A Consistent Regularization Approach for Structured Prediction

Carlo Ciliberto [*,1] *cciliber@mit.edu*
Alessandro Rudi [*,1,2] *ale_rudi@mit.edu*
Lorenzo Rosasco [1,2] *lrosasco@mit.edu*

[1] Laboratory for Computational and Statistical Learning - Istituto Italiano di Tecnologia, Genova, Italy & Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[2] Università degli Studi di Genova, Genova, Italy.
[*] **Equal contribution**

Abstract

We propose and analyze a regularization approach for structured prediction problems. We characterize a large class of loss functions that allows to naturally embed structured outputs in a linear space. We exploit this fact to design learning algorithms using a surrogate loss approach and regularization techniques. We prove universal consistency and finite sample bounds characterizing the generalization properties of the proposed method. Experimental results are provided to demonstrate the practical usefulness of the proposed approach.

1 Introduction

Many machine learning applications require dealing with data-sets having complex structures, e.g. natural language processing, image segmentation, reconstruction or captioning, pose estimation, protein folding prediction to name a few [1, 2, 3]. Structured prediction problems pose a challenge for classic off-the-shelf learning algorithms for regression or binary classification. This has motivated the extension of methods such as support vector machines to structured problems [4]. Dealing with structured prediction problems is also a challenge for learning theory. While the theory of empirical risk minimization provides a very general statistical framework, in practice it needs to be complemented with an ad-hoc analysis for each specific setting. Indeed, in the last few years, an effort has been made to analyze specific structured problems, such as multiclass classification [5], multi-labeling [6], ranking [7] or quantile estimation [8]. A natural question is whether a unifying learning framework can be developed to address a wide range of problems from theory to algorithms.

This paper takes a step in this direction, proposing and analyzing a general regularization approach to structured prediction. Our starting observation is that for a large class of these problems, we can define a natural embedding of the associated loss functions into a linear space. This allows to define a (least squares) surrogate problem of the original structured one, that is cast within a multi-output regularized learning framework [9, 10, 11]. We prove that by solving the surrogate, we are able to recover the exact solution of the original structured problem. The corresponding algorithm essentially generalizes approaches considered in [12, 13, 14, 15, 16]. We study the generalization properties of the proposed approach, establishing universal consistency as well as finite sample bounds.

The rest of this paper is organized as follows: in Sec. 2 we introduce the structured prediction problem in its generality and present our algorithm to approach it. In Sec. 3 we introduce and discuss a surrogate framework for structured prediction, from which we derive our algorithm. In Sec. 4, we analyze the theoretical properties of the proposed algorithm. In Sec. 5 we draw connections with previous work in structured prediction. Sec. 6 reports promising experimental results on a variety of structured prediction problems. Sec. 7 concludes the paper outlining relevant directions for future research.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 A Regularization Approach to Structured prediction

The goal of supervised learning is to learn functional relations $f : \mathcal{X} \to \mathcal{Y}$ between two sets $\mathcal{X}, \mathcal{Y}$, given a finite number of examples. In particular in this work we are interested to *structured prediction*, namely the case where $\mathcal{Y}$ is a set of structured outputs (such as histograms, graphs, time sequences, points on a manifold, etc.). Moreover, structure on $\mathcal{Y}$ can be implicitly induced by a suitable loss $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (such as edit distance, ranking error, geodesic distance, indicator function of a subset, etc.). Then, the problem of structured prediction becomes

$$\underset{f:\mathcal{X}\to\mathcal{Y}}{\text{minimize}} \quad \mathcal{E}(f), \qquad \text{with} \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \triangle(f(x), y) \; d\rho(x, y) \tag{1}$$

and the goal is to find a good estimator for the minimizer of the above equation, given a finite number of (training) points $\{(x_i, y_i)\}_{i=1}^n$ sampled from a unknown probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$. In the following we introduce an estimator $\hat{f} : \mathcal{X} \to \mathcal{Y}$ to approach Eq. (1). The rest of this paper is devoted to prove that $\hat{f}$ it a consistent estimator for a minimizer of Eq. (1).

Our Algorithm for Structured Prediction. In this paper we propose and analyze the following estimator

$$\hat{f}(x) = \underset{y\in\mathcal{Y}}{\text{argmin}} \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i) \quad \text{with} \quad \alpha(x) = (\mathbf{K} + n\lambda I)^{-1}\mathbf{K}_x \in \mathbb{R}^n \tag{Alg. 1}$$

given a positive definite kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and training set $\{(x_i, y_i)\}_{i=1}^n$. In the above expression, $\alpha_i(x)$ is i-th entry in $\alpha(x)$, $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix $\mathbf{K}_{i,j} = k(x_i, x_j)$, $\mathbf{K}_x \in \mathbb{R}^n$ the vector with entires $(\mathbf{K}_x)_i = k(x, x_i)$, $\lambda > 0$ a regularization parameter and I the identity matrix.

