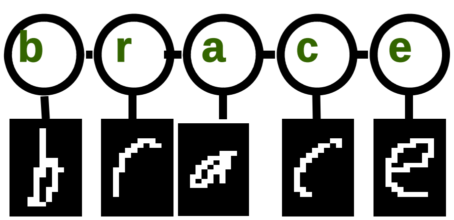
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Max-margin learning of GM

Eric Xing

Lecture 23, Apr 8, 2015



Reading:

Classical Predictive Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\quad\quad \mathcal{Y} \triangleq \{-1, +1\}$
- Predictive function $h(\mathrm{x})$: $y^\star = h(\mathrm{x}) \triangleq \arg\max_{y \in \mathcal{Y}} F(\mathrm{x}, y; \mathbf{w})$
- Examples: $F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$
- Learning: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$

where $\ell(\cdot)$ represents a convex loss, and $R(\mathbf{w})$ is a regularizer preventing overfitting

– **Logistic Regression**
  - **Max-likelihood (or MAP) estimation**

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(y^i | \mathrm{x}^i; \mathbf{w}) + \mathcal{N}(\mathbf{w})$$

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

– **Support Vector Machines (SVM)**
  - **Max-margin learning**

$$\min_{\mathbf{w}, \xi} \quad \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^{N} \xi_i;$$

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y') \geq 1 - \xi_i, \;\; \xi_i \geq 0.$$

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Classical Predictive Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\qquad \mathcal{Y} \triangleq \{-1,+1\}$
- Learning: $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$$

where $\ell(\cdot)$ represents a convex loss, and $R(\mathbf{w})$ is a regularizer preventing overfitting

– Logistic Regression

- Max-likelihood (or MAP) estimation

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(y^i|\mathrm{x}^i; \mathbf{w}) + \mathcal{N}(\mathbf{w})$$

- Corresponds to a Log loss with L2 R

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y'\in\mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

– Support Vector Machines (SVM)

- Max-margin learning

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{N}\xi_i;$$

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta\mathbf{f}_i(y') \geq 1 - \xi_i, \ \xi_i \geq 0.$$

- Corresponds to a hinge loss with L2 R

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y'\in\mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Advantages:

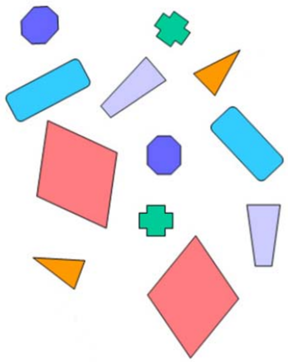
1. Full probabilistic semantics
2. Straightforward Bayesian or direct regularization
3. Hidden structures or generative hierarchy

Advantages:

1. Dual sparsity: few support vectors
2. Kernel tricks
3. Strong empirical results

Structured Prediction Problem

- Unstructured prediction



$$\mathbf{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 0/1 \end{pmatrix}$$

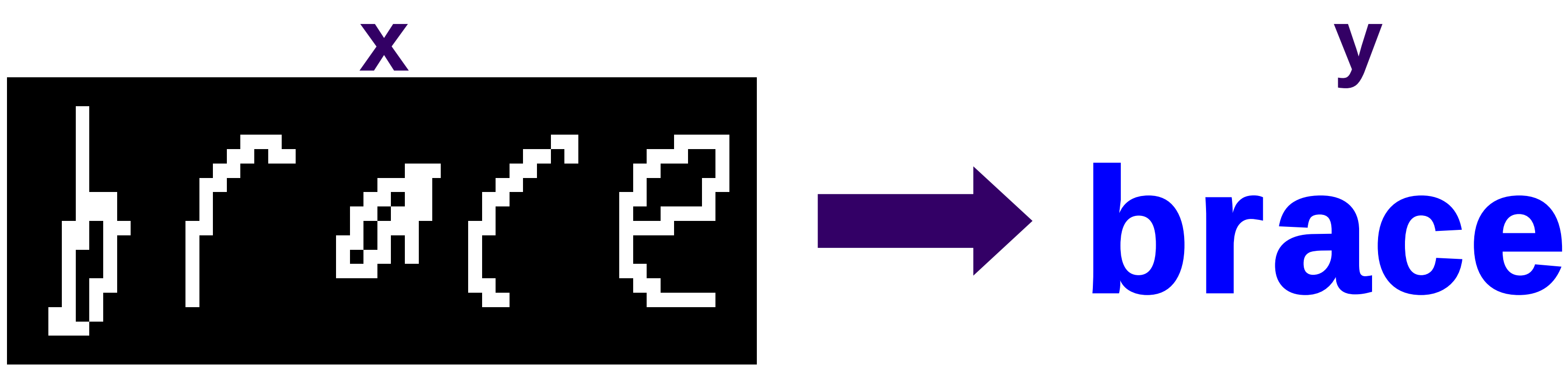
- Structured prediction
  - Part of speech tagging

$\mathbf{x} =$ **"Do you want sugar in it?"** $\Rightarrow$ $\mathbf{y} =$ **<verb pron verb noun prep pron>**

  - Image segmentation

$$\mathbf{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \dots \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \dots \\ \vdots & \vdots & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \dots \\ y_{21} & y_{22} & \dots \\ \vdots & \vdots & \dots \end{pmatrix}$$

OCR example



Sequential structure

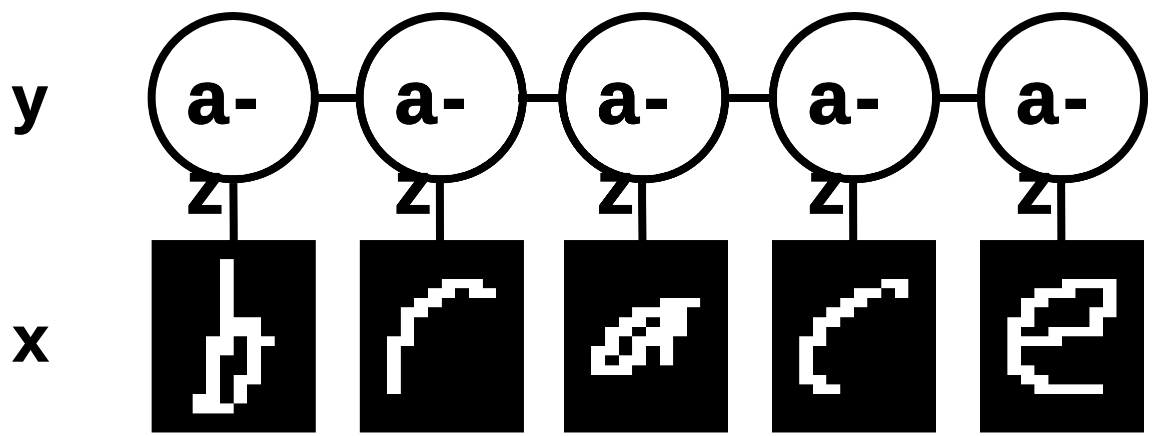
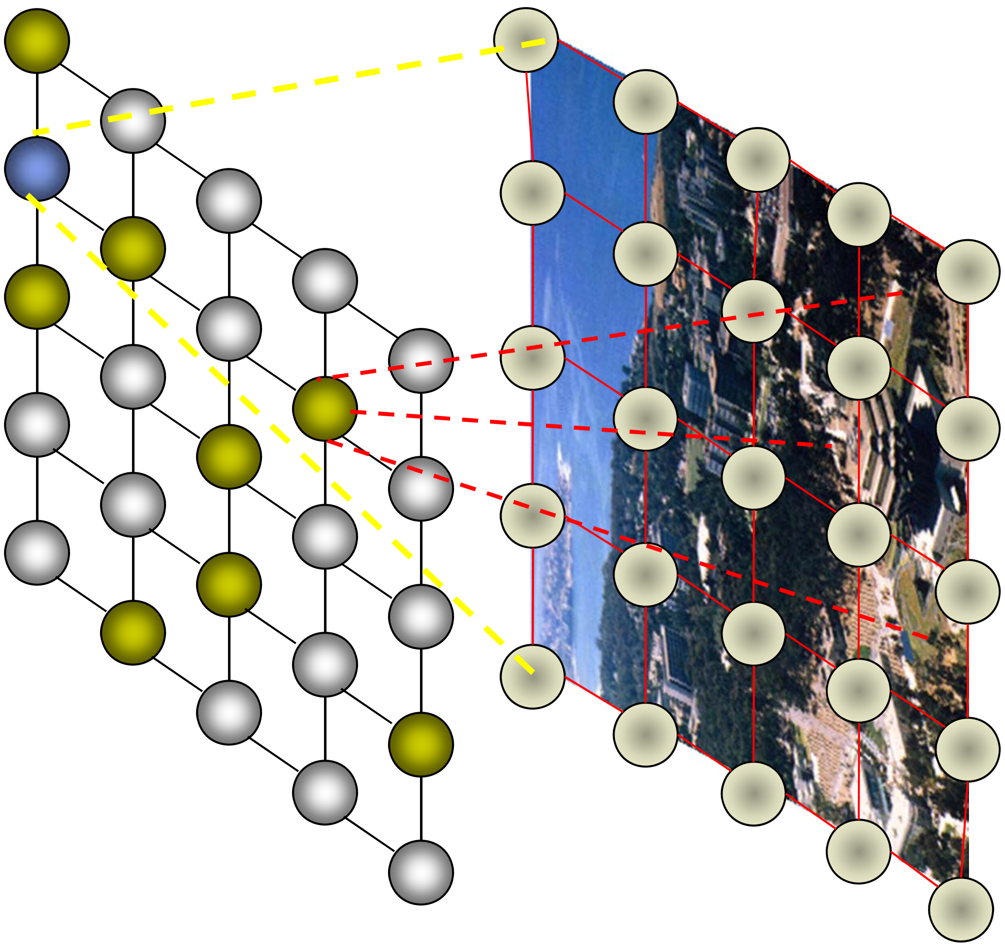


