Leveraging Low-Rank Relations Between Surrogate Tasks in Structured Prediction

Giulia Luise [1] Dimitris Stamos [1] Massimiliano Pontil [1 2] Carlo Ciliberto [1 3]

Abstract

We study the interplay between surrogate methods for structured prediction and techniques from multitask learning designed to leverage relationships between surrogate outputs. We propose an efficient algorithm based on trace norm regularization which, differently from previous methods, does not require explicit knowledge of the coding/decoding functions of the surrogate framework. As a result, our algorithm can be applied to the broad class of problems in which the surrogate space is large or even infinite dimensional. We study excess risk bounds for trace norm regularized structured prediction proving the consistency and learning rates for our estimator. We also identify relevant regimes in which our approach can enjoy better generalization performance than previous methods. Numerical experiments on ranking problems indicate that enforcing low-rank relations among surrogate outputs may indeed provide a significant advantage in practice.

1. Introduction

The problem of structured prediction is receiving increasing attention in machine learning, due to its wide practical importance (Bakir et al., 2007; Nowozin et al., 2011) and the theoretical challenges in designing principled learning procedures (Taskar et al., 2004; 2005; London et al., 2016; Cortes et al., 2016). A key aspect of this problem is the non-vectorial nature of the output space, e.g. graphs, permutations, and manifolds. Consequently, traditional regression and classification algorithms are not well-suited to these settings and more sophisticated methods need to be developed. Among the most well-established strategies for structured prediction are the so-called *surrogate methods* (Bartlett et al., 2006). Within this framework, a coding function is designed to embed the structured output into a linear space, where the resulting problem is solved via standard supervised learning methods. Then, the solution of the surrogate problem is pulled back to the original output space by means of a decoding procedure, which allows one to recover the structured prediction estimator under suitable assumptions. In most cases, the surrogate learning problem amounts to a vector-valued regression in a possibly infinite dimensional space. The prototypical choice for such surrogate estimator is a regularized least squares in a reproducing kernel Hilbert space, as originally considered in (Weston et al., 2003; Cortes et al., 2005; Bartlett et al., 2006) and then explored in (Mroueh et al., 2012; Kadri et al., 2013; Brouard et al., 2016; Ciliberto et al., 2016; Osokin et al., 2017; Rudi et al., 2018; Luise et al., 2018).

The principal goal of this paper is to extend the surrogate approaches to methods that encourage structure among the outputs. Indeed, a large body of work from traditional multitask learning has shown that leveraging the relations among multiple outputs may often lead to better estimators (Alvarez et al., 2012; Argyriou et al., 2008; Caponnetto & De Vito, 2007; Maurer, 2006; Micchelli et al., 2013). However, previous methods that propose to apply multitask strategies to surrogate frameworks (see e.g. Alvarez et al., 2012; Fergus et al., 2010) heavily rely on the explicit knowledge of the encoding function of a surrogate framework. As a consequence they are not applicable when the surrogate space is large or even infinite dimensional.

Contributions. We propose a new algorithm based on low-rank regularization for structured prediction that builds upon the surrogate framework in (Ciliberto et al., 2016; 2017). Differently from previous methods in the literature on surrogate methods, our approach does not require explicit knowledge of the encoding function, by leveraging intrinsic properties of the loss function. In particular, exploiting approaches based on the variational formulation of trace norm regularization (Srebro et al., 2005), we are able to derive an efficient learning algorithm also in the case of infinite dimensional surrogate spaces.

[1] Department of Computer Science, University College London, London, UK [2] Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa [3] Department of Electrical and Electronic Engineering,Imperial College London, London, UK. Correspondence to: Giulia Luise <g.luise.16@ucl.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

We characterize the generalization properties of the proposed estimator by proving excess risk bounds for the corresponding least-squares surrogate estimator that extend previous results (Bach, 2008). In particular, in line with previous work on the topic (Maurer & Pontil, 2013), we identify settings in which the trace norm regularizer can provide significant advantages over standard ℓ_2 regularization. While similar findings have been obtained in the case of a Lipschitz loss, to our knowledge this is a novel result for least-squares regression with trace norm regularization. In this sense, the implications of our analysis extend beyond structured prediction and apply to settings such as collaborative filtering with side information (Abernethy et al., 2009). We evaluate our approach on a number of learning-to-rank problems. In our experiments the proposed method significantly outperforms all competitors, suggesting that enforcing low-rank regularization on the surrogate outputs can be beneficial also in structured prediction settings.

Paper Organization. Sec. 2 reviews surrogate methods and the specific framework adopted in this work. Sec. 3 introduces the proposed approach to trace norm regularization and proves that it does not leverage explicit knowledge of coding and surrogate space. Sec. 4 describes the statistical analysis of the proposed estimator both in a vector-valued and multi-task learning setting. Sec. 5 reports on experiments. Sec. 6 discusses relevant directions for future work.

2. Background

Our proposed estimator belongs to the family of surrogate methods (Bartlett et al., 2006). This section reviews the main ideas behind these approaches.

Surrogate Methods. Surrogate methods are general strategies to address supervised learning problems. Their goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ minimizing the *expected risk* of a distribution ρ on $\mathcal{X} \times \mathcal{Y}$

$$\mathcal{E}(f) := \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y)\, d\rho(x, y), \tag{1}$$

given only n observations $(x_i, y_i)_{i=1}^n$ independently drawn from ρ, which is unknown in practice. Here $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss measuring prediction errors.

Surrogate methods have been conceived to deal with so-called *structured prediction* settings, namely supervised problems where $\mathcal{Y}$ is not a vector space but rather a "structured" set (of e.g. strings, graphs, permutations, points on a manifold, etc.). Surrogate methods have been successfully applied to problems such as classification (Bartlett et al., 2006), multi-labeling (Gao & Zhou, 2013; Mroueh et al., 2012) or ranking (Duchi et al., 2010). They follow an alternative route to standard empirical risk minimization (ERM), which instead consists in directly finding the model that best explains training data within a prescribed hypotheses space.

Surrogate methods are characterized by three phases:

1. **Coding.** Define an embedding $\mathsf{c} : \mathcal{Y} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Map $(x_i, y_i)_{i=1}^n$ to a "surrogate" dataset $(x_i, \mathsf{c}(y_i))_{i=1}^n$.
2. **Learning.** Define a surrogate loss $\mathsf{L} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Learn $\hat{g} : \mathcal{X} \to \mathcal{H}$ via ERM on $(x_i, \mathsf{c}(y_i))_{i=1}^n$.
3. **Decoding.** Define a decoding $\mathsf{d} : \mathcal{H} \to \mathcal{Y}$ and return the estimator $\hat{f} = d \circ \hat{g} : \mathcal{X} \to \mathcal{Y}$.

