A Consistent Regularization Approach for Structured Prediction

Cédric Rommel

Journal Club - CMAP

February 23rd, 2017

Contents

Subject and motivation

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,image reconstruction,

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,image reconstruction,ranking problem, ...

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,image reconstruction,ranking problem, ...

Motivation

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,image reconstruction,ranking problem, ...

Motivation

1. Unifying theoretical framework for SPP,

Article's subject and authors motivation

Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

$$f : \mathcal{X} \to \mathcal{Y}$$

with $\mathcal{Y}$ *structured*.

Examples: natural language parsing,image reconstruction,ranking problem, . . .

Motivation

1. Unifying theoretical framework for SPP,
2. Practical method for majority of these problems.

Problem formulation

The article considers output spaces $\mathcal{Y}$ whose structure is induced by some metric $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and formalizes the problems to be solved as follows:

Problem formulation

The article considers output spaces $\mathcal{Y}$ whose structure is induced by some metric $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and formalizes the problems to be solved as follows:

Main problem (MP)

Estimate

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min}\, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

Problem formulation

The article considers output spaces $\mathcal{Y}$ whose structure is induced by some metric $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and formalizes the problems to be solved as follows:

Main problem (MP)

Estimate

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min}\, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

Rem:

- ρ is unknown;

Problem formulation

The article considers output spaces $\mathcal{Y}$ whose structure is induced by some metric $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and formalizes the problems to be solved as follows:

Main problem (MP)

Estimate

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min}\, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

Rem:

- ρ is unknown;
- using sample $\{(x_i, y_i)\}_{i=1}^n$.

Derivation of a general algorithm

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

- $\forall x \in \mathcal{X}$, $k(x, .) \in H$ and

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

- $\forall x \in \mathcal{X}$, $k(x, .) \in H$ and
- $\forall f \in H$, $f(x) = \langle f, k(x, .) \rangle$,

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

- $\forall x \in \mathcal{X}$, $k(x, .) \in H$ and
- $\forall f \in H$, $f(x) = \langle f, k(x, .) \rangle$,

then H is an RKHS of kernel k.

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

- $\forall x \in \mathcal{X}$, $k(x, .) \in H$ and
- $\forall f \in H$, $f(x) = \langle f, k(x, .) \rangle$,

then H is an RKHS of kernel k.

Rem:

- $k(x, y) = \langle k(x, .), k(y, .) \rangle$,

RKHS recall (1/2)

RKHS definition

Let $\mathcal{X}$ be an arbitrary set and $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ be a Hilbert space.

If $\exists ! k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric, such that

- $\forall x \in \mathcal{X}$, $k(x, \cdot) \in H$ and
- $\forall f \in H$, $f(x) = \langle f, k(x, \cdot) \rangle$,

then H is an RKHS of kernel k.

Rem:

- $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle$,
- A reproducing kernel k is *positive-definite*:

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \tag{1}$$

$\forall n \in \mathbb{N}, x_1, \ldots x_n \in \mathcal{X}, c_1, \ldots, c_n \in \mathbb{R}.$

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle$$

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

where $\psi : \mathcal{X} \to H$ is called *feature map*.

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

where $\psi : \mathcal{X} \to H$ is called *feature map*.

Kernel trick

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

where $\psi : \mathcal{X} \to H$ is called *feature map*.

Kernel trick

$$f(x) = w^\top \psi(x) + b$$

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

where $\psi : \mathcal{X} \to H$ is called *feature map*.

Kernel trick

$$f(x) = w^\top \psi(x) + b \underset{SVM, KRR, \dots}{\leadsto}$$

RKHS recall (2/2)

Moore-Aronszajn theorem

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric positive-definite $\Rightarrow \exists!$ RKHS $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ of kernel k.

Rem:

$\forall x_i, x_j \in \mathcal{X}$,

$$k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle,$$

where $\psi : \mathcal{X} \to H$ is called *feature map*.

Kernel trick

$$f(x) = w^\top \psi(x) + b \underset{SVM, KRR, \ldots}{\rightsquigarrow} \hat{f}(x) = \sum_{i=1}^{n} c_i k(x_i, x)$$

Algorithm description

Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *reproducing kernel*,
- $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = k(x_i, x_j)$,
- $\mathbf{K}_x \in \mathbb{R}^n$ with $(\mathbf{K}_x)_i = k(x, x_i)$.

Algorithm description

Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *reproducing kernel*,
- $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = k(x_i, x_j)$,
- $\mathbf{K}_x \in \mathbb{R}^n$ with $(\mathbf{K}_x)_i = k(x, x_i)$.

