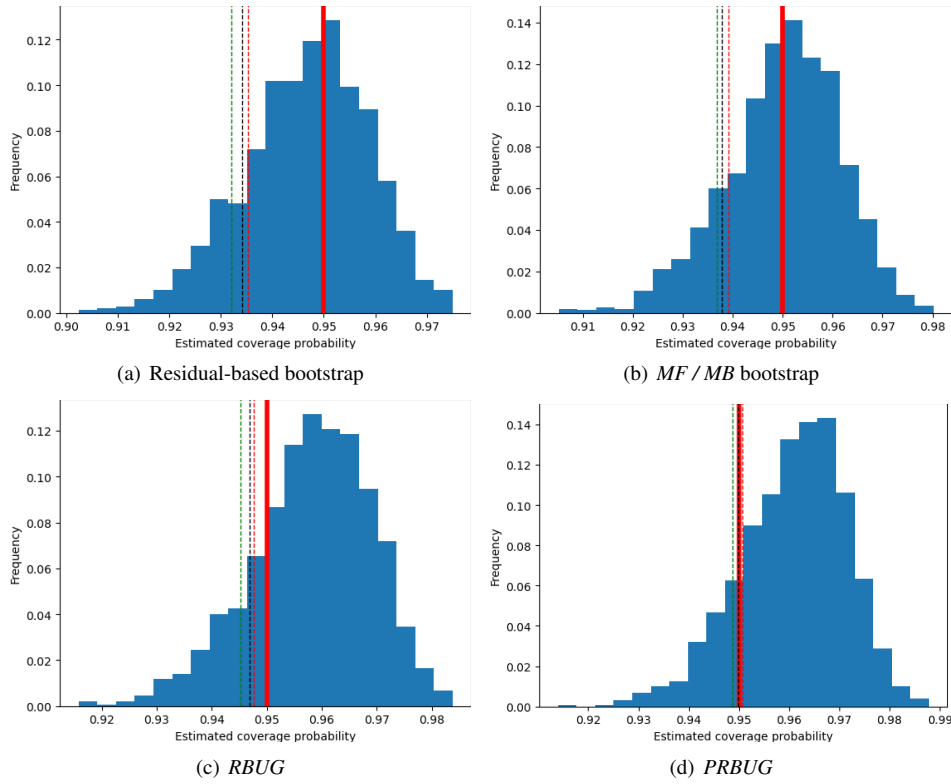


FIG. 3. Histograms for the conditional coverage probabilities of the **Experiment model**. The sample size is 400 and the residuals are generated by i.i.d. Laplace random variables with mean 0 and the scale parameter $1/\sqrt{2}$, which makes the variance 1. The meaning of lines coincide with figure 2.

Residual-based and the MF/MB bootstrap are widely used for prediction in numerous settings like nonparametric/nonlinear regression, quantile regression, time series analysis (regression with dependent errors, autoregression, etc.), and others. We expect our ideas to be applicable in those settings as well; future work will address the details. Furthermore, the case of high-dimensional linear regression is of current interest, i.e., where p is allowed to diverge as $n \rightarrow \infty$; this can also be the subject of future work.

Acknowledgment

We are grateful for valuable suggestions from Dr. Xiaou Pan, Dr. Yiren Wang, the editor and two anonymous reviewers that helped improve the quality of this paper. **Special thanks are due to one of the reviewers who introduced us to Vovk's tolerance region; see remark 1.3.** This research was partially supported by NSF Grant DMS 19-14556.

References

- [1] BAI, Z. D. & YIN, Y. Q. (1993) Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, **21**(3), 1275–1294.
- [2] BARBER, R. F., CANDÈS, E. J., RAMDAS, A. & TIBSHIRANI, R. J. (2021) The limits of distribution-free conditional predictive inference. *Inf. Inference*, **10**(2), 455–482.
- [3] BATES, S., HASTIE, T. & TIBSHIRANI, R. (2021) Cross-validation: what does it estimate and how well does it do it?. *arXiv e-prints*, p. arXiv:2104.00673.
- [4] BERAN, R. (1990) Calibrating prediction regions. *J. Amer. Statist. Assoc.*, **85**(411), 715–723.
- [5] BICKEL, P. J. & FREEDMAN, D. A. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**(6), 1196–1217.
- [6] BILLINGSLEY, P. (1999) *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edn., A Wiley-Interscience Publication.
- [7] ÇINLAR, E. (2011) *Probability and Stochastics*. Springer-Verlag New York, 1 edn.
- [8] CHENG, G. & HUANG, J. Z. (2010) Bootstrap consistency for general semiparametric M -estimation. *Ann. Statist.*, **38**(5), 2884–2915.
- [9] CHERNOZHUKOV, V., WÜTHRICH, K. & ZHU, Y. (2021) Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, **118**(48).
- [10] DE GRYZE, S., LANGHANS, I. & VANDEBROEK, M. (2007) Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, **87**(2), 147–154.
- [11] DOBRIBAN, E. & WAGER, S. (2018) High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.*, **46**(1), 247–279.
- [12] FREEDMAN, D. A. (1981) Bootstrapping regression models. *Ann. Statist.*, **9**(6), 1218–1228.

- [13] GEISSER, S. (1993) *Predictive inference*, vol. 55 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.
- [14] HAHN, M. G. (1977) Conditions for sample-continuity and the central limit theorem. *Ann. Probability*, **5**(3), 351–360.
- [15] HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009) *The Elements of Statistical Learning*. Springer-Verlag New York, 2 edn.
- [16] HORN, R. A. & JOHNSON, C. R. (2013) *Matrix analysis*. Cambridge University Press, Cambridge, second edn.
- [17] JACOD, J. & SHIRYAEV, A. N. (2003) *Limit theorems for stochastic processes*, vol. 288 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edn.
- [18] KOUL, H. L. (2002) *Weighted empirical processes in dynamic nonlinear models*, vol. 166 of *Lecture Notes in Statistics*. Springer-Verlag, New York, Second edition of it Weighted empiricals and linear models [Inst. Math. Statist., Hayward, CA, 1992; MR1218395 (95c:62061)].
- [19] LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. & WASSERMAN, L. (2018) Distribution-Free Predictive Inference for Regression. *J. Amer. Statist. Assoc.*, **113**(523), 1094–1111.
- [20] LEI, J. & WASSERMAN, L. (2014) Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **76**(1), 71–96.
- [21] LIEBERMAN, G. J. & MILLER, JR., R. G. (1963) Simultaneous tolerance intervals in regression. *Biometrika*, **50**, 155–168.
- [22] LIU, H. & YU, B. (2013) Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.*, **7**, 3124–3169.
- [23] LOH, W.-Y. (1987) Calibrating confidence coefficients. *J. Amer. Statist. Assoc.*, **82**(397), 155–162.
- [24] ——— (1991) Bootstrap calibration for confidence interval construction and selection. *Statist. Sinica*, **1**(2), 477–491.
- [25] PAN, L. & POLITIS, D. N. (2016a) Bootstrap prediction intervals for linear, nonlinear and non-parametric autoregressions. *J. Statist. Plann. Inference*, **177**, 1–27.
- [26] ——— (2016b) Bootstrap prediction intervals for Markov processes. *Comput. Statist. Data Anal.*, **100**, 467–494.
- [27] POLITIS, D. N. (2013) Model-free model-fitting and predictive distributions. *TEST*, **22**(2), 183–221.
- [28] ——— (2015) *Model-free prediction and regression*, Frontiers in Probability and the Statistical Sciences. Springer, Cham, A transformation-based approach to inference.
- [29] POLITIS, D. N., ROMANO, J. P. & WOLF, M. (1999) *Subsampling*, Springer Series in Statistics. Springer-Verlag, New York.

- [30] RANGA RAO, R. (1962) Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, **33**, 659–680.
- [31] ROMANO, Y., PATTERSON, E. & CANDES, E. (2019) Conformalized Quantile Regression. in *Advances in Neural Information Processing Systems*, vol. 32, pp. 3543–3553. Curran Associates, Inc.
- [32] SEBER, G. A. F. & LEE, A. J. (2003) *Linear regression analysis*, Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edn.
- [33] SESIA, M. & CANDÈS, E. J. (2020) A comparison of some conformal quantile regression methods. *Stat*, **9**(1), e261, e261 sta4.261.
- [34] SHAFER, G. & VOVK, V. (2008) A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, **9**, 371–421.
- [35] SHAO, J. (1996) Bootstrap model selection. *J. Amer. Statist. Assoc.*, **91**(434), 655–665.
- [36] ——— (2003) *Mathematical statistics*, Springer Texts in Statistics. Springer-Verlag, New York, second edn.
- [37] STEINBERGER, L. & LEEB, H. (2016) Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*.
- [38] STINE, R. A. (1985) Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.*, **80**(392), 1026–1031.
- [39] THOMBS, L. A. & SCHUCANY, W. R. (1990) Bootstrap Prediction Intervals for Autoregression. *J. Amer. Statist. Assoc.*, **85**(410), 486–492.
- [40] TIBSHIRANI, R., JELIAZKOVA, N. & LEDELL, E. (2021) Conformal Inference R Project. <https://github.com/ryantibs/conformal>.
- [41] VILLANI, C. (2009) *Optimal transport*, vol. 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.
- [42] VOVK, V. (2012) Conditional Validity of Inductive Conformal Predictors. in *Proceedings of the Asian Conference on Machine Learning*, ed. by S. C. H. Hoi, & W. Buntine, vol. 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore. PMLR.
- [43] VOVK, V., GAMMERMAN, A. & SHAFER, G. (2005) *Algorithmic learning in a random world*. Springer, New York.
- [44] WALLIS, W. A. (1951) Tolerance intervals for linear regression. in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pp. 43–51. University of California Press, Berkeley and Los Angeles, Calif.
- [45] WANG, Y. & POLITIS, D. N. (2021) Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals. *arXiv preprint arXiv:2109.12156*.

- [46] XIE, H. & HUANG, J. (2009) SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.*, **37**(2), 673–696.
- [47] XU, M., ZHANG, D. & WU, W. B. (2019) Pearson’s chi-squared statistics: approximation theory and beyond. *Biometrika*, **106**(3), 716–723.
- [48] ZHANG, Y. & POLITIS, D. N. (2020) Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap. *arXiv preprint arXiv:2009.08071*, to appear in *Ann.Statist.*
- [49] ZHANG, Y. & POLITIS, D. N. (2021) Supplement to ‘Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees’. .

A. Proof of theorem 4.2 in section 4

Adopt the notations in section 3 and 4 of the paper. Suppose assumption 1 to 4 in section 3 hold true. For any given positive integer $0 < m < \infty$, define $x_m = 2mx - m$ and the stochastic process

$$\tilde{M}_m(x) = \sqrt{n}F'(x_m) \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon} - \frac{1}{n} \sum_{j=1}^n \boldsymbol{\varepsilon}_j \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n (\mathbf{1}_{\varepsilon_j \leq x_m} - F(x_m)) \quad (\text{A.1})$$

Then $\mathbf{E}\tilde{M}_m(x) = 0$ for any given x .

Define the Gaussian process $\mathcal{M}_m(x) \in \mathbf{D}, x \in [0, 1]$ such that

$$\mathbf{E}\mathcal{M}_m(x) = 0, \mathbf{E}\mathcal{M}_m(x)\mathcal{M}_m(z) = \mathcal{V}(x_m, z_m) \text{ for } \forall x, z \in [0, 1] \quad (\text{A.2})$$

here $z_m = 2mz - m$ and \mathcal{M}_m has continuous sample paths almost surely. \mathcal{V} is defined in (3.4) of the paper. The proof of theorem 4.2 has 4 steps:

1. Show the existence of \mathcal{M}_m for any m
2. Prove that $\tilde{M}_m \rightarrow \mathcal{L} \mathcal{M}_m$ (i.e., converges in distribution) under the Skohord topology. Then lemma A.1 below implies that $\tilde{M}_m(x)$'s sample paths will be similar to a continuous function, i.e., $|\tilde{M}_m(x) - \tilde{M}_m(z)|$ can be arbitrarily small as $|x - z| \rightarrow 0$ with probability tending to 1.
3. Prove that the random variable $\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$'s asymptotic distribution will be a normal distribution with mean 0 and variance $\mathcal{U}(x)$ for any $x \in (0, m]$. See (3.5) in the paper for the definition of the superscript $-$.
4. Approximate $\mathcal{S}(x)$ (see (4.2)) by $\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$.

But before presenting the proof, we would like to introduce some useful lemmas.

A.1 Useful lemmas

Suppose random variables A, B satisfy $|A - B| \leq \delta, \delta > 0$. Then $\forall x \in \mathbf{R}, -\mathbf{1}_{x-\delta < B \leq x} \leq \mathbf{1}_{A \leq x} - \mathbf{1}_{B \leq x} \leq \mathbf{1}_{x < B \leq x+\delta}$, which implies

$$\begin{aligned} \mathbf{E}|\mathbf{1}_{A \leq x} - \mathbf{1}_{B \leq x}| &\leq \mathbf{E}|\mathbf{1}_{A \leq x} - \mathbf{1}_{B \leq x}| \times \mathbf{1}_{|A-B| \leq \delta} + \text{Prob}(|A-B| > \delta) \\ &\leq \text{Prob}(|A-B| > \delta) + \text{Prob}(x - \delta < B \leq x + \delta) \end{aligned} \quad (\text{A.3})$$

For any given positive integer r and $\forall t_i \in [0, 1], s_i \in \mathbf{R}, i = 1, 2, \dots, r$, define $t_{i,m} = 2mt_i - m$, then

$$\begin{aligned} &0 \leq \lim_{n \rightarrow \infty} \mathbf{E} \left(\sum_{i=1}^r s_i \tilde{M}_m(t_i) \right)^2 \\ &= \lim_{n \rightarrow \infty} \sigma^2 \sum_{i=1}^r \sum_{j=1}^r s_i s_j F'(t_{i,m}) F'(t_{j,m}) \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 - 2 \mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \overline{\mathcal{X}}_n \right) \\ &\quad - \lim_{n \rightarrow \infty} \sum_{i=1}^r \sum_{j=1}^r s_i s_j \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \overline{\mathcal{X}}_n - 1 \right) \times (F'(t_{i,m})H(t_{j,m}) + F'(t_{j,m})H(t_{i,m})) \\ &\quad + \lim_{n \rightarrow \infty} \sum_{i=1}^r \sum_{j=1}^r s_i s_j (F(\min(t_{i,m}, t_{j,m})) - F(t_{i,m})F(t_{j,m})) = \sum_{i=1}^r \sum_{j=1}^r s_i s_j \mathcal{V}(t_{i,m}, t_{j,m}) \end{aligned} \quad (\text{A.4})$$

Eq.(A.4) implies that $\mathcal{V}(2m \cdot -m, 2m \cdot -m)$ will be the asymptotic covariance function of the stochastic process $\tilde{M}_m(\cdot)$. Moreover, for any real number sequence $\{z_i\}_{i=1,\dots,r}$, the matrix $\{\mathcal{V}(z_i, z_j)\}_{i,j=1,\dots,r}$ is positive semi-definite. From assumption 3, define $\hat{\sigma}^2$ as in section (3.3) of the paper

$$\begin{aligned} \mathbf{E}\hat{\sigma}^2 &\leq \frac{2}{n} \sum_{i=1}^n \mathbf{E} \left(\varepsilon_i - \frac{\sum_{j=1}^n \varepsilon_j}{n} \right)^2 + \frac{2}{n} \sum_{i=1}^n \mathbf{E}((\mathcal{X}_i - \overline{\mathcal{X}}_n)^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon)^2 \\ &\leq 2\sigma^2 + \frac{8M^2\sigma^2}{n} \left\| \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \right\|_2 \end{aligned} \quad (\text{A.5})$$

so $\mathbf{E}\hat{\sigma}^2 = O(1)$. Here $\left\| \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \right\|_2$ is the matrix 2-norm of $\left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1}$. (A.3), (A.4) and (A.5) will be frequently used in the following sections. Then we introduce some lemmas. Lemma A.1 focuses on showing the existence of \mathcal{M}_m and deriving its properties.

LEMMA A.1 Suppose assumption 1 to 4 hold true.

