Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes

Junhyung Park, Krikamol Muandet
Max Plank Institute for Intelligent Systems, Tübingen
CISPA Helmholtz Center for Information Security, Saarbrücken

ALT 2023

Summary

Motivation

Contributions

Vector-valued Learning Problems

- There is a growing literature on learning vector-valued functions:
    - multi-task or multi-output learning;
    - functional response models;
    - kernel conditional mean embeddings;
    - structured prediction;
    - $\vdots$

Vector-valued Learning Problems

- There is a growing literature on learning vector-valued functions:
  - multi-task or multi-output learning;
  - functional response models;
  - kernel conditional mean embeddings;
  - structured prediction;
  - $\vdots$
- [Micchelli and Pontil, 2005], [Alvarez, 2011] – Algorithm for learning vector-valued functions using operator-valued kernels.

Vector-valued Learning Problems

- There is a growing literature on learning vector-valued functions:
  - multi-task or multi-output learning;
  - functional response models;
  - kernel conditional mean embeddings;
  - structured prediction;
  - ⋮
- [Micchelli and Pontil, 2005], [Alvarez, 2011] – Algorithm for learning vector-valued functions using operator-valued kernels.
- [Caponnetto and de Vito, 2007], [Ciliberto et al., 2020], [Cabannes et al., 2021], [Singh et al., 2019] – Learning rates using integral operator techniques for kernel methods.

Vector-valued Learning Problems

- There is a growing literature on learning vector-valued functions:
  - multi-task or multi-output learning;
  - functional response models;
  - kernel conditional mean embeddings;
  - structured prediction;
  - $\vdots$
- [Micchelli and Pontil, 2005], [Alvarez, 2011] – Algorithm for learning vector-valued functions using operator-valued kernels.
- [Caponnetto and de Vito, 2007], [Ciliberto et al., 2020], [Cabannes et al., 2021], [Singh et al., 2019] – Learning rates using integral operator techniques for kernel methods.
- [Yousefi et al., 2018], [Li et al., 2019] – Vector-valued extension of Rademacher complexities.

Empirical Process Theory for Vector-Valued Functions

- Empirical process theory is concerned with the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, and the stochastic process of the form $\{P_n f - Pf : f \in \mathcal{F}\}$.

Empirical Process Theory for Vector-Valued Functions

- Empirical process theory is concerned with the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, and the stochastic process of the form $\{P_n f - Pf : f \in \mathcal{F}\}$.
- For example, we're interested in questions such as whether

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{P} 0.$$

Empirical Process Theory for Vector-Valued Functions

- Empirical process theory is concerned with the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and the stochastic process of the form $\{P_n f - Pf : f \in \mathcal{F}\}$.
- For example, we're interested in questions such as whether

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{P} 0.$$

- For a class $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a Hilbert space, we are interested in questions such as whether

$$\sup_{g\in\mathcal{G}} \left\| \frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}[g(X)] \right\|_{\mathcal{Y}} \xrightarrow{P} 0.$$

Metric Entropy

- Suppose $(\mathcal{Z}, \rho)$ is a metric space. For any $\delta > 0$, the δ-*covering number* of $(\mathcal{Z}, \rho)$, denoted by $N(\delta, \mathcal{Z}, \rho)$, is the minimum number of balls of radius δ with centres in $\mathcal{Z}$ required to cover $\mathcal{Z}$. We define the δ-*entropy* as $H(\delta, \mathcal{Z}, \rho) = \log N(\delta, \mathcal{Z}, \rho)$.

Metric Entropy – Complexity of Function Classes

- For real-valued functions, the following classes of functions have been identified to have good bounds on their metric entropies:
  - Finite-dimensional classes;
  - Classes of smooth functions;
  - Classes of functions of bounded variation;
  - Classes of concave functions;
  - $\vdots$

Metric Entropy – Complexity of Function Classes

- For real-valued functions, the following classes of functions have been identified to have good bounds on their metric entropies:
  - Finite-dimensional classes;
  - Classes of smooth functions;
  - Classes of functions of bounded variation;
  - Classes of concave functions;
  - ⋮
- Other measures of complexity also exist, such as the VC dimension and entropy with bracketing.

Metric Entropy – Complexity of Function Classes

- For real-valued functions, the following classes of functions have been identified to have good bounds on their metric entropies:
    - Finite-dimensional classes;
    - Classes of smooth functions;
    - Classes of functions of bounded variation;
    - Classes of concave functions;
    - ⋮
- Other measures of complexity also exist, such as the VC dimension and entropy with bracketing.
- For vector-valued function classes, investigations on their metric entropies have not received much attention.

Entropy of Vector-Valued Function Classes

Challenges

- If $\mathcal{Y}$ is infinite-dimensional, seemingly trivial function classes such as the class of constant functions onto the unit ball,

$$\mathcal{G} = \left\{ g(x) = y \text{ for all } x \in \mathcal{X} : y \in \mathcal{Y}, \|y\|_{\mathcal{Y}} \leq 1 \right\}$$

have infinite entropy.

Entropy of Vector-Valued Function Classes

Challenges

- If $\mathcal{Y}$ is infinite-dimensional, seemingly trivial function classes such as the class of constant functions onto the unit ball,

$$\mathcal{G} = \left\{ g(x) = y \text{ for all } x \in \mathcal{X} : y \in \mathcal{Y}, \|y\|_{\mathcal{Y}} \leq 1 \right\}$$

  have infinite entropy.
- To have any chance, it is clear that the output range has to be restricted in more than the norm sense.

