Consistency of General Variational Learning Schemes in Banach spaces

S. Salzo[1]

[1] DIBRIS
Università di Genova

Workshop on Algorithms and Dynamics
for Games and Optimization
Playa Blanca, Tongoy, Chile, October 14th-18th, 2013

Joint work with: P.L. Combettes and S. Villa

Outline

The Learning Problem

We are given:

- $\mathcal{X}$ input and $\mathcal{Y}$ output spaces, $\mathcal{Y} \subset \mathrm{Y}$, Y Banach space. (X, Y) is a random variable with value in $\mathcal{X} \times \mathcal{Y}$ and distribution P.
- $(X_i, Y_i)_{i \in \mathbb{N}}$ is a sequence of i.i.d. random variables taking value in $\mathcal{X} \times \mathcal{Y}$ with common distribution P and $Z_n = (X_i, Y_i)_{1 \leq i \leq n}$ is the n-truncated sequence.
- x_i, y_i denote corresponding realizations of the random variables X_i, Y_i. A realizations $z_n = (x_i, y_i)_{1 \leq i \leq n}$ of the random variable Z_n is called a **training set**.
- a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathrm{Y} \to [0, +\infty[$ (least square, p-loss, Hinge loss, logistic loss, Huber's loss, etc.;).

The Learning Problem

The goal is to find a function $f : \mathcal{X} \to \mathsf{Y}$, minimizing the **risk** $R(f) = \mathsf{E}_{\mathsf{P}}\ell(X, Y, f(X))$, that is

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x))\,\mathrm{d}P(x, y)$$

over the space of all measurable functions $\mathcal{M}(\mathcal{X}, \mathsf{Y})$.

The Learning Problem

The goal is to find a function $f : \mathcal{X} \to \mathrm{Y}$, minimizing the **risk** $R(f) = \mathrm{E}_{\mathrm{P}}\ell(X, Y, f(X))$, that is

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x))\,\mathrm{d}P(x, y)$$

over the space of all measurable functions $\mathcal{M}(\mathcal{X}, \mathrm{Y})$. Without any knowledge of P, but using only training sets z_n.

The Learning Problem

The goal is to find a function $f : \mathcal{X} \to \mathrm{Y}$, minimizing the **risk** $R(f) = \mathrm{E}_{\mathrm{P}}\ell(X, Y, f(X))$, that is

$$R(f) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(x, y, f(x))\, dP(y|x) \right) dP_X(x)$$

over the space of all measurable functions $\mathcal{M}(\mathcal{X}, \mathrm{Y})$. Without any knowledge of P, but using only training sets z_n.

The Learning Problem

The goal is to find a function $f : \mathcal{X} \to \mathrm{Y}$, minimizing the **risk** $R(f) = \mathrm{E}_{\mathrm{P}}\ell(X, Y, f(X))$, that is

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \|y - f(x)\|_{\mathrm{Y}}^{p} \, \mathrm{d}P(x, y)$$

over the space of all measurable functions $\mathcal{M}(\mathcal{X}, \mathrm{Y})$. Without any knowledge of P, but using only training sets z_n.

The Learning Problem

The goal is to find a function $f : \mathcal{X} \to \mathrm{Y}$, minimizing the **risk** $R(f) = \mathrm{E}_{\mathrm{P}}\ell(X, Y, f(X))$, that is

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \|y - f(x)\|_{\mathrm{Y}}^{p}\, \mathrm{d}P(x, y)$$

over the space of all measurable functions $\mathcal{M}(\mathcal{X}, \mathrm{Y})$. Without any knowledge of P, but using only training sets z_n.
We are looking for a map (**learning algorithm**)

$$z \in \bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to f_z \in \mathcal{M}(\mathcal{X}, \mathrm{Y})$$

such that the **estimators** f_{Z_n} asymptotically minimize the risk, meaning that $R(f_{Z_n}) \to \inf R(\mathcal{M}(\mathcal{X}, \mathrm{Y}))$ in probability P.

A Learning Algorithm

Regularized Empirical Risk Minimization

We define the **empirical** distribution and risk

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}, \qquad R_n(f, z_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, f(x_i))$$

The problem

$$\min_{f \in \mathcal{M}(\mathcal{X}, \mathrm{Y})} R_n(f, z_n) \tag{1}$$

is **ill-posed** (overfitting). **The solution**: one takes a (sufficiently large) Hilbert space $\mathcal{H} \hookrightarrow \mathcal{M}(\mathcal{X}, \mathrm{Y})$ with embedding $A : \mathcal{H} \to \mathcal{M}(\mathcal{X}, \mathrm{Y})$, consider the restriction of (1) to $\mathcal{H}$ and regularize

$$\min_{u \in \mathcal{H}} R_n(Au, z_n) + \lambda \|u\|_{\mathcal{H}}^2 \qquad (\lambda > 0)$$

A Learning Algorithm

Regularized Empirical Risk Minimization

We define the **empirical** distribution and risk

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}, \qquad R_n(f, z_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, f(x_i))$$

The problem

$$\min_{f \in \mathcal{M}(\mathcal{X}, \mathrm{Y})} R_n(f, z_n) \tag{1}$$

is **ill-posed** (overfitting). **The solution**: one takes a (sufficiently large) Hilbert space $\mathcal{H} \hookrightarrow \mathcal{M}(\mathcal{X}, \mathrm{Y})$ with embedding $A : \mathcal{H} \to \mathcal{M}(\mathcal{X}, \mathrm{Y})$, consider the restriction of (1) to $\mathcal{H}$ and regularize

$$\min_{u \in \mathcal{H}} R_n(Au, z_n) + \lambda \|u\|_{\mathcal{H}}^2 \qquad (\lambda > 0)$$

A Learning Algorithm

Regularized Empirical Risk Minimization

We define the **empirical** distribution and risk

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}, \qquad R_n(f, z_n) = \frac{1}{n}\sum_{i=1}^{n}\ell(x_i, y_i, f(x_i))$$

The problem

$$\min_{f\in\mathcal{M}(\mathcal{X},\mathrm{Y})} R_n(f, z_n) \tag{1}$$

is **ill-posed** (overfitting). **The solution**: one takes a (sufficiently large) Hilbert space $\mathcal{H} \hookrightarrow \mathcal{M}(\mathcal{X}, \mathrm{Y})$ with embedding $A : \mathcal{H} \to \mathcal{M}(\mathcal{X}, \mathrm{Y})$, consider the restriction of (1) to $\mathcal{H}$ and regularize

$$\min_{u\in\mathcal{H}} R_n(Au, z_n) + \lambda\|u\|_{\mathcal{H}}^2 \qquad (\lambda > 0)$$