1. For $\forall 0 < m \in \mathbf{N}$, \exists a Gaussian process \mathcal{M}_m in \mathbf{D} satisfying (A.2) and having continuous sample paths almost surely.
2. For any given $\xi > 0$,

$$\lim_{\delta \rightarrow 0, \xi > 0} \mathbf{P} \left(\sup_{y, z \in [0, 1], |y-z| < \delta} |\mathcal{M}_m(y) - \mathcal{M}_m(z)| > \xi \right) = 0 \quad (\text{A.6})$$

In addition, suppose a sequence of stochastic processes $\tilde{\mathcal{N}}_{m,n} \in \mathbf{D}, n = 1, 2, \dots$ satisfy $\tilde{\mathcal{N}}_{m,n} \rightarrow_{\mathcal{D}} \mathcal{M}_m$ under Skohord topology as $n \rightarrow \infty$. Then $\exists \delta > 0$ such that

$$\limsup_{n \rightarrow \infty} \mathbf{P} \left(\sup_{y, z \in [0, 1], |y-z| < \delta} |\tilde{\mathcal{N}}_{m,n}(y) - \tilde{\mathcal{N}}_{m,n}(z)| \geq \xi \right) \leq \xi \quad (\text{A.7})$$

proof of lemma A.1. From (A.4), for any $t_i \in [0, 1], i = 1, 2, \dots, r$, the random vector $(\mathcal{M}_m(t_1), \dots, \mathcal{M}_m(t_r))^T$ has joint normal distribution with mean 0 and covariance matrix $\{\mathcal{V}(2mt_i - m, 2mt_j - m)\}_{i,j=1,2,\dots,r}$, so the consistency conditions in Kolmogorov extension theorem are satisfied. $\forall 0 \leq t_1 \leq t \leq t_2 \leq 1$,

$$\begin{aligned} \mathbf{E}|\mathcal{M}_m(t) - \mathcal{M}_m(t_1)|^2 |\mathcal{M}_m(t) - \mathcal{M}_m(t_2)|^2 &\leq \frac{1}{2} (\mathbf{E}|\mathcal{M}_m(t) - \mathcal{M}_m(t_1)|^4 + \mathbf{E}|\mathcal{M}_m(t) - \mathcal{M}_m(t_2)|^4) \\ &\leq \frac{3}{2} (\mathbf{E}(\mathcal{M}_m(t) - \mathcal{M}_m(t_1))^2 + \mathbf{E}(\mathcal{M}_m(t) - \mathcal{M}_m(t_2))^2)^2 \end{aligned} \quad (\text{A.8})$$

The last inequality comes from the fact that $\mathcal{M}_m(t) - \mathcal{M}_m(t_1)$ and $\mathcal{M}_m(t) - \mathcal{M}_m(t_2)$ have normal distribution. Define $t_{i,m} = 2mt_i - m$ for $i = 1, 2$. Form assumption 1, \exists a constant $C > 0$ with

$$\begin{aligned} &\mathbf{E}(\mathcal{M}_m(t) - \mathcal{M}_m(t_1))^2 \\ &= \sigma^2 (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T A^{-1} b)(F'(2mt - m) - F'(t_{1,m}))^2 \\ &\quad + F(2mt - m) - F(t_{1,m}) - (F(2mt - m) - F(t_{1,m}))^2 \\ &- 2(\mathcal{X}_f^T A^{-1} b - 1)(F'(2mt - m) - F'(t_{1,m}))(H(2mt - m) - H(t_{1,m})) \leq C(t - t_1) \end{aligned} \quad (\text{A.9})$$

Similarly, $\mathbf{E}(\mathcal{M}_m(t_2) - \mathcal{M}_m(t))^2 \leq C(t_2 - t)$. Then (A.8) implies $\mathbf{E}|\mathcal{M}_m(t) - \mathcal{M}_m(t_1)|^2 |\mathcal{M}_m(t) - \mathcal{M}_m(t_2)|^2 \leq \frac{3}{2}C^2(t_2 - t_1)^2$. Set $\alpha = \beta = 1$ and choose the non-decreasing, continuous function $F(x) = \frac{\sqrt{3}}{\sqrt{2}}Cx$ in eq.(13.15) of Billingsley [6]. (A.9) also implies (13.16) in [6]. From theorem 13.6 in [6], $\exists \mathcal{M}_m \in \mathbf{D}$ satisfying (A.2). According to (A.9),

$$\mathbf{E}(\mathcal{M}_m(t) - \mathcal{M}_m(t_1))^4 \leq 3C^2(t - t_1)^2 \quad (\text{A.10})$$

so theorem 2.3 in Hahn [14] is satisfied by choosing $r = 4$ and the function

$$f(x) = 3C^2x^2 \Rightarrow \int_{[0,1]} x^{-(r+1)/r} f^{1/r}(x) dx = 4(3C^2)^{1/4} < \infty \quad (\text{A.11})$$

In particular, we can choose $\mathcal{M}_m \in \mathbf{D}$ such that $|\mathcal{M}_m(t) - \mathcal{M}_m(t_1)| \leq AH(|t - t_1|)$ almost surely, A is a random variable with $\mathbf{E}A^4 < \infty$, H is a continuous nondecreasing function on $[0, 1]$ such that $H(0) = 0$. This implies \mathcal{M}_m has continuous sample paths almost surely.

We prove (A.6) by

$$\mathbf{P}\left(\sup_{y,z \in [0,1], |y-z| < \delta} |\mathcal{M}_m(y) - \mathcal{M}_m(z)| > \xi\right) \leq \frac{\mathbf{E}A^4}{\xi^4} \times H^4(\delta) \quad (\text{A.12})$$

For any given $\delta > 0$, define a function

$$h_\delta(f) = \sup_{x,y \in [0,1], |x-y| < \delta} |f(x) - f(y)|, \text{ here } f \in \mathbf{D} \quad (\text{A.13})$$

From section 12, Billingsley [6], if $f_n, n = 1, \dots$ converges to f in \mathbf{D} , then \exists strictly increasing mappings $\lambda_n : [0, 1] \rightarrow [0, 1], n = 1, 2, \dots$ such that $\lim_{n \rightarrow \infty} \sup_{x \in [0,1]} |\lambda_n(x) - x| = 0$ and $\lim_{n \rightarrow \infty} \sup_{x \in [0,1]} |f_n(\lambda_n(x)) - f(x)| = 0$; so

$$\begin{aligned} |h_\delta(f_n) - h_\delta(f)| &\leq \sup_{x,y \in [0,1], |x-y| < \delta} |f_n(x) - f_n(y) - f(x) + f(y)| \\ &\leq \sup_{x \in [0,1]} |f_n(x) - f(x)| + \sup_{y \in [0,1]} |f_n(y) - f(y)| \\ &\leq 2\left(\sup_{x \in [0,1]} |f_n(x) - f(\lambda_n^{-1}(x))| + \sup_{x \in [0,1]} |f(\lambda_n^{-1}(x)) - f(x)|\right) \end{aligned} \quad (\text{A.14})$$

If f is continuous on $[0, 1]$, then $\lim_{n \rightarrow \infty} |h_\delta(f_n) - h_\delta(f)| = 0$. For \mathcal{M}_m is continuous almost surely, and \mathbf{R}, \mathbf{D} are Polish spaces (theorem 12.2 in Billingsley [6]), 3.8, page 348 in Jacod and Shiryaev [17] implies $h_\delta(\widetilde{\mathcal{N}}_{m,n}) \rightarrow_{\mathcal{L}} h_\delta(\mathcal{M}_m)$, and theorem 1.9 in Shao [36] implies

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbf{P}\left(\sup_{x,y \in [0,1], |x-y| < \delta} |\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}(y)| \geq \xi\right) \\ \leq \mathbf{P}\left(\sup_{x,y \in [0,1], |x-y| < \delta} |\mathcal{M}_m(x) - \mathcal{M}_m(y)| \geq \xi\right) < \xi \end{aligned} \quad (\text{A.15})$$

for sufficiently small $\delta > 0$. \square

Notably, $\widetilde{\mathcal{N}}_{m,n}$ may not be continuous for finite n . However, if $\widetilde{\mathcal{N}}_{m,n} \rightarrow_{\mathcal{L}} \mathcal{M}_m$, lemma A.1 implies that the discontinuity in $\widetilde{\mathcal{N}}_{m,n}$ should vanish asymptotically. Combine lemma A.1 with (A.3), we derive the following corollary:

COROLLARY A.1 Suppose assumption 1 to 4 hold true. Then for any given $0 < c < 1/4$,

$$\lim_{\delta \rightarrow 0} \sup_{|x-y|+|z-w| < \delta} |\mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq w)| = 0 \quad (\text{A.16})$$

and if $\widetilde{\mathcal{N}}_{m,n} \rightarrow_{\mathcal{L}} \mathcal{M}_m$, then

$$\lim_{n \rightarrow \infty} \sup_{x \in [\frac{1}{2}+c, 1-c], z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z)| = 0 \quad (\text{A.17})$$

Here $(x, z), (y, w) \in [\frac{1}{2}+c, 1-c] \times \mathbf{R}$. See (3.5) in the paper for the definition of the superscript $-$.

proof of corollary A.1. Without loss of generality, assume $z \leq w$. From (A.3), for $\forall \xi > 0$,

$$\begin{aligned} & |\mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq w)| \\ & \leq \mathbf{P}(|\mathcal{M}_m(x) - \mathcal{M}_m(y)| > \xi/2) + \mathbf{P}(|\mathcal{M}_m(1-x) - \mathcal{M}_m(1-y)| > \xi/2) \\ & + \mathbf{P}(z - \xi < \mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq z + \xi) + \mathbf{P}(z < \mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq w) \end{aligned} \quad (\text{A.18})$$

Define $y_m = 2my - m$. From assumption 4, $\min_{y \in [\frac{1}{2}+c, 1-c]} \mathcal{U}(y_m) > 0$ so

$$\begin{aligned} \mathbf{P}(z < \mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq w) &= \Phi\left(\frac{w}{\sqrt{\mathcal{U}(y_m)}}\right) - \Phi\left(\frac{z}{\sqrt{\mathcal{U}(y_m)}}\right) \\ &\leq \frac{\delta}{\min_{y \in [\frac{1}{2}+c, 1-c]} \sqrt{\mathcal{U}(y_m)}} \end{aligned} \quad (\text{A.19})$$

Similarly $\mathbf{P}(z - \xi < \mathcal{M}_m(y) - \mathcal{M}_m(1-y) \leq z + \xi) \leq \frac{2\xi}{\min_{y \in [\frac{1}{2}+c, 1-c]} \sqrt{\mathcal{U}(y_m)}}$. (A.16) is proved by applying lemma A.1 to (A.18).

For $\forall x \in [\frac{1}{2}+c, 1-c]$, define $g_x : \mathbf{D} \rightarrow \mathbf{R} : g_x(f) = f(x) - f^-(1-x)$. We use the same notation as (A.14). If f_n converges to f in \mathbf{D} and f is continuous, $|g_x(f_n) - g_x(f)| \leq |f_n(x) - f(\lambda_n^{-1}(x))| + |f(\lambda_n^{-1}(x)) - f(x)| + \limsup_{t \rightarrow 1-x, t < 1-x} |f_n(t) - f(\lambda_n^{-1}(t))| + \limsup_{t \rightarrow 1-x, t < 1-x} |f(\lambda_n^{-1}(t)) - f(t)|$, which tends to 0 as $n \rightarrow \infty$. Therefore, 3.8, page 348 in Jacod and Shiryaev [17] implies $g_x(\widetilde{\mathcal{N}}_{m,n}) \rightarrow_{\mathcal{L}} g_x(\mathcal{M}_m)$. $\forall \psi > 0, t \in \mathbf{R}$, define $G_0(x) = (1 - \min(1, \max(x, 0))^4)^4$, and $G_{\psi,t}(x) = G_0(\psi x - \psi t)$. From Xu et al. [47], \exists a constant $C > 0$ with

$$\mathbf{1}_{x \leq t} \leq G_{\psi,t}(x) \leq \mathbf{1}_{x \leq t+1/\psi}, \quad \sup_{x,t} |G'_{\psi,t}(x)| \leq C\psi, \quad \sup_{x,t} |G''_{\psi,t}(x)| \leq C\psi^2, \quad \sup_{x,t} |G'''_{\psi,t}(x)| \leq C\psi^3 \quad (\text{A.20})$$

For $\forall \psi > 0$, define the set $\mathcal{A}_\psi = \{G_{\psi,t} : t \in \mathbf{R}\}$. $\forall \delta > 0$, choose $\gamma = \delta/(C\psi)$, then $\forall G_{\psi,t} \in \mathcal{A}_\psi, x, y \in \mathbf{R}$ with $|x-y| < \gamma$, $|G_{\psi,t}(x) - G_{\psi,t}(y)| \leq C\psi|x-y| < \delta \Rightarrow \mathcal{A}_\psi$ is equi-continuous and uniformly bounded by 1. From theorem 3.1 in Rao [30],

$$\lim_{n \rightarrow \infty} \sup_{G_{\psi,t} \in \mathcal{A}_\psi} |\mathbf{E}G_{\psi,t}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x)) - \mathbf{E}G_{\psi,t}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x))| = 0 \quad (\text{A.21})$$

for any fixed $x \in [\frac{1}{2} + c, 1 - c]$. From (A.20),

$$\begin{aligned}
& \mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z) \\
& \leq \mathbf{E}G_{\psi,z} \left(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \right) - \mathbf{E}G_{\psi,z-1/\psi}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x)) \\
& \leq \sup_{G_{\psi,t} \in \mathcal{A}_\psi} |\mathbf{E}G_{\psi,t} \left(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \right) - \mathbf{E}G_{\psi,t}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x))| \\
& \quad + \mathbf{P} \left(z - \frac{1}{\psi} < \mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z + \frac{1}{\psi} \right) \\
& \quad \mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z) \\
& \geq \mathbf{E}G_{\psi,z-1/\psi} \left(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \right) - \mathbf{E}G_{\psi,z}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x)) \\
& \geq - \sup_{G_{\psi,t} \in \mathcal{A}_\psi} |\mathbf{E}G_{\psi,t} \left(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \right) - \mathbf{E}G_{\psi,t}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x))| \\
& \quad - \mathbf{P} \left(z - \frac{1}{\psi} < \mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z + \frac{1}{\psi} \right)
\end{aligned} \tag{A.22}$$

Choose $y = x, z = z + \frac{1}{\psi}, w = z - \frac{1}{\psi}$ in (A.16) and let $\psi \rightarrow \infty$,

$$\limsup_{n \rightarrow \infty} \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z)| = 0 \tag{A.23}$$

Finally, for any given $\xi > 0$, we choose $\frac{1}{2} + c = x_0 < x_1 < \dots < x_M = 1 - c$ and $x_i - x_{i-1} < \delta, i = 1, 2, \dots, M$ with sufficiently small $\delta > 0$. For $\forall x \in [\frac{1}{2} + c, 1 - c]$, $\exists I \in \{0, 1, \dots, M\}$ such that $|x - x_I| < \delta$, and

$$\begin{aligned}
& \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z)| \\
& \leq \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x_I) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I) \leq z)| \\
& + \max_{I=1,2,\dots,M} \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x_I) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I) \leq z) - \mathbf{P}(\mathcal{M}_m(x_I) - \mathcal{M}_m(1-x_I) \leq z)| \\
& \quad + \sup_{z \in \mathbf{R}} |\mathbf{P}(\mathcal{M}_m(x_I) - \mathcal{M}_m(1-x_I) \leq z) - \mathbf{P}(\mathcal{M}_m(x) - \mathcal{M}_m(1-x) \leq z)|
\end{aligned} \tag{A.24}$$

From (A.3), $\forall \xi > 0$,

$$\begin{aligned}
& \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x) \leq z) - \mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x_I) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I) \leq z)| \\
& \leq \mathbf{P} \left(|\widetilde{\mathcal{N}}_{m,n}^-(1-x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I)| > \frac{\xi}{2} \right) + \mathbf{P} \left(|\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}(x_I)| > \frac{\xi}{2} \right) \\
& + 2 \max_{I=1,2,\dots,M} \sup_{z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathcal{N}}_{m,n}(x_I) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I) \leq z) - \mathbf{P}(\mathcal{M}_m(x_I) - \mathcal{M}_m(1-x_I) \leq z)| \\
& \quad + \sup_{z \in \mathbf{R}} \mathbf{P}(z - \xi < \mathcal{M}_m(x_I) - \mathcal{M}_m(1-x_I) \leq z + \xi)
\end{aligned} \tag{A.25}$$

