Large-Margin Structured Prediction via Linear Programming

Zhuoran Wang
Department of Computer Science
University College London
London, WC1E 6BT
United Kingdom

John Shawe-Taylor
Department of Computer Science
University College London
London, WC1E 6BT
United Kingdom

Abstract

This paper presents a novel learning algorithm for structured classification, where the task is to predict multiple and interacting labels (multilabel) for an input object. The problem of finding a large-margin separation between correct multilabels and incorrect ones is formulated as a linear program. Instead of explicitly writing out the entire problem with an exponentially large constraint set, the linear program is solved iteratively via column generation. In this case, the process of generating most violated constraints is equivalent to searching for highest-scored misclassified incorrect multilabels, which can be easily achieved by decoding the structure based on current estimations. In addition, we also explore the integration of column generation and an extragradient method for linear programming to gain further efficiency. The proposed method has the advantages that it can handle arbitrary structures and larger-scale problems. Experimental results on part-of-speech tagging and statistical machine translation tasks are reported, demonstrating the competitiveness of our approach.

1 Introduction

Structured classification is to predict multiple and interacting labels, called multilabels, for a given input object. This kind of problem frequently arises in text, speech and image processing as well as bioinformatics, with examples including sequence labeling, parsing, bipartite matching, etc.

Recent advances in machine learning have applied maximum margin learning techniques to many of the above problems by optimizing a support vector machine (SVM)-style objective function over structured outputs. In comparison with binary classification tasks, a major issue complicating the structured case is that an exponential number of potential incorrect multilabels, i.e. negative examples, exist for every training point. Directly applying an SVM formulation to such problems will yield a quadratic programming (QP) optimization with exponentially many constraints, correspondingly exponentially many variables in the dual form. A solution to this problem was proposed by Altun et al. (2003) and Tsochantaridis et al. (2005) based on the working set method. Whilst Taskar et al. (2003) introduced a novel algorithm for Markov networks, whose spirit is to reformulate the optimization problem into an equivalent problem of size polynomial in the number of cliques in the graphs by decomposing its original dual variables into the so-called marginal dual variables. Alternatively, Bartlett et al. (2004) assumed the dual variables to be generated from a Gibbs distribution of a series of "mini-dual" variables each corresponding to a possible configuration for a clique in the graph, and estimated them based on exponentiated gradient updates. Besides SVM-like optimizations, Taskar et al. (2006) made a further improvement by using a convex-concave saddle-point formulation of the large margin structured estimation, which extended its application to a broader range of combinatorial models whose decoding processes are solvable via convex programming.

However, many real world applications may represent more complex situations. For example, in statistical machine translation (SMT), the word lattice is neither a Markov network nor any convex combinatorial model. Moreover, even an experimental SMT system will involve millions of lexicon entries (labels), for which training the weights usually requires hundreds of thousands of bilingual sentence pairs (training examples). The QP-based methods mentioned above are

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

still impractical for these kinds of tasks.

Thus, in this paper we present a linear programming (LP) approach for large margin learning, which can be applied to general structured prediction problems. In our framework, we still follow the large margin separation hyperplane formulation similar to SVM, but use an L_1-regularization in the objective function. Instead of explicitly writing out the entire problem, the column generation technique is employed to solve it. In this case, it is equivalent to incrementally adding violated constraints which can be detected via decoding the structure with the current estimation. In addition, column generation can be utilized with the extragradient method for LP (Korpelevich, 1976), which, we argue, can be expected to offer further efficiency by providing a proper starting point in every iteration. Compared to previous works, not only does the proposed approach allow arbitrary structures, it also scales better than the QP-based models. In addition, we prove that optimizing the objective function of the proposed model directly minimizes its generalization error bound.

2 Structured Classification

We consider the problem of learning a $\mathbf{w}$-parameterized function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{X}$ and $\mathcal{Y}$ are respectively the input and output spaces. Accordingly, the multilabel prediction $\hat{\mathbf{y}}$ for a given input object $\mathbf{x} \in \mathcal{X}$ is obtained as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}; \mathbf{w}) \tag{1}$$

We define the joint feature mapping $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, and assume that f is from the linear family, as:

$$f(\mathbf{x}, \mathbf{y}; \mathbf{w}) := \mathbf{w}^\top \phi(\mathbf{x}, \mathbf{y}) \tag{2}$$

Then based on a set of training examples $S := \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, m\}$, our goal is to seek the hyperplane $\mathbf{w}$ that separates the positive examples from the negative ones with maximum margin in the $\mathbb{R}^d$ feature space defined by ϕ. This type of problem is formulated by SVM and the extensions (Crammer and Singer, 2001) into the following optimization:

$$\begin{aligned} \max_{\mathbf{w},\gamma} \quad & \gamma \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq \gamma, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ \forall i; \\ & \|\mathbf{w}\|_2 = 1. \end{aligned} \tag{3}$$

where $\Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) = \phi(\mathbf{x}_i, \mathbf{y}_i) - \phi(\mathbf{x}_i, \mathbf{y})$, and γ is the separation margin. This problem can be equivalently transformed to a quadratic program as:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq 1, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ \forall i. \end{aligned} \tag{4}$$

Slack variables ξ_i can be introduced to allow some examples to fail to reach the margin, but only with an associated cost:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{m} \xi_i^2 \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq 1 - \xi_i, \forall \mathbf{y} \neq \mathbf{y}_i, \ \forall i; \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{5}$$

where $C > 0$ is called the regularization coefficient, trading off training errors with the margin.