From a computational perspective, the procedure in Alg. 1 is divided in two steps: a *learning* step where input-dependents weights $\alpha_i(\cdot)$ are computed (which essentially consists in solving a kernel ridge regression problem) and a *prediction* step where the $\alpha_i(x)$-weighted linear combination in Alg. 1 is optimized, leading to a prediction $\hat{f}(x)$ given an input x. The idea of a similar two-steps strategy goes back to standard approaches for structured prediction and was originally proposed in [17], where a "score" function $F(x, y)$ was learned to estimate the "likelihood" of a pair (x, y) sampled from ρ, and then used in $\hat{f}(x) = \text{argmin}_{y\in\mathcal{Y}} -F(x, y)$, to predict the best $\hat{f}(x) \in \mathcal{Y}$ given $x \in \mathcal{X}$. This strategy was extended in [4] for the popular *SVMstruct* and adopted also in a variety of approaches for structured prediction [1, 12, 14].

Intuition. While providing a principled derivation of Alg. 1 for a large class of loss functions is a main contribution of this work, it is useful to first consider the special case where $\triangle$ is induced by a reproducing kernel $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ on the output set, such that

$$\triangle(y, y') = h(y, y) - 2h(y, y') + h(y', y'). \tag{2}$$

This choice of $\triangle$ was originally considered in *Kernel Dependency Estimation (KDE)* [18]. In particular, for the special case of normalized kernels (i.e. $h(y, y) = 1 \;\; \forall y \in \mathcal{Y}$), Alg. 1 essentially reduces to [12, 13, 14] and recalling their derivation is insightful. Note that, since a kernel can be written as $h(y, y') = \langle \psi(y), \psi(y') \rangle_{\mathcal{H}_\mathcal{Y}}$, with $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ a non-linear map into a feature space $\mathcal{H}_\mathcal{Y}$ [19], then Eq. (2) can be rewritten as

$$\triangle(f(x), y') = \|\psi(f(x)) - \psi(y')\|^2_{\mathcal{H}_\mathcal{Y}}. \tag{3}$$

Directly minimizing the equation above with respect to f is generally challenging due to the non linearity ψ. A possibility is to replace $\psi \circ f$ by a function $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ that is easier to optimize. We can then consider the regularized problem

$$\underset{g\in\mathcal{G}}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda \|g\|^2_\mathcal{G} \tag{4}$$

with $\mathcal{G}$ a space of functions[1] $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ of the form $g(x) = \sum_{i=1} k(x, x_i)c_i$ with $c_i \in \mathcal{H}_\mathcal{Y}$ and k a reproducing kernel. Indeed, in this case the solution to Eq. (4) is

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x)\psi(y_i) \quad \text{with} \quad \alpha(x) = (\mathbf{K} + n\lambda I)^{-1}\mathbf{K}_x \in \mathbb{R}^n \tag{5}$$

[1] $\mathcal{G}$ is the reproducing kernel Hilbert space for vector-valued functions [9] with inner product $\langle k(x_i, \cdot)c_i, k(x_j, \cdot)c_j \rangle_\mathcal{G} = k(x_i, x_j)\langle c_i, c_j \rangle_{\mathcal{H}_\mathcal{Y}}$

where the α_i are the same as in Alg. 1. Since we replaced $\triangle(f(x), y)$ by $\|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}}$, a natural question is how to recover an estimator $\hat{f}$ from $\hat{g}$. In [12] it was proposed to consider

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \|\psi(y) - \hat{g}(x)\|^2_{\mathcal{H}_\mathcal{Y}} = \operatorname*{argmin}_{y \in \mathcal{Y}} h(y, y) - 2 \sum_{i=1}^{n} \alpha_i(x) h(y, y_i), \tag{6}$$

which corresponds to Alg. 1 when h is a normalized kernel.

