# Structured Prediction via Implicit Embeddings

Alessandro Rudi

Imaging and Machine Learning, April 1st, Paris

Inria, École normale supérieure
In collaboration with: Carlo Ciliberto, Lorenzo Rosasco, Francis Bach

# Structured Prediction

$$\mathcal{X} \xrightarrow{f} \mathcal{Y}$$

**Image Captioning**
(also Localization
Segmentation
Classification)


→



**Movie Ranking**

NETFLIX
user:127
→


$> \cdots >$

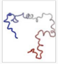
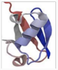

$> \cdots >$



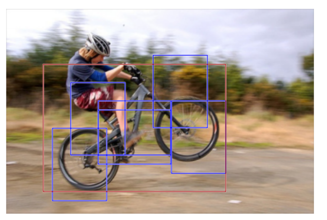
**Speech Recognition**

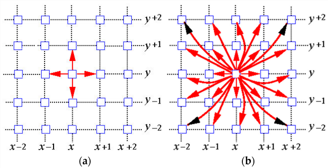

→
"Ok Google"

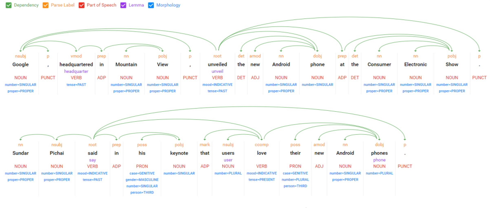
**Protein Folding**

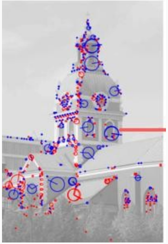

→


# Structured Prediction





# Supervised Learning

- $\mathcal{X}$ input space, $\mathcal{Y}$ output space,
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ loss function,
- $\rho$ probability on $\mathcal{X} \times \mathcal{Y}$.

$$f^\star = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f), \quad \mathcal{E}(f) := \mathbb{E}[\ell(y, f(x))].$$

given only the dataset $(x_i, y_i)_{i=1}^n$ sampled independently from $\rho$.

# Supervised learning: Goal

Given the dataset $(x_i, y_i)_{i=1}^n$ sampled independently from $\rho$, produce $\widehat{f}_n$, such that

## Consistency

$$\lim_{n \to \infty} \mathcal{E}(\widehat{f}_n) = \mathcal{E}(f^\star), \quad a.s.$$

## Learning rates

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \leq c(n), \quad w.h.p.$$

# State of the art: Vector-valued case

$\mathcal{Y}$ is a vector space

- choose suitable $\mathcal{G} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$ (usually a convex function space)
- solve *empirical risk minimization*

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda R(f).$$

- Well known methods: Linear models, generalized linear models, Kernel machines, Kernel SVM. Easy to optimize.
- Consistency and (optimal) learning rates for many losses

# State of the art: Structured case

$\mathcal{Y}$ arbitrary how do we parametrize $\mathcal{G}$ and learn $\widehat{f}$?

## Surrogate approaches

+ Clear theory
- Only for special cases (e.g. classification, ranking, multi-labeling etc.) [Bartlett et al '06, Duchi et al '10, Mroueh et al '12, Gao et al. '13]

## Score learning techniques

+ General algorithmic framework (e.g. StructSVM [Tsochandaridis et al '05])
- Limited Theory ( [McAllester '06] )

# Supervised learning with structure

Is it possible to

(a) have best of both worlds? (general algorithmic framework with clear theory)

(b) learn leveraging the local structure of the input and the output?

We will address (a), (b) using *implicit* embeddings

(related techniques: Cortes et al. 2005; Geurts, Wehenkel, d'Alché Buc '06; Kadri et al. '13; Brouard, Szafranski, d'Alché Buc '16)

# Table of contents

# Structured learning with implicit embeddings

# Characterizing the target function

$$f^\star = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(f(x), y)].$$

# Characterizing the target function

$$f^\star = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(f(x), y)].$$

Pointwise characterization

$$f^\star(x) = \operatorname*{argmin}_{y'\in\mathcal{Y}} \mathbb{E}[\ell(y', y) \mid x]$$

