Structured Prediction 2

Joseph Keshet
Dept. of Computer Science
Bar-Ilan University
Goal

Find a function such that

$$\mathbb{E}[L(y, \hat{y})]$$

is minimized.

- $\hat{y}$ is the predicted sequence (label)
- the expectation is taken over a random draw of $(x, y)$
\[ \hat{y}_w = \arg \max_y w \cdot \phi(x, y) \]

\[ w^* = \arg \min_w \mathbb{E} [L(y, \hat{y}_w)] \]

- \( \hat{y} \) is the predicted sequence (label)
- the expectation is taken over a random draw of \((x, y)\)
Structured Perceptron
Structured Perceptron

\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) \]

\[ w \leftarrow w + \phi(x_i, y_i) - \phi(x_i, \hat{y}) \]
Structured Support Vector Machines (SVMs)
Structured SVM

\[ \mathbf{w}^* = \arg \min_{\mathbf{w}} \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_\mathbf{w}(x))] \]
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]
Structured SVM

\[ \mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \left[ L(\mathbf{y}, \hat{\mathbf{y}}_{\mathbf{w}}(\mathbf{x})) \right] \]

\[ \mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_{\mathbf{w}}(x_i)) + \frac{\lambda}{2} \| \mathbf{w} \|^2 \]

\[ \leq \arg\min_{\mathbf{w}} \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{hinge}}(y_i, \hat{y}_{\mathbf{w}}(x_i)) + \frac{\lambda}{2} \| \mathbf{w} \|^2 \]
Structured SVM

\[ w^* = \operatorname{arg\,min}_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \operatorname{arg\,min}_w \frac{1}{M} \sum_{i=1}^M L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \leq \operatorname{arg\,min}_w \frac{1}{M} \sum_{i=1}^M \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

surrogate hinge-loss
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} \left[ L(y, \hat{y}_w(x)) \right] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \leq \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

surrogate hinge-loss

\[ \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) = \max_{\hat{y}} \left[ L(y_i, \hat{y}) - w \cdot \phi(x_i, y_i) + w \cdot \phi(x_i, \hat{y}) \right] \]
Structured SVM

\[ \hat{y}_L = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) + L(y_i, \hat{y}) \]

\[ w \leftarrow w(1 - \eta \lambda) + \eta \left[ \phi(x_i, y_i) - \phi(x_i, \hat{y}_L) \right] \]
Structured SVM

Margin generalization bound

$$\mathbb{E}[L(y, \hat{y})] \leq \frac{1}{M} \sum_{i} \max_{y : w \cdot \phi(x_i, y_i) - w \cdot \phi(x_i, y) \leq H(y, y_i)} L(y_i, y) + O\left(\frac{1}{\sqrt{M}}, \sqrt{\|w\|}, \frac{1}{\sqrt{\delta}}, \ldots\right)$$

This involves both the Hamming distance (as a margin requirement) and the evaluation metric function (Taskar et al. 2003; McAllester 2007)

No guarantee for the risk (expected measure of performance)
Conditional Random Fields
(CRFs)
$w^* = \arg \min_w \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))]$
\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]
\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[
 w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} - \log P_w(y_i|x_i) + \frac{\lambda}{2} \|w\|^2
\]

where \( P_w(y|x) = \frac{1}{Z} \exp(w \cdot \phi(x, y)) \)
$$w \leftarrow w(1 - \eta \lambda) + \eta \left[ \phi(x_i, y_i) - \mathbb{E}_{\hat{y} \sim P_w(y|x_i)} [\phi(x_i, \hat{y})] \right]$$
\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} - \log P_w(y_i | x_i) + \frac{\lambda}{2} \| w \|^2 \]

CRF ignores the measure of performance
Direct loss minimization
Direct loss minimization

