# Orthogonal Bootstrap: Efficient Simulation of Input Uncertainty

**Kaizhao Liu** [1] **Jose Blanchet** [2] **Lexing Ying** [3] **Yiping Lu** [4 5]

## Abstract

Bootstrap is a popular methodology for simulating input uncertainty. However, it can be computationally expensive when the number of samples is large. We propose a new approach called **Orthogonal Bootstrap** that reduces the number of required Monte Carlo replications. We decomposes the target being simulated into two parts: the *non-orthogonal part* which has a closed-form result known as Infinitesimal Jackknife and the *orthogonal part* which is easier to be simulated. We theoretically and numerically show that Orthogonal Bootstrap significantly reduces the computational cost of Bootstrap while improving empirical accuracy and maintaining the same width of the constructed interval.

## 1. Introduction

The input uncertainty problem (Lam & Qian, 2018) manifests as the propagation of statistical noise from input models, typically derived and calibrated from data, into the subsequent output analysis. This noise can significantly impact the accuracy and reliability of the resulting conclusions, necessitating robust strategies for quantification and mitigation. Bootstrap (Stine, 1985; Efron, 1992a; Efron & Tibshirani, 1994) is a non-parametric method that uses random resampling with replacement to estimate this uncertainty. In this paper, we consider using Bootstrap to estimate the mean and variance of a function with input uncertainty, which has wide application in debiasing the functional estimation (Quenouille, 1949; Efron, 1982; Adams et al., 1971; Cordeiro et al., 2014; Etter & Ying, 2020; Jiao & Han, 2020; Koltchinskii & Zhilova, 2021; Koltchinskii, 2022; Zhou et al., 2021a; Etter & Ying, 2021; Ma & Ying, 2022), improving worst group generalization (Sagawa et al., 2020; Nguyen et al., 2022) and, most influential, constructing the confidence interval (Tukey, 1958; Stine, 1985; Efron, 1992a; Efron & Tibshirani, 1994).

Despite its benefits, Bootstrap is computationally demanding. Ideally, to simulate the Bootstrap estimation perfectly, an infinite number of Monte Carlo replications of resamples is required. In practice, only a finite number of Monte Carlo replications can be actually performed, and this introduce an error in the Bootstrap estimation (that can be characterized by its variance), which we call the *simulation error*. Another source of error in the Bootstrap estimation arises from the randomness in the data. The former can be controlled by increasing the number of Monte Carlo replications, while the latter can not be controlled if the samples are given. To obtain a reasonable Bootstrap estimation, the number of Monte Carlo replications required should at least ensure that the simulation error is smaller than the error arising from the data. In most applications, the number of Monte Carlo replications required scales up with the number of data points (Lam & Qian, 2018; 2022), where each replication can involve expensive optimization procedures to refit models, making it extremely challenging to scale Bootstrap resampling to large datasets. This drawback also occurs in bagging (Wager et al., 2014).

To address these issues, we propose an alternative method called **Orthogonal Bootstrap**. We use semiparametric techniques to reduce the number of required Monte Carlo replications for stochastic simulation under input uncertainty. Our method is inspired by recently proposed double/orthogonal machine learning (Foster & Syrgkanis, 2019; Chernozhukov et al., 2018; 2022b;a) which utilize the influence function (Cook & Weisberg, 1982; Efron, 1992b) to debias parameter estimates in the presence of nuisance parameters. Suppose we want to estimate the uncertainty of the output using Bootstrap. We can regard the Bootstrap method as a two-stage estimator. In the first stage, we generate resample distributions by resampling the original data with replacement and calculate the output for each resampled distribution. In the second stage, we simply calculate the variance of the outputs from the first stage. We consider the first stage to be "nuisance" because our primary interest is in the final

[*]Equal contribution [1]Department of Mathematics, Peking University, Beijing, China [2]Department of Management Science and Engineering, Stanford University [3]Department of Mathematics, Stanford University [4]Courant Institute of Mathematical Sciences, New York University [5]Department of Industrial Engineering and Management Sciences, Northwestern University. Correspondence to: Kaizhao Liu <mrzt@stu.pku.edu.cn>, Yiping Lu <yiping.lu@northwestern.edu>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

uncertainty, and we do not need to know the simulation output for each resampled distribution. Motivated by this, we show in this paper that we can reduce the simulation error of Bootstrap resampling by dealing with the non-orthogonal and orthogonal parts separately. The non-orthogonal part enjoys a closed form result using the influence function (Rousseeuw et al., 2011; Cook & Weisberg, 1980; Koh & Liang, 2017), also known as Infinitesimal Jackknife (IJ) (Jaeckel, 1972; Giordano et al., 2019b;a; Lu et al., 2020; Alaa & Van Der Schaar, 2020; Abad et al., 2022). For modern machine learning, the influence function can be calculated efficiently using implicit Hessian-vector products and is much faster than retraining the model (Cook & Weisberg, 1980; Koh & Liang, 2017; Giordano et al., 2019b). Note that when only calculating the non-orthogonal part, *i.e.* using the IJ method, the final variance estimate tends to be conservative (Efron & Stein, 1981; Efron, 1992b). In Orthogonal Bootstrap, we further simulate the orthogonal part to correct the IJ estimation. Thus, our method provides the same result as the Bootstrap method in contrast to the biased Jackknife estimator. This enables our method to enjoy the accuracy and effectiveness of the Bootstrap method such as higher order coverage for confidence interval construction (Hall, 1986; Efron, 1992b; Hall, 2013) while enjoying similar computational cost as the IJ method. We also remark here that, interestingly, our method can also be understood as using IJ as a control variate in the original Bootstrap method.

### 1.1. Related Work

Our paper uses a similar but not identical setting as in (Lam & Qian, 2022) for the input uncertainty problem. The authors also investigated the simulation effort required by performing Bootstrap. They showed how the total required simulation effort can be reduced from an order bigger than the data size in the conventional approach to an order independent of the data size via subsampling. However, their setting involves two layer of simulation in Bootstrap, while we consider simple plug-in estimator so we consider only one layer of simulation. Interestingly, although their subsampling techniques do not involve influence functions, they leveraged influence functions when proving theoretical results.

A relevant baseline for our paper is the recently proposed Cheap Bootstrap (Lam, 2022). The method also aims to maintain desirable statistical guarantees of confidence interval coverage with minimal resampling effort as low as one Monte Carlo replication. However, Cheap Bootstrap enlarges the confidence interval $t_{B,1-\alpha/2}/z_{1-\alpha/2} > 1$ times, where $t_{B,1-\alpha/2}$ is the $(1-\alpha/2)-$quantile of $t_B$, the student $t-$distribution with degree of freedom $B$ where $B$ is the number of Monte Carlo replications, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)-$quantile of standard normal. When the number of Monte Carlo replications is constrained, the Cheap Bootstrap will provide a confidence interval with a huge width mean. By contrast, our proposed Orthogonal Bootstrap provides a confidence interval with the same expected width mean as the Standard Bootstrap, even when the number of Monte Carlo replications is constrained. At the same time, the Orthogonal Bootstrap can be used beyond confidence interval construction. It can be used for all settings when the Bootstrap technique is applied, for example, bias reduction for functional estimation (Quenouille, 1949; Efron, 1982; 1992a; Jiao & Han, 2020; Koltchinskii & Zhilova, 2021; Ma & Ying, 2022).

Another relevant paper to ours is (Kline & Santos, 2012), which proposed a higher order approximation (up to order $O(n^{-1})$) to the Wild Bootstrap (Wu, 1986; Liu, 1988) . However, (Kline & Santos, 2012)'s methodology can only be used to construct confidence intervals for *linear estimators* while orthogonal bootstrap works without assuming any structure of the output functional. Recently, (Zhou et al., 2021a) proposed Higher-Order Statistical Expansion (HODSE) based on the closed-form representation of the Jackknife estimator of ideal degenerate expansion of the target statistical functional. However, these types of constructions can only be used for certain statistical models, such as smooth function-of-mean model or linear regression. In contrast, our method functions as a general procedure for Bootstrap methods that doesn't depend on either the linear structure of the estimator or the smooth function-of-mean model. We achieve this by considering the Taylor expansion over the input distribution as a control variate for the Bootstrap estimator.

### 1.2. Contribution

Our contributions are summarized as follows:

- We propose Orthogonal Bootstrap, a brand new Bootstrap method that provides the same expected simulation result as the original Standard Bootstrap but with reduced simulation effort via separately treating the non-orthogonal part and orthogonal part.
- Theoretically, we show that the Orthogonal Bootstrap can provably reduce the simulation error so that the required number of Monte Carlo replications required decreased from $\Omega(n)$ to $O(1)$, under the mild assumption that the performance measure has a continuous Fréchet derivative under the Kernel Maximum Mean Discrepancy (MMD) distance. As far as the authors know, we are the first paper to link the Bootstrap simulation error with differentiability in Kernel MMD.
- Empirically, we show that Orthogonal Bootstrap can significantly improve the result on both simulated and real datasets when the number of Monte Carlo replica-

tions is limited.

### 1.3. Organization of the Paper

We organize our paper as follows. In Section 2, we demonstrate how Orthogonal Bootstrap can be used to reduce the bias of a statistical estimator. In Section 3, we demonstrate how Orthogonal Bootstrap can be used to quantify the uncertainty of a statistical estimator. In Section 4, we present some numerical examples on both simulated and real datasets.

### 1.4. Notation

Throughout the paper, we adopt the following notation conventions (Lam & Qian, 2022). The symbol $n$ represents the sample size. The notation $F$ refers to a distribution, $\hat{F}$ stands for the empirical distribution created through sampling from $F$, and $\hat{F}^b$ signifies the empirical distribution created through bootstrap resampling from $\hat{F}$. In particular, superscripts are used to distinguish bootstrapped distributions and subscripts to denote distinct input distributions; for instance, $\hat{F}_i^b$ denotes the empirical distribution constructed from the $b$-th bootstrap resample of the $i$-th empirical distribution $\hat{F}_i$. For a random variable $X_{i,j}$ with double subscripts, the first subscript denotes distinct input distributions, and the second subscript denotes diverse samples drawn from the $i$-th input distribution. We use $\mathbb{E}_*$ and $\mathrm{Var}_*$ to denote the expectation and variance over the bootstrap resamples from the data, conditional on the original data $\hat{F}$. That is, $\mathbb{E}_*$ and $\mathrm{Var}_*$ only accounts for the randomness in simulation, *i.e.* the simulation error. We use $\mathbb{E}$ and Var to denote the expectation and variance over the data distribution $F$. We also use the standard $O_p$ notations: $\Theta_p(\cdot)$, $O_p(\cdot)$, $\Omega_p(\cdot)$ in probability statement about the data distribution $F$, to only hide constants that do not depend on $n$.

## 2. Debiasing via Orthogonal Bootstrap

In this section, we study the problem of estimating function/functional values when uncertainty exists on the input value (Quenouille, 1949; Efron, 1982; 1992a; Jiao & Han, 2020; Koltchinskii & Zhilova, 2021; Koltchinskii, 2022; Zhou et al., 2021a; Etter & Ying, 2020; 2021; Ma & Ying, 2022). We first formulate the problem of simulating with input uncertainty (Song et al., 2014; Lam & Qian, 2018), then describe the motivation of our Orthogonal Bootstrap method.

Suppose one aim to estimate a generic performance measure $\phi(F_1, \cdots, F_m)$ depend on $m$ independent input distributions $F_1, \cdots, F_m$. We consider the setting when the input distribution is unknown and only $n_i$ i.i.d. generated data $\{X_{i,1}, \cdots, X_{i,n_i}\}$ from distribution $F_i$ is available, forming the empirical distributions $\hat{F}_i := \sum_{j=1}^{n_i} \delta_{X_{i,j}}(x)/n_i$.

The Bootstrap (Efron & Tibshirani, 1994) methods simulates $B \in \mathbb{Z}^+$ Monte Carlo replications. In each Monte Carlo replication, resampling with replacement is performed independently and uniformly from $\{X_{i,1}, \cdots, X_{i,n_i}\}$, repeated $n_i$ times. This process yields sets $\{X_{i,1}^b, \cdots, X_{i,n_i}^b\}$ for each replications $b = 1, \cdots, B$. Then according to the bootstrap principle, we can estimate $\mathbb{E}\phi(\hat{F}_1, \cdots, \hat{F}_m)$ by the bootstrap mean

$$\mathbb{E}_*\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) = \lim_{B\to\infty} \frac{1}{B}\sum_{i=1}^{B} \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b),$$

where we recall from the notation section that $\mathbb{E}_*$ is the expectation with respect to resampling conditional on the original data and $\hat{F}_i^b = \sum_{j=1}^{n_i} \delta_{X_{i,j}^b}(x)/n_i$ is the resampled empirical distribution.

The simulated bootstrap mean can be used for debiasing the naive estimation of function/functional values when uncertainty exists on the input value (Quenouille, 1949; Efron, 1982; 1992a; Jiao & Han, 2020; Koltchinskii & Zhilova, 2021; Koltchinskii, 2022; Zhou et al., 2021a; Etter & Ying, 2020; 2021; Ma & Ying, 2022). The debiasing procedure appears in many applications, including inverse a noisy elliptic system (Etter & Ying, 2020; 2021), optimal stopping (Zhou et al., 2021b), online learning (Chen et al., 2022) and stochastic optimization (Ma & Ying, 2022; Li, 2020). (Ma & Ying, 2022) showed that the estimation can be improved from the naive estimator with $\Omega(n)$ times of bootstrap simulation. Bootstrap function/functional estimates in real applications is computationally expensive, as it requires retraining a machine learning model. Therefore, bootstrapping a large number of times (proportional to the number of data points) is impractical, and Bootstrap a limited number of times can lead to high variance and bad performance (Lam & Qian, 2018).

In this paper, we propose methods to reduce resampling effort while maintaining the expected accuracy. To achieve this, suppose $\nabla\phi$ is the von Mises derivative (defined in Section A.1 in the appendix) of the statistical functional $\phi$. Then we decompose the bootstrap simulation target into the non-orthogonal part and the orthogonal part

$$\begin{aligned}
\mathbb{E}_*\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) &= \underbrace{\mathbb{E}_*\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]}_{=0} \\
&+ \mathbb{E}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) - \nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right) \\
&= \mathbb{E}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) - \underbrace{\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]}_{\text{Control Variate}}\right)
\end{aligned}$$

To compute $\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m]$, we use the influence function technique (Cook & Weisberg, 1980; Koh & Liang, 2017; Giordano et al., 2019a). Noting that typically $\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)[\tilde{F}_1 - \hat{F}_1, \cdots, \tilde{F}_m - \hat{F}_m]$ is a linear functional in $(\tilde{F}_1, \cdots, \tilde{F}_m)$, and the Riesz

representation theorem implies the existence of a mean zero, finite variance functions $(\mathcal{I}_1(X_1), \cdots, \mathcal{I}_m(X_m))$ such that $\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)[\tilde{F}_1 - \hat{F}_1, \cdots, \tilde{F}_m - \hat{F}_m] = \sum_{i=1}^m \mathbb{E}_{X_i \sim \tilde{F}_i} \mathcal{I}_i(X_i)$ (Cook & Weisberg, 1980; Serfling, 2009; Van der Vaart, 2000). As the influence function (the non-orthogonal part) is mean zero, *i.e.* $\mathbb{E}_* \nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m] = 0$, it doesn't need to be simulated. We can also understand the non-orthogonal part as a control variate (Asmussen & Glynn, 2007) to reduce the simulation error of bootstrap simulation. The whole idea is summarized in Figure 1.

Applying the above idea to bias correction, we can simulate the bias $\phi(F_1, \cdots, F_m) - \mathbb{E}\phi(\hat{F}_1, \cdots, \hat{F}_m)$ by

$$\phi(\hat{F}_1, \cdots, \hat{F}_m) - \frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}^b - \sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{I}_i^{\phi}(x_{i,j}^b))$$

The whole procedure is summarized in Algorithm 1.

**Algorithm 1** Debiasing via Orthogonal Bootstrap

**Input**: A generic performance measure $\phi(F_1, \cdots, F_m)$, i.i.d samples $\{X_{i,1}, \cdots, X_{i,n_i}\} \in \mathbb{R}^{d_i}$ of $F_i$, and influence function $\mathcal{I}_i^{\phi}$ of $\phi$ respect to $\hat{F}_i$

**Output**: Estimation of $\phi(F_1, \cdots, F_m)$

$\hat{\phi} \leftarrow \phi(\hat{F}_1, \cdots, \hat{F}_m)$, where $\hat{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\delta_{X_{i,j}}$

**for** b=1:B **do**

  **for** i=1:m **do**

    Sample $\{x_{i,1}^b, \cdots, x_{i,n_i}^b\}$ i.i.d from $\hat{F}_i$

  **end for**

  $\hat{\phi}^b \leftarrow \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)$, where $\hat{F}_i^b = \frac{1}{n_i}\sum_{j=1}^{n_i}\delta_{x_{i,j}^b}$

**end for**

Estimate $\phi(F_1, \cdots, F_m)$ by

$$2\hat{\phi} - \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{I}_i^{\phi}(x_{i,j}^b)\right) \quad (1)$$

### 2.1. Provable Improvement of Orthogonal Bootstrap

In this section, we show our Orthogonal Bootstrap method can provably reduce the required Monte Carlo replications under mild assumption. For simplicity, we assume $n_1 = n_2 = \cdots = n_m = n$ in our theoretical investigation. We assume our simulation functional has a continuous Fréchet gradient in tht under the Kernel Maximum Mean Discrepancy (MMD) distance (Muandet et al., 2017). For background in kernel mean embeddings, we refer to Section A.3 in the appendix. We first rewrite the simulation functional in terms of the kernel mean embeddings, *i.e.*

$$\phi(F_1, \cdots, F_m) = h(\mu_1(F_1), \cdots, \mu_m(F_m)) \quad (2)$$

where $\mu_i$ are kernel mean embeddings using kernel $k_i$ (Muandet et al., 2017), $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$ where $\mathcal{H}_i$ is the reproducing kernel Hilbert space respect to the kernel $k_i$ and $h : \mathcal{H} \to \mathbb{R}$ is a functional on $\mathcal{H}$. Denote $\mu := \mu_1 \times \cdots \mu_m$ and $F := F_1 \times \cdots F_m$ for simplicity.

