# Large-Margin Structured Prediction via Linear Programming

Zhuoran Wang[1] John Shawe-Taylor[1] Sándor Szedmák[2]

[1]Computer Science, University College London
[2]Electronics and Computer Science, University of Southampton

May 13, 2009

# Structured Prediction

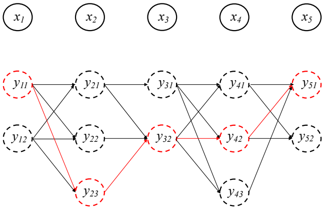
- Each (multi-label) output contains multiple (micro-)labels

# Structured Prediction

- Each (multi-label) output contains multiple (micro-)labels
- Micro-labels interacts each other

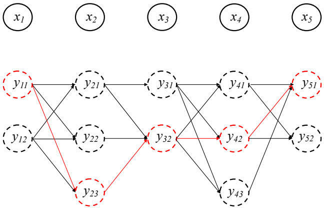
# Structured Prediction

- Each (multi-label) output contains multiple (micro-)labels
- Micro-labels interacts each other
- Example: sequence labeling (HMM)



# Structured Prediction

- Each (multi-label) output contains multiple (micro-)labels
- Micro-labels interacts each other
- Example: sequence labeling (HMM)




- More examples: parsing tree, bipartite matching, hierarchical classification, etc

# Structured Prediction (Cont.)

- Predict multi-label $\mathbf{y} = y_1, y_2, \ldots, y_l$ for an input object $\mathbf{x}$.

# Structured Prediction (Cont.)

- Predict multi-label $\mathbf{y} = y_1, y_2, \ldots, y_l$ for an input object $\mathbf{x}$.
- Formally, given input and output space $\mathcal{X}$ and $\mathcal{Y}$, learn a $\mathbf{w}$-parameterized function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that the prediction $\hat{\mathbf{y}} \in \mathcal{Y}$ for an arbitrary $\mathbf{x} \in \mathcal{X}$ is derived by:

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

# Structured Prediction (Cont.)

- Predict multi-label $\mathbf{y} = y_1, y_2, \ldots, y_l$ for an input object $\mathbf{x}$.
- Formally, given input and output space $\mathcal{X}$ and $\mathcal{Y}$, learn a $\mathbf{w}$-parameterized function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that the prediction $\hat{\mathbf{y}} \in \mathcal{Y}$ for an arbitrary $\mathbf{x} \in \mathcal{X}$ is derived by:

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

- Assume $f$ is from the linear family, and define the joint feature mapping $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. Then we have:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \mathbf{w}^\top \Phi(\mathbf{x}, \mathbf{y})$$

# Structured Prediction (Cont.)

- Predict multi-label $\mathbf{y} = y_1, y_2, \ldots, y_l$ for an input object $\mathbf{x}$.
- Formally, given input and output space $\mathcal{X}$ and $\mathcal{Y}$, learn a $\mathbf{w}$-parameterized function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that the prediction $\hat{\mathbf{y}} \in \mathcal{Y}$ for an arbitrary $\mathbf{x} \in \mathcal{X}$ is derived by:

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

- Assume $f$ is from the linear family, and define the joint feature mapping $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. Then we have:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \mathbf{w}^\top \Phi(\mathbf{x}, \mathbf{y})$$

- Seek the $\mathbf{w}$-parameterized hyperplane separating the positive and negative training examples $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ with large margin.

# Existing Techniques

- Structured Perceptron [Collins, 2002]

# Existing Techniques

- Structured Perceptron [Collins, 2002]
- Margin Infused Relaxed Algorithm (MIRA) [Crammer *et al.*, 2006]

# Existing Techniques

- Structured Perceptron [Collins, 2002]
- Margin Infused Relaxed Algorithm (MIRA) [Crammer *et al.*, 2006]
- SVM-type Algorithms

# Existing Techniques

- Structured Perceptron [Collins, 2002]
- Margin Infused Relaxed Algorithm (MIRA) [Crammer *et al.*, 2006]
- SVM-type Algorithms
  - Hidden Markov Support Vector Machines [Altun *et al.*, 2003] and extensions [Tsochantaridis *et al.*, 2005]

# Existing Techniques

- Structured Perceptron [Collins, 2002]
- Margin Infused Relaxed Algorithm (MIRA) [Crammer *et al.*, 2006]
- SVM-type Algorithms
  - Hidden Markov Support Vector Machines [Altun *et al.*, 2003] and extensions [Tsochantaridis *et al.*, 2005]
  - Max-Margin Markov Networks [Taskar *et al.*, 2003]

# Existing Techniques

- Structured Perceptron [Collins, 2002]
- Margin Infused Relaxed Algorithm (MIRA) [Crammer *et al.*, 2006]
- SVM-type Algorithms
  - Hidden Markov Support Vector Machines [Altun *et al.*, 2003] and extensions [Tsochantaridis *et al.*, 2005]
  - Max-Margin Markov Networks [Taskar *et al.*, 2003]
  - Combinatorial Models [Taskar *et al.*, 2004,2005,2006]

# Large-Margin Separation

- SVM-style formulation:

$$\max_{\mathbf{w},\gamma} \quad \gamma$$
$$\text{s.t.} \quad \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq \gamma, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ i = 1, \ldots, m;$$
$$\|\mathbf{w}\|_2 = 1.$$

# Large-Margin Separation

- SVM-style formulation:

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

- Equivalent form:

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

where $\Delta\Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) = \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})$.