\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) \]

\[ \hat{y}_{\epsilon L} = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) + \epsilon L(y_i, \hat{y}) \]

\[ w \gets w + \frac{\eta}{\epsilon} [\phi(x_i, \hat{y}) - \phi(x_i, \hat{y}_{\epsilon L})] \]

(McAllester, Hazan, and Keshet, 2010)
Direct loss minimization

\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) \]

\[ \hat{y}_{\epsilon L} = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) + \epsilon L(y_i, \hat{y}) \]

\[ w \leftarrow w + \frac{\eta}{\epsilon} \left[ \phi(x_i, \hat{y}) - \phi(x_i, \hat{y}_{\epsilon L}) \right] \]

(McAllester, Hazan, and Keshet, 2010)
Goal

\[ \mathbf{w}^* = \operatorname{arg\,min}_w \mathbb{E} [L(y, \hat{y}_w)] \]

(McAllester, Hazan, and Keshet, 2010)
Goal

\[ w^* = \arg \min_w \mathbb{E} [L(y, \hat{y}_w)] \]

minimize the expectation by finding the gradient \( \nabla_w \)

(McAllester, Hazan, and Keshet, 2010)
Theorem

Let

$$\hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x, \hat{y})$$

$$\hat{y}_{\epsilon L} = \arg \max_{\hat{y}} \ w \cdot \phi(x, \hat{y}) + \epsilon L(y, \hat{y})$$

then

$$\nabla_w \mathbb{E}[L(y, \hat{y})] = \lim_{\epsilon \to 0} \frac{\mathbb{E}[\phi(x, \hat{y}_{\epsilon L}) - \phi(x, \hat{y})]}{\epsilon}$$

(McAllester, Hazan, and Keshet, 2010)
\[ \Delta w \cdot \nabla_w \mathbb{E} [L(y, \hat{y}_w)] = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E} [L(y, \hat{y}_w + \varepsilon \Delta w) - L(y, \hat{y}_w)] \]
\[
\frac{d}{dx} f(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon \Delta x) - f(x)}{\epsilon \Delta x}
\]

**definition of the loss gradient**

\[
\Delta w \cdot \nabla_w \mathbb{E} [L(y, \hat{y}_w)] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E} [L(y, \hat{y}_w + \epsilon \Delta w) - L(y, \hat{y}_w)]
\]
\[ \frac{d}{dx} f(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon \Delta x) - f(x)}{\epsilon \Delta x} \]

\[ \Delta \mathbf{w} \cdot \nabla_{\mathbf{w}} \mathbb{E}[L(\mathbf{y}, \hat{\mathbf{y}}_{\mathbf{w}})] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}[L(\mathbf{y}, \hat{\mathbf{y}}_{\mathbf{w}} + \epsilon \Delta \mathbf{w}) - L(\mathbf{y}, \hat{\mathbf{y}}_{\mathbf{w}})] \]

definition of the loss gradient
determined by an integral over the decision boundary
The definition of the loss gradient is:

\[
\frac{d}{dx} f(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon \Delta x) - f(x)}{\epsilon \Delta x}
\]

The theorem determined by an integral over the decision boundary is:

\[
\Delta w \cdot \nabla_w \mathbb{E}[L(y, \hat{y}_w)] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}[L(y, \hat{y}_w + \epsilon \Delta w) - L(y, \hat{y}_w)]
\]

\[
\nabla_w \mathbb{E}[L(y, \hat{y}_w)] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}[\phi(x, \hat{y}_{\epsilon L}) - \phi(x, \hat{y}_w)]
\]
The theorem determined by an integral over the decision boundary is:

\[
\nabla_w \mathbb{E}[L(y, \hat{y}_w)] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}[\phi(x, \hat{y}_{\epsilon L}) - \phi(x, \hat{y}_w)]
\]