A Learning Algorithm

Regularized Empirical Risk Minimization

$$u_{n,\lambda}(z_n) \in \operatorname*{argmin}_{u \in \mathcal{H}} R_n(Au, z_n) + \lambda \|u\|^2_{\mathcal{H}}, \quad (z_n \in (\mathcal{X} \times \mathcal{Y})^n)(\lambda > 0)$$

The issue of **consistency**: choose $\lambda_n \to 0$ such that the risk of the estimators $Au_{n,\lambda_n}(Z_n)$ converges (in probability) to the minimal risk as the number of samples goes to infinity.

$$\mathrm{P}\Big[R(A\,u_{n,\lambda_n}(Z_n)) - \inf R(\mathcal{M}(\mathcal{X},\mathrm{Y})) > \delta\Big] \to 0 \qquad (\forall\, \delta > 0)$$

A Learning Algorithm
Regularized Empirical Risk Minimization

$$u_{n,\lambda}(z_n) \in \operatorname*{argmin}_{u \in \mathcal{H}} R_n(Au, z_n) + \lambda \|u\|_{\mathcal{H}}^2, \quad (z_n \in (\mathcal{X} \times \mathcal{Y})^n)(\lambda > 0)$$

The issue of **consistency**: choose $\lambda_n \to 0$ such that the risk of the estimators $Au_{n,\lambda_n}(Z_n)$ converges (in probability) to the minimal risk as the number of samples goes to infinity.

$$\mathrm{P}\Big[R\big(A\, u_{n,\lambda_n}(Z_n)\big) - \inf R(\mathcal{M}(\mathcal{X}, \mathrm{Y})) > \delta \Big] \to 0 \qquad (\forall\, \delta > 0)$$

Our Contribution

- Constrained Risk Minimization.

$$\min_{f \in \mathcal{C}} R(f)$$

where $\mathcal{C} \subset \mathcal{M}(\mathcal{X}, \mathrm{Y})$ is a pointwise constraint.

- The General Variational Learning scheme:

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \mathcal{F}} R_n(Au, z_n) + \lambda J(u), \qquad (\lambda > 0)$$

Consistency: $A\, u_{n,\lambda_n}(Z_n) \to \inf R(\mathcal{C})$ in (outer) probability.

Our Contribution

- Constrained Risk Minimization.

$$\min_{f \in \mathcal{C}} R(f)$$

where $\mathcal{C} \subset \mathcal{M}(\mathcal{X}, \mathrm{Y})$ is a pointwise constraint.

- The General Variational Learning scheme:

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \mathcal{F}} R_n(Au, z_n) + \lambda\, J(u), \qquad (\lambda > 0)$$

Consistency: $A\, u_{n,\lambda_n}(Z_n) \to \inf R(\mathcal{C})$ in (outer) probability.

Our Contribution

We prove *consistency* and *rates* of the empirical risk minimization algorithm in the following **extended** scenario:

- ✓ constraint sets $\mathcal{C} \subset \mathcal{M}(\mathcal{X}, \mathrm{Y})$ (pointwise defined);
- ✓ *Banach spaces* $\mathcal{F} \hookrightarrow \mathcal{M}(\mathcal{X}, \mathrm{Y})$ (instead of Hilbert spaces);
- ✓ *general regularizers* $J : \mathcal{F} \to [0, +\infty]$, totally convex on bounded sets (instead of the square of the norm), even with extended values;
- ✓ *inexactness* in the computation of minimizers.

Our Contribution

Case studies:

- $\mathbb{K}$ a countable set. Consistency holds for

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \ell^r(\mathbb{K})} R_n(Au, z_n) + \lambda\big(\|u\|_r^r + H(u)\big)$$

  with $1 < r < +\infty$ and $H\colon \ell^r(\mathbb{K}) \to [0, +\infty]$ proper, convex l.s.c.

- Nonparametric regression in L^p, $1 < p < +\infty$,

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \|(Au)(x_i) - y_i\|_{\mathsf{Y}}^p + \lambda J(u)$$

Our Contribution

Case studies:

- $\mathbb{K}$ a countable set. Consistency holds for

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \ell^r(\mathbb{K})} R_n(Au, z_n) + \lambda\big(\|u\|_r^r + H(u)\big)$$

  with $1 < r < +\infty$ and $H \colon \ell^r(\mathbb{K}) \to [0, +\infty]$ proper, convex l.s.c.

- Nonparametric regression in L^p, $1 < p < +\infty$,

$$u_{n,\lambda}(z_n) \in \varepsilon_\lambda - \operatorname*{argmin}_{u \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \|(Au)(x_i) - y_i\|_{\mathsf{Y}}^p + \lambda J(u)$$

Geometry of Banach spaces

Definition

The Banach space $\mathcal{B}$ is said to be of (Rademacher) *type* $q \in [1,2]$ if $\exists\, T_q > 0$, so that for every $(u_i)_{1\le i\le n} \in \mathcal{B}^n$ and $n \in \mathbb{N}$

$$\left(\int_0^1 \Big\| \sum_{i=1}^n r_i(t) u_i \Big\|^q \mathrm{d}t\right)^{1/q} \le T_q \Big(\sum_{i=1}^n \|u_i\|^q \Big)^{1/q}$$

The r_n's are the Rademacher functions, i.e. $r_n : [0,1] \to \{0,1\}$, $r_n(t) = \mathrm{sign}(\sin(2^n \pi t))$ for every $t \in [0,1]$ and $n \in \mathbb{N}$.

Definition

$\mathcal{B}$ is said to have modulus of convexity (smoothness) of power type $q \in [1,+\infty[$ if $\exists\, c_q > 0$ (resp. $b_q > 0$) such that $\delta_{\mathcal{B}}(\varepsilon) \ge c_q \varepsilon^q$ $\forall\, \varepsilon \in]0,2]$ (resp. $\rho_{\mathcal{B}}(\tau) \le b_q \tau^q$ $\forall\, \tau > 0$).

Geometry of Banach spaces

One can prove (Lindenstrauss-Tzafriri '79 or Beauzamy '85):

- if $1 < p < +\infty$, then L^p has modulus of convexity of power type $\max\{p, 2\}$ and modulus of smoothness of power type $\min\{p, 2\}$.
- the power type of the modulus of convexity of a Banach space is necessarily ≥ 2, and that of smoothness is ≤ 2
- modulus of smoothness of power type $q \implies$ type q.
- the notion of (Redamacher) type is weaker than that of uniform smoothness of power type, in particular it does not implies reflexivity.