# Large-Margin Separation (Cont.)

- Soft margin:

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

# $L_1$-Regularized Optimization

- Modifying SVM formulation with $L_1$-norm regularization:

$$\begin{aligned} \max_{\mathbf{w},\gamma} \quad & \gamma \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq \gamma, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ \ i = 1, \ldots, m; \\ & \|\mathbf{w}\|_1 = 1; \ \ \mathbf{w} \geq \mathbf{0}. \end{aligned}$$

# $L_1$-Regularized Optimization

- Modifying SVM formulation with $L_1$-norm regularization:

$$\begin{aligned} \max_{\mathbf{w},\gamma} \quad & \gamma \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq \gamma, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ i = 1, \ldots, m; \\ & \|\mathbf{w}\|_1 = 1; \ \mathbf{w} \geq \mathbf{0}. \end{aligned}$$

- Equivalent form:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}) \geq 1, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ i = 1, \ldots, m; \\ & \mathbf{w} \geq \mathbf{0}. \end{aligned}$$

# $L_1$-Regularized Optimization (Cont.)

- Soft margin:

$$\begin{aligned} \max_{\mathbf{w},\boldsymbol{\xi},\gamma} \quad & \gamma - D\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{y}) \geq \gamma - \xi_i,\ \forall \mathbf{y} \neq \mathbf{y}_i,\ i = 1,\ldots,m; \\ & \|\mathbf{w}\|_1 = 1;\ \mathbf{w} \geq \mathbf{0};\ \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned}$$

# $L_1$-Regularized Optimization (Cont.)

- Soft margin:

$$\begin{aligned}\max_{\mathbf{w},\boldsymbol{\xi},\gamma} \quad & \gamma - D\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{y}) \geq \gamma - \xi_i,\ \forall \mathbf{y} \neq \mathbf{y}_i,\ i = 1,\ldots,m; \\ & \|\mathbf{w}\|_1 = 1;\ \mathbf{w} \geq \mathbf{0};\ \boldsymbol{\xi} \geq \mathbf{0}.\end{aligned}$$

- Equivalent form:

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

# $L_1$-Regularized Optimization (Cont.)

- Soft margin:

$$\max_{\mathbf{w},\boldsymbol{\xi},\gamma} \quad \gamma - D\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{y}) \geq \gamma - \xi_i, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ i = 1,\ldots,m;$$

$$\|\mathbf{w}\|_1 = 1; \ \mathbf{w} \geq \mathbf{0}; \ \boldsymbol{\xi} \geq \mathbf{0}.$$

- Equivalent form:

$$\min_{\mathbf{w},\boldsymbol{\xi}} \quad \|\mathbf{w}\|_1 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \mathbf{w}^\top \Delta\Phi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{y}) \geq 1 - \xi_i, \ \forall \mathbf{y} \neq \mathbf{y}_i, \ i = 1,\ldots,m;$$

$$\mathbf{w} \geq \mathbf{0}; \ \boldsymbol{\xi} \geq \mathbf{0}.$$

- The latter is more convenient and efficient to handle in practical computations.

# Column Generation

**Algorithm** 1: LP-based training with column generation

input: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$

$\mathbf{w} \leftarrow \mathbf{1}, \boldsymbol{\xi} \leftarrow \mathbf{0}, \mathbf{H} \leftarrow (\ ), \mathbf{M} \leftarrow (\ )$

repeat

  for $i \leftarrow 1$ to $m$

    $\hat{\mathbf{y}} \leftarrow \arg\max_{\mathbf{y} \neq \mathbf{y}_i} \mathbf{w}^\top \phi(\mathbf{x}_i, \mathbf{y})$

    if $\mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}) < 1 - \xi_i$

      $h \leftarrow \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}})^\top$

      $\mathbf{H} \leftarrow \begin{pmatrix} \mathbf{H} \\ h \end{pmatrix}, \mathbf{M} \leftarrow \begin{pmatrix} \mathbf{M} \\ \delta_i{}^* \end{pmatrix}$

    end if

  end for

$$(\mathbf{w}, \boldsymbol{\xi}) \leftarrow \begin{array}{ll} \min & \mathbf{1}^\top \mathbf{w} + C\mathbf{1}^\top \boldsymbol{\xi} \\ \text{s.t.} & \mathbf{H}\mathbf{w} \geq \mathbf{1} - \mathbf{M}\boldsymbol{\xi}; \\ & \mathbf{w} \geq \mathbf{0}; \boldsymbol{\xi} \geq \mathbf{0}. \end{array}$$

until convergence

return $\mathbf{w}$

$^*$ $\delta_i$ denotes the row vector with the $i$th component 1 and all the others 0.

