CONSISTENT LEARNING BY COMPOSITE PROXIMAL THRESHOLDING

Saverio Salzo
Università degli Studi di Genova

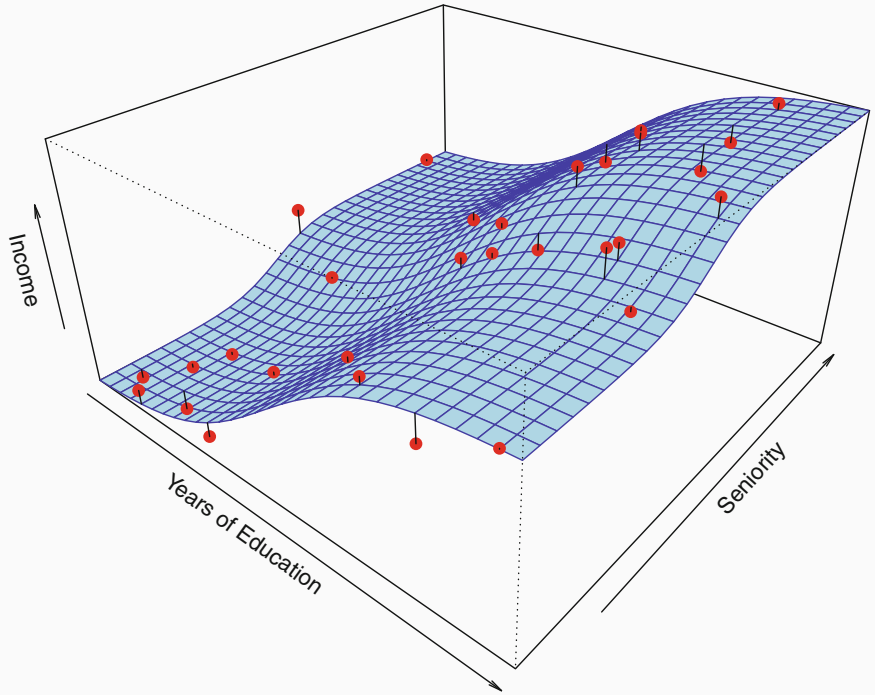
Optimization in Machine learning, vision and image processing
Université Paul Sabatier, Toulouse
6-7 October 2015

OUTLINE

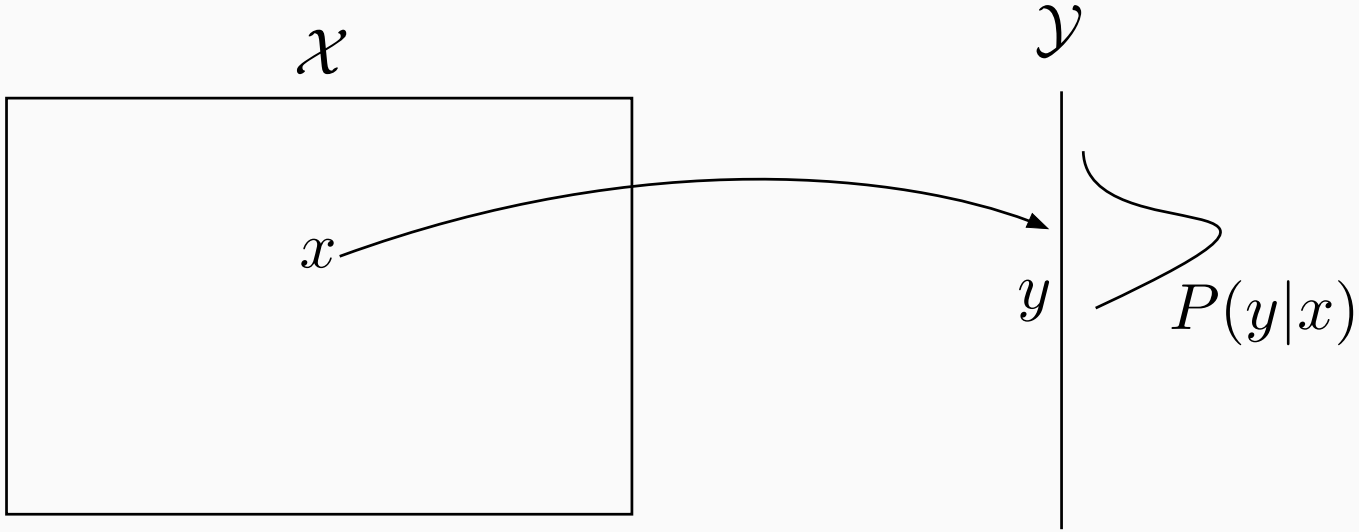
1. Introduction
   - The learning problem
   - Constructing the estimators
   - Our contribution
2. Statistical Analysis
   - Proving consistency (a sketch)
   - The main result
3. Algorithm
   - Description
   - Inexact proximity operators
   - The main result
4. Further developments

INTRODUCTION

NONPARAMETRIC REGRESSION WITH RANDOM DESIGN



Input space $\mathcal{X} = \mathrm{range}(\text{Years of Education}) \times \mathrm{range}(\text{Seniority}) \subset \mathbb{R}^2$.
Output space $\mathcal{Y} = \mathrm{range}(\text{Income}) \subset \mathbb{R}$.
$z_n = (x_i, y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ are i.i.d. realizations of a random variable (X, Y) distributed according to P.



The distribution P describes a "fuzzy" function.
The goal is to estimate a function $f : \mathcal{X} \to \mathcal{Y}$ such that

$$(\forall x \in \mathcal{X}) \qquad f(x) \text{ is chosen according to } P(y|x)$$

For instance one could take $f = \mathrm{E}(Y|X)$, the conditional mean.

THE FORMAL SETTING

We are given

- an input space $(\mathcal{X}, \mathfrak{A})$, an output space $\mathcal{Y} \subset \mathbb{R}$ bounded interval, and a random variable (X, Y) with values in $\mathcal{X} \times \mathcal{Y}$ with distribution P.
- a sequence $(X_i, Y_i)_{i \in \mathbb{N}}$ of i.i.d. random variables taking values in $\mathcal{X} \times \mathcal{Y}$ with common distribution P.
- n observations $z_n = (x_i, y_i)_{1 \leq i \leq n}$ of the random variables $Z_n = (Y_i, Y_i)_{1 \leq i \leq n}$, which constitutes the training set.
- a (large) closed convex constraint set $\mathcal{C} \subset L^2(P_{\mathcal{X}})$ encoding prior information.

The goal is to approach $f_{\mathcal{C}}$, the regression function on $\mathcal{C}$, which is the minimizer of the *L^2-risk*

$$\begin{aligned} R\colon L^2(P_{\mathcal{X}}) &\to \mathbb{R}_+ \\ f &\mapsto \int_{\mathcal{X}\times\mathbb{R}} |y - f(x)|^2 \,\mathrm{d}P(x,y) = \mathrm{E}|Y - f(X)|^2 \end{aligned}$$

over the constraint set $\mathcal{C}$, without any knowledge of P, but using only training sets z_n.

The goal is to approach $f_{\mathcal{C}}$, the regression function on $\mathcal{C}$, which is the minimizer of the *L^2-risk*

$$R\colon L^2(P_{\mathcal{X}}) \to \mathbb{R}_+$$
$$f \mapsto \int_{\mathcal{X}\times\mathbb{R}} |y - f(x)|^2 \,\mathrm{d}P(x,y) = \mathrm{E}|Y - f(X)|^2$$

over the constraint set $\mathcal{C}$, without any knowledge of P, but using only training sets z_n.

A learning algorithm is a map

$$\Lambda\colon \bigcup_{n\in\mathbb{N}} (\mathcal{X}\times\mathcal{Y})^n \to \mathcal{C}$$

such that the estimators $\Lambda(Z_n)$ asymptotically converge to $f_{\mathcal{C}}$, meaning that $\|\Lambda(Z_n) - f_{\mathcal{C}}\|_{L^2(P_{\mathcal{X}})} \to 0$ (in probability or a.s.) as $n \to +\infty$.

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We adopt a linear model:

$$f_u = \sum_{k \in \mathbb{K}} \langle u \mid e_k \rangle \phi_k \text{ (pointwise)}, \quad u \in \ell^2(\mathbb{K})$$

where:

- $(e_k)_{k \in \mathbb{K}}$ is the canonical basis of $\ell^2(\mathbb{K})$ and $(\phi_k)_{k \in \mathbb{K}}$ is a countable dictionary of bounded functions in $\mathcal{M}(\mathcal{X}, \mathbb{R})$ such that

$$\sum_{k \in \mathbb{K}} |\phi_k(x)|^2 \leq \kappa^2 \qquad (\forall x \in \mathcal{X})$$

- the coefficients $\langle u \mid e_k \rangle$ are constrained in given intervals $C_k \subset \mathbb{R}$, thus defining the constraint

$$\mathcal{C} = \overline{\left\{ f_u \,\middle|\, u \in \ell^2(\mathbb{K}) \cap \prod_{k \in \mathbb{K}} C_k \right\}}$$

Our estimators will be in $\mathcal{C}$! Moreover, $\mathcal{H}_K = \left\{ f_u \,\middle|\, u \in \ell^2(\mathbb{K}) \right\}$ is a RKHS with feature map $\Phi \colon \mathcal{X} \to \ell^2(\mathbb{K}) \colon x \mapsto (\phi_k(x))_{k \in \mathbb{K}}$ and kernel $K(x, x') = \langle \Phi(x) \mid \Phi(x') \rangle$.

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We solve a composite regularized least squares regression problem

$$\hat{u}_{n,\lambda}(z_n) = \operatorname{argmin}_{u \in \ell^2(\mathbb{K})} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f_u(x_i))^2 + \lambda \sum_{k \in \mathbb{K}} g_k(\langle u \mid e_k \rangle) \right\}$$

$$g_k \colon \mathbb{R} \to \mathbb{R}, \qquad g_k = \iota_{C_k} + \sigma_{D_k} + \eta_k |\cdot|^r \quad (\forall\, k \in \mathbb{K})$$

where $0 \in C_k$, $D_k \subset \mathbb{R}$ closed intervals, $\eta_k \geq \eta > 0$, and $r \in \,]1, 2]$.

The learning algorithm

$$z_n \mapsto f_{\hat{u}_{n,\lambda}(z_n)}$$

The indicator function of C_k and the support function of D_k are:

$$\iota_{C_k}(\xi) = \begin{cases} 0 & \text{if } \xi \in C_k \\ +\infty & \text{if } \xi \notin C_k \end{cases} \qquad \sigma_{D_k}(\xi) = \begin{cases} \xi \sup D_k & \text{if } \xi \geq 0 \\ \xi \inf D_k & \text{if } \xi < 0 \end{cases}$$

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We solve a composite regularized least squares regression problem

$$\hat{u}_{n,\lambda}(z_n) = \operatorname{argmin}_{u \in \ell^2(\mathbb{K})} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f_u(x_i))^2 + \lambda \sum_{k \in \mathbb{K}} g_k(\langle u \mid e_k \rangle) \right\}$$

$$g_k \colon \mathbb{R} \to \mathbb{R}, \qquad g_k = \iota_{C_k} + \sigma_{D_k} + \eta_k |\cdot|^r \quad (\forall\, k \in \mathbb{K})$$

where $0 \in C_k$, $D_k \subset \mathbb{R}$ closed intervals, $\eta_k \geq \eta > 0$, and $r \in \,]1, 2]$.

