Asymptotics of the Bootstrap via Stability with Applications to Inference with Model Selection

Morgane Austern
Microsoft Research
morgane.austern@gmail.com

Vasilis Syrgkanis
Microsoft Research
vasy@microsoft.com

Abstract

One of the most commonly used methods for forming confidence intervals is the empirical bootstrap, which is especially expedient when the limiting distribution of the estimator is unknown. However, despite its ubiquitous role in machine learning, its theoretical properties are still not well understood. Recent developments in probability have provided new tools to study the bootstrap method. However, they have been applied only to specific applications and contexts, and it is unclear whether these techniques are applicable to the understanding of the consistency of the bootstrap in machine learning pipelines. In this paper, we derive general stability conditions under which the empirical bootstrap estimator is consistent and quantify the speed of convergence. Moreover, we propose alternative ways to use the bootstrap method to build confidence intervals with coverage guarantees. Finally, we illustrate the generality and tightness of our results by examples of interest for machine learning including for two-sample kernel tests after kernel selection and the empirical risk of stacked estimators.

1 Introduction

Bootstrap resampling [26], has been one of the most popular techniques for measuring the uncertainty of a statistic, primarily due to its simple algorithmic definition and its conveniency with dealing with opaque statistical procedures that produce a test statistic. For this reason, uncertainty quantification based on bootstrap resampling has been a staple in the machine learning community, starting from the early days of the field [40] and continuing into the deep learning and SVM era [39, 30, 32, 17]. Despite it's widespread use, general conditions for the consistency of the bootstrap for complex non-linear statistics is generally not fully explored and hence, it is not clear when the bootstrap method will accurately capture uncertainty in machine learning pipelines, especially when model selection procedures are involved.

Our goal is to understand the distribution of a large class of non-linear statistic $\hat{\theta}_n := g_n(X_1, \dots, X_n)$, as the samples X_i are drawn from their unknown distribution. Examples of such statistics could be the out-of-sample risk of a machine learning predictor, or the maximum-mean-discrepancy of a two-sample kernel test, or the prediction of a machine learning model at a given point. One approach to approximating this distribution is the empirical bootstrap: sample new observations $Z_1, \dots, Z_n$ independently and uniformly from $\{X_1, \dots, X_n\}$ (with replacement), and define $\hat{\theta}_n^{\text{boot}}$ as the value of the estimator taken at the bootstrap sample $\hat{\theta}_n^{\text{boot}} := g_n(Z_1, \dots, Z_n)$. This procedure can be repeated many times in order to estimate the conditional distribution of $\hat{\theta}_n^{\text{boot}}$. If this distribution is approximately the same as the distribution of $\hat{\theta}_n$, as the sample size n grows, we say that the bootstrap method is consistent. Consistency of the bootstrap can be subsequently utilized to construct confidence intervals for the quantity of interest.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Classical consistency proofs for the bootstrap, either require that the limiting distribution of the statistic be Gaussian, or even more stringently that the statistic is asymptotically linear. Notably, when $\hat{\theta}_n$ is asymptotically normal those intervals are known to be consistent under general conditions; see e.g. [33, 5, 18]. Moreover, classical conditions for such statements to hold are that the statistic is Hadamard differentiable and smooth with respect to the underlying distribution of the random variables. However, these properties can be violated by test statistics that are implicitly defined via machine learning data analysis pipelines and especially when the model selection is involved. Thus it is crucial to provide more generally applicable sufficient conditions for the validity of the bootstrap.

Notably, recent breakthrough results in statstics by Chatterjee [13], have shown that the Lindenberg technique that has been long used to establish Central Limit Theorems, is more widely applicable and can be leveraged to show limit distributional consistency of two random variables more broadly. Notably, this intuition has already been exploited to show that the bootstrap method is consistent in particular applications in the econometric literature, such as the construction of uniform confidence bands. However, these proofs are tailored to the particular application of interest and do not provide general characterizations.

Our main contribution is to provide a set of sufficient stability conditions that imply the consistency of the bootstrap and which go well beyond the existing general conditions via the use of Chatterjee's generalized Lindenberg approach and a smart path interpolation technique. We then apply these general results to derive inference in two machine learning applications: i) bootstrapping the test statistic in two-sample kernel tests, after model selection on the kernel and sample re-use, ii) bootstrapping the risk of a stacked estimator with sample re-use.

Roughly, our sufficient conditions impose unilateral stability properties on the statistic, reminiscent of the stability conditions required for classical concentration inequalities, such as McDiarmid's inequality. In particular, we assume that the functions (g_n) are approximable by three-times differentiable functions whose first, second and third order partial derivatives, with respect to a single observation, taken at $(X_1, \dots, X_n)$, are of respective order $o(n^{-1/3})$, $o(n^{-1/2})$ and $o(n^{-1})$. These conditions assure that the value of $g_n(X_1, \dots, X_n)$ is not oversensitive to the value of a single observation (see Section 3 for a formal exposition). Exploiting these assumptions, we exactly characterize the limiting distribution of the bootstrap estimator $\hat{\theta}_n^{\mathrm{boot}}$, compare it to the distribution of the original statistic $\hat{\theta}_n$ and study how fast the distribution of $\hat{\theta}_n^{\mathrm{boot}}$ converges. Notably, we also discover that when the mean of the observations X_1 is unknown, then the bootstrap method is in general not consistent and we propose corrections to bootstrap based intervals with guaranteed minimum coverage.

Related litterature The empirical bootstrap method was first introduced in a breakthrough paper by Efron [27]. Other bootstraps methods have since been proposed including the multiplier bootstrap [54], the residual bootstrap [21] or the non-remplacement bootstrap method [47]. A vast literature studies the theoretical properties of those techniques with some of the main results synthesized in the following books [33, 21, 36, 3]. Most relevant to us are studies of the asymptotics of the bootstrap method. The consistency of the bootstrap method for linear statistics, t-statistics, Von-Mises functionals and quantiles has been established in [5, 46, 48] and for U-statistics in [1, 55]. Those results, among others, have been extended to high-dimensional regression and M-estimation [6, 45, 10, 2, 24], misspecified models [49], solutions of estimating equations [11] and to robust estimators [15]. In contrast, other works established the poor performance of the bootstrap method for non smooth statistics [25, 4, 5], or for non-sparse high-dimensional regressions [28].