A concentration inequality in Banach spaces

Let $(\Omega, \mathfrak{A}, \mathrm{P})$ be a probability space, $(\mathcal{B}, \|\cdot\|)$ a separable Banach space of type q, $1 < q \leq 2$ and type-q constant T_q. Let $(\xi_i)_{1 \leq i \leq n}$, $\xi_i : \Omega \to \mathcal{B}$ be independent random variables.

Proposition (Ledoux-Talagrand)

If $(\xi_i)_{1 \leq i \leq n}$ have zero mean, then

$$\mathrm{E}_{\mathrm{P}} \left\| \sum_{i=1}^{n} \xi_i \right\|^q \leq (2T_q)^q \sum_{i=1}^{n} \mathrm{E}_{\mathrm{P}} \|\xi_i\|^q ,$$

Theorem (Hoeffding type inequality)

If $\|\xi_i(\omega)\| \leq B$ P-a.s. for all $i = 1, \dots, n$, then, for all $\tau > 0$

$$P\left[\left\| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathrm{E}_{\mathrm{P}} \xi_i) \right\| \geq 4B \left(\frac{T_q}{n^{1-1/q}} + \sqrt{\frac{\tau}{2n}} + \frac{\tau}{3n} \right) \right] \leq e^{-\tau}.$$

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

  $$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

  $\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$
- Total convexity on bounded sets:

  $$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$
- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t \to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

  $$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

  $\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$
- Total convexity on bounded sets:

  $$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$

- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t\to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

$$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

$\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$

- Total convexity on bounded sets:

$$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$

- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t\to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

  $$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

  $\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$

- Total convexity on bounded sets:

  $$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$

- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t \to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

$$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

$\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$

- Total convexity on bounded sets:

$$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$

- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t\to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

- Total convexity at $u_0 \in \operatorname{dom} J$:

$$(\forall\, u \in \operatorname{dom} J) \quad J(u) - J(u_0) \geq J'(u_0, u - u_0) + \psi(u_0; \|u - u_0\|_{\mathcal{F}})\,.$$

$\psi(u_0, \cdot) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$, $\psi(u_0, 0) = 0$ and $\psi(u_0, t) > 0$ if $t > 0$

- Total convexity on bounded sets:

$$(\forall\, \rho > 0)(\forall\, t > 0) \quad \psi_\rho(t) = \inf_{\|u_0\|_{\mathcal{F}} \leq \rho} \psi(u_0; t) > 0$$

- $\hat{\psi}_\rho(t) := \psi_\rho(t)/t$ is increasing, $\lim_{t\to 0} \hat{\psi}_\rho(t) = 0$.
- $(\hat{\psi}_\rho)^\natural(s) = \sup\{\hat{\psi}_\rho \leq s\}$ the greatest quasi-inverse of $\hat{\psi}_\rho$ is increasing and $\operatorname{dom}(\hat{\psi}_\rho)^\natural = [0, +\infty[$.
- ψ_0 is the modulus of total convexity at zero.
- total convexity on bounded sets $\iff$ uniform convexity on bounded sets (Zalinescu).

Totally convex functions

Example: Let $r \in]1, +\infty[$ and $\mathcal{F}$ be a uniformly convex Banach space with modulus of convexity of power type $q \in [2, +\infty[$ and set $J = \|\cdot\|_{\mathcal{F}}^r$. Then for every $\rho > 0$ and $t > 0$

$$\psi_\rho(t) \geq \begin{cases} \dfrac{K_r c_q}{2^r} t^r & \text{if } r \geq q \\ \dfrac{r}{q} \dfrac{K_r c_q}{2^q} \dfrac{t^q}{(\rho + t)^{q-r}} & \text{if } r < q\,, \end{cases}$$

- For $r > 1$, $\ell^r(\mathbb{K})$ is uniformly convex with modulus of convexity of power type $q = \max\{2, r\}$. Hence $\|\cdot\|_r^r$ is uniformly convex if $r \geq 2$, and only totally convex on bounded sets if $1 < r < 2$.

Totally convex functions

Example: Let $r \in]1, +\infty[$ and $\mathcal{F}$ be a uniformly convex Banach space with modulus of convexity of power type $q \in [2, +\infty[$ and set $J = \|\cdot\|_{\mathcal{F}}^r$. Then for every $\rho > 0$ and $t > 0$

$$\psi_\rho(t) \geq \begin{cases} \dfrac{K_r c_q}{2^r} t^r & \text{if } r \geq q \\ \dfrac{r}{q} \dfrac{K_r c_q}{2^q} \dfrac{t^q}{(\rho + t)^{q-r}} & \text{if } r < q\,, \end{cases}$$

- For $r > 1$, $\ell^r(\mathbb{K})$ is uniformly convex with modulus of convexity of power type $q = \max\{2, r\}$. Hence $\|\cdot\|_r^r$ is uniformly convex if $r \geq 2$, and only totally convex on bounded sets if $1 < r < 2$.

Reproducing Kernel Banach spaces

Definition

A Banach space of functions $\mathcal{W} \subset \mathrm{Y}^{\mathcal{X}}$ such that the evaluation functionals $ev_x : \mathcal{W} \to \mathrm{Y}$ are (linear) continuous for all $x \in \mathcal{X}$.

A way to generate RKBS'.

Proposition

Let $\mathcal{F}$ a Banach space and $A : \mathcal{F} \to \mathrm{Y}^{\mathcal{X}}$ a linear operator continuous for the topology of the point-wise convergence. Then $\operatorname{Im} A$ can be endowed with a norm which make it a RKBS and A a partial isometry.

The associated feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathrm{Y}^*, \mathcal{F}^*)$

$$(\forall u \in \mathcal{F})(\forall x \in \mathcal{X}) \quad (Au)(x) = \gamma(x)^* u$$

γ is measurable if and only if $\operatorname{Im} A \subset \mathcal{M}(\mathcal{X}, \mathrm{Y})$.

Reproducing Kernel Banach spaces

Definition

A Banach space of functions $\mathcal{W} \subset \mathsf{Y}^{\mathcal{X}}$ such that the evaluation functionals $ev_x : \mathcal{W} \to \mathsf{Y}$ are (linear) continuous for all $x \in \mathcal{X}$.

A way to generate RKBS'.

