A CONSISTENT REGULARIZATION APPROACH FOR STRUCTURED PREDICTION

Cédric Rommel

Journal Club - CMAP

February 23rd, 2017
1 Subject and motivation

2 Derivation of a general algorithm

3 Statistical properties

4 Experiments
A Consistent Regularization Approach for Structured Prediction

Cédric Rommel

Subject and motivation

Derivation of a general algorithm

Statistical properties

Experiments

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

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

with \( \mathcal{Y} \) *structured*. 
Article's subject and authors

Motivation

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

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

with \( \mathcal{Y} \) *structured*.

**Examples:** natural language parsing,
Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

\[ f : X \rightarrow Y \]

with \( Y \) *structured*.

**Examples**: natural language parsing, image reconstruction,
Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

with \( \mathcal{Y} \) *structured*.

**Examples**: natural language parsing, image reconstruction, ranking problem, . . .
Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

with \( \mathcal{Y} \) structured.

**Examples:** natural language parsing, image reconstruction, ranking problem, ...
Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

with \( \mathcal{Y} \) *structured*.

**Examples**: natural language parsing, image reconstruction, ranking problem, ...
Article tackles *Structured Prediction Problems* (SPP) where we want to estimate

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

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

**Main problem (MP)**

Estimate

$$f^* \in \arg \min_{f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} \Delta(f(x), y) d\rho(x, y).$$
The article considers output spaces $\mathcal{Y}$ whose structure is induced by some metric $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and formalizes the problems to be solved as follows:

**Main problem (MP)**

Estimate

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

*Rem:*

- $\rho$ is unknown;
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

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

*Rem:*

- $\rho$ is unknown;
- using sample $\{(x_i, y_i)\}_{i=1}^n$. 
A Consistent Regularization Approach for Structured Prediction

Cédric Rommel

Subject and motivation
Derivation of a general algorithm
Statistical properties
Experiments

Derivation of a general algorithm
RKHS DEFINITION

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

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

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

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

$\langle f, k(x, \cdot) \rangle = \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j)$, \( (1) \)

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

RKHS recall (2/2)
RKHS DEFINITION

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

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

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

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 $\exists! k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 $k$. 

RKHS recall (1/2)
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:
- $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle$. 

RKHS recall (2/2)
RKHS definition

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

If $\exists! k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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:

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

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

\[ k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \text{ symmetric positive-definite} \Rightarrow \exists! \text{ RKHS} \]

\[ H \subset \mathcal{F}(\mathcal{X}, \mathbb{R}) \text{ of kernel } k. \]
Moore-Aronszajn Theorem

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

Rem:

\[ \forall x_i, x_j \in \mathcal{X}, \]

\[ k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle \]
**Moore-Aronszajn Theorem**

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

*Rem:*

\[ \forall x_i, x_j \in \mathcal{X}, \]

\[ k(x_i, x_j) = \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \langle \psi(x_i), \psi(x_j) \rangle, \]
**Moore-Aronszajn theorem**

\[ k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \text{ symmetric positive-definite} \Rightarrow \exists! \text{ RKHS} \]

\[ H \subset \mathcal{F}(\mathcal{X}, \mathbb{R}) \text{ of kernel } k. \]

**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.
Moore-Aronszajn theorem

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

\text{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 \( \psi : \mathcal{X} \rightarrow H \) is called \textit{feature map}.

Kernel trick
A Consistent Regularization Approach for Structured Prediction

Cédric Rommel

Subject and motivation

Derivation of a general algorithm

Statistical properties

Experiments

RKHS recall (2/2)

**MOORE-ARONSZAJN THEOREM**

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

*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 \( \psi : \mathcal{X} \rightarrow H \) is called feature map.

**Kernel trick**

\[ f(x) = w^\top \psi(x) + b \]
**Moore-Aronszajn Theorem**

\[ k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \] 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 \( \psi : \mathcal{X} \rightarrow H \) is called feature map.