Example 1 (One Vs All). *$\mathcal{Y} = \{1, \dots, T\}$ set of T classes and ℓ the 0-1 loss: 1) The coding is $\mathsf{c} : \mathcal{Y} \to \mathcal{H} = \mathbb{R}^T$ with $\mathsf{c}(i) = e_i$ the vector of all 0s but 1 at the i-th entry. 2) $\hat{g} : \mathcal{X} \to \mathbb{R}^T$ is learned by minimizing a surrogate loss $\mathsf{L} : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ (e.g. least-squares). 3) The classifier is $\hat{f}(x) = \mathsf{d}(\hat{g}(x))$, with decoding $\mathsf{d}(v) = \operatorname{argmax}_{i=1}^T \{v_i\}$.*

A key element of surrogate methods is the choice of the loss L. Indeed, since $\mathcal{H}$ is linear (e.g. $\mathcal{H} = \mathbb{R}^T$ in Ex. 1), if L is convex it is possible to learn $\hat{g}$ efficiently by means of standard ERM. However, this opens the question of characterizing how the surrogate risk

$$\mathcal{R}(g) = \int \mathsf{L}(g(x), \mathsf{c}(y))\, d\rho(x, y) \tag{2}$$

is related to the original risk $\mathcal{E}(f)$. In particular let $f_* : \mathcal{X} \to \mathcal{Y}$ and $g_* : \mathcal{X} \to \mathcal{H}$ denote the minimizers of respectively $\mathcal{E}(f)$ and $\mathcal{R}(g)$. We require:

- **Fisher Consistency.** $\mathcal{E}(\mathsf{d} \circ g_*) = \mathcal{E}(f_*)$.
- **Comparison Inequality.** For any $g : \mathcal{X} \to \mathcal{H}$,

$$\mathcal{E}(\mathsf{d} \circ g) - \mathcal{E}(f_*) \leq \sigma(\mathcal{R}(g) - \mathcal{R}(g_*)), \tag{3}$$

with $\sigma : \mathbb{R} \to \mathbb{R}_+$ a continuous nondecreasing function, such that $\sigma(0) = 0$.

Fisher consistency guarantees the coding/decoding framework to be coherent with the original problem. The comparison inequality suggests to focus the theoretical analysis on $\hat{g}$, since learning rates for $\hat{g}$ directly lead to learning rates for $\hat{f} = \mathsf{d} \circ \hat{g}$.

SELF Framework. A limiting aspect of surrogate methods is that they are often tailored around individual problems. An exception is the framework in (Ciliberto et al., 2016), which provides a general strategy to identify coding, decoding and surrogate space for a variety of learning problems. The key condition in this settings is for the loss ℓ to be SELF:

Definition 1 (SELF). *A function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a* Structure Encoding Loss Function (SELF) *if there exist a separable Hilbert space $\mathcal{H}_y$, a continuous map $\psi : \mathcal{Y} \to \mathcal{H}_y$ and $V : \mathcal{H}_y \to \mathcal{H}_y$ a bounded linear operator, such that for all $y, y' \in \mathcal{Y}$*

$$\ell(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_y}. \tag{4}$$

The condition above is very general (albeit not always trivial to verify in practice): as shown in (Ciliberto et al., 2016; 2017), most loss functions used in machine learning are SELF (e.g. regression, robust estimation, classification, ranking, etc.).

We can design surrogate frameworks "around" a SELF ℓ, by choosing (*Coding*) the map $\mathsf{c} = \psi : \mathcal{Y} \to \mathcal{H}_y$, the least-squares (*Surrogate loss*) $\mathsf{L}(h, h') = \|h - h'\|_{\mathcal{H}_y}^2$ and (*Decoding*) $\mathsf{d} : \mathcal{H}_y \to \mathcal{Y}$ defined for any $h \in \mathcal{H}_y$ as

$$\mathsf{d}(h) = \operatorname{argmin}_{y \in \mathcal{Y}} \ \langle \psi(y), V h \rangle_{\mathcal{H}_y} \tag{5}$$

for any $h \in \mathcal{H}_y$. The resulting is a sound surrogate framework as summarized by the theorem below.

Theorem 1 (Thm. 2 in (Ciliberto et al., 2016))**.** *Let ℓ be SELF and $\mathcal{Y}$ a compact set. Then, the SELF framework introduced above is Fisher consistent. Moreover, it satisfies a comparison inequality with $\sigma(\cdot) = \mathsf{q}_\ell \sqrt{\cdot}$, where $\mathsf{q}_\ell = \|V\| \sup_{y \in \mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_y}$.*

Loss trick. A key aspect of the SELF framework is that, in practice, the resulting algorithm *does not require explicit knowledge of the coding/decoding and surrogate space* (only needed for the theoretical analysis). To see this, let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_y = \mathbb{R}^T$ and consider the parametrization $g(x) = Gx$ of functions $g : \mathcal{X} \to \mathcal{H}_y$, with $G \in \mathbb{R}^{T \times d}$ a matrix. We can perform Tikhonov regularization to learn the matrix $\hat{G}$ minimizing the (surrogate) empirical risk

$$\min_{g : \mathcal{X} \to \mathcal{H}} \ n^{-1} \sum_{i=1}^n \|g(x_i) - \psi(y_i)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathsf{HS}}^2, \tag{6}$$

where $\|\cdot\|_{\mathsf{HS}}$ is the Hilbert-Schmidt (HS) (or Frobenius) norm regularizer and $\lambda > 0$. Note that $\hat{G}$ can be obtained in closed form and $\hat{g} : \mathcal{X} \to \mathcal{H}_y$ is such that

$$\hat{g}(x) = \hat{G}x = \sum_{i=1}^n \alpha_i(x) \psi(y_i), \quad \text{with} \tag{7}$$

$$\alpha(x) = (\alpha_1(x), \dots, \alpha_n(x))^\top = (K_x + n\lambda I)^{-1} v_x, \tag{8}$$

for every $x \in \mathcal{X}$. Here $K_x \in \mathbb{R}^{n \times n}$ is the empirical kernel matrix of the linear kernel $k(x, x') = x^\top x'$ and $v_x \in \mathbb{R}^n$ is the vector with i-th entry $(v_x)_i = k(x, x_i)$.