# Extragradient Method

- 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:

# Extragradient Method

- 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:
    - $\mathcal{Q}$ and $\mathcal{S}$ are closed and convex.

# Extragradient Method

- 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:
  - $\mathcal{Q}$ and $\mathcal{S}$ are closed and convex.
  - $\pi(\mathbf{u}, \mathbf{v})$ is convex on $\mathbf{u}$ and concave on $\mathbf{v}$ , differentiable and its partial derivatives satisfy the Lipschitz condition on $\mathcal{Q} \times \mathcal{S}$, i.e. there exists a constant $K \geq 0$ such that:

$$\begin{aligned} \|\pi_{\mathbf{u}}(\mathbf{u}, \mathbf{v}) - \pi_{\mathbf{u}}(\mathbf{u}', \mathbf{v}')\|_2 &\leq K(\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{v} - \mathbf{v}'\|_2^2)^{1/2} \\ \|\pi_{\mathbf{v}}(\mathbf{u}, \mathbf{v}) - \pi_{\mathbf{v}}(\mathbf{u}', \mathbf{v}')\|_2 &\leq K(\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{v} - \mathbf{v}'\|_2^2)^{1/2} \end{aligned}$$

# Extragradient Method

- 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:
  - $\mathcal{Q}$ and $\mathcal{S}$ are closed and convex.
  - $\pi(\mathbf{u}, \mathbf{v})$ is convex on $\mathbf{u}$ and concave on $\mathbf{v}$ , differentiable and its partial derivatives satisfy the Lipschitz condition on $\mathcal{Q} \times \mathcal{S}$, i.e. there exists a constant $K \geq 0$ such that:

$$\begin{aligned} \|\pi_{\mathbf{u}}(\mathbf{u}, \mathbf{v}) - \pi_{\mathbf{u}}(\mathbf{u}', \mathbf{v}')\|_2 &\leq K(\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{v} - \mathbf{v}'\|_2^2)^{1/2} \\ \|\pi_{\mathbf{v}}(\mathbf{u}, \mathbf{v}) - \pi_{\mathbf{v}}(\mathbf{u}', \mathbf{v}')\|_2 &\leq K(\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{v} - \mathbf{v}'\|_2^2)^{1/2} \end{aligned}$$

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

# Extragradient Method (Cont.)

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

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

  where $\alpha \geq 0$, and $P_{\mathcal{Q}}$ and $P_{\mathcal{S}}$ are operators projecting their argument onto the corresponding sets.

# Extragradient Method (Cont.)

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

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

  where $\alpha \geq 0$, and $P_{\mathcal{Q}}$ and $P_{\mathcal{S}}$ are operators projecting their argument onto the corresponding sets.

- **Theorem 1.**[Korpelevich, 1976] If assumptions hold and in addition $0 \leq \alpha \leq \frac{1}{K}$, then there exits 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 $t \to \infty$.

# Extragradient Method for LP

- LP in standard form:

Primal:

$$\begin{aligned} \min \quad & \mathbf{c}^\top \mathbf{w} \\ \text{s.t.} \quad & \mathbf{H}\mathbf{w} \geq \mathbf{b};\ \ \mathbf{w} \geq \mathbf{0}. \end{aligned}$$

Dual:

$$\begin{aligned} \max \quad & \mathbf{b}^\top \mathbf{u} \\ \text{s.t.} \quad & \mathbf{H}^\top \mathbf{u} \geq \mathbf{c};\ \ \mathbf{u} \geq \mathbf{0}. \end{aligned}$$

# Extragradient Method for LP

- LP in standard form:

Primal:

$$\min \quad \mathbf{c}^\top \mathbf{w}$$
$$\text{s.t.} \quad \mathbf{H}\mathbf{w} \geq \mathbf{b}; \;\; \mathbf{w} \geq \mathbf{0}.$$

Dual:

$$\max \quad \mathbf{b}^\top \mathbf{u}$$
$$\text{s.t.} \quad \mathbf{H}^\top \mathbf{u} \geq \mathbf{c}; \;\; \mathbf{u} \geq \mathbf{0}.$$

- Solve LP by finding the saddle point of its Lagrange function:

$$\min_{\mathbf{w}\geq\mathbf{0}} \max_{\mathbf{u}\geq\mathbf{0}} \mathcal{L}(\mathbf{w},\mathbf{u}) = \mathbf{c}^\top\mathbf{w} + \mathbf{b}^\top\mathbf{u} - \mathbf{u}^\top\mathbf{H}\mathbf{w}$$

# Extragradient Method for LP

- LP in standard form:

Primal:

$$\min \quad \mathbf{c}^\top \mathbf{w}$$
$$\text{s.t.} \quad \mathbf{H}\mathbf{w} \geq \mathbf{b};\ \ \mathbf{w} \geq \mathbf{0}.$$

Dual:

$$\max \quad \mathbf{b}^\top \mathbf{u}$$
$$\text{s.t.} \quad \mathbf{H}^\top \mathbf{u} \geq \mathbf{c};\ \ \mathbf{u} \geq \mathbf{0}.$$