The learning algorithm

$$z_n \mapsto f_{\hat{u}_{n,\lambda}(z_n)}$$

The **empirical risk** and the **regularizer**

$$\hat{R}_n(f_u) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f_u(x_i))^2 \qquad G(u) = \sum_{k \in \mathbb{K}} g_k(\langle u \mid e_k \rangle)$$

This is a regularized empirical risk minimization learning algorithm.

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We solve a composite regularized least squares regression problem

$$\hat{u}_{n,\lambda}(z_n) = \operatorname{argmin}_{u \in \ell^2(\mathbb{K})} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f_u(x_i))^2 + \lambda \sum_{k \in \mathbb{K}} g_k(\langle u \mid e_k \rangle) \right\}$$

$$g_k \colon \mathbb{R} \to \mathbb{R}, \qquad g_k = \iota_{C_k} + \sigma_{D_k} + \eta_k |\cdot|^r \quad (\forall\, k \in \mathbb{K})$$

where $0 \in C_k$, $D_k \subset \mathbb{R}$ closed intervals, $\eta_k \geq \eta > 0$, and $r \in \,]1, 2]$.

The learning algorithm

$$z_n \mapsto f_{\hat{u}_{n,\lambda}(z_n)}$$

This model encompass:

- ridge regression (Hoerl and Kennard, '70) $g_k = |\cdot|^2$
- elastic net (Zou, Hastie, '05; De Mol et al., '09) $g_k = \omega_k |\cdot| + \eta |\cdot|^2$
- bridge regression (Frank and Friedman '93) $g_k = |\cdot|^r$, $1 < r < 2$.

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We minimize the regularized empirical risk by proximal gradient algorithms: Let $0 < \gamma < \lambda/(2\kappa^2)$ and define $(u_m)_{m\in\mathbb{N}}$

> for all $m \in \mathbb{N}$ and all $k \in \mathbb{K}$
>
> $$\langle u_{m+1} \,|\, e_k \rangle = \mathrm{prox}_{\gamma g_k}\left(\langle u_m \,|\, e_k \rangle - \frac{2\gamma}{n\lambda} \sum_{i=1}^{n} \big(f_{u_m}(x_i) - y_i\big)\phi_k(x_i) \right)$$

where $\mathrm{prox}_{\gamma g_k}(\xi) = \mathrm{argmin}_{t\in\mathbb{R}} \left\{ \gamma g_k(t) + (1/2)(t-\xi)^2 \right\}$.

THE LEARNING ALGORITHM: HOW WE BUILD ESTIMATORS

We minimize the regularized empirical risk by proximal gradient algorithms: Let $0 < \gamma < \lambda/(2\kappa^2)$ and define $(u_m)_{m\in\mathbb{N}}$

> for all $m \in \mathbb{N}$ and all $k \in \mathbb{K}$
>
> $$\langle u_{m+1} \,|\, e_k \rangle = \operatorname{prox}_{\gamma g_k}\left(\langle u_m \,|\, e_k \rangle - \frac{2\gamma}{n\lambda} \sum_{i=1}^{n} \left(f_{u_m}(x_i) - y_i \right) \phi_k(x_i) \right)$$

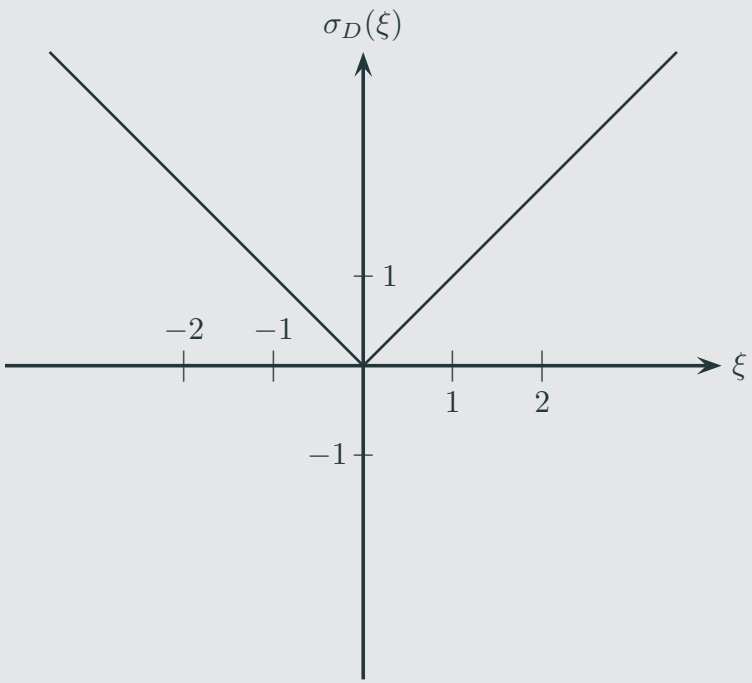
where $\operatorname{prox}_{\gamma g_k}(\xi) = \operatorname{argmin}_{t\in\mathbb{R}} \left\{ \gamma g_k(t) + (1/2)(t-\xi)^2 \right\}$.

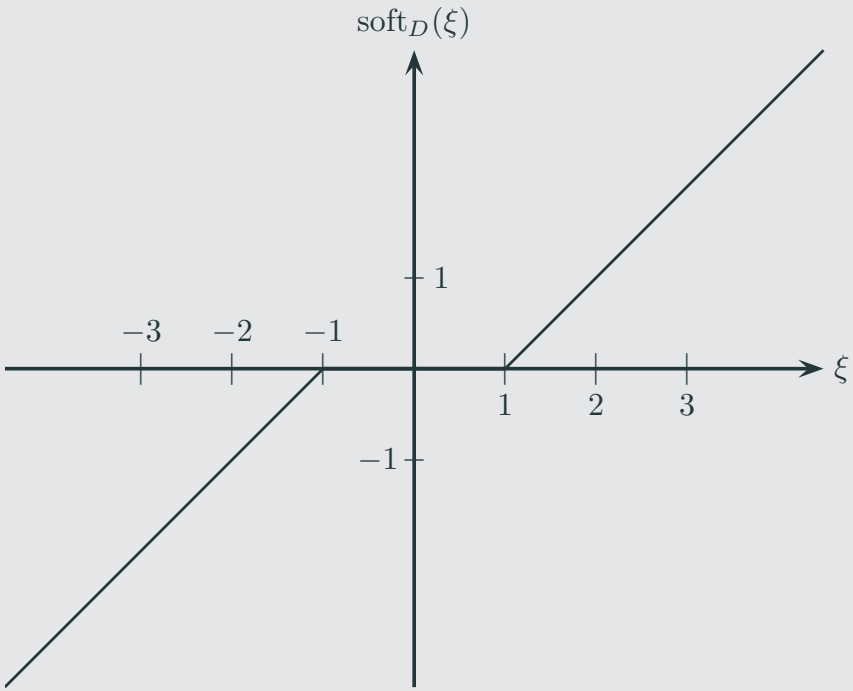
Example

If $g_k = \iota_{C_k}$, $\operatorname{prox}_{g_k} = \pi_{C_k}$. If $g_k = \sigma_{D_k}$, then

$$\operatorname{prox}_{g_k}(\xi) = \operatorname{soft}_{D_k}(\xi) = \begin{cases} \xi - \inf D_k & \text{if } \xi \le \inf D_k \\ 0 & \text{if } \xi \in D_k \\ \xi + \sup D_k & \text{if } \xi \ge \inf D_k \end{cases}$$

Example





$$D = [-1, 1], \quad \sigma_D = |\cdot|$$

THE LEARNING ALGORITHM: MOTIVATION

When the regression function has sparse representation a further goal besides prediction is to identify the set of corresponding relevant features.

To this respect in our learning algorithm:

- estimators have finite support (sparsity is promoted).

THE LEARNING ALGORITHM: MOTIVATION

When the regression function has sparse representation a further goal besides prediction is to identify the set of corresponding relevant features.

To this respect in our learning algorithm:

- estimators have finite support (sparsity is promoted).
- positiveness and boundedness of the coefficients can be enforced.

THE LEARNING ALGORITHM: MOTIVATION

When the regression function has sparse representation a further goal besides prediction is to identify the set of corresponding relevant features.

To this respect in our learning algorithm:

- estimators have finite support (sparsity is promoted).
- positiveness and boundedness of the coefficients can be enforced.
- estimators keep the same grouping selection properties of the elastic-net estimators, but they are possibly not affected by the double shrinkage phenomenon (Zou and Hastie '05).

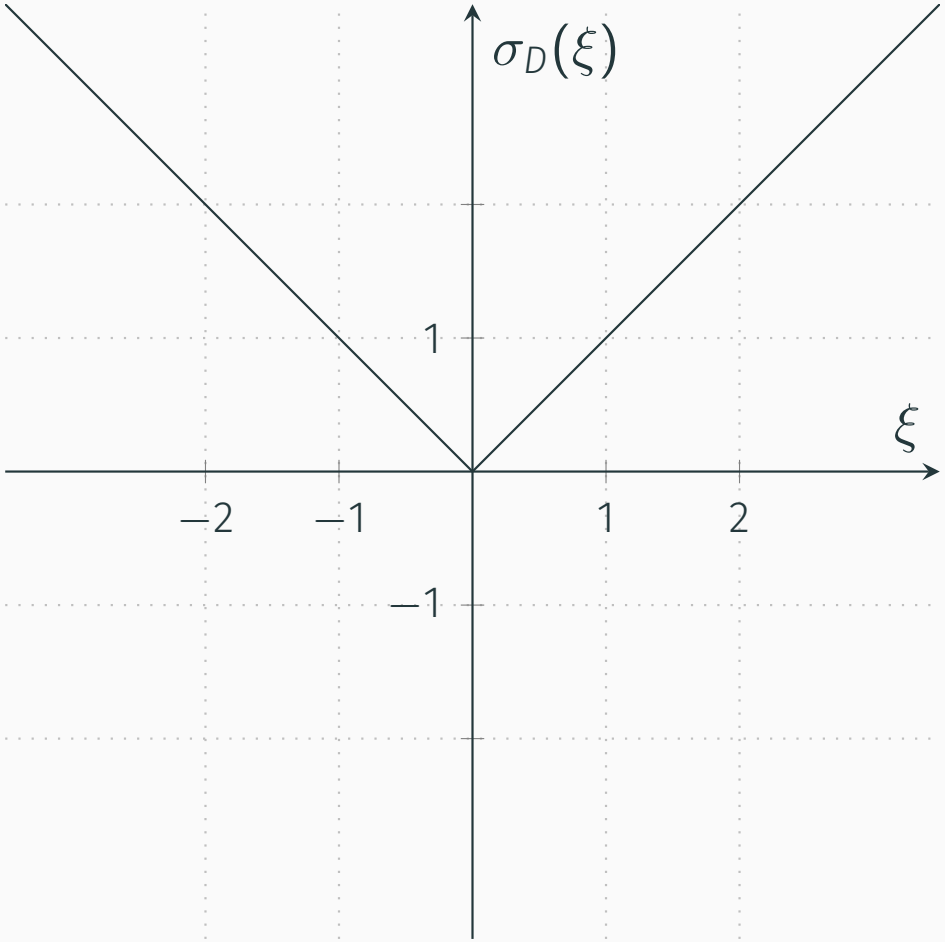
THE LEARNING ALGORITHM: MOTIVATION

When the regression function has sparse representation a further goal besides prediction is to identify the set of corresponding relevant features.

