Consistent Structured Prediction with Max-Min Margin Markov Networks

Alex Nowak-Vila [1] **Francis Bach** [1] **Alessandro Rudi** [1]

Abstract

Max-margin methods for binary classification such as the support vector machine (SVM) have been extended to the structured prediction setting under the name of max-margin Markov networks ($\mathrm{M}^3\mathrm{N}$), or more generally structural SVMs. Unfortunately, these methods are statistically inconsistent when the relationship between inputs and labels is far from deterministic. We overcome such limitations by defining the learning problem in terms of a "max-min" margin formulation, naming the resulting method max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$). We prove consistency and finite sample generalization bounds for $\mathrm{M}^4\mathrm{N}$ and provide an explicit algorithm to compute the estimator. The algorithm achieves a generalization error of $O(1/\sqrt{n})$ for a total cost of $O(n)$ projection-oracle calls (which have at most the same cost as the max-oracle from $\mathrm{M}^3\mathrm{N}$). Experiments on multi-class classification, ordinal regression, sequence prediction and ranking demonstrate the effectiveness of the proposed method.

1. Introduction

Many classification tasks in machine learning lie beyond the classical binary and multi-class classification settings. In those tasks, the output elements are structured objects made of interdependent parts, such as sequences in natural language processing (Smith, 2011), images in computer vision (Nowozin & Lampert, 2011), permutations in ranking or matching problems (Caetano et al., 2009) to name just a few (BakIr et al., 2007). The structured prediction setting has two key properties that makes it radically different from multi-class classification, namely, the exponential growth of the size of the output space with the number of its parts, and the cost-sensitive nature of the learning task, as prediction mistakes are not equally costly. In sequence prediction, for instance, the number of possible outputs grows exponentially with the length of the sequences, and predicting a sequence with one incorrect character is better than predicting the whole sequence wrong.

Classical approaches in binary classification such as the *non-smooth* support vector machine (SVM), and the *smooth* logistic and quadratic plug-in classifiers have been extended to the structured setting under the name of max-margin Markov networks ($\mathrm{M}^3\mathrm{N}$) (Taskar et al., 2004) (or more generally structural SVM (SSVM) (Tsochantaridis et al., 2005)), conditional random fields (CRFs) (Lafferty et al., 2001) and quadratic surrogate (QS) (Ciliberto et al., 2016; 2019), respectively. Theoretical properties of CRF and QS are well-understood. In particular, it is possible to obtain finite-sample generalization bounds of the resulting estimator on the cost-sensitive structured loss (Nowak-Vila et al., 2019a). Unfortunately, these guarantees are not satisfied by $\mathrm{M}^3\mathrm{N}$s even though the method is based on an upper bound of the loss. More precisely, it is known that the upper bound can be not tight (and lead to inconsistent estimation) when the relationship between input and output labels is far from deterministic (Liu, 2007), which it is essentially always the case in structured prediction due to the exponentially large output space. This means that the estimator does not converge to the minimizer of the problem leading to inconsistency.

Recently, a line of work (Fathony et al., 2016; 2018a;b;c) proposed a consistent method based on an adversarial game formulation on the structured problem. However, their analysis does not allow to get generalization bounds and their proposed algorithm is specific for every setting with at least a complexity of $O(n^2)$ to obtain optimal statistical error when learning from n samples. In this paper, we derive this method in the generic structured output setting from first principles coming from the binary SVM. We name this method max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$), as it is based on a correction of the max-margin of $\mathrm{M}^3\mathrm{N}$ to a 'max-min' margin. The proposed algorithm has essentially the same complexity as state-of-the-art methods for $\mathrm{M}^3\mathrm{N}$ on the regularized empirical risk minimization problem, but it comes with consistency guarantees and finite sample generalization bounds on the discrete structured prediction loss, with constants that are polynomial in the number of parts

[1]INRIA - Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University. Correspondence to: Alex Nowak-Vila <alex.nowak-vila@inria.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

of the structured object and do not scale as the size of the output space. More precisely, the algorithm requires a constant number of projection-oracles at every iteration, each of them having at most the same cost as the max-oracle of $\mathrm{M}^3\mathrm{N}$. We also provide experiments on multiple tasks such as multi-class classification, ordinal regression, sequence prediction and ranking, showing the effectiveness of the algorithm. We make the following contributions:

- We introduce max-min margin Markov networks ($\mathrm{M}^4\mathrm{N}$) in Definition 3.1 and prove consistency, linear calibration and finite sample generalization bounds for the regularized ERM estimator in Thms. 3.2, 3.3 and 3.4, respectively.
- We generalize the BCFW algorithm (Lacoste-Julien et al., 2013) used for $\mathrm{M}^3\mathrm{N}$s to $\mathrm{M}^4\mathrm{N}$s and solve the max-min oracle iteratively with projection oracle calls using Saddle Point Mirror Prox (Nemirovski, 2004). We prove bounds on the expected duality gap of the regularized ERM problem in Theorem 5.1 and statistical bounds in Theorem 5.2.
- In Section 6, we perform a thorough experimental analysis of the proposed method on classical unstructured and structured prediction settings.

2. Surrogate Methods for Classification

In this section, we review the first principles underlying surrogate methods starting from binary classification and moving into structured prediction. We put special attention to the difference between plug-in (e.g., logistic) and direct (e.g., SVM) classifiers to show that while there is a complete picture in the binary setting, existing direct classifiers in structured prediction lack the basic properties of binary SVMs. The first goal of this paper is to complete this picture in the structured output setting.

2.1. A Motivation from Binary Classification

Let $\mathcal{Y} = \{-1, 1\}$ and $(x_1, y_1), \dots, (x_n, y_n)$ be n input-output pairs sampled from a distribution ρ. The goal in binary classification is to estimate a binary-valued function $f^\star : \mathcal{X} \to \mathcal{Y}$ that minimizes the classification error

$$\mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho}\, 1(f(x) \neq y).$$

We can avoid working with binary-valued functions by considering instead real-valued functions $g : \mathcal{X} \to \mathbb{R}$ and use the prediction model $f(x) = d \circ g(x) \coloneqq \operatorname{sign}(g(x))$ (Bartlett et al., 2006) where d stands for *decoding*. The resulting problem reads

$$g^\star \in \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}} \mathcal{E}(d \circ g). \tag{1}$$

Unfortunately, directly estimating a $g^\star$ from (1) is intractable for many classes of functions (Arora et al., 1997).