2.1 LP Formulation

As mentioned above, in structured classification problems, the direct use of QP (5) is infeasible, as there will be too many potential negative examples yielding too large a constraint set. Before addressing a solution to this issue, first we make a slight modification to it by replacing the L_2-norm in the objective function with an L_1-norm to make it an LP problem. In addition, we constrain $\mathbf{w}$ to be non-negative, which simplifies its solution, as will be shown later. That is:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \|\mathbf{w}\|_1 + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq 1 - \xi_i, \forall \mathbf{y} \neq \mathbf{y}_i, \ \forall i; \\ & \mathbf{w} \geq \mathbf{0}; \ \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{6}$$

2.2 A More Complex Case

In some special problems, of which an example of concern is SMT, the situation will be a bit more complex. Firstly, in SMT the word lattice given by a source sentence $\mathbf{x}$ may have several paths yielding the same translation $\mathbf{y}$. So here we take each path denoted by y as a multilabel, rather than the output translation $\mathbf{y}$. In addition, for a given source sentence, there might be several reference translations. Moreover, it can not be guaranteed that the reference translations are among the paths in a word lattice. Therefore, in this case we will have to use the so-called pseudo-references that are the best translations we can obtain from the word lattice (Liang et al., 2006; Tillmann and Zhang, 2006) as positive examples.

We use Y to denote the set of the paths that yield good outputs (pseudo-references), whilst $\overline{Y}$ denotes the set of those leading to bad ones. One solution could be that we separate with a large margin each positive example $(\mathbf{x}, y)$, where $y \in Y$, from those negative examples $(\mathbf{x}, \bar{y})$, where $\bar{y} \in \overline{Y}$ and $\bar{y}$ has an internal structure close to y. Then the problem can be rewritten as:

$$\begin{aligned}\min_{\mathbf{w},\boldsymbol{\xi}} \quad & \|\mathbf{w}\|_1 + C\sum_{i=1}^{m}\xi_i \qquad (7)\\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi\left(\mathbf{x}_i, \arg\min_{y\in Y_i}\vartheta(y,\bar{y}), \bar{y}\right) \geq 1-\xi_i,\\ & \forall \bar{y}\in \overline{Y}_i,\ \forall i;\ \mathbf{w}\geq \mathbf{0};\ \boldsymbol{\xi}\geq \mathbf{0}.\end{aligned}$$

where we use $\vartheta(y,\bar{y})$ to denote some metric to measure the closeness of the internal structures of y and $\bar{y}$. This formulation works by enforcing at least parts of the pseudo-references to be highly ranked in contrast to the rest of the paths in every word lattice. The insight behind this is to control the variety inside the problem.

2.3 A More General Case

Alternatively, we could also try to separate all the good multilabels from the bad ones with a large margin:

$$\begin{aligned}\min_{\mathbf{w},\boldsymbol{\xi}} \quad & \|\mathbf{w}\|_1 + C\sum_{i=1}^{m}\xi_i \qquad (8)\\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, y, \bar{y}) \geq 1-\xi_i,\ \forall \bar{y}\in \overline{Y}_i, y\in Y_i,\ \forall i;\\ & \mathbf{w}\geq \mathbf{0};\ \boldsymbol{\xi}\geq \mathbf{0}.\end{aligned}$$

This gives a more general form of the structured classification problems, of which LP (6) and LP (7) can be taken as two special cases.

3 Column Generation

To solve such LP problems that have exponentially large constraint sets, column generation (CG) provides a practical solution without requiring the entire constraint set explicitly to be available. It starts from an initial point, incrementally adds the most violated constraints into a working set and solves the current relaxed subproblem, referred to as restricted master problem, until an optimum is achieved or some approximate stopping criterion is reached.

For convenience of expression, we rewrite our LP problems into matrix form:

$$\begin{aligned}\min_{\mathbf{w},\boldsymbol{\xi}} \quad & \mathbf{1}^\top\mathbf{w} + C\mathbf{1}^\top\boldsymbol{\xi} \qquad (9)\\ \text{s.t.} \quad & \mathbf{H}\mathbf{w}\geq \mathbf{1}-\mathbf{M}\boldsymbol{\xi};\ \mathbf{w}\geq\mathbf{0};\ \boldsymbol{\xi}\geq\mathbf{0}.\end{aligned}$$

where $\mathbf{H} = \left(\Delta\phi(\mathbf{x}_i, y, \bar{y})^\top_{\bar{y}\in\overline{Y}_i, y\in Y_i, 1\leq i\leq m}\right)$, $\mathbf{1}$ denotes the vector with components 1, $\boldsymbol{\xi}$ is the vector representation of the slack variables with the ith element ξ_i, and $\mathbf{M}$ is the matrix matching each ξ_i to its corresponding rows in $\mathbf{H}$. For future discussions, here we also give its dual form with dual variables $\boldsymbol{\lambda}$, as:

$$\begin{aligned}\max_{\boldsymbol{\lambda}} \quad & \mathbf{1}^\top\boldsymbol{\lambda} \qquad (10)\\ \text{s.t.} \quad & \mathbf{H}^\top\boldsymbol{\lambda}\leq\mathbf{1};\ \mathbf{M}^\top\boldsymbol{\lambda}\leq C\mathbf{1};\ \boldsymbol{\lambda}\geq\mathbf{0}.\end{aligned}$$