**Assumption 2.1.** There exists kernel mean embeddings $\mu_i : \mathcal{F}_i \to \mathcal{H}_i$ which maps $F_i$ into $\mathcal{H}_i$ for $i = 1, \cdots, m$. Moreover, for all $i = 1, \cdots, m$, $\mathbb{E}k_i(X_i, X_i)^4 < \infty$ and $\mathbb{E}k_i(X_i, Y_i)^4 < \infty$ where $X_i, Y_i$ are independent samples from $F_i$.

For a functional on $\mathcal{H}$, we say that it is of class $C^1$ if its Fréchet derivative exists and is continuous.

**Assumption 2.2.** The non-constant functional $h : \mathcal{H}_1 \times \cdots \times \mathcal{H}_m \to \mathbb{R}$ is of class $C^1$ and its derivative is Lipschitz in the sense that $|Dh(x_1)(v) - Dh(x_2)(v)| \leq L\|x_1 - x_2\|_{\mathcal{H}}\|v\|_{\mathcal{H}}$. Moreover, $\|\partial_i h(\mu(F))^4\|_{\mathcal{H}_i} < \infty$.

Under the above two assumption on the performance measure 2, we have

**Theorem 2.3.** *Let $X_{ob}$ be the Orthogonal Bootstrap estimator defined in Equation (1) and $X_{sb}$ be the Standard Bootstrap estimator defined by $X_{sb} := 2\hat{\phi} - \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^b$. Under Assumption 2.1 and Assumption 2.2, if the number of Monte Carlo replications $B \geq Cn^{\alpha}$ for some absolute constant $C > 0$ and $\alpha \geq 0$, then the simulation error for the Orthogonal Bootstrap estimator satisfies $Var_*(X_{ob}) = O_p(\frac{1}{n^{2+\alpha}})$ and the simulation error for the Standard Bootstrap estimator satisfies $Var_*(X_{sb}) = \Theta_p(\frac{1}{n^{1+\alpha}})$.*

The theorem is proved by combining Theorem B.6 and Theorem B.8 with Theorem B.18 in the appendix. Here we provide an informal proof to illustrate our idea.

**Informal Proof:** The complexity of bootstrap simulation is related to the variance of the following random variable

$$\xi = \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) - \nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right].$$

The random variable is at the scale of $O_p(\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m)^2$, *i.e.* $O_p(n^{-1})$. Thus its variance is at scale $O_p(n^{-2})$. Thus to achieve simulation error of order $O_p(n^{-\alpha})$, one needs $O(n^{\alpha-2})$ Monte Carlo replications. □

*Remark* 2.4. To debias a functional estimation, one typically needs the simulation error to be $O_p(n^{-2})$ (see Theorem 2 and its proof in (Ma & Ying, 2022)). According to our theorem, Standard Bootstrap needs $\Omega(n)$ Monte Carlo replications but Orthogonal Bootstrap only needs $O(1)$ Monte Carlo replications.

## 3. Variance Estimation via Orthogonal Bootstrap

In this section, we discuss how the idea of Orthogonal Bootstrap can be used to accelerate bootstrap simulation for

estimating the variance, which provides a versatile non-parametric method for constructing confidence intervals and prediction intervals without detailed model knowledge. Following the same setting of the previous section, suppose one aims to estimate a generic performance measure $\phi(F_1, \cdots, F_m)$. To quantify the uncertainty of the plug-in estimator $\phi(\hat{F}_1, \cdots, \hat{F}_m)$, one needs to know its variance, which can be simulated by $\text{Var}_*(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b))$ according to the bootstrap principle (recall that Var without subscript denotes the simulation variance conditioned on the original data). Similar to the previous section, we decompose the variance into the variance of the non-orthogonal part, the variance of the orthogonal part, and their cross-covariance as in Equation (3) as follows.

$$\begin{aligned}\text{Var}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)\right) &= \underbrace{\text{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right)}_{\text{closed form}} \\ &+ \text{Var}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) - \nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right) \\ &+ 2\text{Cov}_*\Big(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right], \\ &\quad \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) - \nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\Big)\end{aligned} \tag{3}$$

The variance of the non-orthogonal part enjoys a closed-form representation using the influence function as follows.

**Lemma 3.1.** *$\text{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right)$ $= \sum_{i=1}^m \frac{1}{n_i^2}\sum_{j=1}^{n_i}(\mathcal{I}_i^\phi(X_{i,j}))^2$*

For the proof, see Lemma B.3 in the appendix. Based on this observation and the orthogonal and non-orthogonal decomposition, we propose Algorithm 2. We calculate the non-orthogonal part (influence function) using the closed-form representation and simulate the remainder part using the Monte-Carlo method.

Under the same assumption as debiasing, Orthogonal Bootstrap provably improves the required Monte Carlo replications when simulating the variance.

**Theorem 3.2.** *For simplicity, we assume $n_1 = n_2 = \cdots = n_m = n$. Let $X_{ob}$ be the Orthogonal Bootstrap estimator defined in Equation (4) and $X_{sb}$ be the Standard Bootstrap estimator defined by $X_{sb} := \frac{1}{B}\sum_{b=1}^B(\hat{\phi}^b - \overline{\phi})^2$. Under Assumption 2.1 and Assumption 2.2, if the number of Monte Carlo replications $B \geq Cn^\alpha$ for some absolute constant $C > 0$ and $\alpha \geq 0$, then the simulation error for the Orthogonal Bootstrap estimator satisfies $\text{Var}_*(X_{ob}) = O_p(\frac{1}{n^{3+\alpha}})$ and the simulation error for the Standard Bootstrap estimator satisfies $\text{Var}_*(X_{sb}) = \Theta_p(\frac{1}{n^{2+\alpha}})$.*

*Remark* 3.3. To achieve relative consistency in variance estimation, one typically needs the simulation error to be $O_p(n^{-3})$ (see Theorem 1 in (Lam & Qian, 2022)). According to Theorem 3.2, the Standard Bootstrap needs $\Omega(n)$ Monte Carlo replications but the Orthogonal Bootstrap only needs $O(1)$ Monte Carlo replications.

The theorem is proved by combining Theorem B.10 and Theorem B.12 with Theorem B.18 in the appendix.

When the sample size $n$ is small, our Orthogonal Bootstrap estimator for variance can obtain negative values sometimes. An improved Orthogonal Bootstrap estimator (Algorithm 3 in the appendix) should be used in this case. We note here that the improved Orthogonal Bootstrap estimator enjoys the same theoretical properties (Theorem B.15 in the appendix) as the original Orthogonal Bootstrap estimator. See Section B.4 in the appendix for details.

## 4. Numerical Examples

In this section, we test the numerical performances of our Orthogonal Bootstrap and compare it with the Standard Bootstrap. For confidence and prediction interval examples, we also compared our method with the recently proposed Cheap Bootstrap method (Lam, 2022). We demonstrated that our Orthogonal Bootstrap achieved significantly smaller bias and higher empirical coverage probability over the original Bootstrap method when the number of Monte Carlo replications is small. Although Cheap Bootstrap (Lam, 2022) can also provide comparable empirical coverage probability to our method, the mean width of the interval constructed by Cheap Bootstrap is much longer. Our Orthogonal Bootstrap method provides interval with higher empirical coverage probability but achieves the same expected width as the original Bootstrap. The experiment details are left in Section C in the appendix.

### 4.1. Debiasing

We consider four numerical examples following (Ma & Ying, 2022). For all of the examples, we compute the the average bias (BIAS) of the estimates across 1000 experiments. We run the naive estimator without debiasing, Standard Bootstrap, and our Orthogonal Bootstrap for a small number of Monte Carlo replications $B = 2, 3, 4, 5, 6, 7, 8, 9, 10$. To ensure robustness and reliability, we repeat the bootstrap procedure ten times and provide insights through the reporting of quantiles at the 5th, 50th, and 95th percentiles. The results is shown in Figure 2.

**Function of Mean** We simulate the function-of-mean model, namely estimating $\phi = g(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$ for a $d$-dimensional random vector $\mathbf{X}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is a function. We have i.i.d random sample $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ of random vector $\mathbf{X}$. We consider an ellipsoidal estimation problem $g(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_2^2$ and a fourth-order polynomial estimation problem $g(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_2^4$. The underlying distribution is set to be $\mathcal{N}(0.2\mathbf{1_d}, I_d)$. We use $d = 25$ and a sample size $n = 100$. The influence function of $g(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|^2$ is $\mathcal{I}^g(\mathbf{x}) = 2(\mathbf{x}^T\hat{\boldsymbol{\mu}} - \|\hat{\boldsymbol{\mu}}\|^2)$, and the influence of $g(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|^4$ follows by the chain rule.

**Algorithm 2** Variance Estimation via Orthogonal Bootstrap

**Input**: A generic performance measure $\phi(F_1, \cdots, F_m)$, i.i.d samples $\{X_{i,1}, \cdots, X_{i,n_i}\} \in \mathbb{R}^{d_1}$ of $F_i$, and influence function $\mathcal{I}_i^{\phi}$ of $\phi$ respect to $\hat{F}_i$.

**Output**: Estimation of Var$\phi(\hat{F}_1, \cdots, \hat{F}_m)$.

$\hat{\phi} \leftarrow \phi(\hat{F}_1, \cdots, \hat{F}_m)$, where $\hat{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}}$

**for** b=1:B **do**

  **for** i=1:m **do**

    Sample $\{x_{i,1}^b, \cdots, x_{i,n_i}^b\}$ i.i.d from $\hat{F}_i$

  **end for**

  $\hat{\phi}^b \leftarrow \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)$, where $\hat{F}_i^b = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{x_{i,j}^b}$

  Calculate $\hat{\mathcal{I}}^b = \sum_{i=1}^{m} \frac{1}{n_i}\sum_{j=1}^{n_i} \mathcal{I}_i^{\phi}(\tilde{x}_{i,j})$

**end for**

Calculate $\overline{\phi - \mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}^b - \hat{\mathcal{I}}^b)$, $\overline{\mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}\hat{\mathcal{I}}^b$.

Estimate $\text{Var}_{F_1,\cdots,F_m}\phi(\hat{F}_1, \cdots, \hat{F}_m)$ by $S^2$, where

$$S^2 = \sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\mathcal{I}_i^{\phi}(X_{i,j}))^2 + \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right). \tag{4}$$

*Figure 1.* We consider modeling the relationship between resampled distribution and simulation output as nuisance estimation in orthogonal statistical learning. In Orthogonal Bootstrap, we use linear modeling for the nuisance estimation and only focus on the simulation of the orthogonal part (*i.e.* the residual of linear modeling) to reduce the simulation error.

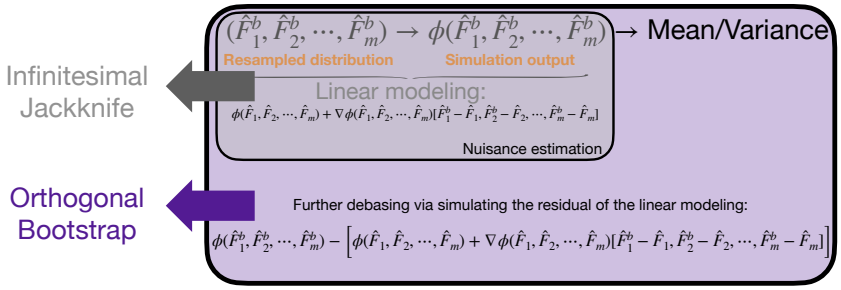


*Table 1.* 95% confidence interval performances with different Bootstrap methods: Standard Bootstrap, Cheap Bootstrap (Lam, 2022) and Orthogonal Bootstrap. We also added the Infinite Jackknife baseline here. The Standard Bootstrap is significantly under-coverage when $B = 2$ to $B = 5$. For example, the Standard Bootstrap method only achieved 59% coverage for estimating the variance of folded normal when only 2 Monte Carlo replications are applied. In contrast, Orthogonal Bootstrap can achieve the target coverage when $B$ is small. Compared with Cheap Bootstrap (Lam, 2022), Orthogonal Bootstrap does not enlarge the length of the constructed confidence interval. Our method also outperforms the Infinitesimal Jackknife.

| Method | $B$ | Variance of Folded normal | | Variance of Double exponential | | Correlation of Bivariate Lognormal | | Linear Regression | |
|---|---|---|---|---|---|---|---|---|---|
| | | Coverage | Width (st. dev.) | Coverage | Width (st. dev.) | Coverage | Width (st. dev.) | Coverage | Width (st. dev.) |
| Standard Bootstrap | 2 | 0.591 | 0.043(0.033) | 0.580 | 0.302(0.238) | 0.566 | 0.112(0.094) | 0.632 | 0.361(0.260) |
| Cheap Bootstrap (Lam, 2022) | 2 | 0.956 | 0.145(0.077) | 0.952 | 1.085(0.606) | 0.936 | 0.385(0.239) | 0.955 | 1.199(0.615) |
| Orthogonal Bootstrap | 2 | 0.952 | 0.076(0.008) | 0.949 | 0.552(0.077) | 0.911 | 0.194(0.063) | 0.946 | 0.615(0.076) |
| Standard Bootstrap | 5 | 0.838 | 0.063(0.024) | 0.834 | 0.454(0.180) | 0.788 | 0.160(0.079) | 0.840 | 0.533(0.191) |
| Cheap Bootstrap (Lam, 2022) | 5 | 0.946 | 0.096(0.032) | 0.945 | 0.674(0.240) | 0.918 | 0.245(0.111) | 0.954 | 0.779(0.250) |
| Orthogonal Bootstrap | 5 | 0.954 | 0.076(0.007) | 0.950 | 0.549(0.076) | 0.913 | 0.195(0.066) | 0.950 | 0.623(0.039) |
| Standard Bootstrap | 10 | 0.906 | 0.069(0.018) | 0.896 | 0.510(0.143) | 0.855 | 0.178(0.073) | 0.905 | 0.576(0.144) |
| Cheap Bootstrap (Lam, 2022) | 10 | 0.951 | 0.084(0.021) | 0.950 | 0.610(0.160) | 0.933 | 0.215(0.082) | 0.958 | 0.693(0.156) |
| Orthogonal Bootstrap | 10 | 0.950 | 0.076(0.007) | 0.955 | 0.548(0.074) | 0.929 | 0.189(0.058) | 0.954 | 0.624(0.028) |
| Infinitesimal Jackknife | | 0.937 | 0.076(0.008) | 0.931 | 0.548(0.075) | 0.899 | 0.191(0.058) | 0.942 | 1.679(0.150) |

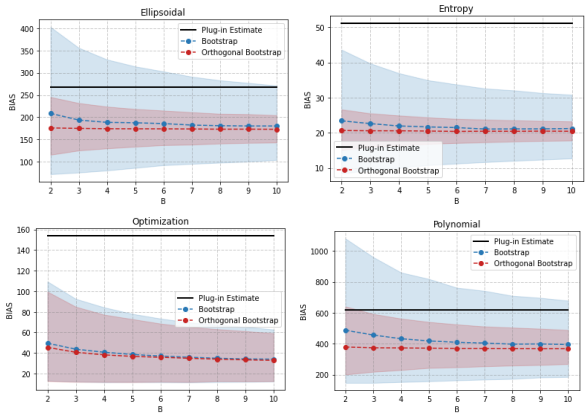


*Figure 2.* Orthogonal Bootstrap can significantly reduce the simulation output for the examples shown in (Ma & Ying, 2022) when the number of Monte Carlo replications is limited. The $x$-axis represents the number of Monte Carlo replications and $y$-axis denotes the bias produced by the estimation. The shaded area represents the 90% quantile interval for repeated simulations. Orthogonal Bootstrap can significantly reduce the simulation error.

**Entropy** We consider estimating entropy for discrete probability distributions, namely estimating $\phi = -\sum_{i=1}^{d} p_i \ln p_i$ where $p = (p_1, \cdots, p_d)$ satisfies $p_i > 0$ and $\sum_{i=1}^{d} p_i = 1$. We generate the groundtruth distribution $p$ from symmetric Dirichlet distribution with parameter $\alpha_i = 1$ for all $i = 1, \cdots, d$. Noisy observations of $p$ are the empirical distributions of single samples from $p$. We use a sample size $n = 1000$ and dimension $d = 100$. The influence function is $\mathcal{I}^{\phi}(x) = \sum_{i=1}^{d}(x_i - \hat{p}_i)(\log \hat{p}_i + 1)$.

**Constrained Optimization Problem** The debiasing method can also be applied to optimization problems with randomness. Here we consider estimating $\phi = \arg\min_{\mathbf{x}} \mathbf{x}^T B\mathbf{x}$ under the constrain $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x} \in \mathbb{R}^p$, $B \in \mathbb{R}^{p\times p}$ is a positive definite matrix, $A \in \mathbb{R}^{d\times p}$, and $\mathbf{b} \in \mathbb{R}^d$ while we only have access to its noisy observations, i.e. $\mathbf{b} \sim \mathcal{N}(\mathbf{b}^*, I_{d\times d})$. We sample $\mathbf{b}^*$ uniformly from the unit sphere $\mathbb{S}^{d-1}$. We use a sample size $n = 10$, dimension $d = 100$ and $p = 200$.