Several recent breakthrough papers studied the consistency of the bootstrap method, both empirical and wild, for the maximum of high-dimensional centered averages with the dimension taken to be growing exponentially fast with the sample size. Notably [18, 19] established the consistency of the bootstrap and Gaussian approximation method when respectively $\log(p_n n)^{7/8} = o(n^{1/8})$ and $\log(p_n n)^{7/6} = o(n^{1/6})$ hold. A series of works have strengthen those results: [20, 23, 41] established the consistency of the multiplier and empirical bootstrap when $\log(pn)^{5/4} = o(n^{1/4})$, [43] established a quasi $\sqrt{n}^{-1}$ rate for the wild bootstrap, [22] built slightly conservative confidence sets with guaranteed coverage under the conditions that $\log(p) = o(n)$ and [16] proved that similar results hold for high-dimensional U-statistics. Those works use a combination of the Stein method, Edgeworth expansions, Lindeberg's method [13] and the Slepian smart interpolation path. We note that the limiting distributions of those statistics are in general not Gaussian [23]. In contrast, our results apply more broadly and are not limited to the study of maximums of centered empirical

averages. We notably apply our results to machine learning estimators that are smoothed arg-minima of an objective function. Other works have studied the accuracy of the bootstrap method for specific statistics whose distributions are known to be asymptotically not Gaussian such as: the operator norm in high dimensions [44, 34, 37], sampled eigenvalues of random matrices in high and moderate dimensions [29] or M-estimators having cube root convergence [9]. The main contrast between this series of work and ours is that, rather than studying the bootstrap method for one specific statistic or application, we seek to establish the asymptotics of the bootstrap method under universal conditions on the estimators (g_n). Our proof builds on a breakthrough method proposed by Chatterjee [13] that generalized the Lindeberg method to a general technique for comparing the expectations of $f(X_{1:n})$ and $f(Y_{1:n})$ of a large class of functions f.

2 Problem Statement

Let (X_i^n) be a triangular array of independent and identically distributed (i.i.d) processes with observations X_i^n taking value in $\mathbb{R}^{d_n}$. Moreover, let $X^n = (X_1^n, \dots, X_n^n)$ denote its n-th row. Consider an estimator $\hat{\theta}_n := g_n(X^n)$, where $g_n : \times_{l=1}^n \mathbb{R}^{d_n} \to \mathbb{R}$ is a measurable function, that we will typically refer to as a *statistic*, and let (g_n) denote the sequence of measurable functions as n grows. To evaluate the performance of this estimator and build confidence intervals, we need to approximate its distribution. In this work, we will analyze the empirical bootstrap method.

Empirical bootstrap Bootstrap samples $Z^n = (Z_1^n, \dots, Z_n^n)$ are sampled with replacement from the observations $\{X_1^n, \dots, X_n^n\}$. This implies that conditionally on X^n the coordinates of Z^n are distributed i.i.d, with $Z_i^n \mid X^n \sim \text{unif}\left(\{X_1^n, \dots, X_n^n\}\right)$, for all $i \in [n]$.

Consistency metric and bootstrap consistency Throughout the paper we denote with $Y^n = (Y_1^n, \dots, Y_n^n)$ an independent copy of X^n. The bootstrap method is said to be consistent for (g_n) if conditionally on X^n the distribution of $g_n(Z^n)$ well-approximates the distribution of $g_n(Y^n)$, as $n \to \infty$. To make this statement rigorous we introduce a metric on the space of probability distributions. First, we define the class of three times continuously differentiable measurable functions with bounded third-order derivatives:

$$\mathcal{F} := \left\{ h \in C^3(\mathbb{R}) \mid \sup_{x \in \mathbb{R}} \left| h^{(i)}(x) \right| \leq 1, \ \ \forall\, 1 \leq i \leq 3 \right\};$$

Given this, we define the distance on the space of probability measures, as the maximum mean discrepancy, where test functions range over the class $\mathcal{F}$:

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{h \in \mathcal{F}} \mathbb{E}_{X \sim \mu, Y \sim \nu} \left[h(X) - h(Y) \right].$$

We remark that $d_{\mathcal{F}}$ is a metric on the space of probability measures of real-valued random variables. Notably for two probability distributions μ and ν if $d_{\mathcal{F}}(\mu, \nu) = 0$ then those distributions are the same $\nu = \mu$. Moreover, the topology defined by $d_{\mathcal{F}}$ is finer than the weak convergence topology. Indeed, for a sequence of distributions (ν_n) if we have $d_{\mathcal{F}}(\nu_n, \mu) \to 0$ then (μ_n) converges weakly to μ: $\nu_n \xrightarrow{d} \mu$. Finally we remark that this metric is related to the classical Levy-Prokhorov distance on probability spaces [7].
Moreover, we use the shorthand notation:

$$d_{\mathcal{F}}\left(\mu, \nu \mid \mathcal{E}\right) := \sup_{h \in \mathcal{F}} \mathbb{E}_{X \sim \mu, Y \sim \nu} \left[h(X) - h(Y) \mid \mathcal{E} \right].$$

We say that *the empirical bootstrap method is consistent for* (g_n) if:

$$d_{\mathcal{F}}\left(g_n(Z^n), g_n(Y^n) \mid X^n\right) \xrightarrow{p} 0.$$

Centering discrepancy and centered bootstrap consistency Notably, an individual bootstrap sample $Z_1^n \mid X^n$, has a slightly different mean $\mathbb{E}\left[Z_1^n \mid X^n\right] = \bar{X}^n := \frac{1}{n} \sum_{i \leq n} X_i^n$, than the one of X_1^n. As we will see this small difference plays a crucial role in determining the consistency of the bootstrap and for this reason it will be useful to define artificially centered versions of the random variables (Z_i^n) and (Y_i^n). A centered bootstrap sample

$$\tilde{Z}_i^n := Z_i^n - \left(\bar{X}^n - \mathbb{E}\left[X_1^n\right]\right)$$

is a bootstrap sample that has been re-centered to artificially have the same mean as X_1^n. Moreover, denote with $\tilde{Y}_i^n$ a corrected version of Y_i^n, artificially re-centered to have the same mean as Z_1^n, i.e.:

$$\tilde{Y}_i^n := Y_i^n + \bar{X}^n - \mathbb{E}\left[X_1^n\right].$$

We say that *the centered bootstrap is consistent for* (g_n) if:

$$d_{\mathcal{F}}\left(g_n(\tilde{Z}^n), g_n(Y^n) \mid X^n\right) \xrightarrow{p} 0.$$

From metric consistency to confidence intervals with nominal coverage We can compare the confidence intervals of two random variables X and Y in terms of their mutual distance $d_{\mathcal{F}}(X, Y)$ (proof in Appendix M.1).