The discussion above provides an intuition on how Alg. 1 is derived but raises also a few questions. First, it is not clear if and how the same strategy could be generalized to loss functions that do not satisfy Eq. (2). Second, the above reasoning hinges on the idea of replacing $\hat{f}$ with $\hat{g}$ (and then recovering $\hat{f}$ by Eq. (6)), however it is not clear whether this approach can be justified theoretically. Finally, we can ask what are the statistical properties of the resulting algorithm. We address the first two questions in the next section, while the rest of the paper is devoted to establish universal consistency and generalization bounds for algorithm Alg. 1.

3 Surrogate Framework and Derivation

To derive Alg. 1 we consider ideas from surrogate approaches [20, 21, 7] and in particular [5]. The idea is to tackle Eq. (1) by substituting $\triangle(f(x), y)$ with a "relaxation" $L(g(x), y)$ on a space $\mathcal{H}_\mathcal{Y}$, that is easy to optimize. The corresponding surrogate problem is

$$\operatorname*{minimize}_{g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}} \quad \mathcal{R}(g), \qquad \text{with} \qquad \mathcal{R}(g) = \int_{\mathcal{X} \times \mathcal{Y}} L(g(x), y) \; d\rho(x, y), \tag{7}$$

and the question is how a solution g^* for the above problem can be related to a minimizer f^* of Eq. (1). This is made possible by the requirement that there exists a *decoding* $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$, such that

$$\textit{Fisher Consistency:} \quad \mathcal{E}(d \circ g^*) = \mathcal{E}(f^*), \tag{8}$$

$$\textit{Comparison Inequality:} \quad \mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq \varphi(\mathcal{R}(g) - \mathcal{R}(g^*)), \tag{9}$$

hold for all $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is such that $\varphi(s) \to 0$ for $s \to 0$. Indeed, given an estimator $\hat{g}$ for g^*, we can "decode" it considering $\hat{f} = d \circ \hat{g}$ and use the *excess risk* $\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)$ to control $\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)$ via the comparison inequality in Eq. (9). In particular, if $\hat{g}$ is a data-dependent predictor trained on n points and $\mathcal{R}(\hat{g}) \to \mathcal{R}(g^*)$ when $n \to +\infty$, we automatically have $\mathcal{E}(\hat{f}) \to \mathcal{E}(f^*)$. Moreover, if φ in Eq. (9) is known explicitly, generalization bounds for $\hat{g}$ are automatically extended to $\hat{f}$.

Provided with this perspective on surrogate approaches, here we revisit the discussion of Sec. 2 for the case of a loss function induced by a kernel h. Indeed, by assuming the surrogate $L(g(x), y) = \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}}$, Eq. (4) becomes the empirical version of the surrogate problem at Eq. (7) and leads to an estimator $\hat{g}$ of g^* as in Eq. (5). Therefore, the approach in [12, 14] to recover $\hat{f}(x) = \operatorname{argmin}_y L(g(x), y)$ can be interpreted as the result $\hat{f}(x) = d \circ \hat{g}(x)$ of a suitable decoding of $\hat{g}(x)$. An immediate question is whether the above framework satisfies Eq. (8) and (9). Moreover, we can ask if the same idea could be applied to more general loss functions.

In this work we identify conditions on $\triangle$ that are satisfied by a large family of functions and moreover allow to design a surrogate framework for which we prove Eq. (8) and (9). The first step in this direction is to introduce the following assumption.

Assumption 1. *There exists a separable Hilbert space $\mathcal{H}_\mathcal{Y}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$, a continuous embedding $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and a bounded linear operator $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$, such that*

$$\triangle(y, y') = \langle \psi(y), V \psi(y') \rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall y, y' \in \mathcal{Y} \tag{10}$$

Asm. 1 is similar to Eq. (3) and in particular to the definition of a reproducing kernel. Note however that by not requiring V to be positive semidefinite (or even symmetric), we allow for a surprisingly wide range of functions beyond kernel functions. Indeed, below we give some examples of functions that satisfy Asm. 1 (see supplementary material Sec. C for more details):

Example 1. The following functions of the form $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1:

1. *Any loss on $\mathcal{Y}$ of finite cardinality*. Several problems belong to this setting, such as Multi-Class Classification, Multi-labeling, Ranking, predicting Graphs (e.g. protein foldings).
2. *Regression and Classification Loss Functions*. Least-squares, Logistic, Hinge, ϵ-insensitive, τ-Pinball.
3. *Robust Loss Functions*. Most loss functions used for *robust estimation* [22] such as the absolute value, Huber, Cauchy, German-McLure, "Fair" and $L_2 - L1$. See [22] or the supplementary material for their explicit formulation.
4. *KDE*. Loss functions $\triangle$ induced by a kernel such as in Eq. (2).
5. *Distances on Histograms/Probabilities*. The χ^2 and the Hellinger distances.
6. *Diffusion distances on Manifolds*. The squared diffusion distance induced by the heat kernel (at time $t > 0$) on a compact Reimannian manifold without boundary [23].

