On the Asymptotic Distribution of the Minimum Empirical Risk

Jacob Westerhout [1] **TrungTin Nguyen** [1] **Xin Guo** [1] **Hien Duy Nguyen** [2 3]

Abstract

Empirical risk minimization (ERM) is a foundational framework for the estimation of solutions to statistical and machine learning problems. Characterizing the distributional properties of the minimum empirical risk (MER) provides valuable tools for conducting inference and assessing the goodness of model fit. We provide a comprehensive account of the asymptotic distribution for the order-$\sqrt{n}$ blowup of the MER under generic and abstract assumptions, and present practical conditions under which our theorems hold. Our results improve upon and relax the assumptions made in previous works. Specifically, we provide asymptotic distributions for MERs for non-independent and identically distributed data, and when the loss functions may be discontinuous or indexed by non-Euclidean spaces. We further present results that enable the application of these asymptotics for statistical inference. Specifically, the construction of consistent confidence sets using the bootstrap and consistent hypothesis tests using penalized model selection. We illustrate the utility of our approach by applying our results to neural network problems.

1. Introduction

Empirical risk minimization (ERM) is among the most foundational paradigms of machine learning (ML). ERM considers approximating

$$\inf_{x\in\mathcal{X}} \mathbb{E}(l(x,Z)),$$

by instead computing

$$\inf_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n} l(x,Z_i),$$

where the data Z_i has the same distribution as Z, and the loss function $l(\cdot, Z)$ is indexed by parameter $x \in \mathcal{X}$.

ERM appears as a fundamental topic in texts such as Vapnik (1998), Vidyasagar (2003), and Shalev-Shwartz & Ben-David (2014). A great variety of ML methods, from linear and logistic regression to maximum likelihood estimation, support vector machines, and even deep neural networks, can be characterized as ERM problems. The study of ERM problems is also fundamental in statistics and econometric theory, taking on guises such as extremum estimation (Amemiya, 1985; Gourieroux & Monfort, 1995), M-estimation (Serfling, 1980; van der Vaart & Wellner, 2023), and minimum contrasts estimation (Dacunha-Castelle & Duflo, 1986; Bickel & Doksum, 2015), among other names.

The primary objects of interest when studying ERM problems are the empirical risk minimizer (ERM) and the minimum of the empirical risk (MER). Typical problems are the convergence of the ERM and MER to their target values, often referred to as consistency (e.g., Vapnik 1998, Ch. 3 and van der Vaart & Wellner 2023, Sec. 3.3); the finite-sample concentration of mass and in expectation around their targets, often studied as oracle inequalities (e.g., Koltchinskii 2011, Ch. 4 and Cucker & Zhou 2007, Ch. 3); and the asymptotic convergence of blowup sequences of ERMs and their functions to limiting distributions (e.g., van der Vaart & Wellner 2023, Sec. 3.3).

Of relatively less interest has been the study of the limiting distribution of the MER, which dates back to the original work of Wilks (1938), who provided conditions under which the order-n blowup (or simply, n-blowup; defined in Section 3) of the maximum likelihood (ML) converges to a χ^2 random variable. This forms the basis for likelihood ratio tests, for comparing the difference between the log-likelihoods of two competing models. Generalizations of such results are frequently sought and studied to provide hypothesis tests and uncertainty quantification in general statistical settings. Examples of developments in this vein include the works of Vuong (1989), who demonstrates the distributional convergence of the n-blowup to a weighted

[1]School of Mathematics and Physics, The University of Queensland, St Lucia, QLD 4072, Australia [2]School of Computing, Engineering, and Mathematical Sciences, La Trobe University, Bundoora, VIC 3086, Australia [3]Institute of Mathematics for Industry, Kyushu University, Nishi Ward, Fukuoka 819-0395, Japan. Correspondence to: Hien Nguyen <hien@imi.kyushu-u.ac.jp>.

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

sum of χ^2 variables. Asymptotic distributions of n-blowups of non-likelihood MERs were further considered in Shapiro (1989) and Gourieroux & Monfort (1995, Ch. 18).

A shortcoming of the n-blowup asymptotics for MERs is the requirement for strong regularity conditions. Typically, such results require, for example, that the hypothesis space is indexed by a Euclidean space, that the empirical risk be differentiable, and that the limiting or expected risk function be uniquely minimized with a non-singular Hessian at the minimizer (see, e.g., Vuong, 1989 and Shapiro et al., 2021, Sec. 5.1.3). Such assumptions are undesirable in the modern setting where risks may be non-differentiable, have multiple possibly connected minima, and whose parameters may be defined on functional spaces. It is notable that n-blowup asymptotic distributions can be obtained in some special cases in such settings, but the analyses are typically bespoke and laborious, as per the works of Fan et al. (2001), Azaïs et al. (2009), and Dalalyan & Collier (2012).

Although $\sqrt{n}$-blowup asymptotics had been of some interest, the study of such limiting distributions had previously been conducted under the same restrictive assumptions as above (cf. Vuong, 1989). In Shapiro (1991) and following works, including Shapiro (2000) and Shapiro et al. (2021, Ch. 5), a general method for deriving the $\sqrt{n}$-blowup asymptotic distribution of MERs is developed for risks of hypotheses indexed by Euclidean parameters. Although, originally intended for providing guarantees for sample average approximation (SAA) methods in the stochastic programming setting, the results are widely applicable and make minimal assumptions on the risks, requiring only that the empirical risks are computed using independent and identically distributed (IID) data, that risks are Lipschitz continuous on a compact parameter space and the existence of certain moments. Compared to the n-blowup theory, these assumptions are milder and easier to verify.

The proof technique used in Shapiro (1991) and subsequent works relies primarily on a central limit theorem for continuous functions that guarantees the limit of the $\sqrt{n}$-blowup of the empirical risk converges to a bounded continuous Gaussian process (e.g., Dudley, 1999, Thm. 6.3.3). Then, the application of a Danskin-type theorem is used to obtain the directional derivative of the infimum function on the class of continuous functions on a compact set (e.g., Bonnans, 2019, Prop. 5.42), together with an appropriate delta method.

Via recent developments in Danskin-type theorems for infima functions on classes of bounded (and not necessarily continuous) functions by Römisch (2014), Carcamo et al. (2020), and Firpo et al. (2023), we can now provide broad generalizations of the available theory that allow us to obtain $\sqrt{n}$-blowup asymptotic distributions of the MER in situations including when the empirical risks are computed from dependent data, when the risks are discontinuous, and when hypothesis classes are indexed by functional spaces. These generalizations provide new tools for hypothesis testing and uncertainty quantification for broad classes of ML and statistical problems along with novel techniques for model selection.

Aside from our contributions, we note that the results of Shapiro (1991) have progressed in other directions. For instance, Royset & Szechtman (2013) has considered the asymptotic distribution and convergence rates of the MER when computed using an iterative algorithm whose number of iterations increases with sample size. Other results in this direction are summarized by Kim et al. (2015). In recent works, Banholzer et al. (2022) has explored the rates of almost sure and in mean convergence under various assumptions to complement the results of Shapiro (1991). Similar almost sure results, along with moderate deviation principles, are also obtained by Gao & Yiu (2023).

To summarize, our contributions are as follows:

1. We combine the recent developments in Danskin-type theorems of Carcamo et al. (2020) to derive asymptotic distributions of $\sqrt{n}$-blowups of MERs and related quantities under a uniform central limit theorem (cf. van der Vaart & Wellner, 2023) and boundedness assumptions.

2. We demonstrate the use of the modified bootstraps of Fang & Santos (2019) and Hong & Li (2020) to consistently sample from the limiting distributions of the MERs and illustrate how to use such bootstrap samples for conducting hypothesis testing and uncertainty quantification.

3. We propose novel model selection and hypothesis testing routines for drawing inference in general ML and statistical settings, and elaborate on the implementation of these methods in mixture of experts models, and neural network problems.

2. Formal problem setup

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ denote our underlying probability space and $\mathbb{E}$ the expectation operator on this space. Let Z_i be data taking values in a metric space $\mathcal{Z}$; i.e., $Z_i : \Omega \to \mathcal{Z}$ is measurable with respect to (w.r.t.) the Borel σ-algebra on $\mathcal{Z}$. We assume that $(Z_i)_{i\in\mathbb{N}}$ is identically distributed and let Z denote any random variable with the same distribution as each Z_i. We can allow each Z_i to be defined on its own probability space but this offers no increased generality by van der Vaart & Wellner (2023, Ch. 1.3; Ex. 4).

Let $\mathcal{X}$ denote a (non-empty) parameter or hypothesis space (not necessarily Euclidean) and let $l : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ denote a loss function. Typically $\mathcal{X}$ indexes some function class that

is being fit to the data. We denote the empirical or sample risk associated to a finite sample of size n $(Z_i)_{i\in[n]}$ by

$$\hat{f}_n(x,\omega) = \frac{1}{n}\sum_{i=1}^{n} l(x, Z_i(\omega)).$$

We denote its expectation, which we call the expected risk, by $f(x) = \mathbb{E}[l(x,Z)]$.

The primary object of interest in our study is the $\sqrt{n}$-blowup of sequence of minimum empirical risks (MERs)

$$\hat{\psi}_n(\omega) = \inf_{x\in\mathcal{X}} \hat{f}_n(x,\omega),$$

around the minimum expected risk

$$\psi^* = \inf_{x\in\mathcal{X}} f(x).$$

We denote the sets of expected risk minimizers and ϵ-minimizers, respectively, by

$$\mathcal{S} = \arg\min_{x\in\mathcal{X}} f(x), \text{ and } \mathcal{S}^\epsilon = \{x\in\mathcal{X} : f(x) \le \psi^* + \epsilon\}.$$

We similarly denote the corresponding quantities for the empirical risk:

$$\mathcal{S}_n(\omega) = \arg\min_{x\in\mathcal{X}} \hat{f}_n(x,\omega), \text{ and}$$

$$\mathcal{S}_n^\epsilon(\omega) = \left\{x\in\mathcal{X} : \hat{f}_n(x,\omega) \le \hat{\psi}_n(\omega) + \epsilon\right\}.$$

Note that in general $\mathcal{S}$ and $\mathcal{S}_n$ may be empty. Lastly, let

$$\mathcal{H} = \{z \mapsto l(x,z) : \forall x\in\mathcal{X}\}. \tag{1}$$

3. Technical preliminaries

Notation We denote the normal distribution with mean $\mu\in\mathbb{R}$ and variance $\sigma^2\in\mathbb{R}^+$ (or random variable with such distribution) by $\mathrm{N}(\mu,\sigma^2)$.

For each $p\in[1,\infty)$, we denote the $\mathcal{L}^p(\mathbb{P})$ norm of Z by $\|Z\|_p = \{\mathbb{E}|Z|^p\}^{1/p}$ and say that $Z\in\mathcal{L}^p(\mathbb{P})$ if $\|Z\|_p < \infty$. We say that Z is tight if for every $\epsilon>0$ there exists a compact set $\mathcal{K}\subseteq\mathcal{Z}$ so that $\mathbb{P}_Z(\mathcal{K}) \ge 1-\epsilon$. For any $n\in\mathbb{N}$ we write $[n] = \{1,2,\ldots,n\}$.

Outer expectation Define the outer expectation of a (potentially non-measurable) function $U:\Omega\to\bar{\mathbb{R}}$ as

$$\mathbb{E}^*U = \inf\{\mathbb{E}V : V\ge U, V \text{ is measurable}, \mathbb{E}V \text{ exists}\},$$

where $\bar{\mathbb{R}} = \mathbb{R}\cup\{-\infty,\infty\}$. Further, we define the outer probability of an arbitrary set $\mathcal{A}\subseteq\Omega$ by $\mathbb{P}^*(\mathcal{A}) = \mathbb{E}^*(1_\mathcal{A})$, were $1_\mathcal{A}(\omega) = 1$ if $\omega\in\mathcal{A}$ and $1_\mathcal{A}(\omega) = 0$, otherwise. Similarly, we define $\mathbb{P}_*(\mathcal{A}) = -\mathbb{E}^*(-1_\mathcal{A})$. We say that a sequence of real functions $(U_n)_{n\in\mathbb{N}}$ is $o_{\mathbb{P}*}(1)$ or write $U_n \xrightarrow{\mathbb{P}*} 0$ if for every $\epsilon>0$, $\mathbb{P}^*(|U_n|>\epsilon)\to 0$ (cf. van der Vaart & Wellner, 2023, Sec. 1.2).

Dependence concepts We say $(Z_n)_{n\in\mathbb{N}}$ is stationary if for any indices $i_1,\ldots,i_m$, and integer $n\ge 1$, the random vector $(Z_{i_1},\ldots,Z_{i_m})$ has the same joint distribution as $(Z_{n+i_1},\ldots,Z_{n+i_m})$. Further, write $\mathfrak{Z}_1^n = \sigma(Z_i : i\le n)$, as the σ-algebra generated by random variables $\{Z_i, i\in[n]\}$, and $\mathfrak{Z}_{n+k}^\infty = \sigma(Z_i : i\ge n+k)$. As per Bradley (2005), we define the kth β-mixing coefficient as $\beta(k) = \sup_{n\in\mathbb{N}}\beta\left(\mathfrak{Z}_1^n, \mathfrak{Z}_{n+k}^\infty\right)$, where

$$\beta\left(\mathfrak{Z}_1^n, \mathfrak{Z}_{n+k}^\infty\right) = \sup\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{|\mathbb{P}(\mathcal{A}_i\cap\mathcal{B}_j) - \mathbb{P}(\mathcal{A}_i)\mathbb{P}(\mathcal{B}_j)|}{2}$$

with the supremum taken over all pairs of finite partitions, $(\mathcal{A}_i)_{i\in[I]}$ and $(\mathcal{B}_j)_{j\in[J]}$, of Ω, such that $\mathcal{A}_i\in\mathfrak{Z}_1^n$ and $\mathcal{B}_j\in\mathfrak{Z}_{n+k}^\infty$, for each $i\in[I]$ and $j\in[J]$. The sequence $(Z_n)_{n\in\mathbb{N}}$ is said to be β-mixing if $\lim_{k\to\infty}\beta(k) = 0$.

Convergence concepts The weak convergence of random elements $(F_n)_{n\in\mathbb{N}}$ to F is defined in the Hoffmann-Jørgensen sense (see van der Vaart & Wellner, 2023, Section 1.3) and is denoted by $F_n \rightsquigarrow F$. In particular we allow weak convergence of non-measurable objects. For $\tau_n\to\infty$, we say that $\tau_n F_n$ is the τ_n-blowup of the sequence $(F_n)_{n\in\mathbb{N}}$.

For any set $\mathcal{A}$, $\ell^\infty(\mathcal{A}) = \{h : \mathcal{A}\to\mathbb{R} : \sup_{x\in\mathcal{A}}|h(x)| < \infty\}$ is the space of bounded functions on $\mathcal{A}$. Note that functions $h\in\ell^\infty(\mathcal{A})$ are not equivalent classes and may not be measurable if $\mathcal{A}$ has a measure. Note that this differes from the space $\mathcal{L}^\infty$ of essentially bounded functions.

We say that $\mathcal{H}$ has a finite envelope if there exists $\bar{H}:\mathcal{Z}\to\mathbb{R}$, such that $|h(z)|\le\bar{H}(z)$, for each $z\in\mathcal{Z}$ and $h\in\mathcal{H}$. Let $F_n:\Omega\to\ell^\infty(\mathcal{H})$ be given by

$$(F_n(\omega))(h) = \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) - \mathbb{E}h(Z)\right\}.$$

We say that $\mathcal{H}$ is $\mathbb{P}$-Donsker (or just Donsker when there is no ambiguity) if $(Z_i)_{i\in\mathbb{N}}$ are IID, $\sup_{h\in\mathcal{H}}|h(z) - \mathbb{E}_Z h| < \infty$, for each $z\in\mathcal{Z}$, and $F_n\rightsquigarrow F$ for $F:\Omega\to\ell^\infty(\mathcal{H})$, a tight zero-mean Gaussian process with covariance

$$\mathbb{E}[F(h)F(g)] = \mathbb{E}\{[h(Z) - \mathbb{E}h(Z)][g(Z) - \mathbb{E}g(Z)]\}.$$

Donsker classes can be thought of as sets of functions that admit a uniform central limit and are the primary subject of Dudley (1999) and van der Vaart & Wellner (2023, Ch. 2).

Delta method and derivatives Let $\mathcal{U}$ and $\mathcal{V}$ be normed vector spaces and $g:\mathcal{U}\to\mathcal{V}$. We say that g is Hadamard directionally differentiable at x in direction $\eta\in\mathcal{U}$ if for any sequences $(t_n)_{n\in\mathbb{N}}\subset(0,\infty)$ and $(\eta_n)_{n\in\mathbb{N}}\subset\mathcal{U}$ with $t_n\to 0$ and $\eta_n\to\eta$,

$$g'_x(\eta) = \lim_{n\to\infty}\frac{g(x+t_n\eta_n) - g(x)}{t_n}$$

is well defined. When $g'_x(\eta)$ exists for every $\eta \in \mathcal{X}$, we say that g is Hadamard directionally differentiable at x. When $g'_x(\eta)$ exists for every η in some $\mathcal{A} \subseteq \mathcal{X}$, then we say g is Hadamard directionally differentiable at x tangentially to $\mathcal{A}$. If $\eta \mapsto g'_x(\eta)$ is linear, we further say that g is Hadamard differentiable at x. See Schirotzek (2007) for a comprehensive treatment of differentiation concepts.