Convex surrogate methods. The source of intractability of minimizing the classification error (1) comes from the discreteness and non-convexity of the loss. The idea of surrogate methods (Bartlett et al., 2006) is to consider a *convex surrogate loss* $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ such that $g^\star$ can be written as

$$g^\star = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}} \mathcal{R}(g) \coloneqq \mathbb{E}_{(x,y)\sim\rho}\, S(g(x), y). \tag{2}$$

In this case, $g^\star$ can be tractably estimated from n samples over a family of functions $\mathcal{G}$ using regularized ERM. The resulting estimator g_n has the form

$$g_n = \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} S(g(x_i), y_i) + \frac{\lambda_n}{2}\|g\|_{\mathcal{G}}^2, \tag{3}$$

where $\lambda_n > 0$ is the regularization parameter and $\|\cdot\|_{\mathcal{G}}$ is the norm associated to the hypothesis space $\mathcal{G}$. If not stated explicitly, our analysis of the surrogate method holds for any function space, such as reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950) or neural networks (LeCun et al., 2015), where we lose global theoretical convergence guarantees of problem (3).

The classical theoretical requirements of such a surrogate strategy are *Fisher consistency* (i) and a *comparison inequality* (ii):

$$\begin{aligned} &\text{(i)} \quad \mathcal{E}(f^\star) = \mathcal{E}(d \circ g^\star) \\ &\text{(ii)} \quad \zeta(\mathcal{E}(d \circ g) - \mathcal{E}(f^\star)) \leq \mathcal{R}(g) - \mathcal{R}(g^\star), \end{aligned}$$

for all measurable functions g, where $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ is such that $\zeta(\varepsilon) \to 0$ when $\varepsilon \to 0$. Note that Condition (i) is equivalent to (1). Condition (ii) is needed to prove consistency results, to show that $\mathcal{R}(g) \to \mathcal{R}(g^\star)$ implies $\mathcal{E}(d \circ g) \to \mathcal{E}(f^\star)$. The existence of ζ satisfying (ii) is derived from (i) and the continuity and lower boundedness of $S(v, y)$, see Thm. 3 by (Zhang, 2004). Even though the explicit form of ζ is not needed for a consistency analysis, it is necessary to prove finite sample generalization bounds, as it is the mathematical object relating the suboptimality of the surrogate problem to the suboptimality of the original task. Note that the larger $\zeta(\varepsilon)$, the better.

Plug-in classifiers. It is known that (i) is satisfied for any function $g^\star$ that continuously depends on the conditional probability $\rho(1|x)$ as $g^\star(x) \coloneqq t(\rho(1|x))$, where $t : \mathbb{R} \to \mathbb{R}$ is a suitable continuous bijection of the real line[1]. In this case, Eq. (2) can be satisfied using *smooth* losses. Some examples are the logistic loss $\log(1+e^{-yv})$, the squared hinge

[1] It must satisfy $(u - 1/2)t(u - 1/2) \geq 0$ for all $u \in \mathbb{R}$.

loss $\max(0, 1-yv)^2$ and the exponential loss e^{-yv}. In this case, the convexity and smoothness of $S(\cdot, y)$ imply that (ii) is satisfied with $\zeta(\varepsilon) \sim \varepsilon^2$ (Bartlett et al., 2006). Combining this with standard convergence results of regularized ERM estimators g_n on RKHS, the resulting statistical rates are of the form $\mathbb{E}\,\mathcal{E}(d \circ g_n) - \mathcal{E}(f^\star) \sim \|g^\star\|_{\mathcal{G}} n^{-1/4}$. Even if the binary learning problem is easy, $g^\star$ can be highly non-smooth away from the decision boundary, resulting in large $\|g^\star\|_{\mathcal{G}}$. It is known that the dependence on the number of samples can be improved under low noise conditions (Audibert & Tsybakov, 2007).

Support vector machines (SVM). Plug-in classifiers indirectly estimate the conditional probability as $\rho(1|x) = t^{-1}(g^\star(x))$, which is more than just falling in the right binary decision set. SVMs directly tackle the classification task by estimating $g^\star := f^\star = \operatorname{sign}(\rho(1|x) - 1/2)$. In this case, the *non-smooth* hinge loss $S(v, y) = \max(0, 1 - yv)$ satisfies (2). Moreover, (ii) is satisfied with $\zeta(\varepsilon) = \varepsilon$ and statistical rates are of the form $\mathbb{E}\,\mathcal{E}(d \circ \widehat{g}_n) - \mathcal{E}(f^\star) \sim \|f^\star\|_{\mathcal{G}} n^{-1/2}$. Note that $f^\star$ is piece-wise constant on the support of ρ, but it can be shown $f^\star \in \mathcal{G}$, (i.e., $\|f^\star\|_{\mathcal{G}} < \infty$), for standard hypothesis spaces $\mathcal{G}$ such as Sobolev spaces with input space $\mathbb{R}^d$ and smoothness $s > d/2$ under low noise conditions (Pillaud-Vivien et al., 2018b).

2.2. Structured Prediction Setting

In binary classification, the output data are naturally embedded in $\mathbb{R}$ as $\mathcal{Y} = \{-1, 1\} \subset \mathbb{R}$. However, as this is not necessarily the case in structured prediction, it is classical (Taskar et al., 2005) to represent the output with an embedding $\varphi : \mathcal{Y} \to \mathbb{R}^k$ encoding the parts structure with $k \ll |\mathcal{Y}|$. Let $g : \mathcal{X} \to \mathbb{R}^k$ and define the following linear prediction model

$$f(x) = d \circ g(x) := \arg\max_{y \in \mathcal{Y}} \ \varphi(y)^\top g(x). \tag{4}$$

The above decoding (4) corresponds to the classical linear prediction model over factorized joint features $\Phi(x, y) = \varphi(y) \otimes \Phi(x)$ when $g(x)$ is linear in some input features $\Phi(x)$ (BakIr et al., 2007). The form in (4) is required to perform the consistency analysis but the algorithm developed in Section 5 can be readily extended to joint features that do not factorize. Non-linear prediction models have been recently proposed by (Belanger & McCallum, 2016), but this is out of the scope of this paper.

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function between structured outputs encoding the cost-sensitivity of predictions. For instance, it is common to take L to be the Hamming loss over the parts of the structured object. The goal in structured prediction is to estimate $f^\star : \mathcal{X} \to \mathcal{Y}$ that minimizes the *expected risk*:

$$\mathcal{E}(f) = \mathbb{E}_{(x,y)\sim\rho} L(f(x), y). \tag{5}$$

Loss-decoding compatibility. It is classical to assume that the loss decomposes over the structured output parts (Joachims, 2006). This can be generalized as the following affine decomposition of the loss (Ramaswamy et al., 2013; Nowak-Vila et al., 2019b)

$$L(y, y') = \varphi(y)^\top A \varphi(y') + a, \tag{6}$$

for a matrix $A \in \mathbb{R}^{k \times k}$ and scalar $a \in \mathbb{R}$. Indeed, assumption (6) together with the tractability of (4) is essentially equivalent to the tractability of *loss-augmented inference* in structural SVMs (Joachims, 2006). For the sake of notation, we drop the constant a and work with the 'centered' loss $L(y, y') - a$. We provide some examples below.