The definition of the loss gradient is:

\[
\frac{d}{dx} f(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon \Delta x) - f(x)}{\epsilon \Delta x}
\]
To show that the two integrals are the same:

\[ w \cdot [\phi(x, y_1) - \phi(x, y_2)] = 0 \]

\[ [L(y, y_2) - L(y, y_1)] \]

\[ \hat{y}_w = y_1 \]

\[ \hat{y}_w + \epsilon \Delta w = y_2 \]

\[ \epsilon \Delta w \cdot [\phi(x, y_1) - \phi(x, y_2)] \]

\[ \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E} [L(y, \hat{y}_w + \epsilon \Delta w) - L(y, \hat{y}_w)] \]
To show that the two integrals are the same:

\[ w \cdot [\phi(x, y_1) - \phi(x, y_2)] = 0 \]

\[ \Delta w \cdot [\phi(x, y_1) - \phi(x, y_2)] \]

\[ \hat{y}_w = y_1 \]
\[ \hat{y}_{\epsilon L} = y_2 \]

\[ \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Delta w \mathbb{E}[\phi(x, \hat{y}_{\epsilon L}) - \phi(x, \hat{y}_w)] \]
To show that the two integrals are the same:

\[ \Delta w \cdot [\phi(x, y_1) - \phi(x, y_2)] \cdot [L(y, y_2) - L(y, y_1)] \]

both equal this
Algorithm

**Input:** training set \( S = \{(x_1, y_1), \ldots, (x_M, y_M)\} \)

**Initialize:** \( w_0 = 0 \)

**For** each example \((x_m, y_m)\)

**Predict:** \( \hat{y} = \underset{\hat{y}}{\text{arg max}} \ w_{m-1} \cdot \phi(x_m, \hat{y}) \)

**Predict:** \( \hat{y}_L = \underset{\hat{y}_L}{\text{arg max}} \ w_{m-1} \cdot \phi(x_m, \hat{y}_L) + \epsilon L(y_m, \hat{y}_L) \)

**Update:** \( w_m = w_{m-1} - \eta \frac{\phi(x_m, \hat{y}_L) - \phi(x_m, \hat{y})}{\epsilon} \)

**Output** Choose \( w_m \) which attains the lowest cost on a held-out set.
Direct loss minimization

\[ w \leftarrow w - \eta \left[ \nabla_w \mathbb{E}[L(y, \hat{y}_w(x))] \right]_i \]

\[ w \leftarrow w + \phi(x_i, \hat{y}) - \phi(x_i, \hat{y}_{\epsilon L}) \]

(McAllester, Hazan, and Keshet, 2010)
Structured ramp loss
DIRECT RISK MINIMIZATION
DIRECT RISK MINIMIZATION

+ 

REGULARIZATION
DIRECT RISK MINIMIZATION + REGULARIZATION =
DIRECT RISK MINIMIZATION + REGULARIZATION = ramp loss
Ramp loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))] \]

(McAllester, and Keshet, 2011)
Ramp loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

(McAllester, and Keshet, 2011)
Ramp loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{ramp}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \|w\|^2 \]

(McAllester, and Keshet, 2011)
Ramp loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{ramp}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \hat{L}_{ramp}(y_i, \hat{y}_w(x_i)) = \max_{\hat{y}} [L(y_i, \hat{y}) - w \cdot \phi(x_i, \hat{y})] + \max_{\hat{y}} [w \cdot \phi(x_i, \hat{y})] \]

(McAllester, and Keshet, 2011)
Motivation

Direct cost update

\[ w_m = w_{m-1} - \eta \frac{1}{\epsilon} \left[ \phi(x_m, \hat{y}_L) - \phi(x_m, \hat{y}) \right] \]

Structured ramp loss

\[ w^* = \arg\min_w \frac{1}{2} \|w\|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} \left[ w \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) \right] - \max_{\hat{y}} \left[ w \cdot \phi(x_m, \hat{y}) \right] \]

Structured SVM

\[ w^* = \arg\min_w \frac{1}{2} \|w\|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} \left[ w \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) \right] - w \cdot \phi(x_m, y_m) \]