Proposition 1 *Let X and Y be two real-valued random variables and $\mathcal{E}$ any random event. Let $\epsilon > 0$ be a constant then for any Borel set $A \in \mathcal{B}(\mathbb{R})$ the following holds:*

$$P(X \in A_{6\epsilon} \mid \mathcal{E}) \geq P(Y \in A \mid \mathcal{E}) - \frac{d_{\mathcal{F}}(X, Y \mid \mathcal{E})}{\epsilon^3},$$

where we wrote $A_\epsilon := \{x \in \mathbb{R} \mid \exists y \in A \text{ s.t } |x - y| \leq \epsilon\}$. Moreover, if $[a, b]$ is a confidence interval at level $1 - \alpha$ for $Y - \mathbb{E}[Y \mid \mathcal{E}]$, conditional on $\mathcal{E}$, then we have:

$$P\left(X - \mathbb{E}[X \mid \mathcal{E}] \in [a - 6\epsilon, b + 6\epsilon] \mid \mathcal{E}\right) \geq 1 - \alpha - \frac{2d_{\mathcal{F}}(X, Y \mid \mathcal{E})}{\epsilon^3}.$$

For instance, suppose that we care about estimating $\theta_n := \mathbb{E}[g_n(Y^n)]$. Then the bootstrap method, if consistent, can be used to build consistent confidence intervals for θ_n. Indeed since we can estimate the conditional distribution of $\theta_n^{\text{bootstrap}} := g_n(\tilde{Z}^n)$, by drawing sufficiently many bootstrap sub-samples, we can find $C^{\alpha,n}$ such that

$$P\left(\hat{\theta}_n^{\text{bootstrap}} - \mathbb{E}\left[\hat{\theta}_n^{\text{bootstrap}} \mid X^n\right] \in C^{\alpha,n} \mid X^n\right) = 1 - \alpha.$$

Then, if we write $\hat{\theta}_n := g_n(Y^n)$, using the consistency of the bootstrap method we obtain that:

$$\liminf_{\epsilon \downarrow 0} \liminf_{n \to \infty} P\left(\hat{\theta}_n - \theta_n \in C_\epsilon^{\alpha,n}\right) \geq 1 - \alpha.$$

Therefore, confidence intervals built using the bootstrap method achieve asymptotically nominal level of confidence. We note that prior works (e.g. [18, 19]), typically provide a slightly stronger statement that $\liminf_{n\to\infty} P(\hat{\theta}_n - \theta_n \in C^{\alpha,n}) \geq 1 - \alpha$, by proving anti-concentration results on the limit distribution of $\hat{\theta}_n - \theta_n$. Such anti-concentration, allows one to argue that the mass of the random variable $\hat{\theta}_n - \theta_n$ contained in $C_\epsilon^{\alpha,n}$ converges to the mass contained in $C^{\alpha,n}$ as $\epsilon \downarrow 0$ and thereby, $\liminf_{\epsilon\downarrow 0} \liminf_{n\to\infty} P(\hat{\theta}_n - \theta_n \in C_\epsilon^{\alpha,n}) = \liminf_{n\to\infty} P(\hat{\theta}_n - \theta_n \in C^{\alpha,n})$. Given that these results typically require stronger conditions on the statistic and many times Gaussian limits, we omit this step in this work and note that a slightly weaker, albeit still practically useful, statement on coverage is achievable in a more general setup.

From metric consistency to p-values Alternatively, suppose that we want to test if a specific null hypothesis (H_0^n) holds against the alternative (H_1^n). To do so we compute a test statistic $T_n(X_{1:n})$ and determine a rejection region $\mathcal{R}^n$. A crucial quantity to estimate is the p-value $P\left(T_n(X_{1:n}^n) \in \mathcal{R}^n | H_0\right)$. The bootstrap method, if consistent, allows us to upper-bound the p-value by enlarging the rejection region with an infinitesimally small quantity. Indeed according to Proposition 1 if the bootstrap method is consistent then we have

$$\lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} P\left(T_n(Z_{1:n}^n) \in \mathcal{R}_\epsilon^n | X^n\right) \geq \limsup_{n\to\infty} P\left(T_n(X_{1:n}^n) \in \mathcal{R}^n\right).$$

2.1 Notations and definitions

For a scalar random variable X we denote with $\|X\|_{L_p}$, the L_p-norm: $\|X\|_{L_p} := \mathbb{E}[X^p]^{1/p}$. Moreover, for vector $x \in \mathbb{R}^d$, we denote with $\|x\|_p$, the ℓ_p vector norm: $\|x\|_p = \left(\sum_{i=1}^d x_i^p\right)^{1/p}$. For simplicity, given a sequence (x_i), with $x_i \in \mathbb{R}^d$ and a constant $c \in \mathbb{R}^d$, we shorthand

$$x_{1:n} := (x_1, \ldots, x_n), \quad x_{1:n} + c := (x_1 + c, \ldots, x_n + c), \quad c x_{2:n} := (c, x_2, \ldots, x_n).$$

We denote the k-th coordinate of $x_i \in \mathbb{R}^d$ as $x_{i,k}$. For a function $f : \times_{l=1}^n \mathbb{R}^{d_n} \to \mathbb{R}$ and a random variable X taking values in $\mathbb{R}^{d_n}$, we designate $f(\cdot + X)$ the random function: $x_{1:n} \to f(x_{1:n} + X)$.

Lindenberg path interpolation Let $Z^{n,i}$ and $Z^{n,i,x}$ be the following interpolating processes between Z^n and $\tilde{Y}^n$:

$$\begin{aligned} Z^{n,i} &:= \left(\tilde{Y}_1^n, \dots, \tilde{Y}_i^n, Z_{i+1}^n, \dots, Z_n^n\right) \\ Z^{n,i,x} &:= \left(\tilde{Y}_1^n, \dots, \tilde{Y}_{i-1}^n, x, Z_{i+1}^n, \dots, Z_n^n\right) \end{aligned}$$

Higher-order derivatives and bounds If a function f is three-times differentiable then we let:

$$\begin{aligned} \partial_{i,k} f(x_{1:n}) &:= \partial_{x_{i,k}} f(x_{1:n}) \\ \partial^2_{i,k_{1:2}} f(x_{1:n}) &:= \partial_{x_{i,k_1}} \partial_{x_{i,k_2}} f(x_{1:n}) \\ \partial^3_{i,k_{1:3}} f(x_{1:n}) &:= \partial_{x_{i,k_1}} \partial_{x_{i,k_2}} \partial_{x_{i,k_3}} f(x_{1:n}) \end{aligned}$$