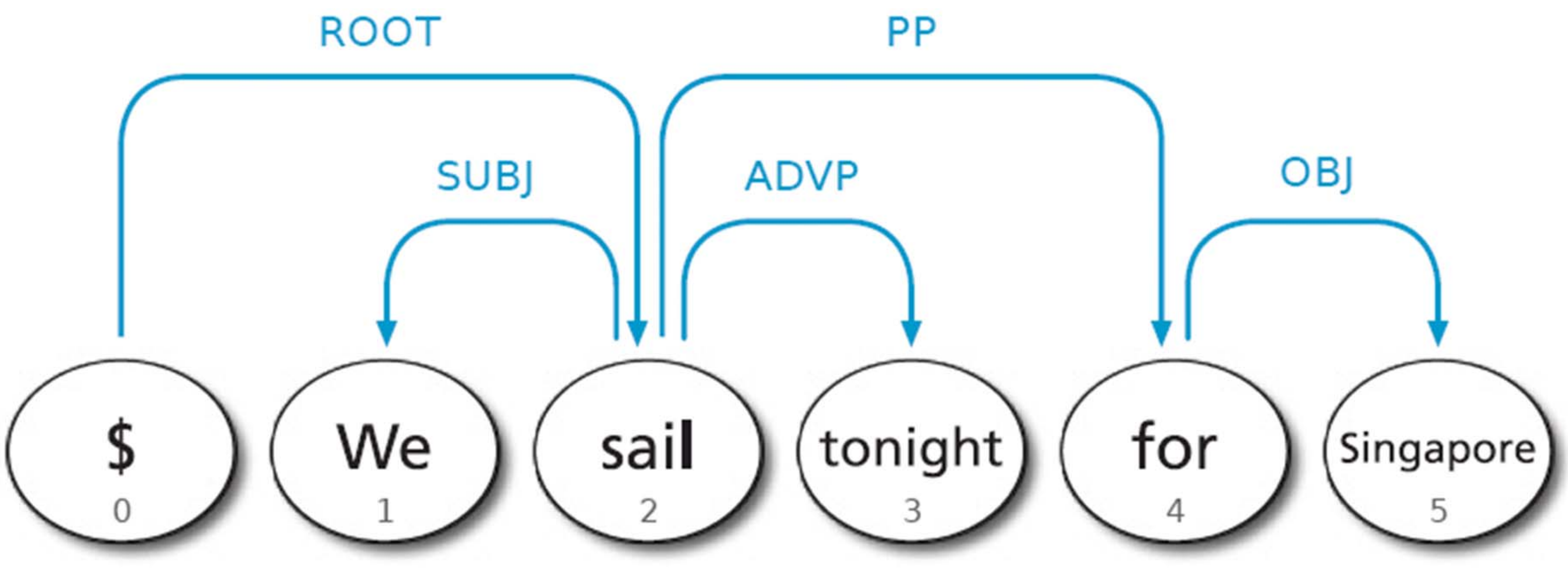
Image Segmentation



$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Jointly segmenting/annotating images
- Image-image matching, image-text matching
- Problem:
  - Given structure (feature), learning $\vec{\theta}$
  - Learning sparse, interpretable, **predictive** structures/features

Dependency parsing of Sentences



Challenge:
Structured outputs, and globally constrained to be a valid tree

Structured Prediction Graphical Models

- **Input and output space** $\mathcal{X} \triangleq \mathbb{R}_{X_1} \times, \ldots, \mathbb{R}_{X_K} \quad \mathcal{Y} \triangleq \mathbb{R}_{Y_1} \times, \ldots, \mathbb{R}_{Y_{K'}}$

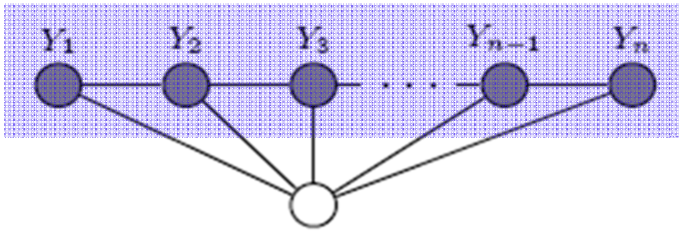
- Conditional Random Fields (CRFs) (Lafferty et al 2001)
  - Based on a Logistic Loss (LR)
  - Max-likelihood estimation (point-estimate)

$$\mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$$

- **Max-margin Markov Networks (M³Ns) (Taskar et al 2003)**
  - **Based on a Hinge Loss (SVM)**
  - **Max-margin learning (point-estimate)**

$$\mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \max_{\mathbf{y}'} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}') - \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}', \mathbf{y})$$

- **Markov properties are encoded in the feature functions $\mathbf{f}(\mathbf{x}, \mathbf{y})$**



$X = X_1, \ldots, X_{n-1}, X_n$

Structured Prediction Graphical Models

- Conditional Random Fields (CRFs) (Lafferty et al 2001)
  - Based on a Logistic Loss (LR)
  - Max-likelihood estimation (point-estimate)

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + R(\mathbf{w})$$

- **Max-margin Markov Networks (M^3Ns) (Taskar et al 2003)**
  - **Based on a Hinge Loss (SVM)**
  - **Max-margin learning (point-estimate)**

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \max_{\mathbf{y}'} \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}') - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + \ell(\mathbf{y}',\mathbf{y}) + R(\mathbf{w})$$

Challenges:

- **SPARSE "Interpretable" prediction model**
- **Prior information of structures**
- **Latent structures/variables**
- **Time series and non-stationarity**
- **Scalable to large-scale problems (e.g., 10^4 input/output dimension)**

Comparing to unstructured predictive models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\quad \mathcal{Y} \triangleq \{-1, +1\}$
- Learning: $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$$

where $\ell(\cdot)$ represents a convex loss, and $R(\mathbf{w})$ is a regularizer preventing overfitting

– **Logistic Regression**
  - **Max-likelihood (or MAP) estimation**

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(y^i | \mathrm{x}^i; \mathbf{w}) + \mathcal{N}(\mathbf{w})$$

  - **Corresponds to a Log loss with L2 R**

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

– **Support Vector Machines (SVM)**
  - **Max-margin learning**

$$\min_{\mathbf{w}, \xi} \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^{N} \xi_i;$$