Since $\sup_{x \in [\frac{1}{2} + c, 1 - c]} \mathbf{P} \left(|\widetilde{\mathcal{N}}_{m,n}(x) - \widetilde{\mathcal{N}}_{m,n}(x_I)| > \frac{\xi}{2} \right)$ and $\sup_{x \in [\frac{1}{2} + c, 1 - c]} \mathbf{P} \left(|\widetilde{\mathcal{N}}_{m,n}^-(1-x) - \widetilde{\mathcal{N}}_{m,n}^-(1-x_I)| > \frac{\xi}{2} \right)$

are less or equal to $\mathbf{P}\left(\sup_{y,z \in [0,1], |y-z| < \delta} |\widetilde{\mathcal{N}}_{m,n}(y) - \widetilde{\mathcal{N}}_{m,n}(z)| > \frac{\xi}{2}\right)$, (A.7), (A.16) and (A.23) imply (A.17). \square

The second lemma focuses on showing the asymptotic continuity of the residuals' empirical process in the real world and in the bootstrap world. Define the stochastic processes

$$\widehat{\alpha}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{\widehat{\varepsilon}_i \leq x} - F(x)) \text{ and } \widetilde{\alpha}^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{\varepsilon_i^* \leq x} - \widehat{F}(x)) \quad (\text{A.26})$$

Here $\widehat{\varepsilon}_i$ and \widehat{F} are defined in (3.2). $\varepsilon_i^*, i = 1, 2, \dots, n$ are i.i.d. random variables generated from \widehat{F} . In algorithm 4.1, ε_i^* serves as the bootstrapped residuals. Define two assistant processes

$$\widetilde{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i \leq x} \text{ and } \widetilde{\alpha}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{\varepsilon_i \leq x} - F(x)), \text{ here } \forall x \in \mathbf{R} \quad (\text{A.27})$$

The notation O_p and o_p have the same meaning as definition 1.9 in Shao [36], i.e., two random variable sequences $X_n, Y_n, n = 1, 2, \dots$ satisfy $X_n = O_p(Y_n)$ if for $\forall t > 0$, \exists a constant C_t such that $\text{Prob}(|X_n| \geq C_t |Y_n|) \leq t$ for $n = 1, 2, \dots$. $X_n = o_p(Y_n)$ if $X_n/Y_n \rightarrow_p 0$ as $n \rightarrow \infty$.

LEMMA A.2 Suppose assumption 1 to 4 hold true. Then for any given $\xi > 0$ and $-\infty < r \leq s < \infty$, $\exists \delta > 0$ such that

$$\limsup_{n \rightarrow \infty} \mathbf{P}\left(\sup_{x,y \in [r,s], |x-y| < \delta} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| > \xi\right) < \xi \quad (\text{A.28})$$

Besides, $\exists \delta > 0$ and $N > 0$ such that $\forall n \geq N$,

$$\mathbf{P}\left(\left\{\mathbf{P}^*\left(\sup_{x,y \in [r,s], |x-y| < \delta} |\widetilde{\alpha}^*(x) - \widetilde{\alpha}^*(y)| > \xi\right) > \xi\right\}\right) < \xi \quad (\text{A.29})$$

proof of lemma A.2. From assumption 4, F is strictly increasing in \mathbf{R} . From lemma 4.1 and 4.2, Bickel and Freedman [5], \exists independent random variables $U_i, i = 1, 2, \dots$ with uniform distribution on $[0, 1]$, a Brownian bridge B and a constant C such that

$$\mathbf{P}\left(\sup_{x \in [0,1]} \left|\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{U_i \leq x} - x) - B(x)\right| \geq C \log(n)/\sqrt{n}\right) \leq C \log(n)/\sqrt{n} \quad (\text{A.30})$$

and $\forall 0 < \delta < 1/2, \xi > 0$,

$$\begin{aligned} & \mathbf{E} \sup_{x,y \in [0,1], |x-y| < \delta} |B(x) - B(y)| \leq C(-\delta \log(\delta))^{1/2} \\ \Rightarrow & \mathbf{P}\left(\sup_{x,y \in [0,1], |x-y| < \delta} |B(x) - B(y)| > \xi\right) \leq \frac{C(-\delta \log(\delta))^{1/2}}{\xi} \end{aligned} \quad (\text{A.31})$$

We choose $\varepsilon_i = F^{-1}(U_i), i = 1, 2, \dots, n$ (ε_i has distribution F according to page 150, Billingsley [6]),

$$\widetilde{\alpha}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{U_i \leq F(x)} - F(x)) \Rightarrow \mathbf{P}\left(\sup_{x \in \mathbf{R}} |\widetilde{\alpha}(x) - B(F(x))| \geq C \log(n)/\sqrt{n}\right) \leq C \log(n)/\sqrt{n} \quad (\text{A.32})$$

From assumption 3,

$$\max_{i=1,\dots,n} \frac{1}{n} \mathcal{X}_i^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_i \leq \frac{1}{n} M^2 \left\| \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \right\|_2 = O(1/n) \quad (\text{A.33})$$

and $\mathbf{E} \|(\mathcal{X}^T \mathcal{X})^{1/2}(\hat{\beta} - \beta)\|_2^2 = p\sigma^2$ implies $\|(\mathcal{X}^T \mathcal{X})^{1/2}(\hat{\beta} - \beta)\|_2 = O_p(1)$. Define

$$\hat{\lambda} = \frac{1}{n} \sum_{j=1}^n \tilde{\varepsilon}_j = \frac{1}{n} \sum_{j=1}^n \varepsilon_j - \overline{\mathcal{X}}_n^T (\hat{\beta} - \beta) \quad (\text{A.34})$$

here $\tilde{\varepsilon}_i$ and $\overline{\mathcal{X}}_n$ are defined in (3.2). With this definition we have $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \hat{\lambda}$. Besides,

$$\begin{aligned} \mathbf{E} \hat{\lambda}^2 &\leq 2\mathbf{E} \left(\frac{1}{n} \sum_{j=1}^n \varepsilon_j \right)^2 + 2\mathbf{E} \left(\overline{\mathcal{X}}_n^T (\hat{\beta} - \beta) \right)^2 = \frac{2\sigma^2}{n} + \frac{2\sigma^2 \overline{\mathcal{X}}_n^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \overline{\mathcal{X}}_n}{n} \\ &\Rightarrow \hat{\lambda} = O_p(1/\sqrt{n}) \end{aligned} \quad (\text{A.35})$$

Define

$$\tilde{\alpha}^\dagger(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{\tilde{\varepsilon}_i \leq x} - F(x)) \Rightarrow \hat{\alpha}(x) = \tilde{\alpha}^\dagger(x + \hat{\lambda}) + \sqrt{n}(F(x + \hat{\lambda}) - F(x)) \quad (\text{A.36})$$

From theorem 6.2.1 in Koul [18],

$$\sup_{x \in \mathbf{R}} |\tilde{\alpha}^\dagger(x) - \tilde{\alpha}(x) - \sqrt{n}F'(x)\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta)| = o_p(1) \quad (\text{A.37})$$

Therefore,

$$\begin{aligned} &\sup_{x \in \mathbf{R}} |\hat{\alpha}(x) - \tilde{\alpha}(x) - \sqrt{n}F'(x)\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta) - \sqrt{n}F'(x)\hat{\lambda}| \\ &\leq \sup_{x \in \mathbf{R}} |\tilde{\alpha}^\dagger(x) - \tilde{\alpha}(x) - \sqrt{n}F'(x)\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta)| + \sup_{x \in \mathbf{R}} |\tilde{\alpha}(x + \hat{\lambda}) - \tilde{\alpha}(x)| \\ &+ \sup_{x \in \mathbf{R}} \sqrt{n}|(F'(x + \hat{\lambda}) - F'(x))\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta)| + \sup_{x \in \mathbf{R}} \sqrt{n}|F(x + \hat{\lambda}) - F(x) - F'(x)\hat{\lambda}| \end{aligned} \quad (\text{A.38})$$

From assumption 1 and 3 and Taylor's theorem, $\sup_{x \in \mathbf{R}} \sqrt{n}|(F'(x + \hat{\lambda}) - F'(x))\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta)|$ and $\sup_{x \in \mathbf{R}} \sqrt{n}|F(x + \hat{\lambda}) - F(x) - F'(x)\hat{\lambda}|$ have order $O_p(1/\sqrt{n})$. From (A.32), with probability tending to 1,

$$\sup_{x \in \mathbf{R}} |\tilde{\alpha}(x + \hat{\lambda}) - \tilde{\alpha}(x)| \leq \frac{2C \log(n)}{\sqrt{n}} + \sup_{x \in \mathbf{R}} |B(F(x + \hat{\lambda})) - B(F(x))| \quad (\text{A.39})$$

F is uniform continuous according to assumption 1, so

$$\sup_{x \in \mathbf{R}} |\tilde{\alpha}(x + \hat{\lambda}) - \tilde{\alpha}(x)| = o_p(1) \Rightarrow \sup_{x \in \mathbf{R}} |\hat{\alpha}(x) - \tilde{\alpha}(x) - \sqrt{n}F'(x)\overline{\mathcal{X}}_n^T(\hat{\beta} - \beta) - \sqrt{n}F'(x)\hat{\lambda}| = o_p(1) \quad (\text{A.40})$$

For any given $-\infty < r \leq s < \infty$ and sufficiently small $\delta > 0$,

$$\begin{aligned} & \sup_{x,y \in [r,s], |x-y| < \delta} |\hat{\alpha}(x) - \hat{\alpha}(y)| \leq \sup_{x,y \in [r,s], |x-y| < \delta} |\tilde{\alpha}(x) - \tilde{\alpha}(y)| \\ & + \sup_{x,y \in [r,s], |x-y| < \delta} \sqrt{n} |(F'(x) - F'(y)) \times (\mathcal{X}_n^T(\hat{\beta} - \beta) + \hat{\lambda})| + o_p(1) \end{aligned} \quad (\text{A.41})$$

From assumption 1, (A.32) and (A.31), we prove (A.28).

Define the function $\hat{\phi}(x) = \inf\{t | x \leq \hat{F}(t)\}$, $x \in [0, 1]$. Page 150, Billingsley [6] implies $\hat{\phi}(x) \leq t \Leftrightarrow x \leq \hat{F}(t)$. If U has uniform distribution on $[0, 1]$, then $\hat{\phi}(U)$ has distribution \hat{F} . Without loss of generality, we choose $\varepsilon_i^* = \hat{\phi}(U_i)$, $i = 1, 2, \dots, n$; (A.30) implies

$$\tilde{\alpha}^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{U_i \leq \hat{F}(x)} - \hat{F}(x)) \Rightarrow \mathbf{P}^* \left(\sup_{x \in \mathbf{R}} |\tilde{\alpha}^*(x) - B(\hat{F}(x))| \geq C \log(n) / \sqrt{n} \right) \leq C \log(n) / \sqrt{n} \quad (\text{A.42})$$

From assumption 3

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i \leq x + \mathcal{X}_i^T(\hat{\beta} - \beta) + \hat{\lambda}} \leq \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i \leq x + M \|\hat{\beta} - \beta\|_2 + |\hat{\lambda}|} = \tilde{F}(x + M \|\hat{\beta} - \beta\|_2 + |\hat{\lambda}|) \quad (\text{A.43})$$

Similarly, $\hat{F}(x) \geq \tilde{F}(x - M \|\hat{\beta} - \beta\|_2 - |\hat{\lambda}|)$. For any given $\omega > 0$, we can find $C_\omega > 0$ with $\mathbf{P}(\|\hat{\beta} - \beta\|_2 > \frac{C_\omega}{2\sqrt{n}}) < \omega$ and $\mathbf{P}(|\hat{\lambda}| > \frac{MC_\omega}{2\sqrt{n}}) < \omega$ for any n . From Glivenko - Cantelli theorem and dominated convergence theorem, $\lim_{n \rightarrow \infty} \mathbf{P}(\sup_{x \in \mathbf{R}} |\tilde{F}(x) - F(x)| > \omega) = 0$. If $\|\hat{\beta} - \beta\|_2 \leq \frac{C_\omega}{2\sqrt{n}}$, $|\hat{\lambda}| \leq \frac{MC_\omega}{2\sqrt{n}}$ and $\sup_{x \in \mathbf{R}} |\tilde{F}(x) - F(x)| \leq \omega$, then for any given $-\infty < r \leq s < \infty$, $\delta > 0$, $-\omega + F(x - \frac{MC_\omega}{\sqrt{n}}) \leq \hat{F}(x) \leq \omega + F(x + \frac{MC_\omega}{\sqrt{n}})$, and

$$\sup_{r \leq x \leq y \leq s, y-x < \delta} \hat{F}(y) - \hat{F}(x) \leq 2\omega + \sup_{r \leq x \leq y \leq s, y-x < \delta} F(y + \frac{MC_\omega}{\sqrt{n}}) - F(x - \frac{MC_\omega}{\sqrt{n}}) \quad (\text{A.44})$$

For any given $-\infty < r \leq s < \infty$ and $\xi > 0$, we choose sufficiently small $\omega, \delta > 0$ and define $\zeta = 2\omega + \sup_{r \leq x \leq y \leq s, y-x < \delta} F(y + \frac{MC_\omega}{\sqrt{n}}) - F(x - \frac{MC_\omega}{\sqrt{n}})$,

$$\begin{aligned} & \sup_{x,y \in [r,s], |x-y| < \delta} |\tilde{\alpha}^*(x) - \tilde{\alpha}^*(y)| \leq 2 \sup_{x \in \mathbf{R}} |\tilde{\alpha}^*(x) - B(\hat{F}(x))| + \sup_{x,y \in [r,s], |x-y| < \delta} |B(\hat{F}(x)) - B(\hat{F}(y))| \\ & \leq 2 \sup_{x \in \mathbf{R}} |\tilde{\alpha}^*(x) - B(\hat{F}(x))| + \sup_{x,y \in [-\omega + F(r - \frac{MC_\omega}{\sqrt{n}}), \omega + F(s + \frac{MC_\omega}{\sqrt{n}})], |x-y| \leq \zeta} |B(x) - B(y)| \\ & \Rightarrow \mathbf{P}^* \left(\sup_{x,y \in [r,s], |x-y| < \delta} |\tilde{\alpha}^*(x) - \tilde{\alpha}^*(y)| > \xi \right) \\ & \leq \mathbf{P}^* \left(\sup_{x \in \mathbf{R}} |\tilde{\alpha}^*(x) - B(\hat{F}(x))| > \frac{\xi}{4} \right) + \mathbf{P}^* \left(\sup_{x,y \in [0,1], |x-y| \leq \zeta} |B(x) - B(y)| > \frac{\xi}{2} \right) \end{aligned} \quad (\text{A.45})$$

For F is uniform continuous, (A.31) and (A.42) imply (A.29). \square

A.2 Proof of theorem 4.2

The existence of \mathcal{M}_m has been shown in lemma A.1, and this section will complete the remaining steps. *proof of theorem 4.2.*

Prove $\tilde{M}_m \rightarrow_{\mathcal{L}} \mathcal{M}_m$. Here \tilde{M}_m is defined in (A.1).

According to theorem 13.5 in Billingsley [6], it suffices to verify the following conditions:

1. $\forall z_1, \dots, z_k \in [0, 1], (\tilde{M}_m(z_1), \dots, \tilde{M}_m(z_k)) \rightarrow_{\mathcal{L}} (\mathcal{M}_m(z_1), \dots, \mathcal{M}_m(z_k))$ in \mathbf{R}^k . According to Cramér-Wold device (theorem 1.9 in Shao [36]), this condition can be proved by showing

$$\sum_{j=1}^k s_j \tilde{M}_m(z_j) \rightarrow_{\mathcal{L}} \sum_{j=1}^k s_j \mathcal{M}_m(z_j) \quad (\text{A.46})$$

here $s_1, \dots, s_k \in \mathbf{R}$ are any given real numbers.