# Characterizing the target function

$$\tilde{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \mathbb{E}[\ell(y', y) \mid x]$$

$$\begin{aligned}
\mathbb{E}[\ell(\tilde{f}(x), y)] &= \mathbb{E}_x[\mathbb{E}[\ell(\tilde{f}(x), y)|x]] \\
&= \mathbb{E}_x[\inf_{y' \in \mathcal{Y}} \mathbb{E}[\ell(y', y)|x]] \\
&\leq \mathbb{E}[\ell(f(x), y)], \quad \forall f : \mathcal{X} \to \mathcal{Y}.
\end{aligned}$$

Then $\mathcal{E}(\tilde{f}) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f)$ (measurability issues solved via Berge maximum theory for measurable functions).

# Implicit embedding

**A1.** There exists Hilbert space $\mathcal{H}$ and $\psi, \phi : \mathcal{Y} \to \mathcal{H}$, bounded continuous such that

$$\ell(y', y) := \langle \psi(y'), \phi(y) \rangle .$$

**Theorem** (Ciliberto, Rosasco, Rudi '16)
A1 *is satisfied*

1. *for any loss* $\ell$ *when* $\mathcal{Y}$ *discrete space*
2. *for any smooth loss* $\ell$ *when* $\mathcal{Y} \subset \mathbb{R}^d$ *compact*
3. *for any smooth loss* $\ell$ *when* $\mathcal{Y} \subseteq \mathcal{M}$ *with* $\mathcal{M}$ *compact manifold*

# Idea for a unified approach

When **A1** holds

$$f^\star(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \mathbb{E}[\ell(y', y) \mid x]$$

# Idea for a unified approach

When **A1** holds

$$f^\star(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \mathbb{E}[\langle \psi(y'), \phi(y) \rangle \mid x]$$

# Idea for a unified approach

When **A1** holds

$$f^\star(x) = \underset{y' \in \mathcal{Y}}{\operatorname{argmin}} \langle \psi(y'), \mathbb{E}[\phi(y) \mid x] \rangle$$

# Idea for a unified approach

When **A1** holds

$$f^\star(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \langle \psi(y'), \mu^\star(x) \rangle$$

with $\mu^\star(x) = \mathbb{E}[\phi(y)|x]$ conditional expectation of $\phi(y)$ given $x$

# The estimator

Given $\widehat{\mu}$ estimating $\mu^\star$, define

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \langle \psi(y'), \widehat{\mu}(x) \rangle$$

# How to compute $\widehat{\mu}$

$\mu^\star = \mathbb{E}[\phi(y)|x]$ is characterized by

$$\mu^\star = \operatorname*{argmin}_{\mu:\mathcal{X}\to\mathcal{H}} \mathbb{E}[\|\mu(x) - \phi(y)\|^2]$$

# How to compute $\widehat{\mu}$

$\mu^\star = \mathbb{E}[\phi(y)|x]$ is characterized by

$$\mu^\star = \operatorname*{argmin}_{\mu:\mathcal{X}\to\mathcal{H}} \mathbb{E}[\|\mu(x) - \phi(y)\|^2]$$

use standard techniques for vector valued problems. Given $\mathcal{G}$ suitable space of functions

$$\widehat{\mu} = \operatorname*{argmin}_{\mu\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \|\mu(x_i) - \phi(y)\|^2 \ + \ \lambda\|\mu\|^2.$$

# $\mathcal{G}$ space of linear functions

Let $\mathcal{X}$ be a vector space and $\mathcal{G} = \mathcal{X} \otimes \mathcal{H}$, then

$$\widehat{\mu}(x) \quad = \quad \sum_{i=1}^{n} \alpha_i(x)\, \phi(y_i),$$

where

$$\alpha_i(x) := [(K + \lambda n I)^{-1} v(x)]_i,$$

and $v(x) = (x^\top x_1, \ldots x^\top x_n) \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ $K_{i,j} = x_i^\top x_j$.

# non-parametric model

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on $\mathcal{X}$. Denote by $\mathcal{F}$ the *reproducing kernel Hilbert space* induced by $k$ over $\mathcal{X}$. Let $\mathcal{G} = \mathcal{F} \otimes \mathcal{H}$, then

$$\widehat{\mu}(x) \quad = \quad \sum_{i=1}^{n} \alpha_i(x)\, \phi(y_i),$$

where

$$\alpha_i(x) := [(K + \lambda n I)^{-1} v(x)]_i,$$

and $v(x) = (k(x, x_1), \ldots k(x, x_n)) \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ $K_{i,j} = k(x_i, x_j)$.