Proposition

Let $\mathcal{F}$ a Banach space and $A : \mathcal{F} \to \mathsf{Y}^{\mathcal{X}}$ a linear operator continuous for the topology of the point-wise convergence. Then $\operatorname{Im} A$ *can be endowed with a norm which make it a RKBS and A a partial isometry.*

The associated feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathsf{Y}^*, \mathcal{F}^*)$

$$(\forall\, u \in \mathcal{F})(\forall\, x \in \mathcal{X}) \quad (Au)(x) = \gamma(x)^* u$$

γ is measurable if and only if $\operatorname{Im} A \subset \mathcal{M}(\mathcal{X}, \mathsf{Y})$.

Reproducing Kernel Banach spaces

Definition

A Banach space of functions $\mathcal{W} \subset \mathsf{Y}^{\mathcal{X}}$ such that the evaluation functionals $ev_x : \mathcal{W} \to \mathsf{Y}$ are (linear) continuous for all $x \in \mathcal{X}$.

A way to generate RKBS'.

Proposition

Let $\mathcal{F}$ a Banach space and $A : \mathcal{F} \to \mathsf{Y}^{\mathcal{X}}$ a linear operator continuous for the topology of the point-wise convergence. Then $\operatorname{Im} A$ *can be endowed with a norm which make it a RKBS and A a partial isometry.*

The associated feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathsf{Y}^*, \mathcal{F}^*)$

$$(\forall\, u \in \mathcal{F})(\forall\, x \in \mathcal{X}) \quad (Au)(x) = \gamma(x)^* u$$

γ is measurable if and only if $\operatorname{Im} A \subset \mathcal{M}(\mathcal{X}, \mathsf{Y})$.

Reproducing Kernel Banach spaces

Example 1: Let $\mathsf{Y} = \mathbb{R}$ (scalar case) and $\mathcal{F}$ strictly convex and smooth. Then there exists a map $\Phi : \mathcal{X} \to \mathcal{F}$ such that $j_2(\Phi(x)) = \gamma(x)^*$ for every $x \in \mathcal{X}$. Therefore for each $u \in \mathcal{F}$ and $x \in \mathcal{X}$

$$(Au)(x) = \langle u, \gamma(x)^* \rangle_{\mathcal{F},\mathcal{F}^*} = \langle u, j_2(\Phi(x)) \rangle_{\mathcal{F},\mathcal{F}^*} .$$

We can define

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad K(x, x') = \langle \Phi(x), j_2(\Phi(x')) \rangle_{\mathcal{F},\mathcal{F}^*}$$

and it holds $K(x, \cdot) = A\Phi(x) \in \mathcal{W} = \operatorname{Im} A$. The function K is the kernel associated to the feature map γ (or equivalently Φ) and satisfies

(i) $K(x, x) = \|\Phi(x)\|^2_{\mathcal{F}} = \|\gamma(x)\|^2_{\mathcal{F}} \geq 0$.

(ii) $|K(x, x')|^2 \leq K(x, x)K(x', x')$.

Reproducing Kernel Banach spaces

Example 2: Let $\mathbb{K}$ be a countable set, $r, r' \in [1, +\infty]$ with $1/r + 1/r' = 1$ and $(\varphi_k)_{k\in\mathbb{K}}$ a *dictionary* of functions $\varphi_k : \mathcal{X} \to \mathrm{Y}$. Assume that for every $x \in \mathcal{X}$, $(\|\varphi_k(x)\|_{\mathrm{Y}})_{k\in K} \in \ell^{r'}(\mathbb{K})$. Then, the following linear operator is well-defined

$$A : \ell^r(\mathbb{K}) \to \mathrm{Y}^{\mathcal{X}} \qquad (A\beta)(x) = \sum_{k\in\mathbb{K}} \beta_k \varphi_k(x)\,.$$

Then

$$\mathcal{W} = \operatorname{Im} A = \left\{ f \in \mathrm{Y}^{\mathcal{X}} \,\middle|\, (\exists\, \beta \in \ell^r(\mathbb{K}))(\forall\, x \in \mathcal{X}) \quad f(x) = \sum_{k\in\mathbb{K}} \beta_k \varphi_k(x) \right\}$$

and

$$\|f\|_{\mathcal{W}} = \inf \left\{ \|\beta\|_r \,\middle|\, (\exists\, \beta \in \ell^r(\mathbb{K}))(\forall\, x \in \mathcal{X}) \quad f(x) = \sum_{k\in\mathbb{K}} \beta_k \varphi_k(x) \right\}.$$

General assumptions.

- ℓ is convex and locally Lipschitz continuous, that is
$$|\ell(x, y, w) - \ell(x, y, w')| \leq |\ell|_{\rho,1} \|w - w'\|_{\mathsf{Y}}$$
for every $\rho > 0$ and $(w, w') \in \mathsf{Y}^2$, $\|w\|_{\mathsf{Y}}, \|w'\|_{\mathsf{Y}} \leq \rho$.
- $\mathcal{F}$ is a separable Banach space, and $\mathcal{F}^*$ is of (Rademacher) type $q' > 1$ (necessarily $q' \leq 2$);
- the feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathsf{Y}^*, \mathcal{F}^*)$ is measurable and bounded;
- $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \mid (\forall x \in \mathcal{X})(f(x) \in \mathsf{C}(x))\}$, with $\mathsf{C}(x) \subset \mathsf{Y}$ nonempty closed convex for every $x \in \mathcal{X}$ (pointwise constraint).
- $J : \mathcal{F} \to [0, +\infty]$ is a l.s.c. function, totally convex on bounded sets with modulus of total convexity on the ball $B_{\mathcal{F}}(\rho)$ denoted by ψ_ρ and $J(0) = 0$.

General assumptions.

- ℓ is convex and locally Lipschitz continuous, that is

$$|\ell(x,y,w) - \ell(x,y,w')| \leq |\ell|_{\rho,1} \|w - w'\|_{\mathrm{Y}}$$

for every $\rho > 0$ and $(w, w') \in \mathrm{Y}^2$, $\|w\|_{\mathrm{Y}}, \|w'\|_{\mathrm{Y}} \leq \rho$.
- $\mathcal{F}$ is a separable Banach space, and $\mathcal{F}^*$ is of (Rademacher) type $q' > 1$ (necessarily $q' \leq 2$);
- the feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathrm{Y}^*, \mathcal{F}^*)$ is measurable and bounded;
- $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathrm{Y}) \,|\, (\forall x \in \mathcal{X})(f(x) \in \mathrm{C}(x))\}$, with $\mathrm{C}(x) \subset \mathrm{Y}$ nonempty closed convex for every $x \in \mathcal{X}$ (pointwise constraint).
- $J : \mathcal{F} \to [0, +\infty]$ is a l.s.c. function, totally convex on bounded sets with modulus of total convexity on the ball $B_{\mathcal{F}}(\rho)$ denoted by ψ_ρ and $J(0) = 0$.