Algorithm 1

Algorithm description

Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *reproducing kernel*,
- $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = k(x_i, x_j)$,
- $\mathbf{K}_x \in \mathbb{R}^n$ with $(\mathbf{K}_x)_i = k(x, x_i)$.

Algorithm 1

Learning:

$$\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x \in \mathbb{R}^n$$

Algorithm description

Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *reproducing kernel*,
- $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = k(x_i, x_j)$,
- $\mathbf{K}_x \in \mathbb{R}^n$ with $(\mathbf{K}_x)_i = k(x, x_i)$.

Algorithm 1

Learning:

$$\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x \in \mathbb{R}^n$$

Prediction:

$$\hat{f}(x) = \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \Delta(y, y_i)$$

Intuition (1/3)

Special case of *Kernel Dependency Estimation*:

$$\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$$

Intuition (1/3)

Special case of *Kernel Dependency Estimation*:

$$\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$$

where h is a reproducing kernel and $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ its nonlinear feature map.

Intuition (1/3)

Special case of *Kernel Dependency Estimation*:

$$\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$$

where h is a reproducing kernel and $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ its nonlinear feature map.

$$\Delta(f(x), y) = \|\psi(f(x)) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}}$$

Intuition (1/3)

Special case of *Kernel Dependency Estimation*:

$$\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$$

where h is a reproducing kernel and $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ its nonlinear feature map.

$$\Delta(f(x), y) = \|\psi(f(x)) - \psi(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}$$

As $\psi(f(x))$ is hard to minimize wrt to f, we replace it by $g \in \mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ easier to optimize:

Intuition (1/3)

Special case of *Kernel Dependency Estimation*:

$$\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$$

where h is a reproducing kernel and $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ its nonlinear feature map.

$$\Delta(f(x), y) = \|\psi(f(x)) - \psi(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}$$

As $\psi(f(x))$ is hard to minimize wrt to f, we replace it by $g \in \mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ easier to optimize:

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}_{\mathcal{Y}}} + \lambda \|g\|^2_{\mathcal{G}}$$

Intuition (2/3)

If we chose

$$\mathcal{G} = \left\{ g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y}) : g(x) = \sum_{i=1}^{n} c_i k(x, x_i), c_i \in \mathcal{H}_\mathcal{Y} \right\}$$

Intuition (2/3)

If we chose

$$\mathcal{G} = \left\{ g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y}) : g(x) = \sum_{i=1}^{n} c_i k(x, x_i), c_i \in \mathcal{H}_\mathcal{Y} \right\}$$

with k a reproducing kernel, the last problem becomes a *Kernel Ridge Regression* problem and

Intuition (2/3)

If we chose

$$\mathcal{G} = \left\{ g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y}) : g(x) = \sum_{i=1}^{n} c_i k(x, x_i), c_i \in \mathcal{H}_\mathcal{Y} \right\}$$

with k a reproducing kernel, the last problem becomes a *Kernel Ridge Regression* problem and

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x) \psi(y_i),$$

with

$$\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x \in \mathbb{R}^n.$$

Intuition (3/3)

In this case, some references advise to choose $\hat{f}$ as follows:

$$\hat{f}(x) = \arg\min_{y \in \mathcal{Y}} \|\psi(y) - \hat{g}(x)\|^2_{\mathcal{H}_\mathcal{Y}}$$

Intuition (3/3)

In this case, some references advise to choose $\hat{f}$ as follows:

$$\hat{f}(x) = \arg\min_{y \in \mathcal{Y}} \|\psi(y) - \hat{g}(x)\|^2_{\mathcal{H}_\mathcal{Y}}$$

$$= \arg\min_{y \in \mathcal{Y}} h(y, y) - 2 \sum_{i=1}^{n} \alpha_i(x) h(y, y_i)$$

Intuition (3/3)

In this case, some references advise to choose $\hat{f}$ as follows:

$$\hat{f}(x) = \underset{y \in \mathcal{Y}}{\arg\min} \|\psi(y) - \hat{g}(x)\|^2_{\mathcal{H}_\mathcal{Y}}$$

$$= \underset{y \in \mathcal{Y}}{\arg\min} h(y, y) - 2 \sum_{i=1}^{n} \alpha_i(x) h(y, y_i)$$

which is equal to the *prediction* step of *Algorithm 1*

$$\hat{f}(x) = \underset{y \in \mathcal{Y}}{\arg\min} \sum_{i=1}^{n} \alpha_i(x) \Delta(y, y_i)$$

if h is *normalized*.

Questions raised ?

1. And if we don't have $\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$?

Questions raised ?