**Kernel trick**

\[ f(x) = w^\top \psi(x) + b \overset{\sim}{\rightarrow} \text{SVM, KRR, ...} \]
**Moore-Aronszajn Theorem**

\[ k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \text{ symmetric positive-definite} \Rightarrow \exists! \text{ RKHS} \]

\[ H \subset \mathcal{F}(\mathcal{X}, \mathbb{R}) \text{ of kernel } k. \]

**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**

\[ f(x) = w^\top \psi(x) + b \overset{\text{SVM,KRR,\ldots}}{\sim} \hat{f}(x) = \sum_{i=1}^{n} c_i k(x_i, x) \]
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 \( K_{ij} = k(x_i, x_j) \),
- \( \mathbf{K}_x \in \mathbb{R}^n \) with \( (\mathbf{K}_x)_i = k(x, x_i) \).
Algorithm Description

Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ reproducing kernel,
- $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = k(x_i, x_j)$,
- $K_x \in \mathbb{R}^n$ with $(K_x)_i = k(x, x_i)$.

Algorithm 1
Let

• \( \lambda \in \mathbb{R} \),
• \( k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) reproducing kernel,
• \( K \in \mathbb{R}^{n \times n} \) with \( K_{ij} = k(x_i, x_j) \),
• \( K_x \in \mathbb{R}^n \) with \( (K_x)_i = k(x, x_i) \).

Algorithm 1

**Learning:**

\[
\alpha(x) = (K + n\lambda I)^{-1}K_x \in \mathbb{R}^n
\]
Let

- $\lambda \in \mathbb{R}$,
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ reproducing kernel,
- $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = k(x_i, x_j)$,
- $K_x \in \mathbb{R}^n$ with $(K_x)_i = k(x, x_i)$.

**Algorithm 1**

**Learning:**

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

**Prediction:**

$$\hat{f}(x) = \arg\min_{y \in Y} \sum_{i=1}^{n} \alpha_i(x) \Delta(y, y_i)$$
Special case of *Kernel Dependency Estimation*:

\[ \Delta(y, y') = h(y, y) - 2h(y, y') + h(y', y') \]
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 \( \psi : \mathcal{Y} \rightarrow \mathcal{H}_\mathcal{Y} \) its nonlinear feature map.
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 \( \psi : \mathcal{Y} \rightarrow \mathcal{H}_\mathcal{Y} \) its nonlinear feature map.

\[ \Delta(f(x), y) = \| \psi(f(x)) - \psi(y) \|_{\mathcal{H}_\mathcal{Y}}^2 \]
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 $\psi : \mathcal{Y} \rightarrow \mathcal{H}_Y$ its nonlinear feature map.

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

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

\[ \Delta(f(x), y) = \| \psi(f(x)) - \psi(y) \|^2_{\mathcal{H}_Y} \]

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

\[ \min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \| g(x_i) - \psi(y_i) \|^2_{\mathcal{H}_Y} + \lambda \| g \|^2_{\mathcal{G}} \]
If we chose

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

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

with $k$ a reproducing kernel, the last problem becomes a *Kernel Ridge Regression* problem and
If we chose
\[
\mathcal{G} = \left\{ g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_Y) : g(x) = \sum_{i=1}^{n} c_i k(x, x_i), c_i \in \mathcal{H}_Y \right\}
\]

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

\[
\alpha(x) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K}_x \in \mathbb{R}^n.
\]
In this case, some references advise to choose $\hat{f}$ as follows:

$$\hat{f}(x) = \arg \min_{y \in \mathcal{Y}} \| \psi(y) - \hat{g}(x) \|^2_{\mathcal{H},\mathcal{Y}}$$
In this case, some references advise to choose $\hat{f}$ as follows:

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

$$
= \arg \min_{y \in \mathcal{Y}} h(y, y) - 2 \sum_{i=1}^{n} \alpha_i(x) h(y, y_i)
$$
In this case, some references advise to choose $\hat{f}$ as follows:

$$\hat{f}(x) = \arg\min_{y \in \mathcal{Y}} \| \psi(y) - \hat{g}(x) \|_2^2$$

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

which is equal to the prediction step of Algorithm 1

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

if $h$ is normalized.
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 ?
Main problem (MP)

\[ f^* \in \arg\min_{f \in \mathcal{F}(X,Y)} \mathcal{E}(f) = \int_{X \times Y} \Delta(f(x), y) d\rho(x, y). \]
**Main problem (MP)**