To this respect in our learning algorithm:

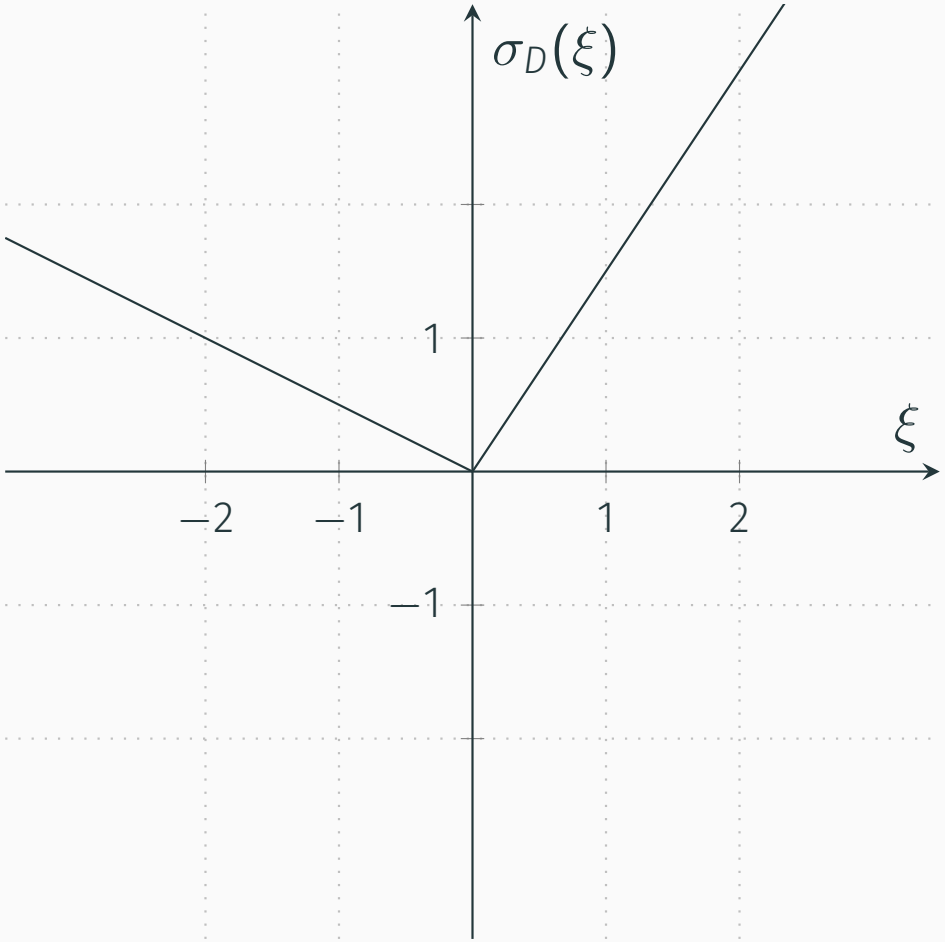
- estimators have finite support (sparsity is promoted).
- positiveness and boundedness of the coefficients can be enforced.
- estimators keep the same grouping selection properties of the elastic-net estimators, but they are possibly not affected by the double shrinkage phenomenon (Zou and Hastie '05).
- the regularizer gives a high flexibility to control the "shape" of the thresholding operation in dependence of the available prior information.