1. And if we don't have $\Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y')$?
2. Is the transfer between $\hat{g}$ and $\hat{f}$ theoretically justified ?

Generalization

Main problem (MP)

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min}\, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

Generalization

Main problem (MP)

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min} \, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

We want to find $\mathcal{L}(g(x), y)$ a relaxation of $\Delta(f(x), y)$ on some space $\mathcal{H}_\mathcal{Y}$ easy to optimize in order to replace (MP) by

Generalization

Main problem (MP)

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\arg\min}\, \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$

We want to find $\mathcal{L}(g(x), y)$ a relaxation of $\Delta(f(x), y)$ on some space $\mathcal{H}_\mathcal{Y}$ easy to optimize in order to replace (MP) by

Surrogate Problem (SP)

$$g^* \in \underset{g \in \mathcal{F}(\mathcal{X},\mathcal{H}_\mathcal{Y})}{\arg\min}\, \mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(g(x), y) d\rho(x, y)$$

Desired properties

(MP) and (SP) will be equivalent if we find an appropriate *decoding* function $d : \mathcal{H}_{\mathcal{Y}} \to \mathcal{Y}$ satisfying

Desired properties

(MP) and (SP) will be equivalent if we find an appropriate *decoding* function $d : \mathcal{H}_{\mathcal{Y}} \to \mathcal{Y}$ satisfying

Fisher Consistency:

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*)$$

Desired properties

(MP) and (SP) will be equivalent if we find an appropriate *decoding* function $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying

Fisher Consistency:

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*)$$

Comparison Inequality:

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq \varphi(\mathcal{R}(g) - \mathcal{R}(g^*))$$

for all $g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y})$, with $\varphi : \mathbb{R} \to \mathbb{R}$ such that $\lim_{s \to 0} \varphi(s) = 0$.

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_{\mathcal{Y}}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$;

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_\mathcal{Y}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ continuous embedding;

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_\mathcal{Y}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ continuous embedding;
- $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ bounded linear operator;

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_{\mathcal{Y}}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ continuous embedding;
- $V : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ bounded linear operator;

such that

$$\forall y, y' \in \mathcal{Y}, \qquad \Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_{\mathcal{Y}}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ continuous embedding;
- $V : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ bounded linear operator;

such that

$$\forall y, y' \in \mathcal{Y}, \qquad \Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

Rem:

- V is not required to be positive definite, nor even symmetric;

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_{\mathcal{Y}}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ continuous embedding;
- $V : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ bounded linear operator;

such that

$$\forall y, y' \in \mathcal{Y}, \qquad \Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

Rem:

- V is not required to be positive definite, nor even symmetric;
- looks like reproducing kernel, but its not;

Surrogate problem equivalence (1/2)

Assumption 1

There is

- $\mathcal{H}_\mathcal{Y}$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$;
- $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ continuous embedding;
- $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ bounded linear operator;

such that

$$\forall y, y' \in \mathcal{Y}, \qquad \Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}}.$$

Rem:

- V is not required to be positive definite, nor even symmetric;
- looks like reproducing kernel, but its not;
- "wide range of functions"

Examples of Δ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...

Examples of Δ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, ϵ-sensitivity, ...

Examples of Δ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, ϵ-sensitivity, ...
- Robust loss functions: Huber, $L2 - L1$, ...

Examples of Δ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, ϵ-sensitivity, ...
- Robust loss functions: Huber, $L2 - L1$, ...
- Kernel Dependency Estimation (loss function from intuition)

Examples of Δ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, ϵ-sensitivity, ...
- Robust loss functions: Huber, $L2 - L1$, ...
- Kernel Dependency Estimation (loss function from intuition)
- Distances on histograms and probabilities

Surrogate problem equivalence (2/2)

Lemma 1

If Δ satisfies *Assumption 1* with ψ bounded, then we have

Surrogate problem equivalence (2/2)

Lemma 1

If Δ satisfies *Assumption 1* with ψ bounded, then we have

$$\mathcal{E}(f) = \int_{\mathcal{X}} \langle \psi(f(x), Vg^*(x)) \rangle_{\mathcal{H}_{\mathcal{Y}}} d\rho_{\mathcal{X}}(x)$$

for all $f : \mathcal{X} \to \mathcal{Y}$,

Surrogate problem equivalence (2/2)

Lemma 1

If Δ satisfies *Assumption 1* with ψ bounded, then we have

$$\mathcal{E}(f) = \int_{\mathcal{X}} \langle \psi(f(x), Vg^*(x)) \rangle_{\mathcal{H}_{\mathcal{Y}}} d\rho_{\mathcal{X}}(x)$$

for all $f : \mathcal{X} \to \mathcal{Y}$,where $g^* : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ solves

$$\min_{g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})} \mathcal{R}(g) = \int_{\mathcal{X} \times \mathcal{Y}} \|g(x) - \psi(y)\|^2_{\mathcal{H}_{\mathcal{Y}}} d\rho(x, y).$$