2. $\mathcal{M}_m(1) - \mathcal{M}_m(1 - \delta) \rightarrow_{\mathcal{L}} 0$ in \mathbf{R} as $\delta \rightarrow 0, \delta > 0$

3. $\exists b \geq 0, a > 1/2$, and a non-decreasing, continuous function G on $[0, 1]$ such that

$$\mathbf{E}|\tilde{M}_m(t) - \tilde{M}_m(s)|^{2b} |\tilde{M}_m(s) - \tilde{M}_m(r)|^{2b} \leq (G(t) - G(r))^{2a} \text{ for } \forall 1 \geq t > s > r \geq 0 \quad (\text{A.47})$$

For the first condition: define $c^T = (c_1, \dots, c_n) = \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T - \frac{1}{n} e^T \Rightarrow c_i = \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_i - \frac{1}{n}$. Here $e = (1, 1, \dots, 1)^T$. Define $z_{j,m} = 2mz_j - m, j = 1, 2, \dots, k$. For any given $s_1, \dots, s_k \in \mathbf{R}$

$$\sum_{j=1}^k s_j \tilde{M}_m(z_j) = \sum_{i=1}^n \left(\left(\sum_{j=1}^k \sqrt{ns_j} F'(z_{j,m}) \right) c_i \varepsilon_i - \frac{1}{\sqrt{n}} \left(\sum_{j=1}^k s_j (\mathbf{1}_{\varepsilon_i \leq z_{j,m}} - F(z_{j,m})) \right) \right) \Rightarrow \mathbf{E} \sum_{j=1}^k s_j \tilde{M}_m(z_j) = 0 \quad (\text{A.48})$$

Form assumption 3 and (5.8.4) in Horn and Johnson [16], we define $Y_i = (\sum_{j=1}^k \sqrt{ns_j} F'(z_{j,m})) c_i \varepsilon_i - \frac{1}{\sqrt{n}} (\sum_{j=1}^k s_j (\mathbf{1}_{\varepsilon_i \leq z_{j,m}} - F(z_{j,m})))$,

$$\begin{aligned} \mathbf{E}Y_i^2 &= n\sigma^2 c_i^2 \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) F'(z_{l,m}) + \frac{1}{n} \sum_{j=1}^k \sum_{l=1}^k s_j s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m}) F(z_{l,m})) \\ &\quad - 2c_i \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) H(z_{l,m}) \\ &\Rightarrow \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbf{E}Y_i^2 = \lim_{n \rightarrow \infty} \sigma^2 \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) F'(z_{l,m}) \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \overline{\mathcal{X}}_n \right) \\ &\quad + \sum_{j=1}^k \sum_{l=1}^k s_j s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m}) F(z_{l,m})) - \lim_{n \rightarrow \infty} 2 \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) H(z_{l,m}) \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \overline{\mathcal{X}}_n - 1 \right) \\ &= \sigma^2 K \times (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T A^{-1} b) + N - 2R \times (\mathcal{X}_f^T A^{-1} b - 1) \quad (\text{A.49}) \end{aligned}$$

here we define

$$K = \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) F'(z_{l,m}), \quad N = \sum_{l=1}^k s_l s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m}) F(z_{l,m})) \quad (\text{A.50})$$

$$\text{and } R = \sum_{j=1}^k \sum_{l=1}^k s_j s_l F'(z_{j,m}) H(z_{l,m})$$

From mean value inequality,

$$\sum_{i=1}^n \mathbf{E}|Y_i|^3 \leq 4k^2 \mathbf{E}|\varepsilon_i|^3 \sum_{j=1}^k |s_j F'(z_{j,m})|^3 \times n\sqrt{n} \sum_{i=1}^n |c_i|^3 + 4k^2 \sum_{j=1}^k |s_j|^3 \times \frac{1}{n\sqrt{n}} \sum_{i=1}^n \mathbf{E}|\mathbf{1}_{\varepsilon_i \leq z_{j,m}} - F(z_{j,m})|^3 \quad (\text{A.51})$$

From assumption 3,

$$\begin{aligned} n\sqrt{n} \sum_{i=1}^n |c_i|^3 &\leq n\sqrt{n} \max_{i=1,2,\dots,n} |c_i| \times \sum_{i=1}^n c_i^2 \\ &\leq \frac{1 + M^2 \|(\mathcal{X}^T \mathcal{X}/n)^{-1}\|_2}{\sqrt{n}} \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \bar{\mathcal{X}}_n \right) \end{aligned} \quad (\text{A.52})$$

which has order $O(1/\sqrt{n})$. $\|(\mathcal{X}^T \mathcal{X}/n)^{-1}\|_2$ is the matrix 2 norm of the matrix $(\mathcal{X}^T \mathcal{X}/n)^{-1}$. If $\sigma^2 K \times (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T A^{-1} b) + N - 2R \times (\mathcal{X}_f^T A^{-1} b - 1) \neq 0$, from Theorem 1.15, Theorem 1.11, and (1.97) in Shao [36],

$$\begin{aligned} \sum_{j=1}^k s_j \tilde{M}_m(z_j) &= \frac{\sum_{j=1}^k s_j \tilde{M}_m(z_j)}{\sqrt{\sum_{i=1}^n \mathbf{E}Y_i^2}} \times \sqrt{\sum_{i=1}^n \mathbf{E}Y_i^2} \\ &\rightarrow_{\mathcal{L}} N(0, \sigma^2 K \times (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T A^{-1} b) + N - 2R \times (\mathcal{X}_f^T A^{-1} b - 1)) \end{aligned} \quad (\text{A.53})$$

On the other hand, if $\sigma^2 K \times (\mathcal{X}_f^T A^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T A^{-1} b) + N - 2R \times (\mathcal{X}_f^T A^{-1} b - 1) = 0$, then $\forall \delta > 0$, from (A.49), $\lim_{n \rightarrow \infty} \mathbf{P}(|\sum_{j=1}^k s_j \tilde{M}_m(z_j)| \geq \delta) \leq \lim_{n \rightarrow \infty} \frac{\mathbf{E}|\sum_{j=1}^k s_j \tilde{M}_m(z_j)|^2}{\delta^2} = 0 \Rightarrow \sum_{j=1}^k s_j \tilde{M}_m(z_j) \rightarrow_{\mathcal{L}} 0$. From theorem 1.9, Shao [36], we prove (A.46) and the first condition.

The second condition: $\forall \xi > 0$,

$$\begin{aligned} \mathbf{P}(|\mathcal{M}_m(1) - \mathcal{M}_m(1 - \delta)| \geq \xi) &\leq \frac{\mathbf{E}|\mathcal{M}_m(1) - \mathcal{M}_m(1 - \delta)|^2}{\xi^2} \\ &\leq \frac{\sigma^2 (\mathcal{X}_f^T A^{-1} \mathcal{X}_f - 2\mathcal{X}_f^T A^{-1} b + 1) (F'(m) - F'(m - 2m\delta))^2}{\xi^2} \\ &\quad + \frac{2|\mathcal{X}_f^T A^{-1} b - 1| \times |F'(m) - F'(m - 2m\delta)| \times |H(m) - H(m - 2m\delta)|}{\xi^2} \\ &\quad + \frac{|F(m) - F(m - 2m\delta)| + |F(m) - F(m - 2m\delta)|^2}{\xi^2} \end{aligned} \quad (\text{A.54})$$

From assumption 1, $\lim_{\delta \rightarrow 0, \delta > 0} \text{Prob}(|\mathcal{M}_m(1) - \mathcal{M}_m(1 - \delta)| \geq \xi) = 0$, and we prove the second condition.

The third condition: we choose $b = a = 1$ in (A.47) and define $t_m = 2mt - m, \forall t$. For $\forall t, s \in [0, 1]$, we define

$$\begin{aligned} \mathcal{A}(t, s) &= \sqrt{n} \left(F'(t_m) - F'(s_m) \right) \times \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon - \frac{1}{n} \sum_{j=1}^n \varepsilon_j \right) \\ \text{and } \mathcal{B}(t, s) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{s_m < \varepsilon_i \leq t_m} - F(t_m) + F(s_m)) \end{aligned} \quad (\text{A.55})$$

From mean value inequality,

$$\begin{aligned} & \mathbf{E}|\tilde{M}_m(t) - \tilde{M}_m(s)|^2 |\tilde{M}_m(s) - \tilde{M}_m(r)|^2 \\ & \leq 4\mathbf{E}(\mathcal{A}(t,s)^2 \mathcal{A}(s,r)^2 + \mathcal{B}(t,s)^2 \mathcal{A}(s,r)^2 + \mathcal{A}(t,s)^2 \mathcal{B}(s,r)^2 + \mathcal{B}(t,s)^2 \mathcal{B}(s,r)^2) \end{aligned} \quad (\text{A.56})$$

From assumption 3, $\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \mathcal{X}_f \rightarrow \mathcal{X}_f^T A^{-1} \mathcal{X}_f$ and $\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \overline{\mathcal{X}}_n \rightarrow \mathcal{X}_f^T A^{-1} b$. Therefore, $\exists C > 0$ such that $|\mathcal{X}_f^T (\frac{\mathcal{X}^T \mathcal{X}}{n})^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T (\frac{\mathcal{X}^T \mathcal{X}}{n})^{-1} \overline{\mathcal{X}}_n| \leq C$ for $\forall n$. Define $c = (c_1, \dots, c_n)^T$, $c_i = \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_i - 1/n$, $\forall 1 \geq t > s > r \geq 0$

$$\begin{aligned} \mathbf{E}\mathcal{A}(t,s)^2 \mathcal{A}(s,r)^2 &= n^2 \left(F'(t_m) - F'(s_m) \right)^2 \left(F'(s_m) - F'(r_m) \right)^2 \mathbf{E}(c^T \varepsilon)^4 \\ &= n^2 \left(F'(t_m) - F'(s_m) \right)^2 \left(F'(s_m) - F'(r_m) \right)^2 \times \left(\mathbf{E}\varepsilon_1^4 \times \sum_{i=1}^n c_i^4 + 3\sigma^4 \sum_{i=1}^n \sum_{j=1, j \neq i}^n c_i^2 c_j^2 \right) \\ &\leq 16m^4 (\mathbf{E}\varepsilon_1^4 + 3\sigma^4) (t-s)^2 (s-r)^2 \times \sup_{x \in \mathbf{R}} |F''(x)|^4 \\ &\quad \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \overline{\mathcal{X}}_n \right)^2 \\ &\leq 16C^2 m^4 (\mathbf{E}\varepsilon_1^4 + 3\sigma^4) (t-r)^2 \times \sup_{x \in \mathbf{R}} |F''(x)|^4 \end{aligned} \quad (\text{A.57})$$

$$\begin{aligned} \mathbf{E}\mathcal{B}(t,s)^2 \mathcal{B}(s,r)^2 &\leq \frac{1}{n} \mathbf{E}(\mathbf{1}_{s_m < \varepsilon_1 \leq t_m} - F(t_m) + F(s_m))^2 (\mathbf{1}_{r_m < \varepsilon_1 \leq s_m} - F(s_m) + F(r_m))^2 \\ &\quad + \mathbf{E}(\mathbf{1}_{s_m < \varepsilon_1 \leq t_m} - F(t_m) + F(s_m))^2 \times \mathbf{E}(\mathbf{1}_{r_m < \varepsilon_1 \leq s_m} - F(s_m) + F(r_m))^2 \\ &\quad + 2(\mathbf{E}(\mathbf{1}_{s_m < \varepsilon_1 \leq t_m} - F(t_m) + F(s_m))(\mathbf{1}_{r_m < \varepsilon_1 \leq s_m} - F(s_m) + F(r_m)))^2 \\ &\leq \frac{3}{n} (F(t_m) - F(s_m))(F(s_m) - F(r_m)) + (F(t_m) - F(s_m))(F(s_m) - F(r_m)) \\ &\quad + 2(F(t_m) - F(s_m))^2 (F(s_m) - F(r_m))^2 \leq 6(F(t_m) - F(r_m))^2 \end{aligned} \quad (\text{A.58})$$

$$\begin{aligned} \mathbf{E}\mathcal{A}(t,s)^2 \mathcal{B}(s,r)^2 &= \left(F'(t_m) - F'(s_m) \right)^2 \times \left(\sum_{i=1}^n c_i^2 \mathbf{E}\varepsilon_i^2 (\mathbf{1}_{r_m < \varepsilon_i \leq s_m} - F(s_m) + F(r_m))^2 \right. \\ &\quad \left. + \sigma^2 (n-1) \sum_{i=1}^n c_i^2 \mathbf{E}(\mathbf{1}_{r_m < \varepsilon_1 \leq s_m} - F(s_m) + F(r_m))^2 + 2 \sum_{i=1}^n \sum_{j=1, j \neq i}^n c_i c_j \times (\mathbf{E}\varepsilon_1 \mathbf{1}_{r_m < \varepsilon_1 \leq s_m})^2 \right) \\ &\leq \sigma^2 \left(F'(t_m) - F'(s_m) \right)^2 \times \left(n \sum_{i=1}^n c_i^2 + 2 \left(\left(\sum_{i=1}^n c_i \right)^2 + \sum_{i=1}^n c_i^2 \right) \right) \end{aligned} \quad (\text{A.59})$$

Notice that $n \sum_{i=1}^n c_i^2 = \mathcal{X}_f^T (\frac{\mathcal{X}^T \mathcal{X}}{n})^{-1} \mathcal{X}_f + 1 - 2\mathcal{X}_f^T (\frac{\mathcal{X}^T \mathcal{X}}{n})^{-1} \overline{\mathcal{X}}_n$ and $\sum_{i=1}^n c_i = \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \overline{\mathcal{X}}_n - 1$, (A.47) is satisfied by choosing $G(x) = C'x$ with a sufficiently large constant C' . Then we prove $\tilde{M}_m \rightarrow_{\mathcal{L}} \mathcal{M}_m$. In particular, for any given $0 < r < s < \infty$, choose sufficiently large integer $m > s + 1$, from corollary A.1

$$\sup_{x \in [r, s], z \in \mathbf{R}} \left| \mathbf{P} \left(\tilde{M}_m \left(\frac{x+m}{2m} \right) - \tilde{M}_m \left(\frac{-x+m}{2m} \right) \leq z \right) - \mathbf{P} \left(\mathcal{M}_m \left(\frac{x+m}{2m} \right) - \mathcal{M}_m \left(\frac{-x+m}{2m} \right) \leq z \right) \right| = o(1) \quad (\text{A.60})$$

see(3.5) for the definition of the superscript $^-$. Since the random variable $\mathcal{M}_m\left(\frac{x+m}{2m}\right) - \mathcal{M}_m\left(\frac{-x+m}{2m}\right)$ has normal distribution with mean 0 and variance $\mathcal{U}(x)$ (see (3.4)), the proof remains showing that $\mathcal{S}(x)$ approximately equals $\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$.