Moreover, for a potentially random function f we define the constants:

$$\begin{aligned} M_k^n &:= 2 \left\| X_{1,k}^n \right\|_{L_{12}}, \\ D_{k_1}^n(f) &:= M_{k_1}^n \max_{i \le n} \left\| \partial_{i,k_1} f(Z^{n,i,\bar{X}^n}) \right\|_{L_{12}} \\ D_{k_{1:2}}^n(f) &:= M_{k_1}^n M_{k_2}^n \max_{i \le n} \left\| \partial^2_{i,k_{1:2}} f(Z^{n,i,\bar{X}^n}) \right\|_{L_{12}} \\ D_{k_{1:3}}^n(f) &:= M_{k_1}^n M_{k_2}^n M_{k_3}^n \max_{i \le n} \left\| \max_{x \in \left[\bar{X}^n, \tilde{Y}_1^n\right] \cup \left[\bar{X}^n, Z_1^n\right]} \partial^3_{i,k_{1:3}} f(Z^{n,i,x}) \right\|_{L_{12}} \end{aligned}$$

where for any two vectors $a, b \in \mathbb{R}^d$, we denote with $[a, b]$ their convex closure, i.e.

$$[a, b] := \{t\, a + (1 - t)\, b : t \in [0, 1]\}$$

3 Main Results

If the statistics (g_n) were linear, i.e. $g_n(x_{1:n}) = \sum_{i \le n} x_i$, then the influence of a single observation X_1^n, on the estimate $\hat{\theta}_n$, would depend uniquely on the value of the random variable itself, i.e. $g_n(X^n) - g(0X_{2:n}^n) = X_1^n$. This is not the case for non-linear statistics. For instance, if $g_n(x_{1:n}) = \max\left(\sum_{i \le n} x_{i,1}, \sum_{i \le n} x_{i,2}\right)$, then the influence of observation x_1 depends on the relative size of $\sum_{i>2} x_{i,1}$ and $\sum_{i>2} x_{i,2}$. In this paper, we want to study the asymptotics of the bootstrap method for such non-linear statistics, with complex influence functions. To control the degree of non-linearity, we assume that the statistics (g_n) can be approximated by three times differentiable functions.

Assumption 1 ($\mathbb{C}^3$-Approximability) *There exists a sequence of functions (f_n) with $f_n \in \mathbb{C}^3$ s.t.:*

1. *The functions (f_n) approximate the estimators (g_n):*

$$\left\| f_n(Z^n) - g_n(Z^n) \right\|_{L_1} + \left\| f_n(\tilde{Y}^n) - g_n(\tilde{Y}^n) \right\|_{L_1} \xrightarrow{n \to \infty} 0. \tag{H_0}$$

2. *The first, second and third order derivatives are respectively of size $o(n^{-1/3})$, $o(n^{-1/2})$, $o(n^{-1})$:*

$$\begin{aligned} R_{n,1} &:= n^{1/3} \sum_{k_1 \le d_n} D_{k_1}^n(f_n) = o(1) \qquad R_{n,2} := \sqrt{n} \sum_{k_1, k_2 \le d_n} D_{k_{1:2}}^n(f_n) = o(1) \\ R_{n,3} &:= n \sum_{k_1, k_2, k_3 \le d_n} D_{k_{1:3}}^n(f_n) = o(1). \end{aligned} \tag{H_1}$$

To motivate Assumption 1, we present in Appendix C two illustrating examples of simple estimators which fail to satisfy conditions (H_0), (H_1) and for which the bootstrap method is not consistent.

Under Assumption 1 we study the limiting distribution of the bootstrap statistic and establish that it is asymptotically the same as $g_n(\tilde{Y}^n)$ (proof in Appendix L).

Theorem 1 *Let $(g_n : \times_{l=1}^{n} \mathbb{R}^{d_n} \to \mathbb{R})$ be a sequence of measurable functions. Let (X_i^n) be a triangular array of i.i.d processes such that $X_1^n \in L_{12}$. Under Assumption 1, there is a constant K independent of n such that:*

$$\left\| d_{\mathcal{F}} \left(g_n(Z^n),\, g_n(\tilde{Y}^n) \mid X^n \right) \right\|_{L_1} \leq \epsilon_n := \left\{ \begin{array}{l} \left\| g_n(\tilde{Y}^n) - f_n(\tilde{Y}^n) \right\|_{L_1} + \left\| g_n(Z^n) - f_n(Z^n) \right\|_{L_1} \\ + K \left(R_{n,1}^2 \max\left\{ \frac{1}{n^{1/6}}, R_{n,1} \right\} + R_{n,3} + R_{n,2} \right) \end{array} \right\} \to 0.$$

Remark 1 *We remark that the theorem also holds under slightly modified stability conditions. See Theorem 8 in the appendix for more details. Moreover, the hypothesis that (X_i^n) is an i.i.d process can also be relaxed to assuming that the process (X_i^n) is exchangeable. See Theorem 10 for more details in the appendix. Finally, note that Theorem 1 can also be extended to random estimators (g_n), such as ones obtained by stochastic optimization methods (e.g SGD). See Theorem 10 for more details in the appendix.*

Remark 2 *We note that Assumption 1 controls how stable the function g_n is to the change of one random variable for example X_1. Many concentration inequalities, such as the Mcdiarmid inequality, impose conditions on similar quantities. To be able to derive a central limit theorem one would need to make additional assumptions regulating how much this change $g_n(X_1, X_2, \ldots, X_n) - g_n(0, X_2, \ldots, X_n)$ depends on the other random variables $X_2, \ldots, X_n$ [14]. The function $h_n \to \max(0, \frac{1}{\sqrt{n}} \sum_{i \leq n} X_i - \mathbb{E}(X_1))$ is an example of a statistic satisfying Assumption 1 but is not asymptotically normal.*

Remark 3 *When the mean of the observations $\mathbb{E}(X_1^n)$ is known, we propose in Appendix E an alternative bootstrap method that exploits this information, the centered-bootstrap method, and prove that it is consistent for $g_n(Y_{1:n}^n)$. This is useful for example for estimating p-values for hypothesis testing.*

Theorem 1 guarantees that we can use the bootstrap method to estimate the distribution of $g_n(\tilde{Y}^n)$, which implies that it can also be used to build confidence intervals for $\mathbb{E}\left[g_n(\tilde{Y}^n) \mid X^n \right]$.