Good properties of decoding function

Theorem 2

If Δ satisfies *Assumption 1* and $\mathcal{Y}$ is compact, then for any $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ measurable and for $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ such that

Good properties of decoding function

Theorem 2

If Δ satisfies *Assumption 1* and $\mathcal{Y}$ is compact, then for any $g : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ measurable and for $d : \mathcal{H}_{\mathcal{Y}} \to \mathcal{Y}$ such that

$$d(h) = \arg\min_{y \in \mathcal{Y}} \langle \psi(y), Vh \rangle_{\mathcal{H}_{\mathcal{Y}}},$$

Good properties of decoding function

Theorem 2

If Δ satisfies *Assumption 1* and $\mathcal{Y}$ is compact, then for any $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ measurable and for $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ such that

$$d(h) = \underset{y \in \mathcal{Y}}{\arg\min} \langle \psi(y), Vh \rangle_{\mathcal{H}_\mathcal{Y}},$$

we have

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*)$$
$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq c_\Delta \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)},$$

with $c_\Delta = \|V\| \max_{y \in \mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$.

Good properties of decoding function

Theorem 2

If Δ satisfies *Assumption 1* and $\mathcal{Y}$ is compact, then for any $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ measurable and for $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ such that

$$d(h) = \arg\min_{y \in \mathcal{Y}} \langle \psi(y), Vh \rangle_{\mathcal{H}_\mathcal{Y}},$$

we have

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*)$$
$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq c_\Delta \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)},$$

with $c_\Delta = \|V\| \max_{y \in \mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$.

Rem: Any SPP on $\mathcal{Y}$ of finite cardinality satifies *Th. 2*

Derivation of Algorithm 1

Lemma 3

If Δ verify *Assumption 1*, $\mathcal{Y}$ is compact and $\hat{g} \in \mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ solves the *Empirical surrogate problem*

Derivation of Algorithm 1

Lemma 3

If Δ verify *Assumption 1*, $\mathcal{Y}$ is compact and $\hat{g} \in \mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y})$ solves the *Empirical surrogate problem*

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda \|g\|^2_{\mathcal{G}}$$

Derivation of Algorithm 1

Lemma 3

If Δ verify *Assumption 1*, $\mathcal{Y}$ is compact and $\hat{g} \in \mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ solves the *Empirical surrogate problem*

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}_{\mathcal{Y}}} + \lambda \|g\|^2_{\mathcal{G}}$$

then

$$\forall x \in \mathcal{X}, \qquad \hat{f} = d \circ \hat{g} = \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \Delta(y, y_i)$$

with

$$\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x.$$

Derivation of Algorithm 1

Lemma 3

If Δ verify *Assumption 1*, $\mathcal{Y}$ is compact and $\hat{g} \in \mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{H_Y})$ solves the *Empirical surrogate problem*

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \psi(y_i)\|_{\mathcal{H_Y}}^2 + \lambda \|g\|_{\mathcal{G}}^2$$

then

$$\forall x \in \mathcal{X}, \qquad \hat{f} = d \circ \hat{g} = \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \Delta(y, y_i)$$

with

$$\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x.$$

Rem: $\hat{f}$ does not depend on ψ nor on V: *loss trick* $\simeq$ *kernel trick*.

Statistical properties

Universal consistency

Theorem 4

Let $n \in \mathbb{N}$ and ρ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$.

Universal consistency

Theorem 4

Let $n \in \mathbb{N}$ and ρ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If Δ satisfies *Assumption 1*, $\mathcal{X}$ and $\mathcal{Y}$ are compact, k is a *continuous universal kernel*

Universal consistency

Theorem 4

Let $n \in \mathbb{N}$ and ρ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If Δ satisfies *Assumption 1*, $\mathcal{X}$ and $\mathcal{Y}$ are compact, k is a *continuous universal kernel*, and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/4}$, then

Universal consistency

Theorem 4

Let $n \in \mathbb{N}$ and ρ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If Δ satisfies *Assumption 1*, $\mathcal{X}$ and $\mathcal{Y}$ are compact, k is a *continuous universal kernel*, and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/4}$, then

$$\lim_{n \to +\infty} \mathcal{E}(\hat{f}_n) = \mathcal{E}(f^*) \qquad \text{with probability } 1.$$

Universal consistency

Theorem 4

Let $n \in \mathbb{N}$ and ρ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If Δ satisfies *Assumption 1*, $\mathcal{X}$ and $\mathcal{Y}$ are compact, k is a *continuous universal kernel*, and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/4}$, then

$$\lim_{n \to +\infty} \mathcal{E}(\hat{f}_n) = \mathcal{E}(f^*) \qquad \text{with probability } 1.$$

Rem: First universal consistency result for general form of SPP, including $\mathcal{Y}$ of infinite cardinality.