(McAllester and Keshet, 2012)
Motivation

Direct cost update

\[
\mathbf{w}_m = \mathbf{w}_{m-1} - \eta \frac{1}{\epsilon} \left[ \phi(x_m, \hat{y}_L) - \phi(x_m, \hat{y}) \right]
\]

Structured ramp loss

\[
\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} [ \mathbf{w} \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) ] - \max_{\hat{y}} [ \mathbf{w} \cdot \phi(x_m, \hat{y}) ]
\]

Structured SVM

\[
\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} [ \mathbf{w} \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) ] - \mathbf{w} \cdot \phi(x_m, y_m)
\]

(McAllester and Keshet, 2012)
Motivation

Direct cost update

\[
\mathbf{w}_m = \mathbf{w}_{m-1} - \eta \frac{1}{\epsilon} \left[ \phi(x_m, \hat{y}_L) - \phi(x_m, \hat{y}) \right]
\]

Structured ramp loss

\[
\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} \left[ \mathbf{w} \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) \right] - \max_{\hat{y}} \left[ \mathbf{w} \cdot \phi(x_m, \hat{y}) \right]
\]

Structured SVM

\[
\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{M} \sum_m \max_{\hat{y}_L} \left[ \mathbf{w} \cdot \phi(x_m, \hat{y}_L) + L(\hat{y}_L, y_m) \right] - \mathbf{w} \cdot \phi(x_m, y_m)
\]

(McAllester and Keshet, 2012)
Ramp loss

\[ \hat{y} = \arg\max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) \]

\[ \hat{y}_L = \arg\max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) + L(y_i, \hat{y}) \]

\[ \Delta w \propto \phi(x_i, \hat{y}) - \phi(x_i, \hat{y}_L) - \lambda w \]

(McAllester, and Keshet, 2011)
Ramp loss

\[ \hat{y} = \arg \max_{\hat{y}} w \cdot \phi(x_i, \hat{y}) \]

\[ \hat{y}_L = \arg \max_{\hat{y}} w \cdot \phi(x_i, \hat{y}) + L(y_i, \hat{y}) \]

\[ \Delta w \propto \phi(x_i, \hat{y}) - \phi(x_i, \hat{y}_L) - \lambda w \]

(McAllester, and Keshet, 2011)
Consistency

\[ \lim_{m \to \infty} L_{\text{ramp}}(w_m) = \inf_w \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))] \]

(McAllester, and Keshet, 2011)
Structured probit loss
Structured probit loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

(Keshet, McAllester, Hazan, 2011)
Structured probit loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

(Keshet, McAllester, Hazan, 2011)
Structured probit loss

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{probit}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

(Keshet, McAllester, Hazan, 2011)
Structured probit loss

\[ w^* = \arg \min_w \mathbb{E}_{(x, y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{probit}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \hat{L}_{\text{probit}}(y_i, \hat{y}_w(x_i)) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} [L(y_i, \hat{y}_w + \epsilon(x_i))] \]

(Keshet, McAllester, Hazan, 2011)
Update rule

\[ \nabla_w \left[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon \left[ L(y_i, f_w + \epsilon(x_i)) \right] + \frac{\lambda}{2} \| w \|^2 \right] \]

= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon \left[ \epsilon L(y_i, f_w + \epsilon(x_i)) \right] + \lambda w

\[ \Delta w \propto -\eta \mathbb{E}_\epsilon \left[ \epsilon L(y_i, \hat{y}_w + \epsilon(x_i)) \right] - \lambda w \]