Prove $\mathcal{S}(x)$ approximately equals $\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m^-\left(\frac{-x+m}{2m}\right)$

Recall the definition

$$\mathcal{S}(x) = \sqrt{n} \left(\mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq x) - \mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x) \right) \quad (\text{A.61})$$

here the conditional probability \mathbf{P}^* is defined in definition 1.1. Since $\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta} = \xi - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon$, here ξ is a random variable being independent of ϵ and having the same distribution as ϵ_1 . We have

$$\begin{aligned} \mathbf{P}^*(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq x) &= \mathbf{P}^*\left(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon \leq \xi \leq x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon\right) \\ &= F\left(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon\right) - F\left(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon\right) \end{aligned} \quad (\text{A.62})$$

On the other hand, we have $\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^* = \xi^* - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon^*$, here $\epsilon^* = (\epsilon_1^*, \dots, \epsilon_n^*)^T$ and ξ^*, ϵ^* are independent with distribution \hat{F} (see algorithm 4.1). Take the conditional distribution, we have

$$\begin{aligned} \mathbf{P}^*(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x) &= \begin{cases} \mathbf{E}^* \text{Prob} \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \mid \mathcal{Y}, \epsilon^* \right) & \text{for fixed design} \\ \mathbf{E}^* \text{Prob} \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \mid \mathcal{Y}, \mathcal{X}, \mathcal{X}_f, \epsilon^* \right) & \text{for random design} \end{cases} \\ &= \mathbf{E}^* \hat{F} \left(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon^* \right) - \mathbf{E}^* \hat{F}^- \left(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon^* \right) \end{aligned} \quad (\text{A.63})$$

Choose $m > s + 1$ and define $\hat{\alpha}(x) = \sqrt{n}(\hat{F}(x) - F(x))$ (the same as in (A.26)). $\forall x \in [r, s]$, from

Taylor's theorem

$$\begin{aligned}
\mathcal{S}(x) &= \sqrt{n}(F(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}) - F(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon})) \\
&\quad - \sqrt{n} \left(\mathbf{E}^* \left(\widehat{F}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{F}^-(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) \right) \right) \\
&= \left(F'(x) - F'(-x) \right) \times \sqrt{n} \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon} \\
&\quad + \frac{F''(\eta_1) - F''(\eta_2)}{2} \times \sqrt{n} (\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon})^2 \\
&\quad - \mathbf{E}^* (\widehat{\alpha}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}(x)) \\
&\quad + \mathbf{E}^* (\widehat{\alpha}^-(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}^-(-x)) \\
&\quad - \widehat{\alpha}(x) + \widehat{\alpha}^-(-x) - \sqrt{n} \mathbf{E}^* (F(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - F(x)) \\
&\quad + \sqrt{n} \mathbf{E}^* (F(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - F(-x)) \\
\Rightarrow \sup_{x \in [r, s]} |\mathcal{S}(x) - \left(\widetilde{M}_m \left(\frac{x+m}{2m} \right) - \widetilde{M}_m^- \left(\frac{-x+m}{2m} \right) \right)| &\leq \sqrt{n} (\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon})^2 \times \sup_{x \in \mathbf{R}} |F''(x)| \\
&\quad + \sup_{x \in [r, s]} |\mathbf{E}^* (\widehat{\alpha}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}(x))| \\
&\quad + \sup_{x \in [r, s]} |\mathbf{E}^* (\widehat{\alpha}^-(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}^-(-x))| \\
&\quad + \sup_{x \in [r, s]} |\widehat{\alpha}(x) - \widetilde{\alpha}(x) - \frac{F'(x)}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i| + \sup_{x \in [r, s]} |\widehat{\alpha}^-(-x) - \widetilde{\alpha}^-(-x) - \frac{F'(-x)}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i| \\
&\quad + \sqrt{n} \mathbf{E}^* (\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*)^2 \times \sup_{x \in \mathbf{R}} |F''(x)|
\end{aligned} \tag{A.64}$$

here η_1, η_2 are two arbitrary real numbers. From lemma A.2, for any given $\xi > 0$, $\exists 1/2 > \delta > 0$ such that for sufficiently large n , $\mathbf{P} \left(\sup_{x, y \in [-m, m], |x-y| < \delta} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| \leq \xi \right) > 1 - \xi$. If $\sup_{x, y \in [-m, m], |x-y| < \delta} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| \leq \xi$, then $\forall x \in [r, s]$,

$$\begin{aligned}
&|\mathbf{E}^* (\widehat{\alpha}^-(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}^-(-x))|, |\mathbf{E}^* (\widehat{\alpha}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*) - \widehat{\alpha}(x))| \\
&\leq \sqrt{n} \mathbf{P}^* (|\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*| > \delta) + \xi \leq \xi + \frac{\widehat{\sigma}^2}{\sqrt{n} \delta^2} \mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f
\end{aligned} \tag{A.65}$$

$\widehat{\sigma}^2$ is defined in (3.3). Also notice that

$$\sqrt{n} \mathbf{E}^* (\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \boldsymbol{\varepsilon}^*)^2 = \frac{\widehat{\sigma}^2}{\sqrt{n}} \mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f \tag{A.66}$$

Since $\mathbf{E} \widehat{\sigma}^2 \leq 4\sigma^2 + \frac{4\sigma^2 M^2}{n} \|(\frac{\mathcal{X}^T \mathcal{X}}{n})^{-1}\|_2 + 2\mathbf{E} \widehat{\lambda}^2 = O(1)$, combine with (A.38) we have $\forall \xi > 0$,

$$\mathbf{P} \left(\sup_{x \in [r, s]} |\mathcal{S}(x) - \left(\widetilde{M}_m \left(\frac{x+m}{2m} \right) - \widetilde{M}_m^- \left(\frac{-x+m}{2m} \right) \right)| > \xi \right) \rightarrow 0 \tag{A.67}$$

Finally, from (A.3) and corollary A.1, $\forall \delta > 0$,

$$\begin{aligned}
& \sup_{x \in [r, s], y \in \mathbf{R}} \left| \mathbf{P}(\mathcal{S}(x) \leq y) - \Phi\left(\frac{y}{\sqrt{\mathcal{U}(x)}}\right) \right| \\
& \leq \sup_{x \in [r, s]} \mathbf{P}\left(\left|\mathcal{S}(x) - \widetilde{\mathcal{M}}_m\left(\frac{x+m}{2m}\right) - \widetilde{\mathcal{M}}_m^-\left(\frac{-x+m}{2m}\right)\right| > \delta\right) \\
& + 3 \sup_{x \in [r, s], y \in \mathbf{R}} \left| \mathbf{P}\left(\widetilde{\mathcal{M}}_m\left(\frac{x+m}{2m}\right) - \widetilde{\mathcal{M}}_m^-\left(\frac{-x+m}{2m}\right) \leq y\right) - \Phi\left(\frac{y}{\sqrt{\mathcal{U}(x)}}\right) \right| \\
& + \sup_{x \in [r, s], y \in \mathbf{R}} \left(\Phi\left(\frac{y+\delta}{\sqrt{\mathcal{U}(x)}}\right) - \Phi\left(\frac{y-\delta}{\sqrt{\mathcal{U}(x)}}\right) \right)
\end{aligned} \tag{A.68}$$

From assumption 4, we prove (4.3). \square

B. Proofs of theorems in section 5

The Wasserstein distance can be used to quantify the difference between two probability distributions. We refer chapter 6, Villani [41] for a detail introduction. Lemma B.1 below bounds the Wasserstein distance between the distribution $T(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i - \bar{\varepsilon} \leq x}$ and $F(x) = \mathbf{P}(\varepsilon_1 \leq x)$, $x \in \mathbf{R}$. Here $\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$.

LEMMA B.1 Suppose assumption 1 and 2, then

$$\liminf_{n \rightarrow \infty} \mathbf{E}^* |X - Y|^2 = 0 \text{ almost surely} \tag{B.1}$$

The infimum is taken over all random variables $(X, Y) \in \mathbf{R}^2$ such that $\mathbf{P}^*(X \leq x) = T(x)$ and $\mathbf{P}^*(Y \leq x) = F(x)$.

Proof. From assumption 1, Gilvenko-Cantelli theorem, and the strong law of large number (e.g., theorem 1.13 in Shao [36]),

$$\limsup_{n \rightarrow \infty} |T(x) - F(x)| \leq \limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i \leq x} - F(x) \right| + \limsup_{n \rightarrow \infty} |F(x + \bar{\varepsilon}) - F(x)| = 0 \text{ almost surely} \tag{B.2}$$

From the strong law of large number, $\lim_{n \rightarrow \infty} \int_{\mathbf{R}} x^2 dT = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 - \lim_{n \rightarrow \infty} \bar{\varepsilon}^2 = \sigma^2$ almost surely. Choose $x_0 = 0$ in definition 6.8, Villani [41]. From proposition 5.7, page 112 in Çinlar [7] and theorem 6.9, Villani [41], we prove (B.1). \square

Recall (5.6) of the paper, the stochastic process $\widehat{\mathcal{S}}(x)$ is defined as

$$\begin{aligned}
\widehat{\mathcal{M}}(x) &= \sqrt{n} \widehat{F} \left(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{j=1}^n e_j^* \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathbf{1}_{e_j^* \leq x} \\
&\text{and } \widehat{\mathcal{S}}(x) = \widehat{\mathcal{M}}(x) - \widehat{\mathcal{M}}(-x)
\end{aligned} \tag{B.3}$$

Lemma B.2 ensures that $\widehat{\mathcal{S}}$ (defined in (5.6) of the paper) has the same asymptotic distribution as \mathcal{S} (defined in (4.2), also see theorem 4.2).

LEMMA B.2 Suppose assumption 1 to 4 hold true. Then for any given $0 < r < s < \infty, \xi > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\sup_{x \in [r, s]} \sup_{y \in \mathbf{R}} |\mathbf{P}^* (\widehat{\mathcal{S}}(x) \leq y) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right)| > \xi \right) = 0 \quad (\text{B.4})$$

here Φ is the cumulative distribution function of a standard normal random variable

Recall that $\Phi^{-1}(\alpha)$ is the α -th quantile of Φ . For any given $0 < r < s < \infty$ and $\xi > 0$, lemma B.2 implies with probability tending to 1, $\forall 2\xi < 1 - \gamma < 1 - \xi, r \leq x \leq s$,

$$\begin{aligned} \mathbf{P}^* \left(\widehat{\mathcal{S}}(x) \leq \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma - 2\xi) \right) - (1 - \gamma - 2\xi) &\leq \xi \\ \Rightarrow d_{1-\gamma}^*(x) &\geq \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma - 2\xi) \\ \mathbf{P}^* \left(\widehat{\mathcal{S}}(x) \leq \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma + \xi) \right) - (1 - \gamma + \xi) &\geq -\xi \\ \Rightarrow d_{1-\gamma}^*(x) &\leq \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma + \xi) \end{aligned} \quad (\text{B.5})$$

see (5.7) for the definition of $d_{1-\gamma}^*(x)$.

Suppose the integer $m > s + 1$. In (A.60) we show the stochastic process $\tilde{M}_m \left(\frac{x+m}{2m} \right) - \tilde{M}_m^- \left(\frac{-x+m}{2m} \right)$ (defined in (A.1)) has an asymptotic distribution $\Phi \left(\cdot / \sqrt{\mathcal{U}(x)} \right)$. So the remaining problem involves approxi-

imating the distribution of $\widehat{\mathcal{S}}(x)$ by the distribution of $\tilde{M}_m \left(\frac{x+m}{2m} \right) - \tilde{M}_m^- \left(\frac{-x+m}{2m} \right)$.

Proof of lemma B.2. From lemma B.1, almost surely for $\forall 1/4 > \delta > 0, \exists N > 0$ such that $\forall n \geq N$, there exists a random vector $(e_1^\dagger, \varepsilon_1^\dagger) \in \mathbf{R}^2$ such that $\mathbf{P}^*(e_1^\dagger \leq x) = T(x)$ (defined in lemma B.1) and $\mathbf{P}^*(\varepsilon_1^\dagger \leq x) = F(x)$. Moreover, $\mathbf{E}^*(\varepsilon_1^\dagger - e_1^\dagger)^2 < \delta^9$. We generate n i.i.d. observations $(e_i^\dagger, \varepsilon_i^\dagger)$, $i = 1, 2, \dots, n$ and define $e^\dagger = (e_1^\dagger, \dots, e_n^\dagger)^T$ as well as $\varepsilon^\dagger = (\varepsilon_1^\dagger, \dots, \varepsilon_n^\dagger)^T$. Suppose $m > s + 1$ and define

$$\begin{aligned} \tilde{\mathcal{M}}_m^\dagger(x) &= \sqrt{n} F'(x_m) \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^\dagger - \frac{1}{n} \sum_{j=1}^n e_j^\dagger \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n (\mathbf{1}_{e_j^\dagger \leq x_m} - T(x_m)) \\ \tilde{M}_m^\dagger(x) &= \sqrt{n} F'(x_m) \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^\dagger - \frac{1}{n} \sum_{j=1}^n \varepsilon_j^\dagger \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n (\mathbf{1}_{\varepsilon_j^\dagger \leq x_m} - F(x_m)) \end{aligned} \quad (\text{B.6})$$

here $x \in [0, 1]$, $x_m = 2mx - m$. With this definition we have $\mathbf{P}^*(\tilde{\mathcal{M}}_m^\dagger(x) \leq y) = \mathbf{P}(\tilde{M}_m^\dagger(x) \leq y)$ for any x, y .

For any given $1/4 > \xi > 0$ and $x \in (\frac{1}{2}, 1]$,

$$\begin{aligned}
& \mathbf{P}^* \left(|(\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x)) - (\widetilde{M}_m^\dagger(x) - \widetilde{M}_m^{\dagger-}(1-x))| > 3\xi \right) \\
& \leq \mathbf{P}^* \left(\sqrt{n} |F'(x_m) - F'(-x_m)| \times |\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T (\varepsilon^\dagger - e^\dagger)| > \xi \right) \\
& \quad + \mathbf{P}^* \left(|F'(x_m) - F'(-x_m)| \times \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i^\dagger - e_i^\dagger) \right| > \xi \right) \\
& + \mathbf{P}^* \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{-x_m \leq e_i^\dagger \leq x_m} - T(x_m) + T^-(x_m) - \mathbf{1}_{-x_m \leq \varepsilon_i^\dagger \leq x_m} + F(x_m) - F(-x_m) \right| > \xi \right) \quad (\text{B.7}) \\
& \leq \frac{(F'(x_m) - F'(-x_m))^2 \mathbf{E}^*(\varepsilon_1^\dagger - e_1^\dagger)^2}{\xi^2} \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 \right) \\
& \quad + \frac{4}{\xi^2} \mathbf{E}^*(\mathbf{1}_{e_1^\dagger \leq x_m} - T(x_m) - \mathbf{1}_{\varepsilon_1^\dagger \leq x_m} + F(x_m))^2 \\
& \quad + \frac{4}{\xi^2} \mathbf{E}^*(\mathbf{1}_{e_1^\dagger < -x_m} - T^-(x_m) - \mathbf{1}_{\varepsilon_1^\dagger < -x_m} + F(-x_m))^2
\end{aligned}$$

Notice that

$$\mathbf{E}^*(\mathbf{1}_{e_1^\dagger \leq x_m} - T(x_m) - \mathbf{1}_{\varepsilon_1^\dagger \leq x_m} + F(x_m))^2 \leq 2\mathbf{E}^*(\mathbf{1}_{e_1^\dagger \leq x_m} - \mathbf{1}_{\varepsilon_1^\dagger \leq x_m})^2 + 2 \sup_{x \in \mathbf{R}} |T(x) - F(x)|^2 \quad (\text{B.8})$$

from (A.3),

$$\begin{aligned}
\mathbf{E}^* |\mathbf{1}_{e_1^\dagger \leq x_m} - \mathbf{1}_{\varepsilon_1^\dagger \leq x_m}| & \leq \mathbf{P}^*(|e_1^\dagger - \varepsilon_1^\dagger| > \xi) + F(x_m + \xi) - F(x_m - \xi) \\
& \leq \frac{\delta^9}{\xi^2} + \sup_{x \in \mathbf{R}} (F(x) - F(x - 2\xi)) \quad (\text{B.9})
\end{aligned}$$

From dominated convergence theorem

$$\begin{aligned}
& \mathbf{E}^*(\mathbf{1}_{e_1^\dagger < -x_m} - T^-(x_m) - \mathbf{1}_{\varepsilon_1^\dagger < -x_m} + F(-x_m))^2 \\
& = \lim_{h \rightarrow \infty} \mathbf{E}^*(\mathbf{1}_{e_1^\dagger \leq -x_m - \frac{1}{h}} - T(-x_m - \frac{1}{h}) - \mathbf{1}_{\varepsilon_1^\dagger \leq -x_m - \frac{1}{h}} + F(-x_m - \frac{1}{h}))^2 \quad (\text{B.10}) \\
& \leq \frac{2\delta^9}{\xi^2} + 2 \sup_{x \in \mathbf{R}} (F(x) - F(x - 2\xi)) + 2 \sup_{x \in \mathbf{R}} |T(x) - F(x)|^2
\end{aligned}$$

therefore, from (A.3), assumption 4 and (A.60)

$$\begin{aligned}
& \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^* \left(\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x_m)}} \right)| \\
& \leq \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} \mathbf{P}^* \left(|(\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x)) - (\widetilde{M}_m^\dagger(x) - \widetilde{M}_m^{\dagger-}(1-x))| > 3\xi \right) \\
& \quad + 3 \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^* \left(\widetilde{M}_m^\dagger(x) - \widetilde{M}_m^{\dagger-}(1-x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x_m)}} \right)| \quad (\text{B.11}) \\
& \quad + \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} \left(\Phi \left(\frac{y+3\xi}{\sqrt{\mathcal{U}(x_m)}} \right) - \Phi \left(\frac{y-3\xi}{\sqrt{\mathcal{U}(x_m)}} \right) \right) \\
& \Rightarrow \lim_{n \rightarrow \infty} \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^* \left(\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x_m)}} \right)| = 0 \text{ almost surely}
\end{aligned}$$