THE LEARNING ALGORITHM: MOTIVATION



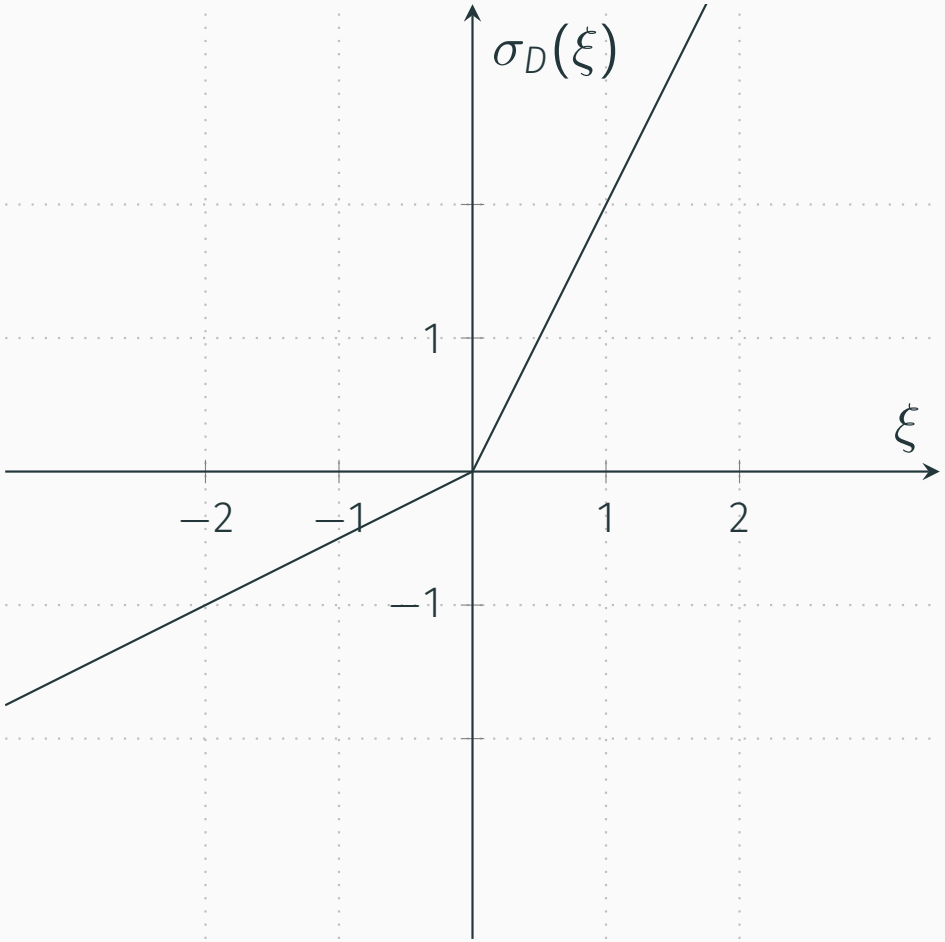
$$D = [-1, 1]$$

THE LEARNING ALGORITHM: MOTIVATION



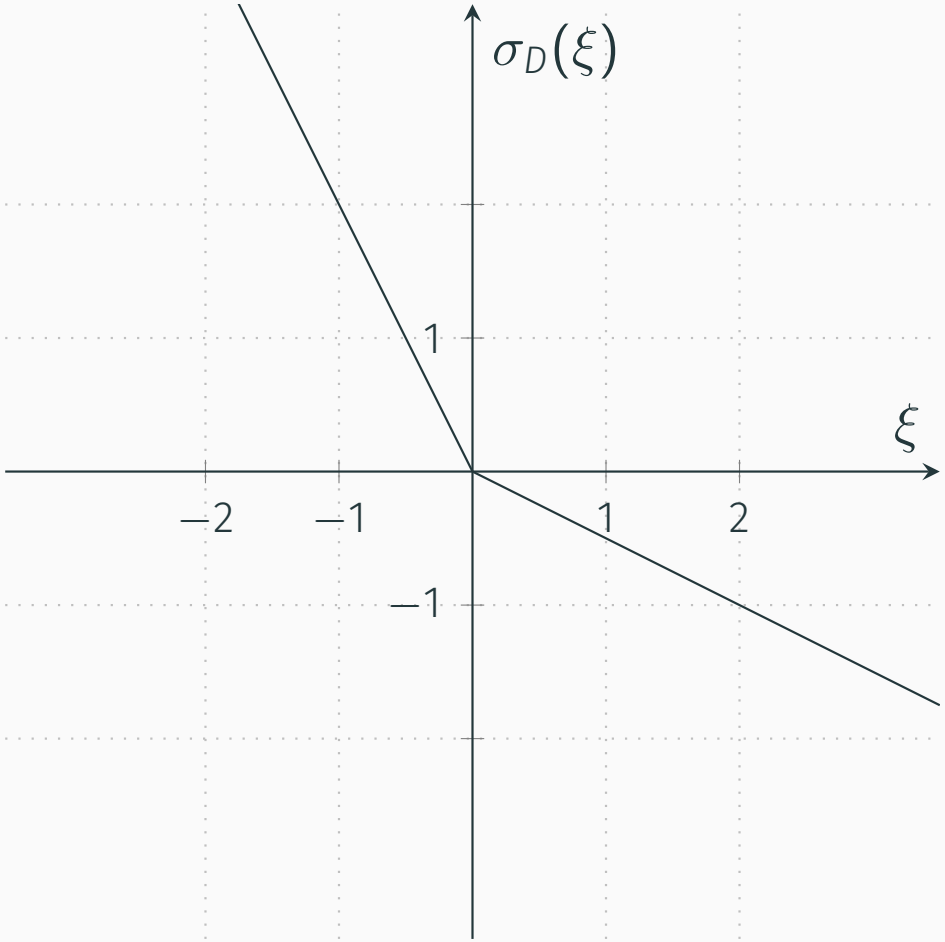
$$D = [-0.5, 1.5]$$

THE LEARNING ALGORITHM: MOTIVATION



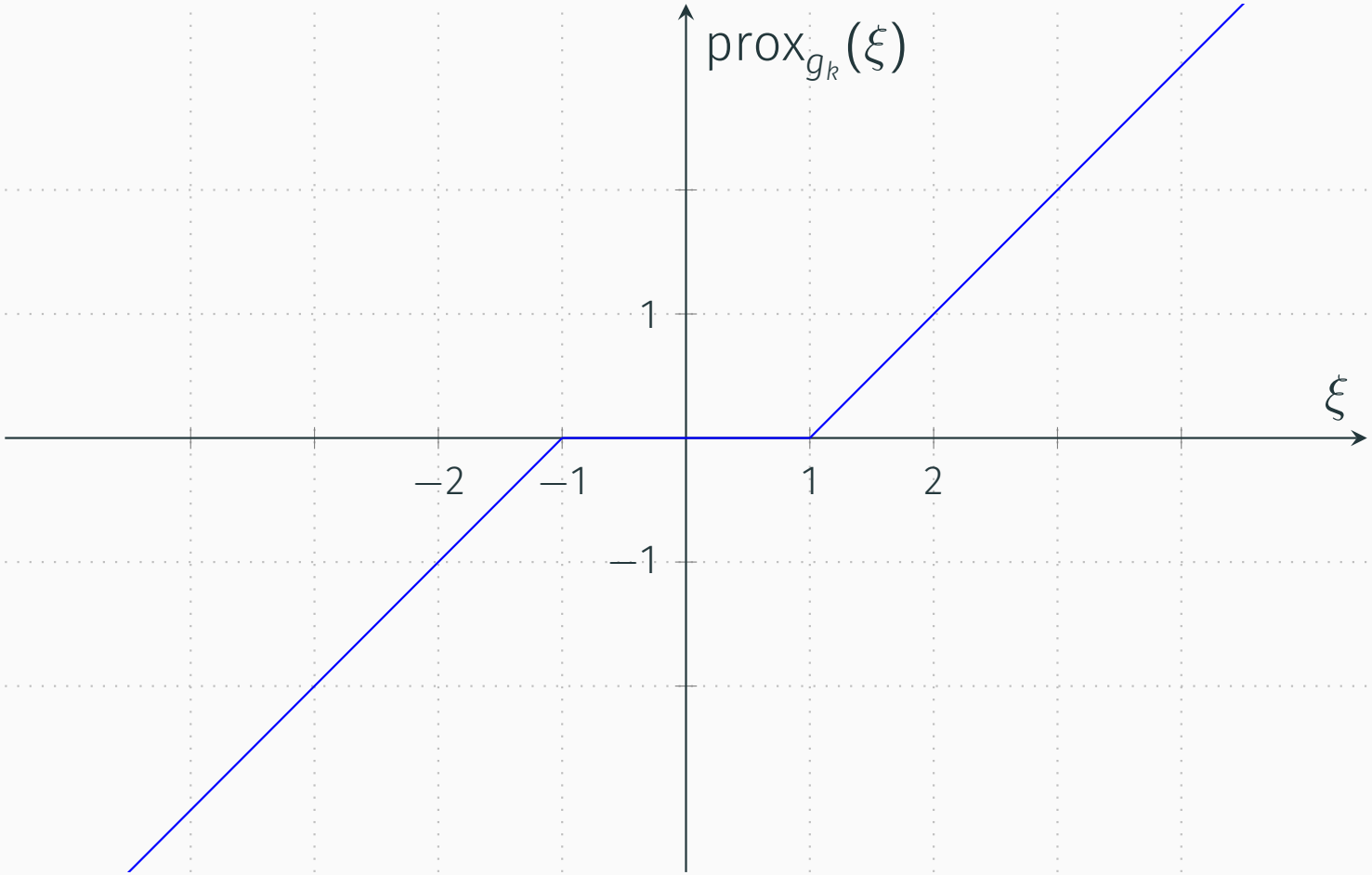
$$D = [0.5, 2]$$

THE LEARNING ALGORITHM: MOTIVATION



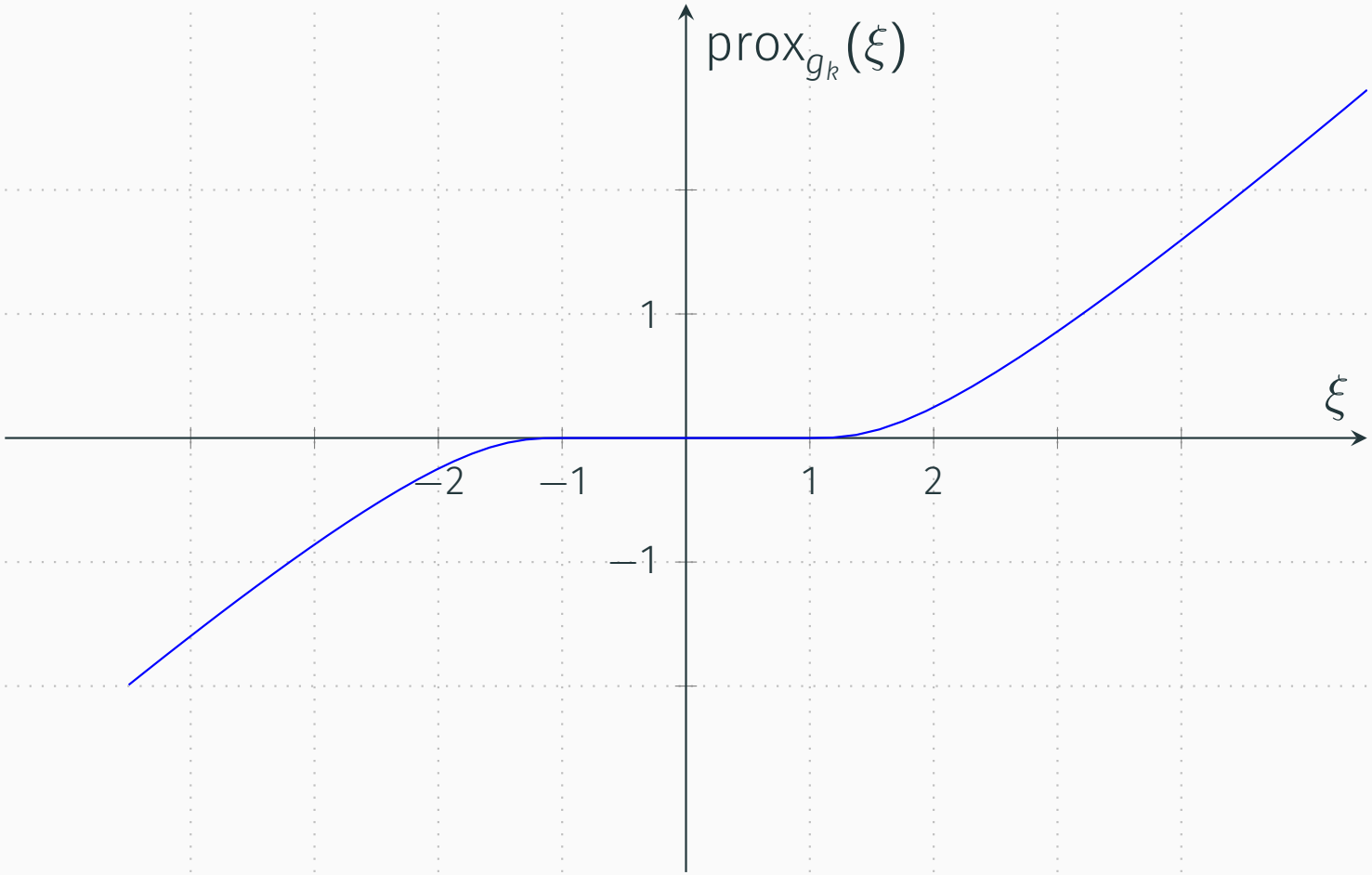
$$D = [-2, -0.5]$$

THE LEARNING ALGORITHM: MOTIVATION



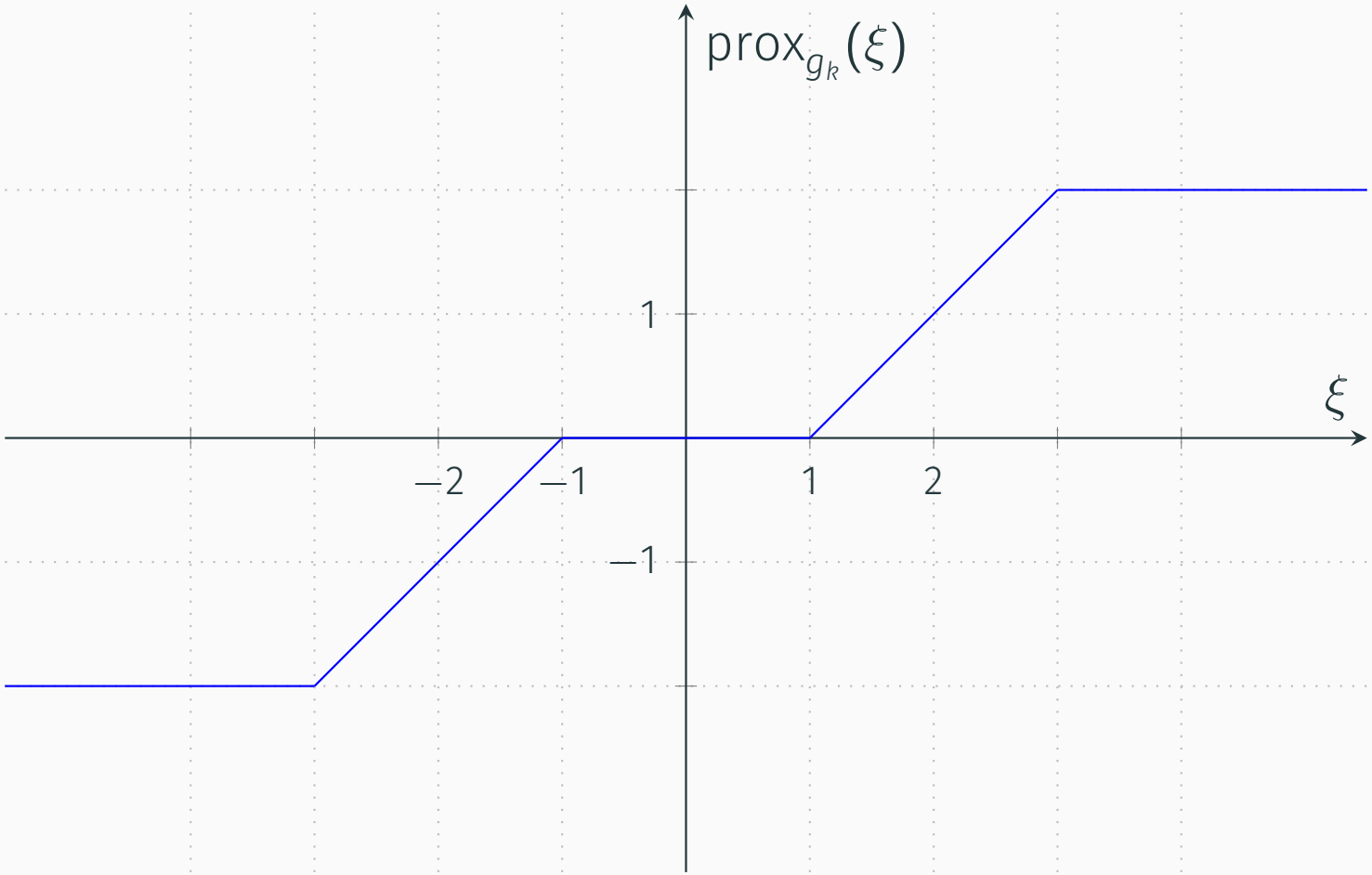
$$C_k = \mathbb{R}, D_k = [-1, 1], \eta_k = 0$$

THE LEARNING ALGORITHM: MOTIVATION



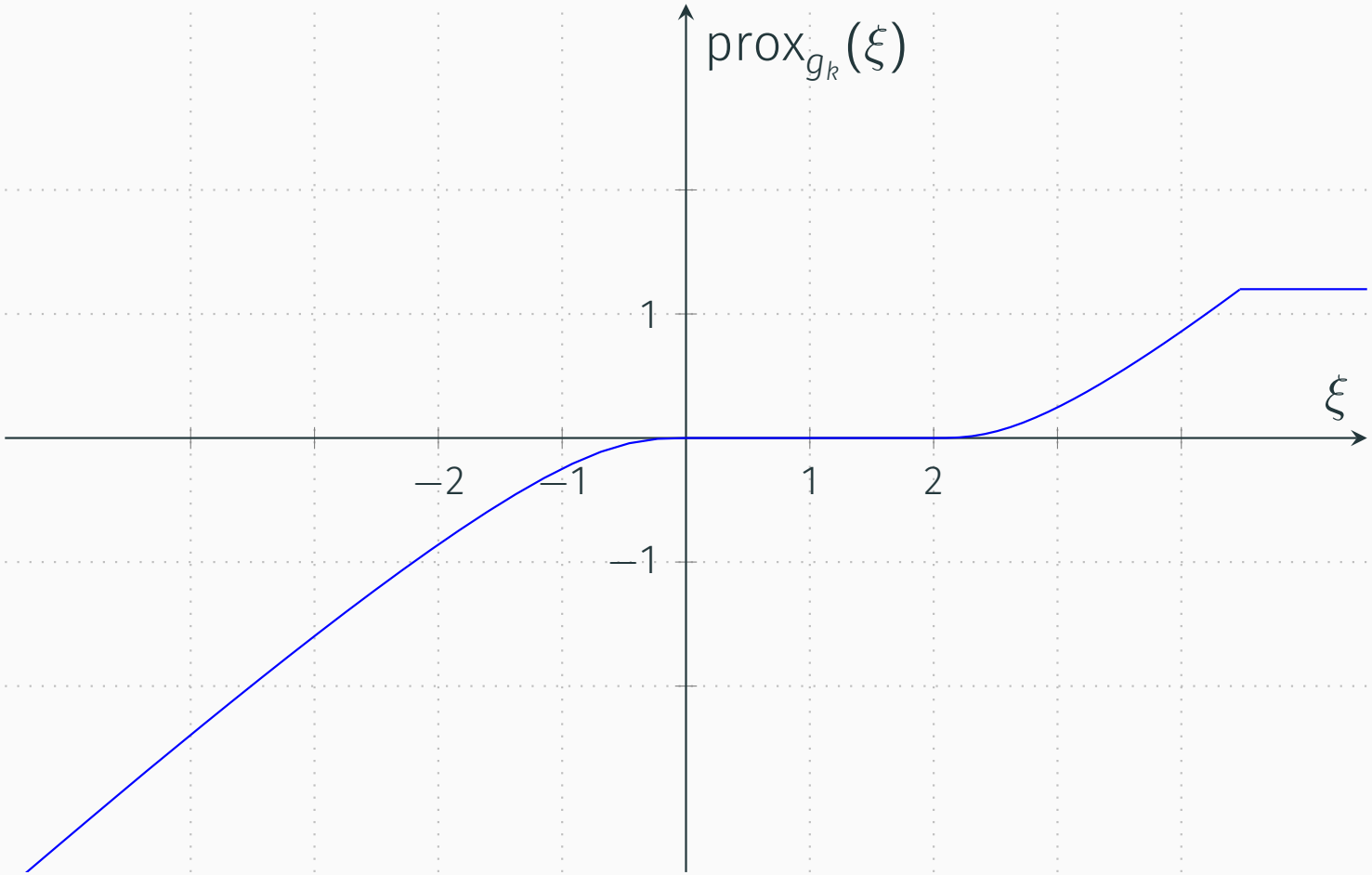
$$C_k = \mathbb{R}, D_k = [-1, 1], \eta_k = 0.9, r = 4/3$$