The following Danskin-type result is instrumental in our analyses. It is a minor extension of the results of Römisch (2014, Prop. 1) and Carcamo et al. (2020, Thm. 2.1).

Theorem 3.1. *Let $\mathcal{A}$ be an arbitrary set and $\mathcal{B} \subseteq \mathcal{A}$. Define $\iota : \ell^\infty(\mathcal{A}) \to \mathbb{R}$ by $\iota(g) = \inf_{\mathcal{B}} g$. Then ι is Hadamard directionally differentiable at any $g \in \ell^\infty(\mathcal{A})$, and for each direction $\eta \in \ell^\infty(\mathcal{A})$,*

$$\iota'_g(\eta) = \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(g,\epsilon,\mathcal{B})} \eta(x),$$

where $\mathcal{S}(g, \epsilon, \mathcal{B}) = \{x \in \mathcal{B} : g(x) \leq \iota(g) + \epsilon\}$.

Some refinements of Theorem 3.1 under topological assumptions on $\mathcal{X}$ are considered by Carcamo et al. (2020), along with conditions for the (full) Hadamard differentiability of ι. We combine the result above with the following directionally differentiable form of the delta method by Römisch (2014). For a generic set $\mathcal{A}$ and generic $F : \Omega \to \ell^\infty(\mathcal{A})$, we write $\inf_{\mathcal{A}} F : \Omega \to \mathbb{R}$ to mean $\inf_{\mathcal{A}} F(\omega) = \inf_{x \in \mathcal{A}} F(x, \omega)$.

Fact 3.2. *Let $\mathcal{U}, \mathcal{V}$ be normed vector spaces and $\mu \in \mathcal{U}$. Let $g : \mathcal{U} \to \mathcal{V}$ be Hadamard directionally differentiable at μ.*

For each $n \in \mathbb{N}$, let $X_n : \Omega \to \mathcal{U}$ be maps and $X : \Omega \to \mathcal{U}$ be measurable w.r.t. the Borel σ-algebra on $\mathcal{U}$. Assume that $\tau_n \to \infty$ and

$$\tau_n(X_n - \mu) \rightsquigarrow X.$$

Then,

$$\tau_n(g(X_n) - g(\mu)) \rightsquigarrow g'_\mu(X), \text{ and}$$
$$\tau_n(g(X_n) - g(\mu)) - g'_\mu(\tau_n(X_n - \mu)) = o_{\mathbb{P}^*}(1).$$

4. Main results

For comparison, we begin with the $\sqrt{n}$-blowup theorem of Shapiro et al. (2021, Thm 5.7), which we generalize. Let $d \in \mathbb{N}$ and make the following assumptions:

A1 $\mathcal{X} \subset \mathbb{R}^d$ is compact, and there exists an $\bar{x} \in \mathcal{X}$, such that $\mathbb{E}\left[l(\bar{x}, Z)^2\right] < \infty$.

A2 There exists a measurable $L : \Omega \to [0, \infty)$, such that $\mathbb{E}\left[L^2\right] < \infty$, and a.s. in ω, for every $x, x' \in \mathcal{X}$,

$$|l(x, Z(\omega)) - l(x', Z(\omega))| \leq L(\omega) \|x - x'\|.$$

Fact 4.1. *If $(Z_i)_{i \in \mathbb{N}}$ are IID, and A1 and A2 hold, then*

$$\sqrt{n}\left(\hat{\psi}_n - \psi^*\right) \rightsquigarrow \inf_{\mathcal{S}} F, \text{ and}$$

$$\hat{\psi}_n = \inf_{x \in \mathcal{S}} \hat{f}_n(x) + o_{\mathbb{P}}\left(n^{-1/2}\right),$$

where F is a zero-mean Gaussian process indexed by $\mathcal{X}$ with covariance between $F(x)$ and $F(x')$:

$$\mathbb{E}\left\{[l(x, Z) - f(x)]\,[l(x', Z) - f(x')]\right\}, \tag{2}$$

for each $x, x' \in \mathcal{X}$. In particular, if $\mathcal{S} = \{x^*\}$, then

$$\sqrt{n}\left(\hat{\psi}_n - \psi^*\right) \rightsquigarrow \mathrm{N}\left(0, \sigma_*^2\right),$$

for $\sigma_*^2 = \mathbb{E}\left[F(x^*)^2\right]$.

Note the undesirable requirements that the hypothesis be indexed by a compact Euclidean space and data being IID, along with the smoothness assumption A2, are present to to invoke a central limit theorem for continuous functions (cf. Dudley, 1999, Thm. 6.3.3). To relax the requirements of Fact 4.1, we present the following abstract result by combining Theorem 3.1 and Fact 3.2. Make the assumptions:

B1 Assume $f \in \ell^\infty(\mathcal{X})$, and there is a $\mathbb{P}$-a.s. set $\Omega_0 \subseteq \Omega$ such that for n sufficiently large $\hat{f}_n(\cdot, \omega) \in \ell^\infty(\mathcal{X})$, $\forall \omega \in \Omega_0$.

B2 There exists $\tau_n \to \infty$ such that

$$\tau_n(\hat{f}_n - f) \rightsquigarrow F,$$

for Borel measurable $F : \Omega \to \ell^\infty(\mathcal{X})$.

Theorem 4.2. *If B1 and B2 hold, then*

$$\tau_n\left(\hat{\psi}_n - \psi^*\right) \rightsquigarrow \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} F(x), \text{ and} \tag{3}$$

$$\hat{\psi}_n = \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} \left\{\hat{f}_n(x) - f(x) + \psi^*\right\} + o_{\mathbb{P}^*}\left(\tau_n^{-1}\right). \tag{4}$$

Remark 4.3. If additionally $\mathcal{X}$ is compact, f is lower semi-continuous, and the sample paths of F are lower semi-continuous, then the weak limit reduces to $\inf_{\mathcal{S}} F$ See Appendix E for details.

The abstraction of B1 and B2 allows for extensive flexibility in obtaining conclusions (3) and (4). We give sufficient conditions in the sequel.

Because of the possibility of pathological sample paths of F it is very hard for $\tau_n(\hat{\psi}_n - \psi^*)$ converge to a Gaussian limit. The following result gives one sufficient condition.

Corollary 4.4. *Assume B1 and B2 hold. If $\mathcal{S}^\epsilon = \mathcal{S}$ for $\epsilon > 0$ small enough, and the function $x \mapsto l(x, \cdot)$ is constant on $\mathcal{S}$, then*

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow F(x^*)$$

for some $x^* \in \mathcal{S}$. If F is Gaussian then so is $F(x^*)$.

Sufficient conditions for B2 The following result provides our most anticipated use case, facilitated by many sufficient conditions; see e.g. van der Vaart & Wellner (2023, Sec. 2.5).

Corollary 4.5. *If B1 is true, and $\mathcal{H}$ is Donsker, then (3) and (4) hold with $\tau_n = \sqrt{n}$, where F is a zero-mean Gaussian process indexed by $\mathcal{X}$ with covariance (2).*

Donsker classes have many known sufficient conditions. Here we present one condition. For others see (Dudley, 1999; van der Vaart & Wellner, 2023). The idea is that if a class of functions is not very 'complicated' then it will be Donsker. One way to measure complexity is to use bracketing numbers.

When $\mathcal{H}$ is a subset of a vector space with norm $\|\cdot\|$ and $l, u \in \mathcal{H}$, we say that $[l, u] = \{h : l \leq h \leq u\}$ is an ϵ-bracket if $\|u - l\| < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|)$ is the minimum number of ϵ-brackets required to cover $\mathcal{H}$. The following result is found in Dudley (1999, Thm. 7.2.1).

Fact 4.6. *The class $\mathcal{H} \subset \mathcal{L}^2(\mathbb{P}_Z)$ is Donsker if*

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_2)} d\epsilon < \infty. \tag{5}$$

Interestingly, our results enable analysis of classes of functions that admit Donsker properties under limited dependence assumptions. The following result from Dedecker & Louhichi (2002, Thm 5.2) provides conditions when $(Z_i)_{i\in\mathbb{N}}$ are β-mixing. Let $p \in (2, \infty)$ and assume:

C1 The set $\mathcal{X}$ is Polish, $\mathcal{H} \subseteq \mathcal{L}^2(\mathbb{P}_Z)$ is such that

$$\int_0^1 \sqrt{\log N_{[]}\left(\epsilon, \mathcal{H}, \|\cdot\|_p\right)} d\epsilon < \infty,$$

and $\sup_{h\in\mathcal{H}} |h(z) - \mathbb{E}_Z h| < \infty$, for each $z \in \mathcal{Z}$.

C2 $\sum_{k=1}^{\infty} k^{2/(p-2)} \beta(k) < \infty$.

Fact 4.7. *Let $(Z_i)_{i\in[n]}$ be a stationary sequence and suppose that C1 and C2 are satisfied. Then $\sqrt{n}\left(\hat{f}_n - f\right) \rightsquigarrow F$, where F is a tight zero-mean Gaussian process.*

Note that there exists numerous examples where C2 is satisfied, such as when $(Z_i)_{i\in\mathbb{N}}$ is m-dependent in the sense that Z_{n+m+1}, is independent $(Z_i)_{i\in[n]}$ for every $n \in \mathbb{N}$. In such case, $\beta(k) = 0$, for all $k > m$. Other processes, including autoregressive sequences, can also be proved to satisfy C2 (cf. Doukhan, 1995, Sec. 2.4).

Computing the bracketing number of any function class is difficult, however there are many known upper bounds. For example, upper bounds are known for convex function classes (van der Vaart & Wellner, 2023, Thm 2.7.14), monotone function classes (van der Vaart & Wellner, 2023, Thm 2.7.9), or function classes with Holder-derivatives (van der Vaart & Wellner, 2023, Cor 2.7.2, Cor 2.7.3).

There are similar results for parametric classes. When $(\mathcal{X}, d)$ is a metric space and $|l(x, z) - l(x', z)| \leq \bar{H}(z)\, d(x, x')$ is satisfied $\forall x, x' \in \mathcal{X}$, for some fixed $\bar{H} : \mathcal{Z} \to \mathbb{R}$, then for any norm $\|\cdot\|$, we have $N_{[]}\left(2\epsilon \left\|\bar{H}\right\|, \mathcal{H}, \|\cdot\|\right) \leq N(\epsilon, \mathcal{X}, d)$, where $N(\epsilon, \mathcal{X}, d)$ is the minimum number of balls of radius ϵ required to cover $\mathcal{X}$ (cf. Kosorok, 2008, Thm. 9.23).

5. Statistical inference

In order for limits of the form $\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow F^*$ to be of practical utility, we need a method to approximate F^*. When F^* is Gaussian, under reasonable conditions, it is possible obtain the convergence result:

$$\tau_n\left(\hat{\psi}_n - \psi^*\right)/\hat{\sigma}_n \rightsquigarrow \mathrm{N}(0, 1),$$

for sample variance $\hat{\sigma}_n^2$ of $\hat{f}_n(x)$, for some $x \in \mathcal{S}_n$; see Theorems C.3 and C.4. Given some restrictions, Corollaries 4.4 and E.2 can be used to obtain the Gaussianity of F^*. However, in general it is not possible to directly approximate the limiting process using sample means and variances (cf. Kosorok, 2008, p. 19).

Standard bootstrapping procedures only work under very restrictive conditions. In our context, Fang & Santos (2019, Thm. 3.1) states that when $\mathcal{H}$ is Donsker, many bootstrapping procedures (including the non-parametric bootstrap) are consistent precisely when the inf map is (fully) Hadamard differentiable on the support tangentially to the image measure of F (the Donsker limit). Theorem 3.1 only gives directional differentiability and conditions for ι to be Hadamard differentiable are very restrictive. Our current best results for conditions under which the bootstrap is consistent are:

- If F has lower semi-continuous sample paths, $\mathcal{X}$ is compact, and f lower semi-continuous and bounded then we require $x \mapsto l(x, \cdot)$ to be constant on $\mathcal{S}$. See Theorem B.2.
- If F and f are bounded then we require that for ϵ small enough, $\mathcal{S}^\epsilon = \mathcal{S}$, and on $\mathcal{S}$, $x \mapsto l(x, \cdot)$ is constant. See Theorem B.4.

Fang & Santos (2019) have given a framework that slightly modifies the bootstrap and allows for consistent approximation of F^*.

Most bootstrapping procedures can be equivalently considering as drawing weights $(W_i)_{i\in[n]}$ from some distribution,

giving a bootstrapped empirical risk of

$$f_n^{\mathrm{b}}(x) = \left(\sum_{i=1}^n W_i\right)^{-1} \sum_{i=1}^n W_i l(x, Z_i) \tag{6}$$

For $\mathcal{Y}$ a metric space we let $\mathrm{BL}_1(\mathcal{Y}) = \{g : \mathcal{Y} \to [0,1] : \mathrm{Lip}(g) \leq 1\}$, where $\mathrm{Lip}(g)$ refers to the (smallest) Lipschitz constant of g. We write $\mathcal{M}(\Omega, \mathcal{Y})$ to denote the set of Borel measurable functions from Ω to $\mathcal{Y}$. We write $\mathcal{B}^{\mathcal{A}}$ to denote all functions from $\mathcal{A}$ to $\mathcal{B}$.

For any σ-subalgebra $\mathfrak{A} \subseteq \mathfrak{F}$ the outer conditional expectation $\mathbb{E}^*(\cdot|\mathfrak{A}) : \mathbb{R}^\Omega \to \mathcal{L}^1(\Omega)$ is defined via

$$\mathbb{E}^*(f|\mathfrak{A}) = \inf\left\{\mathbb{E}(g|\mathfrak{A}) \,\middle|\, g \geq f, \text{ and } \mathbb{E}(g) \text{ exists}\right\},$$

where the infimum is defined using standard partial ordering of random variables ($f \leq g$ if and only if $f(\omega) \leq g(\omega)$ for almost all ω). Similarly define $\mathbb{E}_*(f|\mathfrak{A}) = -\mathbb{E}^*(-f|\mathfrak{A})$.

We then define $d_{\mathrm{BL}}^n : \mathcal{Y}^\Omega \times \mathcal{M}(\Omega, \mathcal{Y}) \to \overline{\mathbb{R}}^\Omega$ via

$$d_{\mathrm{BL}}^n(U, V) = \sup_{g \in \mathrm{BL}_1(\mathcal{Y})} |\mathbb{E}^*(g(U)|(Z_i)_{i\in[n]}) - \mathbb{E}(g(V))|.$$

where the supremum depends on (Z_i) and is again taken w.r.t. the standard ordering on random variables.

Following the notation of Kosorok (2008) we say that

$$\tau_n(f_n^b - \hat{f}_n)|(Z_i)_{i\in[n]} \overset{\mathbb{P}^*}{\rightsquigarrow} F,$$

if and only if

$$d_{\mathrm{BL}}^n(\tau_n(f_n^b - \hat{f}_n), F) \xrightarrow{\mathbb{P}^*} 0, \text{ and}$$

$$\mathbb{E}^*(g(\tau_n(f_n^b - \hat{f}_n)) - \mathbb{E}_*(g(\tau_n(f_n^b - \hat{f}_n)) \to 0,$$

for all $g \in \mathrm{BL}_1(\mathcal{X})$. Convergence in d_{BL}^n is sufficient to generate asymptotically correct quantiles when F has continuous distribution function (cf. Bücher & Kojadinovic, 2019, Lem. 4.2).

In order to then approximate F^* we use functions $i_n : \ell^\infty(\mathcal{X}) \times \Omega \to \mathbb{R}$ such that

$$i_n(\tau_n(f_n^b - \hat{f}_n), \cdot)|(Z_i)_{i\in[n]} \overset{\mathbb{P}^*}{\rightsquigarrow} F^*.$$

We consider two possible forms of i_n:

1. From Fang & Santos (2019) and Hong & Li (2018),

$$\hat{\iota}_{s_n,n}(\eta, \omega) = s_n^{-1}(\inf_{\mathcal{X}}(\hat{f}_n(\omega) + s_n\eta) - \hat{\psi}_n(\omega)).$$

2. Modified from Firpo et al. (2023),

$$\tilde{\iota}_{t_n,n}(\eta, \omega) = \inf_{x \in \mathcal{S}_n^{t_n}(\omega)}(\eta).$$

Make the following assumptions:

D1 $\tau_n(\hat{f}_n - f) \rightsquigarrow F$ for some tight, Gaussian $F : \Omega \to \ell^\infty(\mathcal{X})$.

D2 $\tau_n(f_n^b - \hat{f}_n)|(Z_i)_{i\in[n]} \overset{\mathbb{P}^*}{\rightsquigarrow} F$ and $\tau_n(f_n^b - \hat{f}_n)$ is asymptotically measurable (c.f. van der Vaart & Wellner, 2023, def 1.3.7).

D3 $\tau_n(f_n^b - \hat{f}_n)$ is a measurable function of the weights $(W_i)_{i\in[n]}$ for fixed $(Z_i)_{i\in[n]}$.

D4 The weights are chosen independent of the data.