To compute the influence function, we do not solve the quadratic programming problem directly and differentiate. Instead, we utilize a general procedure that can be applied to optimization problems lacking an explicit expression for their solutions by differentiating the KKT conditions of the constrained optimization problem. The details can be found Section C.1.1 in the appendix.

### 4.2. Confidence Interval Construction

In this section, we aim to construct the confidence interval (Hall, 1986; Hall & Martin, 1988; Efron, 1992a) with minimal resampling effort using our Orthogonal Bootstrap technique. The confidence interval constructed by the Standard Bootstrap method is

$$[\hat{\phi} - z_{1-\alpha/2}\sqrt{\mathrm{Var}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)\right)}, \hat{\phi} + z_{1-\alpha/2}\sqrt{\mathrm{Var}_*\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)\right)}],$$

where $z_{1-\alpha/2}$ being the $(1-\alpha/2)-$quantile of the standard normal, $\hat{\phi} = \phi(\hat{F}_1, \cdots, \hat{F}_m)$ is the plug-in estimator of $\phi(F_1, \cdots, F_m)$ and $\hat{F}_i = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_{i,j}}$. We use Orthogonal Bootstrap to estimate $\mathrm{Var}_*(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b))$. The details of the algorithm for confidence interval construction is provided in Algorithm 4 in the appendix.

**Elementary Examples** Following (Lam, 2022), we consider a folded standard normal (*i.e.*,$|N(0,1)|$) and double exponential with rate 1 (*i.e.* $\mathrm{Sgn}\times\mathrm{Exp}(1)$, where $\mathrm{Sgn} = +1$ or $-1$ with equal probability and is independent with $\mathrm{Exp}(1)$). We aim to provide confidence intervals for their variance. The two setups have explicit ground truth, namely $1 - 2/\pi$ and 2, respectively. The influence function of variance is $\mathcal{I}(x) = (x - \hat{\mu})^2 - \hat{\sigma}^2$, where $\hat{\mu}$ is the empirical mean and $\hat{\sigma}^2$ is the empirical variance. The third example, also follows (Lam, 2022), is estimating the correlation of bivariate lognormal (*i.e.*$(e^{Z_1}, e^{Z_2})$, where $(Z_1, Z_2)$ is the bivariate normal with mean zero, unit variance and correlation 0.5). The ground truth of this example is also known as $(e^{3/2} - e)/(e^2 - e)$. The influence function of correlation can be determined by the influence of covariance and the chain rules. We use a sample size $n = 1000$ for all three examples. We run 1000 independent simulations and report the empirical coverage and the mean width of the confidence interval. We run the Standard Bootstrap, Cheap Bootstrap (Lam, 2022) and Orthogonal Bootstrap using a small number of Monte Carlo replications $B = 2, 5, 10$. We set the confidence level $(1 - \alpha)$ as 95%. For each setting, we repeat our experiments 1000 times and report the empirical coverage, interval width mean and standard deviation. The result is shown in Table 1.

**Regression Problem** We apply our method to a linear regression problem in (Sengupta et al., 2016; Lam, 2022). We aim to fit a model $Y = \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$ where we set dimension $d = 50$ and use data $\{(x_{1,i}, \cdots, x_{d,i}, Y_i)\}_{i=1}^{n}$ of size $n = 5000$ to fit the model. We set $\log(X_i) \sim N(0,1)$ and $\epsilon \sim 10 * N(0,1)$ as the data generating process. The influence function is easy to calculate in this case as the estimator is a $M$-estimator. For example, (Cook & Weisberg, 1980; Koh & Liang, 2017) calculate the influence function for M-estimator. Specifically, consider the influence function of the M-estimator $\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$ , where $z_i$ represents training data and labels, $L(z, \theta)$ is the loss function for data $z$ and parameter $\theta$. The influence

*Table 2.* 95% prediction interval performances with different Bootstrap methods: Standard Bootstrap, Cheap Bootstrap (Lam, 2022), and Orthogonal Bootstrap. Results are averaged over 3 random seeds. Our method can achieve 95% coverage with the minimum times of Bootstrap without enlarging the prediction interval length.

| Method | $B$ | Yacht Hydrodynamics | | Energy Efficiency | | Kin8nm | |
|---|---|---|---|---|---|---|---|
| | | Coverage | Width | Coverage | Width | Coverage | Width |
| Standard Bootstrap | 2 | 0.86 | 37.02 | 0.932 | 16.29 | 0.9455 | 0.6216 |
| Cheap Bootstrap (Lam, 2022) | 2 | 1.00 | 82.26 | 1.00 | 36.03 | 1.00 | 1.3746 |
| Orthogonal Bootstrap | 2 | 0.99 | 45.88 | 0.951 | 15.54 | 0.9463 | 0.6224 |
| Standard Bootstrap | 5 | 0.86 | 35.21 | 0.962 | 17.27 | 0.9455 | 0.6154 |
| Cheap Bootstrap (Lam, 2022) | 5 | 0.86 | 46.42 | 0.993 | 22.77 | 0.9772 | 0.8110 |
| Orthogonal Bootstrap | 5 | 0.97 | 42.08 | 0.962 | 17.40 | 0.9455 | 0.6158 |

function for the parameter $\theta$ at point $z$ is

$$\mathcal{I}^{\theta}(z) = \frac{d \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta)}{d\epsilon}\Big|_{\epsilon=0} \quad (5)$$
$$= -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z,\hat{\theta}),$$

where $H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta}^2 L(z_i,\hat{\theta})$ is the empirical hessian of the loss function and is positive definite by assumption. As shown in (Koh & Liang, 2017; Alaa & Van Der Schaar, 2020), the influence function can be solved efficiently via inverse hessian vector product combined with the modern auto-grad systems like TensorFlow (Abadi, 2016) and Pytorch (Paszke et al., 2019). The calculation of the influence function is much faster than retraining the model. In this example, we run the Standard Bootstrap, Cheap Bootstrap (Lam, 2022), and our Orthogonal Bootstrap using a small number of Monte Carlo replications $B = 2, 5, 10$ and report the empirical coverage, interval width mean, and standard deviation for the first coefficient in Table 1.

### 4.3. Prediction Interval Construction for Real Data

In this section, we use Orthogonal Bootstrap to accelerate the pivot Bootstrap method (Contarino et al., 2022) to construct prediction intervals for neural networks. Following (Alaa & Van Der Schaar, 2020), we conduct our experiments on 3 UCI benchmark datasets for regression: yacht hydrodynamics, energy efficiency (Dua & Graff, 2017) and kin8nm. We partition these datasets into distinct training and testing subsets. For the training datasets, we employ two-layer neural networks with a hidden dimension of 100 and a hyperbolic tangent (tanh) activation function. Subsequently, we apply Standard Bootstrap, Cheap Bootstrap, and our proposed Orthogonal Bootstrap to simulate the variance of the resampling step in the pivot Bootstrap method (Contarino et al., 2022). The construction details for prediction intervals can be found in Algorithm 5 in the appendix. and the training procedures specific to each dataset We set the target coverage to be $(1-\alpha) = 0.95$. We report the empirical coverage and interval width mean in Table 2, where the empirical coverage is the percentage of test data points which fall into the corresponding prediction interval. Our reported results are averaged over three random seeds to ensure robustness and reliability.

## 5. Conclusion and Discussion

In summary, our paper introduces the concept of Orthogonal Bootstrap, a novel technique that streamlines the process of bootstrap resampling. By separately treating the non-orthogonal (influence function) and orthogonal parts, akin to utilizing Infinitesimal Jackknife estimator as a control variate, we effectively reduce the number of required Monte Carlo replications. This innovation allows our method to maintain the higher-order coverage properties associated with traditional Bootstrap methods while simultaneously decreasing the computational burden. It's important to note that the control variate significantly reduces the computational costs involved in the simulation process while it doesn't alter the expected length of the constructed confidence interval. In essence, Orthogonal Bootstrap presents a practical and efficient solution for improving the accuracy and speed of Bootstrap resampling, making it a valuable tool for statistical analysis and inference.

## Impact Statement

Our paper contributes to fast and accurate quantification of the uncertainty for larger scale machine learning problems, which giving policy maker an idea of reliability of the AI prediction and is vital for risk management and help to be responsible in contexts where AI decisions have significant ethical implications. The theoretical results presented in the paper will have no ethical impact.

## References

Abad, J., Bhatt, U., Weller, A., and Cherubin, G. Approximating full conformal prediction at scale via influence functions. *arXiv preprint arXiv:2202.01315*, 2022.

Abadi, M. Tensorflow: learning functions at scale. In

*Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, pp. 1–1, 2016.

Adams, J., Gray, H., and Watkins, T. An asymptotic characterization of bias reduction by jackknifing. *The Annals of Mathematical Statistics*, pp. 1606–1612, 1971.

Alaa, A. and Van Der Schaar, M. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*, pp. 165–174. PMLR, 2020.

Asmussen, S. and Glynn, P. W. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

Chang, K.-C. *Methods in nonlinear analysis*, volume 10. Springer, 2005.

Chen, N., Gao, X., and Xiong, Y. Debiasing samples from online learning using bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pp. 8514–8533. PMLR, 2022.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

Chernozhukov, V., Newey, W. K., and Singh, R. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b.

Contarino, A., Kabban, C. S., Johnstone, C., and Mohd-Zaid, F. Constructing prediction intervals with neural networks: An empirical evaluation of bootstrapping and conformal inference methods. *arXiv preprint arXiv:2210.05354*, 2022.

Cook, R. D. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

Cook, R. D. and Weisberg, S. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

Cordeiro, G. M., Cribari-Neto, F., et al. *An introduction to Bartlett correction and bias reduction*. Springer, 2014.

Dayal, S. A converse of taylor's theorem for functions on banach spaces. *Proceedings of the American Mathematical Society*, 65(2):265–273, 1977.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Efron, B. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer, 1992a.

Efron, B. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):83–111, 1992b.

Efron, B. and Stein, C. The jackknife estimate of variance. *The Annals of Statistics*, pp. 586–596, 1981.

Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

Etter, P. and Ying, L. Operator augmentation for general noisy matrix systems. *arXiv preprint arXiv:2104.11294*, 2021.

Etter, P. A. and Ying, L. Operator augmentation for noisy elliptic systems. *arXiv preprint arXiv:2010.09656*, 2020.

Fernholz, L. T. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

Filippova, A. Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory of Probability & Its Applications*, 7(1):24–57, 1962.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Giordano, R., Jordan, M. I., and Broderick, T. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1139–1147. PMLR, 2019b.

Hall, P. On the bootstrap and confidence intervals. *The Annals of Statistics*, pp. 1431–1452, 1986.

Hall, P. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

Hall, P. and Martin, M. A. On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671, 1988.

Jaeckel, L. A. *The infinitesimal jackknife*. Bell Telephone Laboratories, 1972.

Jiao, J. and Han, Y. Bias correction with jackknife, bootstrap, and taylor series. *IEEE Transactions on Information Theory*, 66(7):4392–4418, 2020.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., and Robins, J. M. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9):1341–1356, 2011. doi: 10.1109/TNN.2011.2162110.

Kline, P. and Santos, A. A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1 (1):23–41, 2012.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Koltchinskii, V. Estimation of smooth functionals in high-dimensional models: bootstrap chains and gaussian approximation. *The Annals of Statistics*, 50(4):2386–2415, 2022.

Koltchinskii, V. and Zhilova, M. Estimation of smooth functionals in normal models: bias reduction and asymptotic efficiency. *The Annals of Statistics*, 49(5):2577–2610, 2021.

Lam, H. A cheap bootstrap method for fast inference. *arXiv preprint arXiv:2202.00090*, 2022.

Lam, H. and Qian, H. Subsampling variance for input uncertainty quantification. In *2018 Winter Simulation Conference (WSC)*, pp. 1611–1622. IEEE, 2018.

Lam, H. and Qian, H. Subsampling to enhance efficiency in input uncertainty quantification. *Operations Research*, 70(3):1891–1913, 2022.

Li, S. Debiasing the debiased lasso with bootstrap. *Electronic Journal of Statistics*, 14(1):2298–2337, 2020.

Liu, R. Y. Bootstrap procedures under some non-iid models. *The annals of statistics*, 16(4):1696–1708, 1988.

Lu, Z., Ie, E., and Sha, F. Uncertainty estimation with infinitesimal jackknife, its distribution and mean-field approximation. *arXiv preprint arXiv:2006.07584*, 2020.

Ma, C. and Ying, L. Correcting convexity bias in function and functional estimate. *arXiv preprint arXiv:2208.07996*, 2022.

Martens, J. Deep learning via hessian-free optimization. In *ICML*, volume 27, pp. 735–742, 2010.

Mises, R. v. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Nguyen, T. H., Zhang, H. R., and Nguyen, H. L. Improved worst-group robustness via classifier retraining on independent splits. *arXiv preprint arXiv:2204.09583*, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Quenouille, M. H. Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pp. 483–484. Cambridge University Press, 1949.

REEDS, J. I. On the definition of von mises functionals. *Ph. D thesis, Harvard University*, 1976.

Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., and Stahel, W. A. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Sengupta, S., Volgushev, S., and Shao, X. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.

Serfling, R. J. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Song, E., Nelson, B. L., and Pegden, C. D. Advanced tutorial: Input uncertainty quantification. In *Proceedings of the Winter Simulation Conference 2014*, pp. 162–176. IEEE, 2014.

Stine, R. A. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392): 1026–1031, 1985.

Tukey, J. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

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

Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

Wu, C.-F. J. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

Zhou, F., Li, P., and Zhang, C.-H. High-order statistical functional expansion and its application to some nonsmooth problems. *arXiv preprint arXiv:2112.15591*, 2021a.

Zhou, Z., Wang, G., Blanchet, J., and Glynn, P. W. Unbiased optimal stopping via the muse. *arXiv preprint arXiv:2106.02263*, 2021b.

## Organization of the Appendix

The appendix is structured as follows. In Section A, we provide some background on the von Mises expansion of the statistical functional (Serfling, 2009), kernel mean embeddings (Muandet et al., 2017), and multivariate calculus on Banach spaces (Chang, 2005). Similar to the development of (Lam & Qian, 2022), we propose general assumptions that theoretically guarantee the success of our method in Section B. We subsequently offer a range of illustrative examples where this assumption holds in the setting of kernel mean embeddings. Finally we present the details of our experiment and several additional experiment results in Section C.

## A. Preliminaries

### A.1. The Von Mises Expansion

In this section, we present the Von Mises expansion, a distributional analog of the Taylor expansion applied for a statistical functional $T$. Given two points $F$ and $G$ in a collection $\mathcal{F}$ of distributions, we consider the taylor expansion of the statistic function $T$ over the line segment in $\mathcal{F}$ joining $F$ and $G$ consists of the set of distribution functions $\{(1-\lambda)F + \lambda G, 0 \leq \lambda \leq 1\}$, *i.e.*

$$\begin{aligned} T(G) - T(F) &= d_1 T(F; G-F) + \frac{1}{2!} d_2 T(F; G-F) + \cdots, \\ &= \sum_{k=1}^{m} \frac{1}{k!} d_k T(F; G-F) + \frac{1}{(m+1)!} \frac{d^{m+1}}{d\lambda^{(m+1)}} T(F + \lambda(G-F))|_{\lambda^*} \end{aligned} \tag{6}$$

where $d_k(F; G-F)$ is the $k$-th order von Mises diffferential of $T$ at $F$ in the direction of $G$ to be

$$d_k T(F; G-F) = \frac{d^k}{d\lambda^k} T(F + \lambda(G-F))|_{\lambda=0+}.$$

Note that the von Mises differential is defined in Gateaux's manner (see Definition 1.1.2 in (Chang, 2005)). In typical cases, the $k$-th order von Moses differential $d_k T(F; G-F)$ is always $k$-linear (Fernholz, 2012; Abad et al., 2022), *i.e.* there exists a function $T_k[x_1, \cdots, x_k], (x_1, \cdots, x_k) \in \mathbb{R}^k$ such that

$$d_k T(F; G-F) = \int \cdots \int T_k[F; x_1, \cdots, x_k] \prod_{i=1}^{k} d[G(x_i) - F(x_i)]$$

holds for all $G$ (Dayal, 1977) (e.g. Lemma A.4). Following Lemma 6.3.2.A/B in (Serfling, 2009), we use the V-statistics representation of $d_k T(F; F_n - F)$, we can have the following von Mises expansion (Mises, 1947; Filippova, 1962; REEDS, 1976; Van der Vaart, 2000; Serfling, 2009)

$$\theta(G) = \theta(F) + \mathbb{E}_G \phi_1(X) + \frac{1}{2}\mathbb{E}_G \phi_2(X_1, X_2) + \cdots = \theta(F) + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbb{E}_G \phi_k(X_1, \cdots, X_k)$$

The function $\phi_1(x)$ is known as the influence function of $\theta$ and similarly $\phi_k(x_1, \cdots, x_k)$ is the $k$-th order influence function defined as

$$\phi_k(x_1, \cdots, x_k) = \frac{d}{ds_1}|_{s_1=0} \cdots \frac{d}{ds_k}|_{s_k=0} \theta((1 - \sum s_i)F + \sum s_i \delta_{x_i})$$

Now we document some lemma that are useful for our theoretical development. We consider a statistical functional resembling the form of the $k$-th order von Mises differential.