- Solve LP by finding the saddle point of its Lagrange function:

$$\min_{\mathbf{w}\geq\mathbf{0}} \max_{\mathbf{u}\geq\mathbf{0}} \mathcal{L}(\mathbf{w},\mathbf{u}) = \mathbf{c}^\top \mathbf{w} + \mathbf{b}^\top \mathbf{u} - \mathbf{u}^\top \mathbf{H}\mathbf{w}$$

- Update rules:

$$\begin{aligned}
\bar{\mathbf{w}}^k &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^k - \alpha(\mathbf{c} - \mathbf{H}^\top \mathbf{u}^k)) \\
\bar{\mathbf{u}}^k &= P_{\mathbf{u}\geq\mathbf{0}}(\mathbf{u}^k + \alpha(\mathbf{b} - \mathbf{H}\mathbf{w}^k)) \\
\mathbf{w}^k &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^k - \alpha(\mathbf{c} - \mathbf{H}^\top \bar{\mathbf{u}}^k)) \\
\mathbf{u}^k &= P_{\mathbf{u}\geq\mathbf{0}}(\mathbf{u}^k + \alpha(\mathbf{b} - \mathbf{H}\bar{\mathbf{w}}^k))
\end{aligned}$$

where step size $0 < \alpha < \|2\mathbf{H}\|_F^{-\frac{1}{2}}$.

# Extragradient Method for LP

- LP in standard form:

Primal:
$$\min \quad \mathbf{c}^\top \mathbf{w}$$
$$\text{s.t.} \quad \mathbf{H}\mathbf{w} \geq \mathbf{b}; \ \ \mathbf{w} \geq \mathbf{0}.$$

Dual:
$$\max \quad \mathbf{b}^\top \mathbf{u}$$
$$\text{s.t.} \quad \mathbf{H}^\top \mathbf{u} \geq \mathbf{c}; \ \ \mathbf{u} \geq \mathbf{0}.$$

- Solve LP by finding the saddle point of its Lagrange function:

$$\min_{\mathbf{w}\geq\mathbf{0}} \max_{\mathbf{u}\geq\mathbf{0}} \mathcal{L}(\mathbf{w}, \mathbf{u}) = \mathbf{c}^\top \mathbf{w} + \mathbf{b}^\top \mathbf{u} - \mathbf{u}^\top \mathbf{H}\mathbf{w}$$

- Update rules:

$$\begin{aligned}
\bar{\mathbf{w}}^k &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^k - \alpha(\mathbf{c} - \mathbf{H}^\top \mathbf{u}^k)) \\
\bar{\mathbf{u}}^k &= P_{\mathbf{u}\geq\mathbf{0}}(\mathbf{u}^k + \alpha(\mathbf{b} - \mathbf{H}\mathbf{w}^k)) \\
\mathbf{w}^k &= P_{\mathbf{w}\geq\mathbf{0}}(\mathbf{w}^k - \alpha(\mathbf{c} - \mathbf{H}^\top \bar{\mathbf{u}}^k)) \\
\mathbf{u}^k &= P_{\mathbf{u}\geq\mathbf{0}}(\mathbf{u}^k + \alpha(\mathbf{b} - \mathbf{H}\bar{\mathbf{w}}^k))
\end{aligned}$$

where step size $0 < \alpha < \|2\mathbf{H}\|_F^{-\frac{1}{2}}$.

- Converge geometrically.

# Extragradient Method for LP (Cont.)

- Apply to our problem, the Lagrange function is:

$$\min_{\mathbf{u}=(\mathbf{w},\boldsymbol{\xi})} \max_{\mathbf{v}=\boldsymbol{\lambda}} \quad \pi(\mathbf{u},\mathbf{v}) = \mathbf{1}^\top \mathbf{w} + C\mathbf{1}^\top \boldsymbol{\xi} + \boldsymbol{\lambda}^\top \mathbf{1} - \boldsymbol{\lambda}^\top \mathbf{M}\boldsymbol{\xi} - \boldsymbol{\lambda}^\top \mathbf{H}\mathbf{w}$$

$$\text{s.t.} \quad \mathcal{Q} = \{\mathbf{u} = (\mathbf{w},\boldsymbol{\xi}) | \mathbf{w} \geq \mathbf{0}, \boldsymbol{\xi} \geq \mathbf{0}\};$$

$$\mathcal{S} = \{\mathbf{v} = \boldsymbol{\lambda} | \boldsymbol{\lambda} \geq \mathbf{0}\}.$$