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y') \geq 1 - \xi_i, \; \xi_i \geq 0.$$

  - **Corresponds to a hinge loss with L2 R**

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Structured models

$$h(\mathbf{x}) = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{x})}{\arg\max}\ s(\mathbf{x}, \mathbf{y})$$ ← **scoring function**

space of feasible outputs

Assumptions:

$$score(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_p \mathbf{w}^\top \mathbf{f}(\mathbf{x}_p, \mathbf{y}_p)$$

linear combination of features

sum of part scores:

- **index *p* represents a part in the structure**

Large Margin Estimation

- Given training example ($\mathbf{x}$, $\mathbf{y}^*$), we want:

$$\arg\max_{\mathbf{y}} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{y}^*$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) > \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{y} \neq \mathbf{y}^*$$

$$\boxed{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \gamma\,\ell(\mathbf{y}^*, \mathbf{y}) \quad \forall \mathbf{y}}$$

- **Maximize margin** γ
- **Mistake weighted margin** $\gamma\ell(\mathbf{y}^*, \mathbf{y})$

$\ell(\mathbf{y}^*, \mathbf{y}) = \sum_i I(y_i^* \neq y_i)$ **# of mistakes in y**

***Taskar et al. 03**

Large Margin Estimation

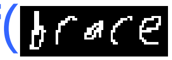
- Recall from SVMs:
  - Maximizing margin γ is equivalent to minimizing the square of the L2-norm of the weight vector **w**:

- New objective function:

$$\min_{\mathbf{w}} \quad \frac{1}{2}||\mathbf{w}||^2$$
$$s.t. \;\; \mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i') + \ell(\mathbf{y}_i, \mathbf{y}_i'), \quad \forall i, \mathbf{y}_i' \in \mathcal{Y}_i$$

OCR Example

- We want:

  argmax$_{\textbf{word}}$ $\mathbf{w}^T \mathbf{f}($  **, word) = “brace”**

- Equivalently:

  $\mathbf{w}^T \mathbf{f}($  **,“brace”) >** $\mathbf{w}^T \mathbf{f}($  **,“aaaaa”)**

  $\mathbf{w}^T \mathbf{f}($  **,“brace”) >** $\mathbf{w}^T \mathbf{f}($  **,“aaaab”)**

  …

  $\mathbf{w}^T \mathbf{f}($  **,“brace”) >** $\mathbf{w}^T \mathbf{f}($  **,“zzzzz”)**

a lot!

Min-max Formulation

- Brute force enumeration of constraints:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y}), \quad \forall \mathbf{y}$$

  - The constraints are exponential in the size of the structure

- Alternative: min-max formulation
  - add only the most violated constraint

$$\mathbf{y}' = \arg\max_{\mathbf{y} \neq \mathbf{y}^*} [\mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}) + \ell(\mathbf{y}^i, \mathbf{y})]$$

$$\text{add to QP}: \ \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}^i) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}') + \ell(\mathbf{y}^i, \mathbf{y}')$$

  - Handles more general loss functions
  - Only polynomial # of constraints needed

Min-max Formulation

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\mathbf{y} \neq \mathbf{y}*} \ \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y})$$

- Key step: convert the maximization in the constraint from discrete to continuous
  - This enables us to plug it into a QP

$$\max_{\mathbf{y} \neq \mathbf{y}^*} \ \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y}) \Longleftrightarrow \max_{\mathbf{z} \in \mathcal{Z}} \ (\mathbf{F}^\top \mathbf{w} + \ell)^\top \mathbf{z}$$

discrete optim. **continuous optim.**

- How to do this conversion?
  - Linear chain example in the next slides →

$\mathbf{y} \Rightarrow \mathbf{z}$ map for linear chain structures

OCR example: $\mathbf{y} =$ 'ABABB';

$\mathbf{z}$'s are the indicator variables for the corresponding classes (alphabet)

$$
\begin{array}{c}
\mathrm{A} \\ \mathrm{B} \\ : \\ \mathrm{B}
\end{array}
\quad
z_1(m) = \begin{bmatrix} 1 \\ 0 \\ : \\ 0 \end{bmatrix}
\quad
z_2(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix}
\quad
z_3(m) = \begin{bmatrix} 1 \\ 0 \\ : \\ 0 \end{bmatrix}
\quad
z_4(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix}
\quad
z_5(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix}
$$

$$
\begin{array}{c}
\mathrm{A} \\ \mathrm{B} \\ : \\ \mathrm{B}
\end{array}
\quad
z_{12}(m,n) = \begin{bmatrix} 0 & 1 & . & 0 \\ 0 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
\quad
z_{23}(m,n) = \begin{bmatrix} 0 & 0 & . & 0 \\ 1 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
\quad
z_{34}(m,n) = \begin{bmatrix} 0 & 1 & . & 0 \\ 0 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
\quad
z_{45}(m,n) = \begin{bmatrix} 0 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
$$

$$
\begin{bmatrix} \mathrm{A} & \mathrm{B} & . & \mathrm{B} \end{bmatrix}
\quad
\begin{bmatrix} \mathrm{A} & \mathrm{B} & . & \mathrm{B} \end{bmatrix}
\quad
\begin{bmatrix} \mathrm{A} & \mathrm{B} & . & \mathrm{B} \end{bmatrix}
\quad
\begin{bmatrix} \mathrm{A} & \mathrm{B} & . & \mathrm{B} \end{bmatrix}
$$

$\mathbf{y} \Rightarrow \mathbf{z}$ map for linear chain structures

Rewriting the maximization function in terms of indicator variables:

$$\left.\begin{aligned}\max_{\mathbf{z}} \quad & \sum_{j,m} z_j(m)\left[\mathbf{w}^\top \mathbf{f}_{\text{node}}(\mathbf{x}_j, m) + \ell_j(m)\right] \\ + & \sum_{jk,m,n} z_{jk}(m,n)\left[\mathbf{w}^\top \mathbf{f}_{\text{edge}}(\mathbf{x}_{jk}, m, n) + \ell_{jk}(m,n)\right]\end{aligned}\right\} (\mathbf{F}^\top \mathbf{w} + \boldsymbol{\ell})^\top \mathbf{z}$$

$z_k(n)$

0	1	0	0

$z_j(m)$

0
0
1
0

0	0	0	0
0	0	0	0
0	1	0	0
0	0	0	0

$z_{jk}(m,n)$

$$\left.\begin{aligned} & z_j(m) \geq 0; \; z_{jk}(m,n) \geq 0; \\ \text{normalization} \quad & \sum_m z_j(m) = 1 \\ \text{agreement} \quad & \sum_n z_{jk}(m,n) = z_j(m) \end{aligned}\right\} \mathbf{A}\mathbf{z} = \mathbf{b}$$

$$\max_{\mathbf{A}\mathbf{z}=\mathbf{b}} \; (\mathbf{F}^\top \mathbf{w} + \boldsymbol{\ell})^\top \mathbf{z}$$

Min-max formulation

- Original problem:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$
$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\mathbf{y}} \ \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y})$$

- Transformed problem:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$
$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\substack{\mathbf{z} \geq 0; \\ \mathbf{A}\mathbf{z} = \mathbf{b};}} \mathbf{q}^\top \mathbf{z} \quad \text{where } \mathbf{q}^\top = \mathbf{w}^\top \mathbf{F} + \ell^\top$$

  - Has integral solutions **z** for chains, trees
  - Can be fractional for untriangulated networks

Min-max formulation

- Using strong Lagrangian duality:
  (beyond the scope of this lecture)

$$\max_{\substack{\mathbf{z}\geq 0; \\ \mathbf{A}\mathbf{z}=\mathbf{b};}} \mathbf{q}^\top \mathbf{z} \;=\; \min_{\mathbf{A}^\top \mu \geq \mathbf{q}} \mathbf{b}^\top \mu$$

- Use the result above to minimize jointly over **w** and μ:

$$\begin{aligned} \min_{\mathbf{w},\mu} \; & \frac{1}{2}||\mathbf{w}||^2 \\ \text{s.t.} \; & \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}^*) \geq \mathbf{b}^\top \mu; \\ & \mathbf{A}^\top \mu \geq \mathbf{q}; \end{aligned}$$