Theorem 5.1. *Assume that B1 and D1–D4 are satisfied.*

- *If $s_n \to 0$ and $\tau_n s_n \to \infty$, then*

$$d_{\mathrm{BL}}^n\left(\hat{\iota}_{s_n,n}(\tau_n(f_n^b - \hat{f}_n)), \lim_{\epsilon\downarrow 0}\inf_{S^\epsilon} F\right) \overset{\mathbb{P}^*}{\to} 0.$$

- *If $t_n \to 0$ with $\tau_n t_n \to \infty$, then*

$$d_{\mathrm{BL}}^n\left(\tilde{\iota}_{t_n,n}(\tau_n(f_n^b - \hat{f}_n)), \lim_{\epsilon\downarrow 0}\inf_{S^\epsilon} F\right) \overset{\mathbb{P}^*}{\to} 0.$$

The following result provides our most anticipated use case

Corollary 5.2. *If B1 is satisfied with $\mathcal{H}$ Donsker and $(W_i)_{i\in\mathbb{N}}$ corresponding to the nonparametric bootstrap, then the conclusion of Theorem 5.1 holds with $\tau_n = \sqrt{n}$.*

It is well known that for non-IID data that the standard bootstrap tends to fail, see for example Singh (1981, Rem. 2.1) or Liu & Singh (1992). To handle the non-IID case we consider the moving block boostrap (MBB). Rather than resampling the data with replacement, we draw n/l blocks of l contiguous data points out of the possible $n - l + 1$ blocks of data. Such a procedure preserves the dependency structure much more than the standard bootstrap.

The MBB satisfies D4, satisfies D3 by Lemma A.5 and using Bühlmann (1995, Thm. 1), D1–D2 are satisfied under the assumptions:

E1 The β-mixing coefficients of the data satisfy $\beta(k) \leq \exp(-c_1 k)$, for some $c_1 > 0$.

E2 The block length l satisfies $l(n) = O\left(n^{1/2-\epsilon}\right)$, for some $\epsilon \in (0, 1/2)$.

E3 $\mathcal{H}$ has envelope $\bar{H} \in \mathcal{L}^p(\mathbb{P}_Z)$, for some $p > 4$, whereby for some constants $c_2, c_3 > 0$, $N_{[]}\left(\epsilon, \mathcal{H}, \|\cdot\|_p\right) \leq c_2\epsilon^{-c_3}$.

E4 $\mathcal{X}$ is Souslin in the sense that it is an analytic subset of a compact metric space, with Borel σ-algebra $\mathfrak{B}(\mathcal{X})$ (cf. Dellacherie & Meyer, 1975, Def. 16), and $l(\cdot, Z) : \mathcal{X} \times \Omega \to \mathbb{R}$ is jointly measurable on $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{F}$.

We note that E1–E3 are much stronger than the non-bootstrap counterpart C1 and C2 for Theorem 4.7, especially the higher moment requirement and fast mixing rate. We know of no alternatives that make bracketing assumptions, however the Vapnik–Chervonenkis (VC) result of Radulović (2002, thrm 2.5) provides an alternative under stronger entropy, but weaker mixing rates and moments assumptions.

Model selection We can use the asymptotic limits of Theorem 4.2 to conduct model selection by approximating the quantiles of the limiting distribution. This requires a minor modification of Theorem 4.2.

Let $(\mathcal{X}_k)_{k\in[m]}$ be a sequence of m parameter spaces defining corresponding model spaces $(\mathcal{H}_k)_{k\in[m]}$, where $\mathcal{H}_k$ is defined as per (1) with $\mathcal{X}$ replaced by $\mathcal{X}_k$. Let $\psi_k^* = \inf_{x_k\in\mathcal{X}_k} f(x_k)$ be the minimum expected risk obtained by models in $\mathcal{H}_k$. Similarly we define $\hat{\psi}_{k,n}(\cdot) = \inf_{x_k\in\mathcal{X}_k} \hat{f}_n(x_k,\cdot)$.

Theorem 5.3. *Let $\mathcal{X}_0, \mathcal{X}_1$ be arbitrary sets. Assume B1 and B2 are satisfied with $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ and further assume $\psi_0^* = \psi_1^*$. Then*

$$\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \rightsquigarrow F^*, \text{ where}$$
$$F^* = \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}(f,\epsilon,\mathcal{X}_1)} F(x) - \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}(f,\epsilon,\mathcal{X}_0)} F(x),$$

and $\mathcal{S}(\cdot,\cdot,\cdot)$ is as defined in Theorem 3.1.

The following result is useful for model selection.

Corollary 5.4. *Let $\mathcal{X}_0, \mathcal{X}_1$ be sets. Assume B1, B2 are satisfied with $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$. If $\psi_1^* = \psi_0^*$, we have for any $\alpha \in [0,1]$,*

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\hat{\psi}_{n,1} \leq \hat{\psi}_{n,0} + \frac{c_\alpha}{\tau_n} \right) \leq \alpha, \text{ where}$$
$$c_\alpha = \sup\left\{ c \in \overline{\mathbb{R}} : \mathbb{P}(F^* \leq c) \leq \alpha \right\},$$

and F^* is as defined in Theorem 5.3.

This then provides a method for testing the null hypothesis $\mathrm{H}_0 : \psi_0^* = \psi_1^*$ against the alternative $\mathrm{H}_1 : \psi_1^* > \psi_0^*$ at any size $\alpha \in [0,1]$ by rejecting H_0 if $\hat{\psi}_{n,1} > \hat{\psi}_{n,0} + c_\alpha/\tau_n$.

Such results cannot be used to choose favourably between any 2 models with the same minimum expected risk. However, using our results, we can infer the optimal hypothesis, with the minimum *complexity*, within a set of competing hypotheses. With this goal in mind, we let $k \in [m]$ index the model classes in order of complexity with larger k corresponding to higher complexity. For example, k could be the order of polynomials that form the hypothesis space.

The aim is to estimate the least complex model within the class of models with optimal performance:

$$k^* = \min\left\{ \arg\min_{k\in[m]} \psi_k^* \right\}.$$

Towards this end, we construct a penalized empirical risk-based estimator as per the information criteria of Akaike (1974) and Schwarz (1978). Namely, we estimate k^* by

$$\hat{K}_n = \min\left\{ \arg\min_{k\in[m]} \left(\hat{\psi}_{k,n} + P_{k,n} \right) \right\},$$

where $(P_{k,n})_{k\in[m]}$ is a sequence of penalty functions, possibly depending on $(Z_i)_{i\in[n]}$. Following the usual approach, as espoused in Claeskens & Hjort (2008, Ch. 4) and Baudry (2015), we propose conditions under which $\hat{K}_n$ is a consistent estimator of k^*, in the sense that, as $n \to \infty$, $\mathbb{P}_*\left(\hat{K}_n = k^*\right) \to 1$.

Assume that $\tau_n \to \infty$ and make the following assumptions for each $k \in [m]$:

F1 $\tau_n\left(\hat{\psi}_{k,n} - \psi_k^*\right)$ is asymptotically bounded in probability in the sense that $\forall\delta > 0\ \exists M \in \mathbb{R}$ such that

$$\liminf_{n\to\infty} \mathbb{P}_*\left(\tau_n\left(\hat{\psi}_{k,n} - \psi_k^*\right) < M \right) \geq 1 - \delta$$

F2 $P_{k,n} > 0$, $P_{k,n} = o_{\mathbb{P}*}(1)$, and $\tau_n\{P_{l,n} - P_{k,n}\} \xrightarrow{\mathbb{P}*} \infty$, for every $l > k$.

Here, for any sequence of maps $h_n : \Omega \to \mathbb{R}$ we say that $h_n \xrightarrow{\mathbb{P}*} \infty$ if $\forall M \in \mathbb{R}$, $\mathbb{P}_*(h_n > M) \to 1$.

Proposition 5.5. *If F1 and F2 hold for each $k \in [m]$, then K_n is a consistent estimator for k^*.*

To make our result concrete, we note that by Lemma A.4 F1 is satisfied whenever $\tau_n\left(\hat{\psi}_{k,n} - \psi_k^*\right)$ converge in distribution for each $k \in [m]$. Namely, if the hypotheses of Theorem 4.2 are satisfied for each $k \in [m]$, then F1 holds. One then selects an appropriate sequence $(P_{k,n})_{k\in[m]}$ that satisfies F2 to enable the conclusion of Proposition 5.5.

Conditions F1 and F2 broadly generalises the consistency theory of Sin & White (1996) and Baudry (2015) who consider only models indexed by Euclidean spaces with unique minimizers and strong differentiability properties.

6. Incremental hypothesis spaces

We have previously assumed that the parameter space is independent of the sample size. This is often not true in high dimensional problems and so we now relax this assumption. Let $\mathcal{X}_n$ denote the (non-random) parameter space indexed by the sample size n.

Make the following assumption:

G1 $\emptyset \neq \mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{X}_n$.

We have the following extension of the delta method.

Theorem 6.1. *Let $\mathcal{U}, \mathcal{V}$ be normed vector spaces and $\mu \in \mathcal{U}$. $\forall n \in \mathbb{N}$ let $g_n, h_n : \mathcal{U} \to \mathcal{V}$ and let $(\tau_n) \subseteq \mathbb{R}^+$ be such that $\tau_n \to \infty$. For each $n \in \mathbb{N}$ let $X_n : \Omega \to \mathcal{U}$ be maps and let $X : \Omega \to \mathcal{U}$ be Borel measurable. Assume $\forall \eta \in \mathcal{U}$ and $\forall (\eta_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ with $\eta_n \to \eta$,*

$$D_\mu(\eta) = \lim_{n\to\infty} \tau_n[g_n(\mu + \eta_n/\tau_n) - h_n(\mu)] \tag{7}$$

is well defined. Then if $\tau_n(X_n - \mu) \rightsquigarrow X$, we have

$$\tau_n(g_n(X_n) - h_n(\mu)) \rightsquigarrow D_\mu(X), \text{ and}$$
$$\tau_n(g_n(X_n) - h_n(\mu)) - D_\mu(\tau_n(X_n - \mu)) = o_{\mathbb{P}*}(1).$$

Equation (7) is an extension of the idea of Hadamard differentiability to a sequence of functions. The next two results show when this is true for the infima maps. When $\mathcal{X}$ has a topology we let $\mathrm{lsc}(\mathcal{X})$ to be the space of lower semi-continuous functions $g : \mathcal{X} \to \mathbb{R}$ and equip it with the topology of uniform convergence (c.f. Willard, 2012, Def. 42.8).

Theorem 6.2. *Assume G1 and that $\mathcal{X}$ is a compact topological space. Let $f \in \mathrm{lsc}(\mathcal{X})$ and let $(t_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$ with $t_n \to 0$. Then for any $g_n, g \in \mathrm{lsc}(\mathcal{X})$ with $||g_n - g||_\mathcal{X} \to 0$,*

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_\mathcal{X} f}{t_n} = \inf_\mathcal{S} g, \tag{8}$$
$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}_n} f}{t_n} = \inf_\mathcal{S} g. \tag{9}$$

Theorem 6.3. *Assume G1. Let $f \in \ell^\infty(\mathcal{X})$ and let $(t_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$ with $t_n \to 0$. If $\exists N \in \mathbb{N}$ such that*

$$\inf_{\mathcal{X}_N^c} f > \inf_\mathcal{X} f,$$

then for any $g_n, g \in \ell^\infty(\mathcal{X})$ with $||g_n - g||_\mathcal{X} \to 0$,

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_\mathcal{X} f}{t_n} = \lim_{\epsilon\searrow 0} \inf_{\mathcal{S}^\epsilon} g, \tag{10}$$
$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}_n} f}{t_n} = \lim_{\epsilon\searrow 0} \inf_{\mathcal{S}^\epsilon} g. \tag{11}$$

These results above can be combined upon defining

$$\hat{\phi}_n = \inf_{\mathcal{X}_n} \hat{f}_n, \text{ and } \phi_n^* = \inf_{\mathcal{X}_n} f,$$

and making the following assumptions:

H1 $\mathcal{X}$ is a compact metric space, $f \in \mathrm{lsc}(\mathcal{X}) \cap \ell^\infty(\mathcal{X})$, and there is a $\mathbb{P}$-a.s. set $\Omega_0 \subseteq \Omega$ such that for n sufficiently large $\hat{f}_n(\cdot, \omega) \in \mathrm{lsc}(\mathcal{X}) \cap \ell^\infty(\mathcal{X}), \forall \omega \in \Omega_0$.

H2 There exists $\tau_n \to \infty$ such that

$$\tau_n(\hat{f}_n - f) \rightsquigarrow F,$$

for $F : \Omega \to \mathrm{lsc}(\mathcal{X}) \cap \ell^\infty(\mathcal{X})$ Borel measurable.

Theorem 6.4. *Assume G1, B1 and B2 are satisfied and additionally $\exists N \in \mathbb{N}$ such that $\inf_{\mathcal{X}_N^c} f > \inf_\mathcal{X} f$. Then*

$$\tau_n\left(\hat{\phi}_n - \psi^*\right) \rightsquigarrow \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}^\epsilon} F(x),$$
$$\tau_n\left(\hat{\phi}_n - \phi_n^*\right) \rightsquigarrow \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}^\epsilon} F(x),$$

and

$$\hat{\phi}_n = \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}^\epsilon} \left\{\hat{f}_n(x) - f(x) + \psi^*\right\} + o_{\mathbb{P}*}\left(\tau_n^{-1}\right),$$
$$\hat{\phi}_n = \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}^\epsilon} \left\{\hat{f}_n(x) - f(x) + \phi_n^*\right\} + o_{\mathbb{P}*}\left(\tau_n^{-1}\right).$$

If instead G1, H1 and H2 are true, then

$$\tau_n\left(\hat{\phi}_n - \psi^*\right) \rightsquigarrow \inf_{x\in\mathcal{S}} F(x),$$
$$\tau_n\left(\hat{\phi}_n - \phi_n^*\right) \rightsquigarrow \inf_{x\in\mathcal{S}} F(x),$$

and

$$\hat{\phi}_n = \inf_{x\in\mathcal{S}} \hat{f}_n(x) + o_{\mathbb{P}*}\left(\tau_n^{-1}\right),$$
$$\hat{\phi}_n = \inf_{x\in\mathcal{S}} \left\{\hat{f}_n(x) - \psi^* + \phi_n^*\right\} + o_{\mathbb{P}*}\left(\tau_n^{-1}\right).$$

7. Numerical experiments

7.1. Model selection in Gaussian mixture of experts

We firstly provide empirical evidence towards the guarantees of Proposition 5.5.

Data generating process. We generate an 8-dependent stationary sequence $(Z_i)_{i\in[n+8]}$, $Z_i = (X_i, Y_i)$ for each $i \in [n+8]$, from a Gaussian mixture of experts (GMoE; Jacobs et al., 1991) model, with $k^* = 2$ components. Let $(E_i)_{i\in[n]}$ be IID, where $E_i \sim \mathrm{N}(0,1)$ for each $i \in [n]$, and $X_i \sim \mathrm{Unif}(0,1)$, for each $i \in [8]$. Then, for $i \in [n] \setminus [8]$, $X_i|(E_i)_{i\in[n]} \sim 1_{\{\sum_{j=1}^8 E_{i-j} \le 0\}} \mathrm{Unif}(0, 1/3) + 2 \times 1_{\{\sum_{j=1}^8 E_{i-j} > 0\}} \mathrm{Unif}(1/3, 1)$. Next, we simulate latent labels $L_i|X_i \sim 1 + \mathrm{Ber}(\pi(X_i))$, where $\pi(x) = 1/\{1 + \exp(15x - 7)\}$. Finally, we generate responses $Y_i|(X_i, L_i) \sim \mathrm{N}(\mu_{L_i}(X_i), \sigma^2_{L_i})$, where $\mu_1(x) = -15x + 8$, $\mu_2(x) = 0.4x + 0.6$, $\sigma_1^2 = 0.3^2$ and $\sigma_2^2 = 0.4^2$. See Ho et al. (2022), Nguyen et al. (2022; 2023), and references within, for recent developments regarding the estimation and model selection of GMoE models.

Model selection criteria. For each GMoE with k experts, $k \in [5]$, denote its parameter space by $\mathcal{X}_k$. Following the suggestion of Sin & White (1996), when $\tau_n = \sqrt{n}$, we propose penalties of the form $P_{k,n}^{\mathrm{SWIC}} = [\dim(\mathcal{X}_k)\log(n)]/\sqrt{n}$, defining what we designate the Sin and White information criterion (SWIC), where $\dim(\cdot)$ is the number of parameters for each model. It is easy to verify that $P_{k,n}^{\mathrm{SWIC}}$

satisfies F2. The usual BIC and AIC, with penalties $P_{k,n}^{\mathrm{BIC}} = [\dim(\mathcal{X}_k)\log(n)]/(2n)$ and $P_{k,n}^{\mathrm{AIC}} = \dim(\mathcal{X}_k)/n$, do not satisfy F2. To compute the MERs, we implement the usual Expectation–Maximization algorithm for GMoE models (see, e.g., Chamroukhi et al., 2009). Figure 1 displays the relative performance of the SWIC versus the BIC and AIC, over 50 simulations of size $n = 2000$. We observe that SWIC correctly estimates k^* in all replications, whereas AIC always underestimates the complexity. BIC estimates correctly with high probability (0.72) but often overestimates the complexity.