(Keshet, McAllester, Hazan, 2011)
Update rule

\[ \nabla_w \left[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon [L(y_i, f_{w+\epsilon}(x_i))] + \frac{\lambda}{2} \|w\|^2 \right] \]

\[ = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon [\epsilon L(y_i, f_{w+\epsilon}(x_i))] + \lambda w \]

\[ \Delta w \propto -\eta \mathbb{E}_\epsilon [\epsilon L(y_i, \hat{y}_{w+\epsilon}(x_i))] - \lambda w \]

the loss does not need to be separable

(Keshet, McAllester, Hazan, 2011)
**Update rule**

**phoneme error rate**

\[
\Delta \mathbf{w} \propto -\eta \mathbb{E}_\epsilon \left[ \epsilon \mathcal{L}(\mathbf{y}_i, \hat{\mathbf{y}}_{\mathbf{w}+\epsilon(\mathbf{x}_i)}) \right] - \lambda \mathbf{w}
\]

The loss does not need to be separable

(Keshet, McAllester, Hazan, 2011)
**Update rule**

Pascal VOC

Intersection/union = \[
\frac{\text{true pos class}}{\text{true pos + false pos} + \text{false neg}}
\]

\[\Delta w \propto -\eta \mathbb{E}_\epsilon [\epsilon L(y_i, \hat{y}_{w+\epsilon}(x_i))] - \lambda w\]

the loss does not need to be separable

(Keshet, McAllester, Hazan, 2011)
Update rule

phoneme error rate

\[ \Delta w \propto -\eta \mathbb{E}_\epsilon [\epsilon L(y_i, \hat{y}_{w+\epsilon(x_i)})] - \lambda w \]

the loss does not need to be separable

(Keshet, McAllester, Hazan, 2011)
Update rule

\[ \nabla_w \left[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon \left[ L(y_i, f_{w+\epsilon}(x_i)) \right] + \frac{\lambda}{2} \|w\|^2 \right] \]

\[ = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\epsilon \left[ \epsilon \cdot L(y_i, f_{w+\epsilon}(x_i)) \right] + \lambda w \]

\[ \Delta w \propto -\eta \mathbb{E}_\epsilon \left[ \epsilon \cdot L(y_i, \hat{y}_{w+\epsilon}(x_i)) \right] - \lambda w \]

the loss does not need to be separable

(Keshet, McAllester, Hazan, 2011)
Theorem
With probability at least $1 - \delta$ over the draw of the sample we have the following simultaneously for all $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \leq \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m - 1)} \|w\|^2 + \frac{\lambda}{m - 1} \ln \frac{m}{\delta}$$

(Keshet, McAllester, Hazan, 2011)
Theorem

With probability at least $1 - \delta$ over the draw of the sample we have the following simultaneously for all $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \leq \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m-1)} \|w\|^2 + \frac{\lambda}{m-1} \ln \frac{m}{\delta}$$

(Keshet, McAllester, Hazan, 2011)
With probability at least $1 - \delta$ over the draw of the sample we have the following simultaneously for all $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \leq \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m - 1)} \|w\|^2 + \frac{\lambda}{m - 1} \ln \frac{m}{\delta}$$

unseen data

training data

(Keshet, McAllester, Hazan, 2011)
PAC-Bayesian Theorem

Theorem
With probability at least $1 - \delta$ over the draw of the sample we have the following simultaneously for all $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \lesssim \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m-1)} \|w\|^2 + \frac{\lambda}{m-1} \ln \frac{m}{\delta}$$

(Keshet, McAllester, Hazan, 2011)
Theorem
With probability at least $1 - \delta$ over the draw of the sample we have the following \textit{simultaneously for all} $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \leq \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m - 1)} \|w\|^2 + \frac{\lambda}{m - 1} \ln \frac{m}{\delta}$$

(Keshet, McAllester, Hazan, 2011)
PAC-Bayesian Theorem

Theorem
With probability at least $1 - \delta$ over the draw of the sample we have the following simultaneously for all $w$ and for all $\lambda$

$$L_{\text{probit}}(w) \leq \hat{L}_{\text{probit}}(w) + \frac{\lambda}{2(m - 1)} \|w\|^2 + \frac{\lambda}{m - 1} \ln \frac{m}{\delta}$$

minimized by the gradient $\nabla_w$

(Keshet, McAllester, Hazan, 2011)
Consistency

\[
\lim_{m \to \infty} L_{\text{probit}}(w_m) = \inf_{w} \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))]
\]