Define a random variable $(e_1^*, e_1^\dagger) \in \mathbf{R}^2$ which has probability mass $1/n$ on $(\widehat{\varepsilon}_i, \varepsilon_i - \bar{\varepsilon}), i = 1, 2, \dots, n$. We generate independent random variables $(e_i^*, e_i^\dagger), i = 1, 2, \dots, n$ having the same distribution as (e_1^*, e_1^\dagger) . Define $e^* = (e_1^*, \dots, e_n^*)^T, e^\dagger = (e_1^\dagger, \dots, e_n^\dagger)^T$. With this definition e_1^\dagger still has the cumulative distribution function $T(x)$. Define the stochastic process

$$\widetilde{\mathcal{M}}_m^*(x) = \sqrt{n} F'(x_m) \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right) - \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{1}_{e_i^* \leq x_m} - \widehat{F}(x_m)) \quad (\text{B.12})$$

here $x_m = 2mx - m$. This process uses the same mechanism for generating residuals e^* as in *RBUG* (defined in algorithm 5.1). We have

$$\begin{aligned}
& \mathbf{P}^* \left(|\widetilde{\mathcal{M}}_m^*(x) - \widetilde{\mathcal{M}}_m^{*-}(1-x) - \widetilde{\mathcal{M}}_m^\dagger(x) + \widetilde{\mathcal{M}}_m^{\dagger-}(1-x)| > 3\xi \right) \\
& \leq \frac{|F'(x_m) - F'(-x_m)|^2 \mathbf{E}^*(e_1^* - e_1^\dagger)^2}{\xi^2} \times \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 \right) \\
& \quad + \frac{4}{\xi^2} \mathbf{E}^*(\mathbf{1}_{e_1^* \leq x_m} - \widehat{F}(x_m) - \mathbf{1}_{e_1^\dagger \leq x_m} + T(x_m))^2 \\
& \quad + \frac{4}{\xi^2} \mathbf{E}^*(\mathbf{1}_{e_1^* < -x_m} - \widehat{F}^-(x_m) - \mathbf{1}_{e_1^\dagger < -x_m} + T^-(x_m))^2 \quad (\text{B.13})
\end{aligned}$$

Recall $\overline{\mathcal{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathcal{X}_i$,

$$\begin{aligned}
& \mathbf{E}^*(e_1^* - e_1^\dagger)^2 = \frac{1}{n} \sum_{i=1}^n (\widehat{\varepsilon}_i - \varepsilon_i + \bar{\varepsilon})^2 = \frac{1}{n} \sum_{i=1}^n \left((\mathcal{X}_i - \overline{\mathcal{X}}_n)^T (\widehat{\beta} - \beta) \right)^2 \\
& \mathbf{E} \left((\mathcal{X}_i - \overline{\mathcal{X}}_n)^T (\widehat{\beta} - \beta) \right)^2 = \sigma^2 (\mathcal{X}_i - \overline{\mathcal{X}}_n)^T (\mathcal{X}^T \mathcal{X})^{-1} (\mathcal{X}_i - \overline{\mathcal{X}}_n) \quad (\text{B.14})
\end{aligned}$$

Assumption 3 implies $\mathbf{E}^*(e_1^* - e_1^\dagger)^2 = O_p(1/n)$. From assumption 3 and Cauchy inequality

$$\begin{aligned}\widehat{F}(x) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\varepsilon_i - \bar{\varepsilon} \leq x + (\mathcal{X}_i^* - \bar{\mathcal{X}}_n)^T(\widehat{\beta} - \beta)} \leq T(x + 2M\|\widehat{\beta} - \beta\|_2) \\ \text{and } \widehat{F}(x) &\geq T(x - 2M\|\widehat{\beta} - \beta\|_2)\end{aligned}\tag{B.15}$$

therefore

$$\begin{aligned}\sup_{x \in \mathbf{R}} |\widehat{F}(x) - T(x)| &\leq \sup_{x \in \mathbf{R}} |T(x + 2M\|\widehat{\beta} - \beta\|_2) - T(x)| + \sup_{x \in \mathbf{R}} |T(x - 2M\|\widehat{\beta} - \beta\|_2) - T(x)| \\ &\leq 4 \sup_{x \in \mathbf{R}} |F(x) - T(x)| + 2 \sup_{x \in \mathbf{R}} |F(x + 2M\|\widehat{\beta} - \beta\|_2) - F(x)| + 2 \sup_{x \in \mathbf{R}} |F(x - 2M\|\widehat{\beta} - \beta\|_2) - F(x)|\end{aligned}\tag{B.16}$$

Since

$$\begin{aligned}\mathbf{E}^*(\mathbf{1}_{e_1^* \leq x_m} - \widehat{F}(x_m) - \mathbf{1}_{e_1^\dagger \leq x_m} + T(x_m))^2 \\ \leq \frac{2\mathbf{E}^*(e_1^* - e_1^\dagger)^2}{\xi^2} + 2 \sup_{x \in \mathbf{R}} (T(x + \xi) - T(x - \xi)) + 2 \sup_{x \in \mathbf{R}} |\widehat{F}(x) - T(x)|^2\end{aligned}\tag{B.17}$$

The dominated convergence theorem implies

$$\begin{aligned}\mathbf{E}^*(\mathbf{1}_{e_1^* < -x_m} - \widehat{F}^-(x_m) - \mathbf{1}_{e_1^\dagger < -x_m} + T^-(x_m))^2 \\ = \lim_{h \rightarrow \infty} \mathbf{E}^*(\mathbf{1}_{e_1^* \leq -x_m - \frac{1}{h}} - \widehat{F}(-x_m - \frac{1}{h}) - \mathbf{1}_{e_1^\dagger \leq -x_m - \frac{1}{h}} + T(-x_m - \frac{1}{h}))^2 \\ \leq \frac{2\mathbf{E}^*(e_1^* - e_1^\dagger)^2}{\xi^2} + 2 \sup_{x \in \mathbf{R}} (T(x + \xi) - T(x - \xi)) + 2 \sup_{x \in \mathbf{R}} |\widehat{F}(x) - T(x)|^2\end{aligned}\tag{B.18}$$

and

$$\begin{aligned}\sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^*(\widetilde{\mathcal{M}}_m^*(x) - \widetilde{\mathcal{M}}_m^{*-}(1-x) \leq y) - \Phi\left(\frac{y}{\sqrt{\mathcal{U}(x_m)}}\right)| \\ \leq \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} \mathbf{P}^*(|(\widetilde{\mathcal{M}}_m^*(x) - \widetilde{\mathcal{M}}_m^{*-}(1-x)) - (\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x))| > 3\xi) \\ + 3 \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^*(\widetilde{\mathcal{M}}_m^\dagger(x) - \widetilde{\mathcal{M}}_m^{\dagger-}(1-x) \leq y) - \Phi\left(\frac{y}{\sqrt{\mathcal{U}(x_m)}}\right)| \\ + \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} \left(\Phi\left(\frac{y+3\xi}{\sqrt{\mathcal{U}(x_m)}}\right) - \Phi\left(\frac{y-3\xi}{\sqrt{\mathcal{U}(x_m)}}\right) \right)\end{aligned}\tag{B.19}$$

(B.2), (B.11), and (B.16) imply for $\forall \xi > 0$

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^*(\widetilde{\mathcal{M}}_m^*(x) - \widetilde{\mathcal{M}}_m^{*-}(1-x) \leq y) - \Phi\left(\frac{y}{\sqrt{\mathcal{U}(x_m)}}\right)| > \xi\right) = 0 \tag{B.20}$$

Finally, we adopt the notations in lemma A.2. Recall (5.6) (or (B.3) in this section), define $x_m = 2mx - m$

$$\begin{aligned} \sup_{x \in [0,1]} |\widehat{\mathcal{M}}(x_m) - \widetilde{\mathcal{M}}_m^*(x)| &\leq \frac{\sqrt{n} \sup_{x \in \mathbf{R}} |F''(x)|}{2} \times \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right)^2 \\ &\quad + \sup_{x \in [0,1]} |\widehat{\alpha}(x_m + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^*) - \widehat{\alpha}(x_m)| \end{aligned} \quad (\text{B.21})$$

and

$$\begin{aligned} \sup_{x \in [r,s]} |\widehat{\mathcal{P}}(x) - \left(\widetilde{\mathcal{M}}_m^*\left(\frac{x+m}{2m}\right) - \widetilde{\mathcal{M}}_m^{*-}\left(\frac{-x+m}{2m}\right) \right)| &\leq \sup_{x \in [r,s]} |\widehat{\mathcal{M}}(x) - \widetilde{\mathcal{M}}_m^*\left(\frac{x+m}{2m}\right)| \\ &\quad + \sup_{x \in [r,s]} \lim_{h \rightarrow \infty} |\widehat{\mathcal{M}}(-x - \frac{1}{h}) - \widetilde{\mathcal{M}}_m^*\left(\frac{-x+m}{2m} - \frac{1}{2hm}\right)| \leq 2 \sup_{x \in [0,1]} |\widehat{\mathcal{M}}(x_m) - \widetilde{\mathcal{M}}_m^*(x)| \end{aligned} \quad (\text{B.22})$$

$\forall \xi > 0$, (A.3) implies

$$\begin{aligned} \sup_{x \in [r,s], y \in \mathbf{R}} |\mathbf{P}^* \left(\widehat{\mathcal{P}}(x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right)| &\leq \mathbf{P}^* \left(\sup_{x \in [0,1]} |\widehat{\mathcal{M}}(x_m) - \widetilde{\mathcal{M}}_m^*(x)| > \xi \right) \\ &\quad + \sup_{x \in [r,s], y \in \mathbf{R}} \left(\Phi \left(\frac{y+2\xi}{\sqrt{\mathcal{U}(x)}} \right) - \Phi \left(\frac{y-2\xi}{\sqrt{\mathcal{U}(x)}} \right) \right) \\ &\quad + 3 \sup_{x \in [\frac{r+m}{2m}, \frac{s+m}{2m}], y \in \mathbf{R}} |\mathbf{P}^* \left(\widetilde{\mathcal{M}}_m^*(x) - \widetilde{\mathcal{M}}_m^{*-}(1-x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x_m)}} \right)| \end{aligned} \quad (\text{B.23})$$

From lemma A.2, for any given $\xi > 0$, $\exists \frac{1}{4} > \xi_2 > 0, N > 0$ such that for any $n \geq N$,

$\mathbf{P} \left(\sup_{x,y \in [-m-1, m+1], |x-y| < \xi_2} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| > \frac{\xi}{4} \right) < \xi$. If $\sup_{x,y \in [-m-1, m+1], |x-y| < \xi_2} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| \leq \frac{\xi}{4}$, then

$$\begin{aligned} &\mathbf{P}^* \left(\sup_{x \in [0,1]} |\widehat{\mathcal{M}}(x_m) - \widetilde{\mathcal{M}}_m^*(x)| > \xi \right) \\ &\leq \mathbf{P}^* \left(\frac{\sqrt{n} \sup_{x \in \mathbf{R}} |F''(x)|}{2} \times \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right)^2 > \frac{\xi}{2} \right) \\ &\quad + \mathbf{P}^* \left(\left| \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right| \geq \xi_2 \right) \\ &\leq \left(\frac{\sqrt{n} \sup_{x \in \mathbf{R}} |F''(x)|}{\xi} + \frac{1}{\xi_2^2} \right) \mathbf{E}^* \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right)^2 \end{aligned} \quad (\text{B.24})$$

Since $\mathbf{E}^* \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T e^* - \frac{1}{n} \sum_{i=1}^n e_i^* \right)^2 \leq \frac{2\widehat{\sigma}^2}{n} \left(\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \mathcal{X}_f + 1 \right)$. From (B.19) we prove (B.4). \square

In lemma B.3, we define

$$G^*(x) = \mathbf{P}^* \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \widehat{\beta}^*| \leq x \right), \quad x \in \mathbf{R} \quad (\text{B.25})$$

See algorithm 5.1 for the meaning of \mathcal{Y}_f^* and $\hat{\beta}^*$

LEMMA B.3 Suppose assumption 1 to 4. Then $\forall -\infty < r < s < \infty, \zeta > 0, \exists \delta > 0$ such that

$$\limsup_{n \rightarrow \infty} \mathbf{P} \left(\sup_{x \in [r, s]} \sqrt{n} \left(G^*(x) - G^*\left(x - \frac{\delta}{\sqrt{n}}\right) \right) \geq \zeta \right) < \zeta \quad (\text{B.26})$$

Proof. We adopt the notations in lemma A.2 and recall $\mathcal{Y}_f^* = \mathcal{X}_f^T \hat{\beta} + \xi^*$. By conditioning on ε^* ,

$$\begin{aligned} G^*(x) &= \mathbf{E}^* \mathbf{P}^* \left(\left| \xi^* - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^* \right| \leq x \middle| \varepsilon^* \right) \\ &= \mathbf{E}^* \hat{F}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - \mathbf{E}^* \hat{F}^-(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) \\ &= (F(x) - F(-x)) + \mathbf{E}^* (F(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - F(x)) \\ &\quad - \mathbf{E}^* (F(-x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - F(-x)) + \frac{(\hat{\alpha}(x) - \hat{\alpha}^-(x))}{\sqrt{n}} \\ &\quad + \frac{1}{\sqrt{n}} \mathbf{E}^* (\hat{\alpha}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - \hat{\alpha}(x)) \\ &\quad - \frac{1}{\sqrt{n}} \mathbf{E}^* (\hat{\alpha}^-(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - \hat{\alpha}^-(x)) \end{aligned} \quad (\text{B.27})$$

Therefore, for $\forall \frac{1}{4} > \delta > 0$,

$$\begin{aligned} \sup_{x \in [r, s]} \sqrt{n} \left(G^*(x) - G^*\left(x - \frac{\delta}{\sqrt{n}}\right) \right) &\leq 2\delta \sup_{x \in [-s-1, s+1]} |F'(x)| \\ &\quad + \frac{2 \sup_{x \in \mathbf{R}} |F''(x)| \hat{\sigma}^2}{\sqrt{n}} \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f \right) \\ &\quad + 2 \sup_{x, y \in [-s-1, s+1], |x-y| \leq \frac{\delta}{\sqrt{n}}} |\hat{\alpha}(x) - \hat{\alpha}(y)| \\ &\quad + 4 \sup_{x \in [-s-1, s+1]} \mathbf{E}^* |\hat{\alpha}(x + \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \varepsilon^*) - \hat{\alpha}(x)| \end{aligned} \quad (\text{B.28})$$

From lemma A.2 and (A.65), we prove (B.26). \square

Suppose assumption 1 to 4, from (B.27), (B.28), (B.16), and (B.2),

$$\begin{aligned} \sup_{x > 0} |G^*(x) - (F(x) - F(-x))| &\leq 2 \sup_{x \in \mathbf{R}} |\hat{F}(x) - F(x)| \\ &\quad + \frac{\sup_{x \in \mathbf{R}} |F''(x)| \hat{\sigma}^2}{n} \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f \right) \end{aligned} \quad (\text{B.29})$$

which implies $\forall \xi > 0, \lim_{n \rightarrow \infty} \mathbf{P}(\sup_{x > 0} |G^*(x) - (F(x) - F(-x))| > \xi) = 0$. If $\sup_{x > 0} |G^*(x) - (F(x) - F(-x))| \leq \xi$, by defining $c_{1-\alpha}$ such that $F(c_{1-\alpha}) - F(-c_{1-\alpha}) = 1 - \alpha$,

$$\begin{aligned} G^*(c_{1-\alpha+2\xi}) &\geq 1 - \alpha + \xi, \quad G^*(c_{1-\alpha-2\xi}) \leq 1 - \alpha - \xi \\ \Rightarrow c_{1-\alpha-2\xi} &\leq c_{1-\alpha}^* \leq c_{1-\alpha+2\xi}, \quad \forall 2\xi < \alpha < 1 - 2\xi \end{aligned} \quad (\text{B.30})$$

proof of theorem 5.2. Recall $\Phi^{-1}(\alpha)$ is the α -th quantile of the standard normal distribution. We choose r, s in lemma B.2 as $c_{(1-\alpha)/4}, c_{1-\alpha/4}$, here $F(c_z) - F(-c_z) = z, \forall z \in (0, 1)$. From (B.5), (B.29) and (B.30), for sufficiently small $\xi > 0$, with probability tending to 1,

$$\begin{aligned} d_{1-\gamma}^*(c_{1-\alpha}^*) &\leq \sup_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} d_{1-\gamma}^*(x) \leq \sup_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1-\gamma+\xi) \\ \text{and } d_{1-\gamma}^*(c_{1-\alpha}^*) &\geq \inf_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} d_{1-\gamma}^*(x) \geq \inf_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi) \end{aligned} \quad (\text{B.31})$$