# Extragradient Method for LP (Cont.)

- Apply to our problem, the Lagrange function is:

$$\min_{\mathbf{u}=(\mathbf{w},\boldsymbol{\xi})} \max_{\mathbf{v}=\boldsymbol{\lambda}} \quad \pi(\mathbf{u},\mathbf{v}) = \mathbf{1}^\top \mathbf{w} + C\mathbf{1}^\top \boldsymbol{\xi} + \boldsymbol{\lambda}^\top \mathbf{1} - \boldsymbol{\lambda}^\top \mathbf{M}\boldsymbol{\xi} - \boldsymbol{\lambda}^\top \mathbf{H}\mathbf{w}$$

$$\text{s.t.} \quad \mathcal{Q} = \{\mathbf{u} = (\mathbf{w},\boldsymbol{\xi}) | \mathbf{w} \geq \mathbf{0}, \boldsymbol{\xi} \geq \mathbf{0}\};$$

$$\mathcal{S} = \{\mathbf{v} = \boldsymbol{\lambda} | \boldsymbol{\lambda} \geq \mathbf{0}\}.$$

- The corresponding update rules are:

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

# Extragradient Method with CG

**Algorithm** 2: Extragradient method with column generation

tolerances: $\epsilon_1$, $\epsilon_2$

$\mathbf{w}^0 \leftarrow \mathbf{w}$, $\boldsymbol{\xi}^0 \leftarrow \boldsymbol{\xi}$, $\boldsymbol{\lambda}^0 \leftarrow \boldsymbol{\lambda}$

for $i \leftarrow 1$ to $m$

  if $\mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}) < 1 - \xi_i$

    $\xi_i^0 \leftarrow (1 - \mathbf{w}^\top \Delta\phi(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}))$

    $\boldsymbol{\lambda}^0 \leftarrow \begin{pmatrix} \boldsymbol{\lambda}^0 \\ 0 \end{pmatrix}$

  end if

end for

iteratively update from $((\mathbf{w}^0, \boldsymbol{\xi}^0), \boldsymbol{\lambda}^0)$

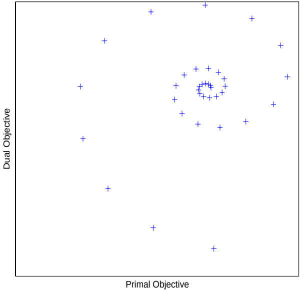
until $\frac{\|(\mathbf{w}^t, \boldsymbol{\xi}^t) - (\mathbf{w}^{t-1}, \boldsymbol{\xi}^{t-1})\|_2}{\|(\mathbf{w}^t, \boldsymbol{\xi}^t)\|_2} < \epsilon_1$ && $\frac{\|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-1}\|_2}{\|\boldsymbol{\lambda}^t\|_2} < \epsilon_1$

&& $0 < \|\mathbf{w}^t\|_1 + C\|\boldsymbol{\xi}^t\|_1 - \|\boldsymbol{\lambda}^t\|_1 < \epsilon_2$

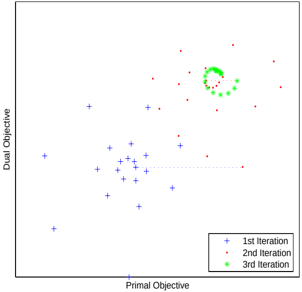
$\mathbf{w} \leftarrow \mathbf{w}^t$, $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi}^t$, $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}^t$

# Extragradient Method with CG (Cont.)

- Visualizations of the extragradient method and the CG process:




Extragradient method




Extragradient method with CG

# Experimental Results 1

- Task: part-of-speech tagging

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test
  - 5 splits

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test
  - 5 splits
- Implementation: C/C++

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test
  - 5 splits
- Implementation: C/C++
- Computing Environment:

# Experimental Results 1

- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test
  - 5 splits
- Implementation: C/C++
- Computing Environment:
  - 8×3.00GHz Intel(R) Xeon(R) CPU

# Experimental Results 1

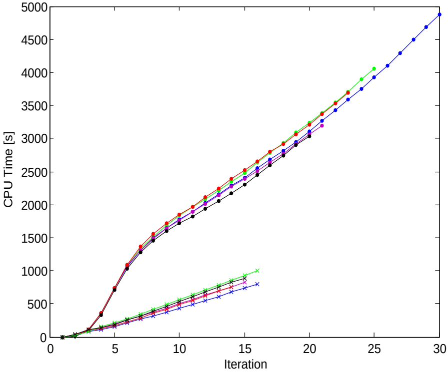
- Task: part-of-speech tagging
- Features: first-order HMM features
- Corpus:
  - 6700 manually tagged sentences from MEDLINE
  - 5700 for training
  - 1000 for test
  - 5 splits
- Implementation: C/C++
- Computing Environment:
  - 8×3.00GHz Intel(R) Xeon(R) CPU
  - 32GB RAM

# Experimental Results 1 (Cont.)

| Model | $Err_{all}$ | $Err_{voc}$ | # CPU Sec. | # Iteration |
|---|---|---|---|---|
| HMM | 20.02±0.29 | 14.44±0.19 | – | – |
| MIRA | 4.91±0.06 | 1.96±0.12 | 9084 | 46 |
| Perceptron | 5.38±0.19 | 2.10±0.07 | 26 | 100 |
| LP-Simplex | 4.94±0.18 | 1.96±0.14 | 3879 | 23 |
| LP-Xgrad | 4.92±0.13 | 1.98±0.12 | 856 | 14 |
| CRF | 4.58±0.14 | 1.81±0.19 | 51403 | 205 |