The Least Squares Loss Surrogate Framework. Asm. 1 implicitly defines the space $\mathcal{H}_\mathcal{Y}$ similarly to Eq. (3). The following result motivates the choice of the least squares surrogate and moreover suggests a possible choice for the decoding.

Lemma 1. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ bounded. Then the expected risk in Eq. (1) can be written as*

$$\mathcal{E}(f) = \int_\mathcal{X} \langle \psi(f(x)), Vg^*(x) \rangle_{\mathcal{H}_\mathcal{Y}} \; d\rho_\mathcal{X}(x) \tag{11}$$

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

$$\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} \; d\rho(x,y). \tag{12}$$

Lemma 1 shows how Eq. (12) arises naturally as surrogate problem. In particular, Eq. (11) suggests to choose the decoding

$$d(h) = \operatorname*{argmin}_{y\in\mathcal{Y}} \; \langle \, \psi(y) \, , \, Vh \, \rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall h \in \mathcal{H}_\mathcal{Y}, \tag{13}$$

since $d \circ g^*(x) = \arg\min_{y\in\mathcal{Y}} \langle \psi(y), Vg^*(x) \rangle$ and therefore $\mathcal{E}(d \circ g^*) \leq \mathcal{E}(f)$ for any measurable $f : \mathcal{X} \to \mathcal{Y}$, leading to Fisher Consistency. We formalize this in the following result.

Theorem 2. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact set. Then, for every measurable $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13), the following holds*

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*) \tag{14}$$

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq c_\triangle \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)}. \tag{15}$$

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

Thm. 2 shows that for all $\triangle$ satisfying Asm. 1, the corresponding surrogate framework identified by the surrogate in Eq. (12) and decoding Eq. (13) satisfies Fisher consistency Eq. (14) and the comparison inequality in Eq. (15). We recall that a finite set $\mathcal{Y}$ is always compact, and moreover, assuming the discrete topology on $\mathcal{Y}$, we have that any $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ is continuous. Therefore, Thm. 2 applies in particular to any structured prediction problem on $\mathcal{Y}$ with finite cardinality.

Thm. 2 suggest to approach structured prediction by first learning $\hat{g}$ and then decoding it to recover $\hat{f} = d \circ \hat{g}$. A natural question is how to choose $\hat{g}$ in order to compute $\hat{f}$ in practice. In the rest of this section we propose an approach to this problem.

Derivation for Alg. 1. Minimizing $\mathcal{R}$ in Eq. (12) corresponds to a vector-valued regression problem [9, 10, 11]. In this work we adopt an empirical risk minimization approach to learn $\hat{g}$ as in Eq. (4). The following result shows that combining $\hat{g}$ with the decoding in Eq. (13) leads to the $\hat{f}$ in Alg. 1.

Lemma 3. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact set. Let $\hat{g} : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be the minimizer of Eq. (4). Then, for all $x \in \mathcal{X}$*

$$d \circ \hat{g}(x) \;=\; \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^n \alpha_i(x) \triangle (y, y_i) \qquad \alpha(x) = (\mathbf{K} + n\lambda I)^{-1} \mathbf{K}_x \in \mathbb{R}^n \tag{16}$$

Lemma 3 concludes the derivation of Alg. 1. An interesting observation is that computing $\hat{f}$ does not require explicit knowledge of the embedding ψ and the operator V, which are implicitly encoded within the loss $\triangle$ by Asm. 1. In analogy to the *kernel trick* [24] we informally refer to such assumption as the "loss trick". We illustrate this effect with an example.

Example 2 (Ranking)**.** *In ranking problems the goal is to predict ordered sequences of a fixed number ℓ of labels. For these problems, $\mathcal{Y}$ corresponds to the set of all ordered sequences of ℓ labels and has cardinality $|\mathcal{Y}| = \ell!$, which is typically dramatically larger than the number n of training examples (e.g. for $\ell = 15$, $\ell! \simeq 10^{12}$). Therefore, given an input $x \in \mathcal{X}$, directly computing $\hat{g}(x) \in \mathbb{R}^{|\mathcal{Y}|}$ is impractical. On the opposite, the loss trick allows to express $d \circ \hat{g}(x)$ only in terms of the n weights $\alpha_i(x)$ in Alg. 1, making the computation of the argmin easier to approach in general. For details on the rank loss $\triangle_{rank}$ and the corresponding optimization over $\mathcal{Y}$, we refer to the empirical analysis of Sec. 6.*

In this section we have shown a derivation for the structured prediction algorithm proposed in this work. In Thm. 2 we have shown how the expected risk of the proposed estimator $\hat{f}$ is related to an estimator $\hat{g}$ via a comparison inequality. In the following we will make use of these results to prove consistency and generalization bounds for Alg. 1.