(Keshet, McAllester, Hazan, 2011)
Extensions

- Different regularization: $\ell_2$, $\ell_1$, more
- Graphical models (HMMs)
- Different perturbation probabilities

\[
\Theta \leftarrow (1 - \eta) \Theta + \eta \lambda \mathbb{E}_\epsilon \left[ \nabla_\epsilon \log \pi(\epsilon) \cdot L(y_i, f_{w+\epsilon}(x_i)) \right]
\]

(for $\pi(\epsilon) = \mathcal{N}(\epsilon; 0, I)$ we have $\nabla_\epsilon \log \pi(\epsilon) = -\epsilon$)

(Hazan, Keshet, Jaakkola, 2013; Hazan, Maji, Keshet, and Jaakkola, 2013)
Extensions

- Different regularization: $\ell_2$, $\ell_1$, more
- Graphical models (HMMs)
- Different perturbation probabilities

\[ \Theta \leftarrow (1 - \eta) \Theta + \eta \lambda \mathbb{E}_{\epsilon} \left[ \nabla_{\epsilon} \log \pi(\epsilon) \cdot L(y_i, f_{w+\epsilon}(x_i)) \right] \]

(for $\pi(\epsilon) = \mathcal{N}(\epsilon; 0, I)$ we have $\nabla_{\epsilon} \log \pi(\epsilon) = -\epsilon$)

(Hazan, Keshet, Jaakkola, 2013; Hazan, Maji, Keshet, and Jaakkola, 2013)
Extensions

- Different regularization: $\ell_2, \ell_1, more$
- Graphical models (HMMs)
- Different perturbation probabilities

$$\Theta \leftarrow (1 - \eta)\Theta + \eta \lambda \mathbb{E}_{\epsilon} \left[ \nabla_{\epsilon} \log \pi(\epsilon) \cdot L(y_i, f_{\mathbf{x}_i + \epsilon}(\mathbf{x}_i)) \right]$$

(for $\pi(\epsilon) = \mathcal{N}(\epsilon; 0, I)$ we have $\nabla_{\epsilon} \log \pi(\epsilon) = -\epsilon$)

(Hazan, Keshet, Jaakkola, 2013; Hazan, Maji, Keshet, and Jaakkola, 2013)
Area Under the ROC Curve (AUC) maximization
Given a speech signal

and a word bought

**Goal**: find if the word is uttered in the speech signal and where
Given a speech signal and a word: bought

**Goal**: find if the word is uttered in the speech signal and where.
\[ \bar{t} = (t_{\text{start}}, t_{\text{end}}) \]

\[ k = \]

\[ \bar{x} = (x_1, x_2, \ldots, \ldots, x_T) \]
speech input signal

\[ k = /bought/ \]

input keyword

Keyword Spotter

\[ f(\bar{x}, k) \]

detection (yes/no)

predicted decision

predicted timing

Tuesday, May 27, 14
speech input signal

input keyword

predicted decision

predicted timing

\[ k = /bought/ \]

\[ \bar{x} \]

\[ \max_{\bar{t}} f(\bar{x}, k, \bar{t}) \]

\[ \bar{t}' \]

detection (yes/no)
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive = \[ \frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}} \]

false positive = \[ \frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}} \]
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive = \[ \frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}} \]

false positive = \[ \frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}} \]

false positive rate

true positive rate

area under curve $A$
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive = \frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}

false positive = \frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}

A = 1
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive = \[\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}\]

false positive = \[\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}\]

A = 0.5
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive = \[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

false positive = \[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]
Common solution I
Common solution I

• Training:
Common solution I

• Training:
  – a model for the event bought
Common solution 1

• Training:
  - a model for the event
  - a model for the world
Common solution I