Min-max formulation

$$
\begin{aligned}
\min_{\mathbf{w},\mu} \quad & \frac{1}{2}||\mathbf{w}||^2 \\
\text{s.t.} \quad & \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{b}^\top \mu; \\
& \mathbf{A}^\top \mu \geq (\mathbf{w}^\top \mathbf{F} + \ell)^\top
\end{aligned}
$$

- Formulation produces compact QP for
  - Low-treewidth Markov networks
  - Associative Markov networks
  - Context free grammars
  - Bipartite matchings
  - Any problem with compact LP inference

Results: Handwriting Recognition

Length: ~8 chars

Letter: 16x8 pixels

10-fold Train/Test

5000/50000 letters

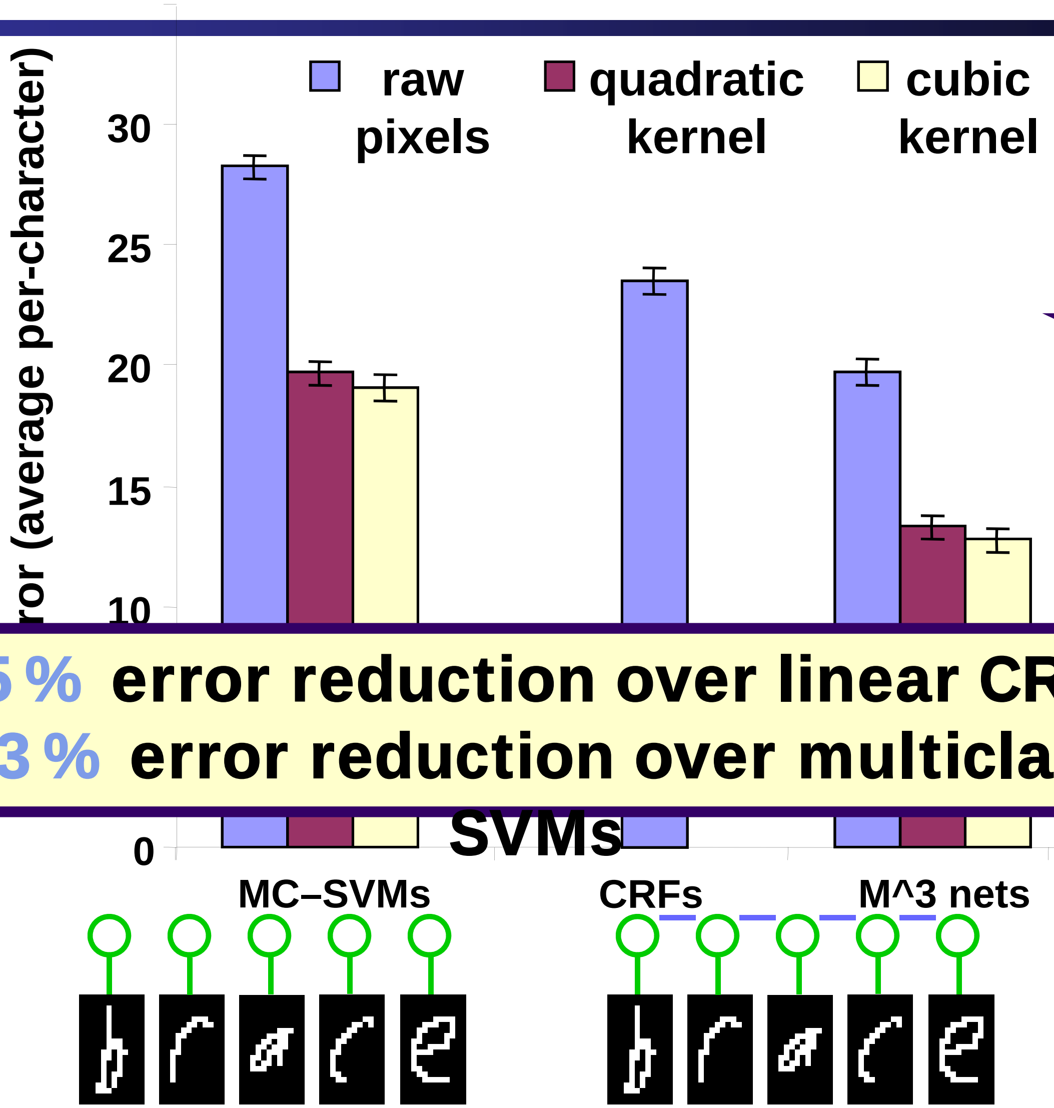
600/6000 words

Models:

Multiclass-SVMs*

CRFs

M^3 nets

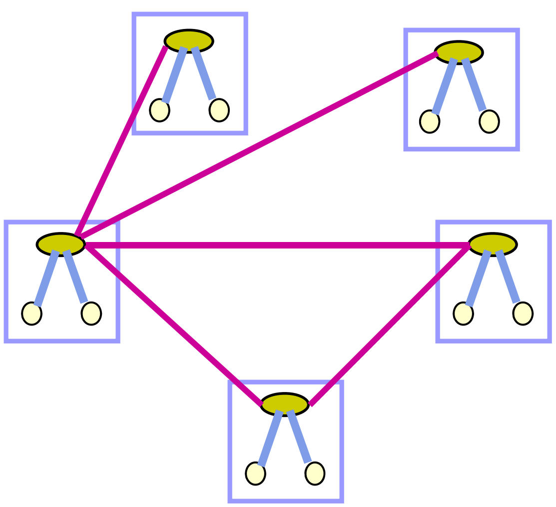


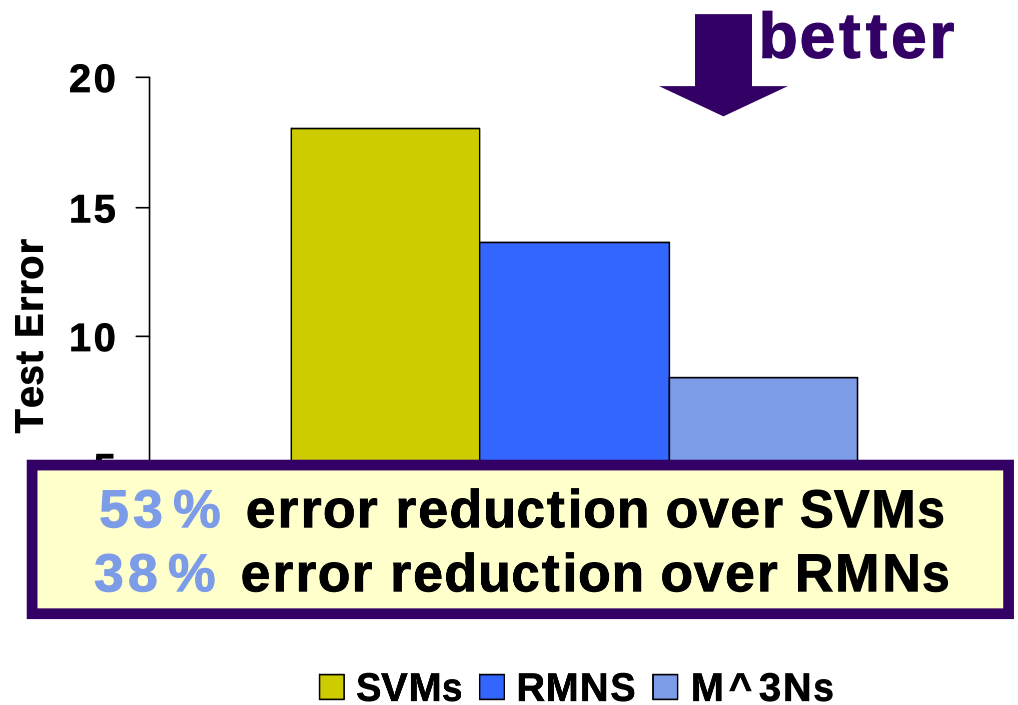
45 % error reduction over linear CRFs
33 % error reduction over multiclass SVMs