Example 2.1 (Structured prediction with factor graphs)**.** *Let $\mathcal{Y} = [R]^M$ be the set of objects made of M parts, each in a vocabulary of size R. In order to model interdependence between different parts, we consider embeddings that decompose over (overlapping) subsets of indices $\alpha \subseteq \{1, \dots, M\}$ (Taskar et al., 2004) as $\varphi(y) = (\varphi_\alpha(y_\alpha))_\alpha$. More precisely, the prediction model corresponds to*

$$\arg\max_{y \in \mathcal{Y}} \ \sum_\alpha \varphi_\alpha(y_\alpha)^\top v_\alpha, \tag{7}$$

where $\varphi_\alpha(y_\alpha) = e_{y_\alpha} \in \mathbb{R}^{R^{|\alpha|}}$ with e_j being the j-th vector of the canonical basis and the dimension of the full-embedding φ is $k = \sum_\alpha R^{|\alpha|} \ll |\mathcal{Y}| = R^M$. It is common (Tsochantaridis et al., 2005) to assume that the loss decomposes additively over the coordinates as $L(y, y') = \frac{1}{M}\sum_{m=1} L_m(y_m, y'_m)$ and so the matrix A associated to the loss decomposition of L is low-rank. Problem (7) *can be solved efficiently for low tree-width structures using the junction-tree algorithm (Wainwright & Jordan, 2008). More specifically, if the objects are sequences with embeddings modelling individual and adjacent pairwise characters, Problem* (7) *can be solved in time $O(MR^2)$ using the Viterbi algorithm (Viterbi, 1967).*

Example 2.2 (Ranking and matching)**.** *The output space is the group of permutations $\mathcal{S}_M$ acting on $\{1, \dots, M\}$. This setting also includes the task of matching the nodes of two graphs of the same size (Caetano et al., 2009). We represent a permutation $\sigma \in \mathcal{S}_M$ using the corresponding permutation matrix $\varphi(\sigma) = P_\sigma \in \mathbb{R}^{M \times M}$. The prediction model corresponds to the linear assignment problem (Burkard et al., 2012)*

$$\arg\max_{\sigma \in \mathcal{S}_M} \ \langle P_\sigma, v \rangle_F, \tag{8}$$

where $v \in \mathbb{R}^{M \times M}$, $\langle \cdot, \cdot \rangle_F$ is the Frobenius scalar product and $k = M^2 \ll |\mathcal{Y}| = M!$. The Hamming loss on permutations satisfies Eq. (6) *as $L(\sigma, \sigma') = \frac{1}{M}\sum_{m=1}^M 1(\sigma(m) \neq \sigma'(m)) = 1 - \frac{1}{M}\langle P_\sigma, P_{\sigma'} \rangle_F$. The linear assignment problem* (8) *can be solved in time $O(M^3)$ using the Hungarian algorithm (Kuhn, 1955).*

Plug-in classifiers in structured prediction. Let $\mu(x) = \mathbb{E}_{y\sim\rho(\cdot|x)}\,\varphi(y)$ be the conditional expectation of the output embedding. Using the fact that $f^\star$ can be characterized pointwise in x as the minimizer in y of $\varphi(y)^\top A\mu(x)$ (Nowak-Vila et al., 2019b), it directly follows that (i) is satisfied for $g^\star(x) = -A\mu(x)$ and, analogously to binary classification, it can be estimated using *smooth* surrogates. Some examples are the quadratic surrogate (QS) $\|v + A\varphi(y)\|_2^2$ (Ciliberto et al., 2016) that estimates $g^\star$ and conditional random fields (CRF) (Lafferty et al., 2001) defined by $\log\big(\sum_{y'\in\mathcal{Y}}\exp v^\top\varphi(y')\big) - v^\top\varphi(y)$ that estimate an invertible continuous transformation of $\mu(x)$. Although CRFs have a powerful probabilistic interpretation, they cannot incorporate the cost-sensitivity matrix A into the surrogate loss, and it must be added a posteriori in the decoding (4) to guarantee consistency. It was shown by (Nowak-Vila et al., 2019a) that these methods satisfy condition (ii) with $\zeta(\varepsilon) \sim \varepsilon^2$ and achieve the analogous statistical rates of binary plug-in classifiers $\sim \|g^\star\|_{\mathcal{G}} n^{-1/4}$.

SVMs for structured prediction. The extension of binary SVM to structured outputs is the structural SVM (SSVM) (Joachims, 2006) (denoted $\mathrm{M^3Ns}$ (Taskar et al., 2004) in the factor graph setting described in Example 2.1). It corresponds to the following surrogate loss

$$S(v,y) = \max_{y'\in\mathcal{Y}}\ \varphi(y)^\top A\varphi(y') + v^\top\varphi(y') - v^\top\varphi(y). \quad (9)$$

In the multi-class case with $\mathcal{Y} = \{1,\dots,k\}$ and $L(y,y') = 1(y \neq y')$ it is also known as the Crammer-Singer SVM (CS-SVM) (Crammer & Singer, 2001) and reads $S(v,j) = \max_{r\neq j}\ 1 + v_r - v_j$. It shares some properties of the binary SVM such as the upper bound property, i.e., $L(d\circ v, y) \leq S(v,y)$ for all $y \in \mathcal{Y}$. However, an important drawback of this loss is that while the upper bound property holds, the minimizer of the surrogate expected risk $g^\star$ and the one of the expected risk $f^\star$ do not coincide when the problem is far from deterministic, as shown by the following Proposition 2.3.

Proposition 2.3 (Inconsistency of CS-SVM (Liu, 2007))**.** *The CS-SVM is Fisher-consistent if and only if for all $x \in \mathcal{X}$, there exists $y \in \{1,\dots,k\}$ such that $\rho(y|x) > 1/2$.*

Note that the consistency condition from Proposition 2.3 is much harder to be met in the structured prediction case as the size of the output space is exponentially large, and it is always satisfied in the binary case (the binary SVM is always consistent). Although there exist consistent extensions of the SVM to the cost-sensitive multi-class setting such as the ones from (Lee et al., 2004; Mroueh et al., 2012), they cannot be naturally extended to the structured setting. In the following section we address this problem by introducing the max-min surrogate and studying its theoretical properties.