**Lemma A.1** (Section 6.3.2 Lemma A, (Serfling, 2009)). *Let $F$ be fixed and $h(x_1, \cdots, x_m)$ be given. A functional of the form*

$$T(G) = \int \cdots \int h(x_1, \cdots, x_m) \prod_{i=1}^{m} \mathrm{d}[G(x_i) - F(x_i)]$$

*can be written as a functional of the form*

$$T(G) = \int \cdots \int \tilde{h}(x_1, \cdots, x_m) \prod_{i=1}^{m} \mathrm{d}[G(x_i)]$$

*where the definition of $\tilde{h}$ depends on $F$. Moreover, we can have $\int \tilde{h}(x_1, \cdots, x_m)\mathrm{d}F(x_i) = 0$ for all $i \in [m]$.*

*Proof.* Take

$$\begin{aligned}\tilde{h}(x_1,\cdots,x_m) =& h(x_1,\cdots,x_m) - \sum_{i=1}^{m}\int h(x_1,\cdots,x_m)\mathrm{d}F(x_i)\\ &+\sum_{i<j}\int\int h(x_1,\cdots,x_m)\mathrm{d}F(x_i)\mathrm{d}F(x_j) - \cdots\\ &+(-1)^m\int\cdots\int h(x_1,\cdots,x_m)\prod_{i=1}^{m}\mathrm{d}F(x_i),\end{aligned}$$

then $\tilde{h}$ satisfies all properties claimed in the lemma. □

Using this alternative representation, we can obtain two lemma which gives the rate of the moment of the $V$-statistics type of functional in Lemma A.1.

**Lemma A.2** (Section 6.3.2 Lemma B, (Serfling, 2009))**.** *Suppose that* $\mathbb{E}_F\{h(X_1,\cdots,X_{i_m})^2\} < \infty$ *for all* $1 \le i_1,\cdots,i_m \le m$*. Then*

$$\mathbb{E}_F\left\{\left(\int\cdots\int h(x_1,\cdots,x_m)\prod_{i=1}^{m}d[\hat{F}(x_i)-F(x_i)]\right)^2\right\} = O(n^{-m}), \tag{7}$$

*where* $\hat{F}$ *is the empirical distribution for i.i.d observations* $X_1, X_2, \cdots, X_n$ *of distribution function* $F$*.*

We replicate the proof of Lemma A.2 here, for we use the same proof technique to prove Lemma A.3

*Proof.* Let $\tilde{h}$ be defined as in Lemma A.1. The left-hand side of (7) is given by

$$n^{-2m}\sum_{i_1=1}^{n}\cdots\sum_{i_m=1}^{n}\sum_{j_1=1}^{n}\cdots\sum_{j_m=1}^{n}\mathbb{E}_F\tilde{h}(X_{i_1},\cdots,X_{i_m})\tilde{h}(X_{j_1},\cdots,X_{j_m}).$$

For $\int\tilde{(}X_1,\cdots,X_k)dF(x_i) = 0$ for $1 \le i \le k$, thus the typical term in the upper part may be possibly nonzero only if the sequence of indices $i_1,\cdots,i_m,j_1,\cdots,j_m$ contains each member at least twice. The number of such terms is clearly $O(n^m)$, and by the assumption that $\mathbb{E}_F\{h(X_1,\cdots,X_{i_m})^2\} < \infty$, we know (7) holds. □

**Lemma A.3.** *Suppose that* $\mathbb{E}_F\{h(X_1,\cdots,X_{i_m})^4\} < \infty$ *for all* $1 \le i_1,\cdots,i_m \le m$*. Then*

$$\mathbb{E}_F\left\{\left(\int\cdots\int h(x_1,\cdots,x_m)\prod_{i=1}^{m}d[\hat{F}(x_i)-F(x_i)]\right)^4\right\} = O(n^{-2m}), \tag{8}$$

*where* $\hat{F}$ *is the empirical distribution for i.i.d observations* $X_1, X_2, \cdots, X_m$ *of distribution function* $F$*.*

*Proof.* The left-hand side of (8) is given by

$$n^{-4m}\sum_{i_1^1=1}^{n}\cdots\sum_{i_m^1=1}^{n}\sum_{i_1^2=1}^{n}\cdots\sum_{i_m^2=1}^{n}\sum_{i_3^1=1}^{n}\cdots\sum_{i_m^3=1}^{n}\sum_{i_1^4=1}^{n}\cdots\sum_{i_m^4=1}^{n}\mathbb{E}\prod_{j=1}^{4}h(X_{i_1^j},\cdots,X_{i_m^j})$$

For $\int h(X_1,\cdots,X_k)dF(x_i) = 0$ for $1 \le i \le k$, thus the typical term in the upper part may be possibly nonzero only if $i_1^1,\cdots,i_m^1,i_1^2,\cdots,i_m^2,i_1^3,\cdots,i_m^3,i_1^4,\cdots,i_m^4$ contains each member at least twice. The number of such terms is clearly $O(n^{2m})$, and by the assumption that $\mathbb{E}_F\{h(X_1,\cdots,X_{i_m})^4\} < \infty$, we know (8) holds. □

The Von Mises Expansion exists for many functionals, for example, divergence (Serfling, 2009; Kandasamy et al., 2014) and (regularized) M-estimation (Serfling, 2009; Giordano et al., 2019b;a). For the Taylor expansion 6 to be rigorous, from Lemma A.2, we showed that it is suffices to show that $n^{m/2}\sup_{0\le\lambda\le1}\left|\frac{d^{m+1}}{d\lambda^{m+1}}T(F+\lambda(F_n-F))\right| \xrightarrow{p} 0$, or to bound the remainder $n^{m/2}R_{mn} = n^{m/2}\left(T(F_n) - T(F) - \sum_{k=1}^{m}\frac{1}{k!}d_k(F;F_n-F)\right) \xrightarrow{p} 0$.

### A.2. Expansion on Normed Space

An alternative functional derivative can be defined via Fréchet derivative if the space of distribution is equipped with a norm. Let $\mathcal{D} := \{\Delta : \Delta = c(G - H)|c \in \mathbb{R}, G \in \mathcal{F}, H \in \mathcal{F}\}$ be the linear space generated by differences . Let $\mathcal{D}$ be equipped with a norm $\|\cdot\|$. The first order Fréchet derivative $T(F; G - F)$ is a linear functional which satisfies

$$\lim_{G \to F} \frac{|T(G) - T(F) - T(F; G - F)|}{\|G - F\|} = 0 \tag{9}$$

if it exists.

The following lemma shows that if the first order Fréchet derivative exists, then the first order von Mises differential exists and is linear.

**Lemma A.4** (Section 6.2.2, Lemma A, (Serfling, 2009)). *Suppose that $T$ has a Fréchet derivative at $F$ with respect to $\|\cdot\|$, then for any $G$, $d_1 T(F; G - F)$ exists and*

$$d_1 T(F; G_F) = T(F; G - F).$$

**Lemma A.5** (Section 6.2.2, Lemma B, (Serfling, 2009)). *Let $T$ have a differential at $F$ with respect to $\|\cdot\|$. Let $\{X_i\}_{i=1}^n$ be i.i.d. sample from $F$ and $\hat{T}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. If $\sqrt{n}\|F_n - F\| = O(1)$, then $\sqrt{n}(T(\hat{F}) - T(F) - T(F; \hat{F} - F)) \to^p 0$.*

We can define higher order Fréchet derivatives in a similar way (Chang, 2005). The second order derivative is defined to be a bilinear mapping which satisfies

$$\|T(G) - T(F) - T(F; G - F) - \frac{1}{2}T(F; G - F, G - F)\| = o(\|G - F\|^2) \tag{10}$$

if it exists, and the $m$th-order derivatives at $F$ are defined successively by

$$\|T(G) - T(F) - \sum_{i=1}^{m} \frac{1}{m!} T(F; \underbrace{G - F, \cdots, G - F}_{m \text{ times}})\| = o(\|G - F\|^m) \tag{11}$$

if they exist.

### A.3. Kernel Mean Embeddings

Suppose $\mathcal{X}$ is a fixed nonempty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a real-valued positive definite kernel function associated with the reproducing kernel Hilbert space $\mathcal{H}$. Recall the reproducing kernel property says that the point-wise evaluation of $f \in \mathcal{H}$ can be expressed by its inner product with the kernel. Suppose $\mathcal{F}$ consists of distributions over $\mathcal{X}$. Recall the definition of kernel mean embeddings (Definition 3.1 in (Muandet et al., 2017))

$$\mu : \mathcal{F} \to \mathcal{H}, \quad F \mapsto \int k(\mathbf{x}, \cdot) \mathrm{d}F(\mathbf{x}).$$

The following lemma states the condition that a distribution $F$ must satisfy to be embedded into $\mathcal{H}$.

**Lemma A.6** (Lemma 3.1, (Muandet et al., 2017)). *If $\mathbb{E}_{X \sim F}\sqrt{k(X, X)} < \infty$, then $\mu(F) \in \mathcal{H}$ and $\mathbb{E}_F(g(X)) = \langle g, \mu(F)\rangle_{\mathcal{H}}$.*

It is easy to see that if $F$ and $G$ can be embedded into $\mathcal{H}$, so is the linear space generated by them. Therefore, we can equip $\mathcal{D}$ with $\|\cdot\|_{\mathcal{H}}$ by embedding $\mathcal{D}$ into $\mathcal{H}$. This is also know as kernel maximum mean discrepancy (kernel MMD).

The next lemma states that with high probability the empirical distribution is close to the original distribution in kernel MMD.

**Lemma A.7** (Theorem 3.4, (Muandet et al., 2017)). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive definite kernel on a separable topological space $\mathcal{X}$ with $\sup_{x \in \mathcal{X}} k(x, x) \le C < \infty$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sqrt{n}\|F_n - F\|_{\mathcal{H}} \le \sqrt{C} + \sqrt{2C \log \frac{1}{\delta}}$$

For further information on kernel mean embeddings, we refer the readers to (Muandet et al., 2017).

### A.4. Multivariate Expansion on Banach Spaces

In this section, we delve into the foundations of multivariate calculus within the context of Banach spaces. We begin by delving into the fundamentals of calculus within the framework of a Banach space (Chang, 2005). Let $X$ denote a Banach space, and consider an open set $U \subset X$. Let $f : U \to \mathbb{R}$ be a map. In parallel with our treatment of statistical functionals, we define the Fréchet derivatives at a point $x \in U$ as elaborated earlier (for example, Definition 9). We denote them by $Df(x)$, $D^2 f(x)$, and $D^m f(x)$ respectively.

Let's now explore the extension of multivariate calculus to the setting of multiple Banach spaces. Suppose we have $m$ Banach spaces $X_1, \ldots, X_m$ and denote the norm of $X_i$ as $\|\cdot\|_i$. Then the direct product of these Banach spaces, denoted as $X := X_1 \times \cdots \times X_m$, forms a Banach space itself. This space is equipped with the direct product norm $\|(x_1, \cdots, x_m)\| := \max_i \|x_i\|_i$, where $x_i \in X_i$ ($i \in \{1, \cdots, m\}$).

Let $U \subset X$ be an open set. For a map $f : U \to \mathbb{R}$, we can define the Fréchet derivative at a point $x \in U$, referring to it as the **total derivative** of $f$ at $x$. We denote it by $Df(x)$. Also, for any $x = (x_1, \cdots, x_m) \in U$ and any $i \in \{1, \cdots, m\}$, we can define the $i$-th **partial derivative** of $f$ at $x$ to be the Fréchet derivative of $f$ with respect to $x_i$ while holding the other variables fixed, provided the limit exists. We denote it by $\frac{\partial f}{\partial x_j}(x)$. By definition this is a linear functional on $X_i$.

If all partial derivative exist and are continuous, then $f$ is said to be of class $C^1$. In this case, we can differentiate first-order partial derivatives to obtain second-order partial derivatives

$$\frac{\partial^2 f}{\partial x_k \partial x_j}(x) := \frac{\partial}{\partial x_k}\left(\frac{\partial f}{\partial x_j}\right)$$

if they exists. Continuing this way leads to higher-order partial derivatives. A function $f$ is said to be **of class $\mathbf{C^k}$** if all the partial derivatives of $f$ of order less than or equal to $k$ exist and are continuous.

The above definitions generalize important concepts of multivariate calculus to product Banach spaces. Many desirable properties of basic multivariate calculus also generalize to this setting. For example, we have the usual chain rule for total derivatives and the equality of mixed partial derivatives. For the readers' convenience, we state the equality for mixed partial derivatives briefly as follows. Suppose $f$ is of class $C^k$, then the partial derivatives exist up to order $k$, and the mixed partial derivatives of any order are independent of the order of differentiation. The proofs are easily generalized from the corresponding proofs in standard multivariate calculus.

For a multivariate function, the partial derivatives is easier to calculate than the total derivatives. The good news is that we can use the partial derivatives to determine the total derivative. Specifically, we have

**Theorem A.8.** *Let $f$ be differentiable at $x \in U$. Then all of the partial derivatives of $f$ at $x$ exist, and*

$$Df(x) = \left(\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_m}(x)\right)$$

*Proof.* For $v = (v_1, \cdots, v_m)$ with norm samll enough such that $x + v \in U$, let $R(v) = f(x+v) - f(x) - Df(x)(v)$. The fact that $f$ is differentiable at $x$ implies that $f(v)/\|v\|$ goes to zero as $\|v\| \to 0$. The $i$-th partial derivative of $f$ at $x$ exists because it is exactly the Fréchet derivative when setting $v_j = 0$ for all $j \neq i$, i.e.,

$$\frac{\partial f}{\partial x_i}(x)(v_i) = Df(x)(v_i).$$

As $Df(x)(v)$ is linear in $v$, we obtain the desired result. □

We can also generalize the multivariate Taylor's theorem to Banach spaces. In order to express it concisely, it helps to introduce some shorthand notation. For any $n$ tuple $I = (i_1, \cdots, i_n)$ of indices, and

$$\partial^I = \frac{\partial^m}{\partial x_{i_1} \cdots \partial x_{i_n}}$$

$$(x - a)^I = (x_{i_1} - a_{i_1}) \cdots (x_{i_n} - a_{i_n}).$$

Then we have

**Theorem A.9.** *Suppose $f$ is of class $C^{k+1}$ for some $k \geq 0$, then*

$$f(x) = f(x_0) + \sum_{i=1}^{k} \frac{1}{i!} \sum_{I:|I|=i} \partial^I f(x_0)(x - x_0)^I + \epsilon(x),$$

*where* $\epsilon(x) = \frac{1}{k!} \sum_{I:|I|=k+1} (x - x_0)^I \int_0^1 (1-t)^k \partial^I f(x_0 + t(x - x_0)) \mathrm{d}t$

*Proof.* The proof can be carried out by first generalizing Theorem A.8 to higher-order derivatives and then use Theorem 1.1.10 in (Chang, 2005). □

As an example, let $\mathcal{H}_i$ be Hilbert spaces. Consider the following function defined via the tensorized inner product

$$f(x_1, \cdots, x_m) = \langle h, x_1 \otimes \cdots \otimes x_m \rangle_{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_m}$$

then

$$\frac{\partial f}{\partial x_j}(x)(v_j) = \langle h, x_1 \otimes \cdots \otimes x_{j-1} \otimes v_j \otimes x_{j+1} \otimes \cdots \otimes x_m \rangle_{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_m}$$

and

$$\frac{\partial^2 f}{\partial x_i \partial x_j}[v_1, v_2] = \begin{cases} \langle h, x_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes v_j \otimes \cdots \otimes x_m \rangle_{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_m} & i \neq j \\ 0 & i = j \end{cases}$$

We can derive higher order derivatives similarly. Moreover, any derivative of order strictly larger than $m$ is equal to zero, for example,

$$\partial^{m+1} f = 0$$

Hence we have by Theorem A.9 that

$$f(x) = f(x_0) + \sum_{i=1}^{m} \frac{1}{i!} \sum_{I:|I|=i} \partial^I f(x_0)(x - x_0)^I \tag{12}$$

## B. Proof of Main Results

Following (Lam & Qian, 2022), we first verify the improvement of our Orthogonal Bootstrap under the following smoothness assumption of our target performance measure. Without explicit statement, we assume $n_1 = n_2 = \cdots = n_m = n$ for simplicity.

**Assumption B.1** (Smoothness at True Input Model, Assumption 3 (Lam & Qian, 2022))**.**

$$\phi(F_1, \cdots, F_m) = \phi(\hat{F}_1, \cdots, \hat{F}_m) + \sum_{i=1}^{m} \int \phi_i(x) \mathrm{d}\hat{F}_i(x) + \delta$$

satisfies $\mathbb{E}[\delta^2] = o(n^{-1})$, $\mathrm{Var}_{X_i \sim F_i}[\phi_i(X_i)] > 0$, and $\mathbb{E}_{X_i \sim F_i}[\phi_i^4(X_i)] < \infty$ for all $i \in [m]$.

This assumption guarantees the existence of non-degenerate influence functions with respect to the true model. The first condition $\mathbb{E}[\delta^2] = o(n^{-1})$ guarantees that the error of the approximation by influence functions is negligible when the number of data $n$ is large. Indeed, the influence function term is asymptotically of order $\Theta_p(n^{-\frac{1}{2}})$ by the central limit theorem, whereas the error $\delta$ is implied to be $o_p(n^{-\frac{1}{2}})$. The second condition $\mathrm{Var}_{X_i \sim F_i}[\phi_i(X_i)] > 0$ says that the influence function $\phi_i$ are non-degenerate. The last assumption is needed to control the variance of the influence function at the empirical model, see Corollary B.4.

**Assumption B.2** (Smoothness at Empirical Input Model, Assumption 4 (Lam & Qian, 2022))**.**

$$\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) = \phi(\hat{F}_1, \cdots, \hat{F}_m) + \sum_{i=1}^{m} \int \mathcal{I}_i^{\phi}(x) \mathrm{d}\hat{F}_i^b(x) + \epsilon$$

satisfies $\mathbb{E}_*[\epsilon^4] = O_p(n^{-4})$ and $\mathbb{E}[(\mathcal{I}_i^{\phi} - \phi_i)^4(X_{i,1})] = o(1)$ for all $i \in [m]$.