Algorithm 1: Column Generation for LP

```
1   input: $\{(\mathbf{x}_i, y_i) | y_i \in Y_i,\ i = 1,\ldots,m\}$
2   $\mathbf{w} \leftarrow \mathbf{1}, \boldsymbol{\xi} \leftarrow \mathbf{0}, \mathbf{H} \leftarrow (\ ), \mathbf{M} \leftarrow (\ )$
3   repeat
4       for $i \leftarrow 1$ to $m$
5           $\bar{y}_i \leftarrow \arg\max_{y\in\overline{Y}_i} \mathbf{w}^\top\phi(\mathbf{x}_i, y)$
6           $y_i \leftarrow \begin{cases} \arg\min_{y\in Y_i}\vartheta(y,\bar{y}_i) & : \text{LP (7)} \\ \arg\min_{y\in Y_i}\mathbf{w}^\top\phi(\mathbf{x}_i, y) & : \text{LP (8)} \end{cases}$
7           if $\mathbf{w}^\top\Delta\phi(\mathbf{x}_i, y_i, \bar{y}_i) < 1 - \xi_i$
8               $h \leftarrow \Delta\phi(\mathbf{x}_i, y_i, \bar{y}_i)^\top$
9               $\mathbf{H} \leftarrow \begin{pmatrix}\mathbf{H}\\ h\end{pmatrix}, \mathbf{M} \leftarrow \begin{pmatrix}\mathbf{M}\\ \delta_i\end{pmatrix}$ [1]
10          end if
11      end for
12      $(\mathbf{w}, \boldsymbol{\xi}) \leftarrow \begin{array}{rl} \min & \mathbf{1}^\top\mathbf{w} + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{s.t.} & \mathbf{H}\mathbf{w} \geq \mathbf{1} - \mathbf{M}\boldsymbol{\xi}; \\ & \mathbf{w}\geq\mathbf{0};\ \boldsymbol{\xi}\geq\mathbf{0}. \end{array}$
13  until convergence
14  return $\mathbf{w}$
```

Algorithm 1 illustrates the process of solving a LP-based structured classification problem using the column generation method. Note here, to balance between the number of constraints to be added into the working set and the number of times the LP subproblems must be solved, in each iteration we generate a most violated constraint for each training example, instead of the most violated constraint over the entire training set. The '$\arg\max$' operation in Line 5 usually can be achieved by doing a k-best decoding of the structure based on $\mathbf{w}$ and seeking the top ranked incorrect multilabel. For simpler problems that have a unique correct multilabel output for each input, such as part-of-speech tagging, $k = 2$ is sufficient.

4 Extragradient Method for LP

To solve the LP (sub)problems in Line 12 of Algorithm 1, the extradradient method proposed by Korpelevich (1976) is utilized. We start from a brief overview of it.

Let $\mathcal{Q}\subset\mathbb{R}^m$ and $\mathcal{S}\subset\mathbb{R}^n$ be two subsets of Euclidean space, and $\pi(\mathbf{u},\mathbf{v})$ be a real-valued function, where $\mathbf{u}\in\mathcal{Q}$ and $\mathbf{v}\in\mathcal{S}$. We assume that:

1. *$\mathcal{Q}$ and $\mathcal{S}$ are closed and convex.*

2. *$\pi(\mathbf{u},\mathbf{v})$ is convex in $\mathbf{u}$ and concave in $\mathbf{v}$, differentiable, and its partial derivatives satisfy the Lipschitz condition on $\mathcal{Q}\times\mathcal{S}$, i.e. there exists a con-*

[1] δ_i denotes a row vector with its ith element 1 and all the others 0.

stant $K \geq 0$ such that:

$$\|\pi_{\mathbf{u}}(\mathbf{u},\mathbf{v}) - \pi_{\mathbf{u}}(\mathbf{u}',\mathbf{v}')\|_2 \leq K(\|\mathbf{u}-\mathbf{u}'\|_2^2 + \|\mathbf{v}-\mathbf{v}'\|_2^2)^{\frac{1}{2}}$$
$$\|\pi_{\mathbf{v}}(\mathbf{u},\mathbf{v}) - \pi_{\mathbf{v}}(\mathbf{u}',\mathbf{v}')\|_2 \leq K(\|\mathbf{u}-\mathbf{u}'\|_2^2 + \|\mathbf{v}-\mathbf{v}'\|_2^2)^{\frac{1}{2}}$$

3. *The set of saddle points $\mathcal{U}^* \times \mathcal{V}^*$ of $\pi(\mathbf{u},\mathbf{v})$ on $\mathcal{Q} \times \mathcal{S}$ is nonempty.*

The extragradient method finds saddle points of $\pi(\mathbf{u},\mathbf{v})$ by the following update rules:

$$\begin{aligned}
\bar{\mathbf{u}}^t &= P_{\mathcal{Q}}(\mathbf{u}^t - \alpha\pi_{\mathbf{u}}(\mathbf{u}^t,\mathbf{v}^t)) \qquad (11)\\
\bar{\mathbf{v}}^t &= P_{\mathcal{S}}(\mathbf{v}^t + \alpha\pi_{\mathbf{v}}(\mathbf{u}^t,\mathbf{v}^t))\\
\mathbf{u}^{t+1} &= P_{\mathcal{Q}}(\mathbf{u}^t - \alpha\pi_{\mathbf{u}}(\bar{\mathbf{u}}^t,\bar{\mathbf{v}}^t))\\
\mathbf{v}^{t+1} &= P_{\mathcal{S}}(\mathbf{v}^t + \alpha\pi_{\mathbf{v}}(\bar{\mathbf{u}}^t,\bar{\mathbf{v}}^t))
\end{aligned}$$

where $\alpha \geq 0$, and $P_{\mathcal{Q}}$ and $P_{\mathcal{S}}$ are operators projecting their arguments onto the corresponding sets. Then Korpelevich (1976) proved the following theorem:

Theorem 4.1 *If assumptions 1–3 hold and $0 \leq \alpha \leq \frac{1}{K}$, then there exists a saddle point $(\mathbf{u}^*,\mathbf{v}^*) \in \mathcal{U}^* \times \mathcal{V}^*$ such that $(\mathbf{u}^t,\mathbf{v}^t) \to (\mathbf{u}^*,\mathbf{v}^*)$ when $t \to \infty$.*

Getting back to our problem LP (9) and LP (10), the extragradient method solves them by finding the saddle point of their Lagrange function:

$$\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\xi}}\max_{\boldsymbol{\lambda}} \quad & \mathbf{1}^\top\mathbf{w} + C\mathbf{1}^\top\boldsymbol{\xi} + \boldsymbol{\lambda}^\top\mathbf{1} - \boldsymbol{\lambda}^\top\mathbf{M}\boldsymbol{\xi} - \boldsymbol{\lambda}^\top\mathbf{H}\mathbf{w}\\
\text{s.t.} \quad & \mathbf{w} \geq \mathbf{0};\ \boldsymbol{\xi} \geq \mathbf{0};\\
& \boldsymbol{\lambda} \geq \mathbf{0}. \qquad (12)
\end{aligned}$$

The corresponding update rules are:

$$\begin{aligned}
\bar{\mathbf{w}}^t &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^t - \alpha(\mathbf{1} - \mathbf{H}^\top\boldsymbol{\lambda}^t)) \qquad (13)\\
\bar{\boldsymbol{\xi}}^t &= P_{\boldsymbol{\xi}\geq\mathbf{0}}(\boldsymbol{\xi}^t - \alpha(C\mathbf{1} - \mathbf{M}^\top\boldsymbol{\lambda}^t))\\
\bar{\boldsymbol{\lambda}}^t &= P_{\boldsymbol{\lambda}\geq\mathbf{0}}(\boldsymbol{\lambda}^t + \alpha(\mathbf{1} - \mathbf{M}\boldsymbol{\xi}^t - \mathbf{H}\mathbf{w}^t))\\
\mathbf{w}^{t+1} &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^t - \alpha(\mathbf{1} - \mathbf{H}^\top\bar{\boldsymbol{\lambda}}^t))\\
\boldsymbol{\xi}^{t+1} &= P_{\boldsymbol{\xi}\geq\mathbf{0}}(\boldsymbol{\xi}^t - \alpha(C\mathbf{1} - \mathbf{M}^\top\bar{\boldsymbol{\lambda}}^t))\\
\boldsymbol{\lambda}^{t+1} &= P_{\boldsymbol{\lambda}\geq\mathbf{0}}(\boldsymbol{\lambda}^t + \alpha(\mathbf{1} - \mathbf{M}\bar{\boldsymbol{\xi}}^t - \mathbf{H}\bar{\mathbf{w}}^t))
\end{aligned}$$

where the step size α can be estimated by $(2\|\mathbf{H}\|_F^2 + 2\|\mathbf{M}\|_F^2)^{-\frac{1}{2}}$, and $\|\cdot\|_F$ denotes the Frobenius norm.

4.1 Extragradient method with CG

When applied to solve our restricted master problems in Algorithm 1, this extragradient method can be slightly modified to offer a practical solution for large-scale training tasks as follows. The proofs of Korpelevich (1976) suggest that the iterative process (13) can be started from an arbitrary feasible point for the primal and the dual problems LP (9) and LP (10), which, after finding a basis in a number of initial steps, will go along a spiral curve and converge to the optimum with the speed of a geometric progression. Thus we could intuitively expect that it will converge faster if started from a feasible point closer to the solution. We replace Line 12 of Algorithm 1 with Algorithm 2, where in each iteration the previous solution is used to find a proper starting point for the extragradient method. Moreover, further efficiency can be gained by not solving the LP subproblems exactly, but to a tolerance (Line 10 of Algorithm 2) leaving the final solution to be appropriately tightened when the optimum is achieved. Since at each time Line 5 of Algorithm 2 actually gives a feasible point for the entire primal problem LP (9), whilst if all those elements of $\boldsymbol{\lambda}$ that correspond to the constraints out of the working set are regarded as 0, $\boldsymbol{\lambda}$ is also a feasible point for the dual problem LP (10), Algorithm 2 will approach the global optimum when the whole process converges.

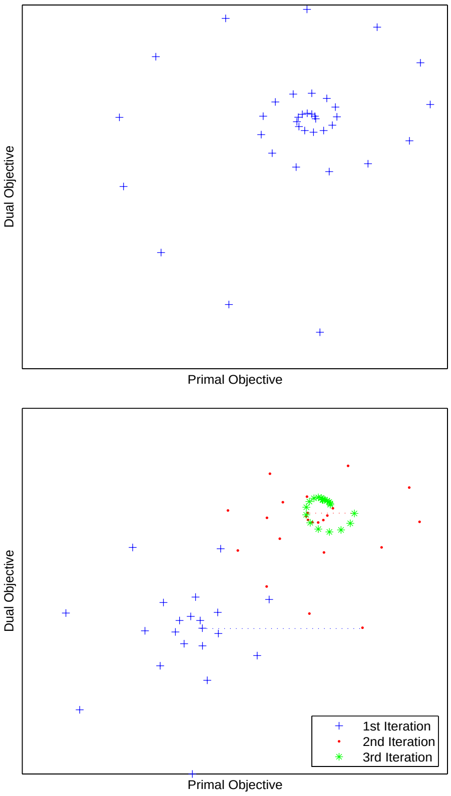

Figure 1: A Visual Demonstration of the Extragradient Method for LP with Column Generation: the top figure shows the process of solving an entire LP problem with the extragradient method; the bottom figure illustrates the column generation procedure to solve the same problem according to Algorithm 2.