# Algorithm and properties

# Explicit representation of $\widehat{f}$

When $\widehat{\mu}$ is a non-parametric model, then

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \langle \psi(y'), \widehat{\mu}(x) \rangle$$

# Explicit representation of $\widehat{f}$

When $\widehat{\mu}$ is a non-parametric model, then

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \left\langle \psi(y'), \sum_{i=1}^{n} \alpha_i(x)\phi(y_i) \right\rangle$$

# Explicit representation of $\widehat{f}$

When $\widehat{\mu}$ is a non-parametric model, then

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \left\langle \psi(y'), \phi(y_i) \right\rangle$$

# Explicit representation of $\widehat{f}$

When $\widehat{\mu}$ is a non-parametric model, then

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \ell(y', y_i).$$

# Explicit representation of $\widehat{f}$

When $\widehat{\mu}$ is a non-parametric model, then

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \ell(y', y_i).$$

No need to know $\mathcal{H}, \phi, \psi$ to run the algorithm!

# Recap

- Given $\ell$ satisfying **A1**
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, kernel on $\mathcal{X}$

The proposed estimator has the form

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \ell(y', y_i),$$

with $\alpha_i(x) := [(K + \lambda n I)^{-1} v(x)]_i$, and $v(x) = (k(x, x_1), \ldots k(x, x_n)) \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ $K_{i,j} = k(x_i, x_j)$.

# Recap

- Given $\ell$ satisfying **A1**
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, kernel on $\mathcal{X}$

The proposed estimator has the form

$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \ell(y', y_i),$$

with $\alpha_i(x) := [(K + \lambda n I)^{-1} v(x)]_i$, and $v(x) = (k(x, x_1), \ldots k(x, x_n)) \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$ $K_{i,j} = k(x_i, x_j)$.

- Applicable to a wide family of problems (no need to know $\mathcal{H}, \phi, \psi$)
- Only optimization on $\mathcal{Y}$ and not on $\{f : \mathcal{X} \to \mathcal{Y}\} = \mathcal{Y}^{\mathcal{X}}$
- Generalization properties?

**Theorem (Comparison inequality)**
*Let $\ell$ satisfy* A1*. For any $\widehat{\mu} : \mathcal{X} \to \mathcal{H}$,*

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star) \leq 2c_\psi \sqrt{\mathbb{E}[\|\widehat{\mu}(x) - \mu^\star(x)\|^2]}.$$

*with* $c_\psi = \sup_{y' \in \mathcal{Y}} \|\psi(y)\|$.

# Consistency of $\widehat{f}$

Theorem (Universal consistency - Ciliberto, Rosasco, Rudi '16)

*Let $\ell$ satisfy* A1 *and $k$ be a universal kernel. Let $\lambda = n^{-1/4}$, then*

$$\lim_{n \to \infty} \mathcal{E}(\widehat{f}) = \mathcal{E}(f^\star),$$

*with probability* 1

# Learning rates of $\widehat{f}$

**Theorem (Rates - Ciliberto, Rosasco, Rudi '16)**
*Let $\ell$ satisfy* A1 *and $\mu^\star \in \mathcal{G}$. Let $\lambda = n^{-1/2}$, then*

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star) \;\; \leq \;\; 2c_\psi \, n^{-1/4}, \qquad w.h.p.$$

# Check point

We provide a framework for structured prediction with

- theoretical guarantees as empirical risk minimization
- explicit algorithm applicable on wide family of problems $(\mathcal{Y}, \ell)$
- some important existing algorithms are covered by this framework (not seen here)

Case studies:

- ranking with different losses (Korba, Garcia, d'Alché-Buc '18)
- Output Fisher Embeddings (Djerrab, Garcia, Sangnier, d'Alché-Buc '18)
- $\mathcal{Y} =$ manifolds, $\ell =$ geodesic distance (Ciliberto et al. 18)
- $\mathcal{Y} =$ probability space, $\ell =$ wasserstein distance (Luise et al. 18)