This assumption guarantees the existence of non-degenerate influence functions with respect to the empirical model. The moment condition on the remainder is needed for controlling the variance of our orthogonal bootstrap estimator. Since a particular empirical model is chosen from the set of empirical models generated by the true model, the condition is described in terms of stochastic order. If the performance measure is sufficiently smooth, for example second order von Mises differential exists, then the second order differential is of order $\Theta_p(\frac{1}{n})$. Therefore expecting $\epsilon$ to be of order $O_p(\frac{1}{n})$ is reasonable. The last assumption entails the observation that the empirical distributions $\hat{F}_i$ converges to true ones $F_i$ as the data size $n$ grows and hence the empirical influence functions $\mathcal{I}_i^{\phi}$ are expect to approach the influence functions $\phi_i$ associated with the true input distributions.

### B.1. Variance of Non-Orthogonal Part

When the influence functions at the empirical model exist, we can obtain a closed form formula for the variance of the non-orthogonal part. As this result is used in the algorithm, we provide the full result and do not assume $n_1 = \cdots = n_m = n$ here.

**Lemma B.3.** *If the influence functions exist and are* 1*-linear, i.e.*

$$\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right] = \sum_{i=1}^{m} \int \mathcal{I}_i^{\phi}(x) \mathrm{d}\hat{F}_i^b(x),$$

*then we have*

$$\mathit{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right) = \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\mathcal{I}_i^{\phi}(X_{i,j}))^2 \tag{13}$$

*Proof.* Notice that $\mathbb{E}_{X_i \sim \hat{F}_i^b} \mathcal{I}_i^{\phi}(X_i) = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathcal{I}_i^{\phi}(X_{i,k}^b)$, where $X_{i,k}^b \sim \hat{F}_i$. Thus we have

$$\begin{aligned}
&\mathrm{Var}_*(\sum_{i=1}^{m} \mathbb{E}_{X_i \sim \hat{F}_i^b} \mathcal{I}_i^{\phi}(X_i)) = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \mathrm{Cov}_*(\mathbb{E}_{X_{i_1} \sim \hat{F}_{i_1}^b} \mathcal{I}_{i_1}^{\phi}(X_{i_1}), \mathbb{E}_{X_{i_2} \sim \hat{F}_{i_2}^b} \mathcal{I}_{i_2}^{\phi}(X_{i_2})) \\
&= \sum_{i=1}^{m} \mathrm{Var}_*(\mathbb{E}_{X_i \sim \hat{F}_i^b} \mathcal{I}_{i_1}^{\phi}(X_i)) + \sum_{i \neq j} \mathrm{Cov}_*(\mathbb{E}_{X_i \sim \hat{F}_i^b} \mathcal{I}_i^{\phi}(X_i), \mathbb{E}_{X_j \sim \hat{F}_j^b} \mathcal{I}_j^{\phi}(X_j))
\end{aligned}$$

As $\hat{F}_i^b$ is independent of $\hat{F}_j^b$, the second term in the above equation equals to zero. Therefore,

$$\begin{aligned}
&\mathrm{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right) = \sum_{i=1}^{m} \mathrm{Var}_*(\mathbb{E}_{X_i \sim \hat{F}_i^b} \mathcal{I}_i^{\phi}(X_i)) \\
&= \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{k,j} \mathrm{Cov}_*(\mathcal{I}_i^{\phi}(X_{i,k}^b), \mathcal{I}_i^{\phi}(X_{i,j}^b)) = \sum_{i=1}^{m} \frac{1}{n_i} \mathrm{Var}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))
\end{aligned}$$

where the last equality holds by the independence between $X_{i,k}^b$ and $X_{i,j}^b$ $(j \neq k)$. Note that as $\mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b)) = 0$ by definition,

$$\mathrm{Var}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b)) = \frac{1}{n_i} \sum_{j=1}^{n} (\mathcal{I}_i^{\phi}(X_{i,j}))^2,$$

thus we obtain the desired result. □

From now on we continue to assume $n_1 = \cdots = n_m = n$. We aim to determine the order of the random variable $\mathrm{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right)$ under Assumption B.1 and Assumption B.2. As the random variable is always positive, we use the first moment method, *i.e.* the Markov's inequality.

Now we can prove the following two results.

**Corollary B.4.** *Under Assumption B.1 and Assumption B.2, we have*

$$\mathit{Var}_*\left(\nabla\phi(\hat{F}_1, \cdots, \hat{F}_m)\left[\hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m\right]\right) = \Theta_p(\frac{1}{n}) \tag{14}$$

*Proof.* Let

$$\xi = \mathrm{Var}_* \left( \nabla \phi(\hat{F}_1, \cdots, \hat{F}_m) \left[ \hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m \right] \right),$$

then $\xi > 0$ is a random variable with respect to the probability distributions $F_1, \cdots, F_m$. The expectation of $\xi$ is $\Theta(\frac{1}{n})$ as

$$\mathbb{E}\xi = \frac{1}{n} \sum_{i=1}^{m} \mathbb{E}(\mathcal{I}_i^{\phi}(X_{i,1}))^2,$$

where we combine $0 < \mathrm{Var}_{X_i \sim F_i}[\phi_i(X_i)] < \infty$ and $\mathbb{E}[(\mathcal{I}_i^{\phi} - \phi_i)^2(X_{i,1})] = o(1)$ to show that $0 < \mathbb{E}(\mathcal{I}_i^{\phi}(X_{i,1}))^2 < \infty$. Via a standard first moment argument using Markov's inequality, we obtain $\xi = \Theta_p(\frac{1}{n})$. □

**Lemma B.5.** *Under Assumption B.1 and Assumption B.2, we have*

$$\mathit{Var}_* \left( \nabla \phi(\hat{F}_1, \cdots, \hat{F}_m) \left[ \hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m \right] \right)^2 = \Theta_p(\frac{1}{n^2}) \tag{15}$$

*Proof.* Let

$$\eta = \mathbb{E}_* \left( \nabla \phi(\hat{F}_1, \cdots, \hat{F}_m) \left[ \hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m \right] \right)^4,$$

then $\eta > 0$ is a random variable with respect to the probability distributions $F_1, \cdots, F_m$. Recall that

$$\nabla \phi(\hat{F}_1, \cdots, \hat{F}_m) \left[ \hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m \right] = \frac{1}{n} \sum_{i=1}^{m} \sum_{k=1}^{n} \mathcal{I}_i^{\phi}(X_{i,k}^b).$$

Note that $\mathbb{E}_{X_i \sim \hat{F}_i} \mathcal{I}_i^{\phi}(X_i) = 0$, we have

$$\begin{aligned}
\eta =& \frac{1}{n^4} \mathbb{E}_* \sum_{i_1,k_1,i_2,k_2,i_3,k_3,i_4,k_4} \mathcal{I}_{i_1}^{\phi}(X_{i_1,k_1}^b) \mathcal{I}_{i_2}^{\phi}(X_{i_2,k_2}^b) \mathcal{I}_{i_3}^{\phi}(X_{i_3,k_3}^b) \mathcal{I}_{i_4}^{\phi}(X_{i_4,k_4}^b) \\
=& \frac{1}{n^4} \mathbb{E}_* \sum_{i,k_1,k_2,k_3,k_4} \mathcal{I}_i^{\phi}(X_{i,k_1}^b) \mathcal{I}_i^{\phi}(X_{i,k_2}^b) \mathcal{I}_i^{\phi}(X_{i,k_3}^b) \mathcal{I}_i^{\phi}(X_{i,k_4}^b) \\
&+ \frac{3}{n^4} \sum_{i_1 \neq i_2,k_1,k_2,k_3,k_4} \mathbb{E}_* \mathcal{I}_{i_1}^{\phi}(X_{i_1,k_1}^b) \mathcal{I}_{i_1}^{\phi}(X_{i_1,k_2}^b) \mathbb{E}_* \mathcal{I}_{i_2}^{\phi}(X_{i_2,k_3}^b) \mathcal{I}_{i_2}^{\phi}(X_{i_2,k_4}^b) \\
=& \frac{3(n-1)}{n^3} \sum_{i=1}^{m} (\mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^2)^2 + \frac{1}{n^3} \sum_{i=1}^{m} \mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^4 + \frac{3}{n^2} \sum_{i \neq j} \mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^2 \mathbb{E}_*(\mathcal{I}_j^{\phi}(X_{j,1}^b))^2
\end{aligned}$$

by a calculation similar to Lemma B.9. Thus

$$\begin{aligned}
&\mathrm{Var}_* \left( \nabla \phi(\hat{F}_1, \cdots, \hat{F}_m) \left[ \hat{F}_1^b - \hat{F}_1, \cdots, \hat{F}_m^b - \hat{F}_m \right] \right)^2 \\
=& \eta - \xi^2 \\
=& \frac{(2n-3)}{n^3} \sum_{i=1}^{m} (\mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^2)^2 + \frac{1}{n^3} \sum_{i=1}^{m} \mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^4 + \frac{2}{n^2} \sum_{i \neq j} \mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^2 \mathbb{E}_*(\mathcal{I}_j^{\phi}(X_{j,1}^b))^2.
\end{aligned}$$

As $\mathbb{E}_{X_i \sim F_i}[\phi_i^4(X_i)] < \infty$ and $\mathbb{E}_{X_i \sim F_i}[(\mathcal{I}_i^{\phi} - \phi_i)^4(X_i)] = o(1)$, we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}_*(\mathcal{I}_i^{\phi}(X_{i,1}^b))^2]^2 = \mathbb{E}[\frac{1}{n} \sum_{j=1}^{n} (\mathcal{I}_i^{\phi}(X_{i,j}))^2]^2 \\
=& \frac{1}{n^2} \mathbb{E} \sum_{j,k} (\mathcal{I}_i^{\phi}(X_{i,j}))^2 (\mathcal{I}_i^{\phi}(X_{i,k}))^2 = O(1),
\end{aligned}$$

$\mathbb{E}(\mathcal{I}_i^\phi(X_i))^2 = O(1)$, and $\mathbb{E}(\mathcal{I}_i^\phi(X_i))^4 = O(1)$.

The expectation of $\text{Var}_*\left(\nabla\phi(\hat{F}_1,\cdots,\hat{F}_m)\left[\hat{F}_1^b-\hat{F}_1,\cdots,\hat{F}_m^b-\hat{F}_m\right]\right)^2$ is $\Theta(\frac{1}{n^2})$ as

$$\begin{aligned}
&\mathbb{E}\text{Var}_*\left(\nabla\phi(\hat{F}_1,\cdots,\hat{F}_m)\left[\hat{F}_1^b-\hat{F}_1,\cdots,\hat{F}_m^b-\hat{F}_m\right]\right)^2\\
=&\frac{(2n-3)}{n^3}\sum_{i=1}^m\mathbb{E}(\mathbb{E}_*(\mathcal{I}_i^\phi(X_{i,1}^b))^2)^2+\frac{2}{n^2}\mathbb{E}\sum_{i\neq j}\mathbb{E}_*(\mathcal{I}_i^\phi(X_{i,1}^b))^2\mathbb{E}_*(\mathcal{I}_j^\phi(X_{j,1}^b))^2+\frac{1}{n^3}\sum_{i=1}^m\mathbb{E}\mathbb{E}_*(\mathcal{I}_i^\phi(X_{i,1}^b))^4\\
=&\frac{(2n-3)}{n^3}\sum_{i=1}^m\mathbb{E}(\mathbb{E}_*(\mathcal{I}_i^\phi(X_{i,1}^b))^2)^2+\frac{2}{n^2}\sum_{i\neq j}\mathbb{E}(\mathcal{I}_i^\phi(X_i))^2\mathbb{E}(\mathcal{I}_j^\phi(X_j))^2+\frac{1}{n^3}\sum_{i=1}^m\mathbb{E}(\mathcal{I}_i^\phi(X_i))^4.
\end{aligned}$$

Combining the above argument and using Markov's inequality again, we get $\eta = \Theta_p(\frac{1}{n^2})$. □

### B.2. Improvement of Orthogonal Bootstrap when Simulating the Mean

**Theorem B.6.** *Under Assumption B.2, consider the Orthogonal Bootstrap debiasing estimator defined in Equation (1)*

$$X := 2\hat{\phi}-\frac{1}{B}\sum_{b=1}^B(\hat{\phi}^b-\sum_{i=1}^m\frac{1}{n}\sum_{j=1}^n\mathcal{I}_i^\phi(x_{i,j}^b)).$$

*If the number of Monte Carlo replications $B \geq Cn^\alpha$ for some absolute constant $C > 0$ and $\alpha \geq 0$, then the simulation error $Var_*(X) = O_p(\frac{1}{n^{2+\alpha}})$.*

*Proof.* The simulation variance of our debiasing estimator is

$$\text{Var}_*X=\frac{1}{B}\text{Var}_*(\hat{\phi}^b-\sum_{i=1}^m\frac{1}{n}\sum_{j=1}^n\mathcal{I}_i^\phi(x_{i,j}^b)).$$

By Assumption B.2, $\hat{\phi}^b=\hat{\phi}+\sum_{i=1}^m\frac{1}{n}\sum_{j=1}^n\mathcal{I}_i^\phi(x_{i,j}^b)+\epsilon$ where $\text{Var}_*\epsilon = O_p(\frac{1}{n^2})$. Therefore

$$\text{Var}_*X=\frac{1}{B}\text{Var}_*\epsilon=O_p(\frac{1}{n^2B})=O_p(\frac{1}{n^{2+\alpha}}).$$

□

It is also easy to obtain a conditional central limit theorem for our estimator if we further assume $\text{Var}_*\epsilon > 0$.

**Theorem B.7.** *Under Assumption B.2, and further assuming that $Var_*\epsilon > 0$. Consider the Orthogonal Bootstrap debiasing estimator defined in Equation (1)*

$$X := 2\hat{\phi}-\frac{1}{B}\sum_{b=1}^B(\hat{\phi}^b-\sum_{i=1}^m\frac{1}{n}\sum_{j=1}^n\mathcal{I}_i^\phi(x_{i,j}^b)).$$

*Then conditioning on the input data $\hat{F}_1,\cdots,\hat{F}_m$, we have*

$$\sqrt{B}\frac{X-\mathbb{E}_*X}{\sqrt{Var_*\epsilon}}\to\mathcal{N}(0,1)$$

*as $B$ tends to infinity. Moreover, we have $Var_*\epsilon = O_p(\frac{1}{n^2})$.*

For comparison, we provide the result for bootstrap here.

**Theorem B.8.** *Under Assumption B.1 and Assumption B.2, consider the Standard Bootstrap debiasing estimator defined by*

$$X := 2\hat{\phi} - \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^b.$$

*Then conditioning on the input data $\hat{F}_1, \cdots, \hat{F}_m$, we have*

$$\frac{\sqrt{B}X - \mathbb{E}_* X}{\sqrt{Var_*\hat{\phi}^b}} \to \mathcal{N}(0,1)$$

*as $B$ tends to infinity. Moreover, we have $Var_*\hat{\phi}^b = \Theta_p(\frac{1}{n})$.*

*Proof.* The simulation variance of $X$ is $\text{Var}_* X = \frac{1}{B}\text{Var}_*\hat{\phi}^b$. By Assumption B.2, we have $\hat{\phi}^b = \hat{\phi} + \sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\mathcal{I}_i^{\phi}(x_{i,j}^b) + \epsilon$ where $\text{Var}_*\epsilon = O_p(\frac{1}{n^2})$. Combined with Corollary B.4, we have

$$\text{Var}_*\left[\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\mathcal{I}_i^{\phi}(x_{i,j}^b)\right] = \Theta_p(\frac{1}{n}).$$

Therefore by Cauchy's inequality

$$\text{Var}_*(\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\mathcal{I}_i^{\phi}(x_{i,j}^b) + \epsilon) = \Theta_p(\frac{1}{n}) + o_p(\frac{1}{n}) = \Theta_p(\frac{1}{n}).$$

An application of central limit theorem yields the result. □

These two theorems show that conditioning on the input data, the variance of the Standard Bootstrap estimator due to simulation is at least $n$ times larger than the variance of the Orthogonal Bootstrap estimator due to simulation. This provides a separation result for the two estimators, showing that our estimator is strictly better under certain smoothness assumptions of the performance measure.

### B.3. Improvement of Orthogonal Bootstrap when Simulating the Variance

We first begin with a lemma that calculate the variance of the sample covariance matrix.