Corollary 1 *Let $(g_n : \times_{l=1}^{n} \mathbb{R}^{d_n} \to \mathbb{R})$ be a sequence of measurable symmetric functions. Let (X_i^n) be a triangular array of i.i.d processes such that $X_1^n \in L_{12}$. Assume that (g_n) and (X_i^n) satisfy all the conditions of Theorem 1. Then there is a constant K independent of n such that for ϵ_n as defined in Theorem 1:*

$$\left\| d_{\mathcal{F}} \left(g_n(Z^n) - \mathbb{E}\left[g_n(Z^n) \mid X^n \right],\, g_n(\tilde{Y}^n) - \mathbb{E}\left[g_n(\tilde{Y}^n) \mid X^n \right] \mid X^n \right) \right\|_{L_1} \leq 2\epsilon_n \to 0.$$

The distribution we are interested in is that of $g_n(Y^n)$ rather than $g_n(\tilde{Y}^n)$. Moreover, the shape of the confidence intervals of $g_n(\tilde{Y}^n)$ can be arbitrary compared to the ones of $g_n(Y^n)$, i.e. they are not systematically larger or smaller. This is illustrated in Appendix D by a series of examples. Therefore in Appendix F we propose conditions that guarantee that the two distributions are asymptotically identical and we prove in Appendix G that those conditions are tight.

When those conditions are not met we propose the use of the bootstrap method to build adjusted confidence intervals that are guaranteed to have at least (but not necessarily equal to) some minimum asymptotic coverage. In Appendix H we propose an alternative method to do so by exploiting the bootstrap method for slightly shifted observations. Moreover, in Appendix I we assume that the mean $\mathbb{E}(X_1^n)$ belongs to a certain subset A_n and build robust confidence intervals with a guaranteed coverage level for all potential values of the mean.

According to Corollary 1 the bootstrap method can be used to build consistent confidence intervals for $\mathbb{E}(g_n(\tilde{Y}^n) \mid X^n)$. Therefore if we can bound the distance from $\mathbb{E}(g_n(\tilde{Y}^n) \mid X^n)$ to $\mathbb{E}(g_n(Y^n))$ we can use the bootstrap method to build confidence intervals on the latter. To do so we exploit the fact that under mild conditions $\sqrt{n}\left[\bar{X}^n - \mathbb{E}(X_1^n)\right]$ is approximately normal. We assume that the function $x \to \mathbb{E}\left[g_n(Y^n + x) \right]$ is α-Holder and that the moments of X_1^n are bounded. More formally,

suppose that there is a sequence (C_n) and a constant b such that

$$\left|\mathbb{E}\left[g_n\left(Y^n+\frac{x}{\sqrt{n}}\right)-g_n(Y^n)\right]\right| \leq C_n \max_{k\leq d_n}|x_k|^\alpha, \quad \forall x\in\mathbb{R}^{d_n} \tag{H_3}$$

$$\min_{j\leq d_n}\|X^n_{1,j}\|_{L_3}\geq b, \quad \frac{\log(d_n)^{7/6}\|\sup_{k\leq d_n}|X^n_{1,k}|\|^4_{L_4}}{n^{1/6}}=o(1)$$

Theorem 2 *Let $(g_n:\times_{l=1}^n\mathbb{R}^{d_n}\to\mathbb{R})$ be a sequence of measurable functions satisfying Assumption 1 and (H_3). Denote Σ_n the variance-covariance matrix of X_1^n and (N^n) to be a sequence of Gaussian vectors distributed as $N^n\sim N(0,\Sigma_n)$. Let $\beta>0$ be a real; write $t^{\beta/2}_{g,n}$, and $t^{\beta/2}_{\mathrm{b},n}(X^n)$ as quantities satisfying*

$$P\left(\left|g_n(Z^n)-\mathbb{E}(g_n(Z^n)|X^n)\right|\geq t^{\beta/2}_{\mathrm{b},n}(X^n)\mid X^n\right)\leq\beta/2;$$

$$P\left(\max_k\left|N^n_k\right|\geq(t^{\beta/2}_{g,n})^{\frac{1}{\alpha}}{C_n}^{-\frac{1}{\alpha}}\right)\leq\beta/2.$$

Then the following holds:

$$\limsup_{\delta\downarrow 0}\limsup_n P\left(\mathbb{E}(g_n(Y^n))\notin\left[g_n(Z^n)-t^{\beta/2}_{\mathrm{b},n}-t^{\beta/2}_{g,n}-\delta,\ \ g_n(Z^n)+t^{\beta/2}_{\mathrm{b},n}+t^{\beta/2}_{g,n}+\delta\right]\right)\leq\beta.$$

See Appendix N.6 for proof of Theorem 2. We present in Appendix H an illustrative example.

4 P-value of a Two-Sample Kernel Test

In this subsection, we show how the bootstrap method can be used to obtain consistent p-values for kernel two sample tests. Given two independent i.i.d processes $(X^n_{i,1})$ and $(X^n_{i,2})$ taking value in $\mathcal{X}_n\subset\mathbb{R}^{d_n}$, the goal of two-sample tests is to determine if the two set of observations $(X^n_{i,1})$ and $(X^n_{i,2})$ are sampled from the same distribution. For ease of notations, we designate by $\mu_{n,1}$ and $\mu_{n,2}$ respectively the distribution of the first sample $(X^n_{i,1})\overset{i.i.d}{\sim}\mu_{n,1}$ and of the second sample $(X^n_{i,2})\overset{i.i.d}{\sim}\mu_{n,2}$; and we want test if the null hypothesis holds

$$(H^n_0):\mu_{n,1}=\mu_{n,2}$$

against the alternative

$$(H^n_1):\mu_{n,1}\neq\mu_{n,2}.$$

A popular method to do so are non-parametric kernel two samples tests [42, 31, 52, 51].

Let $\mathcal{F}_n$ be a class of functions from $\mathcal{X}_n$ into $\mathbb{R}$. If the two distributions are the same $\mu_{n,1}=\mu_{n,2}$ then we have:

$$\sup_{f\in\mathcal{F}_n}\left|\mathbb{E}(f(X^n_{1,1}))-\mathbb{E}(f(X^n_{1,2}))\right|=0.$$

Moreover if $\mathcal{F}_n$ is dense in the space of bounded continuous functions then the converse also holds. The main difficulty therefore consists of choosing the set $\mathcal{F}_n$ to be big enough to differentiate between the distributions $\mu_{n,1}$ and $\mu_{n,2}$ but structured enough that we can estimate of $\sup_{f\in\mathcal{F}_n}\left|\mathbb{E}(f(X^n_{1,1}))-\mathbb{E}(f(X^n_{1,2}))\right|$. To do so, we choose a reproducing kernel space $\mathcal{H}_n$ with kernel $K_n:\mathcal{X}_n\times\mathcal{X}_n\to\mathbb{R}$ and set the class of functions $\mathcal{F}_n$ to be the unit ball of $\mathcal{H}_n$. Different choices of kernels will lead to various level of power for our test especially for structured or high dimensional data. The goal is to choose the kernel that is the most likely to maximize the power of the test.

