Efficient Training of Structured SVMs via Soft Constraints

Ofer Meshi
TTI Chicago

Nathan Srebro
TTI Chicago

Tamir Hazan
University of Haifa, Israel

Abstract

Structured output prediction is a powerful framework for jointly predicting interdependent output labels. Learning the parameters of structured predictors is a central task in machine learning applications, however, training the model from data often becomes computationally expensive. Several methods have been proposed to exploit the model structure, or decomposition, in order to obtain efficient training algorithms. In particular, methods based on linear programming relaxation, or dual decomposition, decompose the prediction task into multiple simpler prediction tasks and enforce agreement between overlapping predictions. In this work we observe that relaxing these agreement constraints and replacing them with soft constraints yields a much easier optimization problem. Based on this insight we propose an alternative training objective, analyze its theoretical properties, and derive an algorithm for its optimization. Our method, based on the Frank-Wolfe algorithm, achieves significant speedups over existing state-of-the-art methods without hurting prediction accuracy.

1 Introduction

Structured output prediction is an effective framework to reason about real-life problems, since it provides the means to map data instances to meaningful labels. By accounting for the correlations between labels, prediction accuracy can be improved in a wide range of applications. In the setting of supervised learning, the parameters of structured predictors are learned from training data. In particular, in the *Structured SVM* framework, the learning objective is formulated as regularized structured hinge loss minimization [2, 11, 12]. Despite the convexity of the structured SVM objective function, finding the optimal parameters of these models is computationally expensive, since it requires comparing training labels to predicted labels. For some specific models (e.g., tree-structured graphs, matchings, and submodular scores), exact prediction can be done efficiently, however, in general computing the objective or the gradient exactly is intractable. Therefore, one usually resorts to approximate inference algorithms [cf. 11, 6, 3]. One family of such approximations that has proved quite successful is based on *linear programming (LP) relaxation*, or *dual decomposition*. In this approach the intractable prediction task is decomposed into simpler prediction tasks, and consistency among overlapping predictions is enforced. Although tractable, these algorithms are often very expensive when used as a subroutine within the learning algorithm.

In this work we propose a novel training algorithm for structured SVMs. Our main insight is that the consistency constraints between overlapping predictions significantly complicate the training objective. Instead, we suggest to enforce these constraints in a soft manner, by introducing a penalty term that accounts for constraint violation.

2 Learning with Soft Constraints

We begin by reviewing the framework of structured SVMs [11, 12]. Consider the task of jointly predicting a set of discrete labels $y = (y_1, \ldots, y_n)$, given a data-instance vector x. Our goal is to learn the parameters $w \in \mathbb{R}^d$ of the linear prediction rule $y(x; w) = \operatorname{argmax}_y w^\top \phi(x, y)$, where $\phi(x, y) \in \mathbb{R}^d$ maps data-label pairs to a feature vector. In this supervised learning setting, w is learned from training data $\{(x^{(m)}, y^{(m)})\}_{m=1}^M$ by reducing the regularized structured hinge loss:

$$\min_w \ \frac{\lambda}{2}\|w\|^2 + \frac{1}{M}\sum_m \max_y \left[w^\top \phi(x^{(m)}, y) + \Delta(y, y^{(m)})\right] - w^\top \phi(x^{(m)}, y^{(m)}) \,, \tag{1}$$

where $\Delta(y, y^{(m)})$ is a task-loss that measures the cost of predicting y when the true output is $y^{(m)}$.

Since the space of possible outputs y is exponential in the number of output variables n, the maximization over outputs will generally not be possible by naive enumeration. In many applications, it is common to assume that the score function *decomposes* into simpler score functions with respect to subsets of indexes $\alpha \subset \{1, ..., n\}$, namely $w^\top \phi(x, y) = \sum_\alpha w_\alpha^\top \phi_\alpha(x, y_\alpha)$. Such decomposed scoring function considers only (possibly overlapping) subsets of output variables $y_\alpha = \{y_i\}_{i \in \alpha}$. Assuming that the task loss Δ decomposes in a similar manner, one can write the maximization problems for prediction and training in the form $\max_y \sum_\alpha \theta_\alpha(y_\alpha)$. Since this problem is generally NP-hard, some kind of approximation will be necessary.