Set-Up

- Let $\mathcal{X}$ be any input domain, and let the output space $\mathcal{Y}$ be a *(not necessarily finite-dimensional) Hilbert space.*

Set-Up

- Let $\mathcal{X}$ be any input domain, and let the output space $\mathcal{Y}$ be a *(not necessarily finite-dimensional) Hilbert space.*
- Let $\mathcal{G}$ be a class of Bochner-integrable functions $g : \mathcal{X} \to \mathcal{Y}$.

Set-Up

- Let $\mathcal{X}$ be any input domain, and let the output space $\mathcal{Y}$ be a *(not necessarily finite-dimensional) Hilbert space.*
- Let $\mathcal{G}$ be a class of Bochner-integrable functions $g : \mathcal{X} \to \mathcal{Y}$.
- Let X be a random variable taking values in $\mathcal{X}$, and $X_1, X_2, \ldots$ i.i.d. copies of it.

Fractal Dimensions

- Let E be a subset of $(\mathcal{Z}, \rho)$. The *upper box-counting dimension* of E is

$$\tau_{\text{box}}(E) := \limsup_{\delta \to 0} \frac{H(\delta, E, \rho)}{-\log \delta}.$$

Fractal Dimensions

- Let E be a subset of $(\mathcal{Z}, \rho)$. The *upper box-counting dimension* of E is

$$\tau_{\mathrm{box}}(E) := \limsup_{\delta \to 0} \frac{H(\delta, E, \rho)}{-\log \delta}.$$

- A subset E of $(\mathcal{Z}, \rho)$ is said to be (M, τ)-*homogeneous* (or simply *homogeneous*) if the intersection of E with any closed ball of radius R can be covered by at most $M\left(\frac{R}{r}\right)^{\tau}$ closed balls of smaller radius r.

Main Results

- Let $m, d \in \mathbb{N}$, $B \subset \mathcal{Y}$ and $\mathcal{X}$ the unit cube in $\mathbb{R}^d$.
- Let $\mathcal{G}_B^m$ be the set of m-times differentiable functions $g : \mathcal{X} \to \mathcal{Y}$ such that:
  - partial derivatives $D^p g : \mathcal{X} \to \mathcal{Y}$ of orders $[p] \leq m$ exist everywhere on the interior of $\mathcal{X}$, and
  - $D^p g(x) \in B$ for all $x \in \mathcal{X}$ and $[p] \leq m$.

Main Results

- Let $m, d \in \mathbb{N}$, $B \subset \mathcal{Y}$ and $\mathcal{X}$ the unit cube in $\mathbb{R}^d$.
- Let $\mathcal{G}_B^m$ be the set of m-times differentiable functions $g : \mathcal{X} \to \mathcal{Y}$ such that:
    - partial derivatives $D^p g : \mathcal{X} \to \mathcal{Y}$ of orders $[p] \leq m$ exist everywhere on the interior of $\mathcal{X}$, and
    - $D^p g(x) \in B$ for all $x \in \mathcal{X}$ and $[p] \leq m$.

Theorem 4

Let $B \subset \mathcal{Y}$ be totally bounded and (M, τ_{asd})-homogeneous. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d, M and τ_{asd} such that

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}}.$$

Main Results

Theorem 5

Let B be a subset of $\mathcal{Y}$ with finite upper box-counting dimension τ_{box}. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d and τ_{box} such that

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right).$$

Main Results

Theorem 5

Let B be a subset of $\mathcal{Y}$ with finite upper box-counting dimension τ_{box}. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d and τ_{box} such that

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right).$$

Theorem 6

Let B be a subset of $\mathcal{Y}$ with $N(\epsilon, B, \|\cdot\|_{\mathcal{Y}}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $M, \tau_{\text{exp}} > 0$. Then for sufficiently small $\delta > 0$, there is some constant K depending on K_B, m, d, M and τ_{exp} such that

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\left(\frac{d}{m}+\tau_{\text{exp}}\right)}.$$

Applications

- Uniform law of large numbers of $\mathcal{G}_B^m$ for B satisfying any of the previous theorems.

Applications

- Uniform law of large numbers of $\mathcal{G}_B^m$ for B satisfying any of the previous theorems.
- Regression with smooth functions, where the output space itself consists of smooth (real-valued) functions, or any other real-valued function classes with appropriately bounded entropies.

Applications

- Uniform law of large numbers of $\mathcal{G}_B^m$ for B satisfying any of the previous theorems.
- Regression with smooth functions, where the output space itself consists of smooth (real-valued) functions, or any other real-valued function classes with appropriately bounded entropies.
- Kernel conditional mean embeddings, where the outputs consist of functions taking values in an RKHS.

Summary

- Despite the growth of literature on vector-valued learning, empirical process theory only exists for real-valued functions.

Summary

- Despite the growth of literature on vector-valued learning, empirical process theory only exists for real-valued functions.
- Our work attempts to make some first steps in developing empirical process theory for vector-valued functions.

Summary

- Despite the growth of literature on vector-valued learning, empirical process theory only exists for real-valued functions.
- Our work attempts to make some first steps in developing empirical process theory for vector-valued functions.
- Future directions:
  - entropy of function classes other than those of smooth functions;
  - infinite-dimensional input spaces;
  - uniform central limit theorems;
  - lower bounds... and many more.