• Training:
  – a model for the event
  – a model for the world

• Inference:
Common solution I

• Training:
  – a model for the event
  – a model for the world

• Inference:
  – compare + bought +
Common solution I

• Training:
  – a model for the event
  – a model for the world

• Inference:
  – compare + bought
  – with
Common solution II
Common solution II

- Training:
Common solution II

- Training:
  - models for the basic elements
    - aa
    - ae
    - ...
    - zh
Common solution II

• Training:
  – models for the basic elements aa ae ... zh

• Inference:
Common solution II

- **Training:**

  - models for the basic elements: aa, ae, ..., zh

- **Inference:**

  - compare: * + bcl b ao t + *

Tuesday, May 27, 14
Common solution II

• Training:
  – models for the basic elements  
    {aa, ae, ... zh}

• Inference:
  – compare  
    * + bcl b ao t + *
  – with *

Tuesday, May 27, 14
Common solution

• There is no algorithm for event spotting (the length of the input signal is not fixed as well as the length of the input event)

• There is no algorithm to maximize the area under the ROC curve
Goal

Find the parameters $w$ of the function

$$\max_{t} f_{w}(\bar{x}, k, \bar{t})$$

such that the expected area under the ROC curve (AUC) is maximized.

We assume there is a fixed, but unknown distribution from which $(\bar{x}, k, \bar{t}, \text{yes/no})$ are drawn.
Maximizing area under ROC (AUC)

For every event (keyword) $k$ define two sets of input signals (speech utterances):

$\mathcal{X}_k^+$

$\mathcal{X}_k^-$
Maximizing area under ROC (AUC)

By definition of the area under the ROC (Bamber, 1975; Hanley and McNeil, 1982):

\[
A = \mathbb{P} \left[ \max_{\mathcal{t}} f_w(\bar{x}^+, k, \mathcal{t}) > \max_{\mathcal{t}} f_w(\bar{x}^-, k, \mathcal{t}) \right]
\]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

By definition of the area under the ROC (Bamber, 1975; Hanley and McNeil, 1982):

\[
A = \mathbb{P} \left[ \max_t f_w(\bar{x}^+, k, \bar{t}) > \max_t f_w(\bar{x}^-, k, \bar{t}) \right]
\]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

By definition of the area under the ROC (Bamber, 1975; Hanley and McNeil, 1982):

\[
A = \mathbb{P} \left[ \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right]
\]

\[
\mathbf{w}^* = \arg \max_{\mathbf{w}} \mathbb{P} \left[ \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right]
\]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

By definition of the area under the ROC (Bamber, 1975; Hanley and McNeil, 1982):

$$A = \mathbb{P} \left[ \max_t f_w(\bar{x}^+, k, \bar{t}) > \max_t f_w(\bar{x}^-, k, \bar{t}) \right]$$

$$w^* = \arg \max_w \mathbb{P} \left[ \max_t f_w(\bar{x}^+, k, \bar{t}) > \max_t f_w(\bar{x}^-, k, \bar{t}) \right]$$

$$= \arg \max_w \mathbb{E} \left[ \delta \left( \max_t f_w(\bar{x}^+, k, \bar{t}) > \max_t f_w(\bar{x}^-, k, \bar{t}) \right) \right]$$