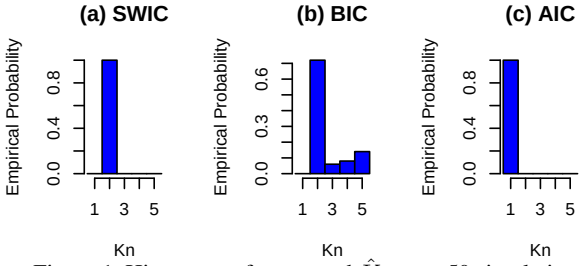

Figure 1. Histogram of computed $\hat{K}_n$ over 50 simulations.

7.2. Neural Network

Here we seek to numerically verify the ability of the bootstrap procedures to generate asymptotically correct quantiles. We do this by using a model for which ψ^* is analytically computable and test if the bootstrap procedures can generate confidence interval (CI) with the correct coverage.

To generate the model we consider a binary classification feedforward neural network (NN) with 1 input node and 1 hidden layer, consisting of 3 nodes with ReLU activation. We first fix a NN and generate n IID replicates $(Z_i)_{i\in\mathbb{N}}$ of $Z = (X, Y)$, where $X \sim \mathrm{Unif}(0, 1)$ and Y is the output of the NN, with input X, flipped 30% of the time.

We fit a NN with the same configuration to the data. Via the data generating process, we know that the minimum classification loss is $\psi^* = 0.3$, by Lemma D.1. These networks were fit by minimising the classification loss using the 'particleswarm' global optimizer in MATLAB. Full details are given in Appendix D.

We seek to compute 90% CIs for the classification loss, using the standard nonparametric bootstrap and the two consistent procedures of Theorem 5.1. Figure 2 shows the coverage of these procedures. Note that for moderately large samples, all methods provide conservative coverage.

Our choice of NN configuration characterizes a class, $\mathcal{F}$, of binary output functions, which is Donsker if $\mathcal{F}$ is a measurable VC subgraph class (cf. Dudley, 1999, Cor. 10.1.5). Since the classification loss of the NN can be evaluated with only a finite number of logical comparison and elementary arithmetic operations, the fact that $\mathcal{F}$ is a VC subgraph class then follows via Anthony & Bartlett (1999, Thm. 8.14). Theorem 5.1 implies that our methods should correctly provide 90% coverage if the limiting distribution of the MER is continuous, while there is minimal support for the standard nonparametric bootstrap in this setting.

Figure 3 shows that the widths of the CIs are of reasonable sizes for moderate amount of data. For this numerical experiment, all bootstrap procedures achieved the nominal coverage. All methods do not take the same amount of computation resources, however (see Figure 4). The method based on Firpo et al. (2023) was considerably faster than the others due to the amortisation property of not requiring refits of the NN for each bootstrap resample.

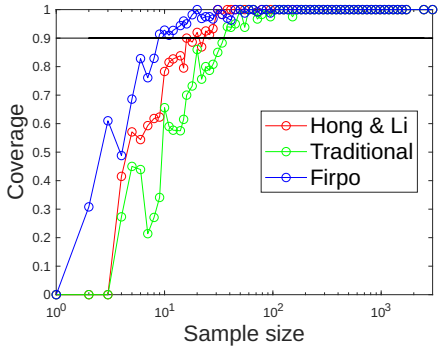

Figure 2. Coverage of nominally 90% bootstrap CIs for various sample sizes.

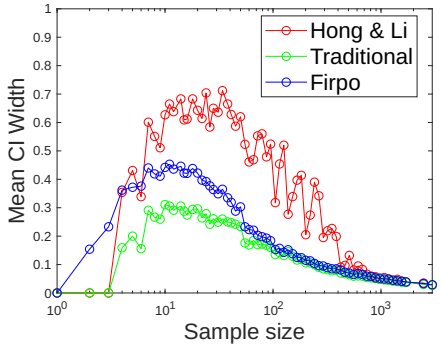

Figure 3. Mean widths of the 90% bootstrap CIs for various sample sizes.

8. Conclusion

We have reported on a comprehensive set of tools for characterizing the asymptotic distribution of MERs along with protocols for model selection and statistical inference, based on these theoretical results. Practical regularity conditions for implementing our methods and example applications are provided to illustrate the utility of our results. Further directions of study will involve better understanding the properties of the limiting distributions of MERs and how these properties interact with various bootstrap methods.

Acknowledgements

We thank the Reviewers and Area Chair whose advice helped to improve our manuscript. All authors acknowledge funding from the ARC grant: DP230100905.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

Amemiya, T. *Advanced econometrics*. Harvard University Press, 1985.

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

Azaïs, J.-M., Gassiat, É., and Mercadier, C. The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM: Probability and Statistics*, 13:301–327, 2009.

Banholzer, D., Fliege, J., and Werner, R. On rates of convergence for sample average approximations in the almost sure sense and in mean. *Mathematical Programming*, pp. 1–39, 2022.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Baudry, J.-P. Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics*, 9:1041–1077, 2015.

Bickel, P. J. and Doksum, K. A. *Mathematical statistics: basic ideas and selected topics*, volume 1. CRC Press, 2015.

Bonnans, J. F. *Convex and stochastic optimization*. Springer, 2019.

Bradley, R. C. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Bücher, A. and Kojadinovic, I. A note on conditional versus joint unconditional weak convergence in bootstrap consistency results. *Journal of Theoretical Probability*, 32 (3):1145–1165, 2019.

Bühlmann, P. The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Processes and their Applications*, 58:247–265, 1995.

Carcamo, J., Cuevas, A., and Rodriguez, L.-A. Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli*, 26:2143–2175, 2020.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22:593–602, 2009. Publisher: Elsevier.

Claeskens, G. and Hjort, N. L. *Model selection and model averaging*. Cambridge University Press, 2008.

Cucker, F. and Zhou, D. X. *Learning theory: an approximation theory viewpoint*. Cambridge University Press, 2007.

Dacunha-Castelle, D. and Duflo, M. *Probability and Statistics: Volume II*. Springer, 1986.

Dalalyan, A. and Collier, O. Wilks' phenomenon and penalized likelihood-ratio test for nonparametric curve registration. In *Artificial Intelligence and Statistics*, pp. 264–272. PMLR, 2012.

Dedecker, J. and Louhichi, S. Maximal inequalities and empirical central limit theorems. In *Empirical Process Techniques for Dependent Data*, pp. 137–159. Springer, 2002.

Dellacherie, C. and Meyer, P.-A. *Probabilities and Potential*. Elsevier, 1975.

Doukhan, P. *Mixing: properties and examples*. Springer, 1995.

Dudley, R. M. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

Fan, J., Zhang, C., and Zhang, J. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of statistics*, 29:153–193, 2001.

Fang, Z. and Santos, A. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86: 377–412, 2019.

Firpo, S., Galvao, A. F., and Parker, T. Uniform inference for value functions. *Journal of Econometrics*, 235:1680–1699, 2023.

Gao, M. and Yiu, K.-F. C. Moderate deviations and invariance principles for sample average approximations. *SIAM Journal on Optimization*, 33:816–841, 2023.

Gourieroux, C. and Monfort, A. *Statistics and Econometric Models*, volume 2. Cambridge University Press, 1995.

Ho, N., Yang, C.-Y., and Jordan, M. I. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23:1–81, 2022.

Hong, H. and Li, J. The numerical delta method. *Journal of Econometrics*, 206:379–394, 2018.

Hong, H. and Li, J. The numerical bootstrap. *The Annals of Statistics*, 48:397–412, 2020.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. Publisher: MIT Press.

Kim, S., Pasupathy, R., and Henderson, S. G. A guide to sample average approximation. *Handbook of simulation optimization*, pp. 207–243, 2015.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Springer, 2011.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Liu, R. Y. and Singh, K. Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, pp. 225–248. Wiley, 1992.

Nguyen, H., Nguyen, T., and Ho, N. Demystifying Softmax Gating Function in Gaussian Mixture of Experts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*, 16:4742 – 4822, 2022.

Papanastassiou, N. A note on convergence of sequences of functions. *Topology and its Applications*, 275:107017, 2020.

Radulović, D. On the bootstrap and empirical processes for dependent sequences. In *Empirical Process Techniques for Dependent Data*, pp. 345–364. Springer, 2002.

Römisch, W. Delta method, infinite dimensional. *Wiley StatsRef: Statistics Reference Online*, 2014.

Royset, J. O. and Szechtman, R. Optimal budget allocation for sample average approximation. *Operations Research*, 61:762–776, 2013.

Schirotzek, W. *Nonsmooth analysis*. Springer Science & Business Media, 2007.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

Serfling, R. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., Hoboken, 1980.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shapiro, A. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, 17: 841–858, 1989.

Shapiro, A. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

Shapiro, A. Statistical inference of stochastic optimization problems. In *Probabilistic Constrained Optimization: Methodology and Applications*, pp. 282–307. Springer, 2000.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

Sin, C.-Y. and White, H. Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71:207–225, 1996.

Singh, K. On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, pp. 1187–1195, 1981.

van der Vaart, A. and Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 2023.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, 1998.

Vidyasagar, M. *Learning and generalisation: with applications to neural networks*. Springer, London, 2003.

Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, pp. 307–333, 1989.

Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62, 1938.

Willard, S. *General Topology*. Courier Corporation, 2012.

A. Proofs

A.1. Theorem 3.1

We have that for any $g \in \ell^\infty(\mathcal{A})$, $g|_\mathcal{B} \in \ell^\infty(\mathcal{B})$, and if h_n converges to h in $\ell^\infty(\mathcal{A})$, then $h_n|_\mathcal{B}$ converges to $h|_\mathcal{B}$ in $\ell^\infty(\mathcal{B})$. The result is then immediate by Römisch (2014, Prop. 1) with universe $\mathcal{B}$.

A.2. Theorem 4.2

Let ι be as in Theorem 3.1 with $\mathcal{A} = \mathcal{X}$. This theorem gives that ι is Hadamard directionally differentiable on $\ell^\infty(\mathcal{X})$. By assumption $F(\Omega) \subset \ell^\infty(\mathcal{X})$, and we can modify $\tau_n(\hat{f}_n - f)$ on $\Omega \setminus \Omega_0$ (where $\mathbb{P}(\Omega_0) = 1$) so that it is also takes values in $\ell^\infty(\mathcal{X})$. The conditions on the delta method (Fact 3.2) are then satisfied from which we get

$$\tau_n(\iota(\hat{f}_n) - \iota(f)) \rightsquigarrow \iota'_f(F), \text{ and} \tag{12}$$

$$\tau_n(\iota(\hat{f}_n) - \iota(f)) - \iota'_f(\tau_n(\hat{f}_n - f)) = o_{\mathbb{P}^*}(1). \tag{13}$$

By definition, $\iota(\hat{f}_n) = \hat{\psi}_n$ and $\iota(f) = \psi^*$. Substituting in the expression for $\iota'_f(F)$, given in Theorem 3.1, into Equation (12) gives Equation (3). Rearranging Equation (13) gives

$$\begin{aligned}
\tau_n(\hat{\psi}_n - \psi^*) &= \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} (\tau_n(\hat{f}_n(x) - f(x))) + o_{\mathbb{P}*}(1) \\
\implies \hat{\psi}_n &= \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} (\hat{f}_n(x) - f(x)) + \psi^* + o_{\mathbb{P}*}(\tau_n^{-1}) \\
&= \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} (\hat{f}_n(x) - f(x) + \psi^*) + o_{\mathbb{P}*}(\tau_n^{-1}).
\end{aligned}$$

This is precisely Equation (4).

A.3. Corollary 4.4

Theorem 4.2 implies that

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} F(x).$$

If $\mathcal{S}^\epsilon$ is constant and equal to $\mathcal{S}$, for ϵ sufficiently small, then

$$\lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} F(x) = \inf_{x \in \mathcal{S}} F(x).$$

Because l is constant on $\mathcal{S}$, $\hat{f}_n$ and f is as well. Because $\tau_n(\hat{f}_n - f) \rightsquigarrow F$, Lemma A.1 gives that F is constant on $\mathcal{X}$. Hence, for any $x^* \in \mathcal{S}$

$$\inf_{x \in \mathcal{S}} F(x) = F(x^*).$$

Lemma A.1. *Let $\mathcal{X}$ be a set and let $\mathcal{A} \subseteq \mathcal{X}$. Let*

$$\mathcal{C}_\mathcal{A} = \{f \in \ell^\infty(\mathcal{X}) \,:\, |f(\mathcal{A})| = 1\}$$

where $|\cdot|$ denotes set cardinality. Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, $X_n : \Omega \to \mathcal{C}_\mathcal{A}$ be maps, and $X : \Omega \to \ell^\infty(\mathcal{X})$ be Borel measurable. If $X_n \rightsquigarrow X$ then, $X : \Omega \to \mathcal{C}_\mathcal{A}$.

Proof. We first claim that $\mathcal{C}_\mathcal{A}$ is closed. To show this take $(f_n) \subset \mathcal{C}_\mathcal{A}$ with $f_n \to f$ in $\ell^\infty(\mathcal{X})$. We aim to show that $f \in \mathcal{C}_\mathcal{A}$. For the sake of contradiction assume $f \notin \mathcal{C}_\mathcal{A}$. This means $\exists a, b \in \mathcal{A}$ such that $f(a) \neq f(b)$. Without loss of generality assume $f(a) > f(b)$ (else relabel). Let $\epsilon = (f(a) - f(b))/2$. Then for n sufficiently large,

$$|f_n(a) - f(a)| < \epsilon \text{ and } |f_n(b) - f(b)| < \epsilon$$

which implies

$$-\epsilon < f_n(a) - f(a) \text{ and } f_n(b) - f(b) < \epsilon.$$

By definition of ϵ,

$$\frac{f(b)-f(a)}{2} < f_n(a) - f(a) \text{ and } f_n(b) - f(b) < \frac{f(a)-f(b)}{2}$$

and so

$$\frac{f(b)+f(a)}{2} < f_n(a) \text{ and } f_n(b) < \frac{f(a)+f(b)}{2}.$$

That is, for n large enough $f_n(a) > f_n(b)$. However, f_n is constant on $\mathcal{A}$ by assumption so we have a contradiction. It must then be true that $f \in \mathcal{C}_{\mathcal{A}}$ and so $\mathcal{C}_{\mathcal{A}}$ is closed.

By the Portmanteau theorem, $\forall \mathcal{F} \subseteq \ell^\infty(\mathcal{X})$ closed,

$$\limsup_{n\to\infty} \mathbb{P}^*(X_n \in \mathcal{F}) \leq \mathbb{P}(X \in \mathcal{F}).$$

Taking $\mathcal{F} = \mathcal{C}_{\mathcal{A}}$ gives

$$1 = \limsup_{n\to\infty} \mathbb{P}^*(X_n \in C_{\mathcal{A}}) \leq \mathbb{P}(X \in \mathcal{C}_{\mathcal{A}})$$

and so $\mathbb{P}(X \in \mathcal{C}_{\mathcal{A}}) = 1$. This is the required result (up to a possible modification on a null set). □

A.4. Theorem 5.1

Let i_n generically denote one of $\hat{\iota}'_{s_n,n}$ or $\tilde{\iota}'_{t_n,n}$. Our proof strategy is to verify the following assumptions:

(a) $\sqrt{n}(\hat{f}_n - f) \rightsquigarrow \mathbb{G}$, for some tight, Gaunssian $\mathbb{G}$.

(b) $\sqrt{n}(f_n^b - \hat{f}_n)|(Z_i)_{i\in[n]} \overset{\mathbb{P}*}{\rightsquigarrow} \mathbb{G}$.

(c) $\sqrt{n}(f_n^b - \hat{f}_n)$ is asymptotically measurable.

(d) $\sqrt{n}(f_n^b - \hat{f}_n)$ is a measurable function of the weights $(W_i)_{i\in[n]}$ for fixed $(Z_i)_{i\in[n]}$.

(e) The weights are chosen independent of the data

(f) $\forall h_1, h_2 \in \ell^\infty(\mathcal{X})$

$$|i_n(h_1) - i_n(h_2)| \leq C_n \|h_1 - h_2\|_{\mathcal{X}},$$

where $C_n = O_{\mathbb{P}^*}(1)$.

(g) $\forall h \in \ell^\infty(\mathcal{X})$

$$i_n(h) \xrightarrow{\mathbb{P}*} \lim_{\epsilon \searrow 0} \inf_{S^\epsilon} h.$$

With all these assumptions along with the Hadamard directionally differentiability of ι given in Theorem 3.1, the result follows from Fang & Santos (2019, Thm. 3.2).

Properties (a)–(e) are precisely D1-D4.

Properties (f) and (g) need to be verified for each estimator $\hat{\iota}'_{s_n,n}$ and $\tilde{\iota}'_{t_n,n}$, separately.