Figure 1 visualizes this process by observing the objective values of a pair of primal and dual problems. In

Algorithm 2: Extragradient Method with CG

```
1   tolerances: $\epsilon_1$, $\epsilon_2$
2   $\mathbf{w}^0 \leftarrow \mathbf{w}$, $\boldsymbol{\xi}^0 \leftarrow \boldsymbol{\xi}$, $\boldsymbol{\lambda}^0 \leftarrow \boldsymbol{\lambda}$
3   for $i \leftarrow 1$ to $m$
4       if $\mathbf{w}^\top \Delta\phi(\mathbf{x}_i, y_i, \bar{y}_i) < 1 - \xi_i$
5           $\xi_i^0 \leftarrow (1 - \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, y_i, \bar{y}_i))$
6           $\boldsymbol{\lambda}^0 \leftarrow (\boldsymbol{\lambda}^\top, 0)^\top$
7       end if
8   end for
9   iterate process (13) from $((\mathbf{w}^0, \boldsymbol{\xi}^0), \boldsymbol{\lambda}^0)$
10  until $\max\left\{\frac{\|(\mathbf{w}^t,\boldsymbol{\xi}^t)-(\mathbf{w}^{t-1},\boldsymbol{\xi}^{t-1})\|_2}{\|(\mathbf{w}^t,\boldsymbol{\xi}^t)\|_2}, \frac{\|\boldsymbol{\lambda}^t-\boldsymbol{\lambda}^{t-1}\|_2}{\|\boldsymbol{\lambda}^t\|_2}\right\} < \epsilon_1$
        && $\|\mathbf{w}^t\|_1 + C\|\boldsymbol{\xi}^t\|_1 - \|\boldsymbol{\lambda}^t\|_1 < \epsilon_2$
11  $\mathbf{w} \leftarrow \mathbf{w}^t$, $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi}^t$, $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}^t$
```

the top figure, it can be seen that in the bilinear case the extragradient method converges to the optimum solution (where the primal and dual objectives have equal values) along a spiral curve. The bottom one shows that after CG, Line 5 of Algorithm 2 actually makes the starting point for the next iteration shift along the dashed line. Although the actual convergence speed in the next iteration still depends on the seriousness of the current suboptimal solution violating the constraints, i.e. how far the next starting point will be shifted from the diagonal line, the extragradient method converges geometrically, which suggests that it can provide an acceptable approximation in not too many steps, since we tend to solve the subproblems to a given tolerance.

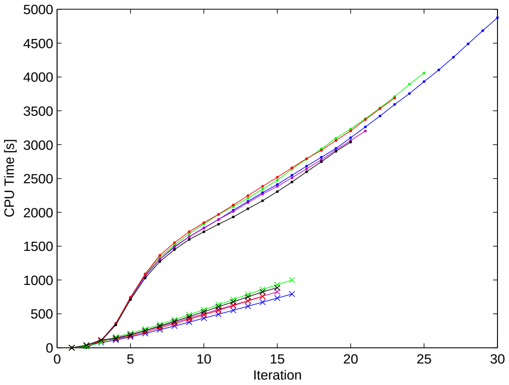

Figure 2: CPU Time Expenses: dual-simplex (dot markers) vs. extragradient (cross markers). (The 5 dot/cross curves are for the 5 repeats of the experiment. Each data point stands for the total CPU time expense up to that iteration. Hence, the number of points on a curve gives the number of iterations spent in total.)

5 Experimental Results

To demonstrate the effect of our method, we test it on two different tasks. In all the following experiments, the algorithms are implemented in C/C++ and run on a cluster machine with 8×3.00GHz Intel(R) Xeon(R) CPUs and 32GB memory.

5.1 Part-of-speech Tagging

We first experiment with the proposed LP method, which we name LP-Struct, on a part-of-speech tagging task, which follows the formulation in LP (6). The corpus we used consists of 6700 manually tagged sentences from a bibliographic database of publications MEDLINE (Smith et al., 2004). We perform random sampling 5 times. At each time, 1000 sentences are selected as the test set, with the remaining 5700 sentences being the training set. The features used for each label in the following experiments are simply its observation and one previous label (i.e. a first-order hidden Markov model). We also compare our LP-Struct to some existing techniques for structured classification problems, including probabilistic hidden Markov model (HMM), conditional random field (CRF) (Lafferty et al., 2001), structured perceptron (Collins, 2002), and single-best MIRA (Crammer et al., 2005). Here, we use an open-source toolkit CRF++[2] to train the CRF and MIRA models. In addition, we compare two versions of LP-Struct that solve the subproblems via the extragradient method and the standard dual-simplex method respectively. The implementation of the dual-simplex method is from LP_SOLVE[3], one of the most efficient open-source LP solvers. The results are shown in Table 1. It can be seen that LP-Struct outperforms HMM and the structured perceptron, and has a result very close to MIRA, but works slightly worse than CRF. However, the training of LP-Struct is significantly faster than MIRA and CRF. In Figure 2 we compare the CPU time and the number of iterations spent in training LP-Struct based on the dual-simplex method and the extragradient method. Not only is the extragradient method much faster than the dual-simplex method, the trend of its time growth is also more stable, as the trend shown in the first a few iterations of the dual-simplex method suggests that it might take too much time to finish some intermediate steps if used for problems of a much larger scale.