THE LEARNING ALGORITHM: MOTIVATION



$$C_k = [-2, 2], D_k = [-1, 1], \eta_k = 0$$

THE LEARNING ALGORITHM: MOTIVATION



$$C_k =]-\infty, 6/5]\,, D_k = [0, 2], \eta_k = 0.9, r = 4/3$$

OUR CONTRIBUTION

We will address two problems:

- consistency of the estimators $f_{\hat{u}_{n,\lambda}(Z_n)}$; meaning that, we can find conditions on a vanishing sequence $(\lambda_n)_{n\in\mathbb{N}}$ such that

$$\|f_{\hat{u}_{n,\lambda_n}(Z_n)} - f_{\mathcal{C}}\|_{L^2(P_{\mathcal{X}})} \to 0 \qquad \text{(in probability or a.s.)}$$

- computation of the estimators $\hat{u}_{n,\lambda}(z_n)$ (for a fixed training set z_n and parameter λ) by mean of an inexact and accelerated forward-backward splitting algorithm.

STATISTICAL ANALYSIS

THE REGRESSION FUNCTION WRT THE CONSTRAINT SET

$f_{\mathcal{C}}$ is the minimizer of the risk over the constraint set $\mathcal{C}$.

Facts

1. Setting $f_* = \mathrm{E}(Y|X)$, we have $R(f) = \|f - f_*\|_{L^2} + \inf_{L^2(P_{\mathcal{X}})} R$

THE REGRESSION FUNCTION WRT THE CONSTRAINT SET

$f_{\mathcal{C}}$ is the minimizer of the risk over the constraint set $\mathcal{C}$.

Facts

1. Setting $f_* = \mathrm{E}(Y|X)$, we have $R(f) = \|f - f_*\|_{L^2} + \inf_{L^2(P_{\mathcal{X}})} R$
2. $f_{\mathcal{C}}$ is the (orthogonal) projection of f_* onto $\mathcal{C}$ in $L^2(\mathcal{X})$;

THE REGRESSION FUNCTION WRT THE CONSTRAINT SET

$f_{\mathcal{C}}$ is the minimizer of the risk over the constraint set $\mathcal{C}$.

Facts

1. Setting $f_* = \mathrm{E}(Y|X)$, we have $R(f) = \|f - f_*\|_{L^2} + \inf_{L^2(P_{\mathcal{X}})} R$
2. $f_{\mathcal{C}}$ is the (orthogonal) projection of f_* onto $\mathcal{C}$ in $L^2(\mathcal{X})$;
3. $(\forall f \in \mathcal{C}) \quad \|f - f_{\mathcal{C}}\|_{L^2}^2 \leq R(f) - \inf_{\mathcal{C}} R$;

THE REGRESSION FUNCTION WRT THE CONSTRAINT SET

$f_{\mathcal{C}}$ is the minimizer of the risk over the constraint set $\mathcal{C}$.

Facts

1. Setting $f_* = \mathrm{E}(Y|X)$, we have $R(f) = \|f - f_*\|_{L^2} + \inf_{L^2(P_{\mathcal{X}})} R$
2. $f_{\mathcal{C}}$ is the (orthogonal) projection of f_* onto $\mathcal{C}$ in $L^2(\mathcal{X})$;
3. $(\forall f \in \mathcal{C}) \quad \|f - f_{\mathcal{C}}\|_{L^2}^2 \leq R(f) - \inf_{\mathcal{C}} R$;
4. $(\forall f \in \mathcal{C})$

$$R(f) - \inf_{\mathcal{C}} R \leq$$
$$2\left[\left(\|f - f_{\mathcal{C}}\|_{L^2} + \sqrt{\inf_{\mathcal{C}} R - \inf_{L^2(\mathcal{X})} R}\right)^2 + \inf_{L^2(\mathcal{X})} R\right]^{1/2} \|f - f_{\mathcal{C}}\|_{L^2}$$

THE REGRESSION FUNCTION WRT THE CONSTRAINT SET

$f_{\mathcal{C}}$ is the minimizer of the risk over the constraint set $\mathcal{C}$.

Facts

1. Setting $f_* = \mathrm{E}(Y|X)$, we have $R(f) = \|f - f_*\|_{L^2} + \inf_{L^2(P_{\mathcal{X}})} R$
2. $f_{\mathcal{C}}$ is the (orthogonal) projection of f_* onto $\mathcal{C}$ in $L^2(\mathcal{X})$;
3. $(\forall f \in \mathcal{C}) \quad \|f - f_{\mathcal{C}}\|_{L^2}^2 \leq R(f) - \inf_{\mathcal{C}} R$;
4. $(\forall f \in \mathcal{C})$

$$R(f) - \inf_{\mathcal{C}} R \leq$$
$$2\Big[\Big(\|f - f_{\mathcal{C}}\|_{L^2} + \sqrt{\inf_{\mathcal{C}} R - \inf_{L^2(\mathcal{X})} R}\Big)^2 + \inf_{L^2(\mathcal{X})} R\Big]^{1/2} \|f - f_{\mathcal{C}}\|_{L^2}$$

It follows that

$$R(\hat{f}_n) - \inf_{\mathcal{C}} R \to 0 \iff \|\hat{f}_n - f_{\mathcal{C}}\|_{L^2(P_{\mathcal{X}})} \to 0\,.$$

Two minimization problems

$$u_\lambda = \mathrm{argmin}_{u \in \ell^2(\mathbb{K})} \left(R(f_u) + \lambda G(u) \right)$$
$$\hat{u}_{n,\lambda}(Z_n) = \mathrm{argmin}_{u \in \ell^2(\mathbb{K})} \left(\hat{R}_n(f_u) + \lambda G(u) \right)$$

Then:

$$\begin{aligned} \|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{\mathcal{C}}\|_{L^2} &\leq \|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{u_\lambda}\|_{L^2} + \|f_{u_\lambda} - f_{\mathcal{C}}\|_{L^2} \\ &\leq \kappa \|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r + \left(R(f_{u_\lambda}) - \inf_{\mathcal{C}} R \right)^{1/2} \end{aligned}$$

We have to control:

- the stochastic term: $\|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r$
- the deterministic term: $R(f_{u_\lambda}) - \inf_{\mathcal{C}} R$

Two minimization problems

$$u_\lambda = \mathrm{argmin}_{u\in\ell^2(\mathbb{K})}\left(R(f_u) + \lambda G(u)\right)$$
$$\hat{u}_{n,\lambda}(Z_n) = \mathrm{argmin}_{u\in\ell^2(\mathbb{K})}\left(\hat{R}_n(f_u) + \lambda G(u)\right)$$

Then:

$$\begin{aligned}\|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{\mathcal{C}}\|_{L^2} &\leq \|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{u_\lambda}\|_{L^2} + \|f_{u_\lambda} - f_{\mathcal{C}}\|_{L^2} \\ &\leq \kappa\|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r + \left(R(f_{u_\lambda}) - \inf_{\mathcal{C}} R\right)^{1/2}\end{aligned}$$

We have to control:

- the stochastic term: $\|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r$
- the deterministic term: $R(f_{u_\lambda}) - \inf_{\mathcal{C}} R = R(f_{u_\lambda}) - \inf_{u\in\mathrm{dom}\, G} R(f_u)$ tends to zero as $\lambda \to 0$ because of a general variational principle (Attouch '96).

Two minimization problems

$$u_\lambda = \mathrm{argmin}_{u\in\ell^2(\mathbb{K})}\left(R(f_u) + \lambda G(u)\right)$$
$$\hat{u}_{n,\lambda}(Z_n) = \mathrm{argmin}_{u\in\ell^2(\mathbb{K})}\left(\hat{R}_n(f_u) + \lambda G(u)\right)$$

Then:

$$\begin{aligned}\|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{\mathcal{C}}\|_{L^2} &\leq \|f_{\hat{u}_{n,\lambda}(Z_n)} - f_{u_\lambda}\|_{L^2} + \|f_{u_\lambda} - f_{\mathcal{C}}\|_{L^2} \\ &\leq \kappa\|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r + \left(R(f_{u_\lambda}) - \inf_{\mathcal{C}} R\right)^{1/2}\end{aligned}$$

We have to control:

- the stochastic term: $\|\hat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r$?
- the deterministic term: $R(f_{u_\lambda}) - \inf_{\mathcal{C}} R = R(f_{u_\lambda}) - \inf_{u\in\mathrm{dom}\, G} R(f_u)$ tends to zero as $\lambda \to 0$ because of a general variational principle (Attouch '96).