3. Max-Min Surrogate Loss

Assume that the loss is not degenerated, i.e., $L(y,y) < L(y,y')$ for all $y, y' \in \mathcal{Y}$ such that $y \neq y'$. In this case, $f^\star(x)$ is the minimizer in y of $\varphi(y)^\top A^\top\varphi(f^\star(x))$, which means that (1) is satisfied by

$$g^\star(x) \coloneqq -A^\top\varphi(f^\star(x)) \in \mathbb{R}^k.$$

Note the analogy with SVMs, where we directly estimate $f^\star$ but now through the representation φ of the structured output, avoiding the full enumeration of $\mathcal{Y}$. We need to find a surrogate function $S(v,y)$ that satisfies Eq. (2) for this $g^\star$. Following the same notation as (Nowak-Vila et al., 2019a), we define the *marginal polytope* (Wainwright & Jordan, 2008) as the convex hull of the embedded output space $\mathcal{M} = \mathrm{hull}(\varphi(\mathcal{Y})) \subset \mathbb{R}^k$.

Definition 3.1 (Max-min surrogate loss)**.** *Define the max-min loss as*

$$S(v,y) \coloneqq \max_{\mu\in\mathcal{M}}\min_{y'\in\mathcal{Y}}\ \varphi(y')^\top A\mu + v^\top\mu - v^\top\varphi(y). \quad (10)$$

The max-min loss is *non-smooth*, *convex* and can be cast as a Fenchel-Young loss (Blondel et al., 2020). More specifically, Eq. (10) can be written as $S(v,y) = \Omega^*(v) + \Omega(\varphi(y)) - v^\top\varphi(y)$ with

$$\Omega(\mu) = -\min_{y'\in\mathcal{Y}}\ \varphi(y')^\top A\mu + 1_{\mathcal{M}}(\mu), \quad (11)$$

where $\Omega(\varphi(y)) = 0$ for all $y \in \mathcal{Y}$, Ω^* denotes the Fenchel-conjugate of Ω, and $1_{\mathcal{M}}(\mu) = 0$ if $\mu \in \mathcal{M}$ and $+\infty$ otherwise.

Note that the dependence on y is only in the linear term $v^\top\varphi(y)$, while for SSVMs (9) it appears in the maximization. Thus, we can study the geometry of the loss through the non-smooth convex function $\Omega^*(v)$ (see Figure 1 for visualizations of some representative unstructured examples). Connections between surrogates (10) and (9) are discussed in Section 4.

3.1. Fisher Consistency

Fisher consistency is provided by the following Theorem 3.2.

Theorem 3.2 (Fisher Consistency (i))**.** *The surrogate loss (10) satisfies (i) for $g^\star(x) = -A^\top\varphi(f^\star(x))$.*

This result has been proven by Fathony et al. (2018a) in the cost-sensitive multi-class case. Our proof of Theorem 3.2 is constructive and based on Fenchel duality.

Sketch of the proof. We want to show that $-A^\top\varphi(f^\star(x))$ is the minimizer of $\mathbb{E}_{y\sim\rho(\cdot|x)}\ S(v,y)$ almost surely for every x. The proof is constructive and based on Fenchel duality, using the Fenchel-Young loss representation of the

max-min surrogate. First, note that the conditional surrogate risk can be written as $\mathbb{E}_{y\sim\rho(\cdot|x)}\ S(v,y) = \Omega^*(v) - v^\top\mu(x)$, where $\mu(x) = \mathbb{E}_{y\sim\rho(\cdot|x)}\,\varphi(y) \in \mathcal{M}$. Second, note that by Fenchel-duality, $\partial_\mu\Omega(\mu(x))$ is the set of minimizers of $\Omega^*(v) - v^\top\mu(x)$. Finally, if we assume that the set of $x \in \mathcal{X}$ such that $\mu(x)$ is in the boundary of $\mathcal{M}$ has measure zero, then

$$-A^\top f^\star(x) \in \partial_\mu\Omega(\mu(x)),$$

where Ω is defined in (11) and we have used that $f^\star(x)$ is the minimizer in y of $\varphi(y)^\top A\mu(x)$. A more detailed proof can be found in Appendix C.1. □

3.2. Comparison Inequality

Fisher consistency is not enough to prove finite-sample generalization bounds on the excess risk $\mathcal{E}(d\circ g) - \mathcal{E}(f^\star)$. For this, we provide in the following Theorem 3.3 an explicit form of the comparison inequality.

Theorem 3.3 (Comparison inequality (ii))**.** *Assume L is symmetric and that there exists $C > 0$ such that for any probability $\alpha \in \Delta_\mathcal{Y}$, it holds that $\alpha_y \geq 1/C$ for $y \in \arg\min_{y\in\mathcal{Y}} \mathbb{E}_{z\sim\alpha} L(y,z)$. Then, the comparison inequality (ii) for the max-min loss* (10) *reads*

$$\mathcal{E}(d\circ g) - \mathcal{E}(f^\star) \leq C(\mathcal{R}(g) - \mathcal{R}(g^\star)).$$

The second condition on the loss states that if y is optimal for x, then its conditional probability is bounded away from zero as $\rho(y|x) \geq 1/C$. This condition is used to obtain a simple quantitative lower bound on the function ζ of (ii) and more tight (albeit less explicit in general) expressions of the constant C can be found in Appendices C.3 and C.4.

Constant C for multi-class. When $L(y,y') = 1(y \neq y')$ with $\mathcal{Y} = \{1,\dots,k\}$, we have that $C = k$, as the minimum conditional probability of an optimal output is $1/k$. The constant for this specific setting was derived independently using a different analysis by Duchi et al. (2018).

Constant C for factor graphs (Example 2.1). For a factor graph with separable embeddings and a decomposable loss $L = \frac{1}{M}\sum_{m=1}^M L_m(y_m, y'_m)$, we have that $C = \max_{m\in[M]} C_m$, where C_m is the constant associated to the individual loss L_m. This is proven in Proposition C.11.

Constant C for ranking and matching (Example 2.2). In this setting, Theorem 3.3 gives $C = M!$, and so the relation between both excess risks is not informative. The problem of exponential constants in the comparison inequality was pointed out by Osokin et al. (2017). We can weaken the assumption and change condition $\alpha_y \geq 1/C$ to

$$\max_{\beta\in\Delta_\mathcal{Y}} \beta_y \ \text{ s.t. } \ \mathbb{E}_{z\sim\beta}\,\varphi(z) = \mathbb{E}_{z'\sim\alpha}\,\varphi(z') \geq 1/C.$$

Under this assumption, we have that $C = M$, thus avoiding the exponentially large size of the output space.

3.3. Generalization of Regularized ERM

In the following Theorem 3.4, we use this result to prove a finite-sample generalization bound on the regularized ERM estimator (3) when the hypothesis space $\mathcal{G}$ is a vector-valued RKHS.