General assumptions.

- ℓ is convex and locally Lipschitz continuous, that is
$$|\ell(x,y,w)-\ell(x,y,w')| \leq |\ell|_{\rho,1}\|w-w'\|_{\mathsf{Y}}$$
for every $\rho > 0$ and $(w,w') \in \mathsf{Y}^2$, $\|w\|_{\mathsf{Y}}, \|w'\|_{\mathsf{Y}} \leq \rho$.
- $\mathcal{F}$ is a separable Banach space, and $\mathcal{F}^*$ is of (Rademacher) type $q' > 1$ (necessarily $q' \leq 2$);
- the feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathsf{Y}^*, \mathcal{F}^*)$ is measurable and bounded;
- $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \mid (\forall x \in \mathcal{X})(f(x) \in \mathsf{C}(x))\}$, with $\mathsf{C}(x) \subset \mathsf{Y}$ nonempty closed convex for every $x \in \mathcal{X}$ (pointwise constraint).
- $J : \mathcal{F} \to [0, +\infty]$ is a l.s.c. function, totally convex on bounded sets with modulus of total convexity on the ball $B_{\mathcal{F}}(\rho)$ denoted by ψ_ρ and $J(0) = 0$.

General assumptions.

- ℓ is convex and locally Lipschitz continuous, that is
$$|\ell(x,y,w) - \ell(x,y,w')| \leq |\ell|_{\rho,1} \|w - w'\|_{\mathsf{Y}}$$
for every $\rho > 0$ and $(w, w') \in \mathsf{Y}^2$, $\|w\|_{\mathsf{Y}}, \|w'\|_{\mathsf{Y}} \leq \rho$.
- $\mathcal{F}$ is a separable Banach space, and $\mathcal{F}^*$ is of (Rademacher) type $q' > 1$ (necessarily $q' \leq 2$);
- the feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathsf{Y}^*, \mathcal{F}^*)$ is measurable and bounded;
- $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \,|\, (\forall x \in \mathcal{X})(f(x) \in \mathsf{C}(x))\}$, with $\mathsf{C}(x) \subset \mathsf{Y}$ nonempty closed convex for every $x \in \mathcal{X}$ (pointwise constraint).
- $J : \mathcal{F} \to [0, +\infty]$ is a l.s.c. function, totally convex on bounded sets with modulus of total convexity on the ball $B_{\mathcal{F}}(\rho)$ denoted by ψ_ρ and $J(0) = 0$.

General assumptions.

- ℓ is convex and locally Lipschitz continuous, that is

$$|\ell(x, y, w) - \ell(x, y, w')| \leq |\ell|_{\rho,1} \|w - w'\|_{\mathrm{Y}}$$

  for every $\rho > 0$ and $(w, w') \in \mathrm{Y}^2$, $\|w\|_{\mathrm{Y}}, \|w'\|_{\mathrm{Y}} \leq \rho$.
- $\mathcal{F}$ is a separable Banach space, and $\mathcal{F}^*$ is of (Rademacher) type $q' > 1$ (necessarily $q' \leq 2$);
- the feature map $\gamma : \mathcal{X} \to \mathcal{L}(\mathrm{Y}^*, \mathcal{F}^*)$ is measurable and bounded;
- $\mathcal{C} = \big\{f \in \mathcal{M}(\mathcal{X}, \mathrm{Y}) \,|\, (\forall\, x \in \mathcal{X})(f(x) \in \mathrm{C}(x))\big\}$, with $\mathrm{C}(x) \subset \mathrm{Y}$ nonempty closed convex for every $x \in \mathcal{X}$ (pointwise constraint).
- $J : \mathcal{F} \to [0, +\infty]$ is a l.s.c. function, totally convex on bounded sets with modulus of total convexity on the ball $B_{\mathcal{F}}(\rho)$ denoted by ψ_ρ and $J(0) = 0$.

A key decomposition

- If $\overline{dom J} = A^{-1}(\mathcal{C})$ and $\mathcal{C} \cap \operatorname{Im} A$ is dense in $\mathcal{C} \cap L^p(\mathcal{X}, P_{\mathcal{X}}; \mathrm{Y})$, then $\inf I(\operatorname{dom} J) = \inf R(\mathcal{C})$.
- An auxiliary (non-stochastic) regularized risk minimization problem is introduced:

$$u_\lambda \in \underset{u \in \mathcal{F}}{\operatorname{argmin}}\, R(Au) + \lambda J(u) \qquad (\lambda > 0)$$

If we set $I = R \circ A$ and $I_n(\cdot, z_n) = R_n(\cdot, z_n) \circ A$, then

$$I(u_{n,\lambda}(Z_n)) - \inf I(\operatorname{dom} J) = \underbrace{I(u_{n,\lambda}(Z_n)) - I(u_\lambda)}_{\text{sample error}} + \underbrace{I(u_\lambda) - \inf I(\operatorname{dom} J)}_{\text{approximation error}}$$

The behavior of both errors is studied separately.

A key decomposition

- If $\overline{dom J} = A^{-1}(\mathcal{C})$ and $\mathcal{C} \cap \operatorname{Im} A$ is dense in $\mathcal{C} \cap L^p(\mathcal{X}, P_{\mathcal{X}}; \mathrm{Y})$, then $\inf I(\operatorname{dom} J) = \inf R(\mathcal{C})$.
- An auxiliary (non-stochastic) regularized risk minimization problem is introduced:

$$u_\lambda \in \operatorname*{argmin}_{u \in \mathcal{F}} R(Au) + \lambda J(u) \qquad (\lambda > 0)$$

If we set $I = R \circ A$ and $I_n(\cdot, z_n) = R_n(\cdot, z_n) \circ A$, then

$$I(u_{n,\lambda}(Z_n)) - \inf I(\operatorname{dom} J) = \underbrace{I(u_{n,\lambda}(Z_n)) - I(u_\lambda)}_{\text{sample error}} + \underbrace{I(u_\lambda) - \inf I(\operatorname{dom} J)}_{\text{approximation error}}$$

The behavior of both errors is studied separately.