**Lemma B.9.** *Let $X_1, \cdots, X_n$ be i.i.d. sample drawn from a multivariate distribution function $F$ in $\mathbb{R}^p$ with finite fourth moments. Let $X_{ki}$ denote the $k$-th coordinate of the random variable $X_i$, and let $\mu_k := \mathbb{E}X_{ki}$ be its mean. Let $\mu_{kl} := \mathbb{E}(X_{ki} - \mu_k)(X_{li} - \mu_l)$ be the second order central moment. Let $\mu_{klm} := \mathbb{E}(X_{ki} - \mu_k)(X_{li} - \mu_l)(X_{mi} - \mu_m)$ be the third order central moment. Let $\mu_{klmq} := \mathbb{E}(X_{ki} - \mu_k)(X_{li} - \mu_l)(X_{mi} - \mu_m)(X_{qi} - \mu_q)$ be the fourth order central moment.*

*Let the sample covariance matrix be*

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T,$$

*then*

$$Cov(S_{nkl}, S_{nmq}) = \frac{(n-1)^2}{n^3}\alpha_{klmq} + \frac{n-1}{n^3}(\mu_{km}\mu_{lq} + \mu_{kq}\mu_{ml}),$$

*where*

$$\alpha_{klmq} = \mu_{klmq} - \mu_{kl}\mu_{mq}.$$

*Proof.* First of all we have the decomposition

$$S_{nkl} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_{ki} - \mu_k)(X_{li} - \mu_l)}_{:=A_{nkl}} - \underbrace{(\bar{X}_k - \mu_k)(\bar{X}_l - \mu_l)}_{:=B_{nkl}}.$$

Note that the covariance between $S_{nkl}$ and $S_{nmq}$ is

$$\begin{aligned}&\text{Cov}(S_{nkl}, S_{nmq})\\=&\text{Cov}(A_{nkl}, A_{nmq}) - \text{Cov}(A_{nkl}, B_{nmq}) - \text{Cov}(B_{nkl}, A_{nmq}) + \text{Cov}(B_{nkl}, B_{nmq})\end{aligned}$$

For the first term we have

$$\begin{aligned}&\text{Cov}(A_{nkl}, A_{nmq})\\=&\frac{1}{n}\text{Cov}((X_{ki}-\mu_k)(X_{li}-\mu_l), (X_{mi}-\mu_m)(X_{qi}-\mu_q))\\=&\frac{1}{n}\left(\mathbb{E}(X_{ki}-\mu_k)(X_{li}-\mu_l)(X_{mi}-\mu_m)(X_{qi}-\mu_q) - \mu_{kl}\mu_{mq}\right)\\=&\frac{1}{n}\alpha_{klmq}\end{aligned}$$

For the last term we have

$$\begin{aligned}&\text{Cov}(B_{nkl}, B_{nmq})\\=&\text{Cov}(\frac{1}{n^2}\sum_i(X_{ki}-\mu_k)\sum_j(X_{lj}-\mu_l), \frac{1}{n^2}\sum_s(X_{ms}-\mu_k)\sum_t(X_{qt}-\mu_q))\\=&\frac{1}{n^4}\sum_{i,j,s,t}\mathbb{E}(X_{ki}-\mu_k)(X_{lj}-\mu_l)(X_{ms}-\mu_m)(X_{qt}-\mu_q)\\&-\frac{1}{n^4}\mathbb{E}\sum_{i,j}(X_{ki}-\mu_m)(X_{lj}-\mu_l)\mathbb{E}\sum_{s,t}(X_{ms}-\mu_m)(X_{qt}-\mu_q)\\=&\frac{n(n-1)}{n^4}(\mu_{kl}\mu_{mq}+\mu_{km}\mu_{lq}+\mu_{kq}\mu_{ml}) - \frac{1}{n^2}\mu_{kl}\mu_{mq}\\&+\frac{1}{n^3}\mathbb{E}(X_k-\mu_k)(X_l-\mu_l)(X_m-\mu_k)(X_q-\mu_q)\\=&\frac{n-1}{n^3}(\mu_{km}\mu_{lq}+\mu_{kq}\mu_{ml}) + \frac{1}{n^3}\alpha_{klmq}\end{aligned}$$

For the cross term we have

$$\begin{aligned}&\text{Cov}(A_{nkl}, B_{nmq})\\=&\text{Cov}(\frac{1}{n}\sum_i(X_{ki}-\mu_k)(X_{li}-\mu_l), \frac{1}{n^2}\sum_s(X_{ms}-\mu_k)\sum_t(X_{qt}-\mu_q))\\=&\frac{1}{n^3}\sum_{i,s,t}\mathbb{E}(X_{ki}-\mu_k)(X_{li}-\mu_l)(X_{ms}-\mu_m)(X_{qt}-\mu_q)\\&-\frac{1}{n^3}\mathbb{E}\sum_i(X_{ki}-\mu_k)(X_{li}-\mu_l)\mathbb{E}\sum_{s,t}(X_{ms}-\mu_s)(X_{qt}-\mu_q)\\=&\frac{n(n-1)}{n^3}\mu_{kl}\mu_{mq} + \frac{1}{n^2}\mu_{klmq} - \frac{1}{n}\mu_{kl}\mu_{mq}\\=&\frac{1}{n^2}\alpha_{klmq}\end{aligned}$$

Therefore

$$\text{Cov}(S_{nkl}, S_{nmq}) = \frac{(n-1)^2}{n^3}\alpha_{klmq} + \frac{n-1}{n^3}(\mu_{km}\mu_{lq}+\mu_{kq}\mu_{ml})$$

□

**Theorem B.10.** *Under Assumption B.2, consider the Orthogonal Bootstrap variance estimator defined in Equation (4)*

$$X := \frac{1}{n^2}\sum_{i=1}^m\sum_{j=1}^n(\mathcal{I}_i^\phi(X_{i,j}))^2 + \frac{1}{B}\sum_{b=1}^B\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^B\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right).$$

*If the number of Monte Carlo replications $B \geq Cn^\alpha$ for some absolute constant $C > 0$ and $\alpha \geq 0$, then $Var_*(X) = O_p(\frac{1}{n^{3+\alpha}})$.*

*Proof.* Let $Y^b := \hat{\phi}^b - \hat{\mathcal{I}}^b$ and $Z^b := \hat{\phi}^b$. Regard $(Y^b, Z^b)$ as a two dimensional vector, then its sample covariance matrix is

$$S_B = \begin{pmatrix} \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 & \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right) \\ \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right) & \frac{1}{B}\sum_{b=1}^B \left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right)^2 \end{pmatrix},$$

so $S_{B11} := \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2$ and $S_{B12} := \frac{2}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right)$. Then the variance of $X$ is

$$\text{Var}_* X = \text{Var}_*(S_{B11} + S_{B12}) = \text{Var}_* S_{B11} + 2\text{Cov}_*(S_{B11}, S_{B12}) + \text{Var}_* S_{B12}.$$

By Lemma B.9

$$\text{Var}_* S_{B11} = \frac{(B-1)^2}{B^3}\alpha_{1111} + \frac{2(B-1)}{B^3}\mu_{11}^2,$$
$$\text{Var}_* S_{B12} = \frac{(B-1)^2}{B^3}\alpha_{1212} + \frac{2(B-1)}{B^3}\mu_{12}^2,$$

and

$$\text{Cov}(S_{B11}, S_{B12}) = \frac{(B-1)^2}{B^3}\alpha_{1112} + \frac{2(B-1)}{B^3}\mu_{11}\mu_{12}.$$

Therefore

$$\text{Var}_* X = \frac{(B-1)^2}{B^3}(\alpha_{1111} + \alpha_{1212} - 2\alpha_{1112}) + \frac{2(B-1)}{B^3}(\mu_{11} - \mu_{12})^2$$

By definition and recalling that $\mathbb{E}_*\hat{\mathcal{I}}^b = 0$, $\text{Var}_*\hat{\mathcal{I}}^b = \Theta_p(\frac{1}{n})$ (Corollary B.4), and $\text{Var}_*(\hat{\mathcal{I}}^b)^2 = \Theta_p(\frac{1}{n^2})$ (Lemma B.5),

$$\mu_{11} = \text{Var}_*\epsilon^b = O_p(\frac{1}{n^2})$$
$$\mu_{12} = \text{Var}_*\epsilon^b + \mathbb{E}_*\epsilon^b\hat{\mathcal{I}}^b = O_p(\frac{1}{n^{\frac{3}{2}}})$$
$$\mu_{1111} = \mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^4 = O_p(\frac{1}{n^4})$$
$$\mu_{1212} = \mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^4 + 2\mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^3\hat{\mathcal{I}}^b + \mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^2(\hat{\mathcal{I}}^b)^2 = O_p(\frac{1}{n^3})$$
$$\mu_{1112} = \mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^4 + \mathbb{E}_*(\epsilon - \mathbb{E}_*\epsilon)^3\hat{\mathcal{I}}^b = O_p(\frac{1}{n^{\frac{7}{2}}})$$

Therefore when $B \geq Cn^\alpha$, $\text{Var}_* X = O_p(\frac{1}{n^{3+\alpha}})$. □

**Theorem B.11.** *Under Assumption B.2, consider the Orthogonal Bootstrap variance estimator defined in Equation (4)*

$$X := \frac{1}{n^2}\sum_{i=1}^m\sum_{j=1}^n (\mathcal{I}_i^\phi(X_{i,j}))^2 + \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right).$$

*Then conditioning on input data $\hat{F}_1, \cdots, \hat{F}_m$, we have*

$$\sqrt{B}\frac{X - \mathbb{E}_* X}{\sqrt{S^2}} \to \mathcal{N}(0, 1)$$

*as $B$ tends to infinity. Moreover, $S^2 = O_p(\frac{1}{n^3})$.*

For comparison, we provide the result for Standard Bootstrap here.

**Theorem B.12.** *Under Assumption B.1 and Assumption B.2, consider the Standard Bootstrap variance estimator*

$$X := \frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}^b - \overline{\phi})^2.$$

*Then conditioning on the input data $\hat{F}_1, \cdots, \hat{F}_m$, we have*

$$\sqrt{B}\frac{X - \mathbb{E}_* X}{\sqrt{S^2}} \to \mathcal{N}(0,1).$$

*Moreover, we have $S^2 = \Theta_p(\frac{1}{n^2})$.*

*Proof.* By Lemma B.9, the variance of $X$ is

$$\text{Var}_* X = \frac{(B-1)^2}{B^3}\left[\text{Var}_*(\hat{\phi}^b - E_*\hat{\phi}^b)^2\right] + \frac{2(B-1)}{B^3}\text{Var}_*^2\hat{\phi}^b.$$

Now we need to specify $\text{Var}_*(\hat{\phi}^b - E_*\hat{\phi}^b)^2$. By Assumption B.2,

$$\hat{\phi}^b = \hat{\phi} + \hat{\mathcal{I}}^b + \epsilon^b$$

where $\hat{\mathcal{I}}^b = \sum_{i=1}^m \frac{1}{n}\sum_{j=1}^n \mathcal{I}_i^{\phi}(x_{i,j}^b)$ and $\mathbb{E}_*(\epsilon^b)^4 = O_p(\frac{1}{n^4})$. We have already obtained

$$\text{Var}_*(\hat{\mathcal{I}}^b) = \Theta_p(\frac{1}{n}),$$

and

$$\text{Var}_*(\hat{\mathcal{I}}^b)^2 = \Theta_p(\frac{1}{n^2}),$$

in Corollary B.4 and Lemma B.5 respectively.

Combining the above estimate, we obtain $\text{Var}_*(\hat{\phi}^b - E_*\hat{\phi}^b)^2 = \Theta_p(\frac{1}{n^2})$. Moreover, $\text{Var}_*^2\hat{\phi}^b$ is also $\Theta_p(\frac{1}{n^2})$. Therefore we can apply the central limit theorem when we are considering the large $B$ limit, conditioning on the input data. □

Similar to the situation of debiasing, these two theorems show that conditioning on the input data, the variance of the Standard Bootstrap estimator due to simulation is at least $n$ times larger than the variance of the Orthogonal Bootstrap estimator due to simulation when estimating variance.

### B.4. Improved Variance Estimation

Occasionally, our Orthogonal Bootstrap estimator for variance can yield negative values, particularly when the sample size $n$ is small. In such instances, the resulting confidence intervals constructed using the Orthogonal Bootstrap method become devoid of meaningful interpretation.

However, this issue can be effectively addressed through a slight modification to our estimator, without compromising the fundamental theoretical properties of our method. We refer to this enhanced version as the **Improved Orthogonal Bootstrap**.

The core idea behind the Improved Orthogonal Bootstrap is that when simulation results produce negative values, we seamlessly transition to utilizing the Infinitesimal Jackknife method. The algorithm for the Improved Orthogonal Bootstrap is summarized in Algorithm 3. Importantly, we emphasize that this improvement preserves the desirable properties of the original Orthogonal Bootstrap method.

Now we assert that our improved Orthogonal Bootstrap still has the same favorable properties as original Orthogonal Bootstrap. Crucially, it's worth noting that the occurrence of negative simulation results is exceedingly rare, and as such, has a negligible impact on the overall algorithm's performance.

**Algorithm 3** Variance Estimation via Improved Orthogonal Bootstrap

**Input**: A generic performance measure $\phi(F_1, \cdots, F_m)$, i.i.d samples $\{X_{i,1}, \cdots, X_{i,n_i}\} \in \mathbb{R}^{d_1}$ of $F_i$, and influence function $\mathcal{I}_i^{\phi}$ of $\phi$ respect to $\hat{F}_i$.

**Output**: Estimation of $\text{Var}_{F_1,\cdots,F_m}\phi(\hat{F}_1, \cdots, \hat{F}_m)$.

$\hat{\phi} \leftarrow \phi(\hat{F}_1, \cdots, \hat{F}_m)$, where $\hat{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\delta_{X_{i,j}}$
**for** b=1:B **do**
  **for** i=1:m **do**
    Sample $\{x_{i,1}^b, \cdots, x_{i,n_i}^b\}$ i.i.d from $\hat{F}_i$
  **end for**
  $\hat{\phi}^b \leftarrow \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)$, where $\hat{F}_i^b = \frac{1}{n_i}\sum_{j=1}^{n_i}\delta_{x_{i,j}^b}$
  $\hat{\mathcal{I}}^b = \sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{I}_i^{\phi}(\tilde{x}_{i,j})$
**end for**
$\overline{\phi - \mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}^b - \hat{\mathcal{I}}^b), \overline{\mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}\hat{\mathcal{I}}^b$.
Construct the $1-\alpha$-confidence interval as

$$[\hat{\phi} - z_{1-\alpha/2}S, \hat{\phi} + z_{1-\alpha/2}S],$$

where

$$S^2 = \begin{cases} S_1^2 & \text{if} \quad S_1^2 \geq 0 \\ S_2^2 & \text{if} \quad S_1^2 < 0 \end{cases}, \tag{16}$$

$$S_1^2 = \sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\mathcal{I}_i^{\phi}(X_{i,j}))^2 + \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right), \tag{17}$$

$$S_2^2 = \sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\mathcal{I}_i^{\phi}(X_{i,j}))^2, \tag{18}$$

and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal.

**Lemma B.13** (Chebyshev's Inequality)**.** *For a random variable $X$ with finite variance $VarX$ and for $t > \mathbb{E}X$, we have*

$$\mathbb{P}(X > t) \leq \frac{VarX}{(t - \mathbb{E}X)^2}.$$

**Lemma B.14.** *For a random variable $X \geq 0$,*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)\mathrm{d}t.$$

**Theorem B.15.** *Under the same assumptions as Theorem B.2, consider the Orthogonal Bootstrap debiasing estimator defined in Algorithm 3. If the number of Monte Carlo replications $B \geq Cn^\alpha$ for some absolute constant $C > 0$ and $\alpha \geq 0$, then $Var_*(X) = O_p(\frac{1}{n^{3+\alpha}})$.*

*Proof.* Denote $\mathcal{I} := \frac{1}{n^2}\sum_{i=1}^m \sum_{j=1}^n (\mathcal{I}_i^\phi(X_{i,j}))^2 > 0$, and

$$X := \frac{1}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^B \left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right).$$

As $S^2 > 0$, we have $\mathbb{E}_*(S^2) = \mathcal{I} + \mathbb{E}_*(X) > 0$ and $\mathcal{I}$ is of order $O_p(\frac{1}{n})$. Now our improved estimator is

$$\tilde{S} := \mathcal{I} + X1_{X > -\mathcal{I}}.$$

For any $t \geq \mathcal{I}$, by Lemma B.13 we have

$$\mathbb{P}_*(X < -t) \leq \frac{\mathrm{Var}_* X}{(t + \mathbb{E}_* X)^2}.$$

As $-X1_{X<-\mathcal{I}} \geq 0$, by Lemma B.14,

$$\begin{aligned}
\mathbb{E}_*(X1_{X<-\mathcal{I}}) &= -\int_0^\infty \mathbb{P}_*(X1_{X<-\mathcal{I}} < -t)\mathrm{d}t \\
&= -\int_0^{\mathcal{I}} \mathbb{P}_*(X < -\mathcal{I})\mathrm{d}t - \int_{\mathcal{I}}^\infty \mathbb{P}_*(X < -t)\mathrm{d}t \\
&\geq -\frac{\mathcal{I}\mathrm{Var}_* X}{(\mathcal{I} + \mathbb{E}_* X)^2} - \frac{\mathrm{Var}_* X}{\mathcal{I} + \mathbb{E}_* X}.
\end{aligned}$$

With $\mathbb{E}_* X = O_p(\frac{1}{n^{\frac{3}{2}}})$ and $\mathrm{Var}_* X = O_p(\frac{1}{Bn^3})$, we get $\mathbb{E}_* X1_{X<-\mathcal{I}} = O_p(\frac{1}{Bn^2})$. Further,

$$\mathrm{Var}_*\tilde{S} \leq \mathrm{Var}_* X + (\mathbb{E}_* X)^2 - (\mathbb{E}_* X1_{X>-\mathcal{I}})^2 \leq \mathrm{Var}_* X + 2\mathbb{E}_* X\mathbb{E}_* X1_{X<-\mathcal{I}}.$$

As $\mathbb{E}_* X = O_p(\frac{1}{n^{\frac{3}{2}}})$ and $\mathbb{E}_* X1_{X<-\mathcal{I}} = O_p(\frac{1}{Bn^2})$, $\mathrm{Var}_*\tilde{S}$ is still of order $O_p(\frac{1}{Bn^3})$, which is the same as the order of $\mathrm{Var}_* S$. □

### B.5. Verifying the Assumptions using Kernel Mean Embedding

Now we investigate cases where we can theoretically prove Assumption B.1 and Assumption B.2.