One approach to relax the hard learning problem of Eq. (1) replaces the intractable loss for each training example by its (relaxed) dual [10, 8, 5]. Dividing the subsets α into singletons, denoted by $i = 1, ..., n$, and high-order subsets c (also called "factors"), the resulting training problem is:

$$G: \quad \min_{w,\delta} g(w, \delta) \ := \frac{\lambda}{2}\|w\|^2 + \frac{1}{M}\sum_m \sum_\alpha \max_{y_\alpha} \hat{\theta}_\alpha^{(m)}(y_\alpha; w, \delta) \,, \tag{2}$$

$$\text{where} \quad \hat{\theta}_\alpha^{(m)}(y_\alpha; w, \delta) = \begin{cases} \theta_i^{(m)}(y_i; w) + \sum_{c: i \in c} \delta_{ci}^{(m)}(y_i) & \alpha \in \{i\} \\ \theta_c^{(m)}(y_c; w) - \sum_{i: i \in c} \delta_{ci}^{(m)}(y_i) & \alpha \in \{c\} \end{cases}$$

$$\theta_\alpha^{(m)}(y_\alpha; w) = w_\alpha^\top \left(\phi_\alpha(x^{(m)}, y_\alpha) - \phi_\alpha(x^{(m)}, y_\alpha^{(m)})\right) + \Delta(y_\alpha, y_\alpha^{(m)}) \quad \forall \alpha \in \{c, i\}$$

In this formulation $\hat{\theta}_\alpha$ is a *reparameterization* of the model scores θ_α, where the variables $\delta_{ci}^{(m)}(y_i)$ serve to encourage agreement between factor and variable maximizers. For each sample m there exists such variable for a factor c, variable $i \in c$, and assignment y_i [9].

It is well known [cf. 11] that the dual problem associated with problem G takes the form:

$$F: \quad \max_{\mu \in \mathcal{M}_\mathcal{L}^\times} f(\mu) \ := \ \mu^\top \ell - \frac{\lambda}{2}\|\Psi\mu\|^2 \,, \tag{3}$$

where μ is the set of dual variables, $\Psi_{m,\alpha,y_\alpha} = \frac{1}{\lambda M}\left(\phi_\alpha(x^{(m)}, y_\alpha^{(m)}) - \phi_\alpha(x^{(m)}, y_\alpha)\right)$ is a column vector in $\mathbb{R}^d$, and $\ell_{m,\alpha,y_\alpha} = \frac{1}{M}\Delta(y_\alpha, y_\alpha^{(m)})$ is a scalar. In this formulation the dual variables $\mu_\alpha^{(m)}(y_\alpha)$ can be interpreted as the marginal probability of the subset assignment y_α [see, e.g., 11, 13]. Furthermore, the constraint set $\mathcal{M}_\mathcal{L}^\times$, known as the *local marginal polytope*, is a product domain which enforces agreement between local marginals within each training example:

$$\mathcal{M}_\mathcal{L}^{(m)} = \left\{ \mu^{(m)} \geq 0 : \begin{array}{ll} \mu_c^{(m)}(y_i) = \mu_i^{(m)}(y_i) & \forall c, i \in c, y_i \\ \sum_{y_\alpha} \mu_\alpha^{(m)}(y_\alpha) = 1 & \forall \alpha = \{c, i\} \end{array} \right\} \,, \tag{4}$$

where $\mu_c^{(m)}(y_i) = \sum_{y_{c \setminus i}} \mu_c^{(m)}(y_c)$ is the marginal of the variable assignment y_i taken from the factor marginal $\mu_c^{(m)}$.