The study of the approximation error (weak case)

We consider $u_\lambda \in \varepsilon(\lambda)\text{-}\operatorname{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Proposition

1. *If* $\lim_{\lambda \to 0} \varepsilon(\lambda) = 0$, *then* $\lim_{\lambda \to 0} I(u_\lambda) = \inf I(\operatorname{dom} J)$.
2. *If* J *is totally convex and* $\{0\} = \operatorname{argmin} J \cap \operatorname{dom} I$, *then*

$$(\forall \lambda \in \,]0, +\infty[) \quad \|u_\lambda\|_{\mathcal{F}} \leq \psi_0^\natural\left(\frac{I(0) - \inf I(\operatorname{dom} J) + \varepsilon(\lambda)}{\lambda}\right).$$

The study of the approximation error (strong case)

We consider $u_\lambda \in \varepsilon(\lambda)\text{-}\operatorname{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Proposition (Attouch '96)

If $\varepsilon(\lambda)/\lambda \to 0$, J is coercive and $S = \operatorname{argmin}_{\operatorname{dom} J} I \neq \varnothing$, then $(u_\lambda)_{\lambda>0}$ is bounded. Moreover

1. *if $\lambda_n \to 0$ and $u_{\lambda_n} \rightharpoonup u^\dagger$ for some $u^\dagger \in \mathfrak{F}$, then $u^\dagger \in \operatorname{argmin}_{u \in S} J(u)$*
2. $\lim_{\lambda\to 0} J(u_\lambda) = \inf J(S)$
3. $\lim_{\lambda\to 0} \frac{1}{\lambda}\big(I(u_\lambda) - \inf I(\operatorname{dom} J)\big) = 0$.
4. *If J is strictly quasiconvex, then $u^\dagger$ is uniquely determined and $u_\lambda \rightharpoonup u^\dagger$ as $\lambda \to 0$.*
5. *If J is totally convex on bounded sets, then $u_\lambda \to u^\dagger$ as $\lambda \to 0$.*

A General Representer Theorem

We consider $u_\lambda \in \operatorname{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Theorem (Representer)

For all $\lambda > 0$, there exists $h_\lambda \in L^{p'}(\mathcal{X} \times \mathcal{Y}, P; \mathrm{Y}^*)$ such that

$$h_\lambda(x, y) \in \partial\ell(x, y, Au_\lambda(x))$$

$$-\mathrm{E}_P[\gamma h_\lambda] \in \lambda \partial J(u_\lambda)\,,$$

where $\gamma h_\lambda : \mathcal{X} \times \mathcal{Y} \to \mathcal{F}^*$, $(\gamma h_\lambda)(x, y) = \gamma(x) h_\lambda(x, y)$.

Moreover, if $p = 1$, $\|h_\lambda\|_\infty \leq c_\ell$; If $p = \infty$ and ℓ is locally Lipschitz continuous, we have $\|h_\lambda\|_\infty \leq |\ell|_{\kappa\rho_\lambda, 1}$, where $\kappa = \|\gamma\|_\infty$ and $\rho_\lambda > 0$ is any number such that $\rho_\lambda > \|u_\lambda\|$.

A General Representer Theorem

We consider $u_\lambda \in \operatorname{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Theorem (Representer)

For all $\lambda > 0$, there exists $h_\lambda \in L^{p'}(\mathcal{X} \times \mathcal{Y}, P; \mathrm{Y}^*)$ such that

$$h_\lambda(x,y) \in \partial \ell(x,y,Au_\lambda(x))$$

$$u_\lambda = j_{r'}\big(-\mathrm{E}_P[\gamma h_\lambda]/(r\lambda)\big), \quad J = \|\cdot\|_{\mathcal{F}}^r$$

where $\gamma h_\lambda : \mathcal{X} \times \mathcal{Y} \to \mathcal{F}^*$, $(\gamma h_\lambda)(x,y) = \gamma(x) h_\lambda(x,y)$.

Moreover, if $p = 1$, $\|h_\lambda\|_\infty \le c_\ell$; If $p = \infty$ and ℓ is locally Lipschitz continuous, we have $\|h_\lambda\|_\infty \le |\ell|_{\kappa\rho_\lambda,1}$, where $\kappa = \|\gamma\|_\infty$ and $\rho_\lambda > 0$ is any number such that $\rho_\lambda > \|u_\lambda\|$.

A General Representer Theorem

We consider $u_\lambda \in \operatorname{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Theorem (Representer)

For all $\lambda > 0$, there exists $h_\lambda \in L^{p'}(\mathcal{X} \times \mathcal{Y}, P; \mathrm{Y}^*)$ such that

$$h_\lambda(x, y) \in \partial\ell(x, y, Au_\lambda(x))$$

$$u_\lambda = j_{r'}\Big(\sum_{i=1}^{n} \gamma(x_i)\alpha_i\Big), \qquad \alpha_i = -1/(nr\lambda)h(x_i, y_i) \in \mathrm{Y}^* .$$

where $\gamma h_\lambda : \mathcal{X} \times \mathcal{Y} \to \mathcal{F}^*$, $(\gamma h_\lambda)(x, y) = \gamma(x)h_\lambda(x, y)$.

Moreover, if $p = 1$, $\|h_\lambda\|_\infty \leq c_\ell$; If $p = \infty$ and ℓ is locally Lipschitz continuous, we have $\|h_\lambda\|_\infty \leq |\ell|_{\kappa\rho_\lambda,1}$, where $\kappa = \|\gamma\|_\infty$ and $\rho_\lambda > 0$ is any number such that $\rho_\lambda > \|u_\lambda\|$.

A Stability Theorem

We consider $u_\lambda \in \mathrm{argmin}_{\mathcal{F}}(I + \lambda J)$, for all $\lambda > 0$.

Theorem (Stability)

Suppose further that J is totally convex with modulus of total convexity $\psi(u, \cdot)$ at u.
Then, for every distribution $\tilde{P}$ on $\mathcal{X} \times \mathcal{Y}$ and any $\tilde{u}_\lambda \in \mathcal{F}$ with $d(0, \partial(\tilde{I} + \lambda J)(\tilde{u}_\lambda)) \leq \varepsilon$, we have

$$\hat{\psi}(u_\lambda, \|\tilde{u}_\lambda - u_\lambda\|_{\mathcal{F}}) \leq \frac{1}{\lambda}\|\mathrm{E}_{\tilde{P}}[\gamma h_\lambda] - \mathrm{E}_P[\gamma h_\lambda]\|_{\mathcal{F}^*} + \frac{\varepsilon}{\lambda},$$

where $\hat{\psi}(u, t) = \psi(u, t)/t$.