# Experimental Results 1 (Cont.)

- Dual-Simplex vs. Extragradient




• Dual-Simplex Method
× Extragradient Method

# Statistical Machine Translation

- More complex situations:

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence
  - Many paths in a word lattice may lead to a same translation

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence
  - Many paths in a word lattice may lead to a same translation
  - Correct translation may not be achieved by decoder

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence
  - Many paths in a word lattice may lead to a same translation
  - Correct translation may not be achieved by decoder
- Possible solutions:

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence
  - Many paths in a word lattice may lead to a same translation
  - Correct translation may not be achieved by decoder
- Possible solutions:
  - Taking each path $y$ as a potential multi-label output, but not the final translation $\mathbf{y}$

# Statistical Machine Translation

- More complex situations:
  - Many possible translations exist for a given source sentence
  - Many paths in a word lattice may lead to a same translation
  - Correct translation may not be achieved by decoder
- Possible solutions:
  - Taking each path $y$ as a potential multi-label output, but not the final translation $\mathbf{y}$
  - Using pseudo-references (with inner alignment structures) as positive examples

# More General Formulations

- Separating negative examples from closest positive examples (I):

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

# More General Formulations

- Separating negative examples from closest positive examples (I):

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

- Separating all negative examples from all positive examples (II):

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

# Experimental Results 2

- Task: purely-discriminative training for SMT

# Experimental Results 2

- Task: purely-discriminative training for SMT
- Corpus: Canada Hansard Senate Debates corpus

# Experimental Results 2

- Task: purely-discriminative training for SMT
- Corpus: Canada Hansard Senate Debates corpus
- Baseline system: Moses

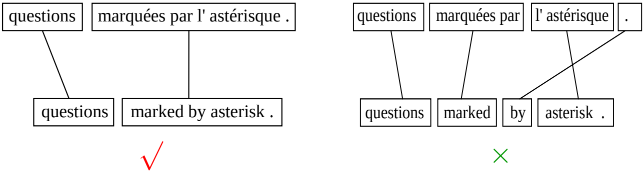
# Experimental Results 2

- Task: purely-discriminative training for SMT
- Corpus: Canada Hansard Senate Debates corpus
- Baseline system: Moses
- Features:

| Blanket Features | | | Discriminative Features | | |
|---|---|---|---|---|---|
| distortion log-prob. | 1 | | phrase distortions | 213,191 | |
| –orientation-based | | ×3 | –orientation-based | | ×3 |
| –forward-backward | | ×2 | –forward-backward | | ×2 |
| translation log-prob. | 1 | | phrase translations | 213,191 | |
| –bidirectional | | ×2 | –bidirectional | | ×2 |
| lexicon weight | 1 | | LM uni-grams | 78,400 | |
| –bidirectional | | ×2 | –backoff weights | 78,400 | |
| tri-gram LM log-prob. | 1 | | LM bi-grams | 1,544,378 | |
| word penalty | 1 | | –backoff weights | 1,544,378 | |
| phrase penalty | 1 | | LM tri-grams | 1,593,959 | |
| distortion distance | 1 | | | | |
| Total: | 14 | | Total: | 7,925,811 | |

# Experimental Results 2 (Cont.)

- Pseudo-reference extraction:
  - Decode top 10,000-best lists
  - Keep all paths yielding translations
  - Filter out those with bad inner alignments (open questions)
    - Artificial rules
    - Statistically significant tests



# Experimental Results 2 (Cont.)

- Results with all features

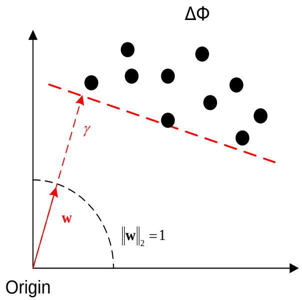
| | LP (I) | LP (II) | Baseline |
|---|---|---|---|
| BLEU (%) | 32.53 | 32.30 | 31.69 |
| NIST | 8.06 | 8.19 | 7.94 |

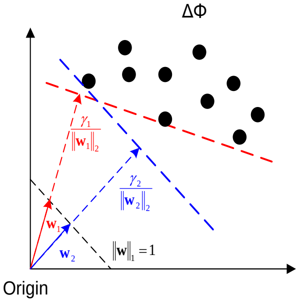
- Effects of different features

| LP (I): Blanket + | DLM | DTM | DLM+DTM | DD+DLM+DTM |
|---|---|---|---|---|
| BLEU (%) | 33.00 | 31.55 | 32.79 | 32.53 |
| NIST | 8.12 | 7.89 | 8.15 | 8.06 |
| LP (II): Blanket + | DLM | DTM | DLM+DTM | DD+DLM+DTM |
| BLEU (%) | 33.80 | 31.47 | 32.87 | 32.30 |
| NIST | 8.11 | 7.80 | 7.98 | 8.19 |

# Approximate Large-Margin Separation

- $L_2$-regularization vs. $L_1$-regularization:




Unit ball: $\|\mathbf{w}\|_2 = 1$ Unit simplex: $\|\mathbf{w}\|_1 = 1$

# Generalization Bound Analysis

- **Proposition 1.**

# Generalization Bound Analysis

- **Proposition 1.**
  - 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, $\{((\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}, 1) | \hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m \cup \{((\mathbf{x}_i, \hat{\mathbf{y}}, \mathbf{y}_i), -1) | \hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m$.