Crammer & Singer 01

Results: Hypertext Classification

- **WebKB dataset**
  - **Four CS department websites:** 1300 pages/3500 links
  - **Classify each page:** faculty, course, student, project, other
  - **Train on three universities/test on fourth**





*Taskar et al 02

MLE versus max-margin learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian or direct regularization
  - Hidden structures or generative hierarchy

- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Support vector property, sound theoretical guarantee with limited samples
  - Kernel tricks

- Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)
  - Model averaging $\hat{y} = \text{sign} \int p(\mathbf{w}) F(x; \mathbf{w}) \, d\mathbf{w}$ $\quad (y \in \{+1, -1\})$
  - The optimization problem (binary classification)

$$\min_{p(\Theta)} KL(p(\Theta) || p_0(\Theta))$$

$$\text{s.t.} \int p(\Theta)[y_i F(x; \mathbf{w}) - \xi_i] \, d\Theta \geq 0, \forall i,$$

where Θ *is the parameter* $\mathbf{w}$ *when* ξ *are kept fixed or the pair* $(\mathbf{w}, \xi)$ *when we want to optimize over* ξ

Maximum Entropy Discrimination Markov Networks

- Structured MaxEnt Discrimination (SMED):

$$\mathrm{P1}: \quad \min_{p(\mathbf{w}),\xi} \boxed{KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)}$$
$$\text{s.t. } \; p(\mathbf{w}) \in \mathcal{F}_1, \; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

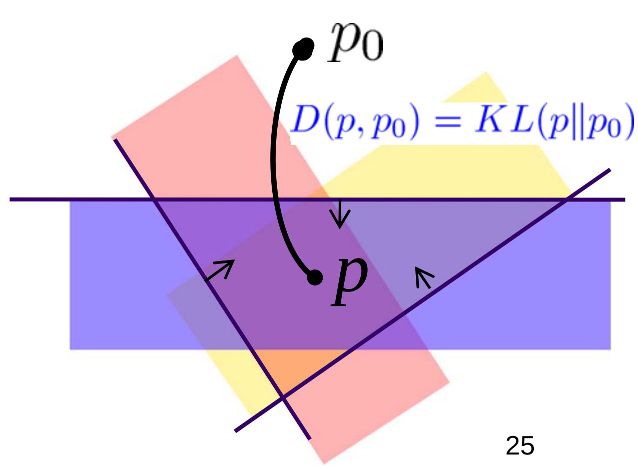
- Feasible subspace of weight distribution:

$$\mathcal{F}_1 = \{p(\mathbf{w}) : \boxed{\int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, d\mathbf{w} \geq -\xi_i,} \; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

expected margin constraints.

- Average from distribution of M^3Ns

$$h_1(\mathbf{x}; p(\mathbf{w})) = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x}, \mathbf{y}; \mathbf{w}) d\mathbf{w}$$



Solution to MaxEnDNet

- Theorem:
  - Posterior Distribution:

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\Big\{ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\Big\}$$

  - Dual Optimization Problem:

$$\begin{aligned} \mathrm{D1}: \quad & \max_{\alpha} \ -\log Z(\alpha) - U^{\star}(\alpha) \\ & \text{s.t.} \ \ \alpha_i(\mathbf{y}) \geq 0, \ \forall i, \ \forall \mathbf{y}, \end{aligned}$$

$U^{\star}(\cdot)$ is the conjugate of the $U(\cdot)$, i.e., $U^{\star}(\alpha) = \sup_{\xi} \big(\sum_{i,\mathrm{y}} \alpha_i(\mathrm{y})\xi_i - U(\xi) \big)$

Gaussian MaxEnDNet (reduction to M³N)

- Theorem
  - Assume $F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}), U(\xi) = C\sum_i \xi_i$, and $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0,I)$
    - Posterior distribution: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I)$, where $\mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})$
    - Dual optimization:

$$\boxed{\begin{aligned} &\max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})\|^2 \\ &\text{s.t.} \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C;\ \alpha_i(\mathbf{y}) \geq 0,\ \forall i,\ \forall \mathbf{y}, \end{aligned}}$$

M³N

    - Predictive rule:

$$h_1(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w})F(\mathbf{x},\mathbf{y};\mathbf{w})\,d\mathbf{w} = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \mu_{\mathbf{w}}^\top \mathbf{f}(\mathbf{x},\mathbf{y})$$

- Thus, MaxEnDNet subsumes M³Ns and admits all the merits of max-margin learning
- Furthermore, MaxEnDNet has at least three advantages …

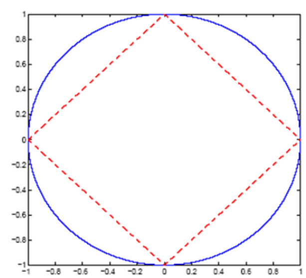
Three Advantages

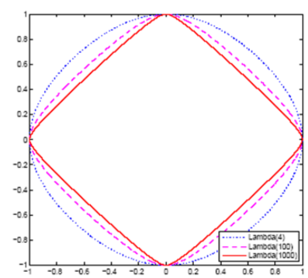
- An averaging Model: PAC-Bayesian prediction error guarantee (Theorem 3)

$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\Big(\sqrt{\frac{\gamma^{-2}KL(p||p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\Big).$$

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M^{3}N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M^{3}N)

$$\min_{\mu,\xi}\ \sqrt{\lambda}\sum_{k=1}^{K}\Big(\sqrt{\mu_k^2+\frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big) + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.}\quad \mu^\top \Delta\mathbf{f}_i(\mathbf{y}) \ge \Delta\ell_i(\mathbf{y}) - \xi_i;\ \xi_i \ge 0,\quad \forall i,\ \forall \mathbf{y} \ne \mathbf{y}^i.$$





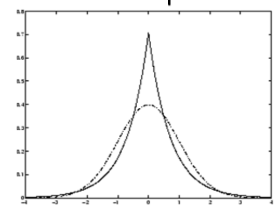
- Integrating Generative and Discriminative principles (next class)
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

Laplace MaxEnDNet (primal sparse M^3N)

(Zhu and Xing, ICML 2009)

- Laplace Prior:

$$p_0(\mathbf{w}) = \prod_{k=1}^{K} \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|w_k|} = (\frac{\sqrt{\lambda}}{2})^K e^{-\sqrt{\lambda}\|\mathbf{w}\|}$$

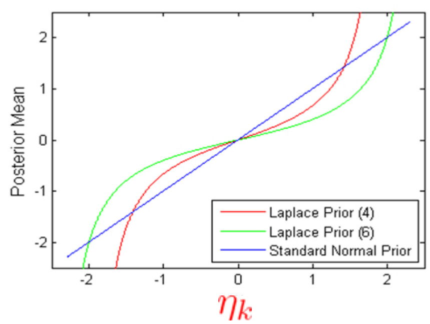


- Corollary 4:
  - Under a Laplace MaxEnDNet, the posterior mean of parameter vector $\mathbf{w}$ is:

$$\forall k,\ \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$

where the vector η is a linear combination of "support vectors":

$$\eta = \sum_{\alpha} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y})$$



- The Gaussian MaxEnDNet and the regular M^3N has no such shrinkage
  - there, we have

$$\langle \mathbf{w} \rangle_p = \eta \iff \forall k,\ \langle w_k \rangle_p = \eta_k$$

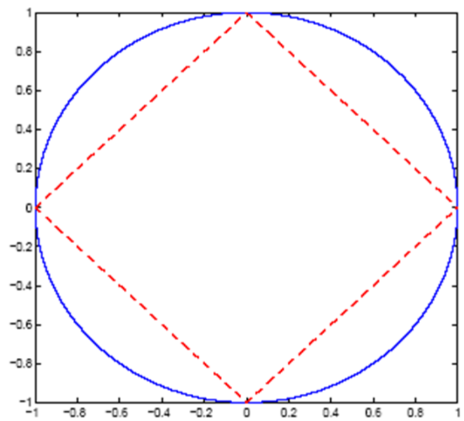
LapMEDN vs. L_2 and L_1 regularization

$$\min_{\mu,\xi} \ |\mu| + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \ \ \mu^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i; \ \xi_i \geq 0, \ \ \forall i, \ \forall \mathbf{y} \neq \mathbf{y}^i.$$