Let $\left(K_{\theta_k}(\cdot,\cdot)\right)_{k\leq p_n}$ be a finite set of potential Kernel candidates. We write for all $i,j\leq n$ and for all $k\leq p_n$

$$H^{\theta_k}_{i,j}:=K_{\theta_k}(X^n_{j,1},X^n_{i,1})+K_{\theta_k}(X^n_{j,2},X^n_{i,2})-K_{\theta_k}(X^n_{j,1},X^n_{i,2})-K_{\theta_k}(X^n_{j,2},X^n_{i,1});$$

and for all subsets $B\subset[\![n]\!]$ we denote $\hat{M}_{\theta_k}(X^n_B):=\frac{1}{|B|^2}\sum_{i,j\in B}H^{\theta_k}_{i,j}$. The idea proposed in [42] is to select the kernel that gives rise to a test with the highest (estimated) power. This is done by selecting

a subset $B_n \subset [\![n]\!]$ and maximizing the following quantity $\hat{\theta}_n^{B_n} := \text{argmax}_{\theta \in \{\theta_1, \ldots, \theta_{p_n}\}} p_\theta(X^n_{B_n})$ where we have set

$$p_\theta(X^n_{B_n}) := \frac{\hat{M}_\theta(X^n_{B_n})}{\frac{4}{|B_n|^3}\sum_{i \in B_n}\left[\sum_{j \in B_n} H^\theta_{i,j}\right]^2 - \frac{4}{|B_n|^4}\left[\sum_{i,j \leq B_n} H^\theta_{i,j}\right]^2 + \lambda_n}$$

where (λ_n) are tuning parameters. Once the kernel is chosen the test statistic is computed on $[\![n]\!] \setminus B_n$ the remaining data: $\frac{1}{n^2}\sum_{i,j \in [\![n]\!] \setminus B_n} H^{\hat{\theta}_n}_{i,j}$. The fact that the kernel is chosen on a different sample than the test statistic is computed on, means that the conditional limiting distribution of the test statistic, under H_0, is known to be a chi-square [42]. Hence one can compute a consistent estimate of the p-value. However under this approach only a portion of the data is used to select the kernel. This could be problematic when dealing with high-dimensional kernels.

We propose a different method that does not require data splitting and uses the bootstrap method to estimate the p-value. The test statistic that we propose is a softmax:

$$\hat{T}_n(X^n) := \sum_{k \leq p_n} \frac{1}{n^2} \sum_{i,j \leq n} H^{\theta_k}_{i,j}\, \omega_k(X_{1:n}), \qquad \text{where } \omega_k(X_{1:n}) := \frac{e^{\beta_n p_{\theta_k}(X^n)}}{\sum_{k' \leq p_n} e^{\beta_n p_{\theta'_k}(X^n)}};$$

and where (β_n) are hyper-parameters. The bigger β_n is the more weight we give to the kernel maximizing $p_\theta(X^n)$.

We note that the distribution of $\hat{T}_n$ is unknown and depends in an intricate fashion on the set of kernels $\{K_{\theta_k},\ k \leq p_n\}$ as well as on p_n. Therefore to be able to compute the p-value we want to estimate its distribution under H_0.

For technical reasons it is convenient to apply the statistic to a random vector whose coordinates are identically distributed. We achieve this by randomly permuting the two samples before passing it to the kernel test statistic (a transformation also conducted in the prior work of [42]). We remark that under the null hypothesis the distribution of $X^n_{i,1}$ and $X^n_{i,2}$ are the same which implies that the samples are interchangeable $\left(X^n_{i,1}, X^n_{i,2}\right) \overset{d}{=} \left(X^n_{i,2}, X^n_{i,1}\right)$. It is therefore natural to compare the distribution of (X^n_i) to the corresponding randomly permuted process. This is the idea behind permutation tests [42]. In general, for an i.i.d random process $(\tilde{X}_i)$ taking value in $\mathbb{R}^2$ we define the process $(\tilde{X}^M_i)$ obtained by randomly permuting the observations $\tilde{X}_{i,1}$ and $\tilde{X}_{i,2}$:

$$\tilde{X}^M_i := \begin{cases} \tilde{X}_i \text{ with probability } 0.5 \\ (\tilde{X}_{i,2}, \tilde{X}_{i,1})^T \text{ with probability } 0.5. \end{cases}$$

We note that this permuted process has identically distributed coordinates, i.e. $\tilde{X}^M_{1,1} \overset{d}{=} \tilde{X}^M_{1,2}$. As a side fact, we note that $d_W\left(\tilde{X}_1, \tilde{X}^M_1\right) \leq d_W\left(\tilde{X}_{1,1}, \tilde{X}_{1,2}\right)$, where d_W is the Wassertein distance, but, more importantly, if the distribution of $(\tilde{X}_i)$ are already in H_0 then its distribution is left invariant by those permutations. We show that the bootstrap gives consistent and asymptotically tight upper-bounds to the p-value even when p_n grows exponentially fast (proof in Appendix Q.1).

Proposition 2 *Let $(X^n_i) := \left((X^n_{i,1}, X^n_{i,2})\right)$ be a triangular array of i.i.d processes. Let $\{K_{\theta_k},\ k \leq p_n\}$ be a sequence of positive definite continuous kernels. We suppose that*

$$\max_{k \leq p_n} \text{tr}(\text{K}_{\theta_\text{k}}) < \infty; \quad \frac{\beta_\text{n} \log(\text{p}_\text{n}) \text{D}^4_\text{n}}{\lambda^2_\text{n}} = \text{o}(\text{n}^{1/6});$$

where we shorthanded $D_n := \max\left(\left\|\sup_{k \leq p_n} K_{\theta_k}(X^M_{1,1}, X^M_{1,1})\right\|_{L_{120}},\ 1\right)$. Let (Y^n_i) be an independent copy of (X^n_i) and (Z^n_i) be bootstrap samples of (X^n_i). We have:

$$\left\| d_{\mathcal{F}}\left(n\hat{T}_n(Z^M_{1:n}), n\hat{T}_n(Y^M_{1:n}) \mid X^n\right)\right\|_{L_1} \to 0.$$

5 Empirical Risk of Smooth Stacked Ensemble Estimator

A ubiquitous and popular approach for model selection and ensembling in machine learning practice is known as stacking [53, 8, 50]. Given a set of trained *base estimators* $\{\hat{\theta}^1, \ldots, \hat{\theta}^{p_n}\}$, for example

representing a fitted neural network, a random forest and a nearest-neighbour estimator, we call the *smooth-stacked estimator* the linear ensemble of those estimators $\{\hat{\theta}^k\}$ weighted by coefficients that are related to the out-of-sample risk of each estimator. An important question: if we use all the samples to estimate the weights of the ensemble, then can we construct confidence intervals on the risk of the ensemble estimator?