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

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

**Main Problem (MP)**

\[ \begin{align*} 
    f^* & \in \arg\min_{f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \mathcal{E}(f) \\
    & = \int_{\mathcal{X} \times \mathcal{Y}} \Delta(f(x), y) d\rho(x, y). 
\end{align*} \]

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

**Surrogate Problem (SP)**

\[ \begin{align*} 
    g^* & \in \arg\min_{g \in \mathcal{F}(\mathcal{X}, \mathcal{H}_Y)} \mathcal{R}(g) \\
    & = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}(g(x), y) d\rho(x, y) 
\end{align*} \]
(MP) and (SP) will be equivalent if we find an appropriate decoding function $d : \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{Y}$ satisfying
(MP) and (SP) will be equivalent if we find an appropriate decoding function \( d : \mathcal{H}_Y \rightarrow Y \) satisfying

**Fisher Consistency:**

\[
\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*)
\]
(MP) and (SP) will be equivalent if we find an appropriate decoding function $d : \mathcal{H}_Y \rightarrow Y$ 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 $g \in \mathcal{F}(X, \mathcal{H}_Y)$, with $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ such that $\lim_{s \rightarrow 0} \varphi(s) = 0$. 
Assumption 1

There is

- $\mathcal{H}_Y$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_Y}$;
**Assumption 1**

There is

- $H_Y$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{H_Y}$;
- $\psi : Y \rightarrow H_Y$ continuous embedding;
Assumption 1

There is

- $\mathcal{H}_Y$ separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_Y}$;
- $\psi : \mathcal{Y} \rightarrow \mathcal{H}_Y$ continuous embedding;
- $V : \mathcal{H}_Y \rightarrow \mathcal{H}_Y$ bounded linear operator;
Assumption 1

There is

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

such that

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

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

Rem:
- \( V \) is not required to be positive definite, nor even symmetric;
**Assumption 1**

There is

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

such that

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

**Rem:**

- $V$ is not required to be positive definite, nor even symmetric;
- looks like reproducing kernel, but it's not;
Assumption 1

There is

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

such that

\[
\forall y, y' \in Y, \quad \Delta(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_Y}.
\]

Rem:

- \( V \) is not required to be positive definite, nor even symmetric;
- looks like reproducing kernel, but it's not;
- "wide range of functions"
Examples of \( \Delta \) verifying Assumption 1

- Loss functions on \( \mathcal{Y} \) of finite cardinality: Multi-class classification, ranking, ...
Examples of $\Delta$ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, $\epsilon$-sensitivity, ...
Examples of $\Delta$ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, $\epsilon$-sensitivity, ...
- Robust loss functions: Huber, $L_2 - L_1$, ...
Examples of $\Delta$ verifying Assumption 1

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

Cédric Rommel

Subject and motivation

Derivation of a general algorithm

Statistical properties

Experiments

Examples of $\Delta$ verifying Assumption 1

- Loss functions on $\mathcal{Y}$ of finite cardinality: Multi-class classification, ranking, ...
- Least-squares, Logistic, Hinge, $\epsilon$-sensitivity, ...
- Robust loss functions: Huber, $L2 - L1$, ...
- Kernel Dependency Estimation (loss function from intuition)
- Distances on histograms and probabilities
**Lemma 1**

If $\Delta$ satisfies *Assumption 1* with $\psi$ bounded, then we have
Lemma 1

If $\Delta$ satisfies Assumption 1 with $\psi$ bounded, then we have

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

for all $f : \mathcal{X} \to \mathcal{Y}$,
Lemma 1

If $\Delta$ satisfies Assumption 1 with $\psi$ bounded, then we have

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

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

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

If $\Delta$ 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 $\Delta$ 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

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

with $c_\Delta = \|V\|_{\max} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$. Rem: Any SPP on $\mathcal{Y}$ of finite cardinality satisfies Th. 2.
**THEOREM 2**

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

Good properties of decoding function

Theorem 2

If $\Delta$ 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 satisfies Th. 2
Theorem 2

If $\Delta$ satisfies Assumption 1 and $\mathcal{Y}$ is compact, then for any $g : \mathcal{X} \rightarrow \mathcal{H}_{\mathcal{Y}}$ measurable and for $d : \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{Y}$ 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 satisfies Th. 2
**Lemma 3**