# Generalization Bound Analysis

- **Proposition 1.**
  - 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, $\{((\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}, 1)|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m \cup \{((\mathbf{x}_i, \hat{\mathbf{y}}, \mathbf{y}_i), -1)|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m$.
  - Suppose $\mathbf{w}$ parameterizes the optimal separating hyperplane passing through the origin for a labeled data set, $\{((\mathbf{x}_i, \mathbf{y}, \hat{\mathbf{y}}), z_i)|z_i \in \{-1, +1\}, i = 1, \ldots, m\}$, aligned such that $\mathbf{y} = \mathbf{y}_i, \hat{\mathbf{y}} \neq \mathbf{y}_i$ for $z_i = 1$, and $\mathbf{y} \neq \mathbf{y}_i, \hat{\mathbf{y}} = \mathbf{y}_i$ for $z_i = -1$. Then $\mathbf{w}$ parameterizes the supporting hyperplane for the unlabeled data set, $\{(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}})|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m$.

# Generalization Bound Analysis

- **Proposition 1.**
  - 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, $\{((\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}, 1)|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m \cup \{((\mathbf{x}_i, \hat{\mathbf{y}}, \mathbf{y}_i), -1)|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m$.
  - Suppose $\mathbf{w}$ parameterizes the optimal separating hyperplane passing through the origin for a labeled data set, $\{((\mathbf{x}_i, \mathbf{y}, \hat{\mathbf{y}}), z_i)|z_i \in \{-1, +1\}, i = 1, \ldots, m\}$, aligned such that $\mathbf{y} = \mathbf{y}_i, \hat{\mathbf{y}} \neq \mathbf{y}_i$ for $z_i = 1$, and $\mathbf{y} \neq \mathbf{y}_i, \hat{\mathbf{y}} = \mathbf{y}_i$ for $z_i = -1$. Then $\mathbf{w}$ parameterizes the supporting hyperplane for the unlabeled data set, $\{(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}})|\hat{\mathbf{y}} \neq \mathbf{y}_i\}_{i=1}^m$.
- **Definition 1.** Define the auxiliary inner product space:

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

  in which the inner product is given by $\langle f, g \rangle = \sum_{\mathbf{z} \in \text{supp}(f)} f(\mathbf{z}) g(\mathbf{z})$.

# Generalization Bound Analysis (Cont.)

- Embed our input spaceinto space $X \times L(X)$ using the mapping $\tau : (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \mapsto ((\mathbf{x}, \mathbf{y}), \frac{1}{C}\delta_{\hat{\mathbf{x}}})$ where $C > 0$ is a constant, and $\delta_{\hat{\mathbf{x}}} \in L(X)$ is defined to be:

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

# Generalization Bound Analysis (Cont.)

- Embed our input spaceinto space $X \times L(X)$ using the mapping $\tau : (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \mapsto ((\mathbf{x}, \mathbf{y}), \frac{1}{C}\delta_{\hat{\mathbf{x}}})$ where $C > 0$ is a constant, and $\delta_{\hat{\mathbf{x}}} \in L(X)$ is defined to be:

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

- For a function $(f, g) \in \mathcal{F} \times L(X)$, define its action on $\tau(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \in X \times L(X)$ as:

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

# Generalization Bound Analysis (Cont.)

- Embed our input spaceinto space $X \times L(X)$ using the mapping $\tau : (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \mapsto ((\mathbf{x}, \mathbf{y}), \frac{1}{C}\delta_{\hat{\mathbf{x}}})$ where $C > 0$ is a constant, and $\delta_{\hat{\mathbf{x}}} \in L(X)$ is defined to be:

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

- For a function $(f, g) \in \mathcal{F} \times L(X)$, define its action on $\tau(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \in X \times L(X)$ as:

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

- For a fixed margin $\gamma$, the slack variables $\xi_i$ in our LP problems can be derived from $\xi_i = \max(0, \gamma - \inf_{\hat{\mathbf{y}} \neq \mathbf{y}_i} f(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}))$.

