On the Asymptotic Distribution of the Minimum Empirical Risk

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

[1]University of Queensland [2]La Trobe University [3]Kyushu University

Problem setup

Aim is to find

$$\psi^* := \inf_{x\in\mathcal{X}} \mathbb{E}(l(x,Z)),$$

where Z is a random variable representing the data and $l(\cdot, Z)$ is a loss function indexed by parameter $x \in \mathcal{X}$. Approximate this problem by solving

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

where the data Z_i has the same distribution as Z.

On the lack of a test set

This is a different paradigm to the usual ML paradigm. ψ^* is a measure of how well a model class could possibly work on a given problem. We treat $\hat{\psi}_n$ as an estimate of ψ^* and hence as an estimate of the optimal performance.

Empirical process background

Let

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

For any $\tau_n \to \infty$, let $F_n : \Omega \times \mathcal{H} \to \mathbb{R}$ be given by

$$F_n(\omega, h) = \tau_n \left\{ \frac{1}{n}\sum_{i=1}^{n} h(Z_i(\omega)) - \mathbb{E}h(Z) \right\}. \tag{2}$$

We will often assume that there is a random variable F such that

$$F_n \rightsquigarrow F$$

where $\rightsquigarrow$ denotes weak convergence (convergence in distribution). That is, we will often assume a type of central limit theorem on functions.

Main result

Define

$$\mathcal{S}^\epsilon = \{x \in \mathcal{X} : \mathbb{E}[l(x,Z)] \le \psi^* + \epsilon\}.$$

If

1. l is bounded
2. $F_n \rightsquigarrow F$ for some bounded Borel measurable F

then

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

and

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

Sufficient conditions

The most common way to generate $F_n \rightsquigarrow F$ is to assume $\mathcal{H}$ is Donsker. This occurs when, $\tau_n = \sqrt{n}$, (Z_i) are iid, $\sup_{h\in\mathcal{H}}|h(z) - \mathbb{E}_Z h| < \infty$, for each $z \in \mathcal{Z}$, and F is a zero-mean Gaussian process with covariance

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

The idea is that if a class of functions is not very 'complicated' then it will be Donsker. Sufficient conditions are known for

1. convex function classes [6, Thm 2.7.14]
2. monotone function classes [6, Thm 2.7.9]
3. function classes with Holder-derivatives [6, Cor 2.7.2, Cor 2.7.3]

When the data is iid, all binary classification, feed forward neural networks are Donsker. Similar results are known for non-iid data.

For more information on Donsker classes see [1] and [6, Ch. 2].

Applications to model selection

Let $\mathcal{X}_1, \mathcal{X}_2$ be a pair of parameter spaces defining corresponding model spaces $\mathcal{H}_1, \mathcal{H}_2$ analogous to (1). Define F_n^1, F_n^2 as in (2) but with $\mathcal{H}$ replaced by $\mathcal{H}_1$ and $\mathcal{H}_2$.

Let

$$\psi_k^* = \inf_{x_k\in\mathcal{X}_k} \mathbb{E}[l(x_k, Z)]$$

be the minimum expected risk obtained by models in $\mathcal{H}_k$ and let

$$\hat{\psi}_{k,n} = \inf_{x_k\in\mathcal{X}_k}\frac{1}{n}\sum_{i=1}^{n} l(x_k, Z_i).$$

We aim to find $k^* = \arg\min_{k\in\{1,2\}} \psi_k^*$ That is, we want to find the model class that can possibly perform the best on the problem.

References

[1] Richard M Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[2] Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86:377–412, 2019.

[3] Sergio Firpo, Antonio F Galvao, and Thomas Parker. Uniform inference for value functions. *Journal of Econometrics*, 235:1680–1699, 2023.

[4] Han Hong and Jessie Li. The numerical delta method. *Journal of Econometrics*, 206:379–394, 2018.

[5] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

[6] AW van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 2023.

Model selection result

If

1. l is bounded
2. $F_n^1 \rightsquigarrow F^1$ and $F_n^2 \rightsquigarrow F^2$ with F^1, F^2 bounded and Borel measurable.

Assuming $\psi_0^* = \psi_1^*$ we have

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

$$F^* = \lim_{\epsilon\searrow 0}\inf_{x\in\mathcal{S}_1(\epsilon)} F(x) - \lim_{\epsilon\searrow 0}\inf_{x\in\mathcal{S}_0(\epsilon)} F(x),$$

and

$$\mathcal{S}_i(\epsilon) = \{x \in \mathcal{X}_i : \mathbb{E}[l(x,Z)] \le \psi_i^* + \epsilon\}.$$

Then for any $\alpha \in [0,1]$,

$$\limsup_{n\to\infty} \mathbb{P}^*\left(\hat{\psi}_{n,1} \le \hat{\psi}_{n,0} + \frac{c_\alpha}{\tau_n}\right) \le \alpha, \text{ where}$$

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

Incremental model space

We now assume that the model space can depend on the number of data points n. Let $\mathcal{X}_n$ such a (non-random) parameter space. For simplicity assume $\emptyset \ne \mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots$ and let $\mathcal{X} = \bigcup_{n=1}^{\infty}\mathcal{X}_n$.

Define

$$\hat{\phi}_n = \inf_{x\in\mathcal{X}_n}\frac{1}{n}\sum_{i=1}^{n} l(x, Z_i), \text{ and } \phi_n^* = \inf_{x\in\mathcal{X}_n}\mathbb{E}[l(x,Z)].$$

There are now 2 quantities of interest: ψ^* and ϕ_n^*.

Main incremental result

If

1. l is bounded
2. $F_n \rightsquigarrow F$ for some bounded and Borel measurable 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

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

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

Further,

$$\tau_n(\phi_n^* - \psi^*) \to 0$$

Difficulties in approximating weak limits

In order for these results to be of any use, we need a way to approximate F. With this we can generate confidences intervals for ψ^*. There are some difficulties

1. Direct approximation of the limiting process using sample means and variances does not yield correct results (cf. [5, p. 19]).
2. Only under very restrictive conditions does the bootstrap work [2, Thm. 3.1] (for example when S^ϵ is constant for ϵ small enough).

We then need new procedures.

Procedure to approximate weak limits

Rather than bootstrapping the data, then performing the minimization, we instead bootstrap some other function of the data. We consider 2 such functions

1. From [2] and [4],
$$\hat{\iota}_{s_n,n}(\eta) = s_n^{-1}\left(\inf_{\mathcal{X}}(\hat{f}_n + s_n\eta) - \hat{\psi}_n\right).$$
2. Modified from [3],
$$\tilde{\iota}_{s_n,n}(\eta) = \inf_{x\in\mathcal{S}_n^{s_n}}(\eta).$$

where

$$S_n^\epsilon = \left\{x \in \mathcal{X} : \frac{1}{n}\sum_{i=1}^{n} l(x, Z_i) \le \hat{\psi}_n + \epsilon\right\}.$$

Consider a bootstrapping procedure which draws weights W_i corresponding the number of times point Z_i was resampled. Let

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

be the bootstrapped empirical risk and let

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

be the standard empirical risk.

Then provided $\mathcal{H}$ is Donsker, $s_n \to 0$ with $s_n\tau_n \to \infty$ and under mild conditions on the boostrapping procedure, quantiles of

$$\hat{\iota}_{s_n,n}(\tau_n(f_n^b - \hat{f}_n)) \qquad \tilde{\iota}_{t_n,n}(\tau_n(f_n^b - \hat{f}_n))$$

are guaranteed to tend to the quantiles of F asymptotically.