Lemma ((Xu and Roach '91) Total convexity on bounded sets)

Let $\rho > 0$, let $u_0 \in \ell^r(\mathbb{K})$, and $u_0^* \in \partial G(u_0)$. Then there exists a universal constant $M > 0$ (depending only on r) such that

$$(\forall\, u \in \ell^r) \quad G(u) - G(u_0) \geq \langle u - u_0 \mid u_0^* \rangle + \eta M \frac{\|u - u_0\|_r^2}{(\|u_0\|_r + \|u - u_0\|_r)^{2-r}}\,.$$

Theorem (Combettes, Salzo and Villa '15)

There exists a measurable function $\Psi_\lambda \colon \mathcal{X} \times \mathcal{Y} \to \ell^2(\mathbb{K})$, which is of type $\Psi_\lambda(x, y) = h_\lambda(x, y)\Phi(x)$, such that

$$-\mathrm{E}_P[\Psi_\lambda] \in \lambda \partial G(u_\lambda),$$

$\|\Psi_\lambda\|_\infty \leq C\|u_\lambda\|_2$, $\|\Psi_\lambda\|_2 \leq 2\kappa(R(f_{u_\lambda}))^{1/2}$, *and* $\forall\, z_n \in (\mathcal{X} \times \mathcal{Y})^n$

$$\eta M \frac{\|\hat{u}_{n,\lambda}(z_n) - u_\lambda\|_r}{(\|u_\lambda\|_r + \|\hat{u}_{n,\lambda}(z_n) - u_\lambda\|_r)^{2-r}} \leq \frac{1}{\lambda}\Big\|\frac{1}{n}\sum_{i=1}^{n} \Psi_\lambda(x_i, y_i) - \mathrm{E}_P[\Psi_\lambda]\Big\|_2.$$

$$P\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\Psi_\lambda(X_i,Y_i)-\mathrm{E}[\Psi_\lambda(X,Y)]\right\|_2\leq\delta(n,\lambda,\tau)\right]\geq 1-e^{-\tau}$$

Theorem (Combettes, Salzo and Villa '15)

Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $]0,+\infty[$ converging to 0. Then the following holds:

1. *If $1/(\lambda_n^{2/r}n^{1/2})\to 0$, then $\|f_{\hat{u}_{n,\lambda}(Z_n)}-f_{\mathcal{C}}\|_{L^2}\to 0$ in probability.*
2. *If $\log n/(\lambda_n^{2/r}n^{1/2})\to 0$, then $\|f_{\hat{u}_{n,\lambda}(Z_n)}-f_{\mathcal{C}}\|_{L^2}\to 0$ P-a.s.*

$$P\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\Psi_\lambda(X_i,Y_i)-\mathrm{E}[\Psi_\lambda(X,Y)]\right\|_2\leq\delta(n,\lambda,\tau)\right]\geq 1-e^{-\tau}$$

Theorem (Combettes, Salzo and Villa '15)

Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $]0,+\infty[$ converging to 0. Then the following holds:

1. *If $1/(\lambda_n^{2/r}n^{1/2})\to 0$, then $\|f_{\hat{u}_{n,\lambda}(Z_n)}-f_{\mathcal{C}}\|_{L^2}\to 0$ in probability.*
2. *If $\log n/(\lambda_n^{2/r}n^{1/2})\to 0$, then $\|f_{\hat{u}_{n,\lambda}(Z_n)}-f_{\mathcal{C}}\|_{L^2}\to 0$ P-a.s.*
3. *Suppose $f_{\mathcal{C}}\in\{f_u\mid u\in\operatorname{dom}G\}$ and set $S=\operatorname{argmin}_{u\in\operatorname{dom}G}R(f_u)$. Then, there exists a unique $u^\dagger\in S$ which minimizes G over S and $f_{u^\dagger}=f_{\mathcal{C}}$. Moreover:*
   - *if $1/(\lambda_n n^{1/2})\to 0$, then $\|\hat{u}_{n,\lambda_n}(Z_n)-u^\dagger\|_r\to 0$ in probability;*
   - *if $\log n/(\lambda_n n^{1/2})\to 0$, then $\|\hat{u}_{n,\lambda_n}(Z_n)-u^\dagger\|_r\to 0$ P-a.s.*

ALGORITHM

THE MINIMIZATION PROBLEM

We aim at minimizing the functional

$$u \in \ell^2(\mathbb{K}) \mapsto \frac{1}{n}\sum_{i=1}^{n}(y_i - f_u(x_i))^2 + \lambda \sum_{k\in\mathbb{K}} g_k(\langle u \mid e_k\rangle)$$

$$g_k \colon \mathbb{R} \to \mathbb{R}, \qquad g_k = \iota_{C_k} + \sigma_{D_k} + \eta_k|\cdot|^r \quad (\forall\, k \in \mathbb{K})$$

which is the sum of a:

- differentiable term $u \mapsto \hat{R}_n(f_u)$ with Lipschitz continuous gradient.
- lower semicontinuous, convex and separable term $\lambda G(u) = \lambda \sum_{k\in\mathbb{K}} g_k(\langle u \mid e_k\rangle)$.

Note that we need convergence in objective values.

THE MINIMIZATION PROBLEM

We aim at minimizing the functional

$$u \in \ell^2(\mathbb{K}) \mapsto \frac{1}{n}\sum_{i=1}^{n}(y_i - f_u(x_i))^2 + \lambda \sum_{k\in\mathbb{K}} g_k(\langle u \mid e_k\rangle)$$

$$g_k \colon \mathbb{R} \to \mathbb{R}, \qquad g_k = \iota_{C_k} + \sigma_{D_k} + \eta_k|\cdot|^r \quad (\forall\, k \in \mathbb{K})$$

which is the sum of a:

- differentiable term $u \mapsto \hat{R}_n(f_u)$ with Lipschitz continuous gradient.
- lower semicontinuous, convex and separable term $\lambda G(u) = \lambda \sum_{k\in\mathbb{K}} g_k(\langle u \mid e_k\rangle)$.

Note that we need convergence in objective values.

This problem can be solved by forward-backward splitting methods. These algorithms have been studied for convex separable regularizers in (Combettes and Pesquet '07; Bredies and Lorenz '08).

AN INEXACT ACCELERATED FBS ALGORITHM

Fix $u_0 = v_0 \in \ell^2(\mathbb{K})$, $\tau_0 = 1$ and define for every $m \in \mathbb{N}$,

$$
\begin{aligned}
&\text{for all } k \in \mathbb{K} \\
&\langle w_m \mid e_k \rangle = \langle v_m \mid e_k \rangle - \frac{2\gamma}{n\lambda} \sum_{i=1}^{n} \big(f_{v_m}(x_i) - y_i \big) \phi_k(x_i) \\
&\langle u_{m+1} \mid e_k \rangle = \operatorname{prox}_{\gamma g_k}(\langle w_m \mid e_k \rangle) \\
\tau_{m+1} &= \frac{1}{2}\Big(1 + \sqrt{1 + 4\tau_m^2}\Big) \\
v_{m+1} &= u_{m+1} + \frac{\tau_m - 1}{\tau_{m+1}}(u_{m+1} - u_m)
\end{aligned}
$$

(Nesterov '83; Guler '92; Beck and Teboulle '09; Villa, Salzo et. al. '13)

AN INEXACT ACCELERATED FBS ALGORITHM

Fix $u_0 = v_0 \in \ell^2(\mathbb{K})$, $\tau_0 = 1$ and define for every $m \in \mathbb{N}$,

$$
\begin{aligned}
&\text{for all } k \in \mathbb{K} \\
&\langle w_m \mid e_k \rangle = \langle v_m \mid e_k \rangle - \frac{2\gamma}{n\lambda} \sum_{i=1}^{n} \big(f_{v_m}(x_i) - y_i\big)\phi_k(x_i) \\
&\langle u_{m+1} \mid e_k \rangle = \pi_{C_k}\Big(\mathrm{prox}_{\gamma\eta_k|\cdot|^r}\big(\mathrm{soft}_{\gamma D_k}\langle w_m \mid e_k \rangle\big)\Big)
\end{aligned}
$$

$$
\tau_{m+1} = \frac{1}{2}\Big(1 + \sqrt{1 + 4\tau_m^2}\Big)
$$

$$
v_{m+1} = u_{m+1} + \frac{\tau_m - 1}{\tau_{m+1}}(u_{m+1} - u_m)
$$

(Nesterov '83; Guler '92; Beck and Teboulle '09; Villa, Salzo et. al. '13)

AN INEXACT ACCELERATED FBS ALGORITHM

Fix $u_0 = v_0 \in \ell^2(\mathbb{K})$, $\tau_0 = 1$ and define for every $m \in \mathbb{N}$,

for all $k \in \mathbb{K}$

$$\langle w_m \mid e_k \rangle = \langle v_m \mid e_k \rangle - \frac{2\gamma}{n\lambda} \sum_{i=1}^{n} \big(f_{v_m}(x_i) - y_i\big)\phi_k(x_i)$$

$$\langle u_{m+1} \mid e_k \rangle = \pi_{C_k}\Big(\mathrm{prox}_{\gamma\eta_k|\cdot|^r}\big(\mathrm{soft}_{\gamma D_k}\langle w_m \mid e_k \rangle\big) + \alpha_{m,k}\Big)$$

$$\tau_{m+1} = \frac{1}{2}\Big(1 + \sqrt{1 + 4\tau_m^2}\Big)$$

$$v_{m+1} = u_{m+1} + \frac{\tau_m - 1}{\tau_{m+1}}(u_{m+1} - u_m)$$

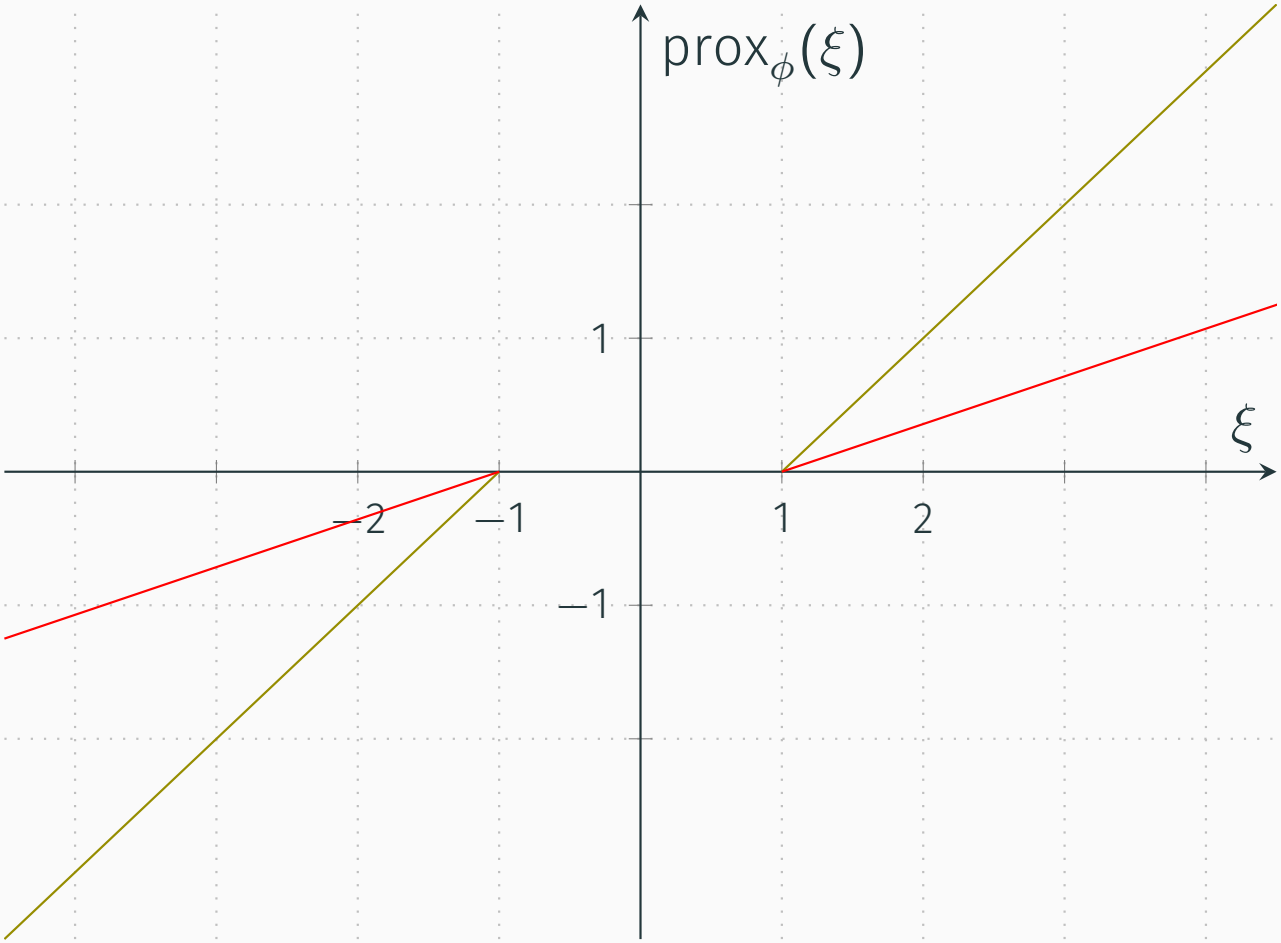
(Nesterov '83; Guler '92; Beck and Teboulle '09; Villa, Salzo et. al. '13)

where:

- $0 < \gamma < \lambda/(2\kappa^2)$
- $(\alpha_{m,k})_{(m,k)\in\mathbb{N}\times\mathbb{K}}$ is a double sequence in $\mathbb{R}$ that accounts for errors in the computation of the proximity operator $\mathrm{prox}_{\gamma\eta_k|\cdot|^r}$ and, for some $c > 0$ and $p \in]3/2, +\infty[$ and $(\nu_k)_{\in\mathbb{K}} \in \ell^1(\mathbb{K})$, satisfies the condition

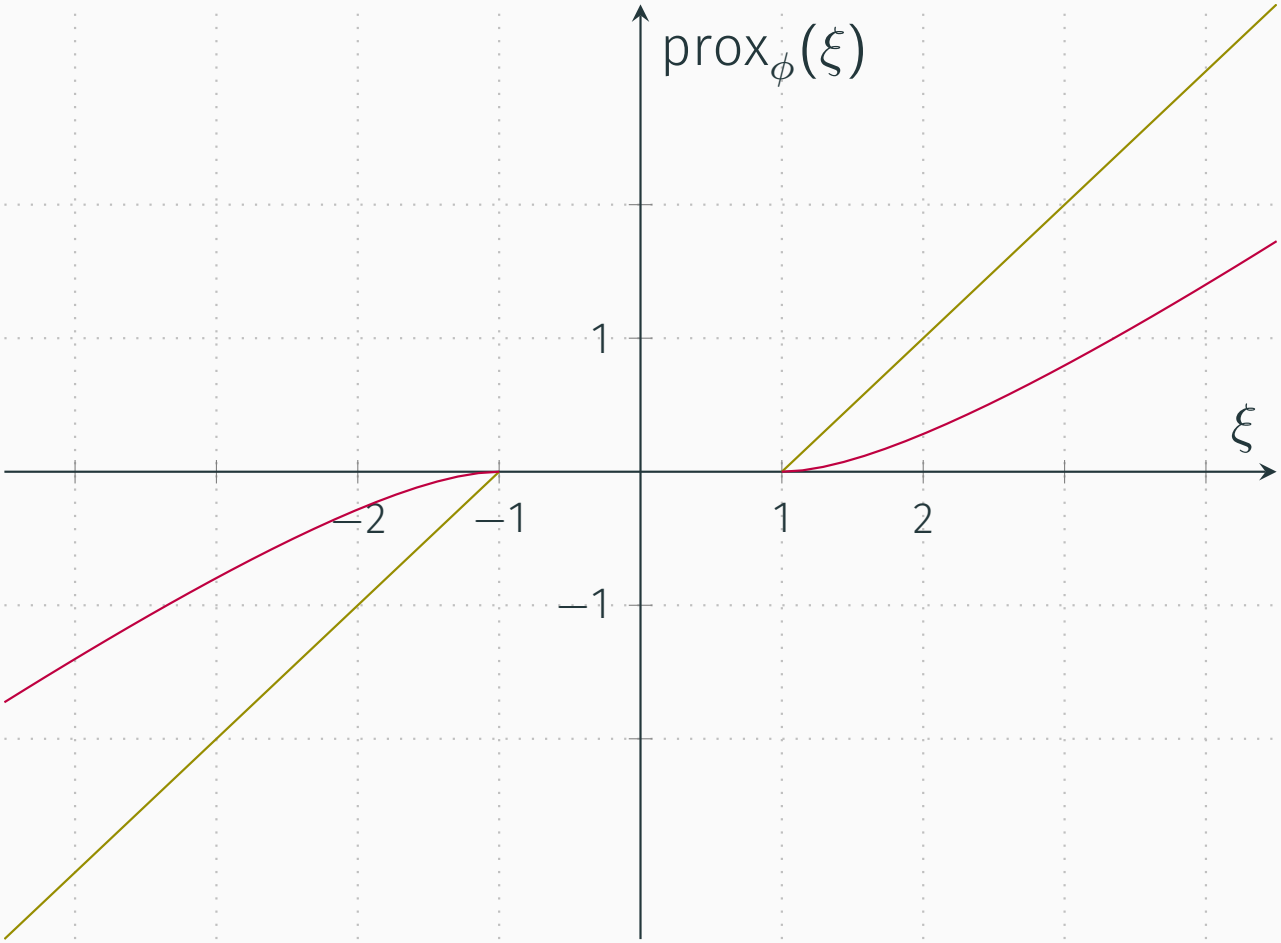
$$|\alpha_{k,m}| \leq \min\left\{\frac{cm^{-2p}\nu_k}{2\gamma\eta_k r(|\langle w_m \mid e_k\rangle| + 1)^{r-1} + 2|\langle w_m \mid e_k\rangle| + 1}, \frac{|\mathrm{soft}_{\gamma D_k}(\langle w_m \mid e_k\rangle)|}{1 + r\gamma\eta_k}, \left(\frac{|\mathrm{soft}_{\gamma D_k}(\langle w_m \mid e_k\rangle)|}{1 + r\gamma\eta_k}\right)^{1/(r-1)}\right\}.$$

PROXIMAL THRESHOLDING: THE ROLE OF THE EXPONENT r



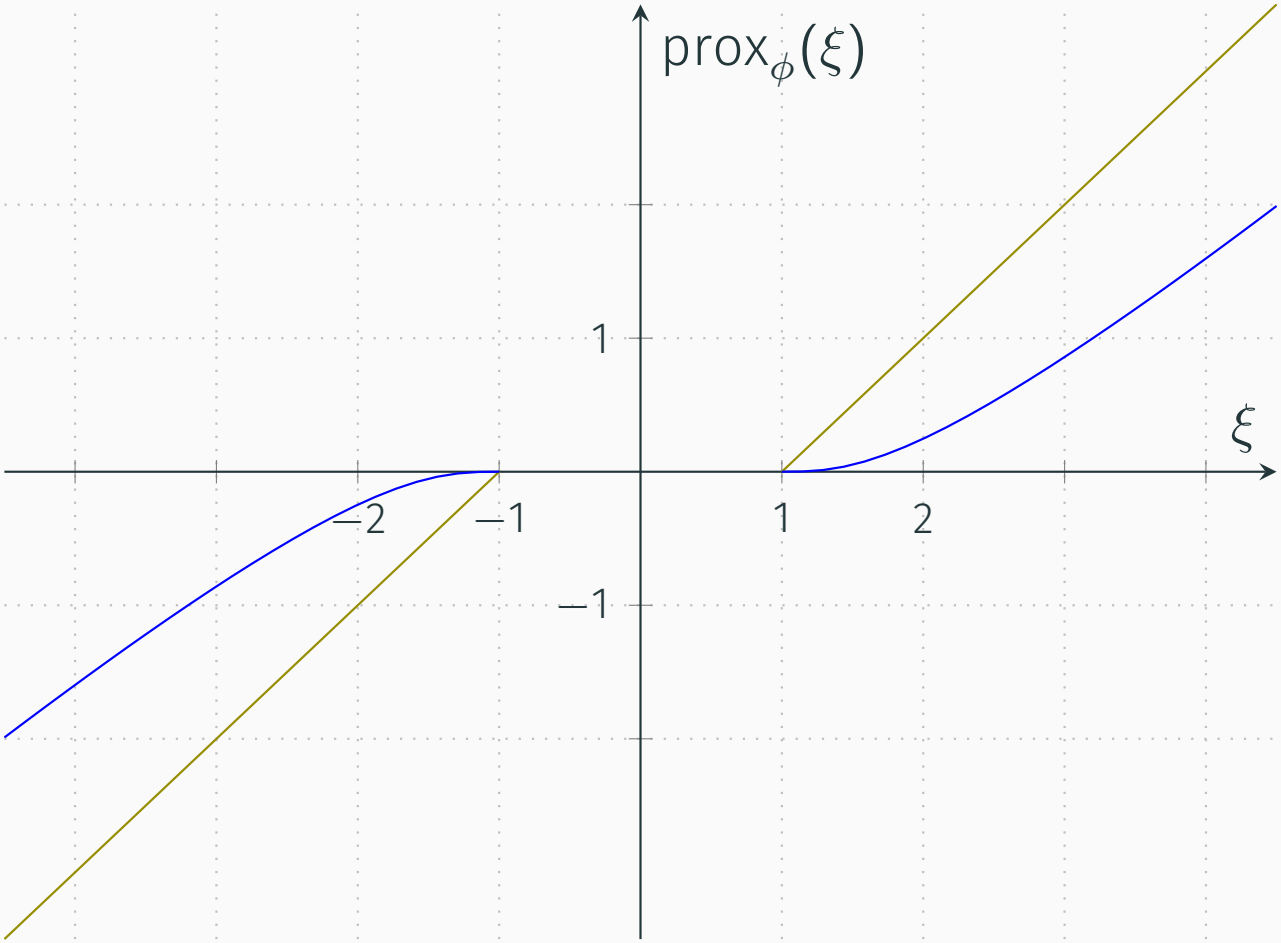
Soft thresholding (olive) and prox of $\phi = |\cdot| + 0.9|\cdot|^2$ (red).

PROXIMAL THRESHOLDING: THE ROLE OF THE EXPONENT r



Soft thresholding (olive) and prox of $\phi = |\cdot| + 0.9|\cdot|^{3/2}$ (purple).

PROXIMAL THRESHOLDING: THE ROLE OF THE EXPONENT r



Soft thresholding (olive) and prox of $\phi = |\cdot| + 0.9|\cdot|^{4/3}$ (blue).

THE PROXIMITY OPERATOR OF POWERS

Let $r \in]1, 2]$, let $\tau > 0$, let $\mu \in \mathbb{R}$, and consider $\mathrm{prox}_{\tau|\cdot|^r} : \mathbb{R} \to \mathbb{R}$.

$$\mathrm{prox}_{\tau|\cdot|^r}(\mu) = \mathrm{sign}(\mu)\xi, \quad \text{where} \quad \xi \geq 0 \text{ and } \xi + r\tau\xi^{r-1} = |\mu| \,.$$

There are several exponents r for which the equation can be explicitly solved: $r \in \{3/2, 4/3, 5/4\}$. However in general this is not possible and one needs to rely on iterative methods, e.g. by bisection.

THE PROXIMITY OPERATOR OF POWERS

Let $r \in]1,2]$, let $\tau > 0$, let $\mu \in \mathbb{R}$, and consider $\mathrm{prox}_{\tau|\cdot|^r} : \mathbb{R} \to \mathbb{R}$.

$$\mathrm{prox}_{\tau|\cdot|^r}(\mu) = \mathrm{sign}(\mu)\xi, \quad \text{where} \quad \xi \geq 0 \text{ and } \xi + r\tau\xi^{r-1} = |\mu|\,.$$

There are several exponents r for which the equation can be explicitly solved: $r \in \{3/2, 4/3, 5/4\}$. However in general this is not possible and one needs to rely on iterative methods, e.g. by bisection.