For $\hat{\iota}_{s_n,n}$, (g) is given by Theorem 3.3 of Hong & Li (2018) and (f) follows by theorem 3.2 of Hong & Li (2018) if the inf map is Lipschitz continuous. This is indeed true as for $h_1, h_2 \in \ell^\infty(\mathcal{X})$,

$$|\inf_{\mathcal{X}} h_1 - \inf_{\mathcal{X}} h_2| \leq \sup_{\mathcal{X}} |h_1 - h_2| = ||h_1 - h_2||_{\mathcal{X}}.$$

For $\tilde{\iota}_{t_n,n}$, (f) follows by a similar argument to the one above, as for any $\mathcal{A} \subseteq \mathcal{X}$ and $h_1, h_2 \in \ell^\infty(\mathcal{X})$,

$$|\inf_{\mathcal{A}} h_1 - \inf_{\mathcal{A}} h_2| \leq \sup_{\mathcal{A}} |h_1 - h_2| \leq ||h_1 - h_2||_{\mathcal{X}}.$$

In particular, the inequality above holds for $\mathcal{A} = \mathcal{S}_n^{t_n}$. Because $\tau_n(\hat{f}_n - f) \rightsquigarrow F$, Lemma A.4 implies that $\tau_n(\hat{f}_n - f)$ is asymptotically bounded in probability in the sense that $\forall \delta > 0$, $\exists \mathcal{B} \subseteq \mathcal{X}$ bounded s.t.

$$\liminf_{n\to\infty} \mathbb{P}_*(X_n \in \mathcal{B}) \geq 1 - \delta$$

(g) then follows by Theorem A.2.

Theorem A.2. *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a non-empty set, and $\hat{f}_n, f : \Omega \to \ell^\infty(\mathcal{X})$. If there is $\tau_n \to \infty$ such that $\tau_n(\hat{f}_n - f)$ is asymptotically bounded in probability, then for any $\epsilon_n \searrow 0$ such that $\epsilon_n \tau_n \to \infty$ and any $h \in \ell^\infty(\mathcal{X})$,*

$$\inf_{\mathcal{S}_n^{\epsilon_n}} h \xrightarrow{\mathbb{P}_*} \lim_{\epsilon \searrow 0} \inf_{S^\epsilon} h.$$

Proof. We shall write $\mathcal{S}(f, \epsilon)$ to mean $\mathcal{S}^\epsilon$ and $\mathcal{S}(\hat{f}_n, \epsilon)$ to mean $\mathcal{S}_n^\epsilon$. Fix $\delta > 0$. We are required to show

$$\lim_{n\to\infty} \mathbb{P}^* \left(\left| \inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h \right| > \delta \right) = 0.$$

We firstly have

$$\begin{aligned}
\mathbb{P}^* \left(\left| \inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h \right| > \delta \right) &= \mathbb{P}^* \left(\left\{ \inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h > \delta \right\} \cup \left\{ \inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h < -\delta \right\} \right) \\
&\leq \mathbb{P}^* \left(\inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h > \delta \right) + \mathbb{P}^* \left(\inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h < -\delta \right).
\end{aligned}$$

It then suffices to show that each of these outer probabilities tend to 0. We have

$$\begin{aligned}
\mathcal{S}(\hat{f}_n, \epsilon_n) &= \{x \in \mathcal{X} \,:\, \hat{f}_n(x) \leq \inf_{\mathcal{X}} \hat{f}_n + \epsilon_n\} \\
&= \{x \in \mathcal{X} \,:\, \hat{f}_n(x) + f(x) - f(x) \leq \inf_{\mathcal{X}} (\hat{f}_n - f + f) + \epsilon_n\} \\
&\subseteq \{x \in \mathcal{X} \,:\, f(x) - \|\hat{f}_n - f\|_{\mathcal{X}} \leq \inf_{\mathcal{X}} f + \|\hat{f}_n - f\|_{\mathcal{X}} + \epsilon_n\} \\
&= \mathcal{S}(f, \epsilon_n + 2\|\hat{f}_n - f\|_{\mathcal{X}})
\end{aligned}$$

We hence get

$$\inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h \geq \inf_{\mathcal{S}(f, \epsilon_n + \|\hat{f}_n - f\|_{\mathcal{X}})} h$$

and so

$$\mathbb{P}^* \left(\inf_{\mathcal{S}(\hat{f}_n, \epsilon_n)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h < -\delta \right) \leq \mathbb{P}^* \left(\inf_{\mathcal{S}(f, \epsilon_n + \|\hat{f}_n - f\|_{\mathcal{X}})} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h < -\delta \right).$$

Then, observe that when $\eta \searrow 0$, we have

$$\inf_{\mathcal{S}(f,\eta)} h \nearrow \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h,$$

and hence, there is some $\eta_\delta > 0$ such that $\forall \eta < \eta_\delta$

$$\inf_{\mathcal{S}(f,\eta)} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h \geq -\delta.$$

By Lemma A.3, $||\hat{f}_n - f||_{\mathcal{X}} = o_{\mathbb{P}^*}(1)$ and because $\epsilon_n \searrow 0$,

$$\mathbb{P}_*(\epsilon_n + 2\|f_n - f\|_{\mathcal{X}} < \eta_\delta) \to 1$$

By definition of η_δ this gives

$$\mathbb{P}^* \left(\inf_{\mathcal{S}(f + \epsilon_n + 2\|\hat{f}_n - f\|_{\mathcal{X}})} h - \lim_{\epsilon \searrow 0} \inf_{S(f,\epsilon)} h < -\delta \right) \to 0.$$

and so

$$\mathbb{P}^*\left(\inf_{\mathcal{S}(\hat{f}_n,\epsilon_n)} h - \lim_{\epsilon\searrow 0}\inf_{S(f,\epsilon)} h < -\delta\right) \to 0.$$

Hence, we are done if we can show that

$$\mathbb{P}^*\left(\inf_{\mathcal{S}(\hat{f}_n,\epsilon_n)} h - \lim_{\epsilon\searrow 0}\inf_{S(f,\epsilon)} h > \delta\right) \to 0.$$

Via a similar argument to above we get

$$\mathcal{S}(\hat{f}_n,\epsilon_n) \supseteq \mathcal{S}(f,\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}})$$

and therefore

$$\inf_{\mathcal{S}(\hat{f}_n,\epsilon_n)} h \leq \inf_{\mathcal{S}(f,\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}})} h.$$

We then have

$$\mathbb{P}^*\left(\inf_{\mathcal{S}(\hat{f}_n,\epsilon_n)} h - \lim_{\epsilon\searrow 0}\inf_{S(f,\epsilon)} h > \delta\right) \leq \mathbb{P}^*\left(\inf_{\mathcal{S}(f,\epsilon_n - 2||\hat{f}_n - f||_{\mathcal{X}})} h - \lim_{\epsilon\searrow 0}\inf_{S(f,\epsilon)} h > \delta\right).$$

Because for any $\eta > 0$,

$$\inf_{\mathcal{S}(f,\eta)} h \leq \lim_{\epsilon\searrow 0}\inf_{S(f,\epsilon)} h$$

the result follows if

$$\mathbb{P}_*(\epsilon_n - 2||\hat{f}_n - f||_{\mathcal{X}} > 0) \to 1.$$

Because $\tau_n(\hat{f}_n - f)$ is asymptotically bounded in probability, $\forall \eta > 0$, $\exists M \in \mathbb{R}$ such that for n large enough

$$\mathbb{P}_*\left(||\hat{f}_n - f||_{\mathcal{X}} < \frac{M}{\tau_n}\right) > 1 - \eta$$

which is equivalent to

$$\mathbb{P}_*\left(\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}} > \epsilon_n - \frac{2M}{\tau_n}\right) > 1 - \eta.$$

It then suffices to show that for n large enough,

$$\epsilon_n - \frac{2M}{\tau_n} > 0$$

as this would give for n large enough

$$\mathbb{P}_*\left(\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}} > \epsilon_n - \frac{2M}{\tau_n}\right) \leq \mathbb{P}_*\left(\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}} > 0\right)$$

and hence give

$$\liminf_{n\to\infty} \mathbb{P}_*\left(\epsilon_n - 2\|\hat{f}_n - f\|_{\mathcal{X}} > 0\right) \geq 1 - \eta$$

The result then follows by sending η to 0. Rearrangement of this expression yields

$$\epsilon_n \tau_n > 2M,$$

which is true for n large enough as $\epsilon_n \tau_n \to \infty$. □

Lemma A.3. *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a normed space, and $\forall n \in \mathbb{N}$ let $X_n : \Omega \to \mathcal{X}$ maps. If there is a sequence $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n \to \infty$ such that $\tau_n X_n$ is asymptotically bounded in probability, then $X_n \xrightarrow{\mathbb{P}_*} 0$*

Proof. We aim to show that $\forall \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}_*(||X_n|| < \epsilon) = 1$$

Fix an $\epsilon > 0$. Then because $\tau_n X_n$ is asymptotically bounded in probability $\forall \delta > 0$, $\exists M > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}_*(||\tau_n X_n|| < M) = \liminf_{n\to\infty} \mathbb{P}_*(||X_n|| < M/\tau_n) \geq 1 - \delta$$

For n large enough

$$\mathbb{P}_*(||X_n|| < M/\tau_n) \leq \mathbb{P}_*(||X_n|| < \epsilon)$$

and so $\forall \delta > 0$

$$\liminf_{n\to\infty} \mathbb{P}_*(||X_n|| < \epsilon) \geq 1 - \delta$$

Sending $\delta \to 0$ gives

$$\liminf_{n\to\infty} \mathbb{P}_*(||X_n|| < \epsilon) \geq 1$$

which is equivalent to

$$\lim_{n\to\infty} \mathbb{P}_*(||X_n|| < \epsilon) = 1$$

This is exactly what we wanted to show. □

Lemma A.4. *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a metric space, $X : \Omega \to \mathcal{X}$ Borel measurable and $\forall n \in \mathbb{N}$ let $X_n : \Omega \to \mathcal{X}$ be maps. If $X_n \rightsquigarrow X$ then $(X_n)_{n\in\mathbb{N}}$ is asymptotically bounded in probability.*

Proof. Fix $\delta > 0$. We have

$$\mathbb{P}\left(X \in \bigcup_{m=1}^{\infty} B(0, m)\right) = 1$$

and so by continuity from below $\exists m \in \mathbb{N}$ such that

$$\mathbb{P}(X \in B(0, m)) \geq 1 - \delta$$

By the Portmanteau theorem we have

$$\liminf_{n\to\infty} \mathbb{P}_*(X_n \in B(0, m)) \geq \mathbb{P}(X \in B(0, m))$$

Combining these 2 inequalities gives the result. □

A.5. Corollary 5.2

D1 is satisfied by definition of Donsker. D2 follows by $\mathcal{H}$ being Donsker by Theorem 3.7.1 of (van der Vaart & Wellner, 2023). For D3, $\sqrt{n}(f_n^b - \hat{f}_n)$ as a function of the weights is simply a linear combination of elements of $\ell^\infty(\mathcal{X})$. Measurability follows by Lemma A.5. D4 is clear.

Lemma A.5. *Let $\mathcal{X}$ be a set. For any $n \in \mathbb{N}$ and fixed $f_1, \ldots, f_n \in \ell^\infty(\mathcal{X})$, let $s : \mathbb{R}^n \to \ell^\infty(\mathcal{X})$ be given by*

$$s(a_1, \ldots, a_n) = \sum_{i=1}^{n} a_i f_i.$$

The s is Borel measurable.

Proof. Because $\mathbb{R}^n$ and $\ell^\infty(\mathcal{X})$ are equipped with their Borel σ-algebras it suffices to show that s is continuous. We will actually show that s is Lipschitz continuous.

$$\begin{aligned}
||s(a) - s(b)||_{\mathcal{X}} &= \left\| \sum_{i=1}^{n} (a_i - b_i) f_i \right\|_{\mathcal{X}} \\
&\leq \sum_{i=1}^{\infty} ||(a_i - b_i) f_i||_{\mathcal{X}} \\
&\leq \sum_{i=1}^{\infty} |a_i - b_i| \, \|f_i\|_{\mathcal{X}} \\
&\leq \max_{i \in [n]} \|f_i\|_{\mathcal{X}} \|a - b\|_1 \\
&\leq n \max_{i \in [n]} \|f_i\|_{\mathcal{X}} \|a - b\|_2
\end{aligned}$$

Hence, s is Lipschitz continuous with Lipschitz constant at most $n \max_{i \in [n]} \|f_i\|_{\mathcal{X}}$. □

A.6. Theorem 5.3

Let $\iota_0, \iota_1 : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ given by

$$\iota_i(f) = \inf_{x \in \mathcal{X}_i} f(x)$$

By Theorem 3.1, ι_i is Hadamard directionally differentiable with derivative

$$\iota'_{i,f}(g) = \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(f, \epsilon, \mathcal{X}_i)} g(x).$$

In particular we have for any $g_n \to g$ in $\ell^\infty(\mathcal{X})$,

$$\tau_n \begin{pmatrix} \iota_1(f + g_n/\tau_n) - \iota_1(f) \\ \iota_0(f + g_n/\tau_n) - \iota_0(f) \end{pmatrix} \to \begin{pmatrix} \iota'_{1,f}(g) \\ \iota'_{0,f}(g). \end{pmatrix}$$

That is, $(\iota_1, \iota_0)^\top$ is Hadamard directionally differentiable. Modify $\hat{f}_n$ on $\Omega \setminus \Omega_0$ so that $\hat{f}_n$ is bounded. All the conditions on the delta method (Fact 3.2) are satisfied. We then get.

$$\tau_n \begin{pmatrix} \hat{\psi}_{1,n} - \psi_1^* \\ \hat{\psi}_{0,n} - \psi_0^* \end{pmatrix} \rightsquigarrow \begin{pmatrix} \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(f, \epsilon, \mathcal{X}_1)} F(x) \\ \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(f, \epsilon, \mathcal{X}_0)} F(x) \end{pmatrix}$$

Apply continuous mapping with the map $e : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $e(x, y) = x - y$ gives

$$\tau_n(\hat{\psi}_{1,n} - \psi_1^* - \hat{\psi}_{0,n} + \psi_0^*) \rightsquigarrow \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(f, \epsilon, \mathcal{X}_1)} F(x) - \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}(f, \epsilon, \mathcal{X}_0)} F(x).$$

Because $\psi_1^* = \psi_0^*$, this is the required result.

A.7. Corollary 5.4

Theorem 5.3 implies that

$$\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \rightsquigarrow F^*.$$

The Portmanteau theorem (see van der Vaart & Wellner, 2023, Thm. 1.3.4) gives this is equivalent to

$$\limsup_{n \to \infty} \mathbb{P}^*(\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \in \mathcal{C}) \leq \mathbb{P}(F^* \in \mathcal{C})$$

for all closed $\mathcal{C} \subseteq \mathbb{R}$. In particular, if we take $\mathcal{C} = (-\infty, c]$, for $c \in \overline{\mathbb{R}}$, we get

$$\limsup_{n \to \infty} \mathbb{P}^*(\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \leq c) \leq \mathbb{P}(F^* \leq c)$$

We then have for any c for which $\mathbb{P}(F^* \leq c) \leq \alpha$,

$$\limsup_{n\to\infty} \mathbb{P}^*(\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \leq c) \leq \alpha.$$

In particular, it holds for the largest c with this property:

$$c_\alpha = \sup\left\{c \in \overline{\mathbb{R}} : \mathbb{P}(F^* \leq c) \leq \alpha\right\}.$$

We have therefore shown that

$$\limsup_{n\to\infty} \mathbb{P}^*(\tau_n(\hat{\psi}_{1,n} - \hat{\psi}_{0,n}) \leq c_\alpha) \leq \alpha.$$

The LHS rearranges to give the required result.

A.8. Proposition 5.5

Let

$$\mathcal{K} = \arg\min_{k\in[m]} \psi_k^*.$$

Because $[m]$ is finite this, as well as k^* and $\hat{K}_n$, are well defined.

Note that by Lemma A.3, F1 implies that $\hat{\psi}_{k,n} \xrightarrow{\mathbb{P}*} \psi_k^*$, for each $k \in [m]$. Together with F2, we get $\forall\epsilon > 0$ and $\forall k \in [m]$, the sets

$$\left\{\omega : |\hat{\psi}_{k,n} - \psi_k^*| < \frac{\epsilon}{2}\right\}, \text{and} \left\{\omega : P_{k,n} < \frac{\epsilon}{2}\right\}$$

have inner probability tending to 1. On the intersection of these events

$$\begin{aligned}\hat{\psi}_{k,n} + P_{k,n} &> \hat{\psi}_{k,n} - P_{k,n} \\ &> \hat{\psi}_{k,n} - \frac{\epsilon}{2} \\ &> \psi_k^* - \epsilon\end{aligned}$$

For $k \notin \mathcal{K}$ we can take $\epsilon = (\psi_k^* - \psi_{k^*}^*)/2$ (note $\epsilon > 0$) to get

$$\begin{aligned}&= \psi_k^* - \frac{\psi_k^* - \psi_{k^*}^*}{2} \\ &= \frac{\psi_k^* + \psi_{k^*}^*}{2} \\ &= \psi_{k^*}^* + \frac{\psi_k^* - \psi_{k^*}^*}{2} \\ &= \psi_{k^*}^* + \epsilon \\ &> \hat{\psi}_{k^*,n} + \frac{\epsilon}{2} \\ &> \hat{\psi}_{k^*,n} + P_{k^*,n}\end{aligned}$$

Because $[m]$ is finite, on

$$\tilde{\Omega}_n := \bigcap_{k\notin\mathcal{K}} \left\{\omega : |\hat{\psi}_{k,n} - \psi_k^*| < \frac{\epsilon}{2}\right\} \cap \left\{\omega : P_{k,n} < \frac{\epsilon}{2}\right\},$$

we have

$$\inf_{k\notin\mathcal{K}} (\hat{\psi}_{k,n} + P_{k,n}) > \hat{\psi}_{k^*,n} + P_{k^*,n}.$$

Because $\tilde{\Omega}_n$ is the finite intersection of sets whose inner probability tends to 1, $\mathbb{P}_*(\tilde{\Omega}_n) \to 1$.