Refinements of the analysis:

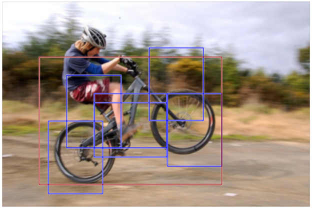
- different derivation (Osokin, Bach, Lacoste-Julien '17; Goh '18)
- determination of the constant $c_\psi$ in terms of $\log |\mathcal{Y}|$ for discrete sets (Nowak, Bach, Rudi '18; Struminsky et al. '18)

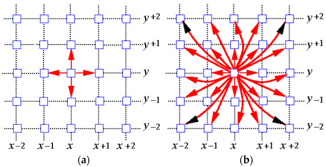
Extensions:

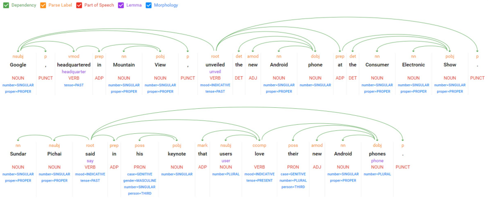
- application to multitask-learning (Ciliberto, Rosasco, Rudi '17)
- beyond least squares surrogate (Nowak, Bach, Rudi '19)
- regularizing with trace norm (Luise, Stamos, Pontil, Ciliberto '19)
- localized structured prediction (Ciliberto, Bach, Rudi '18)

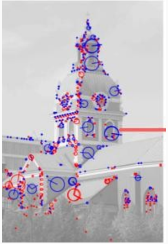
# Leveraging local structure

# Local Structure





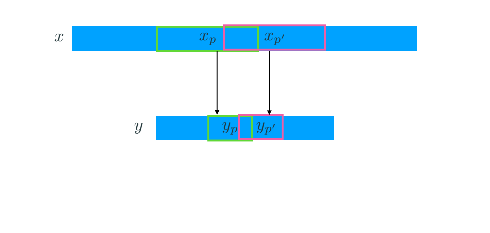



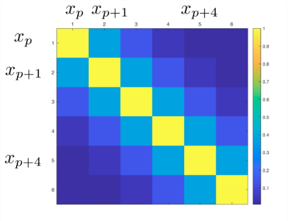
# Parts and locality

We are interested in problems where we have a set of *parts* $P$ that capture:

*Inter*-locality




*Intra*-locality




## Examples

Images: (overlapping) patches of a fixed size, overlapping pyramids on patches, ...

Audio: (overlapping) windows in time/frequency space, ...

# Loss Functions

$$\ell(y', y) = \sum_{p \in P} \ell_0([y']_p, [y]_p)$$

- set $P$ indicizes the parts
- $\ell_0$ loss on parts
- $[y]_p$ is the $p$-th part of $y$

# Examples of loss functions

$$\ell(y', y) = \sum_{p \in P} \ell_0([y']_p, [y]_p)$$

Many losses in computer vision, multilabeling, multitask learning (Ciliberto, Bach, Rudi '18)

### Example (Hamming like loss is implicitly by parts)

Let $\mathcal{Y}$ be space of circular sequences of length $d$. Let $P$ the set of subsequences of length $s < d$.

$$\ell(y', y) = \frac{1}{d} \sum_{i=1}^{d} \bar{\ell}(y'_i, y_i) = \frac{1}{|P|} \sum_{p \in P} \ell_0([y']_p, [y]_p),$$

$$\ell_0([y']_p, [y]_p) = \frac{1}{s} \sum_{i=0}^{s-1} \bar{\ell}(y'_{p+i}, y_{p+i}).$$

# Building the estimator

Assume that $\ell_0$ satisfied **A1**. Then

$$\ell(y', y) = \sum_{p \in P} \langle \psi([y']_p), \phi([y]_p) \rangle \,,$$

and the target function is characterized by

$$f^\star(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{p \in P} \langle \psi([y']_p), \mu^\star(x, p) \rangle \,,$$

with

$$\mu^\star(x, p) = \mathbb{E}[\phi([y]_p) \mid x]$$

conditional expectation of the $p$-th part of $y$, given $x$.