Let us consider the performance measure defined in Equation (2)

$$\phi(F_1, \cdots, F_m) = h(\mu_1(F_1), \cdots, \mu_m(F_m))$$

where $\mu_i$ are kernel mean embeddings using kernels $k_i$, $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$ where $\mathcal{H}_i$ is the reproducing kernel Hilbert space respect to the kernel $k_i$ and $h : \mathcal{H} \to \mathbb{R}$ is a functional on $\mathcal{H}$. Denote $\mu := \mu_1 \times \cdots \mu_m$ and $F = F_1 \times \cdots F_m$ for simplicity. Suppose Assumption 2.1 and Assumption 2.2 holds for the performance measure. For the readers convenience, we restate the two assumptions here.

**Assumption B.16.** There exists kernel mean embeddings $\mu_i : \mathcal{F}_i \to \mathcal{H}_i$ which maps $F_i$ into $\mathcal{H}_i$ for $i = 1, \cdots, m$. Moreover, for all $i = 1, \cdots, m$, $\mathbb{E}k_i(X_i, X_i)^4 < \infty$ and $\mathbb{E}k_i(X_i, Y_i)^4 < \infty$ where $X_i, Y_i$ are independent samples from $F_i$.

It is easy to see if $\mathbb{E}k_i(X_i, X_i)^4 < \infty$ where $X_i$ are samples from $F_i$, then $F_i$ can be embedded into RKHS with kernel $k_i$ by Lemma A.6.

**Assumption B.17.** The non-constant functional $h : \mathcal{H}_1 \times \cdots \times \mathcal{H}_m \to \mathbb{R}$ is of class $C^1$ and its derivative is Lipschitz in the sense that $|Dh(x_1)(v) - Dh(x_2)(v)| \leq L\|x_1 - x_2\|_{\mathcal{H}}\|v\|_{\mathcal{H}}$. Moreover, $\|\partial_i h(\mu(F))^4\|_{\mathcal{H}_i} < \infty$.

The Lipschitz condition is the key ingredient for our result. If $h$ is of class $C^1$ and every partial derivative of $h$ is Lipschitz, then by Theorem A.8 the Lipschitz condition is satisfied. The last assumption is important for controlling the moment of influence function.

**Theorem B.18.** *Under Assumption 2.1 and Assumption 2.2, Assumption B.1 and Assumption B.2 holds for the performance measure given by (2).*

*Proof.* First we prove Assumption B.1. Denote $x := \mu(F)$ and $y := \mu(\hat{F})$. Then as $h$ is of class $C^1$, we can write

$$h(y) = h(x) + Dh(x)[y - x] + \delta,$$

where the remainder term can be controlled by the Lipschitz property of $Dh$. Specifically,

$$\begin{aligned}\delta &= h(y) - h(x) - Dh(x)[y - x] \\ &= \int_0^1 Dh(x + t(y - x))\mathrm{d}t[y - x] - Dh(x)[y - x] \\ &= \int_0^1 (Dh(x + t(y - x)) - Dh(x))\mathrm{d}t[y - x]\end{aligned}$$

so $|\delta| \leq \int_0^1 |Dh(x + t(y - x)) - Dh(x)|\mathrm{d}t[y - x] \leq \frac{L}{2}\|y - x\|_{\mathcal{H}}^2$. Now we control the term $\|y - x\|_{\mathcal{H}}^2$:

$$\|y - x\|_{\mathcal{H}}^2 = \max_i \int\int k_i(y_1, y_2)\mathrm{d}[\hat{F}_i(y_1) - F_i(y_1)]\mathrm{d}[\hat{F}_i(y_2) - F_i(y_2)]$$

As $\mathbb{E}k_i(X_i, X_i)^4 < \infty$ and $\mathbb{E}k_i(X_i, Y_i)^4 < \infty$ where $X_i, Y_i$ are independent samples from $F_i$, by Lemma A.3, we have $\mathbb{E}\|y - x\|_{\mathcal{H}}^8 = O(\frac{1}{n^4})$. Thus $\mathbb{E}\delta^2 = o(\frac{1}{n})$.

Next, by Theorem A.8 we can write

$$Dh(x)(v) = \sum_{i=1}^m \partial_i h(x)(v_i).$$

Noting that we can view a linear functional on a Hilbert space as an element in this Hilbert space via Riesz's representation theorem, we have $\partial_i h(x)(y_i) = \langle \partial_i h(x), y_i \rangle_{\mathcal{H}_i} = \mathbb{E}_{X_i \sim \hat{F}_i} \partial_i h(x)(X_i)$. So the influence function is the centered version of $\mathbb{E}_{X_i \sim \hat{F}_i} \partial_i h(x)(X_i)$, i.e. $\mathbb{E}_{X_i \sim \hat{F}_i} \partial_i h(x)(X_i) - \mathbb{E}_{X_i \sim F_i} \partial_i h(x)(X_i)$. Therefore if we set $\phi_i = \partial_i h(x) - \mathbb{E}_{X_i \sim F_i} \partial_i h(x)(X_i)$, we can obtain

$$\phi(\hat{F}_1, \cdots, \hat{F}_m) = \phi(F_1, \cdots, F_m) + \sum_{i=1}^m \int \phi_i(x)\mathrm{d}\hat{F}_i(x) + \delta.$$

Now the condition $\|\partial_i h(x)^4\|_{\mathcal{H}} < \infty$ implies $\mathbb{E}_{X_i \sim F_i}(\phi_i(X_i))^4 < \infty$, and the condition that $h$ is not constant is equivalent to $\mathrm{Var}_{X_i \sim F_i} \phi_i(X_i) > 0$.

Next we prove Assumption B.2. Denote $y := \mu(\hat{F})$ and $z := \mu(\hat{F}^b)$. Again we can write

$$h(z) = h(y) + Dh(y)[z - y] + \epsilon,$$

where

$$\begin{aligned}\epsilon &= h(z) - h(y) - Dh(y)[z - y] \\ &= \int_0^1 Dh(y + t(z - y))\mathrm{d}t[z - y] - Dh(y)[z - y] \\ &= \int_0^1 (Dh(y + t(z - y)) - Dh(y))\mathrm{d}t[z - y]\end{aligned}$$

so

$$\begin{aligned}|\epsilon| &\leq \int_0^1 |Dh(y+t(z-y)) - Dh(y)|\mathrm{d}t[z-y] \\ &\leq \frac{L}{2}\|z-y\|_{\mathcal{H}}^2\end{aligned}$$

Again we control the term $\|z-y\|_{\mathcal{H}}^2$:

$$\begin{aligned}\|z-y\|_{\mathcal{H}}^2 &= \max_i \frac{1}{n^2}\sum_{j,k}^n \left(k_i(x_{i,j}^b, x_{i,k}^b) - 2k_i(x_{i,j}^b, x_{i,k}) + k_i(x_{i,j}, x_{i,k})\right) \\ &= \max_i \int\int k_i(y_1,y_2)\mathrm{d}[\hat{F}_i^b(y_1) - \hat{F}_i(y_1)]\mathrm{d}[\hat{F}_i^b(y_2) - \hat{F}_i(y_2)]\end{aligned}$$

As $\mathbb{E}k_i(X_i,X_i)^4 < \infty$ and $\mathbb{E}k_i(X_i,Y_i)^4 < \infty$ where $X_i, Y_i$ are independent samples from $F_i$, by Markov's inequality, we have $\mathbb{E}_* k_i(X_i^b, X_i^b)^4 < \infty$ and $\mathbb{E}_* k_i(X_i^b, Y_i^b)^4 < \infty$ with high probability where $X_i^b, Y_i^b$ are independent samples from $\hat{F}_i$. By Lemma A.2 and Lemma A.3, we have $\mathbb{E}_*\|y-x\|_{\mathcal{H}}^4 = O_p(\frac{1}{n^2})$ and $\mathbb{E}_*\|y-x\|_{\mathcal{H}}^8 = O_p(\frac{1}{n^4})$.

Again, by Theorem A.8 we can write

$$Dh(y)(v) = \sum_{i=1}^m \partial_i h(y)(v_i),$$

and we have $\partial_i h(y)(z_i) = \langle \partial_i h(y), z_i\rangle_{\mathcal{H}_i} = \mathbb{E}_{X_i\sim\hat{F}_i^b}\partial_i h(y)(X_i)$. So the influence function is the centered version of $\mathbb{E}_{X_i\sim\hat{F}_i^b}\partial_i h(y)(X_i)$, i.e. $\mathbb{E}_{X_i\sim\hat{F}_i^b}\partial_i h(y)(X_i) - \mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(y)(X_i)$. Therefore if we set $\mathcal{I}_i^\phi = \partial_i h(y) - \mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(y)(X_i)$, we can obtain

$$\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b) = \phi(\hat{F}_1, \cdots, \hat{F}_m) + \sum_{i=1}^m \int \mathcal{I}_i^\phi(x)\mathrm{d}\hat{F}_i^b(x) + \epsilon.$$

Now we prove $\mathbb{E}[(\mathcal{I}_i^\phi - \phi_i)^4(X_{i,1})] = o(1)$. First note that

$$\begin{aligned}&\mathbb{E}[(\mathcal{I}_i^\phi - \phi_i)^4(X_{i,1})] \\ &= \mathbb{E}[(\partial_i h(y) - \mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(y)(X_i) - \partial_i h(x) + \mathbb{E}_{X_i\sim F_i}\partial_i h(x)(X_i))^4(X_{i,1})] \\ &\leq 8\mathbb{E}(\partial_i h(y)(X_{i,1}) - \partial_i h(x)(X_{i,1}))^4 + 8\mathbb{E}(\mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(y)(X_i) - \mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(x)(X_i))^4 \\ &+ 8\mathbb{E}(\mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(x)(X_i) - \mathbb{E}_{X_i\sim F_i}\partial_i h(x)(X_i))^4\end{aligned}$$

We deal with each term separately. For the first term, noting that $\mathbb{E}k_i(X_i,X_i)^4 < \infty$ where $X_i \sim F_i$ and $\mathbb{E}\|y-x\|_{\mathcal{H}}^8 = O(\frac{1}{n^4})$, we have

$$\begin{aligned}\mathbb{E}(\partial_i h(y)(X_{i,1}) - \partial_i h(x)(X_{i,1}))^4 &\leq L^4\mathbb{E}\|y-x\|_{\mathcal{H}}^4\|k_i(\cdot, X_{i,1})\|_{\mathcal{H}}^4 \\ &\leq L^4\sqrt{\mathbb{E}\|y-x\|_{\mathcal{H}}^8\mathbb{E}\|k_i(\cdot,X_{i,1})\|_{\mathcal{H}}^8} \\ &= L^4\sqrt{\mathbb{E}\|y-x\|_{\mathcal{H}}^8\mathbb{E}k_i^4(X_{i,1},X_{i,1})} \\ &= O(\frac{1}{n^2})\end{aligned}$$

For the second term, we calculate $\|y\|_{\mathcal{H}}^2 = \max_i \frac{1}{n^2}\sum_{j,k} k_i(x_{i,j},x_{i,k})$. As $\mathbb{E}k_i(X_i,X_i)^4 < \infty$ and $\mathbb{E}k_i(X_i,Y_i)^4 < \infty$ where $X_i, Y_i$ are independent samples from $F_i$, $\mathbb{E}\|y\|_{\mathcal{H}}^8 < \infty$. So,

$$\begin{aligned}\mathbb{E}(\mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(y)(X_i) - \mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(x)(X_i))^4 &\leq L^4\mathbb{E}\|y-x\|_{\mathcal{H}}^4\|y\|_{\mathcal{H}}^4 \\ &\leq L^4\sqrt{\mathbb{E}\|y-x\|_{\mathcal{H}}^8\mathbb{E}\|y\|_{\mathcal{H}}^8} \\ &= O(\frac{1}{n^2})\end{aligned}$$

For the last term, noting that $\|\partial_i h(x)\|_{\mathcal{H}}^4 < \infty$ and $\mathbb{E}\|y-x\|_{\mathcal{H}}^4 = O(\frac{1}{n^2})$

$$\begin{aligned}\mathbb{E}(\mathbb{E}_{X_i\sim\hat{F}_i}\partial_i h(x)(X_i) - \mathbb{E}_{X_i\sim F_i}\partial_i h(x)(X_i))^4 &= \mathbb{E}(\partial_i h(x)(y-x))^4 \\ &\le \|\partial_i h(x)\|_{\mathcal{H}}^4 \mathbb{E}\|y-x\|_{\mathcal{H}}^4 \\ &= O(\frac{1}{n^2})\end{aligned}$$

Therefore the proof is completed. □

*Remark* B.19. From the proof we can see that the global Lipschitz condition Assumption 2.2 can be relaxed to a Lipschitz condition within a subset of $U \subset \mathcal{H}$

$$U := U_1 \times \cdots \times U_m,$$

where each $U_i$ is the convex hull of $\mu_i(F_i), \mu_i(\hat{F}_i)$ and $\mu_i(\hat{F}_i^b)$.

A specific function that satisfies our assumption is

$$\phi(\mu_1(F_1), \cdots, \mu_m(F_m)) = \sum_{i=1}^m \langle f_i, \mu_i(F_i)\rangle_{\mathcal{H}_i}^2,$$

where $f_i : \mathbb{R} \to \mathbb{R}$ is a function that can be embeded in $\mathcal{H}_i$ and $\|f_i^4\|_{\mathcal{H}_i} < \infty$ for all $i \in [m]$. We have

$$\partial_i \phi(\mu(F)) = 2\langle f_i, \mu_i(F_i)\rangle f_i,$$

which is nontrivial and Lipschitz continuous with Lipschitz constant $2\|f_i\|_{\mathcal{H}_i}^2$, so $D\phi$ is also Lipschitz continuous by Theorem A.8. Moreover, $\|f_i^4\|_{\mathcal{H}} < \infty$ implies $\|\partial_i\phi(\mu(F))^4\|_{\mathcal{H}_i} < \infty$. A direct consequence of this example is that $(\mathbb{E}X)^2$ can be proved to be simulated by our method efficiently.

Another example is the finite-horizon performance measure proposed in (Lam & Qian, 2022). For convenience we restate the performance measure below. The performance measure is of the form

$$\phi(F_1, \ldots, F_m) = \mathbb{E}_{F_1,\ldots,F_m} h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$$

where $\boldsymbol{X}_i = (X_i(1), \cdots, X_i(T_i))$ represents the $i$-th input process consisting of $T_i$ i.i.d. random variables distributed under $F_i$, each $T_i$ being a deterministic time, and $h$ is a performance function which satisfies Assumption 8 and Assumption 9 in (Lam & Qian, 2022). It can actually be regarded as a special form of our performance measure as $\phi$ can also be written as

$$\phi = \langle h, \mu_1(F_1) \otimes \cdots \otimes \mu_m(F_m)\rangle_{\mathcal{H}_1\otimes\cdots\otimes\mathcal{H}_m}.$$

Noting that $D^s\phi = 0$ where $s = \sum_{i=1}^m T_i + 1$, by Theorem A.9 we can see that the local Lipschitz condition is satisfied.

## C. Additional Experiments

### C.1. Debiasing

In Figure 3 we provide the root mean square error (RMSE) for the four debiasing tasks in Section 4.1. The precise definition of RMSE and BIAS is

$$\text{RMSE} = \sqrt{\sum_{i=1}^{1000}(\hat{\phi}-\phi)^2}, \quad \text{BIAS} = \sum_{i=1}^{1000}|\hat{\phi}-\phi|.$$

We also provide the median (50% percentile) of ten bootstrap resampling procedure in Table 3.

#### C.1.1. Calculation of Influence Function for Constrained Optimization Problems

In this section we provide a method of calculating the influence function for constrained optimization problem with randomness.

