Theoretical Computer Science 519 (2014) 88–102

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Accelerated training of max-margin Markov networks with kernels

Xinhua Zhang [a,*], Ankan Saha [b], S.V.N. Vishwanathan [c]

[a] Department of Computing Science, University of Alberta, Edmonton, AB T6G2E8, Canada
[b] Department of Computer Science, University of Chicago, Chicago, IL 60637, USA
[c] Department of Statistics and Computer Science, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Keywords:
Convex optimization
Max-margin models
Kernel method
Graphical models

ABSTRACT

Structured output prediction is an important machine learning problem both in theory and practice, and the max-margin Markov network (M^{3}N) is an effective approach. All state-of-the-art algorithms for optimizing M^{3}N objectives take at least $O(1/\epsilon)$ number of iterations to find an ϵ accurate solution. Nesterov [1] broke this barrier by proposing an excessive gap reduction technique (EGR) which converges in $O(1/\sqrt{\epsilon})$ iterations. However, it is restricted to Euclidean projections which consequently requires an intractable amount of computation for each iteration when applied to solve M^{3}N. In this paper, we show that by extending EGR to Bregman projection, this faster rate of convergence can be retained, and more importantly, the updates can be performed efficiently by exploiting graphical model factorization. Further, we design a kernelized procedure which allows all computations per iteration to be performed at the same cost as the state-of-the-art approaches.

1. Introduction

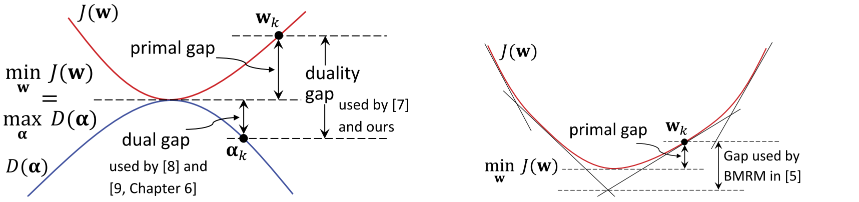
In the supervised learning setting, one is given a training set of labeled data points and the aim is to learn a function which predicts labels on unseen data points. Sometimes the label space has a rich internal structure which characterizes the combinatorial or recursive inter-dependencies of the application domain. It is widely believed that capturing these dependencies is critical for effectively learning with *structured output*. Examples of such problems include sequence labeling, context free grammar parsing, and word alignment. However, parameter estimation is generally hard even for simple linear models, because the size of the label space is potentially exponentially large (see e.g. [2]). Therefore it is crucial to exploit the underlying conditional independence assumptions for the sake of computational tractability. This is often done by defining a graphical model on the output space, and exploiting the underlying graphical model factorization to perform efficient computations.

Research in structured prediction can broadly be categorized into two tracks: Optimizing conditional likelihood in an exponential family results in conditional random fields (CRFs) [3], and a maximum margin approach leads to max-margin Markov networks (M^{3}Ns) [4]. Unsurprisingly, these two approaches share many commonalities: First, they both minimize a regularized risk with a square norm regularizer. Second, they assume that there is a joint feature map $\boldsymbol{\phi}$ which maps $(\mathbf{x},\mathbf{y})$ to a feature vector in $\mathbb{R}^p$.[1] Third, they assume a label loss $\ell(\mathbf{y},\mathbf{y}^i;\mathbf{x}^i)$ which quantifies the loss of predicting label $\mathbf{y}$ when the correct label of input $\mathbf{x}^i$ is $\mathbf{y}^i$. Finally, they assume that the space of labels $\mathcal{Y}$ is endowed with a graphical model

[*] Corresponding author.
E-mail addresses: xinhua2@cs.ualberta.ca (X. Zhang), ankans@cs.uchicago.edu (A. Saha), vishy@stat.purdue.edu (S.V.N. Vishwanathan).

[1] We discuss kernels and associated feature maps into a Reproducing Kernel Hilbert Space (RKHS) in Section 4.3.

0304-3975/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.tcs.2013.09.026



(a) Primal gap, dual gap, and duality gap (b) BMRM gap (and similarly for SVM-Struct)

Fig. 1. Illustration of stopping criterion monitored by various algorithms; convergence rates are stated with respect to these stopping criterion. $D(\boldsymbol{\alpha})$ is the Lagrange dual of $J(\mathbf{w})$, and $\min_{\mathbf{w}} J(\mathbf{w}) = \max_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha})$. Neither the primal gap nor the dual gap is actually measurable in practice since $\min_{\mathbf{w}} J(\mathbf{w})$ (and $\max_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha})$) is unknown. BMRM (right) therefore uses a measurable upper bound of the primal gap. SVM-Struct monitors constraint violation, which can be also be translated to an upper bound on the primal gap.