Define

$$\bar{d} = \sup_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1-\gamma+\xi) \text{ and } \underline{d} = \inf_{x \in [c_{1-\alpha-2\xi}, c_{1-\alpha+2\xi}]} \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi) \quad (\text{B.32})$$

From (B.29), with probability tending to 1,

$$c^*(1-\alpha, 1-\gamma) \leq c_{1-\alpha+\frac{\bar{d}}{\sqrt{n}}}^* \leq c_{1-\alpha+\frac{\bar{d}}{\sqrt{n}}+2\xi} \text{ and } c^*(1-\alpha, 1-\gamma) \geq c_{1-\alpha+\frac{\underline{d}}{\sqrt{n}}}^* \geq c_{1-\alpha+\frac{\underline{d}}{\sqrt{n}}-2\xi} \quad (\text{B.33})$$

Define $\bar{c} = c_{1-\alpha+\frac{\bar{d}}{\sqrt{n}}+2\xi}$, and $\underline{c} = c_{1-\alpha+\frac{\underline{d}}{\sqrt{n}}-2\xi}$. From assumption 1 and 4, c_α is continuous in $\alpha \in (0, 1)$; and $\mathcal{U}(x)$ is continuous in \mathbf{R} . Define \mathcal{S} as in (4.2). Then

$$\begin{aligned} &\sqrt{n} \left(\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c^*(1-\alpha, 1-\gamma) \right) - (1-\alpha) \right) \\ &= \mathcal{S}(c_{1-\alpha}) + (\mathcal{S}(c^*(1-\alpha, 1-\gamma)) - \mathcal{S}(c_{1-\alpha})) \\ &+ \sqrt{n} \left(G^*(c^*(1-\alpha, 1-\gamma)) - (1-\alpha + \frac{d_{1-\gamma}^*(c_{1-\alpha}^*)}{\sqrt{n}}) \right) \\ &+ \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma) + \left(d_{1-\gamma}^*(c_{1-\alpha}^*) - \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma) \right) \end{aligned} \quad (\text{B.34})$$

we choose $r = \underline{c}$ and $s = \bar{c}$ in lemma B.3. With probability tending to 1

$$\begin{aligned} &|\sqrt{n} \left(G^*(c^*(1-\alpha, 1-\gamma)) - (1-\alpha + \frac{d_{1-\gamma}^*(c_{1-\alpha}^*)}{\sqrt{n}}) \right)| \\ &\leq \sqrt{n} \left(G^*(c^*(1-\alpha, 1-\gamma)) - G^*\left(c^*(1-\alpha, 1-\gamma) - \frac{1}{n}\right) \right) < \xi \end{aligned} \quad (\text{B.35})$$

We choose a positive integer $m > \bar{c} + 1$. From (A.67) and lemma A.1, with probability tending to 1,

$$\begin{aligned} &|\mathcal{S}(c^*(1-\alpha, 1-\gamma)) - \mathcal{S}(c_{1-\alpha})| \leq \sup_{x \in [\underline{c}, \bar{c}]} |\mathcal{S}(x) - \mathcal{S}(c_{1-\alpha})| \\ &\leq 2 \sup_{x \in [\underline{c}, \bar{c}]} \left| \mathcal{S}(x) - \left(\tilde{M}_m\left(\frac{x+m}{2m}\right) - \tilde{M}_m\left(\frac{-x+m}{2m}\right) \right) \right| + 2 \sup_{y, z \in [0, 1], |y-z| \leq \frac{\bar{c}-\underline{c}}{2m}} |\tilde{M}_m(y) - \tilde{M}_m(z)| \\ &\Rightarrow \text{for } \forall \xi > 0, \limsup_{n \rightarrow \infty} \mathbf{P}(|\mathcal{S}(c^*(1-\alpha, 1-\gamma)) - \mathcal{S}(c_{1-\alpha})| > \xi) < \xi \end{aligned} \quad (\text{B.36})$$

For \mathcal{U} is continuous and

$$\begin{aligned} &|d_{1-\gamma}^*(c_{1-\alpha}^*) - \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma)| \\ &\leq |\bar{d} - \sqrt{\mathcal{U}(c_{1-\alpha})} \Phi^{-1}(1-\gamma)| + |\underline{d} - \sqrt{\mathcal{U}(c_{1-\alpha})} \Phi^{-1}(1-\gamma)| \end{aligned} \quad (\text{B.37})$$

with probability tending to 1, we have for $\forall \xi > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\left| \sqrt{n} \left(\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \hat{\beta}| \leq c^*(1-\alpha, 1-\gamma) \right) - (1-\alpha) \right) - \left(\mathcal{S}(c_{1-\alpha}) + \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma) \right) \right| > \xi \right) = 0 \quad (\text{B.38})$$

On one hand, from theorem 4.2, $\forall \xi > 0$, we choose $Z > 0$ such that $\Phi \left(\frac{Z}{\sqrt{\mathcal{U}(c_{1-\alpha})}} \right) - \Phi \left(-\frac{Z}{\sqrt{\mathcal{U}(c_{1-\alpha})}} \right) > 1 - \xi$, we have $\lim_{n \rightarrow \infty} \mathbf{P}(|\mathcal{S}(c_{1-\alpha})| \leq Z) > 1 - \xi$. On the other hand, for any given $\xi \in \mathbf{R}$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathcal{S}(c_{1-\alpha}) + \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma) + \xi \geq 0 \right) = 1 - \Phi \left(-\Phi^{-1}(1-\gamma) - \frac{\xi}{\sqrt{\mathcal{U}(c_{1-\alpha})}} \right) \quad (\text{B.39})$$

Combine with (B.38), we prove theorem 5.2. \square

Proof of corollary 5.1. From theorem 10.1 in [32] and assumption 3, define $\tilde{\epsilon}_i$ and \tilde{r}_i as in (3.2) and (4.1), $\tilde{r}_i = \tilde{\epsilon}_i / (1 - h_i)$ with $h_i = \mathcal{X}_i^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_i$, and $\exists C > 0$ such that $h_i \leq C/n$ for $i = 1, 2, \dots, n$. From Cauchy inequality, for sufficiently large n

$$\begin{aligned} \hat{r}_i &= \frac{\hat{\epsilon}_i}{1 - h_i} + \frac{1}{n} \sum_{j=1}^n \frac{(h_i - h_j) \tilde{\epsilon}_j}{(1 - h_i)(1 - h_j)} \\ \Rightarrow \sum_{i=1}^n (\hat{r}_i - \hat{\epsilon}_i)^2 &\leq \sum_{i=1}^n \frac{2h_i^2 \tilde{\epsilon}_i^2}{(1 - h_i)^2} + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{(h_i - h_j)^2}{(1 - h_i)^2 (1 - h_j)^2} \sum_{j=1}^n \tilde{\epsilon}_j^2 \leq \frac{4C^2}{n^2} \sum_{i=1}^n \tilde{\epsilon}_i^2 + \frac{16C^2}{n^2} \sum_{j=1}^n \tilde{\epsilon}_j^2 \\ \Rightarrow \mathbf{E} \sum_{i=1}^n (\hat{r}_i - \hat{\epsilon}_i)^2 &\leq \frac{4C^2}{n^2} \sum_{i=1}^n \mathbf{E} \tilde{\epsilon}_i^2 + \frac{16C^2}{n^2} \sum_{j=1}^n \mathbf{E} \tilde{\epsilon}_j^2 \leq \frac{20C^2}{n^2} \times (2n\sigma^2 + 2\sigma^2 \sum_{i=1}^n \mathcal{X}_i^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_i) \end{aligned} \quad (\text{B.40})$$

We define a random vector $(\epsilon_1^*, r_1^*) \in \mathbf{R}^2$ having probability mass $1/n$ on $(\hat{\epsilon}_i, \hat{r}_i), i = 1, 2, \dots, n$. We generate i.i.d. random variables $(\epsilon_i^*, r_i^*), i = 1, 2, \dots, n$ and (ϵ_f^*, ξ^*) with the same distribution as (ϵ_1^*, r_1^*) . We denote $\epsilon^* = (\epsilon_1^*, \dots, \epsilon_n^*)^T$ and $r^* = (r_1^*, \dots, r_n^*)^T$. For any given $0 < r < s < \infty, \xi > 0$, we choose $\delta = C/n^{3/4}$ in (A.3) with C a constant. Then define

$$\mathcal{G}^*(x) = \mathbf{P}^* \left(|\mathcal{Y}_f^* - \mathcal{X}_f^T \hat{\beta}^*| \leq x \right), \quad x \in \mathbf{R} \quad (\text{B.41})$$

here we choose $\hat{\tau} = \hat{r}$ in algorithm 5.1. In other words, \mathcal{G}^* plays the same role as G^* , and the only difference is the mechanism for generating bootstrapped random variables.

$$\begin{aligned} \sup_{x \in [r, s]} |G^*(x) - \mathcal{G}^*(x)| &\leq \mathbf{P}^* \left(\left| |\epsilon_f^* - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon^*| - |r_f^* - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T r^*| \right| > \frac{C}{n^{3/4}} \right) \\ &+ \sup_{x \in [r, s]} \mathbf{P}^* \left(x - \frac{C}{n^{3/4}} < |\epsilon_f^* - \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \epsilon^*| \leq x + \frac{C}{n^{3/4}} \right) \leq \frac{4\sqrt{n}}{C^2} \sum_{i=1}^n (\hat{\epsilon}_i - \hat{r}_i)^2 \\ &+ \frac{4\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \mathcal{X}_f}{C^2 \sqrt{n}} \times \sum_{i=1}^n (\hat{\epsilon}_i - \hat{r}_i)^2 + \sup_{x \in [r, s]} \left(G^*(x + \frac{C}{n^{3/4}}) - G^*(x - \frac{C}{n^{3/4}}) \right) \end{aligned} \quad (\text{B.42})$$

and for sufficiently large n ,

$$\begin{aligned} \sup_{x \geq r} |\mathcal{G}^*(x) - (F(x) - F(-x))| &\leq \frac{4\sqrt{n}}{C^2} \sum_{i=1}^n (\hat{\epsilon}_i - \hat{r}_i)^2 + \frac{4\mathcal{X}_f^T (\mathcal{X}^T \mathcal{X} / n)^{-1} \mathcal{X}_f}{C^2 \sqrt{n}} \times \sum_{i=1}^n (\hat{\epsilon}_i - \hat{r}_i)^2 \\ &+ 3 \sup_{x > 0} |G^*(x) - (F(x) - F(-x))| + \sup_{x \geq r} \left(F\left(x + \frac{C}{n^{3/4}}\right) - F\left(x - \frac{C}{n^{3/4}}\right) \right) \\ &+ \sup_{x \geq r} \left(F\left(-x + \frac{C}{n^{3/4}}\right) - F\left(-x - \frac{C}{n^{3/4}}\right) \right) \end{aligned} \quad (\text{B.43})$$

Lemma B.3 and (B.29) imply $\lim_{n \rightarrow \infty} \mathbf{P} \left(\sqrt{n} \sup_{x \in [r, s]} |G^*(x) - \mathcal{G}^*(x)| > \xi \right) = 0$; and $\lim_{n \rightarrow \infty} \mathbf{P} \left(\sup_{x \geq r} |\mathcal{G}^*(x) - (F(x) - F(-x))| > \xi \right) = 0$.

We define $\widehat{\mathcal{F}}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\hat{r}_i \leq x}$, and $\widehat{\alpha}(x)$ as in lemma A.2. For any given $-\infty < r < s < \infty$, $\xi > 0$, and sufficiently large n , lemma A.2 implies

$$\begin{aligned} \sup_{x \in [r, s]} |\widehat{\mathcal{F}}(x) - \widehat{F}(x)| &\leq \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{|\hat{r}_i - \hat{\epsilon}_i| > \frac{C}{n^{3/4}}} + \sup_{x \in [r, s]} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{x - \frac{C}{n^{3/4}} < \hat{\epsilon}_i \leq x + \frac{C}{n^{3/4}}} \\ &\Rightarrow \mathbf{P} \left(\sqrt{n} \sup_{x \in [r, s]} |\widehat{\mathcal{F}}(x) - \widehat{F}(x)| > \xi \right) \leq \frac{2}{\sqrt{n}\xi} \sum_{i=1}^n \mathbf{P} \left(|\hat{\epsilon}_i - \hat{r}_i| > \frac{C}{n^{3/4}} \right) \\ &+ \mathbf{P} \left(\sup_{x \in [r-1, s+1]} |\widehat{\alpha}(x) - \widehat{\alpha}(x - \frac{2C}{n^{3/4}})| > \frac{\xi}{4} \right) \\ &+ \mathbf{P} \left(\sup_{x \in [r-1, s+1]} F'(x) \times \frac{2C}{n^{1/4}} > \frac{\xi}{4} \right) \Rightarrow \lim_{n \rightarrow \infty} \mathbf{P} \left(\sqrt{n} \sup_{x \in [r, s]} |\widehat{\mathcal{F}}(x) - \widehat{F}(x)| > \xi \right) = 0 \end{aligned} \quad (\text{B.44})$$

Here C is an arbitrary large positive constant.