Applying the SELF decoding in Eq. (5) to $\hat{g}$, we have

$$\hat{f}(x) = \mathsf{d}(\hat{g}(x)) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^n \alpha_i(x) \ell(y, y_i), \tag{9}$$

for all $x \in \mathcal{X}$. This follows by combining the SELF property $\ell(y, y_i) = \langle \psi(y), V \ \psi(y_i) \rangle_{\mathcal{H}_y}$ with $\hat{g}$ in Eq. (7) and the linearity of the inner product. Eq. (9) was originally dubbed "loss trick" since it avoids explicit knowledge of the coding ψ, similarly to the feature map for the kernel trick (Schölkopf et al., 2002).

The characterization of $\hat{f}$ in terms of an optimization problem over $\mathcal{Y}$ (like in Eq. (9)) is a common practice to most structured prediction algorithms. In the literature, such decoding process is referred to as the inference (Nowozin et al., 2011) or pre-image (Brouard et al., 2016; Cortes et al., 2005; Weston et al., 2003) problem. We refer to (Honeine & Richard, 2011; Bakir et al., 2007; Nowozin et al., 2011) for examples on how these problems are addressed in practice.

General Setting. The derivation above holds also when $\mathcal{H}_y$ is infinite dimensional and when using a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$. Let $\mathcal{H}_x$ be the reproducing kernel Hilbert space (RKHS) induced by k and $\phi : \mathcal{X} \to \mathcal{H}_x$ a corresponding feature map (Aronszajn, 1950). We can parametrize $g : \mathcal{X} \to \mathcal{H}_y$ as $g(\cdot) = G\phi(\cdot)$, with $G \in \mathcal{H}_y \otimes \mathcal{H}_x$ the space of Hilbert-Schmidt operators from $\mathcal{H}_x$ to $\mathcal{H}_y$ (the natural generalization of $\mathbb{R}^{d \times T} = \mathbb{R}^d \otimes \mathbb{R}^T$ to the infinite setting). The problem in Eq. (6) can still be solved in closed form analogously to Eq. (7), with now K_x the empirical kernel matrix of k (Caponnetto & De Vito, 2007). This leads to the decoding for $\hat{f}$ as in Eq. (9).

3. Low-Rank SELF Learning

Building upon the SELF framework, we discuss the use of multitask regularizers to exploit potential relations among the surrogate outputs. Our analysis is motivated by observing that Eq. (6) is equivalent to learning multiple (possibly infinitely many) functions

$$\min_{\{g_t\} \in \mathcal{H}_x} \ \frac{1}{n} \sum_{t \in \mathcal{T}} \sum_{i=1}^n (g_t(x_i) - \varphi_t(y_i))^2 + \lambda \|g_t\|_{\mathcal{H}_x}^2, \tag{10}$$

where, given a basis $\{e_t\}_{t \in \mathcal{T}}$ of $\mathcal{H}_y$ with $\mathcal{T} \subseteq \mathbb{N}$, we have denoted $\psi_t(y) = \langle e_t, \psi(y) \rangle_{\mathcal{H}_y}$ for any $y \in \mathcal{Y}$ and $t \in \mathcal{T}$. Indeed, from the literature on vector-valued learning in RKHS (see e.g Micchelli & Pontil, 2005), we have that for $g : \mathcal{X} \to \mathcal{H}_y$ parametrized by an operator $G \in \mathcal{H}_y \otimes \mathcal{H}_x$, any $g_t : \mathcal{X} \to \mathbb{R}$ defined by $g_t(\cdot) = \langle e_t, g(\cdot) \rangle_{\mathcal{H}_y}$, is a function in the RKHS $\mathcal{H}_x$ and, moreover, $\|G\|_{\mathsf{HS}}^2 = \sum_{t \in \mathcal{T}} \|g_t\|_{\mathcal{H}_x}^2$.

The observation above implies that we are learning the surrogate "components" g_t as separate problems or *tasks*, an approach often referred to as "independent task learning" within the multitask learning (MTL) literature (see e.g. Argyriou et al., 2008). In this respect, a more appropriate strategy would be to leverage the potential relations between such components during learning. In particular, we consider the problem

$$\min_{G \in \mathcal{H}_y \otimes \mathcal{H}_x} \ \frac{1}{n} \sum_{i=1}^n \|G\phi(x_i) - \psi(y_i)\|_{\mathcal{H}_y}^2 + \lambda \|G\|_*, \tag{11}$$

where $\|G\|_*$ denotes the trace norm, namely the sum of the singular values of G. Similarly to the ℓ_1-norm on vectors, the trace norm favours sparse (and thus low-rank) solutions. Intuitively, encouraging G to be low-rank reduces the degrees of freedom allowed to the individual tasks g_t. This approach was extensively investigated and successfully applied to several MTL settings, (see e.g. Argyriou et al., 2008; Bach, 2008; Abernethy et al., 2009; Maurer & Pontil, 2013).

In general, the idea of combining MTL methods with surrogate frameworks has already been studied in settings such as classification or multi-labeling (see e.g. Alvarez et al., 2012; Fergus et al., 2010). However, these approaches require to explicitly use the coding/decoding and surrogate space within the learning algorithm. This is clearly unfeasible when $\mathcal{H}_y$ is large or infinite dimensional.

SELF and Trace Norm MTL. In this work we leverage the SELF property outlined in Sec. 2 to derive an algorithm that overcomes the issues above and *does not* require explicit knowledge of the coding map ψ. However, our approach still requires to access the matrix $K_y \in \mathbb{R}^{n\times n}$ of inner products $(K_y)_{ij} = \langle \psi(y_i), \psi(y_j)\rangle_{\mathcal{H}_y}$ between the training outputs. When the surrogate space $\mathcal{H}_y$ is a RKHS, K_y corresponds to an empirical *output kernel matrix*, which can be efficiently computed. This motivates us to introduce the the following assumption.

Assumption 1 (SELF & RKHS)**.** *The loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is SELF with $\mathcal{H}_y$ a RKHS on $\mathcal{Y}$ with reproducing kernel $k_y(y, y') = \langle \psi(y), \psi(y')\rangle_{\mathcal{H}_y}$ for any $y, y' \in \mathcal{Y}$.*

The assumption above imposes an additional constraint on ℓ and thus on the applicability of Alg. 1. However, it was shown in (Ciliberto et al., 2016) that this requirement is always satisfied by any loss when $\mathcal{Y}$ is a discrete set. In this case the output kernel is $k_y(y, y') = \delta_{y=y'}$ the 0-1 kernel. Moreover, it was recently shown that Asm. 1 holds for any smooth ℓ on a compact set $\mathcal{Y}$ by choosing $k_y(y, y') = \exp(-\|y - y\|/\sigma)$ the Abel kernel with hyperparameter $\sigma > 0$ (Luise et al., 2018).