Theorem 3.4 (Generalization of regularized ERM)**.** *Let $\mathcal{G}$ be a vector-valued RKHS, assume $g^\star \in \mathcal{G}$ and let g_n and $\lambda_n = \kappa L \log^{1/2}(1/\delta) n^{-1/2}$ as in* (3)*. Then, with probability $1-\delta$:*

$$\mathcal{E}(d\circ g_n) - \mathcal{E}(f^\star) \leq M\|\varphi(f^\star)\|_\mathcal{G}\sqrt{\frac{\log(1/\delta)}{n}},$$

with $M = \kappa C L\|A\|$. Here, $L = 2\max_{y\in\mathcal{Y}}\|\varphi(y)\|_2$, $\|A\| = \sup_{\|v\|_2\leq 1}\|Av\|_2$, $\kappa = \sup_{x\in\mathcal{X}}\operatorname{Tr} K(x,x)^{1/2}$ is the size of the features and C is the one of Theorem 3.3.

Analogously to the binary case, the multivariate function $\varphi(f^\star)$ is piecewise constant on the support of the distribution ρ. In Theorem D.2 in Appendix D we prove that standard low noise conditions, analogous to the one discussed by Pillaud-Vivien et al. (2018b) for the binary case, are enough to guarantee $\|\varphi(f^\star)\|_\mathcal{G} < \infty$.

4. Comparison with Structural SVM

Max-min as a correction of the Structural SVM. We can re-write the maximization over the discrete output space $\mathcal{Y}$ in the definition of the SSVM (9) as a maximization over its convex hull $\mathcal{M} = \operatorname{hull}(\varphi(\mathcal{Y}))$

$$S(v,y) = \max_{\mu\in\mathcal{M}}\ \varphi(y)^\top A\mu + v^\top\mu - v^\top\varphi(y). \quad (12)$$

Note the similarity between (10) and (12). In particular, the max-min loss differs from the structural SVM in that the maximization is done using $\min_{y'\in\mathcal{Y}}\varphi(y')^\top A\mu$ and not the loss at the observed output y as $\varphi(y)^\top A\mu$. Hence, we can view the max-min surrogate as a *correction* of the SSVM so that basic statistical properties (i) and (ii) hold. Moreover, this connection might be used to properly understand the statistical properties of SSVM. This is left for future work.

Notion of max-min margin. Given $v \in \mathbb{R}^k$ and $y_i \in \mathcal{Y}$, the classical SSVM is motivated by a soft version of the following notion of margin:

$$v^\top\varphi(y_i) - v^\top\varphi(y) \geq L(y_i,y) = \varphi(y_i)^\top A\varphi(y),$$

for all $y \in \mathcal{Y}$, which is equivalent to $v^\top\varphi(y_i) - v^\top\mu \geq \varphi(y_i)^\top A\mu$ for all $\mu \in \mathcal{M}$. However, we have seen in Proposition 2.3 that this condition is too strong and only leads to a consistent method if the problem is nearly deterministic,

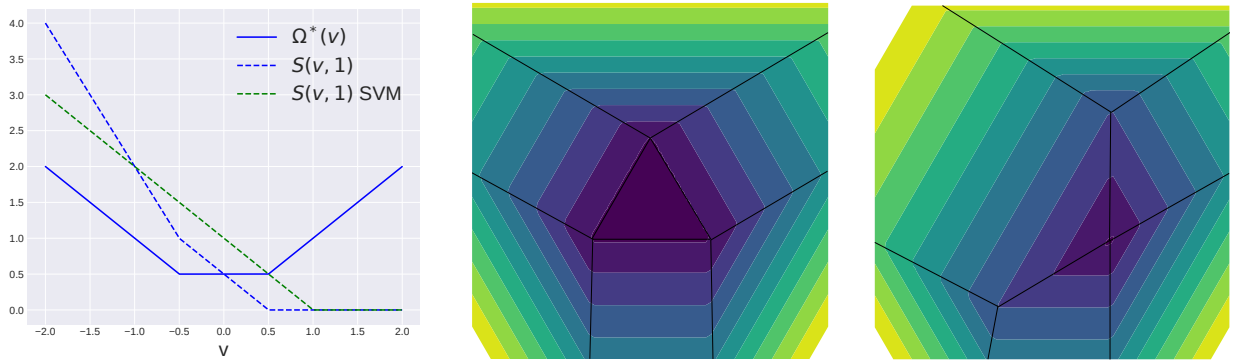

Figure 1. **Left:** The binary max-min loss has two symmetric kinks instead of one as the SVM. **Middle:** $\Omega^*(v)$ in $v^\top 1 = 0$ for multi-class 0-1 loss $1(y \neq y')$ with $k = 3$. **Right:** $\Omega^*(v)$ in $v^\top 1 = 0$ for ordinal regression with the absolute loss $|y - y'|$ with $k = 3$.

i.e., we observe the optimal y with large probability, which, as already mentioned, is generally far from true in structured prediction. The max-min surrogate (10) deals with the case where this strong condition is not met and works with a notion of margin that compares groups of outputs instead of just pairs. We define the max-min margin as

$$v^\top \varphi(y_i) - v^\top \mu \geq \min_{y' \in \mathcal{Y}} \varphi(y')^\top A \mu, \qquad (13)$$

for all $\mu \in \mathcal{M}$. After introducing slack variables in (13) we obtain a soft version of the max-min margin that leads to the max-min regularized ERM problem (3).

5. Algorithm

In this section we derive a dual-based algorithm to solve the max-min regularized ERM problem (3) when the hypothesis space is a RKHS. The algorithm can be easily adapted to the case where g is parametrized using a neural network as commented at the end of Section 5.3.

5.1. Problem Formulation

Let $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R}^k\}$ be a vector-valued RKHS, which we assume of the form $\mathcal{G} = \mathbb{R}^k \otimes \overline{\mathcal{G}}$, where $\overline{\mathcal{G}}$ is a scalar RKHS with associated features $\Phi : \mathcal{X} \to \overline{\mathcal{G}}$. Every function in $\mathcal{G}$ can be written as $g_w(x) = w^\top \Phi(x) \in \mathbb{R}^k$ where $w_j, \Phi(x) \in \overline{\mathcal{G}}$. For the sake of presentation, we assume that $\mathcal{G} = \mathbb{R}^{d \times k}$ is finite dimensional, but our analysis also holds for the infinite dimensional case. The dual **(D)** of the regularized ERM problem (3) for the max-min surrogate loss (10) reads

$$\textbf{(D)} \quad \max_{\mu \in \mathcal{M}^n} \frac{1}{n} \sum_{i=1}^n \min_{y'} \varphi(y')^\top A \mu_i - \frac{\lambda}{2} \|\Phi_n(\mu - \varphi_n)\|_2^2,$$

where $\Phi_n = \frac{1}{\lambda n}(\Phi(x_1), \dots, \Phi(x_n))$ is the $d \times n$ scaled input data matrix and $\varphi_n = (\varphi(y_1), \dots, \varphi(y_n))^\top$ is the $n \times k$ output data matrix. The dual variables map to the primal variables through the mapping $w(\mu) = \frac{1}{\lambda n} \sum_{i=1}^n \Phi(x_i)(\mu_i - \varphi(y_i))^\top$. By strong duality, it holds $w^\star = w(\mu^\star)$. The dual formulation **(D)** is a constrained *non-smooth* optimization problem, where the non-smoothness comes from the first term of the objective function. In order to derive a learning algorithm, we leverage ideas from the block-coordinate Frank-Wolfe algorithm used for SSVMs.