5.2 Statistical Machine Translation

Next we experiment with LP-Struct on SMT, a much more complex problem, based on the formulations in

[2] Available at: http://crfpp.sourceforge.net/

[3] Available at: http://lpsolve.sourceforge.net/5.5/

Table 1: Experimental Results for Part-of-Speech Tagging

Model	Error Rate (%)	Avg. CPU Time (s)	Avg. #Iteration
CRF	4.58±0.14	51403	205
MIRA	4.91±0.06	9084	46
Perceptron	5.38±0.19	26	100
LP-Struct/Simplex	4.94±0.18	3879	23
LP-Struct/Extragradient	4.92±0.13	856	14
HMM	20.02±0.29	–	–

Table 2: Experimental Results for SMT

Training Method	LP (7)	LP (8)	MERT
BLEU (%)	32.35	32.30	31.69
NIST	7.95	8.19	7.94

LP (7) and LP (8). We modify the Moses (Koehn et al., 2007) system to make it fit for our experiments, and do a purely-discriminative training for it on the Senate Debates data set of the Canada Hansards corpus. There are 182K sentence pairs in the training set and 12K and 13K sentence pairs in two separate test sets. Here we only translate in one direction, French to English. Pseudo-references are generated by searching for the highest-BLEU-scored hypotheses in beam search stacks. In addition, to simplify the situation, in our experiments we only allow the adjacent phrases to exchange their positions. The features we use for training include both blanket features and discriminative features, similar to Liang et al.'s (2006) work. Concretely, for blanket features we have forward and backward orientation-based distortion probabilities (6 features), one tri-gram language model probability, bidirectional translation probabilities (2 features) and lexicon weights (2 features), a word penalty, a phrase penalty and a distortion distance penalty, i.e. 14 features in total. We also use all the English trigrams, bilingual phrase pairs and their corresponding distortion orientations extracted from the training set as the discriminative features. Finally, we have in total about 8 millon features. The LP solver here employs the extragradient method introduced above. We test it on the 13K test set and compare the results to the baseline Moses system whose parameters (14 blanket features only) are tuned based on the minimum error-rate training (MERT) (Och, 2003) on the 12K development set. The results are summarized in Table 2, where we find that both of the models LP (7) and LP (8) improve the baseline. Interestingly, LP (7) gains more on the BLEU score, whilst LP (8) gains more on the NIST score. These results are statistical significant according to the approximate randomization test (Riezler and Maxwell III, 2005), with $p < 0.05$.

6 Generalization Bound

This section gives a generalization bound of our LP-based method for structured classification. We start by introducing some notation.

In our problem LP (8), we actually take a triple $(\mathbf{x}, y, \bar{y})$ as a training example. We assume that $(\mathbf{x}, y, \bar{y})$ is generated from the joint space $X := \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$, and $\mathcal{F}$ to be a class of real-valued functions on X, such that $\mathcal{F}(X) := \{f = \langle \mathbf{w}, \Delta\phi(X)\rangle | \Delta\phi : X \to \mathbb{R}^d, \mathbf{w} \in \mathbb{R}^{d+}\}$. Let $\mathcal{D}$ be a distribution on X. The error $\mathrm{err}_{\mathcal{D}}(f)$ of a function $f \in \mathcal{F}$ is defined to be the probability $\mathcal{D}\{(\mathbf{x}, y, \bar{y}) : f(\mathbf{x}, y, \bar{y}) < 0\}$. We can also rewrite our training set in the form of $S = s_1 \cup s_2 \cup \ldots \cup s_m$, where $s_i := \{(\mathbf{x}_i, y_i, \bar{y}_i) | y_i \in Y_i, \bar{y}_i \in \overline{Y}_i\}$.

The proposition from Schölkopf et al. (2001) to elucidate the relationship between single-class classification and binary classification can be adopted here.

Proposition 6.1 *(i) Suppose* $\mathbf{w}$ *parameterizes the supporting hyperplane for the data set* S*. Then* $\mathbf{w}$ *parameterizes the optimal separating hyperplane for the labeled data set,* $\{((\mathbf{x}_i, y_i, \bar{y}_i), 1) | y_i \in Y_i, \bar{y}_i \in \overline{Y}_i, i = 1, \ldots, m\} \cup \{((\mathbf{x}_i, \bar{y}_i, y_i), -1) | y_i \in Y_i, \bar{y}_i \in \overline{Y}_i, i = 1, \ldots, m\}$*. (ii) Suppose* $\mathbf{w}$ *parameterizes the optimal separating hyperplane passing through the origin for a labeled data set,* $\{((\mathbf{x}_i, y_i, y'_i), z_i) | z_i \in \{-1, +1\}, i = 1, \ldots, m\}$*, aligned such that* $y_i \in Y_i, y'_i \in \overline{Y}_i$ *for* $z_i = 1$*, and* $y'_i \in \overline{Y}_i, y'_i \in Y_i$ *for* $z_i = -1$*. Then* $\mathbf{w}$ *parameterizes the supporting hyperplane for the unlabeled data set,* $\{(\mathbf{x}_i, y_i, \bar{y}_i) | y_i \in Y_i, \bar{y}_i \in \overline{Y}_i, i = 1, \ldots, m\}$*.*

Now we can utilize the methodology for SVMs (Cristianini and Shawe-Taylor, 2000) to analyze our model. We start by introducing the following definitions.