Algorithm. Standard methods to solve Eq. (11), such as forward-backward splitting, require to perform the singular value decomposition of the estimator at every iteration (Mazumder et al., 2010). This is prohibitive for large scale applications and, to overcome these drawbacks, algorithms exploiting the variational form

$$\|G\|_* = \frac{1}{2} \inf \Big\{ \|A\|_{\mathsf{HS}}^2 + \|B\|_{\mathsf{HS}}^2 \ : \ G = AB^*,\ r \in \mathbb{N}, \\ A \in \mathcal{H}_y \otimes \mathbb{R}^r,\ \ B \in \mathcal{H}_x \otimes \mathbb{R}^r \Big\},$$

of the trace norm have been considered (see e.g. Srebro et al., 2005). Using this characterization, Eq. (11) is reformulated as the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \|AB^*\phi(x_i) - \psi(y_i)\|_{\mathcal{H}_y}^2 + \lambda\big(\|A\|_{\mathsf{HS}}^2 + \|B\|_{\mathsf{HS}}^2\big), \quad (12)$$

Algorithm 1 LOW-RANK SELF LEARNING

```
Input: $K_x, K_y \in \mathbb{R}^{n\times n}$ empirical kernel matrices for input and output data, $\lambda$ regularizer, $r$ rank, $\nu$ step size, $k$ number of iterations.
Initalize: Sample $M_0, N_0 \in \mathbb{R}^{n\times r}$ randomly.
For $j = 0, \dots, k$:
    $M_{j+1} = (1-\lambda\nu)M_j - \nu(K_x M_j N_j - I)K_y N_j$
    $N_{j+1} = (1-\lambda\nu)N_j - \nu(N_j M_j^\top K_x - I)K_x M_j$
Return: The weighting function $\alpha^{\mathsf{tn}} : \mathcal{X} \to \mathbb{R}^n$
        with $\alpha^{\mathsf{tn}}(x) = N_k M_k^\top v_x$ for any $x \in \mathcal{X}$
        with $v_x \in \mathbb{R}^n$ as in Eq. (8)
```

over the operators $A \in \mathcal{H}_y \otimes \mathbb{R}^r$ and $B \in \mathcal{H}_x \otimes \mathbb{R}^r$, where B^* denotes the adjoint of B and $r \in \mathbb{N}$ is now a further hyperparameter. The functional in Eq. (12) is smooth and methods such as gradient descent can be applied. Interestingly, despite the functional being non-convex, guarantees on the global convergence in these settings have been explored (Journée et al., 2010).

In the SELF setting, minimizing Eq. (12) has the additional advantage that it allows us to derive an analogous of the loss trick introduced in Eq. (9). In particular, the following result shows how each iterate of gradient descent can be efficiently "decoded" into a structured prediction estimator according to Alg. 1.

Theorem 2 (Loss Trick for Trace Norm)**.** *Under Asm. 1, let $M, N \in \mathbb{R}^{n\times r}$ and (A_k, B_k) be the k-th iterate of gradient descent on Eq. (12) from $A_0 = \sum_{i=1}^n \phi(x_i) \otimes M^i$ and $B_0 = \sum_{i=1}^n \psi(y_i) \otimes N^i$, with M^i, N^i denoting the i-th rows of M and N respectively. Let $\hat{g}_k : \mathcal{X} \to \mathcal{H}_y$ be such that $\hat{g}_k(\cdot) = A_k B_k^* \phi(\cdot)$. Then, the structured prediction estimator $\hat{f}_k = \mathsf{d} \circ \hat{g}_k : \mathcal{X} \to \mathcal{Y}$ with decoding d in Eq. (5) is such that*

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

for any $x \in \mathcal{X}$, with $\alpha^{\mathsf{tn}}(x) \in \mathbb{R}^n$ the output of Alg. 1 after k iterations starting from $(M_0, N_0) = (M, N)$.

The result above shows that Alg. 1 offers a concrete algorithm to perform the SELF decoding $\hat{f}_k = \mathsf{d} \circ \hat{g}_k$ of the surrogate function $\hat{g}_k(\cdot) = A_k B_k^* \phi(\cdot)$ obtained after k iterations of gradient descent on Eq. (12). Note that when $\mathcal{H}_y$ is infinite dimensional it would be otherwise impossible to perform gradient descent in practice. In this sense, Thm. 2 can be interpreted as a representer theorem with respect to both inputs and outputs. The details of the proof are reported in Appendix A; the key aspect is to show that every iterate (A_j, B_j) of gradient descent on Eq. (12) is of the form $A_j = \sum_{i=1}^n \phi(x_i) \otimes M^i$ and $B_j = \sum_{i=1}^n \psi(y_i) \otimes N^i$

for some matrices $M, N \in \mathbb{R}^{n \times r}$. Hence, the products $A_j^* A_j = M^\top K_x M$ and $B_j^* B_j = N^\top K_y N$ – used in the optimization – are $r \times r$ matrices that can be efficiently computed in practice, leading to Alg. 1.

We conclude this section by noting that, in contrast to trace norm regularization, not every MTL regularizer fits naturally within the SELF framework.

SELF and other MTL Regularizer. A well-established family of MTL methods consists in replacing the trace norm $\|G\|_*$ with $\operatorname{tr}(GAG^*)$ in Eq. (11), where $A \in \mathcal{H}_y \otimes \mathcal{H}_y$ is a positive definite linear operator enforcing specific relations on the tasks via a deformation of the metric of $\mathcal{H}_y$ (see Micchelli & Pontil, 2005; Jacob et al., 2008; Alvarez et al., 2012, and references therein). While in principle appealing also in surrogate settings, thes approaches present critical computational and modelling challenges for the SELF framework: the change of metric induced by A has a disruptive effect on the loss trick. As a consequence, an equivalent of Thm. 2 does not hold in general (see Appendix A.3 for a detailed discussion).

4. Theoretical Analysis

In this section we study the generalization properties of low-rank SELF learning. Our analysis is indirect since we characterize the learning rates of the *Ivanov* estimator (in contrast to *Tikhonov*, see Eq. (11))

$$\hat{G} = \operatorname*{argmin}_{\|G\|_* \leq \gamma} \frac{1}{n} \sum_{i=1}^{n} \|G\phi(x_i) - \psi(y_i)\|^2_{\mathcal{H}_y}. \tag{13}$$

Indeed, while Tikhonov regularization is typically more convenient from a computational perspective, Ivanov regularization is often more amenable to theoretical analysis since it is naturally related to standard complexity measures for hypotheses spaces such as Rademacher complexity, Covering Numbers or VC dimension (Shalev-Shwartz & Ben-David, 2014). However, the two regularization strategies are *equivalent* in the following sense: for any γ there exists $\lambda(\gamma)$ such that the minimizer of Eq. (11) (Tikhonov) is also a minimizer for Eq. (13) (Ivanov) with constraint γ (and vice-versa). This follows from a standard Lagrangian duality argument leveraging the convexity of the two problems (see e.g. Oneto et al., 2016, or Appendix E for more details). Hence, while our results in the following are reported for the Ivanov estimator from Eq. (13), they apply equivalently to Tikhonov in Eq. (11).