4 Statistical Analysis

In this section we study the statistical properties of Alg. 1 exploiting of the relation between the structured and surrogate problems characterized be the comparison inequality in Thm. 2. We begin our analysis by proving that Alg. 1 is *universally consistent*.

Theorem 4 (Universal Consistency)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1, $\mathcal{X}$ and $\mathcal{Y}$ be compact sets and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous universal reproducing kernel[2]. For any $n \in \mathbb{N}$ and any distribution ρ on $\mathcal{X} \times \mathcal{Y}$ let $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ be obtained by Alg. 1 with $\{(x_i, y_i)\}_{i=1}^n$ training points independently sampled from ρ and $\lambda_n = n^{-1/4}$. Then,*

$$\lim_{n \to +\infty} \mathcal{E}(\hat{f}_n) = \mathcal{E}(f^*) \qquad \text{with probability } 1 \tag{17}$$

Thm. 4 shows that, when the $\triangle$ satisfies Asm. 1, Alg. 1 approximates a solution f^* to Eq. (1) arbitrarily well, given a sufficient number of training examples. To the best of our knowledge this is the first consistency result for structured prediction in the general setting considered in this work and characterized by Asm. 1, in particular for the case of $\mathcal{Y}$ with infinite cardinality (dense or discrete).

The *No Free Lunch* Theorem [25] states that it is not possible to prove uniform convergence rates for Eq. (17). However, by imposing suitable assumptions on the regularity of g^* it is possible to prove generalization bounds for $\hat{g}$ and then, using Thm. 2, extend them to $\hat{f}$. To show this, it is sufficient to require that g^* belongs to $\mathcal{G}$ the reproducing kernel Hilbert space used in the ridge regression of Eq. (4). Note that in the proofs of Thm. 4 and Thm. 5, our analysis on $\hat{g}$ borrows ideas from [10] and extends their result to our setting for the case of $\mathcal{H}_\mathcal{Y}$ infinite dimensional (i.e. when $\mathcal{Y}$ has infinite cardinality). Indeed, note that in this case [10] cannot be applied to the estimator $\hat{g}$ considered in this work (see supplementary material Sec. B.3, Lemma 18 for details).

Theorem 5 (Generalization Bound)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1, $\mathcal{Y}$ be a compact set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a bounded continuous reproducing kernel. Let $\hat{f}_n$ denote the solution of Alg. 1 with n training points and $\lambda = n^{-1/2}$. If the surrogate risk $\mathcal{R}$ defined in Eq. (12) admits a minimizer $g^* \in \mathcal{G}$, then*

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 \, n^{-\frac{1}{4}} \tag{18}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant not depending on n and τ.

The bound in Thm. 5 is of the same order of the generalization bounds available for the least squares binary classifier [26]. Indeed, in Sec. 5 we show that in classification settings Alg. 1 reduces to least squares classification. This opens the way to possible improvements, as we discuss in the following.

[2] This is a standard assumption for universal consistency (see [21]). An example of continuous universal kernel is the Gaussian $k(x, x') = \exp(-\gamma \|x - x'\|^2)$, with $\gamma > 0$.

Remark 1 (Better Comparison Inequality)**.** *The generalization bounds for the least squares classifier can be improved by imposing regularity conditions on ρ via the Tsybakov condition [26]. This was observed in [26] for binary classification with the least squares surrogate, where a tighter comparison inequality than the one in Thm. 2 was proved. Therefore, a natural question is whether the inequality of Thm. 2 could be similarly improved, consequently leading to better rates for Thm. 5. Promising results in this direction can be found in [5], where the Tsybakov condition was generalized to the multi-class setting and led to a tight comparison inequality analogous to the one for the binary setting. However, this question deserves further investigation. Indeed, it is not clear how the approach in [5] could be further generalized to the case where $\mathcal{Y}$ has infinite cardinality.*