*Table 3.* Debiasing performances with different bootstrap methods: Standard Bootstrap and Orthogonal Bootstrap.

| | $B$ | Ellipsoidal | | Polynomial | | Entropy | | Optimization | |
|---|---|---|---|---|---|---|---|---|---|
| | | **RMSE** | **BIAS** | **RMSE** | **BIAS** | **RMSE** | **BIAS** | **RMSE** | **BIAS** |
| Standard Bootstrap | 2 | 7.23 | 197.9 | 17.57 | 501.4 | 0.893 | 24.23 | 1.734 | 49.52 |
| Orthogonal Bootstrap | 2 | 6.63 | 168.9 | 14.92 | 395.8 | 0.836 | 21.15 | 1.598 | 45.44 |
| Standard Bootstrap | 5 | 6.80 | 178.8 | 15.69 | 428.7 | 0.862 | 22.26 | 1.404 | 38.57 |
| Orthogonal Bootstrap | 5 | 6.61 | 167.0 | 14.79 | 388.2 | 0.833 | 20.92 | 1.355 | 36.70 |
| Standard Bootstrap | 10 | 6.68 | 172.3 | 14.99 | 404.8 | 0.840 | 21.53 | 1.281 | 33.96 |
| Orthogonal Bootstrap | 10 | 6.61 | 166.5 | 14.66 | 383.3 | 0.833 | 20.92 | 1.246 | 32.82 |
| Naive Estimator | | 10.30 | 271.1 | 25.84 | 634.7 | 1.802 | 51.18 | 4.973 | 153.8 |

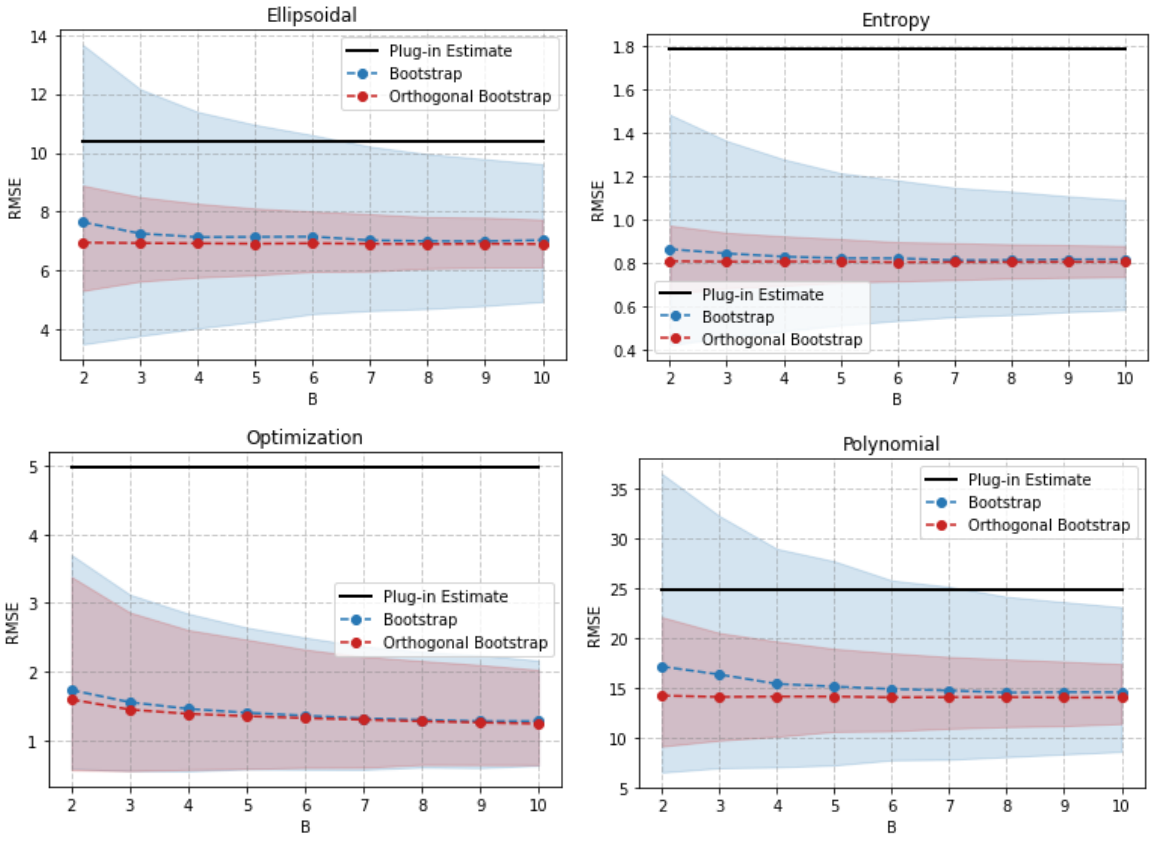


*Figure 3.* Orthogonal Bootstrap can significantly reduce the simulation output for the examples shown in (Ma & Ying, 2022) when the number of Bootstrap resampling is limited. The $x$-axis represents the time of Bootstrap resampling and $y$-axis denotes the root mean square error produced by the estimation. The shaded area represents the 80% quantile interval for repeated simulations. Orthogonal Bootstrap can significantly reduce the simulation variance.

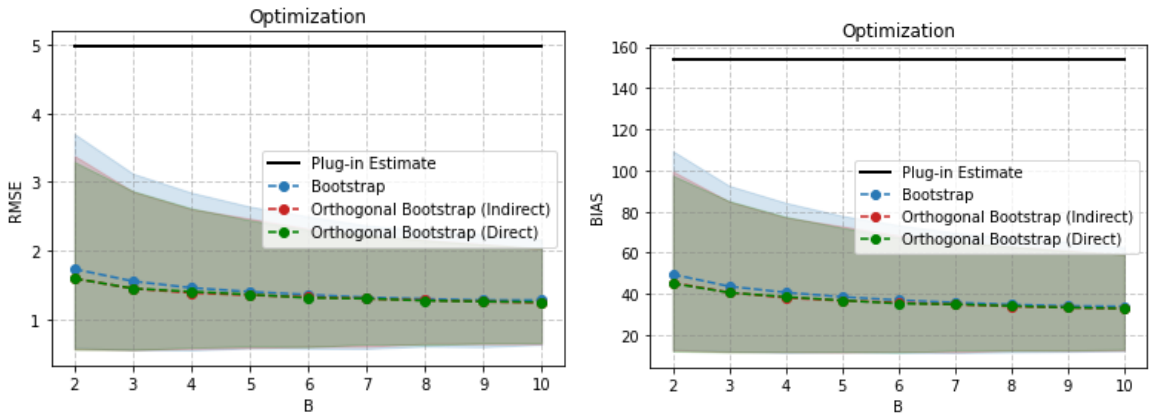


*Figure 4.* Comparison of indirect and direct influence function calculation in the constrained optimization problem.

For a general constrained optimization problem with randomness

$$\begin{aligned} \text{Minimize} \quad & f(x,\xi) \\ \text{subject to} \quad & h_i(x,\xi) \le 0, \quad i = 1,2,\dots,m \\ & g_j(x,\xi) = 0, \quad j = 1,2,\dots,p \\ \text{where} \quad & x \in \mathbb{R}^n \text{ and } \xi \in \mathbb{R}^k \text{ is a random vector,} \end{aligned}$$

consider its Lagrangian

$$L(x,\mu,\nu) = f(x,\xi) + \sum_{i=1}^{m} \mu_i g_i(x,\xi) + \sum_{j=1}^{p} \nu_j h_j(x,\xi)$$

where $\mu_i$ are the dual variables on the equality constraints and $\nu_j \ge 0$ are the dual variables on the inequality constraints. Regard $\xi$ as a fixed parameter for now, the KKT conditions for stationarity, primal feasibility, and complementary slackness are

$$\begin{aligned} & \nabla_x f(x^*,\xi) + \sum_{i=1}^{m} \mu_i^* \nabla_x g_i(x^*,\xi) + \sum_{j=1}^{p} \nu_j^* \nabla_x h_j(x^*,\xi) = 0 \\ & h_j(x^*,\xi) = 0, \quad j = 1,2,\dots,p \\ & \mu_i^* g_i(x^*,\xi) = 0, \quad i = 1,2,\dots,m \end{aligned}$$

where $x^*$, $\mu_i^*$, and $\mu_j^*$ are the optimal primal and dual variables. Now, unfix $\xi$. Taking the differentials of these conditions yields the equation

$$\begin{aligned} & \left( \nabla_{xx} f(x^*,\xi) + \sum_{i=1}^{m} \mu_i^* \nabla_{xx} g_i(x^*,\xi) + \sum_{j=1}^{p} \nu_j^* \nabla_{xx} h_j(x^*) \right) \mathrm{d}x^* + \sum_{i=1}^{m} \nabla_x g_i(x^*) \mathrm{d}\mu_i^* + \sum_{j=1}^{p} \nabla_x h_j(x^*) \mathrm{d}\nu_j^* \\ & + \left( \nabla_{x\xi} f(x^*,\xi) + \sum_{i=1}^{m} \mu_i^* \nabla_{x\xi} g_i(x^*,\xi) + \sum_{j=1}^{p} \nu_j^* \nabla_{x\xi} h_j(x^*) \right) \mathrm{d}\xi = 0 \\ & \nabla_x h_j(x^*,\xi)\mathrm{d}x^* + \nabla_\xi h_j(x^*,\xi)\mathrm{d}\xi = 0, \quad j = 1,2,\dots,p \\ & g_i(x^*,\xi)\mathrm{d}\mu_i^* + \mu_i^* \nabla_x g_i(x^*,\xi)\mathrm{d}x^* + \mu_i^* \nabla_\xi g_i(x^*,\xi)\mathrm{d}\xi = 0, \quad i = 1,2,\dots,m \end{aligned}$$

The optimal primal and dual variables can be calculated via numerical method. Therefore, we can solve the above equation to determine $\frac{\partial x^*}{\partial \xi}$, $\frac{\partial \mu^*}{\partial \xi}$, and $\frac{\partial \nu^*}{\partial \xi}$. If these quantities behave well, then we can combine them with the chain rule to determine the influence function with respect to the distribution of $\xi$.

The optimization problem we have selected for this study has a particularly tractable solution, making it amenable to differentiation and, consequently, to the computation of the influence function. In Figure 4, we present a comparative analysis of these two approaches for calculating the influence function. As anticipated, calculating the influence function via the explicit solution of the constrained optimization problem yields the same result in comparison to the implicit way of determining the influence function. Therefore, for optimization problem without a tractable solution, our implicit way of determining the influence function can be powerful.

### C.2. Confidence Interval Construction

The confidence interval constructed by the bootstrap method is

$$[\hat{\phi} - z_{1-\alpha/2}\sqrt{\text{Var}\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)\right)}, \hat{\phi} + z_{1-\alpha/2}\sqrt{\text{Var}\left(\phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)\right)}],$$

where $z_{1-\alpha/2}$ being the $(1-\alpha/2)$-quantile of the standard normal, $\hat{\phi} = \phi(\hat{F}_1, \cdots, \hat{F}_m)$ is the plug-in estimator of $\phi(F_1, \cdots, F_m)$ and $\hat{F}_i = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_{i,j}}$. Therefore using Orthogonal Bootstrap to simulate the variance, we arrive at Algorithm 4 for constructing confidence interval.

**Algorithm 4** Confidence Interval Construction via Orthogonal Bootstrap

**Input**: A generic performance measure $\phi(F_1, \cdots, F_m)$, i.i.d samples $\{X_{i,1}, \cdots, X_{i,n_i}\} \in \mathbb{R}^{d_1}$ of $F_i$, influence function $\mathcal{I}_i^{\phi}$ of $\phi$ respect to $\hat{F}_i$, and confidence level $\alpha$.
**Output**: $(1-\alpha)$ prediction interval of $\phi(F_1, \cdots, F_m)$.

$\hat{\phi} \leftarrow \phi(\hat{F}_1, \cdots, \hat{F}_m)$, where $\hat{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}}$
**for** b=1:B **do**
  **for** i=1:m **do**
    Sample $\{x_{i,1}^b, \cdots, x_{i,n_i}^b\}$ i.i.d from $\hat{F}_i$
  **end for**
  $\hat{\phi}^b \leftarrow \phi(\hat{F}_1^b, \cdots, \hat{F}_m^b)$, where $\hat{F}_i^b = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{x_{i,j}^b}$
  Calculate $\hat{\mathcal{I}}^b = \sum_{i=1}^{m} \frac{1}{n_i}\sum_{j=1}^{n_i} \mathcal{I}_i^{\phi}(\tilde{x}_{i,j})$
**end for**
Calculate $\overline{\phi - \mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}^b - \hat{\mathcal{I}}^b)$, $\overline{\mathcal{I}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}\hat{\mathcal{I}}^b$.
Construct the $1-\alpha$-prediction interval as $[\hat{\phi} - z_{1-\alpha/2}S, \hat{\phi} + z_{1-\alpha/2}S]$, where

$$S^2 = \sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\mathcal{I}_i^{\phi}(X_{i,j}))^2 + \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)^2 + \frac{2}{B}\sum_{b=1}^{B}\left(\hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}}\right)\left(\hat{\mathcal{I}}^b - \overline{\mathcal{I}}\right), \tag{19}$$

and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal.

### C.3. Real Data

#### C.3.1. Prediction Interval Construction via Orthogonal Bootstrap

Suppose that the input data and target data, $\boldsymbol{X}$ and $\boldsymbol{y}$, are observed, and a neural network is trained on these data to produce an estimated regression function $\hat{f}$. Then $B$ Monte Carlo replications are collected, and a neural network is trained on each to produce regression functions $f^b$, $b = 1, \ldots, B$. Now, suppose a new feature vector, $\boldsymbol{x}_{test}$, is observed and it is desired to provide a point estimate and a prediction interval (PI) for its unknown target value, $y_{test}$.

Prediction interval is slightly different from confidence interval because the model possess inherent irreducible error (Contarino et al., 2022; Khosravi et al., 2011). To construct prediction interval, besides the term $S^2$ in confidence interval construction, we need to add an additional term $\sigma^2$ for the irreducible error of the model. Specifically, we follow the pivot bootstrap method described in (Contarino et al., 2022). The irreducible error $\sigma^2$ can be estimated from the residuals of the out-of-sample predictions. For a bootstrap resample $\{\boldsymbol{X}_i^b\}_{i=1}^n$, the corresponding set of out-of-bag observations is

$\{\boldsymbol{X}_i | \boldsymbol{X}_i \notin \{\boldsymbol{X}_i^b\}_{i=1}^n\}$ and is denoted here as $\boldsymbol{X}_{oob}^b$. Then, for each bootstrap resample, $\sigma_b^2$ is:

$$\sigma_b^2 = \frac{\sum_{\boldsymbol{x} \in \boldsymbol{X}_{oob}^b} (y - f^b(\boldsymbol{x}))^2}{n^b}$$

where $n^b$ is the number of out-of-bag resamples. We summarize our algorithm in Algorithm 5.

**Algorithm 5** Prediction Interval Construction via Orthogonal Bootstrap

**Input**: Number of training data $n$, training data $\boldsymbol{X}$ and $\boldsymbol{y}$, test observation $\boldsymbol{x}_{test}$, learning algorithm $L$, desired number of Monte Carlo replications $B$, and desired coverage probability $1 - \alpha$.
**Output**: $(1 - \alpha)$ prediction interval of $y_{test}$.

Train learning algorithm $L$ on $\boldsymbol{X}$ and $\boldsymbol{y}$; denote the trained regressor as $\hat{f}$
$\hat{\phi} \leftarrow \hat{f}(\boldsymbol{x}_{test})$
For every input data $\boldsymbol{X}_i \in \boldsymbol{X}$, calculate the influence function $\mathcal{I}(\boldsymbol{X}_i)$ of $\hat{\phi}$,
**for** b=1:B **do**
  Generate bootstrap sample $(\boldsymbol{X}^b, \boldsymbol{y}^b)$ from $(\boldsymbol{X}, \boldsymbol{y})$
  Determine the out of bag samples $(\boldsymbol{X}_{oob}^b, \boldsymbol{y}_{oob}^b)$
  Train learning algorithm $L$ on $\boldsymbol{X}^b$ and $\boldsymbol{y}^b$; denote the trained regressor as $f^b$
  $\hat{\phi}^b \leftarrow f^b(\boldsymbol{x}_{test})$
  $\hat{\mathcal{I}}^b = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(\boldsymbol{X}_i^b)$
**end for**
$\overline{\phi - \mathcal{I}} \leftarrow \frac{1}{B} \sum_{b=1}^B (\hat{\phi}^b - \hat{\mathcal{I}}^b), \overline{\mathcal{I}} \leftarrow \frac{1}{B} \sum_{b=1}^B \hat{\mathcal{I}}^b$.
Construct the $1 - \alpha$-confidence interval as

$$[\hat{\phi} - z_{1-\alpha/2} \sqrt{S^2 + \sigma^2}, \hat{\phi} + z_{1-\alpha/2} \sqrt{S^2 + \sigma^2}],$$

where

$$S^2 = \begin{cases} S_1^2 & \text{if} \quad S_1^2 \geq 0 \\ S_2^2 & \text{if} \quad S_1^2 < 0 \end{cases}, \tag{20}$$

$$S_1^2 = \frac{1}{n^2} \sum_{i=1}^n (\mathcal{I}(\boldsymbol{X}_i))^2 + \frac{1}{B} \sum_{b=1}^B \left( \hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}} \right)^2 + \frac{2}{B} \sum_{b=1}^B \left( \hat{\phi}^b - \hat{\mathcal{I}}^b - \overline{\phi - \mathcal{I}} \right) \left( \hat{\mathcal{I}}^b - \overline{\mathcal{I}} \right), \tag{21}$$

$$S_2^2 = \frac{1}{n^2} \sum_{i=1}^n (\mathcal{I}(\boldsymbol{X}_i))^2, \tag{22}$$

and $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal.

### C.3.2. DETAILS

In this section we provide the details of the real data experiments in section. For all real data examples, we employ the Adam optimizer with default hyperparameters, with the exception of setting the weight decay to 0.01. The training loss is set to be the squared loss, i.e. $L(\boldsymbol{x}, \theta) = (f_\theta(\boldsymbol{x}) - y)^2$, where $f_\theta(\boldsymbol{x})$ is parameterized by a two-layer neural network with hidden dimension 100 .

For the Yacht dataset, we utilize a batch size of 64 and train for 500 epochs. We use 80% data for training and 20% for testing. For the Energy dataset, we utilize a batch size of 128 and train for 250 epochs. We use 70% data for training and 30% for testing. For the kin8nm dataset, we utilize a batch size of 256 and train for 150 epochs. We use 95% data for training and 5% for testing. The specific choice of hyperparameters serves the dual purpose of ensuring that the neural networks fit the data effectively while also ensuring that the inherent variance of the model remains within a reasonable range compared to the bootstrapped variance.

The influence function of the model parameters $\theta$ is

$$\mathcal{I}^\theta(\boldsymbol{x}) = -H_\theta^{-1} \nabla_\theta L(\boldsymbol{x}, \theta), \tag{23}$$

therefore by the chain rule of derivatives, the influence function at a particular test point $\boldsymbol{x}_{\text{test}}$ is

$$\mathcal{I}^{\boldsymbol{x}_{\text{test}}}(\boldsymbol{x}) = -\nabla_\theta f_\theta(\boldsymbol{x}_{\text{test}})^T H_\theta^{-1} \nabla_\theta L(\boldsymbol{x}, \theta). \tag{24}$$

We use the conjugate gradient method to calculate the influence function (Koh & Liang, 2017) (Martens, 2010).