Fix $k \in \mathcal{K}$ and $\epsilon > 0$. By definition of asymptotically bounded in probability, $\exists M_1, M_2 \in \mathbb{R}$ such that for n large enough,

$$\begin{aligned}\mathbb{P}_*(\tau_n|\hat{\psi}_{k,n} - \psi_k^*| < M_1) &\geq 1 - \frac{\epsilon}{2}, \\ \mathbb{P}_*(\tau_n|\hat{\psi}_{k^*,n} - \psi_{k^*}^*| < M_2) &\geq 1 - \frac{\epsilon}{2}.\end{aligned}$$

which are of course equivalent to

$$\mathbb{P}_*(-M_1 < \tau_n(\hat{\psi}_{k,n} - \psi_k^*) < M_1) \geq 1 - \frac{\epsilon}{2},$$
$$\mathbb{P}_*(-M_2 < \tau_n(\psi_{k^*}^* - \hat{\psi}_{k^*,n}) < M_2) \geq 1 - \frac{\epsilon}{2}.$$

We hence obtain

$$\mathbb{P}_*(-M_1 - M_2 < \tau_n(\hat{\psi}_{k,n} - \psi_k^*) - \tau_n(\hat{\psi}_{k^*,n} - \psi_{k^*}^*) < M_2 + M_1) \geq 1 - \epsilon.$$

By definition of $\mathcal{K}$, $\psi_k^* = \psi_{k^*}^*$ and so the above expression simplifies to

$$\mathbb{P}_*(-M_1 - M_2 < \tau_n(\hat{\psi}_{k,n} - \hat{\psi}_{k^*,n}) < M_2 + M_1) \geq 1 - \epsilon$$
$$\iff \mathbb{P}_*(\tau_n|(\hat{\psi}_{k,n} - \hat{\psi}_{k^*,n})| < M_2 + M_1) \geq 1 - \epsilon.$$

If $k \neq k^*$, then $k > k^*$ and by F2 $\tau_n(P_{k,n} - P_{k^*,n}) \xrightarrow{\mathbb{P}_*} \infty$. By definition for n large enough

$$\mathbb{P}_*(P_{k,n} - P_{k^*,n} > (M_1 + M_2)/\tau_n) \geq 1 - \epsilon.$$

Combining the above results we get for n large enough

$$\mathbb{P}_*\left(|\hat{\psi}_{k,n} - \hat{\psi}_{k^*,n}| < \frac{M_1 + M_2}{\tau_n}\right) \geq 1 - \epsilon,$$
$$\mathbb{P}_*\left(P_{k,n} - P_{k^*,n} > \frac{M_1 + M_2}{\tau_n}\right) \geq 1 - \epsilon.$$

Let

$$\overline{\Omega}_{k,n} = \left\{\omega \,:\, |\hat{\psi}_{k,n} - \hat{\psi}_{k^*,n}| < \frac{M_1 + M_2}{\tau_n}\right\} \cap \left\{\omega \,:\, P_{k,n} - P_{k^*,n} > \frac{M_1 + M_2}{\tau_n}\right\}.$$

Then on $\overline{\Omega}_{k,n}$ we have

$$\hat{\psi}_{k^*,n} - \hat{\psi}_{k,n} < \frac{M_1 + M_2}{\tau_n} < P_{k,n} - P_{k^*,n}$$
$$\implies \hat{\psi}_{k^*,n} + P_{k^*,n} < \hat{\psi}_{k,n} + P_{k,n}$$

Let

$$\overline{\Omega}_n = \bigcap_{k \in \mathcal{K}} \overline{\Omega}_{k,n}$$

Then $\mathbb{P}_*(\overline{\Omega}_n) \geq 1 - 2|\mathcal{K}|\epsilon$ and on this set

$$\hat{\psi}_{k^*,n} + P_{k^*,n} < \inf_{k \in \mathcal{K} \setminus \{k^*\}} (\hat{\psi}_{k,n} + P_{k,n}).$$

Hence, $\mathbb{P}_*(\tilde{\Omega}_n \cap \overline{\Omega}_n) \to 1$ and on $\tilde{\Omega}_n \cap \overline{\Omega}_n$

$$\hat{\psi}_{k^*,n} + P_{k^*,n} < \inf_{k \in [m] \setminus \{k^*\}} (\hat{\psi}_{k,n} + P_{k,n}).$$

That is on $\tilde{\Omega}_n \cap \overline{\Omega}_n$

$$\{k^*\} = \text{argmin}_{k \in [m]} (\hat{\psi}_{k,n} + P_{k,n})$$

and so in particular

$$k^* = \min\left[\text{argmin}_{k \in \mathcal{K}} (\hat{\psi}_{k,n} + P_{k,n})\right].$$

I.e.

$$\mathbb{P}_*(\hat{K}_n = k^*) \to 1$$

as required.

A.9. Theorem 6.1

First proving

$$\tau_n(g_n(X_n) - h_n(\mu)) \rightsquigarrow D_\mu(X) \tag{14}$$

Define $d_n : \mathcal{U} \to \mathcal{V}$ via

$$d_n(x) = \tau_n(g_n(\mu + x/\tau_n) - h_n(\mu))$$

By assumption, for any $\eta_n \to \eta$, $g_n(\eta_n) \to D_\mu(\eta)$. Generalized continuous mapping then gives

$$d_n(\tau_n(X_n - \mu)) \rightsquigarrow D_\mu(X)$$

The expanded form of the LHS is

$$\tau_n(g_n(X_n) - h_n(\mu))$$

This is exactly Equation (14).

Now showing

$$\tau_n(g_n(X_n) - h_n(\mu)) - D_\mu(\tau_n(X_n - \mu)) = o_{\mathbb{P}*}(1) \tag{15}$$

Next define $\tilde{d}_n : \mathcal{U} \to \mathcal{V} \times \mathcal{V}$ via

$$\tilde{d}_n(x) = \begin{pmatrix} d_n(x) \\ D_\mu(x) \end{pmatrix}$$

Because $D_\mu(x)$ is converged to continuously, it must be continuous (Papanastassiou, 2020, Prop. 2.5). By assumption and the continuity of D_μ, for any $x_n \to x$

$$\tilde{d}_n(x_n) \to (D_\mu(x),\, D_\mu(x)).$$

Generalized continuous mapping then gives

$$\tilde{d}_n(\tau_n(X_n - \mu)) \rightsquigarrow \begin{pmatrix} D_\mu(X) \\ D_\mu(X) \end{pmatrix}$$

The expanded form of the LHS is

$$\begin{pmatrix} \tau_n(g_n(X_n) - h_n(\mu)) \\ D_\mu(\tau_n(X_n - \mu)). \end{pmatrix}$$

Applying continuous mapping (see (van der Vaart & Wellner, 2023, thrm 1.3.6)) with the map $s : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$, $s(y_1, y_2) = y_1 - y_2$ gives

$$\tau_n(g(X_n) - g(\mu)) - D_\mu(\tau_n(X_n - \mu)) \rightsquigarrow 0$$

Because weak convergence to a constant implies convergence in outer probability to that constant (see van der Vaart & Wellner, 2023, Lem. 1.10.2) we get

$$\tau_n(g(X_n) - g(\mu)) - D_\mu(\tau_n(X_n - \mu)) \xrightarrow{\mathbb{P}*} 0$$

This is exactly Equation (15).

A.10. Theorem 6.2

In this section, for a general function $h : \mathcal{X} \to \mathbb{R}$ we write

$$\mathcal{S}(h, \epsilon) = \left\{ x \in \mathcal{X} \,:\, h(x) \leq \inf_{\mathcal{X}} h + \epsilon \right\}.$$

First some helpful lemmas.

Lemma A.6. *Let $\mathcal{X}$ be a set and $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$. $\forall n \in \mathbb{N}$ let $f, g, g_n : \mathcal{X} \to \overline{\mathbb{R}}$ be bounded below with $||g_n - g||_{\mathcal{X}} \to 0$ and let $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$. Then*

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f)}{t_n} = \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}}(f)}{t_n}$$

provided either limit exists.

Proof. If

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f)}{t_n}$$

exists then

$$\begin{aligned}\frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f)}{t_n} - ||g_n - g||_{\mathcal{X}} &\leq \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}}(f)}{t_n} \\ &\leq \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f)}{t_n} + ||g_n - g||_{\mathcal{X}}\end{aligned}$$

and so so the result follows by squeeze theorem. If

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}}(f)}{t_n}$$

exists then

$$\begin{aligned}\frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}}(f)}{t_n} - ||g_n - g||_{\mathcal{X}} &\leq \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f)}{t_n} \\ &\leq \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}}(f)}{t_n} + ||g_n - g||_{\mathcal{X}}\end{aligned}$$

and so the result again follows by squeeze theorem. □

Lemma A.7. *Let $\mathcal{X}$ be a set, $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$ with $\mathcal{X}_n \neq \emptyset$, $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$, let $f, g : \mathcal{X} \to \overline{\mathbb{R}}$ be bounded below and let*

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n} f - \inf_{\mathcal{X}} f}{t_n} = 0$$

Then

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}} f}{t_n} = \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n}$$

provided either limit exists.

Proof. If

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}} f}{t_n}$$

exists then we have

$$\begin{aligned}\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}} f}{t_n} &= \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}} f}{t_n} - \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n} f - \inf_{\mathcal{X}} f}{t_n} \\ &= \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n}.\end{aligned}$$

Similarly if

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n}$$

exists we have

$$\begin{aligned}\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n} &= \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n} + \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n} f - \inf_{\mathcal{X}} f}{t_n} \\ &= \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}} f}{t_n}.\end{aligned}$$

□

Lemma A.8. *Let $\mathcal{X}$ be a non-empty, compact, metric space and $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq X$ be such that $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{X}_n$. Let $f \in \mathrm{lsc}(\mathcal{X})$ and let $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$.*

Then

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n} f - \inf_{\mathcal{X}} f}{t_n} = 0$$

Proof. Because f is lsc and $\mathcal{X}$ is compact, by EVT f attains it's minima. Let x_0 be such a minimizer. Then $\exists N \in \mathbb{N}$ s.t. $\forall n > N$, $x_0 \in \mathcal{X}_n$. Hence $\forall n > N$ $\inf_{\mathcal{X}_n} f = \inf_{\mathcal{X}} f$. The result is then immediate. □

Lemma A.9. *Let $\mathcal{X}$ be a compact metric space, and $f, g : \mathcal{X} \to \mathbb{R}$ be lower semi-continuous. Then*

$$\lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g = \inf_{\mathcal{S}} g.$$

Proof. Recall the fact that a lower semi-continuous function achieves its minimum on a compact set and hence $\mathcal{S} \neq \emptyset$. Observe that for any $\epsilon > 0$, $\mathcal{S}^\epsilon \supseteq \mathcal{S}$ and hence

$$\lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g \leq \inf_{\mathcal{S}} g. \tag{16}$$

Because f is lsc, for any $\epsilon > 0$, $\mathcal{S}^\epsilon$ is closed. Because $\mathcal{X}$ is compact we then have $\mathcal{S}^\epsilon$ is compact. For any integer $n \geq 1$, since $\mathcal{S}^{1/n}$ is compact and g is lower semi-continuous, there exists a minimizer x_n,

$$g(x_n) = \inf_{\mathcal{S}^{1/n}} g \leq \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g.$$

Since $\mathcal{X}$ is compact, (x_n) has a converging subsequence, which we assume is just (x_n) itself, without loss of generality. Write $x_0 = \lim x_n$. Since, $x_n \in \mathcal{S}^{1/n}$

$$f(x_n) \leq \inf_{\mathcal{X}} f + \frac{1}{n}$$

and so

$$f(x_0) \leq \liminf_{n\to\infty} f(x_n) \leq \inf_{\mathcal{X}} f$$

This implies $x_0 \in \mathcal{S}$. Note that $(g(x_n))_{n\in\mathbb{N}}$ is a non-decreasing sequence and

$$\inf_{x\in\mathcal{S}(f,0)} g(x) \leq g(x_0) \leq \lim_{n\to\infty} g(x_n) \leq \lim_{\epsilon\searrow 0} \inf_{x\in\mathcal{S}(f,\epsilon)} g(x).$$

This, together with (16), completes the proof. □

Theorem A.10. *Let $\mathcal{X}$ be a non-empty, compact, metric space and $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$ be such that $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{X}_n$. Let $f \in \mathrm{lsc}(\mathcal{X})$ and let $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$.*

For any $g \in \mathrm{lsc}(\mathcal{X})$

$$\inf_{\mathcal{S}(f,0)} g = \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n}$$

Proof. We will show

$$\inf_{\mathcal{S}(f,0)} g \leq \liminf_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n} \text{ and} \tag{17}$$

$$\inf_{\mathcal{S}(f,0)} g \geq \limsup_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g) - \inf_{\mathcal{X}_n} f}{t_n}. \tag{18}$$

The combination of both will show the result.

First showing Equation (17). Take any $j_n > 0$ s.t. $j_n/t_n \to 0$ and take $x \in \mathcal{S}_n(f + t_n g, j_n)$ where

$$\mathcal{S}_n(f, \epsilon) = \left\{ x \in \mathcal{X}_n \,:\, f(x) \leq \inf_{\mathcal{X}_n} f + \epsilon \right\}$$

Then

$$\begin{aligned}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n} &\geq \frac{f(x)+t_ng(x)-j_n-f(x)}{t_n}\\ &= \frac{t_ng(x)-j_n}{t_n}\\ &= g_n(x)-\frac{j_n}{t_n}\\ &\geq \inf_{\mathcal{S}_n(f+t_ng,j_n)} g-\frac{j_n}{t_n}\\ \implies \liminf_{n\to\infty}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n} &\geq \liminf_{n\to\infty}\inf_{\mathcal{S}_n(f+t_ng,j_n)} g\end{aligned}$$

We have

$$\begin{aligned}\mathcal{S}_n(f+t_ng,j_n) &= \left\{x\in\mathcal{X}_n\,:\,f(x)+t_ng(x)\leq \inf_{\mathcal{X}_n}(f+t_ng)+j_n\right\}\\ &\subseteq \left\{x\in\mathcal{X}_n\,:\,f(x)-t_n||g||_{\mathcal{X}}\leq \inf_{\mathcal{X}_n}f+t_n||g||_{\mathcal{X}}+j_n\right\}\\ &= \left\{x\in\mathcal{X}_n\,:\,f(x)\leq \inf_{\mathcal{X}_n}f+2t_n||g||_{\mathcal{X}}+j_n\right\}\\ &= \left\{x\in\mathcal{X}_n\,:\,f(x)\leq \inf_{\mathcal{X}}f+2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f\right\}\\ &= \mathcal{S}(f,2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f)\cap\mathcal{X}_n\\ &\subseteq \mathcal{S}(f,2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f)\end{aligned}$$

Hence,

$$\liminf_{n\to\infty}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n}\geq \liminf_{n\to\infty}\inf_{\mathcal{S}(f,2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f)} g.$$

Because $2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f\searrow 0$

$$\liminf_{n\to\infty}\inf_{\mathcal{S}(f,2t_n||g||_{\mathcal{X}}+j_n+\inf_{\mathcal{X}_n}f-\inf_{\mathcal{X}}f)} g=\lim_{\epsilon\searrow 0}\inf_{\mathcal{S}(f,\epsilon)} g$$

and so

$$\liminf_{n\to\infty}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n}\geq \lim_{\epsilon\searrow 0}\inf_{\mathcal{S}(f,\epsilon)} g.$$

By Lemma A.9

$$\lim_{\epsilon\searrow 0}\inf_{\mathcal{S}(f,\epsilon)} g=\inf_{\mathcal{S}(f,0)} g.$$

Finally, we have

$$\liminf_{n\to\infty}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n}\geq \inf_{\mathcal{S}(f,0)} g.$$

This is precisely Equation (17).

Now proving Equation (18). Fix $s_n>0$, such that $s_n/t_n\to 0$ and take $x\in S_n(f,s_n)$. Then

$$\begin{aligned}\frac{\inf_{\mathcal{X}_n}(f+t_ng)-\inf_{\mathcal{X}_n}f}{t_n} &\leq \frac{f(x)+t_ng(x)-f(x)+s_n}{t_n}\\ &= \frac{t_ng(x)+s_n}{t_n}\\ &= g(x)+\frac{s_n}{t_n}\end{aligned}$$

This is true for any x so

$$\frac{\inf_{\mathcal{X}_n}(f+t_n g)-\inf_{\mathcal{X}_n} f}{t_n} \leq \inf_{\mathcal{S}_n(f,s_n)} g + \frac{s_n}{t_n}$$
$$\implies \limsup_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f+t_n g)-\inf_{\mathcal{X}_n} f}{t_n} \leq \limsup_{n\to\infty} \inf_{\mathcal{S}_n(f,s_n)} g$$

We are then done if we can show that

$$\limsup_{n\to\infty} \inf_{\mathcal{S}_n(f,s_n)} g \leq \lim_{\epsilon\searrow 0} \inf_{\mathcal{S}(f,0)} g$$

Because f is lsc, $\mathcal{S}(f,0)$ is closed. Because $\mathcal{X}$ is compact, $S(f,0)$ is then also compact. Then by EVT we have $\inf_{\mathcal{S}(f,0)} g$ is attained. Let the point of attainment be x^*. We have $\exists N \in \mathbb{N}$ s.t. $\forall n > N$ $x^* \in \mathcal{X}_n$. Then $\forall n > N$

$$\inf_{S(f,0)} g = g(x^*) = \inf_{\mathcal{S}(f,0)\cap\mathcal{X}_n} g \geq \inf_{\mathcal{S}(f,s_n)\cap\mathcal{X}_n} g$$

Hence

$$\limsup_{n\to\infty} \inf_{\mathcal{S}_n(f,s_n)\cap\mathcal{X}_n} \leq \inf_{S(f,0)} g$$

which is the required result. □

Combining all this together we get the following result. This is precisely what we want to show.