Focusing on the dual problem F, our main insight is that part of the difficulty in optimizing this objective stems from the fact that the marginals associated with a training example $\mu^{(m)}$ are coupled together through the agreement constraints in $\mathcal{M}_\mathcal{L}$. Therefore, we next alleviate this complication by relaxing these constraints. In particular, applying the *penalty method* [1] to the agreement constraints

Algorithm 1 Block-coordinate Frank-Wolfe for soft structured SVM

1: Initialize: $w = 0$, $\delta = 0$, $\mu_\alpha^{(m)}(y_\alpha) = \mathbb{1}\{y_\alpha = y_\alpha^{(m)}\}$ for all m, α, y_α
2: **while** not converged **do**
3: Randomly sample a block (m, α)
4: Let $\hat{\theta}_\alpha^{(m)}$ be a reparameterization as in Eq. (2)
5: Let $y_\alpha^* = \operatorname{argmax}_{y_\alpha} \hat{\theta}_\alpha^{(m)}(y_\alpha)$, and let s_α be the corresponding indicator vector
6: Let $\gamma = \begin{cases} \frac{\hat{\theta}_i^\top \left(s_i - \mu_i^{(m)}\right)}{\lambda \|\Psi_{m,i}(s_i - \mu_i^{(m)})\|^2 + \rho N_i \|s_i - \mu_i^{(m)}\|^2} & \alpha \in \{i\} \\ \frac{\hat{\theta}_c^\top \left(s_c - \mu_c^{(m)}\right)}{\lambda \|\Psi_{m,c}(s_c - \mu_c^{(m)})\|^2 - \rho \sum_{i: i \in c} \|A_{ci}(s_c - \mu_c^{(m)})\|^2} & \alpha \in \{c\} \end{cases}$
and clip to $[0, 1]$
7: Update $\mu_\alpha^{(m)} \leftarrow (1 - \gamma)\mu_\alpha^{(m)} + \gamma s_\alpha$
8: Update $w = \Psi\mu$ and $\delta = A\mu$
9: **end while**

in F (see Eq. (3)) means replacing the constraint $\mu_c^{(m)}(y_i) = \mu_i^{(m)}(y_i)$ with a penalty term of the form $\frac{1}{2\rho}\left(\mu_c^{(m)}(y_i) - \mu_i^{(m)}(y_i)\right)^2$ for all $m, c, i \in c, y_i$. Clearly, as $\rho \to 0$ the penalty increases and the solution of the unconstrained problem converges to a solution of the original constrained problem. As before, we can write the resulting problem concisely as:

$$F_\rho : \quad \max_{\mu \in \mathcal{S}^\times} f_\rho(\mu) \;\; := \;\; \mu^\top \ell - \frac{\lambda}{2}\|\Psi\mu\|^2 - \frac{\rho}{2}\|A\mu\|^2 \,, \tag{5}$$

where $(A\mu)_{m,c,i,y_i} = \frac{1}{\rho M}\left(\mu_c^{(m)}(y_i) - \mu_i^{(m)}(y_i)\right)$, and $\mathcal{S}^\times$ is a product domain with per-factor simplex constraints (see Eq. (4)). Intuitively, the additional penalty term serves to "smooth" the boundaries of the local marginal polytope constraints in F, while keeping the feasible domain $\mu \in \mathcal{M}_\mathcal{L}^\times$ unchanged.

The dual problem of F_ρ turns out to be:

$$G_\rho : \quad \min_{w,\delta} g_\rho(w, \delta) \;\; := \;\; g(w, \delta) + \frac{\rho}{2}\|\delta\|^2 \tag{6}$$

This problem is the same as the primal G, except for the additional L_2 term for δ. For this problem we have the dual-primal mapping: $w(\mu) = \Psi\mu$ and $\delta(\mu) = A\mu$, which we will later use.

