Entropy and Margin Maximization for Structured Output Learning

Patrick Pletscher, Cheng Soon Ong, Joachim M. Buhmann

ETH Zurich, Switzerland
Machine Learning Group

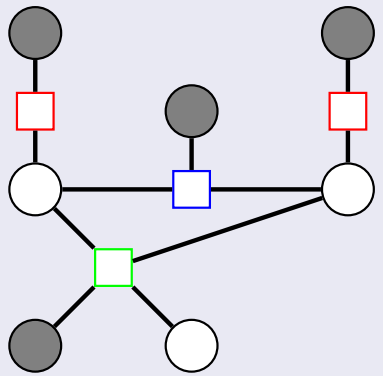
ECML 2010, Barcelona

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Graphical models

Graphical model

- Discrete output variables $\mathbf{y} \in \mathcal{Y}$ ○ and input variables $\mathbf{x} \in \mathcal{X}$ ●.
- Standard factor graph specifying dependencies between variables.
- Factor graph structure and parameterization of the factors assumed to be given.
- Several factors can share the same parameter.



Structured Output Learning

Task: learn parameters of the factors.

Applications

- Computer vision.
- Natural language processing.
- ...

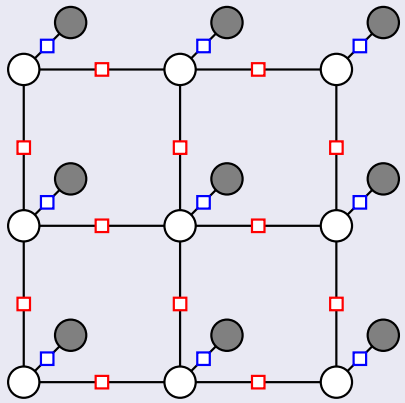
Example: Ising model or binary image segmentation

Neighboring sites should have the same label

- $y_i \in \{0, 1\}$ and $x_i \in \{0, 1\}$.
- Energy given by

$$E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\sum_{i \in \mathcal{V}} w^u(|y_i - x_i|) - \sum_{(i,j) \in \mathcal{E}} w^p(|y_i - y_j|)$$

- Parameters: $\mathbf{w}^u = [0, a]^\mathsf{T}$ and $\mathbf{w}^p = [0, b]^\mathsf{T}$.



Rewrite the energy

Introduce ($\delta_c(z) = 1$ if $z = c$, 0 otherwise):

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \sum_{i \in \mathcal{V}} \delta_1(|y_i - x_i|) \\ \sum_{(i,j) \in \mathcal{E}} \delta_1(|y_i - y_j|) \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Energy of a configuration: $E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle$

Structured Output Prediction and Learning

One additional assumption: Linearity

Assume that the individual factor parameterizations are linear.

Feature map $\phi(\mathbf{x}, \mathbf{y})$

- A factor graph together with its parametrization implies a joint feature map $\phi(\cdot\,, \cdot\,)$.
- Counts the different configurations of the factors.
- Can have thousands of parameters. Ising model very simple.

Score (negative energy)

The score of a configuration is then $\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle$.

Empirical Risk Minimization

Regularized empirical risk

For a given data set $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ the empirical risk is given by

$$\mathcal{L}_\ell(\mathbf{w}, \mathcal{D}, C) = \sum_{n=1}^{N} \ell(\mathbf{w}, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}) + \frac{C}{2}\|\mathbf{w}\|_2^2,$$

where we employ the Euclidean norm as a regularizer. The *loss* of an example is denoted by $\ell(\mathbf{w}, \mathbf{x}^{(n)}, \mathbf{y}^{(n)})$.

Risk minimization

Choose the parameter with the smallest regularized empirical risk

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \mathcal{L}_\ell(\mathbf{w}, \mathcal{D}, C)$$

This work

Understanding relationship between different losses $\ell(\mathbf{w}, \mathbf{x}, \mathbf{y})$.

Maximum-Margin Learning

Structured SVM (Taskar, Guestrin, and Koller, 2003; Tsochantaridis et al., 2004)

Idea: large margin from all other outputs

Given example $(\mathbf{x}, \mathbf{y})$ we want a large margin from all other outputs $\mathbf{y}'$:

$$\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle - \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}') \rangle \geq \Delta(\mathbf{y}', \mathbf{y}) \quad \forall \mathbf{y}' \in \mathcal{Y} \backslash \mathbf{y}$$

Margin term $\Delta(\mathbf{y}', \mathbf{y})$

- Specifies distance between outputs $\mathbf{y}'$ and $\mathbf{y}$.
- Not all outputs are equally bad.
- $\Delta(\mathbf{y}, \mathbf{y}) = 0$; $\Delta(\mathbf{y}', \mathbf{y}) \geq 0$.