Theorem A.11. *Let $\mathcal{X}$ be a non-empty, compact, metric space and $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq X$ be such that $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{X}_n$. Let $f \in \mathrm{lsc}(\mathcal{X})$ and let $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$.*

For any $g_n, g \in \mathrm{lsc}(\mathcal{X})$ with $||g_n - g||_{\mathcal{X}} \to 0$

$$\inf_{\mathcal{S}(f,0)} g = \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f+t_n g_n)-\inf_{\mathcal{X}} f}{t_n}, \text{ and} \tag{19}$$

$$\inf_{\mathcal{S}(f,0)} g = \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f+t_n g_n)-\inf_{\mathcal{X}_n} f}{t_n}. \tag{20}$$

Proof. Lemma A.6 gives that it suffices to show Gateaux differentiability. Lemma A.8 gives that the conditions on Lemma A.7 are satisfied and hence the Gateaux differentiable forms of Equation (8) and Equation (9) are equivalent. Theorem A.10 gives the Gateaux differentiable form of Equation (9). □

A.11. Theorem 6.3

Again, in this section, for a general function $h : \mathcal{X} \to \mathbb{R}$ we write

$$\mathcal{S}(h,\epsilon) = \left\{ x \in \mathcal{X} \,:\, h(x) \leq \inf_{\mathcal{X}} h + \epsilon \right\}.$$

First some helpful results

Theorem A.12. *Let $\mathcal{X}$ be a set, and $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$. Let $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$, $(j_n) \subseteq \mathbb{R}^+$ with $j_n/t_n \to 0$, and let $f \in \ell^\infty(\mathcal{X})$. If*

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n^c} f - \inf_{\mathcal{X}} f}{t_n} = \infty, \tag{21}$$

then $\forall (g_n)_{n\in\mathbb{N}} \subset \ell^\infty(\mathcal{X})$ uniformly bounded, $\exists N \in \mathbb{N}$ s.t. $\forall n > N$

$$S(f + t_n g_n, j_n) \cap \mathcal{X}_n \neq \emptyset.$$

Proof. Take any $(g_n)_{n\in\mathbb{N}} \subset \ell^\infty(\mathcal{X})$ uniformly bounded and let the uniform bound be C. That is, $\forall n \in \mathbb{N}$, $||g_n||_\infty \leq C$.

Because $j_n/t_n \to 0$, $\exists N_1 \in \mathbb{N}$ such that $\forall n > N_1$, $j_n \leq t_n C$. Similarly, Equation (21) gives that $\exists N_2 \in \mathbb{N}$ such that $\forall n > N_2$

$$\inf_{\mathcal{X}_n^c} f > \inf_{\mathcal{X}} f + 3t_n C$$

Take any $x \in \mathcal{S}(f + t_n g, j_n)$. By definition

$$\begin{aligned} f(x) + t_n g(x) &\leq \inf_{\mathcal{X}}(f + t_n g_n) + j_n \\ &\leq \inf_{\mathcal{X}} f + t_n \sup_{\mathcal{X}} g_n + j_n \\ \implies f(x) &\leq \inf_{\mathcal{X}} f + t_n(\sup_{\mathcal{X}} g_n - g_n(x)) + j_n \\ &\leq \inf_{\mathcal{X}} f + 2t_n C + j_n \end{aligned}$$

If $n > N_1$ then

$$\leq \inf_{\mathcal{X}} f + 3t_n C$$

If additionally $n > N_2$ then

$$< \inf_{\mathcal{X}_n^c} f$$

That is $\forall n > \max\{N_1, N_2\}$, $f(x) < \inf_{\mathcal{X}_n^c} f$ and so $x \in \mathcal{X}_n$. Hence, $\forall n > \max\{N_1, N_2\}$ $x \in \mathcal{S}(f + t_n g_n, j_n) \cap \mathcal{X}_n$ and so in particular $\forall n > \max\{N_1, N_2\}$

$$\mathcal{S}(f + t_n g_n, j_n) \cap \mathcal{X}_n \neq \emptyset$$

as required. □

Corollary A.13. *Let $\mathcal{X}$ be a set, $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$, $f \in \ell^\infty(\mathcal{X})$, $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$, $(j_n) \subseteq \mathbb{R}^+$ with $j_n/t_n \to 0$. If $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and $\exists N \in \mathbb{N}$ such that*

$$\inf_{\mathcal{X}_N^c} f > \inf_{\mathcal{X}} f.$$

Then $\exists N \in \mathbb{N}$ s.t. $\forall n > N$

$$\mathcal{S}(f + t_n g_n, j_n) \cap \mathcal{X}_n \neq \emptyset.$$

Proof. By monotonicity of $\mathcal{X}_n$ we have

$$\inf_{\mathcal{X}_1^c} f \leq \inf_{\mathcal{X}_2^c} f \leq \cdots$$

and so $\exists \epsilon > 0$ s.t. $\forall n > N$

$$\inf_{\mathcal{X}_N^c} f > \inf_{\mathcal{X}} f + \epsilon.$$

Hence,

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n^c} f - \inf_{\mathcal{X}} f}{t_n} \geq \lim_{n\to\infty} \frac{\epsilon}{t_n} = \infty$$

The result is then immediate by Theorem A.12. □

The following result means that conclusion of Corollary A.13 gives us Equation (10).

Theorem A.14. *Let $\mathcal{X}$ be a set, $\forall n \in \mathbb{N}$ let $\mathcal{X}_n \subseteq \mathcal{X}$ with $\mathcal{X}_n \neq \emptyset$, $f \in \ell^\infty(\mathcal{X})$, $(t_n) \subseteq \mathbb{R}^+$ with $t_n \to 0$. Then for any $g \in \ell^\infty(\mathcal{X})$ and $(g_n)_{n\in\mathbb{N}} \subset \ell^\infty(\mathcal{X})$ with $||g_n - g||_{\mathcal{X}} \to 0$ the following are equivalent*

1. $$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}} f}{t_n} = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}(f,\epsilon)} g$$

2. $$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f + t_n g_n)}{t_n} = 0$$

3. $\exists (s_n) \subseteq \mathbb{R}^+$ *with* $s_n/t_n \to 0$ *such that* $\exists N \in \mathbb{N}$ *s.t.* $\forall n > N$

$$S(f + t_n g_n, s_n) \cap \mathcal{X}_n \neq \emptyset$$

Proof. (1) $\iff$ (2)
Theorem 3.1 gives

$$\lim_{n\to\infty} \frac{\inf_{\mathcal{X}}(f + t_n g_n) - \inf_{\mathcal{X}} f}{t_n} = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}(f,\epsilon)} g$$

So (1) happens if and only if

$$\begin{aligned} 0 &= \lim_{n\to\infty} \left| \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}} f}{t_n} - \frac{\inf_{\mathcal{X}}(f + t_n g_n) - \inf_{\mathcal{X}} f}{t_n} \right| \\ &= \lim_{n\to\infty} \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f + t_n g_n)}{t_n}. \end{aligned}$$

This is precisely (2).

(2) $\implies$ (3)
Take

$$s_n = \inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f + t_n g_n) + t_n^2.$$

We always have

$$\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f + t_n g_n) \geq 0$$

so $s_n > 0$. Additionally by assumption $s_n/t_n \to 0$. Finally,

$$S(f + t_n g_n, s_n) \cap \mathcal{X}_n = \left\{ x \in \mathcal{X}_n \,:\, f(x) + t_n g_n(x) \leq \inf_{\mathcal{X}}(f + t_n g_n) + s_n \right\} \tag{22}$$

$$= \left\{ x \in \mathcal{X}_n \,:\, f(x) + t_n g_n(x) \leq \inf_{\mathcal{X}_n}(f + t_n g_n) + t_n^2 \right\} \tag{23}$$

and this is non-empty as ϵ-minimizers are always non-empty.

(3) $\implies$ (2)
By assumption $\exists s_n > 0$ with $s_n/t_n \to 0$ and $S(f + t_n g_n, s_n) \cap \mathcal{X}_n \neq \emptyset$. Hence take $x \in S(f + t_n g_n, s_n) \cap \mathcal{X}_n$. Then

$$0 \leq \frac{\inf_{\mathcal{X}_n}(f + t_n g_n) - \inf_{\mathcal{X}}(f + t_n g_n)}{t_n} \tag{24}$$

$$\leq \frac{f(x) + t_n g_n(x) - \inf_{\mathcal{X}}(f + t_n g_n)}{t_n} \tag{25}$$

$$\leq \frac{f(x) + t_n g_n(x) - f(x) - t_n g_n(x) + s_n}{t_n} \tag{26}$$

$$= \frac{s_n}{t_n}. \tag{27}$$

The result then follows by squeeze theorem. □

To show Equation (11) we note that if $\forall n > N$

$$\inf_{\mathcal{X}_n^c} f > \inf_{\mathcal{X}} f$$

then $\forall n > N$,

$$\inf_{\mathcal{X}_n} f = \inf_{\mathcal{X}} f$$

Equation (11) then follows by Lemma A.7.

B. Directional differentiable results

Theorem 3.1 gives that the infimum map is Hadamard directionally differentiable. Here we investigate conditions for which the infimum is (fully) Hadamard differentiable. Carcamo et al. (2020, Corr 2.4) provides the following result.

Theorem B.1. *Let $\mathcal{X}$ be a compact metric and $\iota : \mathcal{C}(\mathcal{X}) \to \mathbb{R}$ be given by*

$$\iota(f) = \inf_{x \in \mathcal{X}} f(x)$$

Then ι is Hadamard differentiable at any $f \in \mathcal{C}(\mathcal{X})$ tangentially to $\mathcal{C}(\mathcal{X})$ iff $|\mathcal{S}| = 1$ where $|\cdot|$ denotes cardinality.

It is possible to extend this result for our applications because we do not need Hadamard differentiability tangentially to all of $\mathcal{C}(\mathcal{X})$. We can use knowledge of the loss function to give information about the support of the image measure of F and hence reduce the space where we are required to show differentiability. The results presented are in no sense complete but are adequate in our context.

B.1. The continuous case

Theorem B.2. *Let $\mathcal{X}$ be a non-empty compact metric space, f lower semi-continuous and bounded, and $\iota : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ given by*

$$\iota(g) = \inf_{x \in \mathcal{X}} g(x).$$

Let

$$\mathcal{A}_f = \{g \in \operatorname{lsc}(\mathcal{X}) \cap \ell^\infty(\mathcal{X}) : |g(\mathcal{S})| = 1\}.$$

Then ι is Hadamard differentiable tangentially to $\mathcal{A}_f$.

Proof. Theorem 3.1 implies that ι is Hadamard directionally differentiable at f with derivative

$$\iota'_f(g) = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g$$

Because we are only considering the case of differentiability tangentially to $\mathcal{A}_f$, we consider $g \in \mathcal{A}_f \subseteq \operatorname{lsc}(\mathcal{X})$. Lemma A.9 gives that this expression reduces to

$$\iota'_f(g) = \inf_{\mathcal{S}} g$$

Because $|g(\mathcal{S})| = 1$, this expression further reduces to

$$\iota'_f(g) = g(x^*),$$

for any $x^* \in \mathcal{S}$. This equation is clearly linear in g. □

B.2. The bounded case

We first provide an extension of the result in Carcamo et al. (2020, Cor. 2.4) to the case of bounded of functions. This is presented to show the difficulty of getting full differentiability in this case.

Theorem B.3. *For $\mathcal{X}$ be a non-empty set and $\iota : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ given by*

$$\iota(g) = \inf_{x \in \mathcal{X}} g(x)$$

ι is Hadamard differentiable at $f \in \ell^\infty(\mathcal{X})$ if and only if $\exists \tilde{\epsilon} > 0$ such that $\forall \epsilon \in (0, \tilde{\epsilon})$

$$\mathcal{S}^\epsilon = \{\tilde{x}\}$$

for some $\tilde{x}$ depending only on f.

Proof. By Theorem 3.1, ι is Hadamard directionally differentiable at f with derivative

$$\iota'_f(g) = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g.$$

We then have to show that this expression is linear iff $\mathcal{S}^\epsilon$ is eventually a singleton.

Sufficiency is clear as if $\mathcal{S}^\epsilon$ is eventually equal to $\{\tilde{x}\}$

$$\iota'_f(g) = g(\tilde{x}),$$

which is clearly linear.

To show necessity, observe that linearity is really two conditions:

1. $\forall \lambda \in \mathbb{R}, \forall g \in \mathcal{L}^\infty(\mathcal{X}), \iota'_f(\lambda g) = \lambda \iota'_f(g)$.
2. $\forall g, h \in \mathcal{L}^\infty(\mathcal{X}), \iota'_f(g+h) = \iota'_f(g) + \iota'_f(h)$.

It is sufficient to show that one of these conditions imply $\mathcal{S}^\epsilon$ is eventually a singleton as then both conditions together will imply this. Doing this for the first condition.

When $\lambda < 0$ this condition reduces to: $\forall g \in \ell^\infty(\mathcal{X})$

$$\lambda \lim_{\epsilon \searrow 0} \sup_{\mathcal{S}^\epsilon} g = \lambda \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g.$$

It then suffices to show that if $\forall g \in \ell^\infty(\mathcal{X})$

$$\lim_{\epsilon \searrow 0} \sup_{\mathcal{S}^\epsilon} g = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g$$

then $\mathcal{S}^\epsilon$ is eventually a singleton. There are 3 possible cases (as $\mathcal{S}^\epsilon$ is never empty):

1. $\mathcal{S}^\epsilon$ is eventually a singleton.
2. $\mathcal{S}^\epsilon$ is eventually a set containing 2 or more elements.
3. $\mathcal{S}^\epsilon$ is not eventually constant.

Showing that 2 and 3 are not possible.

2 Denote the eventually constant value of $\mathcal{S}^\epsilon$ as $\mathcal{S}$. Take any $y, z \in \mathcal{S}$ with $y \neq z$ and define

$$g(x) = \begin{cases} 1 & x \neq y,\ x \neq z \\ 2 & x = y \\ 0 & x = z \end{cases}$$

g is clearly bounded and

$$2 = \sup_{\mathcal{S}} g = \lim_{\epsilon \searrow 0} \sup_{\mathcal{S}^\epsilon} g(x) \neq \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g(x) = \inf_{\mathcal{S}} g = 0.$$

Hence 2 is not possible.

3 $\mathcal{S}^\epsilon$ not being eventually constant means that $\forall \epsilon > 0, \exists \delta \in (0, \epsilon)$, such that

$$\mathcal{S}^\delta \subset \mathcal{S}^\epsilon.$$

where "$\subset$" denotes proper subset ($\subseteq$ trivially holds always). Let $(\epsilon_n) \subset \mathbb{R}^+$ be a strictly decreasing sequence converging to 0 such that

$$\mathcal{S}^{\epsilon_{n+1}} \subset \mathcal{S}^{\epsilon_n}.$$

On $\mathcal{S}^{\epsilon_{2n}} \setminus \mathcal{S}^{\epsilon_{2n+1}}$ define g as 2. On $\mathcal{S}^{\epsilon_{2n+1}} \setminus \mathcal{S}^{\epsilon_{2n+2}}$ define g as 0. At any other point define g as 1. Then we again have

$$2 = \lim_{\epsilon \searrow 0} \sup_{\mathcal{S}^\epsilon} g \neq \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g = 0.$$

□

The following theorem is of primary interest.

Theorem B.4. *Let $\mathcal{X}$ be a set and $\iota : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ be given by*

$$\iota(g) = \inf_{x \in \mathcal{X}} g(x).$$

Assume that $f \in \ell^\infty(\mathcal{X})$ and that for ϵ sufficiently small, $\mathcal{S}^\epsilon$ is constant equal to $\mathcal{S}$. Define

$$\mathcal{B}_f = \{g \in \ell^\infty(\mathcal{X}) \, : \, |g(\mathcal{S})| = 1\}.$$

Then, ι is Hadamard differentiable tangentially to $\mathcal{B}_f$.

Proof. Theorem 3.1 gives that ι is Hadamard directionally differentiable at f with derivative

$$\iota'_f(g) = \lim_{\epsilon \searrow 0} \inf_{\mathcal{S}^\epsilon} g.$$

Because $\mathcal{S}^\epsilon$ is eventually constant and equal to $\mathcal{S}$, this reduces to

$$\iota'_f(g) = \inf_{\mathcal{S}} g$$

Because we are only interested in the derivative tangentially to $\mathcal{B}_f$, and any $g \in \mathcal{B}_f$ is constant on $\mathcal{S}$, this formula reduces to

$$\iota'_f(g) = g(\tilde{x}),$$

for any $\tilde{x} \in \mathcal{S}$. This is clearly linear. □

C. Normal confidence intervals

Here we for $\mathcal{A}, \mathcal{B}$ subsets of a metric space $(\mathcal{X}, d)$ we define

$$\mathbb{D}(\mathcal{A}, \mathcal{B}) = \sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a, b).$$

Note that $\mathbb{D}$ is not a metric, and indeed we have $\mathbb{D}(\mathcal{A}, \mathcal{B}) = 0$ if and only if $\mathcal{A} \subset \mathrm{cl}(\mathcal{B})$. Under very reasonable conditions it is possible for $\mathbb{D}(\hat{\mathcal{S}}_n, \mathcal{S}) \xrightarrow{\text{a.s.*}} 0$, see Shapiro et al. (2021, Thm 5.4) for some basic results.