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_{w} \mathbb{E} \left[ \delta \left[ \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right] \right] \]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_{w} \mathbb{E} \left[ \delta \left( \max_{\tilde{t}} f_w(\tilde{x}^+, k, \tilde{t}) > \max_{\tilde{t}} f_w(\tilde{x}^-, k, \tilde{t}) \right) \right] \]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_w \mathbb{E} \left[ \delta \left[ \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right] \right] \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \delta \left[ \max_{\bar{t}} f_w(\bar{x}_{i}^+, k_i, \bar{t}) < \max_{\bar{t}} f_w(\bar{x}_{i}^-, k_i, \bar{t}) \right] + \frac{\lambda}{2} \| w \|^2 \]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_w \mathbb{E} \left[ \delta \left( \max_{\tilde{t}} f_w(\bar{x}^+, k, \tilde{t}) > \max_{\tilde{t}} f_w(\bar{x}^-, k, \tilde{t}) \right) \right] \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \delta \left( \max_{\tilde{t}} f_w(\bar{x}_i^+, k_i, \tilde{t}) < \max_{\tilde{t}} f_w(\bar{x}_i^-, k_i, \tilde{t}) \right) + \frac{\lambda}{2} \| w \|^2 \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \left[ 1 - f_w(\bar{x}_i^+, k_i, \tilde{t}_i^+) + \max_{\tilde{t}} f_w(\bar{x}_i^-, k_i, \tilde{t}) \right] + \frac{\lambda}{2} \| w \|^2 \]

(Keshet, Grangier and Bengio, 2009)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_w \mathbb{E} \left[ \delta \left( \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right) \right] \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \delta \left( \max_{\bar{t}} f_w(\bar{x}_i^+, k_i, \bar{t}) < \max_{\bar{t}} f_w(\bar{x}_i^-, k_i, \bar{t}) \right) + \frac{\lambda}{2} \| w \|^2 \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \left[ 1 - f_w(\bar{x}_i^+, k_i, t_i^+) + \max_{\bar{t}} f_w(\bar{x}_i^-, k_i, \bar{t}) \right] + \frac{\lambda}{2} \| w \|^2 \]

\( (\text{Keshet}, \text{Grangier and Bengio, 2009}) \)
Maximizing area under ROC (AUC)

\[ w^* = \arg \max_w \mathbb{E} \left[ \delta \left( \max_{\bar{t}} f_w(\bar{x}^+, k, \bar{t}) > \max_{\bar{t}} f_w(\bar{x}^-, k, \bar{t}) \right) \right] \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \delta \left( \max_{\bar{t}} f_w(\bar{x}^+_i, k_i, \bar{t}) < \max_{\bar{t}} f_w(\bar{x}^-_i, k_i, \bar{t}) \right) + \frac{\lambda}{2} \| w \|^{2} \]

\[ w^* = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \left[ 1 - f_w(\bar{x}^+_i, k_i, \bar{t}^+_i) + \max_{\bar{t}} f_w(\bar{x}^-_i, k_i, \bar{t}) \right] + \frac{\lambda}{2} \| w \|^{2} \]

(Keshet, Grangier and Bengio, 2009)
Clean read speech (TIMIT)

Area under the ROC curve:
0.99 discriminative
0.94 HMM
Read newspaper (WSJ)

Area under the ROC curve:
- 0.94 discriminative
- 0.87 HMM
# Other datasets

<table>
<thead>
<tr>
<th>Dataset Description</th>
<th>HMM</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT (clean, read)</td>
<td>0.941</td>
<td>0.996</td>
</tr>
<tr>
<td>HTIMIT (noisy, read)</td>
<td>0.922</td>
<td>0.949</td>
</tr>
<tr>
<td>WSJ (clean, read)</td>
<td>0.870</td>
<td>0.942</td>
</tr>
<tr>
<td>OGI Stories (noisy)</td>
<td>0.722</td>
<td>0.769</td>
</tr>
<tr>
<td>HUMAINE (accented, spontaneous, and emotionally colored)</td>
<td>0.63</td>
<td>0.800</td>
</tr>
<tr>
<td>Switchboard (phone, conversational)</td>
<td>0.920</td>
<td>0.938</td>
</tr>
</tbody>
</table>

(Keshet, Grangier and Bengio, 2009; Wöllmer, Eyben, Keshet, Graves, Schuller, and Rigoll, 2009; Prabhavalkar, Keshet, Livescu and Fosler-Lussier, 2012)