Loss: maximum violation

$$\ell_{MM}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle + \max_{\mathbf{y}' \in \mathcal{Y}} [\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}') \rangle + \Delta(\mathbf{y}', \mathbf{y})]$$

Maximum-Likelihood Learning

Conditional Random Field (Lafferty, McCallum, and Pereira, 2001)

Conditional log-linear model

Assume a log-linear model

$$P(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})} \exp\left(\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y})\rangle\right),$$

with the partition sum

$$Z(\mathbf{x},\mathbf{w}) = \sum_{\mathbf{y}'\in\mathcal{Y}} \exp(\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y}')\rangle).$$

Loss: negative log-likelihood

The loss is given as the negative log-likelihood of the data

$$\ell_{LL}(\mathbf{w},\mathbf{x},\mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y})\rangle + \log Z(\mathbf{x},\mathbf{w}).$$

Computational aspects: CRF vs. SSVM

Inference

- Summation vs. maximization over an exponentially large set $\mathcal{Y}$.
- Efficient algorithms for loop-free graphical models.
- Submodularity: maximization can be easier than summation.
- General case: both computationally intractable.

Learning

- Both losses convex.
- Max-margin loss is non-differentiable.
- Log-loss: quasi-Newton, e.g. L-BFGS.
- Max-margin: cutting plane algorithms (Tsochantaridis et al., 2004), Bundle Method for Regularized Risk Minimization (Teo et al., 2009).

Inverse temperature β

Include an inverse temperature into posterior

$$P_\beta(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z_\beta(\mathbf{x},\mathbf{w})} \exp\left(\beta\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y})\rangle\right),$$

with

$$Z_\beta(\mathbf{x},\mathbf{w}) = \sum_{\mathbf{y}'\in\mathcal{Y}} \exp(\beta\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y}')\rangle).$$

Resulting loss

Scale resulting log-loss by $1/\beta$:

$$-\frac{1}{\beta}\log P_\beta(\mathbf{y}|\mathbf{x},\mathbf{w}) = -\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y})\rangle + \frac{1}{\beta}\sum_{\mathbf{y}'\in\mathcal{Y}} \exp(\beta\langle \mathbf{w}, \phi(\mathbf{x},\mathbf{y}')\rangle)$$

Does not change anything for now...

Just rescales the regularizer to $C' = C/\beta$.

Including a margin term (1/2)

Conditional probability over outputs

Let us define

$$P_\beta(\mathbf{y}'|\mathbf{y}) = \frac{1}{Z_\beta(\mathbf{y})} \exp(\beta\Delta(\mathbf{y}', \mathbf{y}))$$

Combination

$$P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w}) \propto P_\beta(\mathbf{y}'|\mathbf{x}, \mathbf{w}) P_\beta(\mathbf{y}'|\mathbf{y})$$

Ensuring normalization

$$P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w}) = \frac{1}{Z_\beta(\mathbf{y}, \mathbf{x}, \mathbf{w})} \exp\Big(\beta\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}')\rangle + \beta\Delta(\mathbf{y}', \mathbf{y})\Big),$$

Biased by margin term, probably not a good density estimation.

Including a margin term (2/2)

Final loss

Again we rescale by $1/\beta$ and take the logarithm:

$$\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\big\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \big\rangle + \frac{1}{\beta} \log \sum_{\mathbf{y}' \in \mathcal{Y}} \exp\Big(\beta \big\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}') \big\rangle + \beta \Delta(\mathbf{y}', \mathbf{y})\Big).$$

Soft-max loss ℓ_β

- CRF and SSVM are special cases.
- Two parameters to tweak: β and C.
- Relevant literature: Gimpel and Smith, 2010; Hazan and Urtasun, 2010; Zhang, 2005.