We next justify the use of the softly constrained objective by bounding its difference from the constrained one. In the next theorem we use the following notation. Let $\|w\|_2 \le B$ for all w, let $\|\phi(x, y)\|_2 \le R$ for all (x, y), and let $\Delta(y, y') \le L$ for all (y, y'). Therefore, $\|\theta\|_\infty \le 2BR + L$. In addition, let $|Y_i|$ be the number of states of output variable i, and let $Y_{\max} = \max_i |Y_i|$ denote the maximum over all variables. Finally, let q be the maximal number of factors (including singletons) in any instance.

Theorem 2.1. *Let g_ρ^* be the optimal value of G_ρ, and let g^* be the optimal value of G. Then $g_\rho^* - \frac{\rho}{2}h \le g^* \le g_\rho^*$, where $h = M(8Y_{\max}q(BR + L))^2$.*

This theorem shows that despite using soft constraints, we still have guarantees w.r.t. the original constrained objective. At first glance, the bound in Theorem 2.1 may seem quite loose due to the linear dependence on the number of samples M. However, recall that the difference between g_ρ and g is the L_2 regularization term for δ. Unlike the weight vector w, the number of agreement variables (length of δ) grows with M, and therefore its norm also grows linearly with M. We will later see that this is not a serious limitation of our approach, since in practice the difference between g_ρ and g is not so large, even for relatively high values of ρ.

Given Theorem 2.1, we can obtain a similar result for a near-optimal solution.

Theorem 2.2. *Let μ_ρ^ϵ be a dual solution to F_ρ for which the duality gap is bounded: $D_\rho(\mu_\rho^\epsilon) = g_\rho(w(\mu_\rho^\epsilon), \delta(\mu_\rho^\epsilon)) - f_\rho(\mu_\rho^\epsilon) \le \epsilon$. Then $w(\mu_\rho^\epsilon)$ is $\left(\epsilon + \frac{\rho}{2}h\right)$-optimal for G.*

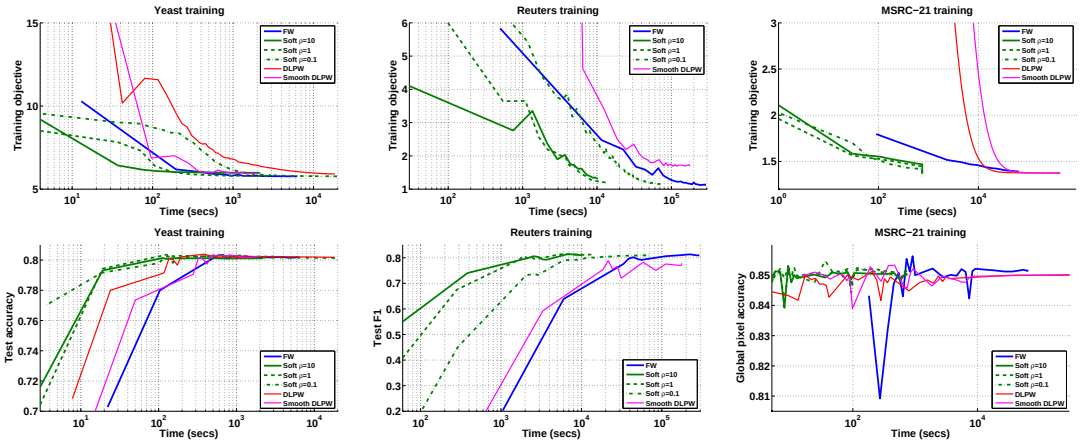

Figure 1: Comparison of training objective (top row) and performance measure (bottom row) as a function of runtime for the Yeast (left column), Reuters (middle column) and MSRC-21 (right column) datasets. In the top row we show the hard-constrained objective value $g(w(\mu), \delta(\mu))$.