5.2. Derivation of the Algorithm

Background on BCFW for $\mathbf{M^3Ns}$. The dual of the SSVM is the same as problem **(D)** but the first term is linear: $\frac{1}{n} \sum_{i=1}^n \varphi(y_i)^\top A \mu_i$, making the dual objective function *smooth*. The BCFW algorithm (Lacoste-Julien et al., 2013) minimizes a linearization of the smooth dual objective function block-wise, using the separability of the compact domain. At each iteration t, the algorithm picks $i \in [n]$ at random, and updates $\mu_i^{(t+1)} = (1 - \gamma)\mu_i^{(t)} + \gamma \bar{\mu}_i^{(t+1)}$ with $\bar{\mu}_i^{(t+1)} := \arg\max_{\mu_i' \in \mathcal{M}} \langle \mu_i', \nabla_{(i)} h(\mu^{(t)}) \rangle$ where h is the dual objective and γ is the step-size. Note that $\bar{\mu}_i^{(t+1)}$ is an extreme point of $\mathcal{M}$ and it can be written as a combinatorial maximization problem over $\mathcal{Y}$ that corresponds precisely to inference (4). In the next subsection, we generalize BCFW to the case where the dual is a sum of a non-smooth and a smooth function such as the dual **(D)** of our problem.

Generalized BCFW (GBCFW) for $\mathbf{M^4N}$. Borrowing ideas from Bach (2015) in the non block-separable case, we only linearize the smooth-part of the function, i.e., the quadratic term. We change the computation of the direction to

$$\bar{\mu}_i^{(t+1)} := \arg\max_{\mu_i' \in \mathcal{M}} \langle \mu_i', \nabla_{(i)} \frac{-\lambda}{2} \|\Phi_n(\mu^{(t)} - \varphi_n)\|_2^2 \rangle + \min_{y'} \varphi(y')^\top A \mu_i' = \mathcal{O}(g_{w(\mu^{(t)})}(x_i)),$$

where the max-min oracle $\mathcal{O} : \mathbb{R}^k \to \mathcal{M}$ is defined as

$$\mathcal{O}(v) = \arg\max_{\mu \in \mathcal{M}} \min_{\nu \in \mathcal{M}} \nu^\top A \mu + v^\top \mu. \qquad (14)$$

Algorithm 1 GBCFW (primal)

Let $w^{(0)} := w_i^{(0)} := 0$
for $t = 0$ **to** T **do**
  Pick i at random in $\{1, \dots, n\}$
  $(\mu_i^\star, \nu_i^\star) \in \mathcal{O}_K(g_{w^{(t)}}(x_i), \mu_i^\star, \nu_i^\star)$
  $w_s := \Phi(x_i)(\mu_i^\star - \varphi(y_i))^\top / (\lambda n)$
  $w_i^{(t+1)} := (1 - \frac{2n}{t+2n}) w_i^{(t)} + \frac{2n}{t+2n} w_s$
  $w^{(t+1)} := w^{(t)} + w_i^{(t+1)} - w_i^{(t)}$
end for

Algorithm 2 SP-MP $\quad (\bar\mu^{(K)}, \bar\nu^{(K)}) \in \mathcal{O}_K(v, \mu^{(0)}\nu^{(0)})$

for $k = 0$ **to** $K - 1$ **do**
  $\mu_{1/2}^{(k+1)} \in \arg\min_{\mu \in \mathcal{M}} -\eta \mu^\top (A^\top \nu^{(k)} + v) + D_{-H}(\mu, \mu^{(k)})$
  $\nu_{1/2}^{(k+1)} \in \arg\min_{\nu \in \mathcal{M}} \eta \nu^\top A \mu^{(k)} + D_{-H}(\nu, \nu^{(k)})$
  $\mu^{(k+1)} \in \arg\min_{\mu \in \mathcal{M}} -\eta \mu^\top (A^\top \nu_{1/2}^{(k+1)} + v) + D_{-H}(\mu, \mu^{(k)})$
  $\nu^{(k+1)} \in \arg\min_{\nu \in \mathcal{M}} \eta \nu^\top A \mu_{1/2}^{(k+1)} + D_{-H}(\nu, \nu^{(k)})$
end for
$\bar\mu^{(K)} := \frac{1}{K} \sum_{k=1}^K \mu^{(k)}$, $\bar\nu^{(K)} := \frac{1}{K} \sum_{k=1}^K \nu^{(k)}$

Note that the mapping $w(\mu)$ between primal and dual variables is affine. Hence, one can write the update of the primal variables without saving the dual variables as detailed in Algorithm 1. The following Theorem 5.1 specifies the required number of iterations of Algorithm 1 to obtain an ε-optimal solution with an approximate oracle (14).

Theorem 5.1 (Convergence of GBCFW with approximate oracle)**.** *Let $\varepsilon > 0$. If the approximate oracle provides an answer with error $\varepsilon/2$, then the final error of Algorithm 1 achieves an expected duality gap of ε when $T = \tilde{O}\big(n + \frac{2R^2}{\lambda\varepsilon} \operatorname{diam}(\mathcal{M})^2\big)$, where R is the maximum norm of the features.*

5.3. Computation of the Max-Min Oracle

The max-min oracle (14) corresponds to a concave-convex bilinear saddle-point problem. We use a standard alternating procedure of ascent and descent steps on the variables μ and ν, respectively. Consider a strongly concave differentiable entropy $H : \mathcal{C} \supset \mathcal{M} \to \mathbb{R}$ defined in a convex set $\mathcal{C}$ containing $\mathcal{M}$ such that $\nabla H(\mathcal{C}) = \mathbb{R}^k$ and $\lim_{\mu \in \partial\mathcal{C}} \|\nabla H(\mu)\| = +\infty$, where $\partial\mathcal{C}$ is the boundary of $\mathcal{C}$. Then, perform Mirror ascent/descent updates using $-H$ as the Mirror map. For instance, if $u = A^\top \nu + v$ is the gradient of (14) w.r.t μ, the update on μ takes the following form:

$$\arg\min_{\mu \in \mathcal{M}} -\eta \mu^\top u + D_{-H}(\mu, \mu^{(t)}), \qquad (15)$$

where $D_{-H}(\mu, \mu') = -H(\mu) + H(\mu') + \nabla H(\mu')^\top (\mu - \mu')$ is the Bregman divergence associated to the convex function $-H$. The resulting ascent/descent algorithm has a convergence rate of $O(t^{-1/2})$, which can be considerably improved to $O(t^{-1})$ with essentially no extra cost by performing four projections instead of two at each iteration. This corresponds to the extra-gradient strategy, called *Saddle Point Mirror Prox* (SP-MP) when using a Mirror map and is detailed in Algorithm 2.