Remark 2 (Other Surrogate Frameworks)**.** *In this paper we focused on a least squares surrogate loss function and corresponding framework. A natural question is to ask whether other loss functions could be considered to approach the structured prediction problem, sharing the same or possibly even better properties. This question is related also to Remark 1, since different surrogate frameworks could lead to sharper comparison inequalities. This seems an interesting direction for future work.*

5 Connection with Previous Work

Binary and Multi-class Classification. It is interesting to note that in classification settings, Alg. 1 corresponds to the least squares classifier [26]. Indeed, let $\mathcal{Y} = \{1, \dots, \ell\}$ be a set of labels and consider the *misclassification* loss $\triangle(y, y') = 1$ for $y \neq y'$ and 0 otherwise. Then $\triangle(y, y') = e_y^\top V e_{y'}$ with $e_i \in \mathbb{R}^\ell$ the i-the element of the canonical basis of $\mathbb{R}^\ell$ and $V = \mathbf{1} - I$, where I is the $\ell \times \ell$ identity matrix and $\mathbf{1}$ the matrix with all entries equal to 1. In the notation of surrogate methods adopted in this work, $\mathcal{H}_\mathcal{Y} = \mathbb{R}^\ell$ and $\psi(y) = e_y$. Note that both Least squares classification and our approach solve the surrogate problem at Eq. (4)

$$\frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - e_{y_i}\|_{\mathbb{R}^\ell}^2 + \lambda \|g\|_\mathcal{G}^2 \tag{19}$$

to obtain a vector-valued predictor $\hat{g} : \mathcal{X} \to \mathbb{R}^\ell$ as in Eq. (5). Then, the least squares classifier $\hat{c}$ and the decoding $\hat{f} = d \circ \hat{g}$ are respectively obtained by

$$\hat{c}(x) = \operatorname*{argmax}_{i=1,\dots,\ell} \hat{g}(x) \qquad\qquad \hat{f}(x) = \operatorname*{argmin}_{i=1,\dots,\ell} V\hat{g}(x). \tag{20}$$

However, since $V = \mathbf{1} - I$, it is easy to see that $\hat{c}(x) = \hat{f}(x)$ for all $x \in \mathcal{X}$.

Kernel Dependency Estimation. In Sec. 2 we discussed the relation between KDE [18, 12] and Alg. 1. In particular, we have observed that if $\triangle$ is induced by a kernel $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ as in Eq. (2) and h is normalized, i.e. $h(y, y) = \kappa \; \forall y \in \mathcal{Y}$, with $\kappa > 0$, then algorithm Eq. (6) proposed in [12] leads to the same predictor as Alg. 1. Therefore, we can apply Thm. 4 and 5 to prove universal consistency and generalization bounds for methods such as [12, 14]. Some theoretical properties of KDE have been previously studied in [15] from a PAC Bayesian perspective. However, the obtained bounds do not allow to control the excess risk or establish consistency of the method. Moreover, note that when the kernel h is not normalized, the "decoding" in Eq. (6) is not equivalent to Alg. 1. In particular, given the surrogate solution g^*, applying Eq. (6) leads to predictors that do not minimize Eq. (1). As a consequence, approaches in [12, 13, 14] are not consistent in the general case.

Support Vector Machines for Structured Output. A popular approach to structured prediction is the *Support Vector Machine for Structured Outputs (SVMstruct)* [4] that extends ideas from the well-known SVM algorithm to the structured setting. One of the main advantages of SVMstruct is that it can be applied to a variety of problems since it does not impose strong assumptions on the loss. In this view, our approach shares similar properties, and in particular allows to consider $\mathcal{Y}$ of infinite cardinality. Moreover, we note that generalization studies for SVMstruct are available [3] (Ch. 11). However, it seems that these latter results do not allow to derive universal consistency of the method.

6 Experiments

In this section we report on preliminary experiments showing the performance of the proposed approach on simulated as well as real structured prediction problems.

	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].

Loss	**KDE [18]** (Gaussian)	**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.