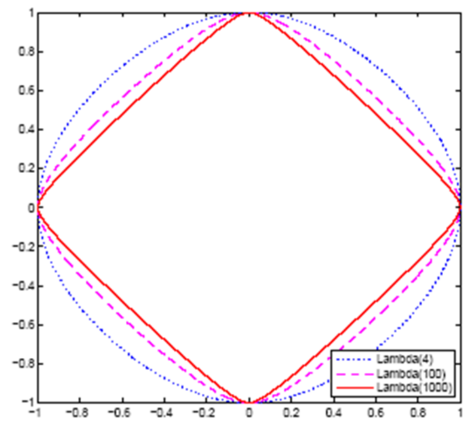
- Corollary 5: LapMEDN corresponding to solving the following primal optimization problem:

$$\min_{\mu,\xi} \ \sqrt{\lambda}\sum_{k=1}^{K}\Big(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big) + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \ \ \mu^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i; \ \xi_i \geq 0, \ \ \forall i, \ \forall \mathbf{y} \neq \mathbf{y}^i.$$

- KL norm: $$\|\mu\|_{KL} \triangleq \sum_{k=1}^{K}\Big(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big)$$



L_1 and L_2 norms



KL norms

Recall Primal and Dual Problems of M³Ns

- Primal problem:

$$\mathrm{P0}\ (\mathrm{M}^3\mathrm{N}):\ \min_{\mathbf{w},\xi}\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i:\ \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0\ ,$$

- Algorithms
  - Cutting plane
  - Sub-gradient
  - …

- Dual problem:

$$\mathrm{D0}\ (\mathrm{M}^3\mathrm{N}):\ \max_{\alpha}\ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\eta^\top\eta$$

$$\text{s.t. } \forall i,\ \forall \mathbf{y}:\ \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C;\ \ \alpha_i(\mathbf{y}) \geq 0.$$

$$\text{where } \eta = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

- Algorithms:
  - SMO
  - Exponentiated gradient
  - …

$$\mathbf{w}^\star = \eta^\star = \sum_{i,\mathbf{y}} \alpha_i^\star(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

- So, M³N is dual sparse!

$$\mathbf{y}^\star = h(\mathbf{x}) \triangleq \arg\max_{y} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$$

Variational Learning of LapMEDN

- Exact primal or dual function is hard to optimize

$$\min_{\mu,\xi} \sqrt{\lambda}\sum_{k=1}^{K}\Big(\sqrt{\mu_k^2+\frac{1}{\lambda}}-\frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big)+C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.}\ \ \mu^\top\Delta\mathbf{f}_i(\mathbf{y})\geq\Delta\ell_i(\mathbf{y})-\xi_i;\ \xi_i\geq 0,\ \ \forall i,\ \forall\mathbf{y}\neq\mathbf{y}^i.$$

$$\max_{\alpha}\ \sum_{i,\mathbf{y}}\alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y})-\sum_{k=1}^{K}\log\frac{\lambda}{\lambda-\eta_k^2}$$
$$\text{s.t.}\ \ \sum_{\mathbf{y}}\alpha_i(\mathbf{y})=C;\ \alpha_i(\mathbf{y})\geq 0,\ \forall i,\ \forall\mathbf{y}.$$

- Use the hierarchical representation of Laplace prior, we get:

$$\begin{aligned} KL(p\|p_0) &= -H(p)-\langle\log\int p(\mathbf{w}|\tau)p(\tau|\lambda)\,\mathrm{d}\tau\,\rangle_p \\ &\leq -H(p)-\langle\int q(\tau)\log\frac{p(\mathbf{w}|\tau)p(\tau|\lambda)}{q(\tau)}\,\mathrm{d}\tau\,\rangle_p \triangleq \mathcal{L}(p(\mathbf{w}),q(\tau)) \end{aligned}$$

- We optimize an upper bound:

$$\min_{p(\mathbf{w})\in\mathcal{F}_1;q(\tau);\xi}\ \mathcal{L}(p(\mathbf{w}),q(\tau))+U(\xi)$$

- Why is it easier?
  - Alternating minimization leads to nicer optimization problems

Keep $q(\tau)$ fixed	**Keep $p(\mathbf{w})$ fixed**
- The effective prior is normal $\forall k:\ p_0(w_k\|\tau_k)=\mathcal{N}(w_k\|0,\langle\frac{1}{\tau_k}\rangle_{q(\tau)}^{-1})$	- Closed form solution of $q(\tau)$ and its expectation $\langle\frac{1}{\tau_k}\rangle_q=\sqrt{\frac{\lambda}{\langle w_k^2\rangle_p}}.$

An M³N optimization problem!

Closed-form solution!

Algorithmic issues of solving M^{3}Ns

- Primal problem:

$$\text{P0 (M}^3\text{N)}: \min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i : \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0\ ,$$

- Algorithms
  - Cutting plane
  - Sub-gradient
  - …

- **Dual problem:**

$$\text{D0 (M}^3\text{N)}: \max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\eta^\top\eta$$

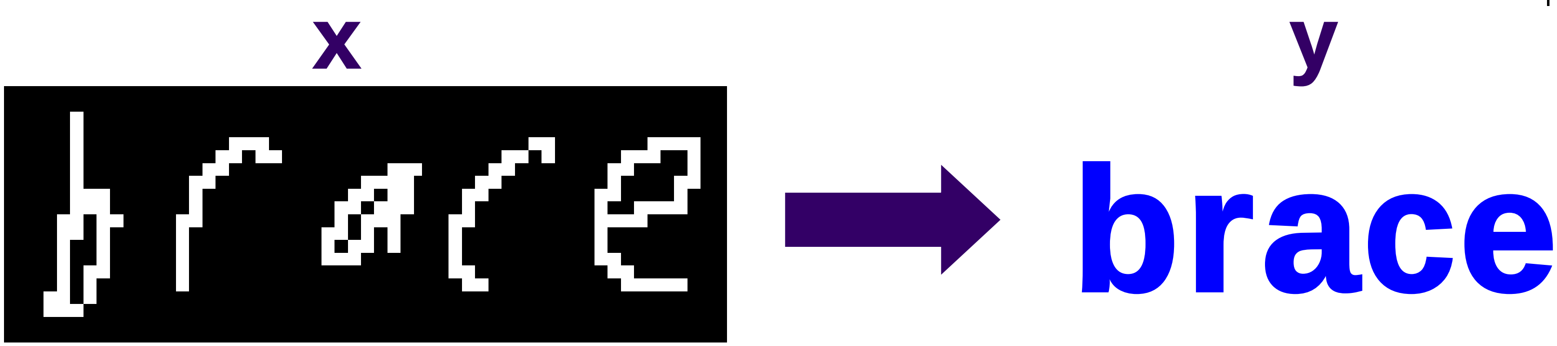
$$\text{s.t. } \forall i,\ \forall \mathbf{y}: \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C;\ \ \alpha_i(\mathbf{y}) \geq 0.$$

$$\text{where } \eta = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

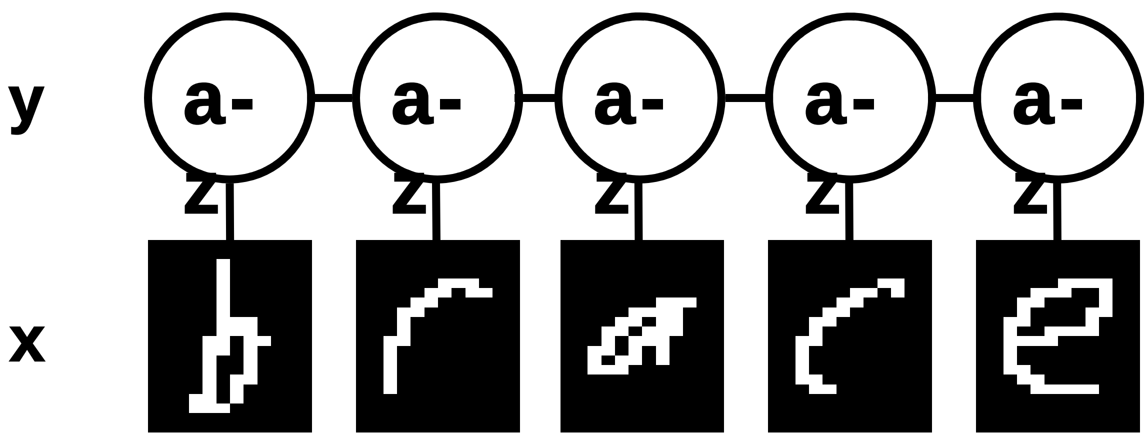
- **Algorithms:**
  - **SMO**
  - **Exponentiated gradient**
  - **…**