If $\Delta$ 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*
Lemma 3

If $\Delta$ 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

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

Rem: $\hat{f}$ does not depend on $\psi$ nor on $V$: loss trick $\approx$ kernel trick.
**Lemma 3**

If $\Delta$ 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

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

with

$$
\alpha(x) = (K + n\lambda I)^{-1}K_x.
$$
Lemma 3

If $\Delta$ 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

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

then

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

with

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

Rem: $\hat{f}$ does not depend on $\psi$ nor on $V$: loss trick $\simeq$ kernel trick.
STATISTICAL PROPERTIES
Theorem 4

Let $n \in \mathbb{N}$ and $\rho$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. 
**Theorem 4**

Let $n \in \mathbb{N}$ and $\rho$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If $\Delta$ satisfies *Assumption 1*, $\mathcal{X}$ and $\mathcal{Y}$ are compact, $k$ is a continuous universal kernel.
Theorem 4

Let $n \in \mathbb{N}$ and $\rho$ an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$. If $\Delta$ 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
THEOREM 4

Let \( n \in \mathbb{N} \) and \( \rho \) an arbitrary distribution on \( \mathcal{X} \times \mathcal{Y} \). If \( \Delta \) 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

\[
\lim_{n \to +\infty} E(\hat{f}_n) = E(f^*) \quad \text{with probability 1.}
\]
Theorem 4

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

$$\lim_{n \to +\infty} \mathbb{E}(\hat{f}_n) = \mathbb{E}(f^*)$$

with probability 1.

Rem: First universal consistency result for general form of SPP, including $\mathcal{Y}$ of infinite cardinality.
Theorem 5

If \( \Delta \) 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

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

holds with probability 1 minus \( 8e^{-\tau} \) for any \( \tau > 0 \), with \( c \) a constant independent of \( n \) and \( \tau \).

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 \( \rho \);
- Not sure if the same strategy would work on infinite setting.
THEOREM 5

If $\Delta$ 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 $g^*$...
**Theorem 5**

If $\Delta$ 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 $g^*$...

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

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with $c$ a constant independant of $n$ and $\tau$. 
THEOREM 5

If $\Delta$ 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 independent of $n$ and $\tau$.

Rem:
- Not possible to prove uniform convergence rates according to ref [25];
**Theorem 5**

If $\Delta$ 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 independent of $n$ and $\tau$.

*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 $\rho$;
**Theorem 5**

If $\Delta$ 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 independent of $n$ and $\tau$.

**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 $\rho$;
- Not sure if the same strategy would work on infinite setting.
Experiments


\[ \Delta_{\text{rank}}(y, y') = \frac{1}{2} \sum_{i,j=1}^{M} \gamma(y')_{ij} \left(1 - \text{sign}(y_i - y_j)\right) \]

<table>
<thead>
<tr>
<th>Rank Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear</strong> [7]</td>
</tr>
<tr>
<td><strong>Hinge</strong> [27]</td>
</tr>
<tr>
<td><strong>Logistic</strong> [28]</td>
</tr>
<tr>
<td><strong>SVM Struct</strong> [4]</td>
</tr>
<tr>
<td><strong>Alg. 1</strong></td>
</tr>
</tbody>
</table>

Table 1: Normalized \( \Delta_{\text{rank}} \) for ranking methods on the MovieLens dataset [29].
Experiments - Digit reconstruction

\[ \Delta_G(y, y') = 1 - k_G(y, y') \]

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

\[ \Delta_R(y, y') = \text{”Recognition” loss wrt SVM classifier} \]

<table>
<thead>
<tr>
<th>Loss</th>
<th>KDE [18] (Gaussian)</th>
<th>Alg. 1 (Hellinger)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta_G)</td>
<td>0.149 ± 0.013</td>
<td>0.172 ± 0.011</td>
</tr>
<tr>
<td>(\Delta_H)</td>
<td>0.736 ± 0.032</td>
<td>0.647 ± 0.017</td>
</tr>
<tr>
<td>(\Delta_R)</td>
<td>0.294 ± 0.012</td>
<td>0.193 ± 0.015</td>
</tr>
</tbody>
</table>

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