Ranking Movies. We considered the problem of ranking movies in the MovieLens dataset [29] (ratings (from 1 to 5) of 1682 movies by 943 users). The goal was to predict preferences of a given user, i.e. an ordering of the 1682 movies, according to the user's partial ratings. We applied Alg. 1 to the ranking problem using the *rank loss* [7]

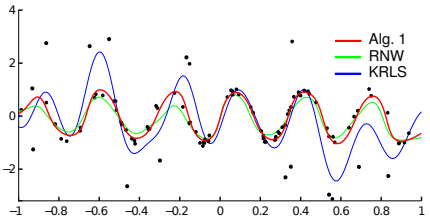
$$\triangle_{rank}(y, y') = \frac{1}{2} \sum_{i,j=1}^{M} \gamma(y')_{ij} \, (1 - \operatorname{sign}(y_i - y_j)), \tag{21}$$

where M is the number of movies, y is a re-ordering of the sequence $1, \ldots, M$. The scalar $\gamma(y)_{ij}$ denotes the costs (or reward) of having movie j ranked higher than movie i. Similarly to [7], we set $\gamma(y)_{ij}$ equal to the difference of ratings provided by user associated to y (from 1 to 5). We chose as k in Alg. 1, a linear kernel on features similar to those proposed in [7], which were computed based on users' profession, age, similarity of previous ratings, etc. Since solving Alg. 1 for $\triangle_{rank}$ is NP-hard (see [7]) we adopted the *Feedback Arc Set approximation (FAS)* proposed in [30] to approximate the $\hat{f}(x)$ of Alg. 1. Results are reported in Tab. 1 comparing Alg. 1 (Ours) with surrogate ranking methods using a Linear [7], Hinge [27] or Logistic [28] loss and Struct SVM [4]. We randomly sampled $n = 643$ users for training and tested on the remaining 300. We performed 5-fold cross-validation for model selection. We report the normalized $\triangle_{rank}$, averaged over 10 trials to account for statistical variability. Interestingly, our approach appears to outperform all competitors, suggesting that Alg. 1 is a viable approach to ranking.

Image Reconstruction with Hellinger Distance. We considered the USPS digits reconstruction experiment originally proposed in [18]. The goal is to predict the lower half of an image depicting a digit, given the upper half of the same image in input. The standard approach is to use a Gaussian kernel k_G on images in input and adopt KDE methods such as [18, 12, 14] with loss $\triangle_G(y, y') = 1 - k_G(y, y')$. Here we take a different approach and, following [31], we interpret an image depicting a digit as an histogram and normalize it to sum up to 1. Therefore, $\mathcal{Y}$ is the unit simplex in $\mathbb{R}^{128}$ (16×16 images) and we adopt the Hellinger distance $\triangle_H$

$$\triangle_H(y, y') = \sum_{i=1}^{M} |(y_i)^{1/2} - (y'_i)^{1/2}| \qquad \text{for} \quad y = (y_i)_{i=1}^{M} \tag{22}$$

to measure distances on $\mathcal{Y}$. We used the kernel k_G on the input space and compared Alg. 1 using respectively $\triangle_H$ and $\triangle_G$. For $\triangle_G$ Alg. 1 correpsponds to [12]. We performed digit reconstruction experiments by training on 1000 examples evenly distributed among the 10 digits of USPS and tested on 5000 images. We performed 5-fold cross-validation for model selection. Tab. 2 reports the performance of Alg. 1 and the KDE methods averaged over 10 runs. Performance are reported according to the Gaussian loss $\triangle_G$ and Hellinger loss $\triangle_H$. Unsurprisingly, methods trained with respect to a specific loss perform better than the competitor with respect to such loss. Therefore, as a further measure of performance we also introduced the "Recognition" loss $\triangle_R$. This loss has to be intended as a measure of how "well" a predictor was able to correctly reconstruct an image for digit recognition purposes. To this end, we trained an automatic digit classifier and defined $\triangle_R$ to be the misclassification error of such classifier when tested on images reconstructed by the two prediction algorithms. This automatic classifier was trained using a standard SVM [24] on a separate subset of USPS images and achieved an average 0.04% error rate on the true 5000 test sets. In this case a clear difference in performance can be observed between using two different loss functions, suggesting that $\triangle_H$ is more suited for the reconstruction problem.



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.

Robust Estimation. We considered a regression problem with many outliers and evaluated Alg. 1 using the Cauchy loss (see Example 1 - (3)) for robust estimation. Indeed, in this setting, $\mathcal{Y} = [-M, M] \subset \mathbb{R}$ is not structured, but the non-convexity of $\triangle$ can be an obstacle to the learning process. We generated a dataset according to the model $y = \sin(6\pi x) + \epsilon + \zeta$, where x was sampled uniformly on $[-1, 1]$ and ϵ according to a zero-mean Gaussian with variance 0.1. ζ modeled the outliers and was sampled according to a zero-mean random variable that was 0 with probability 0.90 and a value uniformly at random in $[-3, 3]$ with probability 0.1. We compared Alg. 1 with the *Nadaraya-Watson* robust estimator (RNW) [32] and kernel ridge regression (KRR) with a Gaussian kernel as baseline. To train Alg. 1 we used a Gaussian kernel on the input and performed predictions (i.e. solved Eq. (16)) using Matlab FMINUNC function for unconstrained minimization. Experiments were performed with training sets of increasing dimension (100 repetitions each) and test set of 1000 examples. 5-fold cross-validation for model selection. Results are reported in Fig. 1, showing that our estimator significantly outperforms the others. Moreover, our method appears to greatly benefit from training sets of increasing size.