The following decomposition rule holds (Combettes and Pesquet '07)

$$g_k = \iota_{C_k} + \eta_k|\cdot|^r + \sigma_{D_k}$$
$$\mathrm{prox}_{\gamma g_k}(\mu) = \pi_{C_k}\big(\mathrm{prox}_{\gamma\eta_k|\cdot|^r}\big(\mathrm{soft}_{\gamma D_k}(\mu)\big)\big)$$

THE PROXIMITY OPERATOR OF POWERS

Let $r \in]1,2]$, let $\tau > 0$, let $\mu \in \mathbb{R}$, and consider $\mathrm{prox}_{\tau|\cdot|^r} : \mathbb{R} \to \mathbb{R}$.

$$\mathrm{prox}_{\tau|\cdot|^r}(\mu) = \mathrm{sign}(\mu)\xi, \quad \text{where} \quad \xi \geq 0 \text{ and } \xi + r\tau\xi^{r-1} = |\mu|\,.$$

There are several exponents r for which the equation can be explicitly solved: $r \in \{3/2, 4/3, 5/4\}$. However in general this is not possible and one needs to rely on iterative methods, e.g. by bisection.

The following decomposition rule holds (Combettes and Pesquet '07)

$$g_k = \iota_{C_k} + \eta_k|\cdot|^r + \sigma_{D_k}$$
$$\mathrm{prox}_{\gamma g_k}(\mu) \neq \pi_{C_k}\left(\mathrm{prox}_{\gamma\eta_k|\cdot|^r}\left(\mathrm{soft}_{\gamma D_k}(\mu)\right) + \alpha_{k,m}\right)\,)$$

The question is to adjust the decomposition rule in the presence of errors.

A NOTION OF INEXACT PROXIMITY OPERATOR

$$u \simeq_\delta \mathrm{prox}_{\eta\varphi}(w) \overset{def}{\iff} u \in \frac{\delta^2}{2\eta} - \mathrm{argmin}_{\xi\in\mathcal{H}} \left\{ \varphi(\xi) + \frac{1}{2\eta}\|\xi - w\|_{\mathcal{H}}^2 \right\}.$$

A NOTION OF INEXACT PROXIMITY OPERATOR

$$u \simeq_\delta \operatorname{prox}_{\eta\varphi}(w) \overset{def}{\iff} u \in \frac{\delta^2}{2\eta} - \operatorname{argmin}_{\xi\in\mathcal{H}} \left\{ \varphi(\xi) + \frac{1}{2\eta}\|\xi - w\|^2_{\mathcal{H}} \right\}.$$

Lemma

Let $\psi, \phi, \sigma \in \Gamma_0^+(\mathbb{R})$ with $\psi(0) = 0$ and σ positively homogeneous. Then

- $s = \operatorname{prox}_{\vartheta|\cdot|^r}(\mu) + \alpha \quad \Longrightarrow \quad s \simeq_\delta \operatorname{prox}_{\vartheta|\cdot|^r}(\mu)$
- $s \simeq_\delta \operatorname{prox}_\psi(\operatorname{prox}_\sigma(\mu)) \implies s \simeq_\delta \operatorname{prox}_{\psi+\sigma}(\mu)$
- $s \simeq_\delta \operatorname{prox}_\phi(\mu)$ *and* $p = \pi_{\mathcal{C}}(s) \implies p \simeq_\delta \operatorname{prox}_{\iota_{\mathcal{C}}+\phi}(\mu)$.

A NOTION OF INEXACT PROXIMITY OPERATOR

$$u \simeq_\delta \operatorname{prox}_{\eta\varphi}(w) \overset{def}{\iff} u \in \frac{\delta^2}{2\eta} - \operatorname{argmin}_{\xi\in\mathcal{H}} \left\{ \varphi(\xi) + \frac{1}{2\eta}\|\xi - w\|_{\mathcal{H}}^2 \right\}.$$

Lemma

Let $\psi, \phi, \sigma \in \Gamma_0^+(\mathbb{R})$ with $\psi(0) = 0$ and σ positively homogeneous. Then

- $s = \operatorname{prox}_{\vartheta|\cdot|^r}(\mu) + \alpha \quad \Longrightarrow \quad s \simeq_\delta \operatorname{prox}_{\vartheta|\cdot|^r}(\mu)$
- $s \simeq_\delta \operatorname{prox}_\psi(\operatorname{prox}_\sigma(\mu)) \Longrightarrow s \simeq_\delta \operatorname{prox}_{\psi+\sigma}(\mu)$
- $s \simeq_\delta \operatorname{prox}_\phi(\mu)$ *and* $p = \pi_{\mathcal{C}}(s) \Longrightarrow p \simeq_\delta \operatorname{prox}_{\iota_{\mathcal{C}}+\phi}(\mu)$.

$$\begin{aligned} g_k &= \iota_{C_k} + \eta_k|\cdot|^r + \sigma_{D_k} \\ \left.\begin{aligned} s &= \operatorname{prox}_{\gamma\eta_k|\cdot|^r}\left(\operatorname{soft}_{\gamma D_k}(\mu)\right) + \alpha_{k,m} \\ p &= \pi_{C_k}(s) \end{aligned}\right\} &\Rightarrow p \simeq_{\delta_{k,m}} \operatorname{prox}_{\gamma g_k}(\mu) \end{aligned}$$

Theorem

Let

$$\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n}(f_u(x_i) - y_i)^2 \quad \textit{and} \quad G(u) = \sum_{k\in\mathbb{K}} g_k(\langle u \mid e_k\rangle).$$

Then there exist a vanishing sequence $(\delta_m)_{m\in\mathbb{N}}$, $\delta_m \leq Cm^{-p}$, *such that*

$$\begin{cases} u_{m+1} \simeq_{\delta_m} \operatorname{prox}_{\gamma G}\left(v_m - \frac{\gamma}{\lambda}\nabla\widehat{F}_n(v_m)\right) \\ v_{m+1} = u_{m+1} + \frac{\tau_m - 1}{\tau_{m+1}}(u_{m+1} - u_m) \end{cases}$$

Hence it follows from (Villa, Salzo et. al. '13) *that the sequence* $(u_m)_{m\in\mathbb{N}}$ *is minimizing for* $\hat{F}_n + \lambda G$ *and* $\|u_m - \hat{u}_{n,\lambda}(z_n)\|_r \to 0$ *as* $m \to +\infty$.

FURTHER DEVELOPMENTS

A MORE GENERAL FRAMEWORK

The statistical analysis can be pursued in more general terms (Combettes, Salzo and Villa '14):

1. The functional model is given by means of a linear operator

$$A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathsf{Y})\colon u \mapsto f_u$$

   continuous w.r.t. the topology of the pointwise convergence, where $\mathcal{F}$ (the feature space) and Y are **Banach spaces**. This way $\mathrm{ran}(A)$ is a pre-RKBS of vector-valued functions.

A MORE GENERAL FRAMEWORK

The statistical analysis can be pursued in more general terms (Combettes, Salzo and Villa '14):

1. The functional model is given by means of a linear operator

$$A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathsf{Y})\colon u \mapsto f_u$$

   continuous w.r.t. the topology of the pointwise convergence, where $\mathcal{F}$ (the feature space) and Y are **Banach spaces**. This way $\operatorname{ran}(A)$ is a pre-RKBS of vector-valued functions.
2. The constraint set is $\mathcal{C} = \big\{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \,\big|\, (\forall x \in \mathcal{X})\, f(x) \in C(x)\big\}$, where, for every $x \in \mathcal{X}$, $C(x) \subset \mathsf{Y}$ is closed convex.

A MORE GENERAL FRAMEWORK

The statistical analysis can be pursued in more general terms (Combettes, Salzo and Villa '14):

1. The functional model is given by means of a linear operator
$$A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathsf{Y})\colon u \mapsto f_u$$
continuous w.r.t. the topology of the pointwise convergence, where $\mathcal{F}$ (the feature space) and Y are **Banach spaces**. This way $\operatorname{ran}(A)$ is a pre-RKBS of vector-valued functions.
2. The constraint set is $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathsf{Y}) \mid (\forall x \in \mathcal{X})\, f(x) \in C(x)\}$, where, for every $x \in \mathcal{X}$, $C(x) \subset \mathsf{Y}$ is closed convex.
3. The risk is based on a **general loss** function
$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x))\mathrm{d}P(x, y)$$

A MORE GENERAL FRAMEWORK

The statistical analysis can be pursued in more general terms (Combettes, Salzo and Villa '14):

1. The functional model is given by means of a linear operator
$$A\colon \mathcal{F} \to \mathcal{M}(\mathcal{X}, \mathrm{Y})\colon u \mapsto f_u$$
continuous w.r.t. the topology of the pointwise convergence, where $\mathcal{F}$ (the feature space) and Y are **Banach spaces**. This way $\operatorname{ran}(A)$ is a pre-RKBS of vector-valued functions.
2. The constraint set is $\mathcal{C} = \{f \in \mathcal{M}(\mathcal{X}, \mathrm{Y}) \mid (\forall x \in \mathcal{X})\, f(x) \in C(x)\}$, where, for every $x \in \mathcal{X}$, $C(x) \subset \mathrm{Y}$ is closed convex.
3. The risk is based on a **general loss** function
$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x))\mathrm{d}P(x, y)$$
4. $G\colon \mathcal{F} \to]-\infty, +\infty]$ is **totally convex**, meaning that, for every $t > 0$ and every $u \in \operatorname{dom} G$,
$$0 < \psi_G(u; t) = \inf\left\{G(v) - G(u) - G'(u; v-u) \,\middle|\, u \in \operatorname{dom} G, \|u - v\| = t\right\}.$$

A MORE GENERAL FRAMEWORK

- We characterized the property of universality w.r.t. $\mathcal{C}$, i.e., when $\inf_{\mathcal{C}} R = \inf_{\mathcal{C} \cap \mathrm{ran}(A)} R$.
- We proved a representer and sensitivity theorem: that is there exists a measurable function $\Psi_\lambda \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{F}^*$ such that $-\mathrm{E}_P[\Psi_\lambda] \in \lambda \partial G(u_\lambda)$ and, for every $z_n \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\frac{\psi_G(u_\lambda, \|\hat{u}_{n,\lambda}(z_n) - u_\lambda\|)}{\|\hat{u}_{n,\lambda}(z_n) - u_\lambda\|} \leq \frac{1}{\lambda} \Big\| \frac{1}{n} \sum_{i=1}^{n} \Psi_\lambda(x_i, y_i) - \mathrm{E}_P[\Psi_\lambda] \Big\|.$$

- We found conditions on the sequence $(\lambda_n)_{n \in \mathbb{N}}$, ensuring consistency, that depend on the **modulus of total convexity** of G, the **Rademacher type** of $\mathcal{F}^*$, and the local Lipschitz constants of the loss function.

AN ALTERNATIVE APPROACH

When restricted to **real-valued functions**, an alternative approach is to use Rademacher complexities.

1. In this case the statistical analysis achieves sharper results when dealing with general loss function.
2. **Open problem**: Rademacher complexity for (reproducing kernel) classes of functions taking values in infinite dimension spaces. A contraction principle is missing in that scenario.

Thank you

SOME REFERENCES

- P. L. Combettes and J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, SIAM J. Optim., vol. 18, pp. 1351–1376, 2007.
- P. L. Combettes, S. Salzo and S. Villa, Consistency of Regularized Learning Schemes in Banach spaces. ArXiv, 2014.
- P. L. Combettes, S. Salzo and S. Villa, Consistent Learning by Composite Proximal Thresholding. ArXiv, 2015.
- S. Villa, S. Salzo, L. Baldassarre, A. Verri Accelerated and Inexact Forward-Backward Algorithms, SIOPT., 16, 2013.
- Z B. Xu and G.F. Roach, Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces , Journal of Mathematical Analysis and Applications, 157, 1991.