We now proceed to present the main result of this section, proving excess risk bounds for the trace norm surrogate estimator. In the following we assume a reproducing kernel k_x on $\mathcal{X}$ and k_y on $\mathcal{Y}$ (according to Asm. 1) and denote $\mathsf{m}_x^2 = \sup_{x \in \mathcal{X}} k_x(x,x)$ and $\mathsf{m}_y^2 = \sup_{y \in \mathcal{Y}} k_y(y,y)$. Moreover, let $C = \mathbb{E}\, \phi(x) \otimes \phi(x)$ the covariance operator over input data sampled from ρ and by $\|C\|_{\mathrm{op}}$ its operator norm, namely its largest singular value. We introduce the following condition.

Assumption 2. *There exists $G_* \in \mathcal{H}_y \otimes \mathcal{H}_x$ with finite trace norm, $\|G_*\|_* < +\infty$, such that $g_*(\cdot) = G_*\phi(\cdot)$ is a minimizer of the risk $\mathcal{R}$ in Eq. (2).*

The assumption above requires the ideal solution of the surrogate problem to belong to the space of hypotheses of the learning algorithm. This is a standard requirement in statistical learning theory in order to characterize the excess risk bounds of an estimator (see e.g. Shalev-Shwartz & Ben-David, 2014).

Theorem 3. *Under Asm. 2, let $\mathcal{Y}$ be a compact set, let $(x_i, y_i)_{i=1}^n$ be a set of n points sampled i.i.d. and let $\hat{g}(\cdot) = \hat{G}\phi(\cdot)$ with $\hat{G}$ the solution of Eq. (13) for $\gamma = \|G_*\|_*$. Then, for any $\delta > 0$*

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \;\leq\; (\mathsf{m}_y + \mathsf{M})\sqrt{\frac{4 \log \frac{\mathsf{r}}{\delta}}{n}} + O(n^{-1}), \tag{14}$$

with probability at least $1 - \delta$, where

$$\mathsf{M} = 2\mathsf{m}_x \|C\|_{\mathrm{op}}^{1/2} \|G_*\|_*^2 + \mathsf{m}_x \mathcal{R}(g_*) \|G_*\|_*, \tag{15}$$

with r a constant not depending on δ, n or G_*.

The proof is detailed in Appendix B. The two main ingredients are: $i)$ the boundedness of the trace norm of G_*, which allows us to exploit the duality between trace and operator norms; $ii)$ recent results on Bernstein's inequalities for the operator norm of random operators between separable Hilbert spaces (Minsker, 2017).

We care to point out that previous results are available in the following settings: (Bach, 2008) shows the convergence in distribution for the trace norm estimator to the minimum risk and (Koltchinskii et al., 2011) shows excess risk bounds in high probability for an estimator which leverages previous knowledge on the distribution (e.g. matrix completion problem). Both are devised for finite dimensional settings. To our knowledge this is the first work proving excess risk bounds in high probability for trace norm regularized least squares. Note that the relevance of Thm. 3 is not limited to structured prediction but it can be also applied to problems such as collaborative filtering with attributes (Abernethy et al., 2009).

Discussion. We now discuss under which conditions trace norm (TN) regularization provides an advantage over standard the Hilbert-Schmidt (HS) one. We refer to Appendix B for a more in-depth discussion on the comparison between the two estimators, while addressing here the key points.

For the HS estimator, excess risk bounds can be derived by imposing the less restrictive assumption $\|G_*\|_{\mathsf{HS}} < +\infty$.

A result analogous to Thm. 3 can be obtained (see Appendix B), with constant M such that

$$\mathsf{M} \;=\; \mathsf{m}_{\mathcal{X}}(\mathsf{m}_{\mathcal{X}}+\|C\|_{\mathrm{op}}^{\frac{1}{2}})\;\|G_*\|_{\mathsf{HS}}^2 \;+\; \mathsf{m}_{\mathcal{X}}\mathcal{R}(g_*)\|G_*\|_{\mathsf{HS}}.$$

This constant is structurally similar to the one for TN (with $\|\cdot\|_{\mathsf{HS}}$ appearing in place of $\|\cdot\|_*$), plus the additional term $\mathsf{m}_{\mathcal{X}}^2\|G\|_{\mathsf{HS}}^2$. We first note that if $\|G_*\|_{\mathsf{HS}} \ll \|G_*\|_*$, the bound offers no advantage with respect to the HS counterpart.

Hence, we focus on the setting where $\|G_*\|_{\mathsf{HS}} \sim \|G_*\|_*$ are of the same order. This corresponds to the relevant scenario where the multiple outputs/tasks encoded by G_* are (almost) linearly dependent. In this case, the constant M associated to the TN estimator can potentially be significantly smaller than the one for HS: while for TN the term $\|G_*\|_*^2$ is mitigated by $\|C\|_{\mathrm{op}}^{1/2}$, for HS the corresponding term $\|G_*\|_{\mathsf{HS}}$ is multiplied by $(\mathsf{m}_{\mathcal{X}} + \|C\|_{\mathrm{op}}^{1/2})$. Note that the operator norm is such that $\|C\|_{\mathrm{op}}^{1/2} \leq \mathsf{m}_{\mathcal{X}}$ but can potentially be *significantly* smaller than $\mathsf{m}_{\mathcal{X}}$. For instance, when $\mathcal{X} = \mathbb{R}^d$, $k_{\mathcal{X}}$ is the linear kernel and training points are sampled uniformly on the unit sphere, we have $\mathsf{m}_{\mathcal{X}} = 1$ while $\|C\|_{\mathrm{op}}^{1/2} = \frac{1}{\sqrt{d}}$.

In summary, trace norm regularization allows to leverage structural properties of the data distribution *provided that the output tasks are related*. This effect can be interpreted as the process of "sharing" information among the otherwise independent learning problems. A similar result to Thm. 3 was proved in (Maurer & Pontil, 2013) for Lipschitz loss functions (and $\mathcal{H}_{\mathcal{Y}}$ finite dimensional). We refer to such work for a more in-depth discussion on the implications of the link between trace norm regularization and operator norm of the covariance operator.