- **Nonlinear Features with Kernels**
  - **Generative entropic kernels [Martins et al, JMLR 2009]**
  - **Nonparametric RKHS embedding of rich distributions [on going]**
- **Approximate decoders for global features**
  - **LP-relaxed Inference (polyhedral outer approx.) [Martins et al, ICML 09, ACL 09]**
  - **Balancing Accuracy and Runtime: Loss-augmented inference**

Experimental results on OCR datasets



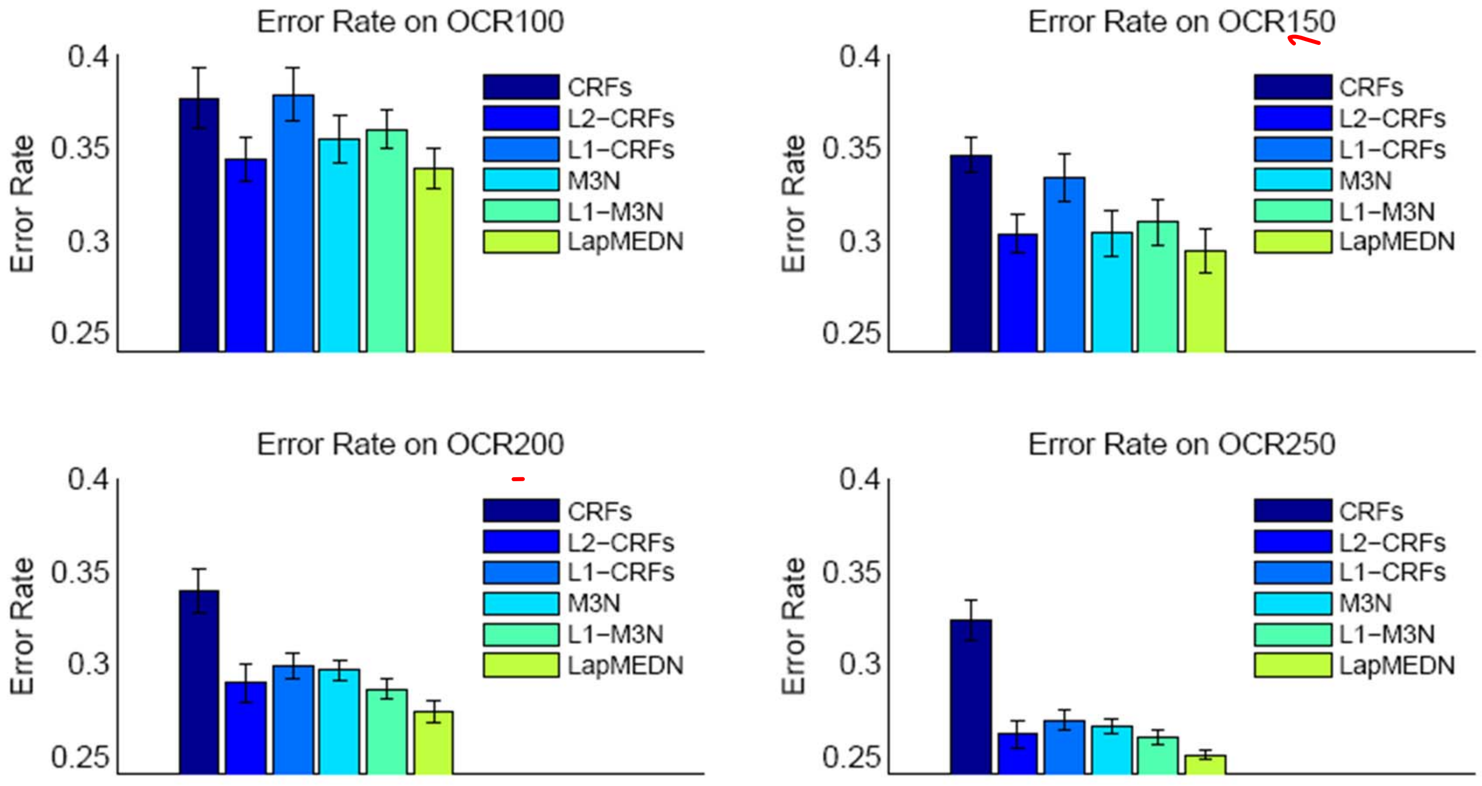
Structured output



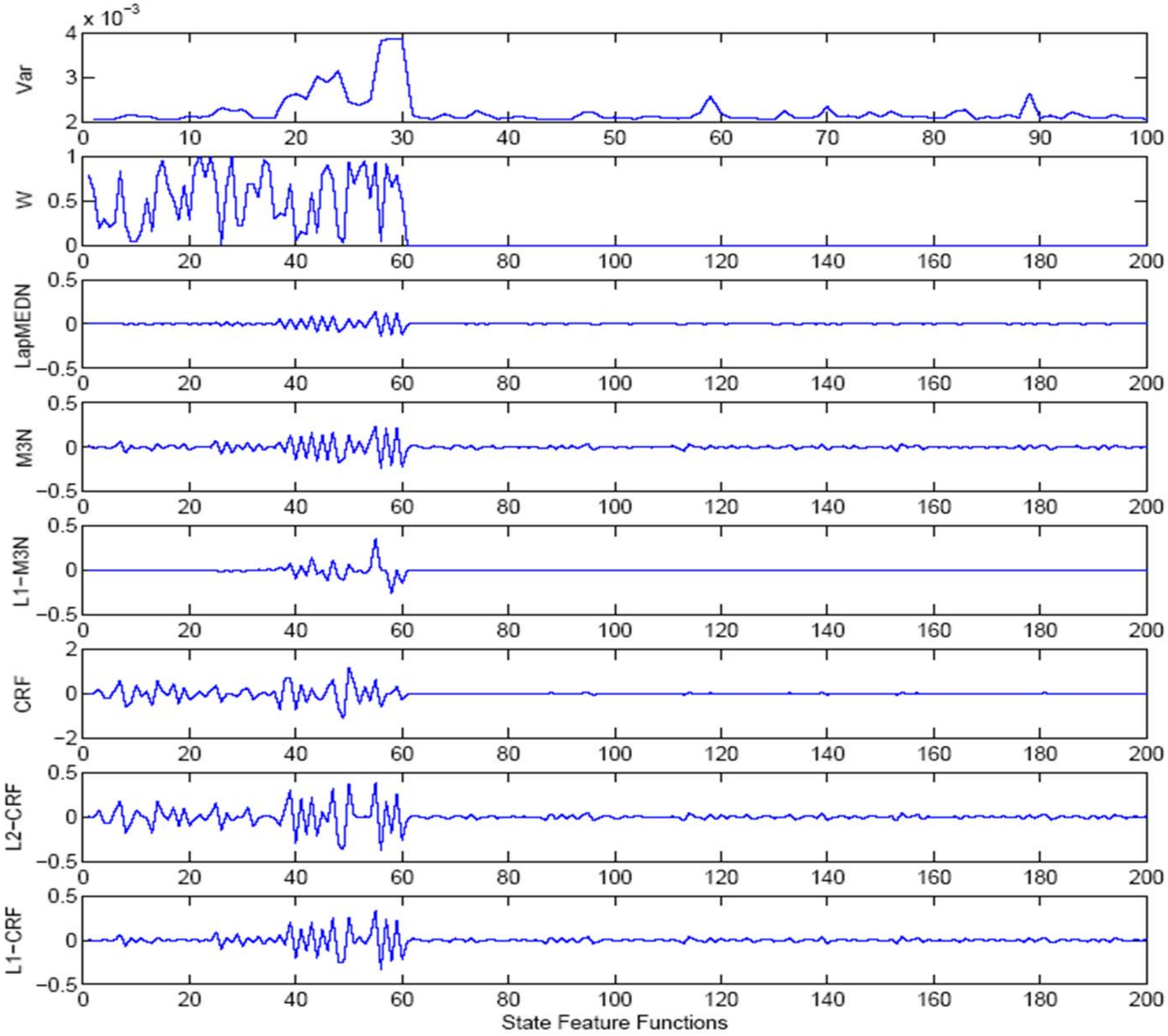
Experimental results on OCR datasets

$$(\mathrm{CRFs},\ L_1-\mathrm{CRFs},\ L_2-\mathrm{CRFs},\ \mathrm{M^3Ns},\ L_1-\mathrm{M^3Ns},\ \text{and LapMEDN})$$

