Exchangeable bootstrap of empirical process suprema: concentration properties and an application to two-sample hypothesis testing

Guillaume Maillard

ENSAI

May 19, 2025

Exchangeable weighted bootstrap

Let $Z = h(P_n, P)$, for ex.

$$h(P_n, P) = |\hat{\theta}_n - \theta| = |\theta(P_n) - \theta(P)|$$

The *bootstrap heuristic* states that

$$\mathcal{L}(h(P_n, P)) \approx \mathcal{L}\left(c_W h(P_n^W, P_n)|P_n\right)$$

where $\mathcal{L}(Y), \mathcal{L}(Y|X) =$ law of Y, resp. law of Y knowing X,

$$P_n^W = \frac{1}{n}\sum_{i=1}^{n} W_i \delta_{X_i}$$

W exchangeable, independent from P_n, $\sum_{i=1}^{n} W_i = n$.

Bootstrapped supremum

Let $\xi_i = W_i - 1$ be exch. s.t. $\sum_{i=1}^n \xi_i = 0$,

$$g(X, \xi) = \sup_{f \in \mathcal{F}} \left\{ n \left(P_n^W - P_n \right) (f) \right\}$$
$$= \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \xi_i f(X_i) \right\}$$

where $X_1, \ldots, X_n$ are independent, $\mathcal{F}$ a function class.

Bootstrapped supremum: the point

By the *bootstrap heuristic*, $c_W g(X,\xi)|X$ estimates

$$\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]$$

in distribution, in particular

$$\overline{g}(X) = \mathbb{E}\left[g(X,\xi)|X\right] \approx \frac{1}{c_W}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]\right\}\right].$$

$\implies$ $\overline{g}(X)$ should concentrate.

Applications

- Two-sample MMD tests (power analysis)
- Confidence regions for the mean in high dimensions
- Model selection (penalties)

Justifying the bootstrap heuristic

Summary of known results:

- Asymptotically true for Donsker $\mathcal{F}$ [6]
- Weak convergence rate for VC $\mathcal{F}$ [2]
- Bounds [4] comparing $\mathbb{E}[g(X,\xi)]$ with

$$M_n = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]\right\}\right]$$

Open question: non-asymptotic *concentration* of $\overline{g}(X)/g(X,\xi)$?

Rademacher complexities

Remark: if ξ is replaced by i.i.d Rademacher ε, g with

$$h(X,\varepsilon) = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i)$$

the equivalent of $\overline{g}$ is

$$\mathrm{Rad}\left(\mathbf{X}\right) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i)\middle| X_1,\ldots,X_n\right]$$

the *conditional Rademacher complexity*, known to concentrate.

A flaw of Rademacher complexity

ε is not centered, so

$$\sum_{i=1}^{n} \varepsilon_i f(X_i) \neq \sum_{i=1}^{n} \varepsilon_i (f(X_i) - \mathbb{E}[f(X_i)])$$

$\implies \mathbb{E}[f(X_i)^2]$ instead of $\mathrm{Var}(f(X_i))$,

No translation invariance.

Self-bounding function

Definition

Let $\mathcal{X}$ be a set and $\phi : \mathcal{X}^n \to \mathbb{R}_+$. ϕ is $(a, b)-$self-bounding if there exists $\phi_i : \mathcal{X}^{n-1} \to \mathbb{R}_+$ $(1 \leq i \leq n)$ s.t. for all $x \in \mathcal{X}$,

- $0 \leq \phi(x) - \phi_i(x_{(i)}) \leq 1$ *for all i*
- $\sum_{i=1}^n \phi(x) - \phi_i(x_{(i)}) \leq af(x) + b$

where $x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Concentration of self-bounding functions

Theorem (BLM 6.20,6.21 [1])

Let $X_1, \dots, X_n$ be independent, valued in $\mathcal{X}$. Let $\phi : \mathcal{X}^n \to \mathbb{R}$ be $(a, b)-$self bounding with $a \geq \frac{1}{3}$. Let $Z = \phi(X_1, \dots, X_n)$, then

$$Z \leq \mathbb{E}[Z] + \sqrt{2t\left(a\mathbb{E}[Z] + b\right)} + t\left(a - \frac{1}{3}\right)$$

$$Z \geq \mathbb{E}[Z] - \sqrt{2t\left(a\mathbb{E}[Z] + b\right)}$$

each with probability at least $1 - e^{-t}$.

Concentration of $\overline{g}$

Theorem (GM,AS)

Let $\kappa = \mathbb{E}[|\xi_1|] < +\infty$. Let $X_1, \ldots, X_n$ be independent, valued in $\mathcal{X}$. Let all $f \in \mathcal{F}$ take values in $[-1;1]$. The function

$$\frac{\overline{g}}{\kappa\left(4 - \frac{3}{n}\right)} : \mathcal{X}^n \to \mathbb{R}_+$$

is $(a, 0)-$self-bounding, with $a = 3 - \frac{2}{n}$.

Remark: Was known for Rad [1, Example 3.12].