Excess Risk Bounds for $\hat{f}$. By combining Thm. 3 with the comparison inequality for the SELF framework (see Thm. 1) we can immediately derive excess risk bounds for $\hat{f} = \mathsf{d} \circ \hat{g}$.

Corollary 4. *Under the same assumptions and notation of Thm. 3, let ℓ be a SELF loss and $\hat{f} = \mathsf{d} \circ \hat{g} : \mathcal{X} \to \mathcal{Y}$. Then, for every $\delta > 0$, with probability not less than $1 - \delta$ it holds that*

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f_*) \;\leq\; \mathsf{q}_\ell \sqrt[4]{\frac{4(\mathsf{m}_{\mathcal{Y}} + \mathsf{M})^2 \log \frac{\mathsf{r}}{\delta}}{n}} + O(n^{-\frac{1}{2}})$$

where M and r are the same constants of Thm. 3 and q_ℓ is as in Thm. 1.

The result above provides comparable learning rates to those of the original SELF estimator (Ciliberto et al., 2016). However, since the constant M corresponds to the one from Thm. 3, whenever trace norm regularization provides an advantage with respect to standard Hilbert-Schmidt regularization on the surrogate problem, such improvement is directly inherited by $\hat{f}$.

We conclude this section by noting that a large body of previous work has been focused on providing *margin-based* bounds for structured predictions (see e.g. Taskar et al., 2004; Cortes et al., 2016). Even if a comparison between the margin-based paradigm and surrogate approaches is non trivial, it is interesting to note that:

1. The *surrogate* bounds in Thm. 3 are comparable to the generalization *margin* bounds (see e.g. (Taskar et al., 2004)) both in terms of rates (i.e. $O(n^{-1/2})$) and some of the key quantities (e.g. $\|w\|^2$ in (Taskar et al., 2004) is related to the $\|g_*\|_{\mathsf{HS}}$ in our setting).

2. While in surrogate settings the comparison inequality in Eq. (3) offers a direct connection with the original structured problem, for margin-based methods the link is less direct: similarly to the binary classification setting, the gap between the margin loss and the original loss could be bridged by a *noise condition* (i.e. when there is low noise at the decision boundary between labels, prediction becomes easier). When explicitly available, this condition might possibly lead to fast rates for the structured problem. Interestingly, it was recently shown in (Thm 3.5 in Nowak et al., 2019) that an improved comparison inequality holds true in low-noise finite settings also for the SELF framework. This might offer a further connection between surrogate and margin-based methods for structured prediction.

4.1. Multitask Learning

So far we have studied trace norm regularization when learning the multiple g_t in Eq. (10) within a vector-valued setting, namely where for any input sample x_i in training we observe *all* the corresponding outputs $\psi_t(y_i)$. This choice was made mostly for notational purposes and the analysis can be extended to the more general setting of nonlinear multitask learning, where separate groups of surrogate outputs could be provided each with its own dataset. We provide here a brief summary of this setting and our results within it, while postponing all details to Appendix C.

Let T be a positive integer. In typical multitask learning (MTL) settings the goal is to learn multiple functions $f_1, \dots, f_T : \mathcal{X} \to \mathcal{Y}$ jointly. While most previous MTL methods considered how to enforce linear relations among tasks, (Ciliberto et al., 2017) proposed a generalization of SELF framework to address nonlinear multitask problems (NL-MTL). In this setting, relations are enforced by means of a constraint set $\mathcal{C} \subset \mathcal{Y}^T$ (e.g. a set of nonlinear constraints that $f_1, \dots, f_T$ need to satisfy simultaneosly). The goal is to minimize the *multi-task excess risk*

$$\min_{f:\mathcal{X}\to\mathcal{C}} \mathcal{E}_T(f), \;\; \mathcal{E}_T(f) = \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{X}\times\mathbb{R}} \ell(f_t(x), y)d\rho_t(x, y),$$

where the ρ_t are unknown probability distributions on $\mathcal{X}\times\mathcal{Y}$, observed via finite samples $(x_{it}, y_{it})_{i=1}^{n_t}$, for $t = 1, \dots, T$. The NL-MTL framework interprets the nonlinear multitask problem as a structured prediction problem where the constraint set $\mathcal{C}$ represents the "structured" output. Assuming ℓ to be SELF with space $\mathcal{H}_y$ and coding ψ, the estimator $\hat{f} : \mathcal{X} \to \mathcal{C}$ then is obtained via the MTL decoding map d_T

$$\hat{f}(x) = \mathsf{d}_T(\hat{g}(x)) := \operatorname*{argmin}_{c\in\mathcal{C}} \sum_{t=1}^{T} \langle \psi(c_t), V\hat{g}_t(x)\rangle, \quad (16)$$

where each $\hat{g}_t(\cdot) = G_t\phi(\cdot) : \mathcal{X} \to \mathcal{H}_y$ is learned independently via surrogate ridge regression like in Eq. (6).

Similarly to the vector-valued case of Eq. (10), we can "aggregate" the operators $G_t \in \mathcal{H}_x \otimes \mathcal{H}_y$ in a single operator G, which is then learned by trace norm regularization as in Eq. (11) (see Appendix C for the definition of G). Then, an analogous of Thm. 2 holds for the corresponding variational formulation of such problem, which guarantees the loss trick to hold (see Appendix A.2 for the corresponding version of Alg. 1).

Also in this setting we study the theoretical properties of the low-rank approach via the estimator obtained via surrogate Ivanov regularization

$$\hat{G} = \operatorname*{argmin}_{\|G\|_*\le\gamma} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{n_t}\sum_{i=1}^{n_t} \|G_t\phi(x_{it}) - \psi(y_{it})\|^2_{\mathcal{H}_y}. \quad (17)$$

We report the result characterizing the excess risk bounds for $\hat{G}$ (see Thm. 7 for the formal version). Note that in this setting the surrogate risk $\mathcal{R}_T$ of G corresponds to the average least-squares surrogate risks of the individual G_t. In the following we denote by $\bar{C} = \frac{1}{T}\sum_{t=1}^{T} C_t$ the average of the input covariance operators $C_t = \mathbb{E}_{x\sim\rho_t}\phi(x)\otimes\phi(x)$ according to ρ_t.

Theorem 5 (Informal)**.** *Under Asm. 2, let $\{x_{it}, y_{it}\}_{t=1}^{n}$ be independently sampled from ρ_t for $t = 1, \dots, T$. Let $\hat{g}(\cdot) = \hat{G}\phi(\cdot)$ with $\hat{G}$ minimizer of Eq. (17). Then $\forall\delta > 0$, with probability at least $1 - \delta$,*

$$\mathcal{R}_T(\hat{g}) - \mathcal{R}_T(g_*) \le \sqrt{\frac{2\mathsf{M}'\log\frac{T\mathsf{r}'}{\delta}}{Tn}} + O((nT)^{-1}),$$

where the constant M' depends on $\|G_*\|_*$, $\|\bar{C}\|^{1/2}_{\rm op}$, $\mathcal{R}_T(g_*)$ and r' a constant independent of δ, n, T, G_*.