Projection for factor graphs (Example 2.1). The entropy in $\mathcal{M}$ defined by the factor graph (Wainwright & Jordan, 2008) can be written explicitly in terms of the entropies of each part $\alpha \subset [M]$ if the factor graph has a junction tree structure (Koller & Friedman, 2009). For instance, in the case of a sequence of length M with unary and adjacent pairwise factors, we have $H(\mu) = \sum_{m=1}^{M-1} H_S(\mu_{m,m+1}) - \sum_{m=1}^{M} H_S(\mu_m)$, where H_S is the Shannon entropy and $\mu_m, \mu_{m,m+1}$ are the unary and pair-wise marginals, respectively. The projection (15) corresponds to marginal inference in CRFs and can be computed using the sum-product algorithm in time $O(MR^2)$. In this case, the complexity of the projection-oracle is the same the one of the max-oracle for SSVMs.

Projection for ranking and matching (Example 2.2). In this setting, the projection using the entropy in $\mathcal{M}$ is known to be #P-complete (Valiant, 1979). Thus, CRFs are essentially intractable in this setting (Petterson et al., 2009). If instead we use the entropy $H(P) = -\sum_{i,j=1}^{M} P_{ij} \log P_{ij}$ defined over the marginals $P \in \mathcal{M}$, the projection can be computed up to precision δ in $O(M^2/\delta)$ iterations using the Sinkhorn-Knopp algorithm (Cuturi, 2013). This can be potentially much cheaper than the max-oracle of SSVMs, which has a cubic dependence in M. The projection with respect to the Euclidean norm has similar complexity but implementation is more involved (Blondel et al., 2017).

Warm-starting the oracles. On the one hand, Algorithm 1 is guaranteed to converge as long as the error incurred in the oracle $\mathcal{O}$ decreases sublinearly with the number of global iterations as $\varepsilon_t \propto n/(t+n)$ (see Appendix F.1). On the other hand, Algorithm 2 can be naturally warm-started because it is an *any-time* algorithm as the step-size η does not depend on the current iteration or a finite horizon. Hence, we are in a setting where a warm-start strategy can be advantageous. More specifically, at every iteration t, we save the pairs $(\mu_i^\star, \nu_i^\star) \in \mathcal{O}(g_{w^{(t)}}(x_i))$ and the next time we revisit the i-th training example we initialize Algorithm 2 with this pair. Even though the formal demonstration of the effectiveness of the strategy is technically hard, we provide a strong experimental argument showing that a constant number of Algorithm 2 iterations are enough to match the allowed error ε_t.

Using the kernel trick. An extension to infinite-dimensional RKHS is straightforward to derive as Algorithm 1 is dual-based. In this case, the algorithm keeps track of the dual variables μ_i for $i = 1, \dots, n$.

Connection to stochastic subgradient algorithms. It is known that (generalized) conditional gradient methods in the dual are formally equivalent to subgradient methods in the primal (Bach, 2015). Indeed, note that w_s in Algorithm 1 is a subgradient of the scaled surrogate loss $S(g_w(x_i), y_i)/\lambda n$. However, the dual-based analysis we provide in this paper allows us to derive guarantees on the expected duality gap and a line-search strategy, which we leave for future work. Viewing Algorithm 1 as a subgradient method is useful when learning the data representation with a neural network. More specifically, both Algorithm 1 and Algorithm 2 remain essentially unchanged by applying the chain rule in the update of w.

5.4. Statistical Analysis of the Algorithm

Finally, the following Theorem 5.2 shows that the full algorithm without the warm-start strategy achieves the same statistical error as the regularized ERM estimator (3) after at most $O(n\sqrt{n})$ projections oracle calls.

Theorem 5.2 (Generalization bound of the algorithm)**.** *Assume the setting of Theorem 3.4. Let $g_{n,T}$ be the T-th iteration of Algorithm 1 applied to problem* (3)*, where each iteration is computed with $K = O(\sqrt{n})$ iterations of Algorithm 2. Then, after $T = O(n)$ iterations, $g_{n,T}$ satisfies the bound of Theorem 3.4 with probability $1 - \delta$.*

As we will show in the next section, in practice a constant number of iterations of Algorithm 2 are enough when using the warm-start strategy. Hence, the total number of required projection-oracles is $O(n)$.

6. Experiments

We perform a comparative experimental analysis for different tasks between $\mathrm{M}^4\mathrm{Ns}$, $\mathrm{M}^3\mathrm{Ns}$ and CRFs optimized with Generalized BCFW + SP-MP (Algorithm 1 + Algorithm 2), BCFW (Lacoste-Julien et al., 2013) and SDCA (Shalev-Shwartz & Zhang, 2013), respectively. All methods are run with our own implementation [2]. We use datasets of the UCI machine learning repository (Asuncion & Newman, 2007) for multi-class classification and ordinal regression, the OCR dataset from Taskar et al. (2004) for sequence prediction and the ranking datasets used by Korba et al. (2018). We use 14 random splits of the dataset into 60% for training, 20% for validation and 20% for testing. We choose the regularization parameter λ in $\{2^{-j}\}_{j=1}^{10}$ using the validation set and show the average test loss on the test sets in Table 1 of the model with the best λ. We use a Gaussian kernel and perform 50 passes on the data and set the number of iterations of Algorithm 2 to 20 and 10 times the length of the sequence for sequence prediction. The results are in Table 1. We perform better than $\mathrm{M}^3\mathrm{Ns}$ in most of the datasets for multi-class classification, ordinal regression and ranking, while we obtain similar results in the sequence dataset with the three methods.

[2] Code in `https://github.com/alexnowakvila/maxminloss`

Effect of warm-start. We perform an experiment tracking the test loss and the average error in the max-min oracle for different iterations of Algorithm 2 with and without warm-starting. The experiments are done in two datasets for ordinal regression and they are shown in Table 2. We observe that both the test loss and average oracle error are lower for the warm-start strategy. Moreover, when warm-starting the final test error barely changes when increasing the iterations past the 50 iteration threshold.