The most straightforward version of stacking is to put all the weight on the model with the smallest out-of-sample risk. Other approaches proposed in practice are to fit a linear regression model using the outputs of each model as an input co-variate to the linear model and using the learned coefficients as coefficients on the ensemble [50].

In this subsection, we analyze a smooth version of stacking, proposed and analyzed experimentally, for instance, in [38], that adds stability to the chosen ensemble, while putting most weight on the best performing model. This ensemble can be viewed as a regularized instance of the linear regression stacking approach where an entropic regularizer is added to the square loss objective. This regularization adds smoothness and stability to the chosen ensemble and allows us to show that the distribution of the ensemble's risk can be estimated with the bootstrap, even if the all the data are used to estimate the weights or fit the base models.

Let (X_i^n) be a triangular array of i.i.d observations taking value in $\mathbb{R}^{d_n}$; and let (m_n) be an increasing sequence. Define $\mathcal{F}_n$ as the space of measurable functions from $\times_{n=1}^{\infty}\mathbb{R}^{d_n}$ to $\mathbb{R}^{d'_n}$. We estimate p_n different estimators $\Omega_n := \left\{\hat{\theta}_n^k(X_{1:m_n}^n),\ k \leq p_n\right\}$ built on the first m_n data-points. Each estimator $\hat{\theta}_n^k$ is a training algorithm that takes as input m_n samples $X_{1:m_n}^n$ and returns a model, which itself is a function from $\mathbb{R}^{d_n}$ to $\mathbb{R}^{d'_n}$. We denote with $\hat{\theta}_n^k(X_{1:m_n}^n)$ the returned model and with $\hat{\theta}_n^k(X_{1:m_n}^n)(x_u) \in \mathbb{R}^{d'_n}$ the evaluation of the model at a point $x_u \in \mathbb{R}^{d_n}$.

The loss of a model at a sample is measured by a common loss function $\mathcal{L}_n : \mathbb{R}^{d_n} \times \mathbb{R}^{d'_n} \to \mathbb{R}$ and the empirical risk of the k-th estimator is computed on all the remaining $n - m_n$ data points as:

$$\mathcal{R}_n^k(x_{1:n}) := \frac{1}{n-m_n}\sum_{u=m_n+1}^{n}\mathcal{L}_n(x_u, \hat{\theta}_n^k(x_{1:m_n})(x_u)).$$

The smooth-stacked estimator is defined as the following ensemble learner

$$\hat{\Theta}_n(x_{1:n})(\cdot) = \sum_{k\leq p_n}\frac{e^{-\beta_n\mathcal{R}_n^k(x_{1:n})}}{\sum_{k'\leq p_n}e^{-\beta_n\mathcal{R}_n^{k'}(x_{1:n})}}\hat{\theta}_n^k(x_{1:m_n})(\cdot).$$

The hyperparameter β_n controls how concentrated the stacked estimator is around the estimator(s) $\hat{\theta}_n^k$ with the lowest empirical error. We denote the empirical risk of an ensemble model $\Theta \in (\mathbb{R}^{d_n} \to \mathbb{R}^{d'_n})$ as

$$\mathcal{R}_\Theta^{\mathrm{s}}(x_{1:n}) := \frac{1}{\sqrt{n-m_n}}\sum_{i=m_n+1}^{n}\mathcal{L}_n\big(x_i, \Theta(x_i)\big).$$

Let (Z^n) be a bootstrap sample of $(X_i^n)_{i\geq m_n}$. We show that the bootstrap method is systematically consistent if and only if $\beta_n = o(\sqrt{n-m_n})$. For simplicity we suppose that the estimators $\hat{\theta}_n^k$ return models that when evaluated at any point $x_u \in \mathbb{R}^{d_n}$ have bounded coordinates; and that the loss function $\mathcal{L}_n$ is smooth, and have bounded partial derivatives in its second argument.

We write the set of all convex combinations of the estimators: $\Omega\left(\{\theta_p,\, p\leq p_n\}\right) := \left\{\sum_{p\leq p_n}\omega_p\theta_p \mid \omega_p \geq 0 \text{ and } \sum_{p\leq p_n}\omega_p = 1\right\}$ and introduce the following notations:

$$T_n := \sup_{\ell\leq d'_n}\left\|\sup_{p\leq p_n}\left|\hat{\theta}_n^p(X_{1:m_n}^n)(X_n^n)_\ell\right|\right\|_{L_\infty} \vee 1,$$

$$L_n := \left\|\sup_{p\leq p_n}\left|\mathcal{L}_n\left(X_n^n, \hat{\theta}_n^p(X_{1:m_n}^n)(X_n^n)\right)\right|\right\|_{L_\infty} \vee \sup_{\ell\leq d'_n}\left\|\sup_{\theta\in\Omega\left(\left\{\hat{\theta}_p(X_{1:m_n}^n),\, p\leq p_n\right\}\right)}\partial_{2,\ell}\left|\mathcal{L}_n(X_n^n, \theta(X_n^n))\right|\right\|_{L_\infty} \vee 1,$$

where by $\partial_{2,\ell}\mathcal{L}_n(x,y)$ we designate $\partial_{y_l}\mathcal{L}_n(x,y)$ We show that if the following hypothesis (H_1^{stacked}) holds then the bootstrap method is asymptotically consistent (proof in Appendix R).

$$\frac{\beta_n d'_n}{\sqrt{n-m_n}} L_n T_n \, e^{\frac{\beta_n}{n-m_n} L_n} \longrightarrow 0. \tag{H_1^{stacked}}$$