Here the constant M' exhibits an analogous behavior to M for Thm. 3 and can lead to significant benefits in the same regimes discussed for the vector-valued setting. Moreover, also in the NL-MTL setting we can leverage a comparison inequality similar to Thm. 1, with constant $\mathsf{q}_{\mathcal{C},\ell,T}$ from (Thm. 5 Ciliberto et al., 2017). As a consequence, we obtain the excess risk bound for our MTL estimator $\hat{f} = \mathsf{d}_T \circ \hat{g}$ of the form

$$\mathcal{E}_T(\hat{f}) - \mathcal{E}_T(f_*) \le \mathsf{q}_{\mathcal{C},\ell,T} \sqrt[4]{\frac{\mathsf{M}'\log\frac{T\mathsf{r}'}{\delta}}{nT}} + O(n^{-\frac{1}{2}}).$$

The constant $\mathsf{q}_{\mathcal{C},\ell,T}$, encodes key structural properties of the constraint set $\mathcal{C}$ and it was observed to potentially provide significant benefits over *linear* MTL methods (see Ex. 1 in the original NL-MTL paper). Since $\mathsf{q}_{\mathcal{C},\ell,T}$ is appearing as a multiplicative factor with respect to M', we could expect our low-rank estimator to provide even further benefits over standard NL-MTL by combining the advantages provided by the *nonlinear relations* between tasks and the *low-rank relations* among the surrogate outputs.

5. Experiments

We evaluated the empirical performance of the proposed method on ranking applications[1], specifically the *pairwise* ranking setting considered in (Duchi et al., 2010; Fürnkranz & Hüllermeier, 2003).

Denote by $\mathcal{D} = \{d_1, \dots, d_N\}$ the full set of *documents* (e.g. movies) that will be ordered according to relevance or preference (i.e. *ranked*). Let $\mathcal{X}$ be the space of *queries* (e.g. users) and assume that for each query $x \in \mathcal{X}$, a subset of the set of the associated *ratings* $\mathbf{y} = \{y_1, \dots, y_N\}$ is given, representing how relevant each document with respect to the query x. Here we assume each label y_i in $\{0, \dots, K\}$. The relation $y_i > y_j$ means that d_i is more relevant than d_j to x and should be assigned a higher rank.

We are interested in learning $f : \mathcal{X} \to \{1, \dots, N\}^N$, which assigns to a given query x a rank (or ordering) of the N object in $\mathcal{D}$. We measure errors according to the (weighted) *pairwise loss*

$$\ell(f(x), \mathbf{y}) = \sum_{j<i=1}^{N} (y_i - y_j)\,\mathrm{sign}(f_j(x) - f_i(x)), \quad (18)$$

with $f_i(x)$ denoting the predicted rank for d_i. Following (Ciliberto et al., 2017), learning to rank with a pairwise loss can be naturally formulated as a nonlinear multitask problem and tackled via structured prediction. In particular we can model the relation between each pair of documents (d_i, d_j) as a function (task) that can take values 1 or -1 depending on whether d_i is more relevant than d_j or vice-versa. Nonlinear constraints in the form of a constraint set $\mathcal{C}$ need to be added to this setting in order to guarantee coherent predictions. This leads to a decoding procedure for Eq. (16) that amounts to solve a minimal feedback arc set problem on graphs (Slater, 1961).

[1] Code at https://github.com/dstamos/LR-SELF

	ml100k	**jester1**	**jester2**	**jester3**	**sushi**
MART	0.499 (±0.050)	0.441 (±0.002)	0.442 (±0.003)	0.443 (±0.020)	0.477 (±0.100)
RankNet	0.525 (±0.007)	0.535 (±0.004)	0.531 (±0.008)	0.511 (±0.017)	0.588 (±0.005)
RankBoost	0.576 (±0.043)	0.531 (±0.002)	0.485 (±0.061)	0.496 (±0.010)	0.589 (±0.010)
AdaRank	0.509 (±0.007)	0.534 (±0.009)	0.526 (±0.001)	0.528 (±0.015)	0.588 (±0.051)
Coordinate Ascent	0.477 (±0.108)	0.492 (±0.004)	0.502 (±0.011)	0.503 (±0.023)	0.473 (±0.103)
LambdaMART	0.564 (±0.045)	0.535 (±0.005)	0.520 (±0.013)	0.587 (±0.001)	0.571 (±0.076)
ListNet	0.532 (±0.030)	0.441 (±0.002)	0.442 (±0.003)	0.456 (±0.059)	0.588 (±0.005)
Random Forests	0.526 (±0.022)	0.548 (±0.001)	0.549 (±0.001)	0.581 (±0.002)	0.566 (±0.010)
SVMrank	0.513 (±0.009)	0.507 (±0.007)	0.506 (±0.001)	0.514 (±0.009)	0.541 (±0.005)
SELF + $\|\cdot\|_{\mathsf{HS}}$	0.312 (±0.005)	0.386 (±0.005)	0.366 (±0.002)	0.375 (±0.005)	0.391 (±0.003)
(Ours) SELF + $\|\cdot\|_*$	**0.156 (±0.005)**	**0.247 (±0.002)**	**0.340 (±0.003)**	**0.343 (±0.003)**	**0.313 (±0.003)**

Table 1. Performance of benchmark approaches and our proposed method on five ranking datasets.

We evaluated our low-rank SELF learning algorithm on the following datasets:

Movielens. Movielens 100k ($mk100k$)[2] consists of ratings (1 to 5) provided by 943 users for a set of 1682 movies. A total of $100,000$ ratings available.

Jester. The Jester[3] datasets consist of user ratings of 100 jokes where ratings range from -10 to 10. Three datasets are available: $jester1$ with $24,983$ users $jester2$ with $23,500$ users and $jester3$ with $24,938$.

Sushi. The Sushi[4] dataset consists of ratings provided by 5000 people on 100 different types of sushi. Ratings ranged from 1 to 5 and only $50,000$ ratings are available. Additional features for users (e.g. gender, age) and sushi type (e.g. style, price) are provided.