7 Conclusions and Future Work

In this work we considered the problem of structured prediction from a Statistical Learning Theory perspective. We proposed a learning algorithm for structured prediction that is split into a learning and prediction step similarly to previous methods in the literature. We studied the statistical properties of the proposed algorithm by adopting a strategy inspired to surrogate methods. In particular, we identified a large family of loss functions for which it is natural to identify a corresponding surrogate problem. This perspective allows to prove a derivation of the algorithm proposed in this work. Moreover, by exploiting a comparison inequality relating the original and surrogate problems we were able to prove universal consistency and generalization bounds under mild assumption. In particular, the bounds proved in this work recover those already known for least squares classification, of which our approach can be seen as a generalization. We supported our theoretical analysis with experiments showing promising results on a variety of structured prediction problems.

A few questions were left opened. First, we ask whether the comparison inequality can be improved (under suitable hypotheses) to obtain faster generalization bounds for our algorithm. Second, the surrogate problem in our work consists of a vector-valued regression (in a possibly infinite dimensional Hilbert space), we solved this problem by plain kernel ridge regression but it is natural to ask whether approaches from the multi-task learning literature could lead to substantial improvements in this setting. Finally, an interesting question is whether alternative surrogate frameworks could be derived for the setting considered in this work, possibly leading to tighter comparison inequalities. We will investigate these questions in the future.

References

[1] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI, IEEE Transactions on*, 32(9):1627–1645, 2010.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on CVPR*, pages 3128–3137, 2015.

[3] Thomas Hofmann Bernhard Schölkopf Alexander J. Smola Ben Taskar Bakir, Gökhan and S.V.N Vishwanathan. *Predicting structured data*. MIT press, 2007.

[4] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *JMLR*, pages 1453–1484, 2005.

[5] Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jacques Slotine. Multiclass learning with simplex coding. In *NIPS*, pages 2798–2806, 2012.

[6] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. *Artificial Intelligence*, 2013.

[7] John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 327–334, 2010.

[8] Ingo Steinwart, Andreas Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[9] Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2004.

[10] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[11] M. Álvarez, N. Lawrence, and L. Rosasco. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. see also http://arxiv.org/abs/1106.6251.

[12] Corinna Cortes, Mehryar Mohri, and Jason Weston. A general regression technique for learning transductions. In *Proceedings of the 22nd international conference on Machine learning*, 2005.

[13] P. Geurts, L. Wehenkel, and F. d'Alché Buc. Kernelizing the output of tree-based methods. In *ICML*, 2006.

[14] H. Kadri, M. Ghavamzadeh, and P. Preux. A generalized kernel approach to structured output learning. *Proc. International Conference on Machine Learning (ICML)*, 2013.

[15] S. Giguère, M. M., K. Sylla, and F. Laviolette. Risk bounds and learning algorithms for the regression approach to structured output prediction. In *ICML*. JMLR Workshop and Conference Proceedings, 2013.

[16] C. Brouard, M. Szafranski, and F. d'Alché Buc. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *JMLR*, 17(176):1–48, 2016.

[17] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

[18] Jason Weston, Olivier Chapelle, Vladimir Vapnik, André Elisseeff, and Bernhard Schölkopf. Kernel dependency estimation. In *Advances in neural information processing systems*, pages 873–880, 2002.

[19] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[20] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[21] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

[22] Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. Springer, 2011.

[23] Richard Schoen and Shing-Tung Yau. *Lectures on differential geometry*, volume 2. International press Boston, 1994.

[24] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[25] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[26] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[27] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

[28] Ofer Dekel, Yoram Singer, and Christopher D Manning. Log-linear models for label ranking. In *Advances in neural information processing systems*, page None, 2004.

[29] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2015.

[30] Peter Eades, Xuemin Lin, and William F Smyth. A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters*, 47(6):319–323, 1993.

[31] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[32] Wolfgang Härdle. Robust regression function estimation. *Journal of Multivariate Analysis*, 14(2):169–180, 1984.