Definition 6.1 *Let* $\mathcal{F}$ *be a family of real-valued functions on domain* X*. Given a sample* $S \in X^l$*, we say that a finite set of functions* $\mathcal{B}$ *covers* $\mathcal{F}$ *at radius* γ *if for all* $f \in \mathcal{F}$*, there exists* $g \in \mathcal{B}$*, such that for each data point* $\mathbf{z} \in S$*,* $|f(\mathbf{z}) - g(\mathbf{z})| < \gamma$*. The covering number of* $\mathcal{F}$ *with respect to* S*, denoted by* $\mathcal{N}(\mathcal{F}, S, \gamma)$*, is the size of the smallest such cover. We also define* *the covering number of $\mathcal{F}$ for any sample of size l, as:*

$$\mathcal{N}(\mathcal{F}, l, \gamma) := \max_{S \in X^l} \mathcal{N}(\mathcal{F}, S, \gamma).$$

In our case, given an $f \in \mathcal{F}$ the margin of an example $(\mathbf{x}, y, \bar{y})$ is defined to be $f(\mathbf{x}, y, \bar{y})$, which corresponds to the binary classification view $zf(\mathbf{x}, y, y')$. We also define the margin of the training set S as $m_S(f) := \min_{(\mathbf{x},y,\bar{y})\in S} f(\mathbf{x}, y, \bar{y})$. Then we have the following theorem derived by Cristianini and Shawe-Taylor (2000).

Theorem 6.1 *Consider thresholding a real-valued function space $\mathcal{F}$ and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathcal{D}$ on X, with probability $1 - \eta$ over the training set S, any function $f \in \mathcal{F}$ that has margin $m_S(f) \geq \gamma$ on S has generalization error no more than $err_{\mathcal{D}}(f) \leq \varepsilon(|S|, \mathcal{F}, \eta, \gamma)$, where:*

$$\varepsilon(|S|, \mathcal{F}, \eta, \gamma) = \frac{2}{|S|}\left(\log_2 \mathcal{N}(\mathcal{F}, 2|S|, \frac{\gamma}{2}) + \log_2 \frac{2}{\eta}\right),$$

provided $|S| > \frac{2}{\varepsilon}$.

Theorem 6.1 can be applied to our soft-margin case as follows. For an input space X, we define the auxiliary inner product space:

$$L(X) := \left\{ g \in \mathbb{R}^X : \begin{array}{l}\text{supp}(g) \text{ is}\\ \text{countable and}\end{array} \sum_{\mathbf{z}\in\text{supp}(g)} g(\mathbf{z})^2 < \infty \right\},$$

where for $g, h \in L(X)$, the inner product is given by $\langle g, h\rangle := \sum_{\mathbf{z}\in\text{supp}(g)} g(\mathbf{z})h(\mathbf{z})$. Now, we embed our input space into space $X \times L(X)$ using the mapping $\tau : (\mathbf{x}, y, \bar{y}) \mapsto ((\mathbf{x}, y, \bar{y}), \frac{1}{C}\delta_{\hat{\mathbf{x}}})$, where $C \in \mathbb{R}^+$ is a constant, and $\delta_{\hat{\mathbf{x}}} \in L(X)$ is defined as:

$$\delta_{\hat{\mathbf{x}}}(\mathbf{x}, y, \bar{y}) := \begin{cases} 1 & \text{if } \mathbf{x} = \hat{\mathbf{x}}; \\ 0 & \text{otherwise.} \end{cases}$$

Hence, for a function $(f, g) \in \mathcal{F} \times L(X)$ we define its action on $\tau(\mathbf{x}, y, \bar{y}) \in X \times L(X)$ to be:

$$(f, g)(\tau(\mathbf{x}, y, \bar{y})) := f(\mathbf{x}, y, \bar{y}) + \frac{1}{C}\langle g, \delta_{\mathbf{x}}\rangle.$$

Now for fixed margin γ, the slack variables ξ_i in LP (8) can be derived from $\xi_i = \max\{0, \gamma - \inf_{y_i\in Y_i, \bar{y}_i \in \overline{Y}_i} f(\mathbf{x}_i, y_i, \bar{y}_i)\}$. By defining $g(S, f, \gamma) \in L(X)$ to be:

$$g(S, f, \gamma) := C \sum_{i=1}^{m} \xi_i \delta_{\mathbf{x}_i},$$

we can make the data separable with margin γ. It is easy to check that $\forall (\mathbf{x}_i, y_i, \bar{y}_i) \in S$: $(f, g)(\tau(\mathbf{x}_i, y_i, \bar{y}_i)) = f(\mathbf{x}_i, y_i, \bar{y}_i) + \xi_i \geq \gamma$, while $\forall (\mathbf{x}, y, \bar{y}) \notin S$: $(f, g)(\tau(\mathbf{x}, y, \bar{y})) = f(\mathbf{x}, y, \bar{y})$. Theorem 6.1 can therefore be translated to the following theorem.

Theorem 6.2 *Consider thresholding a real-valued function space $\mathcal{F}$ and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathcal{D}$ on X, with probability $1 - \eta$ over the training set S, any function $f \in \mathcal{F}$ for which $(f, g) \in \mathcal{G} := \mathcal{F} \times L(X)$ has generalization error no more than $err_{\mathcal{D}}(f) \leq \varepsilon(|S|, \mathcal{G}, \eta, \gamma)$, where*

$$\varepsilon(|S|, \mathcal{G}, \eta, \gamma) = \frac{2}{|S|}\left(\log_2 \mathcal{N}(\mathcal{G}, 2|S|, \frac{\gamma}{2}) + \log_2 \frac{2}{\eta}\right),$$

provided $|S| > \frac{2}{\varepsilon}$, and there is no discrete probability on misclassified training points.