Proposition 3 *Choose (m_n), (β_n) and (p_n) be increasing sequences. Let (X_i^n) be a triangular array of i.i.d observations taking value in $\mathbb{R}^{d_n}$. Set $(\mathcal{L}_n : \mathbb{R}^{d_n} \times \mathbb{R}^{d'_n} \to \mathbb{R})$ to be a sequence of smooth loss functions. Let (Z_i^n) and (Y_i^n) be respectively a bootstrap sample and an independent copy of $(X_{m_n+1}^n, \dots, X_n^n)$. Suppose that the hypothesis (H_1^{stacked}) holds then we have:*

$$\left\| d_{\mathcal{F}}\Big(\mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Z^n_{m_n+1:n}) - \mathbb{E}\big[\mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Z^n_{m_n+1:n})\big|\hat{\Theta}_n\big],\ \mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Y^n_{m_n+1:n}) - \mathbb{E}\big[\mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Y^n_{m_n+1:n})\big|\hat{\Theta}_n\big] \mid X^n \Big) \right\|_{L_1} \to 0;$$

where we have shorthanded $\hat{\Theta}_n := \hat{\Theta}_n(X^n)$. Therefore if we choose $t_{n,\alpha}(X^n)$ to be such that:

$$P\left(\left| \mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Z^n_{m_n+1:n}) - \mathbb{E}\big[\mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Z^n_{m_n+1:n})\big|\hat{\Theta}_n\big] \right| \geq t_{n,\alpha}(X^n) \mid X^n \right) \leq \alpha$$

then the following holds

$$\limsup_{n\to\infty} P\left(\left| \mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Y^n_{m_n+1:n}) - \mathbb{E}\big[\mathcal{R}^{\text{s}}_{\hat{\Theta}_n}(Y^n_{m_n+1:n})\big|\hat{\Theta}_n\big] \right| \geq t_{n,\alpha}(X^n) \mid \hat{\Theta}_n \right) \leq \alpha$$

If β_n grows proportionally to $\beta_n \propto \sqrt{n-m_n}$ then the bootstrap method is not a systematically consistent estimator of the risk of the smooth stacked estimator. We present a simple example illustrating this in Appendix K and establish the asymptotic distribution of the bootstrap. However, we show that using Theorem 2 we can still propose a corrected confidence interval with guaranteed asymptotic coverage.

References

[1] Miguel A Arcones and Evarist Giné. U-processes indexed by vapnik-červonenkis classes of functions with applications to asymptotics and bootstrap of u-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52(1):17–38, 1994.

[2] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

[3] Rudolf Beran and Gilles R Ducharme. *Asympotic theory for bootstrap methods in statistics*. 1991.

[4] Rudolf Beran, Muni S Srivastava, et al. Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13(1):95–115, 1985.

[5] Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6):1196–1217, 11 1981.

[6] Peter J Bickel and David A Freedman. Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, pages 28–48, 1983.

[7] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[8] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[9] Matias D Cattaneo, Michael Jansson, and Kenichi Nagasawa. Bootstrap-based inference for cube root asymptotics. *Econometrica*, 88(5):2203–2219, 2020.

[10] Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[11] Snigdhansu Chatterjee, Arup Bose, et al. Generalized bootstrap for estimating equations. *The Annals of Statistics*, 33(1):414–436, 2005.

[12] Sourav Chatterjee. *Superconcentration and related topics*, volume 15. Springer, 2014.

[13] Sourav Chatterjee et al. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

[14] Sourav Chatterjee et al. A new method of normal approximation. *The Annals of Probability*, 36(4):1584–1610, 2008.

[15] Xi Chen, Wen-Xin Zhou, et al. Robust inference via multiplier bootstrap. *Annals of Statistics*, 48(3):1665–1691, 2020.

[16] Xiaohui Chen et al. Gaussian and bootstrap approximations for high-dimensional u-statistics and their applications. *The Annals of Statistics*, 46(2):642–678, 2018.

[17] Yen-Chi Chen, Y. Samuel Wang, and Elena A. Erosheva. On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. *The Annals of Applied Statistics*, 12(2):846 – 876, 2018.

[18] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[19] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

[20] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:1912.10529*, 2019.

[21] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

[22] Hang Deng. Slightly conservative bootstrap for maxima of sums. *arXiv preprint arXiv:2007.15877*, 2020.

[23] Hang Deng and Cun-Hui Zhang. Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *arXiv preprint arXiv:1705.09528*, 2017.

[24] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

[25] Morris L Eaton and David E Tyler. On wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *The Annals of Statistics*, pages 260–271, 1991.

[26] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[27] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[28] Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235, 2018.

[29] Noureddine El Karoui and Elizabeth Purdom. The non-parametric bootstrap and spectral analysis in moderate and high-dimension. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2115–2124, 2019.

[30] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513–520, 2006.

[31] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[32] Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[33] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[34] Fang Han, Sheng Xu, Wen-Xin Zhou, et al. On gaussian comparison inequality and its application to spectral analysis of large random matrices. *Bernoulli*, 24(3):1787–1833, 2018.

[35] Pawel Hitczenko. Best constants in martingale version of rosenthal's inequality. *The Annals of Probability*, pages 1656–1668, 1990.

[36] Roger W Johnson. An introduction to the bootstrap. *Teaching statistics*, 23(2):49–54, 2001.

[37] Iain M Johnstone and Debashis Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

[38] Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018.

[39] Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9):1341–1356, 2011.

[40] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[41] Yuta Koike. Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Japanese Journal of Statistics and Data Science*, pages 1–41, 2020.

[42] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Dougal J Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020.

[43] Miles E Lopes. Central limit theorem and bootstrap approximation in high dimensions with near square root of n rates. *arXiv preprint arXiv:2009.06004*, 2020.

[44] Miles E Lopes, N Benjamin Erichson, and Michael W Mahoney. Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching. *arXiv preprint arXiv:1909.06120*, 2019.

[45] Enno Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Annals of Statistics*, 17(1):382–400, 1989.

[46] Enno Mammen. *When does bootstrap work?: asymptotic results and simulations*, volume 77. Springer Science & Business Media, 2012.

[47] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

[48] Kesar Singh. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, pages 1187–1195, 1981.

[49] Vladimir Spokoiny, Mayya Zhilova, et al. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015.

[50] J. van der Laan Mark, Polley Eric C, and Hubbard Alan E. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):1–23, 2007.

[51] Li Wenliang, Dougal Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746, 2019.

[52] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.

[53] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241 – 259, 1992.

[54] Chien-Fu Jeff Wu et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

[55] Dixin Zhang. Bayesian bootstraps for u-processes, hypothesis tests and convergence of dirichlet u-processes. *Statistica Sinica*, pages 463–478, 2001.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]