# Generalization Bound Analysis (Cont.)

- Embed our input spaceinto space $X \times L(X)$ using the mapping $\tau : (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \mapsto ((\mathbf{x}, \mathbf{y}), \frac{1}{C}\delta_{\hat{\mathbf{x}}})$ where $C > 0$ is a constant, and $\delta_{\hat{\mathbf{x}}} \in L(X)$ is defined to be:

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

- For a function $(f, g) \in \mathcal{F} \times L(X)$, define its action on $\tau(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \in X \times L(X)$ as:

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

- For a fixed margin $\gamma$, the slack variables $\xi_i$ in our LP problems can be derived from $\xi_i = \max(0, \gamma - \inf_{\hat{\mathbf{y}} \neq \mathbf{y}_i} f(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}))$.
- Define $g_f = g(S, f, \gamma) \in L(\hat{X})$ to be $g_f = C \sum_{i=1}^{m} \xi_i \delta_{\mathbf{x}_i}$. It easy to check:

$$(f, g)(\tau(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})) = \begin{cases} f(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) + \xi_{\mathbf{x}} \geq \gamma & \forall\, (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \in S; \\ f(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) & \forall\, (\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \notin S. \end{cases}$$

# Generalization Bound Analysis (Cont.)

- **Theorem 2.** [Cristianini and Shawe-Taylor, 2000] Consider thresholding a real-valued function space $\mathcal{F}$ and fixed $\gamma \in \mathbb{R}^+$. 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 $(f, g_f) \in \mathcal{G} = \mathcal{F} \times L(X)$ has generalization error no more than

$$\mathrm{err}_{\mathcal{D}}(f) \leq \varepsilon(|S|, \mathcal{F}, \eta, \gamma) = \frac{2}{|S|}\left(\log_2 \mathcal{N}(\mathcal{G}, 2|S|, \frac{\gamma}{2}) + \log_2 \frac{2}{\eta}\right).$$

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

# Generalization Bound Analysis (Cont.)

- **Theorem 2.** [Cristianini and Shawe-Taylor, 2000] Consider thresholding a real-valued function space $\mathcal{F}$ and fixed $\gamma \in \mathbb{R}^+$. 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 $(f, g_f) \in \mathcal{G} = \mathcal{F} \times L(X)$ has generalization error no more than

$$\mathrm{err}_{\mathcal{D}}(f) \leq \varepsilon(|S|, \mathcal{F}, \eta, \gamma) = \frac{2}{|S|}\left(\log_2 \mathcal{N}(\mathcal{G}, 2|S|, \frac{\gamma}{2}) + \log_2 \frac{2}{\eta}\right).$$

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

- Based on our definition $\mathcal{F}(X) = \{f = \langle \mathbf{w}, \Delta\Phi(X)\rangle \, | \mathbf{w} \in \mathbb{R}^{d+}\}$ with respect to a given projection $\Delta\Phi : X \rightarrow \mathbb{R}^d$, the $L_1$-norm of $(f, g_f)$ is then given by:

$$\|(f, g_f)\|_1 = \|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \xi_i.$$

# Generalization Bound Analysis (Cont.)

- **Theorem 2.** [Cristianini and Shawe-Taylor, 2000] Consider thresholding a real-valued function space $\mathcal{F}$ and fixed $\gamma \in \mathbb{R}^+$. 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 $(f, g_f) \in \mathcal{G} = \mathcal{F} \times L(X)$ has generalization error no more than

$$\mathrm{err}_{\mathcal{D}}(f) \leq \varepsilon(|S|, \mathcal{F}, \eta, \gamma) = \frac{2}{|S|}\left(\log_2 \mathcal{N}(\mathcal{G}, 2|S|, \frac{\gamma}{2}) + \log_2 \frac{2}{\eta}\right).$$

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

- Based on our definition $\mathcal{F}(X) = \{f = \langle \mathbf{w}, \Delta\Phi(X)\rangle \mid \mathbf{w} \in \mathbb{R}^{d+}\}$ with respect to a given projection $\Delta\Phi : X \rightarrow \mathbb{R}^d$, the $L_1$-norm of $(f, g_f)$ is then given by:

$$\|(f, g_f)\|_1 = \|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \xi_i.$$

- **Corollary 3.** (Zhang, 2002) If $\max\{\|\Delta\Phi(X)\|_\infty, \frac{1}{C}\} \leq b$ and $\|\mathbf{w}\|_1 + C\sum_{i=1}^{m}\xi_i \leq c$, for the function class $\mathcal{G} = \mathcal{F} \times L(X)$ defined above, we have that

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

# Conclusion

- Advantages:

# Conclusion

- Advantages:
  - Accepting arbitrary structures

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties
- Drawbacks:

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties
- Drawbacks:
  - Sensitive to pseudo-references in the SMT case

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties
- Drawbacks:
  - Sensitive to pseudo-references in the SMT case
  - Force the solution to be too sparse sometimes

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties
- Drawbacks:
  - Sensitive to pseudo-references in the SMT case
  - Force the solution to be too sparse sometimes
- ANSI C code for extragradient LP solver is available at: http://www.cs.ucl.ac.uk/staff/z.wang/

# Conclusion

- Advantages:
  - Accepting arbitrary structures
  - More efficient than QP-based methods
  - More accurate than perceptron
  - Nice generalization properties
- Drawbacks:
  - Sensitive to pseudo-references in the SMT case
  - Force the solution to be too sparse sometimes
- ANSI C code for extragradient LP solver is available at: http://www.cs.ucl.ac.uk/staff/z.wang/
- For reference, see:
  Z. Wang & J. Shawe-Taylor (2009). Large-Margin Structured Prediction via Linear Programming. In *AISTATS 2009*. USA.

# Thank you!

Questions?