Based on our previous definition of $\mathcal{F}(X)$, the L_1-norm of (f, g) is then given by $\|(f, g)\|_1 = \|\mathbf{w}\|_1 + C\|\boldsymbol{\xi}\|_1$. If we assume $\max\{\|\Delta\phi(X)\|_\infty, \frac{1}{C}\} \leq b$ and $\|\mathbf{w}\|_1 + C\|\boldsymbol{\xi}\|_1 \leq c$, we can obtain the following corollary from Zhang's (2002) Theorem 5.

Corollary 6.3 *For the function class $\mathcal{G} := \mathcal{F} \times L(X)$ defined above, we have that:*

$$\log_2 \mathcal{N}(\mathcal{G}, l, \gamma) \leq \frac{36c^2b^2(1 + \ln(d + m))}{\gamma^2} \log_2\left(2\left\lceil \frac{4cb}{\gamma} + 2 \right\rceil l + 1\right).$$

Proof Given a uniform distribution vector $\boldsymbol{\mu} \in \mathbb{R}^{n+}$ with each element $\mu_i = \frac{1}{n}$, it is easy to check that for any $\mathbf{v} \in \mathbb{R}^n$ its weighted relative entropy with respect to $\boldsymbol{\mu}$ defined to be $\text{entro}_{\boldsymbol{\mu}}(\mathbf{v}) := \sum_{i=1}^{n} |v_i| \ln \frac{|v_i|}{\mu_i\|\mathbf{v}\|_1}$ satisfies $\text{entro}_{\boldsymbol{\mu}}(\mathbf{v}) \leq \|\mathbf{v}\|_1 \ln n$. Letting $\mathbf{v} = (\mathbf{w}, \boldsymbol{\xi}) \geq \mathbf{0}$ and inserting the above result into Zhang's (2002) Theorem 5 proves Corollary 6.3. ■

Note here, in our case, for any $\gamma > 0$, at the optimum of LP (8), the quantity $\frac{c}{\gamma} = a$ will be a constant.

It can be seen that minimizing the objective function in LP (8) will directly minimizes the generalization error bound of our method. A similar result is given by Demiriz et al. (2002) for the LP Boosting models. But the latter applies LP to boosting feature selection, which is for another problem domain.

When compared to the PAC-Bayesian bound for L_2-regularized structured classification models due to Bartlett et al. (2004), our bound does not have the logarithmic dependence on the number of labels in the training set, but a logarithmic dependence on the feature dimension, which will be much less than the number of labels in the training set in many practical problems, e.g. part-of-speech tagging, parsing and many other natural language processing tasks. In addition, our bound depends on the potential training sample size $|S|$ that is exponentially large, but not the number of training examples m. As the dependence is $\frac{\log_2 l}{l}$, it will significantly gain over the previous results of Bartlett et al. (2004).

7 Conclusions

In this paper, we present a novel algorithm for structured classification problems, which models the SVM-style large-margin separation problem with a linear program. To handle the exponentially large constraint set, the column generation technique is employed. In addition, we argue that further efficiency can be obtained if the restricted master problems are solved using the extragradient method. We show encouraging results by applying our algorithm to part-of-speech tagging and statistical machine translation tasks. To the best of our knowledge, none of the previous large-margin structured prediction models has been applied to handle such large-scale problems as in our experiments. Furthermore, we prove that the generalization error bound of our model can be directly optimized by minimizing the empirical risk on the training data. However, in the SMT case, a drawback of the proposed method is that it is to some extent sensitive to the quality of pseudo-references, as it tends to spend more iterations on those inseparable (but possibly inconsequential) examples. However, generating highly reliable pseudo-references itself could be a difficult problem as well. To develop a more robust algorithm will be one of our future research directions.

Acknowledgements

This work is supported by the EU project SMART (FP6-033917). We thank Sándor Szedmák for numerous useful discussions on the extragradient method and providing an initial MATLAB implementation of the original algorithm.

References

Y. Altun, I. Tsochantaridis, and T. Hofmann (2003). Hidden markov support vector machines. In *ICML.*

P. L. Bartlett, M. Collins, B. Taskar, and D. A. McAllester (2004) Exponentiated gradient algorithms for large-margin structured classification. In *NIPS.* Longer version available at: http://www.stat.berkeley.edu/˜bartlett/papers/bcmt-lmmsc-04.pdf

M. Collins (2002) Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP.*

K. Crammer and Y. Singer (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2.

K. Crammer, R. McDonald, and F. Pereira (2005). Scalable large-margin online learning for structured classification. Technical report, University of Pennsylvania.

N. Cristianini and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

A. Demiriz, K.P. Bennett and J. Shawe-Taylor (2002). Linear programming boosting via column generation. *Machine Learning* 46(1–3).

P. Koehn et al. (2007). Moses: open source toolkit for statistical machine translation. In *ACL 2007 Demo and Poster Sessions.*

G. Korpelevich (1976). The extragrdient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756.

J.D. Lafferty, A. McCallum, and F.C.N. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML.*

P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar (2006). An end-to-end discriminative approach to machine translation. In *COLING/ACL.*

F.J. Och (2003). Minimum error rate training in statistical machine translation. In *ACL.*

S. Riezler and J.T. Maxwell III (2005). On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Tranlsation and/or Summarization.*

B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7).

L. Smith, T. Rindflesch, and W.J. Wilbur (2004). Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

B. Taskar, C. Guestrin, and D. Koller (2003). Max-margin markov networks. In *NIPS.*

B. Taskar, S. Lacoste-Julien, and M. I. Jordan (2006). Structured prediction, dual extragradient and bregman projections. *Journal of Machine Learning Research*, 7.

C. Tillmann and T. Zhang (2006). A discriminative global training algorithm for statistical MT. In *COLING/ACL.*

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6.

T. Zhang (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2.