Table 1
Comparison of specialized optimization algorithms for training structured prediction models. Primal–dual methods maintain estimation sequences in both primal and dual spaces. Details of the oracle will be discussed in Section 5. The convergence rate highlights the dependence on both ϵ and some "constants" that are often hidden in the O notation: n, λ, and the size of the label space $|\mathcal{Y}|$. The convergence rate of SMO on M^{3}N is derived from [10, Corollary 17], noting the dual problem (26) is so-called pairable. It enjoys linear convergence $O(\log\frac{1}{\epsilon})$ when the dual objective is positive definite (pd), and $O(\frac{1}{\epsilon})$ when it is positive semi-definite (psd). The term G in the convergence rate denotes the maximum L_2 norm of the features vectors $\boldsymbol{\phi}(\mathbf{x}^i, \mathbf{y})$. The convergence rate of extragradient depends on λ in an indirect way.

Optimization algorithm	Primal/dual	Type of gap	Oracle for M^{3}N	Convergence rate	
				CRF	M^{3}N
BMRM [5]	primal	primal gap	max	$O(\frac{1}{\lambda}\log\frac{1}{\epsilon})$	$O(\frac{G^2}{\lambda\epsilon})$
SVM-Struct [6]	primal–dual	constraint violation	max	n/a	$O(\frac{G^2}{\lambda\epsilon})$
Extragradient [7]	primal–dual	duality gap	exp	n/a	$O(\frac{\log\|\mathcal{Y}\|}{\epsilon})$
Exponentiated gradient [8]	dual	dual gap	exp	$O(\frac{1}{\lambda}\log\frac{1}{\epsilon})$	$O(\frac{G^2\log\|\mathcal{Y}\|}{\lambda\epsilon})$
SMO [9, Chapter 6]	dual	dual gap	max	n/a	psd: $O(n\|\mathcal{Y}\|\frac{1}{\lambda\epsilon})$ pd: $O(n\|\mathcal{Y}\|\log\frac{1}{\epsilon})$
Our algorithm	primal–dual	duality gap	exp	n/a	$O(G\sqrt{\frac{\log\|\mathcal{Y}\|}{\lambda\epsilon}})$

structure and that $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})$ and $\ell(\mathbf{y}, \mathbf{y}^i; \mathbf{x}^i)$ factorize according to the cliques of this graphical model. The main difference is in the loss function employed. CRFs minimize the L_2-regularized logistic loss:

$$J(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\log\sum_{\mathbf{y}\in\mathcal{Y}}\exp\bigl(\ell\bigl(\mathbf{y}, \mathbf{y}^i; \mathbf{x}^i\bigr) - \bigl\langle\mathbf{w}, \boldsymbol{\phi}\bigl(\mathbf{x}^i, \mathbf{y}^i\bigr) - \boldsymbol{\phi}\bigl(\mathbf{x}^i, \mathbf{y}\bigr)\bigr\rangle\bigr), \tag{1}$$

where all log in this paper stands for natural basis. In contrast, the M^{3}Ns minimize the L_2-regularized hinge loss

$$J(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\max_{\mathbf{y}\in\mathcal{Y}}\bigl\{\ell\bigl(\mathbf{y}, \mathbf{y}^i; \mathbf{x}^i\bigr) - \bigl\langle\mathbf{w}, \boldsymbol{\phi}\bigl(\mathbf{x}^i, \mathbf{y}^i\bigr) - \boldsymbol{\phi}\bigl(\mathbf{x}^i, \mathbf{y}\bigr)\bigr\rangle\bigr\}. \tag{2}$$

A large body of literature exists on efficient algorithms for minimizing the above objective functions. A summary of existing methods, and their convergence rates (iterations needed to find an ϵ accurate solution) can be found in Table 1. The ϵ accuracy of a solution can be measured in many different ways and different algorithms employ different but somewhat related stopping criterion (see Fig. 1). Some produce iterates $\mathbf{w}_k$ in the primal space and bound the *primal gap* $J(\mathbf{w}_k) - \min_{\mathbf{w}} J(\mathbf{w})$. Some solve the dual problem $D(\boldsymbol{\alpha})$ with iterates $\boldsymbol{\alpha}_k$ and bound the *dual gap* $\max_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) - D(\boldsymbol{\alpha}_k)$. Some bound the *duality gap* $J(\mathbf{w}_k) - D(\boldsymbol{\alpha}_k)$, and still others bound $J(\mathbf{w}_k) - \min_{\mathbf{w}} J_k(\mathbf{w})$ where J_k is a uniform lower bound of J. This must be borne in mind when interpreting the convergence rates in Table 1.

Since (1) is a smooth convex objective, classical methods such as L-BFGS can directly be applied [11]. Specialized solvers also exist. For instance a primal algorithm based on bundle methods was proposed by [5], while a dual algorithm for the same problem was proposed by [8]. Both algorithms converge at $O(\frac{1}{\lambda}\log(1/\epsilon))$ rates to an ϵ accurate solution, and, remarkably, their convergence rates are independent of n (the number of data points), and $|\mathcal{Y}|$ (the size of the label space). It is widely believed in optimization (see e.g. Section 9.3 of [12]) that unconstrained smooth strongly convex objective functions can be minimized in $O(\log(1/\epsilon))$ iterations, and these specialized optimizers also achieve this rate. Although interior point methods can converge in quadratic rates $\log(\log(1/\epsilon))$ which are even faster than $O(\log(1/\epsilon))$, its cost per step is prohibitively high.