Weak consistency theorem

Theorem (part one)

Suppose $\ell(\cdot,\cdot,0)$ *is bounded and set* $\ell_{max} := \|\ell(\cdot,\cdot,0)\|_\infty$, *and* $\kappa = \|\gamma\|_\infty$. *Let* $u_{n,\lambda}(z_n) \in \varepsilon_1(\lambda)\varepsilon_2(\lambda)\text{-argmin}(I_n(\cdot, z_n) + \lambda J)$ *for each* $z_n \in (\mathcal{X} \times \mathcal{Y})^n$. *Then, for every* $\tau > 0$

$$\mathrm{P}^*\Big[\, I(u_{n,\lambda}(Z_n)) - \inf I(\mathrm{dom}\, J) > \eta(n,\tau,\lambda) + I(u_\lambda) - \inf I(\mathrm{dom}\, J) \Big] \leq e^{-\tau},$$

where $\eta(n,\tau,\lambda)$ *is equal to*

$$\kappa|\ell|_{\kappa\rho_\lambda,1}\left\{\varepsilon_1(\lambda) + (\hat{\psi}_{\rho_\lambda})^\natural\left(\frac{\kappa|\ell|_{\kappa\rho_\lambda,1}}{\lambda}\left(\frac{4T_{q'}}{n^{1/q}} + \sqrt{\frac{2\tau}{n}}\right) + \frac{\varepsilon_2(\lambda)}{\lambda}\right)\right\}$$

and $\rho_\lambda = \psi_0^\natural((\ell_{max}+1)/\lambda)$.

Weak consistency theorem

Theorem (part two)

Suppose $\ell(\cdot,\cdot,0)$ *is bounded and set* $\ell_{max} := \|\ell(\cdot,\cdot,0)\|_\infty$, *and* $\kappa = \|\gamma\|_\infty$. *Let* $u_{n,\lambda}(z_n) \in \varepsilon_1(\lambda)\varepsilon_2(\lambda)\text{-argmin}(I_n(\cdot, z_n) + \lambda J)$ *for each* $z_n \in (\mathcal{X} \times \mathcal{Y})^n$. *Moreover if* $(\lambda_n)_{n\in\mathbb{N}}$, *is such that* $\lambda_n \to 0$ *and*

$$L_n\varepsilon_1(\lambda_n) \to 0\,,\ \varepsilon_2(\lambda_n) = O\Big(\frac{L_n}{n^{1/q}}\Big)\,,\ L_n(\hat{\psi}_{\rho_{\lambda_n}})^\natural\Big(\frac{L_n}{\lambda_n n^{1/q}}\Big) \to 0\,,$$

where $L_n = |\ell|_{\kappa\rho_{\lambda_n},1}$, *then*

$$(\forall\,\delta > 0)\qquad \lim_{n\to+\infty} \mathrm{P}^*\Big[\,I(u_{n,\lambda_n}(Z_n)) - \inf I(\mathrm{dom}\,J) > \delta\,\Big] = 0\,.$$

Weak consistency theorem

Sketch of the Proof.

- using the Ekeland's variational principle, $\exists v_{n,\lambda} \in \mathcal{F}$ such that

$$\|u_{n,\lambda}(z_n) - v_{n,\lambda}\|_{\mathcal{F}} \leq \varepsilon_1(\lambda), \qquad d(0, \partial(I_n(\cdot, z_n) + \lambda J))(v_{n,\lambda}) \leq \varepsilon_2(\lambda)$$

- using the Representer and Stability Theorems with $\tilde{P}$ the empirical distribution

$$\|v_{n,\lambda} - u_\lambda\|_{\mathcal{F}} \leq (\hat{\psi}_{\rho_\lambda})^\natural \Big(\frac{1}{\lambda} \Big\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) h_\lambda(x_i, y_i) \Big\|_{\mathcal{F}^*} + \frac{\varepsilon_2(\lambda)}{\lambda} \Big).$$

- using Hoeffding's inequality with $\xi_i = \gamma(X_i) h_\lambda(X_i, Y_i) : \Omega \to \mathcal{F}^*$

$$\mathrm{P}\Big[\Big\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n} \sum_{i=1}^{n} \gamma(X_i) h_\lambda(X_i, Y_i) \Big\|_{\mathcal{F}^*} \leq \kappa |\ell|_{\kappa \rho_\lambda} \delta(n, \tau) \Big] \geq 1 - e^{-\tau}$$

Weak consistency theorem

Sketch of the Proof.

- using the Ekeland's variational principle, $\exists v_{n,\lambda} \in \mathcal{F}$ such that

$$\|u_{n,\lambda}(z_n) - v_{n,\lambda}\|_{\mathcal{F}} \leq \varepsilon_1(\lambda), \qquad d(0, \partial(I_n(\cdot, z_n)+\lambda J))(v_{n,\lambda}) \leq \varepsilon_2(\lambda)$$

- using the Representer and Stability Theorems with $\tilde{P}$ the empirical distribution

$$\|v_{n,\lambda} - u_\lambda\|_{\mathcal{F}} \leq (\hat{\psi}_{\rho_\lambda})^\natural \Big(\frac{1}{\lambda} \Big\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n} \sum_{i=1}^n \gamma(x_i) h_\lambda(x_i, y_i) \Big\|_{\mathcal{F}^*} + \frac{\varepsilon_2(\lambda)}{\lambda} \Big).$$

- using Hoeffding's inequality with $\xi_i = \gamma(X_i) h_\lambda(X_i, Y_i) : \Omega \to \mathcal{F}^*$

$$\mathrm{P}\Big[\Big\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n} \sum_{i=1}^n \gamma(X_i) h_\lambda(X_i, Y_i) \Big\|_{\mathcal{F}^*} \leq \kappa |\ell|_{\kappa\rho_\lambda} \delta(n, \tau) \Big] \geq 1 - e^{-\tau}$$

Weak consistency theorem

Sketch of the Proof.