# Learning $\mu^\star(x, p)$

Analogously to the other case we have

$$\mu^\star = \underset{\mu:\mathcal{X}\times P\to\mathcal{H}}{\operatorname{argmin}} \sum_{p\in P} \mathbb{E}[\|\mu(x,p) - \phi([y]_p)\|^2] + \lambda\|\mu\|^2.$$

Applying empirical risk minimization

$$\widehat{\mu} = \underset{\mu\in\mathcal{G}}{\operatorname{argmin}} \frac{1}{n} \sum_{p\in P} \sum_{i=1}^{n} \|\mu(x,p) - \phi([y]_p)\|^2 + \lambda\|\mu\|^2.$$

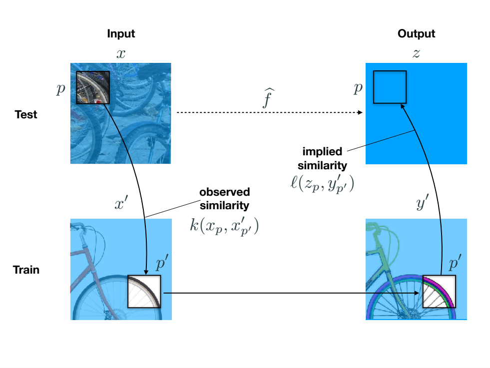
# Non-parametric estimator for $\mu^\star$

Selecting $\mathcal{G} = \mathcal{F} \otimes \mathcal{H}$ with $\mathcal{F}$ a reproducing kernel on $X \times P$, we have

$$\widehat{\mu}(x,p) = \sum_{p' \in P} \sum_{i=1}^{n} \alpha_{i,p'}(x,p)\phi([y_i]'_p),$$

with $\alpha_i(x,p) = [(K + \lambda n P I)^{-1} v(x,p)]_{i,p'}$, $v(x,p)_{i,p'} = k((x,p),(x_i,p'))$ with $v \in \mathbb{R}^{n|P|}$ and $K \in \mathbb{R}^{n|P| \times n|P|}$ with $K_{(i,p'),(j,p'')} = k((x,p'),(x_i,p'')$.

# Final estimator




$$\widehat{f}(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \sum_{p, p' \in P} \sum_{i=1}^{n} \alpha_{i,p'}(x, p) \ell_0([y']_p, [y]_{p'})$$

# Theoretical Properties

- $k((x,p),(x',p')) = k([x]_p,[x']_{p'})$
- $[y]_p$ conditional independent from $x$, given $[x]_p$
- $\mathrm{cov}_k([x]_p,[x]_{p'}) \leq \exp(-\gamma d(p,p'))$, $\gamma > 0$, $d$ distance on parts and $cov_k$ covariance with respect to the kernel $k$

**Theorem (Ciliberto, Bach, Rudi, '18)**

*When $\ell_0$ satisfied* A1 *and under the assumptions above,*

$$\mathbb{E}\,\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star) \leq \left(\frac{c_0 + q_{\gamma,|P|}}{n|P|}\right)^{1/4},$$

*where* $q_{\gamma,|P|} = \frac{1}{|P|}\sum_{p,p'\in P} e^{-\gamma d(p,p')}$.

# Theoretical Properties

Implications: under inter-locality

- and no intra-locality (i.e. $\gamma \approx 0$) then $q_{\gamma,|P|} \approx |P|$ and

$$\mathbb{E}\,\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star) = O(n^{-1/4}).$$

- and intra-locality (i.e. $\gamma \gg 0$) then $q_{\gamma,|P|} = O(1)$ and

$$\mathbb{E}\,\mathcal{E}(\widehat{f}) - \mathcal{E}(f^\star) = O((n|P|)^{-1/4}).$$

# Conclusions

Framework for structured prediction with

- theoretical guarantees as empirical risk minimization
- explicit algorithm applicable on wide family of problems $(\mathcal{Y}, \ell)$
- some important existing algorithms are covered by this framework (not seen here)
- adaptive to local structure

Future work

- wide experimental validation (CV: deblurring and super-resolution)
- generalization to different estimators for $\widehat{\mu}$
- integration with DNN

# Conclusions

Thanks!