Task	Dataset	(d, n, k)	$\mathrm{M}^3\mathrm{N}$	CRF	$\mathrm{M}^4\mathrm{N}$
MC	segment	(19, 2310, 7)	6.64%	6.43%	**6.09%**
	iris	(4, 150, 3)	3.33%	3.08%	3.33%
	wine	(13, 178, 3)	2.56%	2.14%	**2.35%**
	vehicle	(18, 846, 4)	24.6%	25.1%	**24%**
	satimage	(36, 4435, 6)	12.2%	11.5%	**11.9%**
	letter	(16, 15000, 26)	14.6%	13.2%	**13.5%**
	mfeat	(216, 2000, 10)	**3.94%**	4.35%	3.96%
ORD	wisconsin	(32, 193, 5)	**1.24**	1.26	1.26
	stocks	(9, 949, 5)	0.167	0.168	**0.160**
	machine	(6, 208, 10)	0.634	0.628	**0.628**
	abalone	(10, 4176, 10)	0.520	0.526	0.520
	auto	(7, 391, 10)	0.589	0.621	**0.585**
		(d, n, M)			
SEQ	ocr	(128, 6877, 26)	16.2%	16.3%	16.2%
RNK	glass	(9, 214, 6)	17.7%	-	**17.4%**
	bodyfat	(7, 252, 7)	79.6%	-	79.6%
	wine	(13, 178, 3)	5.06%	-	**4.34%**
	vowel	(10, 528, 11)	33.7%	-	**32.2%**
	vehicle	(18, 846, 4)	**14.8%**	-	15.0%

Table 1. Average test losses on the 14 splits for multi-class classification (first), ordinal regression (second), sequence prediction (third) and ranking (forth). We show in percentage the losses for multi-class, sequence prediction and ranking since they are between zero and one. We show in bold the lowest test loss between the direct classifiers $\mathrm{M}^3\mathrm{N}$ and $\mathrm{M}^4\mathrm{N}$.

7. Conclusion

In this paper, we introduced max-min margin Markov networks ($\mathrm{M}^4\mathrm{Ns}$), a method for general structured prediction, that has the same algorithmic and theoretical properties as the regular binary SVM, that is, quantitative convergence bounds through a linear comparison inequality, as well as

Dataset	W-S	$K = 10$	$K = 30$	$K = 50$	$K = 100$
machine	yes	0.42 / 0.57	0.41 / 0.43	0.41 / 0.22	0.41 / 0.13
	no	0.50 / 4.41	0.50 / 2.74	0.44 / 1.25	0.42 / 0.63
auto	yes	0.56 / 1.55	0.55 / 1.29	0.51 / 0.81	0.50 / 0.44
	no	0.61 / 2.66	0.57 / 1.79	0.53 / 0.89	0.51 / 0.47

Table 2. We show the (final ordinal test loss / average oracle error at the last epoch) for M^4Ns trained with 100 passes on data with different iterations of Algorithm 2 with and without warm-starting.

efficient optimization algorithms. Our experiments show its performance on classical structured prediction problems when using RKHS hypothesis spaces. It would be interesting to extend the analysis of the proposed algorithm by rigorously proving the linear dependence in the number of samples when using the warm-start strategy and incorporating a line-search strategy.

Acknowledgements

The authors would like to thank Mathieu Blondel, Martin Arjovsky and Simon Lacoste-Julien for useful discussions. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support the European Research Council (grant SEQUOIA 724063). ANV received support from "La Caixa" Foundation.

References

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

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

Asuncion, A. and Newman, D. UCI machine learning repository, 2007.

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Bach, F. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1): 115–129, 2015.

BakIr, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., and Vishwanathan, S. *Predicting Structured Data*. MIT press, 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.

Belanger, D. and McCallum, A. Structured prediction energy networks. In *International Conference on Machine Learning*, pp. 983–992, 2016.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. *arXiv preprint arXiv:1710.06276*, 2017.

Blondel, M., Martins, A. F., and Niculae, V. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Burkard, R., Dell'Amico, M., and Martello, S. *Assignment problems, revised reprint*, volume 106. Siam, 2012.

Caetano, T. S., McAuley, J. J., Cheng, L., Le, Q. V., and Smola, A. J. Learning graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (6):1048–1058, 2009.

Ciliberto, C., Rosasco, L., and Rudi, A. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pp. 4412–4420, 2016.

Ciliberto, C., Bach, F., and Rudi, A. Localized structured prediction. In *Advances in Neural Information Processing Systems*, pp. 7299–7309, 2019.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

De Loera, J. A., Hemmecke, R., and Köppe, M. *Algebraic and geometric ideas in the theory of discrete optimization*. SIAM, 2012.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Duchi, J., Khosravi, K., and Ruan, F. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems*, pp. 559–567, 2016.

Fathony, R., Asif, K., Liu, A., Bashiri, M. A., Xing, W., Behpour, S., Zhang, X., and Ziebart, B. D. Consistent robust adversarial prediction for general multiclass classification. *arXiv preprint arXiv:1812.07526*, 2018a.

Fathony, R., Behpour, S., Zhang, X., and Ziebart, B. Efficient and consistent adversarial bipartite matching. In *International Conference on Machine Learning*, pp. 1457–1466, 2018b.

Fathony, R., Rezaei, A., Bashiri, M. A., Zhang, X., and Ziebart, B. Distributionally robust graphical models. In *Advances in Neural Information Processing Systems*, pp. 8353–8364, 2018c.

Finocchiaro, J., Frongillo, R., and Waggoner, B. An embedding framework for consistent polyhedral surrogates. *arXiv preprint arXiv:1907.07330*, 2019.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, pp. 427–435, 2013.

Joachims, T. Training linear SVMs in linear time. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 217–226. ACM, 2006.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Korba, A., Garcia, A., and d'Alché Buc, F. A structured prediction approach for label ranking. In *Advances in Neural Information Processing Systems*, pp. 8994–9004, 2018.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 53–61, 2013.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, Y., Lin, Y., and Wahba, G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Liu, Y. Fisher consistency of multicategory support vector machines. In *Artificial Intelligence and Statistics*, pp. 291–298, 2007.

Michelot, C. A finite algorithm for finding the projection of a point onto the canonical simplex of n. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

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

Nemirovski, A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nowak-Vila, A., Bach, F., and Rudi, A. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019a.

Nowak-Vila, 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, 2019b.

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

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.

Petterson, J., Yu, J., McAuley, J. J., and Caetano, T. S. Exponential family graph matching and ranking. In *Advances in Neural Information Processing Systems*, pp. 1455–1463, 2009.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Exponential convergence of testing error for stochastic gradient methods. *Proceedings of the Conference on Learning Theory*, 2018a.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Exponential convergence of testing error for stochastic gradient methods. In *Proceedings of the Conference On Learning Theory*, pp. 250–296, 2018b.

Ramaswamy, H. G. and Agarwal, S. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(1):397–441, 2016.

Ramaswamy, H. G., Agarwal, S., and Tewari, A. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, pp. 1475–1483, 2013.

Ryser, H. J. *Combinatorial mathematics*, volume 14. American Mathematical Soc., 1963.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Smith, N. A. Linguistic structure prediction. *Synthesis lectures on human language technologies*, 4(2):1–274, 2011.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pp. 1545–1552, 2009.

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, 2005.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484, 2005.

Valiant, L. G. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979.

Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.