The main result of this section depends on the following pair of lemmas.

Lemma C.1. *Let $\mathcal{X}$ be a metric space, for each $n \in \mathbb{N}_0$ take $\mathcal{X}_n \subseteq \mathcal{X}$ with $\mathbb{D}(\mathcal{X}_n, \mathcal{X}_0) \to 0$. For each $n \in \mathbb{N}$, let $\sigma_n, \sigma : \mathcal{X} \to \mathbb{R}$ with σ_n converging uniformly to σ. Assume σ is uniformly continuous and constant on $\mathcal{X}_0$ with value $\tilde{\sigma}$. Then for any $x_n \in \mathcal{X}_n$*

$$\sigma_n(x_n) \to \tilde{\sigma}.$$

Proof. The uniform convergence of σ_n to σ implies

$$\lim_{n \to \infty} (\sigma_n(x_n) - \sigma(x_n)) = 0.$$

For any n, let $\tilde{x}_n \in \mathcal{X}_0$ be chosen such that $d(x_n, \tilde{x}_n) \leq \frac{1}{n} + d(x_n, \mathcal{X}_0)$. We have

$$d(x_n, \tilde{x}_n) \leq \frac{1}{n} + \sup_{x \in \mathcal{X}_n} d(x, \mathcal{X}_0) = \frac{1}{n} + \mathbb{D}(\mathcal{X}_n, \mathcal{X}_0) \to 0, \text{ as } n \to \infty.$$

By the uniform continuity of σ we have

$$\lim_{n \to \infty} (\sigma(x_n) - \sigma(\tilde{x}_n)) = 0.$$

and hence

$$\lim_{n \to \infty} (\sigma_n(x_n) - \sigma(\tilde{x}_n)) = 0$$

The proof is complete by noting that $\sigma(\tilde{x}_n) \equiv \tilde{\sigma}$. □

Lemma C.2. *Let $\mathcal{X}$ be a metric space, and for each $n \in \mathbb{N}_0$ take $\mathcal{X}_n \subseteq \mathcal{X}$ with $\mathbb{D}(\mathcal{X}_n, \mathcal{X}_0) \to 0$. For each $n \in \mathbb{N}$ let $\sigma_n, \sigma : \mathcal{X} \to \mathbb{R}$ with σ_n converging uniformly to σ. Assume σ is continuous and constant on $\mathcal{X}_0$ with value $\tilde{\sigma}$, and $\mathcal{X}_0$ compact. Then for any $x_n \in \mathcal{X}_n$*

$$\sigma_n(x_n) \to \tilde{\sigma}.$$

Proof. We conduct the proof by contradiction. Suppose there is a sequence $(x_n \in \mathcal{X}_n)$ such that $\sigma_n(x_n)$ does not converge to $\tilde{\sigma}$. Then there is a subsequence of $(\sigma_n(x_n))$ that stays away from $\tilde{\sigma}$. Without loss of generality, assume that for some $\epsilon > 0$, $\sigma_n(x_n) \notin (\tilde{\sigma} - 2\epsilon, \tilde{\sigma} + 2\epsilon)$ for all n. From the uniform convergence of σ_n to σ, there exists some $N > 0$ such that

$$\sigma(x_n) \notin (\tilde{\sigma} - \epsilon, \tilde{\sigma} + \epsilon), \quad \text{for all } n \geq N. \tag{28}$$

For each n, choose $\tilde{x}_n \in \mathcal{X}_0$ such that

$$d(x_n, \tilde{x}_n) \leq \frac{1}{n} + \inf_{x \in \mathcal{X}_0} d(x_n, x) \leq \frac{1}{n} + \mathbb{D}(\mathcal{X}_n, \mathcal{X}_0) \to 0, \quad \text{as } n \to \infty.$$

Since $\mathcal{X}_0$ is compact, the sequence $(\tilde{x}_n)$ has at least one limit point $\tilde{x}_0 \in \mathcal{X}_0$ which is also a limit point of (x_n). Without loss of generality assume $x_n \to \tilde{x}_0$, which implies $\sigma(x_n) \to \sigma(\tilde{x}_0) = \tilde{\sigma}$, and contradicts (28). The proof is completed. □

We can now give results which allow for the use of standard normal quantiles to generate asymptotically correct CIs. We define the variance process of a random variable $G : \Omega \to \ell^\infty(\mathcal{X})$ as $\sigma^2 : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ via

$$\sigma^2(x) = \mathrm{Var}(G(x)).$$

Theorem C.3. *Make the following assumptions:*

- *H1 and H2 are satisfied with F Gaussian and has mean 0.*
- *$x \mapsto l(x, \cdot)$ is constant on $\mathcal{S}$.*
- *The variance process of F is continuous.*
- *$\mathbb{D}(\mathcal{S}_n, \mathcal{S}) \xrightarrow{\text{a.s.*}} 0$.*
- *The variance process of $\hat{f}_n$ (i.e., $\hat{\sigma}_n^2$) converges uniformly outer almost surely to the variance process of F.*

Then for any $x_n \in \mathcal{S}_n$,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1).$$

Proof. Because f is lower semi-continuous $\mathcal{S}$ is closed. Because $\mathcal{X}$ is compact $\mathcal{S}$ must then be as well. Because l is constant on $\mathcal{S}$, by Lemma A.1, F is as well and so addition σ^2 is as well.

The conditions on Lemma C.2 are satisfied so $\hat{\sigma}_n(x_n) \xrightarrow{\text{a.s*.}} \sigma_*$, where σ_* is the standard deviation of $F(x)$ for some $x \in S$. By Corollary E.2, because F has mean 0

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \mathrm{N}(0, \sigma_*^2).$$

Hence, by Slutsky's theorem,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1)$$

as required. □

Theorem C.4. *Make the following assumptions:*

- *$\mathcal{X}$ is a metric space.*

- *$B1$ and $B2$ hold with F Gaussian and has mean 0.*
- *For sufficiently small ϵ, $\mathcal{S}^\epsilon = \mathcal{S}$ and $x \mapsto l(x, \cdot)$ is constant on $\mathcal{S}$.*
- *The variance process of F is uniformly continuous.*
- *There is some $(\epsilon_n) \subset \mathbb{R}^+$ for which $\mathbb{D}(\mathcal{S}_n^{\epsilon_n}, \mathcal{S}) \xrightarrow{\text{a.s.*}} 0$.*
- *The variance process of $\hat{f}_n$ (i.e., $\hat{\sigma}_n^2$) converges uniformly outer almost surely to the variance process of f.*

Then for any $x_n \in \mathcal{S}_n^{\epsilon_n}$,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1).$$

Proof. By Lemma A.1, F is constant on $\mathcal{S}$ and so σ^2 is as well. The conditions on Lemma C.1 are then satisfied so $\hat{\sigma}_n(x_n) \xrightarrow{\text{a.s*.}} \sigma_*$, where σ_* is the standard deviation of $F(x)$ for some $x \in S$. By Corollary 4.4, because F has mean 0

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \mathrm{N}(0, \sigma_*^2).$$

Hence, by Slutsky's theorem,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1)$$

as required. □

Theorem C.5. *Make the following assumptions:*

- *$\mathcal{X}$ a metric space*
- *B1 and B2 hold with F Gaussian and has mean 0.*
- *For sufficiently small ϵ, $\mathcal{S}^\epsilon = \mathcal{S}$ and $x \mapsto l(x, \cdot)$ is constant on $\mathcal{S}$.*
- *The variance process of F is continuous and $\mathcal{S}$ is compact.*
- *There is some $(\epsilon_n) \subset \mathbb{R}^+$ for which $\mathbb{D}(\mathcal{S}_n^{\epsilon_n}, \mathcal{S}) \xrightarrow{\text{a.s.*}} 0$.*
- *The variance process of $\hat{f}_n$ (i.e., $\hat{\sigma}_n^2$) converges uniformly outer almost surely to the variance process of f.*

Then for any $x_n \in \mathcal{S}_n^{\epsilon_n}$,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1).$$

Proof. By Lemma A.1 F is constant on $\mathcal{S}$ and so σ^2 is as well. The conditions on lemma C.2 are then satisfied so $\hat{\sigma}_n(x_n) \xrightarrow{\text{a.s*.}} \sigma_*$, where σ_* is the standard deviation of $F(x)$ for some $x \in S$. By Corollary 4.4, because F has mean 0

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \mathrm{N}(0, \sigma_*^2).$$

Hence, by Slutsky's theorem,

$$\tau_n \left(\frac{\hat{\psi}_n - \psi^*}{\hat{\sigma}_n(x_n)} \right) \rightsquigarrow \mathrm{N}(0, 1)$$

as required. □

D. Neural network numerical experiment

In this section, we describe the procedure for generating data for the neural network experiment described in Section 7.2. Firstly, we note that generating a neural network with random weights typically resulted in a network which would label all data the same. Instead of weight randomization, we fit a neural network (by minimising cross entropy loss) to the function $\lceil \sin(2\pi x)/2 \rceil$, with 500 replicates of X, sampled from Unif$(0, 1)$. We then flipped the label of 30% of the output. This then became the label.

Write f_n^k to denote the kth sample of f_n^b. For m bootstrap samples we constructed $(1-\alpha)100\%$ CIs by first computing the smallest b for which

$$\frac{1}{m}\left|\left\{k \in [m] \,:\, |i_n(\sqrt{n}(\tilde{f}_n^k - \hat{f}))| \leq b\right\}\right| \geq 1-\alpha,$$

and then defining the CI for ψ^* by

$$\left(\hat{\psi}_n - \frac{b}{\sqrt{n}},\, \hat{\psi}_n + \frac{b}{\sqrt{n}}\right).$$

Throughout our experiment, the number of bootstrap resamples was taken to be $m = \max\{5, \lfloor n/5 \rfloor\}$, where $m = 5$ is only relevant for small sample sizes n.

The bootstrap procedure was replicated multiple times to generate a sample of these CIs. The percentage of the replications for which these CIs included the true value of ψ^* is then referred to as the coverage.

When considering more than 200 data points (i.e., $n \geq 200$) the total number of CIs constructed was 20. For $n < 200$ the bootstrap procedure was fast enough for the methods to run until convergence, in the sense that we continually produced additional CIs, in increments of 20, until the change in coverage was less than 0.01. For this reason, the trend in Figure 4 is stable after after $= 200$ samples. Based on our testing, for $n > 100$ the coverage converged after a very small number of iterations.

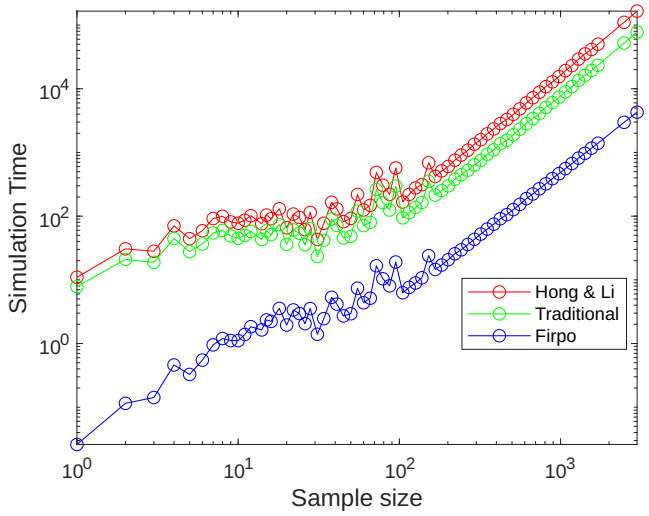

Figure 4. Time taken to estimate coverage probabilities for samples of size n.

D.1. Neural network accuracy

Consider a binary classification problem with the unknown classifier $f^* : \mathcal{X} \to \{\pm 1\}$. Define the classification model via ρ, a joint probability distribution on $\mathcal{X} \times \{\pm 1\}$ with marginal distribution $\rho_{\mathcal{X}}$ on $\mathcal{X}$. We write $(X, Y) \sim \rho$ to mean that ρ characterizes the data generating process of (X, Y). Define the conditional distribution $\rho(y|x)$ via the relationship

$$Y = \varepsilon f^*(X),$$

where ε is independent of X, $\mathbb{P}(\varepsilon = 1) = p$, and $\mathbb{P}(\varepsilon = -1) = 1 - p$ for some $0 \leq p \leq 1$. We assume that f^* is a function in some family $\mathcal{F}$. The following lemma is known in the literature (see, e.g., Bartlett et al., 2006) and we include its proof for completeness.

Lemma D.1. *If $p \geq 1/2$, we have*

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{P}(Y \neq f(X)).$$

Proof. For any $f \in \mathcal{F}$, write

$$\alpha = \alpha_f(x) = \mathbb{P}(f^*(X) \neq f(X) | X = x).$$

Then $0 \leq \alpha \leq 1$. Note that since ε and X are independent, for a.e. $x \in \mathcal{X}$,

$$\begin{aligned}
&\mathbb{P}(\varepsilon f^*(X) \neq f(X) | X = x) \\
=&\mathbb{P}(f^*(X) \neq f(X) | X = x)\mathbb{P}(\varepsilon = 1) + \mathbb{P}(f^*(X) = f(X) | X = x)\mathbb{P}(\varepsilon = -1) \\
=&\alpha p + (1 - \alpha)(1 - p) = (2p - 1)\alpha + 1 - p.
\end{aligned}$$

When $p \geq \frac{1}{2}$, this implies

$$\begin{aligned}
\mathbb{P}(Y \neq f(X)) &= \int_{\mathcal{X}} \mathbb{P}(\varepsilon f^*(X) \neq f(X) | X = x) d\rho_{\mathcal{X}}(x) \\
&\geq \int_{\mathcal{X}} (1 - p) d\rho_{\mathcal{X}}(x) = 1 - p,
\end{aligned}$$

where the equality is achieved at

$$\alpha = \alpha_{f^*}(x) = 0.$$

This completes the proof. □

E. Continuous limit

It is possible to obtain the outcomes of Fact 4.1 under a more general hypothesis.

Corollary E.1. *Assume f is lower semi-continuous and $H2$ holds with the sample paths of F additionally lower semi-continuous. Then*

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \inf_{x \in \mathcal{S}} F(x). \tag{29}$$

If additional f is continuous and $\hat{f}_n$ is almost surely continuous, then

$$\hat{\psi}_n = \inf_{x \in \mathcal{S}} \hat{f}_n(x) + o_{\mathbb{P}*}(\tau_n^{-1}). \tag{30}$$

Proof. The conditions on theorem 4.2 are satisfied and so

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} F(x) \tag{31}$$

$$\hat{\psi}_n = \lim_{\epsilon \searrow 0} \inf_{x \in \mathcal{S}^\epsilon} [\hat{f}_n(x) - f(x) + \psi^*] + o_{\mathbb{P}*}(\tau_n^{-1}) \tag{32}$$

Because F and f are lower semi-continuous, by Lemma A.9, Equation (31) becomes

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \inf_{x \in \mathcal{S}} F(x).$$

This is exactly Equation (29).

Now assume f is continuous and $\hat{f}_n$ is continuous almost surely. We can modify $\hat{f}_n$ on a null so that it is continuous. We then have $\hat{f}_n - f + \psi^*$ is continuous and so in particular it is lower semi-continuous. By Lemma A.9 Equation (32) becomes

$$\begin{aligned}\hat{\psi}_n &= \inf_{x\in S}[\hat{f}_n(x) - f(x) + \psi^*] + o_{\mathbb{P}^*}(\tau_n^{-1}) \\ &= \inf_{x\in S}[\hat{f}_n(x) - \psi^* + \psi^*] + o_{\mathbb{P}^*}(\tau_n^{-1}) \\ &= \inf_{x\in S}\hat{f}_n(x) + o_{\mathbb{P}^*}(\tau_n^{-1}).\end{aligned}$$

This is exactly Equation (30). □

When we have the additional information that the limiting process is lower semi-continuous it is easier to generate conditions that imply the limiting distribution is Gaussian.

Corollary E.2. *Assume $H1$ and $H2$. If $x \mapsto l(x,\cdot)$ is constant on $\mathcal{S}$ then*

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow F(x^*),$$

for any $x^* \in \mathcal{S}$. If F is a Gaussian process then $F(x^*)$ is Gaussian.

Proof. Corollary E.1 gives that

$$\tau_n(\hat{\psi}_n - \psi^*) \rightsquigarrow \inf_{x\in S} F(x).$$

Because l is constant on $\mathcal{S}$, $\hat{f}_n$ and f is as well. Because $\tau_n(\hat{f}_n - f) \rightsquigarrow F$, by Lemma A.1 F is constant on $\mathcal{S}$. Hence for any $x^* \in \mathcal{S}$

$$\inf_{x\in\mathcal{S}} F(x) = F(x^*).$$

□