3 Algorithm

In this section we propose an algorithm for optimizing our alternative dual problem F_ρ (Eq. (5)). In this work we use the *block-coordinate Frank-Wolfe (BCFW)* algorithm, which was introduced by [7]. Applying BCFW to our dual objective f_ρ yields Algorithm 1, where $N_i = |\{c : i \in c\}|$ is the number of factors containing variable i, $\Psi_{m,\alpha}$ is the part in Ψ corresponding to sample m and factor α, and A_{ci} marginalizes μ_c to values of variable $i \in c$.

Algorithm 1 has several compelling properties. First, the update of primal variables (w, δ) in line 8 is computationally cheap since a change in $\mu_\alpha^{(m)}$ affects only w_α and δ variables pertaining to the chosen α (i.e., its neighbors in the factor graph). Second, the algorithm employs simple per-factor maximization oracles. Third, since the *optimal* step-size is computed analytically, there are no hyperparameters to tune. Finally, we use the duality gap as a sound stopping criterion.

Building on the analysis in [7], we can bound the runtime of Algorithm 1 as follows.

Theorem 3.1. *Algorithm 1 obtains an ϵ-optimal solution μ with ϵ-duality-gap $D_\rho(\mu) \leq \epsilon$ after at most $O\left(\frac{q^2}{\epsilon}\left(\frac{1}{\lambda}+\frac{1}{\rho}\right)\right)$ iterations.*

In comparison, for the constrained objective f (Eq. (3)) the rate obtained in [7] is $O\left(\frac{1}{\lambda\epsilon}\right)$, which seems faster. However, each iteration requires calling a maximization oracle for a complete training example ($\mu^{(m)}$ block), while Algorithm 1 requires optimizing only over factor blocks $\mu_\alpha^{(m)}$, which can be much cheaper.

4 Experiments

In this section we compare our algorithm to other state-of-the-art baselines. In particular, we implement Algorithm 1, the BCFW algorithm [7], the DLPW algorithm [8], and a smooth version of DLPW [4]. Notice that DLPW is similar to our method in the sense that the updates are "local" and do not process complete training samples. We conduct experiments on two different domains: multi-label classification (Yeast, Reuters) and image segmentation (MSRC-21). In both cases the model consists of singleton and pairwise scores: $w^\top \phi(x, y) = \sum_i w_i^\top \phi_i(x, y_i) + \sum_{i,j} w_{ij}^\top \phi_{ij}(y_i, y_j)$.

In Figure 1 we observe that our method is upto *two orders of magnitude* faster than the baselines. Our algorithm has cheap local updates and uses the optimal step at each iteration, thereby achieving fast convergence. Moreover, we notice that our algorithm is able to quickly learn a model with high prediction accuracy. We also see that the performance of our method is rather insensitive to the choice of ρ, and even a large value $\rho = 10$ is sufficient to obtain good prediction accuracy.

References

[1] D. Boukari and A. V. Fiacco. Survey of penalty, exact-penalty and multiplier methods from 1968 to 1993. *Optimization*, 32(4):301–334, 1995.

[2] M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, 2002.

[3] T. Finley and T. Joachims. Training structural SVMs when exact inference is intractable. In *Proceedings of the 25th International Conference on Machine learning*, pages 304–311, 2008.

[4] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, pages 838–846. 2010.

[5] N. Komodakis. Efficient training for pairwise or higher order crfs via dual decomposition. In *CVPR*, CVPR '11, pages 1841–1848, 2011.

[6] A. Kulesza and F. Pereira. Structured learning with approximate inference. In *Advances in Neural Information Processing Systems 20*, pages 785–792. 2007.

[7] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61, 2013.

[8] O. Meshi, D. Sontag, T. Jaakkola, and A. Globerson. Learning efficiently with approximate inference via dual losses. In *ICML*, pages 783–790. ACM, 2010.

[9] D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In *Optimization for Machine Learning*, pages 219–254. MIT Press, 2011.

[10] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: a large margin approach. In *ICML*, pages 896–903. ACM, 2005.

[11] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*. MIT Press, 2003.

[12] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, pages 104–112, 2004.

[13] M. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.