- using the Ekeland's variational principle, $\exists v_{n,\lambda} \in \mathcal{F}$ such that

$$\|u_{n,\lambda}(z_n) - v_{n,\lambda}\|_{\mathcal{F}} \leq \varepsilon_1(\lambda), \qquad d(0, \partial(I_n(\cdot, z_n)+\lambda J))(v_{n,\lambda}) \leq \varepsilon_2(\lambda)$$

- using the Representer and Stability Theorems with $\tilde{P}$ the empirical distribution

$$\|v_{n,\lambda} - u_\lambda\|_{\mathcal{F}} \leq (\hat{\psi}_{\rho_\lambda})^\natural \Big(\frac{1}{\lambda} \Big\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n}\sum_{i=1}^n \gamma(x_i) h_\lambda(x_i, y_i) \Big\|_{\mathcal{F}^*} + \frac{\varepsilon_2(\lambda)}{\lambda} \Big).$$

- using Hoeffding's inequality with $\xi_i = \gamma(X_i)h_\lambda(X_i, Y_i) : \Omega \to \mathcal{F}^*$

$$\mathrm{P}\Bigg[\Bigg\| \mathrm{E}_P[\gamma h_\lambda] - \frac{1}{n}\sum_{i=1}^n \gamma(X_i) h_\lambda(X_i, Y_i) \Bigg\|_{\mathcal{F}^*} \leq \kappa |\ell|_{\kappa\rho_\lambda} \delta(n, \tau) \Bigg] \geq 1 - e^{-\tau}$$

Strong consistency theorem

Theorem (part one)

We additionally assume $\operatorname{argmin}_{\operatorname{dom} J} I \neq \varnothing$. *Then* $(u_\lambda)_{\lambda>0}$ *is bounded, and*

$$\mathrm{P}^*\left[\,\|u_{n,\lambda}(Z_n) - u^\dagger\|_{\mathcal{F}} > \eta(n,\tau,\lambda) + \|u_\lambda - u^\dagger\|_{\mathcal{F}}\right] \leq e^{-\tau}$$

$$\mathrm{P}^*\left[\,I(u_{n,\lambda}(Z_n)) - \inf I(\operatorname{dom} J) > \kappa|\ell|_{\kappa\rho_\lambda}\eta(n,\tau,\lambda) + \lambda\right] \leq e^{-\tau}\,,$$

where

$$\eta(n,\tau,\lambda) = \varepsilon_1(\lambda) + (\hat{\psi}_\rho)^\natural\left(\frac{\kappa|\ell|_{\kappa\rho,1}}{\lambda}\left(\frac{4T_{q'}}{n^{1/q}} + \sqrt{\frac{2\tau}{n}}\right) + \frac{\varepsilon_2(\lambda)}{\lambda}\right)$$

and $\rho = \sup_{\lambda>0}\|u_\lambda\|_{\mathcal{F}}$, $\rho_\lambda = \psi_0^\natural((\ell_{max}+1)/\lambda)$.

Strong consistency theorem

Theorem (part two)

We additionally assume $\operatorname{argmin}_{\operatorname{dom} J} I \neq \varnothing$. *Moreover if* $\lambda_n \to 0$ *and*

$$\lambda_n n^{1/q} \to +\infty, \quad \varepsilon_1(\lambda_n) \to 0, \quad \frac{\varepsilon_2(\lambda_n)}{\lambda_n} \to 0$$

then

$$(\forall \delta > 0) \quad \lim_{n \to +\infty} \mathrm{P}^* \Big[\|u_{n,\lambda_n}(Z_n) - u^\dagger\|_{\mathcal{F}} > \delta \Big] = 0$$

Finally if $n^{1/q}\lambda_n / \log n \to +\infty$, *then*

$$\lim_{n \to +\infty} u_{n,\lambda_n}(Z_n) = u^\dagger \quad \mathrm{P} - a.s.$$

The regression function

$$R(f) = \mathrm{E}\|Y - f(X)\|_{\mathrm{Y}}^p = \int_{\mathcal{X}\times\mathcal{Y}} \|y - f(x)\|_{\mathrm{Y}}^p \,\mathrm{d}P(x,y) \quad (1 < p < +\infty)\,.$$

Definition

A function $f_*^{\mathcal{C}} \in \mathcal{M}(\mathcal{X}, \mathrm{Y})$, is called the *p-regression function on* $\mathcal{C}$ if $f_*^{\mathcal{C}} \in \mathcal{C}$ and $R(f_*^{\mathcal{C}}) = \inf R(\mathcal{C})$.

Proposition

The regression function $f_*^{\mathcal{C}}$ exists and for every $f \in \mathcal{C} \cap L^p(\mathcal{X}, P_{\mathcal{X}}; \mathrm{Y})$

$$R(f) - \inf R(\mathcal{C}) \leq C_p \|f - f_*^{\mathcal{C}}\|_p^{\min\{2,p\}} \big(\inf R(\mathcal{C}) + \|f - f_*^{\mathcal{C}}\|_p\big)^{\max\{2,p\}-2},$$

$$\|f_*^{\mathcal{C}} - f\|_p^{\max\{2,p\}} \leq D_p\big(R(f) - \inf R(\mathcal{C})\big) R(f)^{\frac{2-\min\{2,p\}}{p}}\,.$$

Further generalization

Theorem (Xu-Roach '91)

Let $\mathcal{B}$ be Banach space and $p \in]1, +\infty[$. If $\mathcal{B}$ is uniformly convex, then

$$(\forall\, u \in \mathcal{B})(\forall\, \xi \in \mathcal{J}_p(u))(\forall\, v \in \mathcal{B}) \quad \|u+v\|^p - \|u\|^p \geq p\langle \xi, v\rangle + \sigma_p(u,v)$$

where

$$\sigma_p(u,v) = pK_p \int_0^1 \frac{(\|u+tv\| \vee \|u\|)^p}{t} \delta_{\mathcal{B}}\left(\frac{t\|v\|}{2\|u+tv\| \vee \|u\|}\right) \mathrm{d}t\,.$$

and $K_p > 0$ is a constant.

We want to obtain an analogous theorem for the case $\Phi(\|\cdot\|)$, with $\Phi(t) = \int_0^t \phi(s)ds$. This would allow to do nonparametric regression in Orlicz spaces.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.

Conclusion

- We present a General Variational Learning algorithm which constitutes an extension of the regularized ERM under several aspects:
  - ✓ constraints (pointwise positiveness, boundedness);
  - ✓ general loss functions;
  - ✓ general regularization functions (totally convex on bounded sets);
  - ✓ Banach Spaces setting;
- We proved weak and strong consistency theorems;
- We deal with nonparametric regression in L^p spaces;
- The framework is shown to cover the significant case of $\|\cdot\|_r^r$ regularization, with $1 < r \leq 2$.