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then if $\mathcal{R}$ admits a minimizer g^*...

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then if $\mathcal{R}$ admits a minimizer g^*...

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 n^{-1/4}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant independant of n and τ.

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then if $\mathcal{R}$ admits a minimizer g^*...

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 n^{-1/4}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant independant of n and τ.

Rem:

- Not possible to prove uniform convergence rates according to ref [25];

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then if $\mathcal{R}$ admits a minimizer g^*...

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 n^{-1/4}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant independant of n and τ.

Rem:

- Not possible to prove uniform convergence rates according to ref [25];
- Better bounds possible in the case of least-squares classifiers by using Tsybakov condition to regularize ρ;

Generalization bounds

Theorem 5

If Δ satisfies *Assumption 1*, $\mathcal{Y}$ is compact, k is a *bounded* continuous universal kernel and if $\hat{f}_n$ is the estimator obtained through *Alg.1* with n i.i.d. training points with $\lambda_n = n^{-1/2}$, then if $\mathcal{R}$ admits a minimizer g^*...

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 n^{-1/4}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant independant of n and τ.

Rem:

- Not possible to prove uniform convergence rates according to ref [25];
- Better bounds possible in the case of least-squares classifiers by using Tsybakov condition to regularize ρ;
- Not sure if the same strategy would work on infinite setting.

Experiments

Experiements - Ranking Movies

$$\Delta_{rank}(y, y') = \frac{1}{2} \sum_{i,j=1}^{M} \gamma(y')_{ij}(1 - \mathrm{sign}(y_i - y_j))$$

	Rank Loss
Linear [7]	0.430 ± 0.004
Hinge [27]	0.432 ± 0.008
Logistic [28]	0.432 ± 0.012
SVM Struct [4]	0.451 ± 0.008
Alg. 1	$\mathbf{0.396 \pm 0.003}$

Table 1: Normalized $\triangle_{rank}$ for ranking methods on the MovieLens dataset [29].

Experiements - Digit reconstruction

$$\Delta_G(y, y') = 1 - k_G(y, y')$$

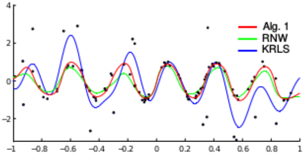
$$\Delta_H(y, y') = \sum_{i=1}^{M} \left| (y_i)^{1/2} - (y_i')^{1/2} \right|, \qquad \text{for} \qquad y = (y_i)_{i=1}^{M}$$

$$\Delta_R(y, y') = \text{"Recongnition" loss wrt SVM classifier}$$

Loss	KDE [18] (Gaussian)	$\boxed{\textbf{Alg. 1}}$ (Hellinger)
$\triangle_G$	$\mathbf{0.149 \pm 0.013}$	0.172 ± 0.011
$\triangle_H$	0.736 ± 0.032	$\mathbf{0.647 \pm 0.017}$
$\triangle_R$	0.294 ± 0.012	$\mathbf{0.193 \pm 0.015}$

Table 2: Digit reconstruction using Gaussian (KDE [18]) and Hellinger loss.

Experiements - Robust Estimation



n	Alg. 1	RNW	KRR
50	**0.39** $\pm$ 0.17	0.45 $\pm$ 0.18	0.62 $\pm$ 0.13
100	**0.21** $\pm$ 0.04	0.29 $\pm$ 0.04	0.47 $\pm$ 0.09
200	**0.12** $\pm$ 0.02	0.24 $\pm$ 0.03	0.33 $\pm$ 0.04
500	**0.08** $\pm$ 0.01	0.22 $\pm$ 0.02	0.31 $\pm$ 0.03
1000	**0.07** $\pm$ 0.01	0.21 $\pm$ 0.02	0.19 $\pm$ 0.02

Figure 1: Robust estimation on the regression problem in Sec. 6 by minimizing the Cauchy loss with Alg. 1 (Ours) or Nadaraya-Watson (Nad). KRLS as a baseline predictor. **Left**. Example of one run of the algorithms. **Right**. Average distance of the predictors to the actual function (without noise and outliers) over 100 runs with respect to training sets of increasing dimension.