We define $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{S}}$ as in (5.6); define $\Lambda(z) = \mathcal{X}_f^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T z - \frac{1}{n} \sum_{i=1}^n z_i$, $\forall z = (z_1, \dots, z_n)^T \in \mathbf{R}^n$; and define

$$\widehat{\mathcal{N}}(x) = \sqrt{n} \widehat{\mathcal{F}}(x + \Lambda(u^*)) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq x}, \quad \widehat{\mathcal{T}}(x) = \widehat{\mathcal{N}}(x) - \widehat{\mathcal{N}}(-x) \quad (\text{B.45})$$

Here $u^* = (u_1^*, \dots, u_n^*)^T$ are i.i.d. random variables generated by drawing from \widehat{r} with replacement. For any given $0 < r < s < \infty$ and $\xi > 0$,

$$\begin{aligned} \mathbf{P}^* \left(\sup_{x \in [r, s]} |\widehat{\mathcal{S}}(x) - \widehat{\mathcal{T}}(x)| > 4\xi \right) &\leq \mathbf{P}^* \left(\sup_{x \in [r, s]} |\widehat{\mathcal{M}}(x) - \widehat{\mathcal{N}}(x)| > 2\xi \right) + \mathbf{P}^* \left(\sup_{x \in [r, s]} |\widehat{\mathcal{M}}(-x) - \widehat{\mathcal{N}}(-x)| > 2\xi \right) \\ &\leq \mathbf{P}^* \left(\sup_{x \in [r, s]} \sqrt{n} |\widehat{F}(x + \Lambda(\epsilon^*)) - \widehat{\mathcal{F}}(x + \Lambda(u^*))| > \xi \right) + \mathbf{P}^* \left(\sup_{x \in [r, s]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq x} \right| > \xi \right) \\ &+ \mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \sqrt{n} |\widehat{F}(-x + \Lambda(\epsilon^*)) - \widehat{\mathcal{F}}(-x + \Lambda(u^*))| > \xi \right) + \mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq -x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq -x} \right| > \xi \right) \end{aligned} \quad (\text{B.46})$$

If $\sup_{x \in [-s-2, s+2]} \sqrt{n} |\widehat{\mathcal{F}}(x) - \widehat{F}(x)| < \xi/4$ and $\sup_{x, y \in [-s-2, s+2], |x-y| < \delta} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| \leq \xi/8$ with $0 < \delta < 1/8$,

$$\begin{aligned}
& \mathbf{P}^* \left(\sup_{x \in [r, s]} \sqrt{n} |\widehat{F}(x + \Lambda(\varepsilon^*)) - \widehat{\mathcal{F}}(x + \Lambda(u^*))| > \xi \right), \mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \sqrt{n} |\widehat{F}(-x + \Lambda(\varepsilon^*)) - \widehat{\mathcal{F}}(-x + \Lambda(u^*))| > \xi \right) \\
& \leq \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} \sqrt{n} |\widehat{F}(x + \Lambda(\varepsilon^*)) - \widehat{F}(x + \Lambda(u^*))| > \xi/2 \right) + \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} \sqrt{n} |\widehat{F}(x + \Lambda(u^*)) - \widehat{\mathcal{F}}(x + \Lambda(u^*))| > \xi/2 \right) \\
& \leq \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} |\widehat{\alpha}(x + \Lambda(\varepsilon^*)) - \widehat{\alpha}(x + \Lambda(u^*))| > \xi/4 \right) + \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} \sqrt{n} |F(x + \Lambda(\varepsilon^*)) - F(x + \Lambda(u^*))| > \xi/4 \right) \\
& + \mathbf{P}^* (|\Lambda(u^*)| > 1) \leq 2\mathbf{P}^* (|\Lambda(\varepsilon^*)| > \delta/4) + 3\mathbf{P}^* (|\Lambda(u^*)| > \delta/4) + \mathbf{P}^* \left(\sup_{x \in [-s-2, s+2]} \sqrt{n} |F'(x)| \times |\Lambda(\varepsilon^*) - \Lambda(u^*)| > \xi/4 \right)
\end{aligned} \tag{B.47}$$

For

$$\begin{aligned}
\mathbf{E}^* \Lambda(\varepsilon^*)^2 & \leq \frac{2\widehat{\sigma}^2}{n} \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 \right) \\
\text{and } \mathbf{E}^* (\Lambda(u^*) - \Lambda(\varepsilon^*))^2 & \leq \frac{2}{n} \left(\mathcal{X}_f^T \left(\frac{\mathcal{X}^T \mathcal{X}}{n} \right)^{-1} \mathcal{X}_f + 1 \right) \times \frac{1}{n} \sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2
\end{aligned} \tag{B.48}$$

(A.5), (B.40), (B.44) and lemma A.2 imply $\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbf{P}^* \left(\sup_{x \in [r, s]} \sqrt{n} |\widehat{F}(x + \Lambda(\varepsilon^*)) - \widehat{\mathcal{F}}(x + \Lambda(u^*))| > \xi \right) > \xi \right) = 0$;

and $\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \sqrt{n} |\widehat{F}(-x + \Lambda(\varepsilon^*)) - \widehat{\mathcal{F}}(-x + \Lambda(u^*))| > \xi \right) > \xi \right) = 0$. On the other hand, we define $\widetilde{\alpha}^*$ as in lemma A.2; from (A.3), for any $0 < \delta < 1/4$,

$$\begin{aligned}
& \mathbf{P}^* \left(\sup_{x \in [r, s]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq x} \right| > \xi \right), \mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq -x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq -x} \right| > \xi \right) \\
& \leq \mathbf{P}^* \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{|e_i^* - u_i^*| > \frac{\delta}{\sqrt{n}}} > \xi/2 \right) + \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{x - \frac{\delta}{\sqrt{n}} < e_i^* \leq x + \frac{\delta}{\sqrt{n}}} > \xi/2 \right) \\
& \leq \frac{2\sqrt{n}}{\xi} \mathbf{P}^* \left(|e_1^* - u_1^*| > \frac{\delta}{\sqrt{n}} \right) + \mathbf{P}^* \left(\sup_{x \in [-s-2, s+2]} |\widetilde{\alpha}^*(x) - \widetilde{\alpha}^*(x - \frac{2\delta}{\sqrt{n}})| > \xi/4 \right) \\
& + \mathbf{P}^* \left(\sup_{x \in [-s-1, s+1]} \sqrt{n} |\widehat{F}(x) - \widehat{F}(x - \frac{2\delta}{\sqrt{n}})| > \xi/4 \right)
\end{aligned} \tag{B.49}$$

Since $\mathbf{P}^* \left(|e_1^* - u_1^*| > \frac{\delta}{\sqrt{n}} \right) \leq \frac{\sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2}{\delta^2}$, (B.40) and lemma A.2 imply

$\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbf{P}^* \left(\sup_{x \in [r, s]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq x} \right| > \xi \right) > \xi \right) = 0$;

and $\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbf{P}^* \left(\sup_{x \in [r/2, s+1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{e_i^* \leq -x} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{1}_{u_i^* \leq -x} \right| > \xi \right) > \xi \right) = 0$. In particular, $\forall \xi > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbf{P}^* \left(\sup_{x \in [r, s]} |\widehat{\mathcal{F}}(x) - \widehat{F}(x)| > \xi \right) > \xi \right) = 0 \tag{B.50}$$

For $\forall \xi > 0$,

$$\begin{aligned} & \sup_{x \in [r, s], y \in \mathbf{R}} |\mathbf{P}^* \left(\widehat{\mathcal{T}}(x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right)| \leq \mathbf{P}^* \left(\sup_{x \in [r, s]} |\widehat{\mathcal{S}}(x) - \widehat{\mathcal{T}}(x)| > \xi \right) \\ & + 3 \sup_{x \in [r, s], y \in \mathbf{R}} |\mathbf{P}^* \left(\widehat{\mathcal{S}}(x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right)| + \sup_{x \in [r, s], y \in \mathbf{R}} \left(\Phi \left(\frac{y + \xi}{\sqrt{\mathcal{U}(x)}} \right) - \Phi \left(\frac{y - \xi}{\sqrt{\mathcal{U}(x)}} \right) \right) \end{aligned} \quad (\text{B.51})$$

Lemma B.2 implies $\lim_{n \rightarrow \infty} \mathbf{P} \left(\sup_{x \in [r, s], y \in \mathbf{R}} |\mathbf{P}^* \left(\widehat{\mathcal{T}}(x) \leq y \right) - \Phi \left(\frac{y}{\sqrt{\mathcal{U}(x)}} \right)| > \xi \right) = 0$.

Suppose $\frac{1}{8} \min(\alpha, 1 - \alpha) > \xi > 0$. From (B.5) and (B.51), with probability tending to 1, $\forall 2\xi < 1 - \gamma < 1 - \xi, r \leq x \leq s, \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma - 2\xi) \leq D_{1-\gamma}^*(x) \leq \sqrt{\mathcal{U}(x)} \times \Phi^{-1}(1 - \gamma + \xi)$. We define $c_z, z \in (0, 1)$ and \underline{d}, \bar{d} as in the proof of theorem 5.2. We choose $r = c_{(1-\alpha)/8} > 0$ in (B.43), with probability tending to 1, $c_{\tau-2\xi} \leq C_\tau^* \leq c_{\tau+\xi}, \forall (1-\alpha)/8 + 2\xi < \tau < 1 - 2\xi$. In particular, this implies $c_{1-\alpha-2\xi} \leq C_{1-\alpha}^* \leq c_{1-\alpha+\xi}$, and $\underline{d} \leq D_{1-\gamma}^*(C_{1-\alpha}^*) \leq \bar{d}$. We choose $r = c_{(1-\alpha)/8}$ and $s = c_{1-\alpha+4\xi}$ in (B.42) and lemma B.3, $C^*(1 - \alpha, 1 - \gamma) \leq C_{1-\alpha+\frac{\bar{d}}{\sqrt{n}}}^* \leq c_{1-\alpha+\frac{\bar{d}+2\xi}{\sqrt{n}}}^*$; and $C^*(1 - \alpha, 1 - \gamma) \geq C_{1-\alpha+\frac{\underline{d}}{\sqrt{n}}}^* \geq c_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}^*$. We define \mathcal{S} and \mathcal{U} as in (4.2) and (3.4), since

$$\begin{aligned} & |\sqrt{n} \left(\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \widehat{\beta}| \leq C^*(1 - \alpha, 1 - \gamma) \right) - (1 - \alpha) \right) - \left(\mathcal{S}(c_{1-\alpha}) + \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1 - \gamma) \right)| \\ & \leq |\sqrt{n} \left(\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \widehat{\beta}| \leq c_{1-\alpha+\frac{\bar{d}+2\xi}{\sqrt{n}}}^* \right) - (1 - \alpha) \right) - \left(\mathcal{S}(c_{1-\alpha}) + \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1 - \gamma) \right)| \\ & + |\sqrt{n} \left(\mathbf{P}^* \left(|\mathcal{Y}_f - \mathcal{X}_f^T \widehat{\beta}| \leq c_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}^* \right) - (1 - \alpha) \right) - \left(\mathcal{S}(c_{1-\alpha}) + \sqrt{\mathcal{U}(c_{1-\alpha})} \times \Phi^{-1}(1 - \gamma) \right)| \end{aligned} \quad (\text{B.52})$$

Replace $c^*(1 - \alpha, 1 - \gamma)$ in (B.34) to (B.36) by $c_{1-\alpha+\frac{\bar{d}+2\xi}{\sqrt{n}}}^*$ and $c_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}^*$, and set $\xi \rightarrow 0$, we prove (5.10). \square

C. Results used in the paper

This paper uses many results from the stochastic process and some results from the optimal transport. Statisticians may not be familiar with them. To make the paper self-contained, this section quotes the frequently used theorems from textbooks and papers. However, we cannot explain the background of each theorem in detail. So we encourage the readers to look through those materials if possible.

LEMMA C.1 (theorem 13.5, Billingsley [6]) Suppose that

$$(X_{t_1}^n, \dots, X_{t_k}^n) \rightarrow_{\mathcal{L}} (X_{t_1}, \dots, X_{t_k}) \quad (\text{C.1})$$

for any points t_i ; that

$$X_1 - X_{1-\delta} \rightarrow_{\mathcal{L}} 0 \text{ as } \delta \rightarrow 0, \delta > 0 \quad (\text{C.2})$$

and that for any $r \leq s \leq t, n \geq 1, \lambda > 0$,

$$\mathbf{P}[|X_s^n - X_r^n| \wedge |X_t^n - X_s^n| \geq \lambda] \leq \frac{1}{\lambda^{4\beta}} [F(t) - F(r)]^{2\alpha} \quad (\text{C.3})$$

where $\beta \geq 0$, $\alpha > 1/2$ and F is a non-decreasing, continuous function on $[0, 1]$. Then $X^n \rightarrow_{\mathcal{L}} X$

LEMMA C.2 (theorem 13.6, Billingsley [6]) There exists in \mathbf{D} (see section 3) a random element with finite-dimensional distributions $\mu_{t_1 \dots t_k}$, provided these distributions are consistent (i.e., satisfy the consistency conditions of Kolmogorov's existence theorem); provided that, for $t_1 \leq t \leq t_2$,

$$\mu_{t_1 t_2}[(u_1, u, u_2) : |u - u_1| \wedge |u_2 - u| \geq \lambda] \leq \frac{1}{\lambda^{4\beta}} (F(t_2) - F(t_1))^{2\alpha} \quad (\text{C.4})$$

where $\beta \geq 0$, $\alpha > 1/2$, and F is a non-decreasing, continuous function on $[0, 1]$; and provided that

$$\lim_{h \rightarrow 0, h > 0} \mu_{t, t+h}[(u_1, u_2) : |u_2 - u_1| \geq \varepsilon] = 0, \quad 0 \leq t \leq 1 \quad (\text{C.5})$$

LEMMA C.3 (theorem 2.3 in Hahn [14]) Let f be a nonnegative function on $[0, 1]$ which is nondecreasing in a neighborhood of 0. Let $X(t)$ be a stochastic process such that for some $r \geq 1$, $\mathbf{E}|X(t) - X(s)|^r \leq f(|t - s|)$. If

$$\int_0^1 y^{-(r+1)/r} f^{1/r}(y) dy < \infty \quad (\text{C.6})$$

then there exists a nondecreasing function ϕ on $[0, 1]$ with $\phi(0) = 0$, which depends only on f , and a random variable A such that $\mathbf{E}|A|^r < \infty$ and

$$|\tilde{X}(s) - \tilde{X}(t)| \leq A\phi(|t - s|) \quad (\text{C.7})$$

Moreover, $\|A\|_r$ is bounded above by a constant depending only on f and ϕ . Here \tilde{X} is a separable version of X .

LEMMA C.4 (3.8, page 348 in Jacod and Shiryaev [17]) Assume that $X^n \rightarrow_{\mathcal{L}} X$ and that $\mathbf{P}(X \in C) = 1$, where C is the continuity set of the function $h : E \rightarrow E'$. Then

- i. If $E' = \mathbf{R}$ and h is bounded, then $\mathbf{E}h(X^n) \rightarrow \mathbf{E}h(X)$;
- ii. If E' is Polish, then $h(X^n) \rightarrow_{\mathcal{L}} h(X)$.

LEMMA C.5 (theorem 3.1 in Rao [30]) Let \mathcal{A} be a class of continuous functions possessing the following properties: 1. \mathcal{A} is uniformly bounded, i.e., \exists a constant $M > 0$ such that $|f(x)| \leq M$ for all $f \in \mathcal{A}$ and all x ; 2. \mathcal{A} is equi-continuous. If μ_n, μ satisfies $\mu_n \rightarrow_{\mathcal{L}} \mu$, then

$$\lim_{n \rightarrow \infty} \sup_{f \in \mathcal{A}} \left| \int f d\mu_n - \int f d\mu \right| = 0 \quad (\text{C.8})$$

LEMMA C.6 (theorem 6.2.1 in Koul [18]) Suppose that the model $\mathcal{Y} = \mathcal{X}\beta + \varepsilon$ holds true. In addition suppose $(\mathcal{X}^T \mathcal{X})^{-1}$ exists, $\max_{i=1, \dots, n} \mathcal{X}_i^T (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}_i = o(1)$ and F has uniform continuous density f . Suppose $\hat{\beta}$ is an estimator of β satisfying

$$|A^{-1}(\hat{\beta} - \beta)|_2 = O_p(1) \quad (\text{C.9})$$

then

$$\sup_{t \in [0, 1]} |W_1(t, \hat{\beta}) - W_1(t, \beta) - q_0(t)\sqrt{n} \times \overline{\mathcal{X}}_n^T A A^{-1}(\hat{\beta} - \beta)| = o_p(1) \quad (\text{C.10})$$

here $q_0(t) = f(F^{-1}(t))$, $W_1(t, s) = \sqrt{n}(H_n(F^{-1}(t), s) - t)$, $H_n(y, s) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\mathcal{Y}_i \leq y + \mathcal{X}_i^T s}$ and $A = (\mathcal{X}^T \mathcal{X})^{-1/2}$. $|\cdot|_2$ is the vector 2 norm in the Euclidean space.

LEMMA C.7 (theorem 6.9, Villani [41]) Let (\mathcal{X}, d) be a Polish space, and $p \in [1, \infty)$. Define $P_p(\mathcal{X})$ as the Borel probability measure on \mathcal{X} with finite moments of order p . Then the Wasserstein distance W_p metrizes the weak convergence. In other words, if $(\mu_k)_{k \in \mathbb{N}} \subset P_p(\mathcal{X})$ is a sequence of measures and $\mu \in P(\mathcal{X})$ is another Borel probability measure on \mathcal{X} , then the statement μ_k converges weakly in $P_p(\mathcal{X})$ to μ and $W_p(\mu_k, \mu) \rightarrow 0$ are equivalent. Here $W_p(\mu_k, \mu)$ is the Wasserstein distance (see lemma B.1). The weakly convergence in $P_p(\mathcal{X})$ means $\exists x_0 \in \mathcal{X}$ such that

$$\mu_k \rightarrow_{\mathcal{L}} \mu \text{ and } \int d(x_0, x)^p d\mu_k \rightarrow \int d(x_0, x)^p d\mu \quad (\text{C.11})$$