We compared our approach to a number of ranking methods: MART (Friedman, 2001), RankNet (Burges et al., 2005), RankBoost (Freund et al., 2003), AdaRank (Xu & Li, 2007), Coordinate Ascent (Metzler & Croft, 2007), LambdaMART (Wu et al., 2010), ListNet, Random Forest. For all methods we used the implementation provided by RankLib[5] library. We also compared with the SVMrank (Joachims, 2006) approach using the implementation made available online by the authors. We also evaluated the original SELF approach in (Ciliberto et al., 2017) (SELF + $\|\cdot\|_{\mathsf{HS}}$).

We used a linear kernel on the input and for each dataset we performed parameter selection using 50% of the available ratings of each user for training, 20% for validation and the remaining for testing. We averaged the performance across 5 trials, each time considering a different random train/validation/test split.

Results. Table 1 reports the average performance of the tested methods across five independent trials. Prediction errors are measured in terms of the pair-wise loss in Eq. (18), normalized between 0 and 1. The performance of both SELF approaches significantly outperforms the competitors, in line with (Ciliberto et al., 2017), where the nonlinear MTL approach based on the SELF framework already improved upon state of the art ranking methods. Our proposed algorithm achieves an even lower prediction error on all datasets. This supports the idea motivating this work that leveraging the low-rank relations can provide significant advantages in practice.

6. Conclusions

This work combines structured prediction methods based on surrogate approaches with multitask learning techniques. Building on a previous framework for structured prediction we derived a trace norm regularization strategy that does not require explicit knowledge of the coding function. This led to a learning algorithm that can be efficiently applied in practice also when the surrogate space is large or infinite dimensional. We studied the generalization properties of the proposed estimator in terms of excess risk bounds for the surrogate learning problem. Our results on trace norm regularization with least-squares loss are, to our knowledge, novel and can be applied also to other settings such as collaborative filtering with side information. Experiments on ranking applications showed that relations between surrogate output can be beneficial in practice.

A question opened by our study is whether other multitask regularizers could be similarly adopted. As mentioned in the paper, even well-established approaches, such as those based on incorporating prior knowledge of the similarity between tasks pairs within the regularizer, do not necessarily extend to our setting. Future work will also investigate whether alternative surrogate loss functions to the canonical least-squares loss could be considered to enforce desirable tasks relations between the surrogate outputs more explicitly.

[2] `http://grouplens.org/datasets/movielens/`

[3] `http://goldberg.berkeley.edu/jester-data/`

[4] `http://www.kamishima.net/sushi/`

[5] `https://sourceforge.net/p/lemur/wiki/RankLib/`

References

Abernethy, J. D., Bach, F. R., Evgeniou, T., and Vert, J. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, Dec 2008.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Bach, F. R. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

Bakir, G., Hofmann, T., Schölkopf, B., Smola, A., Taskar, B., and Vishwanathan, S. Predicting structured data. neural information processing, 2007.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

Brouard, C., Szafranski, M., and D'Alché-Buc, F. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152, 2016.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96. ACM, 2005.

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

Ciliberto, C., Rosasco, L., and Rudi, A. A consistent regularization approach for structured prediction. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4412–4420. 2016.

Ciliberto, C., Rudi, A., Rosasco, L., and Pontil, M. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, pp. 1983–1993, 2017.

Cortes, C., Mohri, M., and Weston, J. A general regression technique for learning transductions. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pp. 153–160, 2005.

Cortes, C., Kuznetsov, V., Mohri, M., and Yang, S. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems*, pp. 2514–2522, 2016.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, pp. 327–334, 2010.

Fergus, R., Bernal, H., Weiss, Y., and Torralba, A. Semantic label sharing for learning with many categories. In *European Conference on Computer Vision*, pp. 762–775. Springer, 2010.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Fürnkranz, J. and Hüllermeier, E. Pairwise preference learning and ranking. In *Proceedings of the 14th European Conference on Machine Learning*, ECML'03, pp. 145–156. Springer-Verlag, 2003.

Gao, W. and Zhou, Z.-H. On the consistency of multi-label learning. *Artificial Intelligence*, 199:22–44, 2013.

Honeine, P. and Richard, C. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2): 77–88, 2011.

Jacob, L., Bach, F., and Vert, J.-P. Clustered multi-task learning: A convex formulation. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, pp. 745–752, 2008.

Joachims, T. Training linear svms in linear time. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

Journée, M., Bach, F., Absil, P.-A., and Sepulchre, R. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

Kadri, H., Ghavamzadeh, M., and Preux, P. A generalized kernel approach to structured output learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 471–479, 2013.

Kollo, T. and von Rosen, D. *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media, 2006.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 10 2011.

London, B., Huang, B., and Getoor, L. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Maurer, A. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

Maurer, A. and Pontil, M. Excess risk bounds for multitask learning with trace norm regularization. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013*, pp. 55–76, 2013.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

Metzler, D. and Croft, W. B. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Micchelli, C. A., Morales, J., and Pontil, M. Regularizers for structured sparsity. *Adv. Comput. Math.*, 38(3):455–489, 2013.

Minsker, S. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127: 111–119, 2017.

Mroueh, Y., Poggio, T., Rosasco, L., and jeacques Slotine, J. Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems 25*, pp. 2789–2797. 2012.

Nowak, A., Bach, F., and Rudi, A. Sharp analysis of learning with discrete losses. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1920–1929, 2019.

Nowozin, S., Lampert, C. H., et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011.

Oneto, L., Ridella, S., and Anguita, D. Tikhonov, ivanov and morozov regularization for support vector machine learning. *Machine Learning*, 103:103–136, 2016.

Osokin, A., Bach, F., and Lacoste-Julien, S. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pp. 302–313, 2017.

Rudi, A., Ciliberto, C., Marconi, G. M., and Rosasco, L. Manifold structured prediction. *Advances in Neural Information Processing Systems*, 2018.

Schölkopf, B., Smola, A. J., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Slater, P. Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312, 1961.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Srebro, N., Rennie, J., and Jaakkola, T. S. Maximum-margin matrix factorization. In Saul, L. K., Weiss, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 17*, pp. 1329–1336. MIT Press, 2005.

Taskar, B., Guestrin, C., and Koller, D. Max-margin markov networks. In *Advances in neural information processing systems*, pp. 25–32, 2004.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pp. 896–903. ACM, 2005.

Weston, J., Chapelle, O., Elisseeff, A., Schölkopf, B., and Vapnik, V. Kernel dependency estimation. In *Advances in Neural Information Processing Systems 15*, pp. 873–880, Cambridge, MA, USA, October 2003. Max-Planck-Gesellschaft, MIT Press.

Wu, Q., Burges, C. J., Svore, K. M., and Gao, J. Adapting boosting for information retrieval measures. *Information Retrieval*, 13 (3):254–270, 2010.

Xu, J. and Li, H. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 391–398. ACM, 2007.