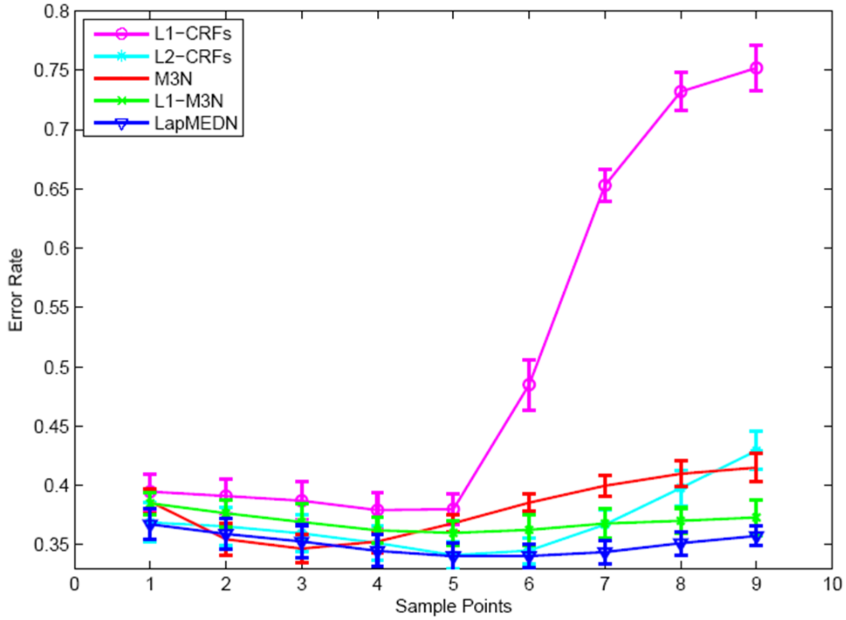
- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.



Feature Selection



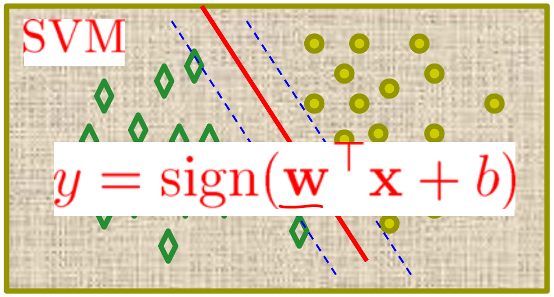
Sensitivity to Regularization Constants



- ***L_1*-CRF and *L_2*-CRF:**
  - **0.001, 0.01, 0.1, 1, 4, 9, 16**
- **M^3N and LapM^3N:**
  - **1, 4, 9, 16, 25, 36, 49, 64, 81**

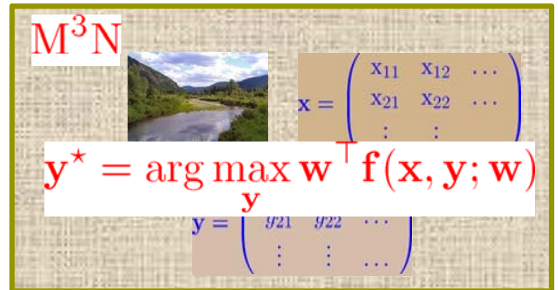
- *L_1*-CRFs are much sensitive to regularization constants; the others are more stable
- LapM^3N is the most stable one

Summary:
Margin-based Learning Paradigms



$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i;$$
$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \ \forall i.$$

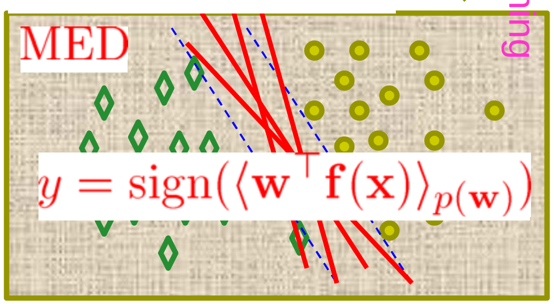
Structured prediction



$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$
$$\text{s.t.} : \ \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i, \ \xi_i \geq 0, \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$

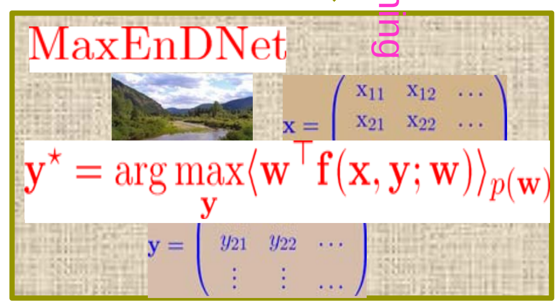
Bayes learning

Bayes learning



$$\min_{p,\xi} KL(p\|p_0) + C\sum_{i=1}^{N} \xi_i;$$
$$\text{s.t.} \quad y_i \langle \mathbf{f}(\mathbf{x}_i) \rangle_{p(\mathbf{w})} \geq 1 - \xi_i, \ \forall i.$$

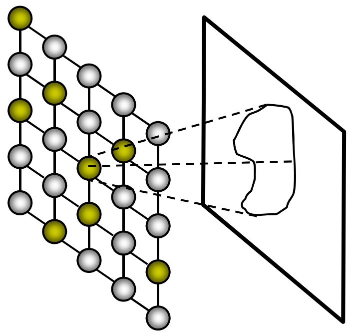
Structured prediction



$$\min_{p(\mathbf{w}),\xi} \ KL(p\|p_0) + U(\xi)$$
$$\text{s.t.} \quad \int p(\mathbf{w})[\Delta F_i(\mathbf{y}; \mathbf{w}) - \Delta\ell_i(\mathbf{y})]\, d\mathbf{w} \geq -\xi_i, \ \xi_i \geq 0,, \ \forall i, \forall \mathbf{y} \neq \mathbf{y}^i.$$

Open Problems

- Unsupervised CRF learning and MaxMargin Learning
  - Only X, but not Y (sometimes part of Y), is available
  - We want to recognize a pattern that is maximally different from the rest!



  - What does margin or conditional likelihood mean in these cases? Given only $\{X_n\}$, how can we define the cost function?

$$\text{margin} = w^T \left(F(y_n, x_n) - F(y'_n, x_n) \right)$$

$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

  - Algorithmic challenge

Remember: Elements of Learning

- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data and other info:
    - Input and output (e.g., continuous, binary, counts, …)
    - Supervised or unsupervised, of a blend of everything?
    - Prior knowledge? Bias?
  - Models and paradigms:
    - BN? MRF? Regression? SVM?
    - Bayesian/Frequents ? Parametric/Nonparametric?
  - Objective/Loss function:
    - MLE? MCLE? Max margin?
    - Log loss, hinge loss, square loss? …
  - Tractability and exactness trade off:
    - Exact inference? MCMC? Variational? Gradient? Greedy search?
    - Online? Batch? Distributed?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**