Proof idea

Let $J \in \{1, \ldots, n\}$ be uniform, let $\tau_{i,J}$ = transposition of i, J.

$$
\begin{aligned}
\overline{g}(x) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{k=1}^{n} \xi_{\tau_{iJ}(k)} f(x_k)\right] \\
&\geq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{k=1}^{n} \mathbb{E}\left[\xi_{\tau_{iJ}(k)} | \xi\right] f(x_k)\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \sum_{k \neq i}^{n} \xi_k f(x_k) + \frac{1}{n} \sum_{j \neq i}^{n} (\xi_i - \xi_j) f(x_j) \right\}\right] \\
&= \overline{g}_i(x_{(i)}).
\end{aligned}
$$

The two-sample testing problem

Let $(X_1, \ldots, X_n), (Y_1, \ldots, Y_m)$ two independent samples from P and Q.

- $H_0 : P = Q$
- $H_1 : P \neq Q$

Pbm: test H_0 vs H_1. *Non-parametric*.

The MMD test

Let $\mathcal{F}$ be a symmetric function class, let

$$T = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{m} \sum_{j=1}^{m} f(Y_j) \right\}.$$

The (general) *MMD test* uses T as test statistic. T estimates

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \int f dP - \int f dQ$$

(*integral probability metric*).

Examples:

- Kolmogorov-Smirnov: $\mathcal{F} = \pm$ indicators of half-lines
- Kernel MMD: $\mathcal{F} =$ Unit ball of RKHS
- Wasserstein 1
- Others (dimension reduction ...)

Threshold using the bootstrap

Let $Z = (X_1, \dots, X_n, Y_1, \dots, Y_m)$. Let $\sigma \in \mathfrak{S}_{n+m}$ a uniform random permutation and

$$\xi = w_\sigma, \;\; w = \Big(\underbrace{\frac{1}{n}, \dots, \frac{1}{n}}_{n \;\; times}, \underbrace{-\frac{1}{m}, \dots, -\frac{1}{m}}_{m \;\; times}\Big).$$

Let $(\xi^{(b)})_{b=1,\dots,B}$ i.i.d copies of ξ, $\hat{q}_B^\alpha$ = empirical $1-\alpha$ quantile of

$$g(Z,\xi), g\left(Z, \xi^{(1)}\right), \dots, g\left(Z, \xi^{(B)}\right)$$

The test which rejects when $T > \hat{q}_B^\alpha$ has level α [5].

How much power?

For $\mathcal{F}$ valued in $[-1;1]$, define

$$\sigma^2_{\mathcal{F}}(R) = \sup_{f \in \mathcal{F}} \left\{ \int f^2 dR - \left(\int f dR \right)^2 \right\}$$

$$M_k(R) = \mathbb{E}_{Z \sim R^{\otimes k}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{k} (f(X_i) - \int f dR) \right].$$

We can show that when $n = m$, if

$$d_{\mathcal{F}}(P, Q) \gg \max \left(\frac{\sigma_{\mathcal{F}}(P)}{\sqrt{n}}, \frac{1}{n} M_n(P), \frac{1}{n} M_n(Q), \frac{1}{n} \right),$$

then the MMD test rejects H_0 with high probability.

Proof

- $T \geq d_{\mathcal{F}}(P, Q) - \varepsilon(n, m)$ by Bernstein's inequality
- Bound $\mathbb{E}[\overline{g}(Z)]$ (Jensen's inequality)
- Concentration of $\overline{g}(Z)$.
- Concentration of $g(z, \xi)$ around $\overline{g}(z)$ (Tolstikhin-Talagrand inequality [3, Theorem 5.6]) - var. parameter $v_+(z)$
- $\implies$ upper bound on $\hat{q}_B^\alpha$ given Z
- Concentration of $v_+(Z)$ (classical arguments).

Bibliography I

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.
Concentration Inequalities: A Nonasymptotic Theory of Independence.
Oxford University Press, 02 2013.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato.
Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings.
Stochastic Processes and their Applications, 126(12):3632–3651, 2016.
In Memoriam: Evarist Giné.

Fang Han.
An introduction to permutation processes (version 0.5), 2024.

Bibliography II

Qiyang Han and Jon A. Wellner.
Convergence rates of least squares regression estimators with heavy-tailed errors.
The Annals of Statistics, 47(4):2286 – 2319, 2019.

Joseph P. Romano and Michael Wolf.
Exact and approximate stepdown methods for multiple hypothesis testing.
Journal of the American Statistical Association, 100(469):94–108, 2005.

A. W. van der Vaart and Jon A. Wellner.
Weak Convergence and Empirical Processes: With Applications to Statistics.
Springer International Publishing, 2023.

Concentration of $g(x, \xi)$?

Theorem

Let $\xi = w_\sigma$ *where* $w \in [a; b]^n$, $\sum_{i=1}^n w_i = 0$, $\sigma \in \mathfrak{S}_n$ *random.*
With prob. $\geq 1 - e^{-t}$,

$$g(x, \xi) \leq \overline{g}(x) + 2(b - a)\sqrt{t L_n v_+(x)}$$

where

$$L_n = 2.32 \log n + \frac{4n}{n-1} + \frac{2}{n}$$

$$v_+(x) = \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \left(f(x_i) - \frac{1}{n} \sum_{k=1}^n f(x_k) \right)^2 \right\}$$

Comments

- Recovers Tolstikhin-Talagrand inequality [3] with $8 \to L_n$
- The trick $\xi \to \xi_\sigma | \xi$ allows for general ξ
- Proof: *Method of exchangeable pairs* (S. Chatterjee)
- aka Stein's method for concentration inequalities
- Here the pair is $(\sigma, \sigma \circ \tau_{I,J})$ (I, J uniform)