Special case: binary classification

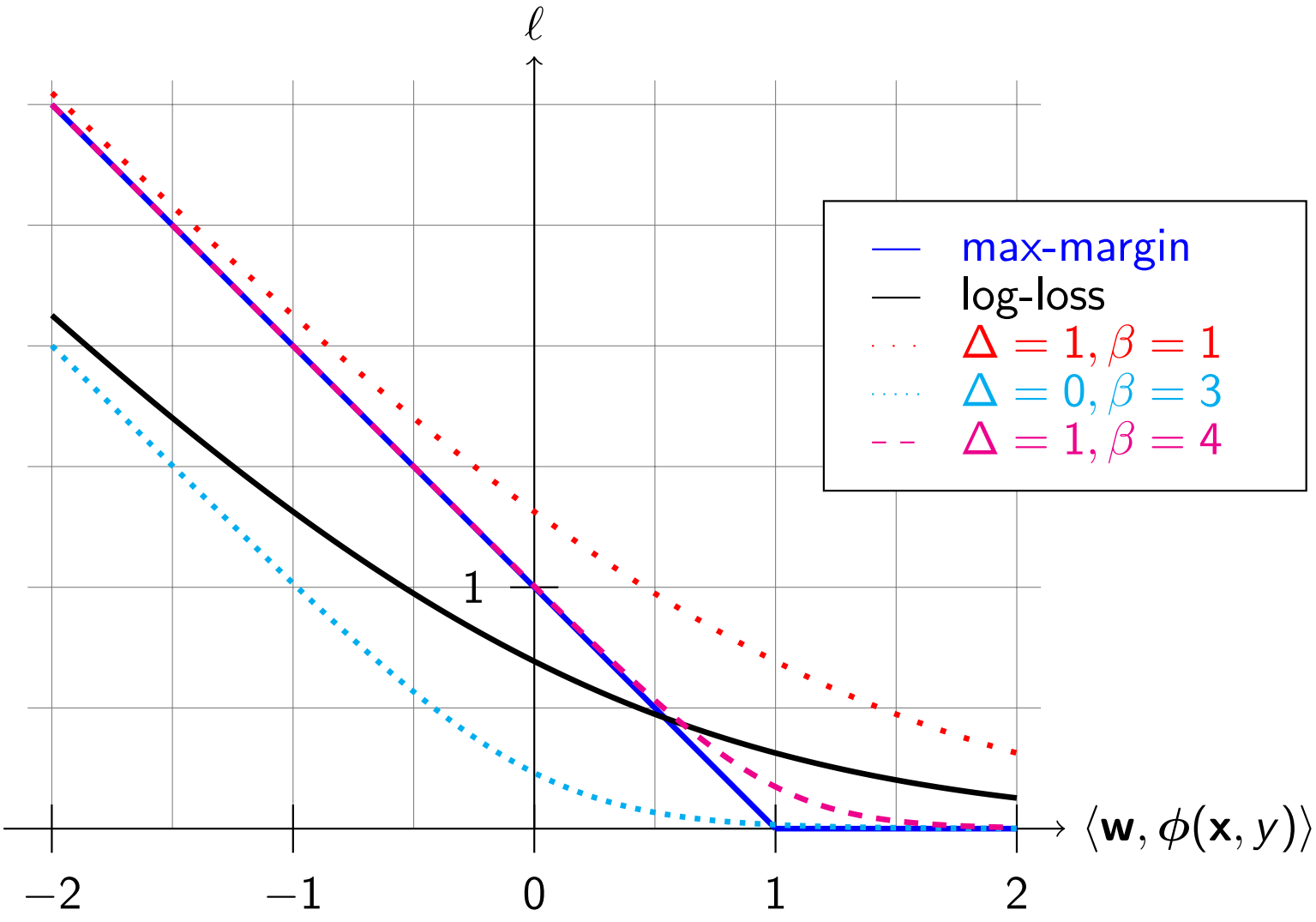


Figure: $\ell_\beta(\mathbf{w}, \mathbf{x}, y)$ for different β compared to log-loss and max-margin loss.

Limit case $\beta \to \infty$: SSVM

Max-margin loss

- For $\beta \to \infty$ only the maximally scoring $\mathbf{y}$ contribute to the partition sum.
- Similar to norms $\|\cdot\|_p$ for $p \to \infty$.
- Thus: recover max-margin:

$$\ell_\infty(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle + \max_{\mathbf{y}' \in \mathcal{Y}} \Big[\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}') \rangle + \Delta(\mathbf{y}', \mathbf{y}) \Big].$$

Posterior aspect

- Concentrate $P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w})$ on all the outputs $\mathbf{y}'$ with the maximum score.
- There can exist several maximum outputs.

But what about CRFs?

Margin term $\Delta(\mathbf{y}', \mathbf{y})$

- Margin term $\Delta(\mathbf{y}', \mathbf{y})$ in the loss $\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y})$.
- Not immediately obvious how to get rid of this term for some β settings.
- Is needed in order to identify CRFs as another special case.

Answered by the dual

Have a look at the dual of the empirical risk minimization with the loss $\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y})$.

Dual view

similar to Collins et al., 2008

The dual minimization problem is given by

$$\min_{\mathbf{u}} \quad \frac{1}{2C}\mathbf{u}^\top \mathbf{A}\mathbf{u} - \mathbf{b}^\top \mathbf{u} + \frac{1}{\beta}\sum_{n=1}^{N}\sum_{\mathbf{y}\in\mathcal{Y}} u_{n,\mathbf{y}} \log u_{n,\mathbf{y}}$$

$$\text{s.t.} \quad u_{n,\mathbf{y}} \geq 0 \quad \text{and} \quad \sum_{\mathbf{y}\in\mathcal{Y}} u_{n,\mathbf{y}} = 1 \ \forall \mathbf{y}, n.$$

$\mathbf{A}$ is given by $A_{(n_1,\mathbf{y}),(n_2,\mathbf{y}')} = \langle \mathbf{g}_{n_1,\mathbf{y}}, \mathbf{g}_{n_2,\mathbf{y}'} \rangle$. With

$$\mathbf{g}_{n,y} = \phi(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}) - \phi(\mathbf{x}^{(n)}, \mathbf{y}) \quad \text{and} \quad b_{n,\mathbf{y}} = \Delta(\mathbf{y}, \mathbf{y}^{(n)}).$$

A total of $N \cdot |\mathcal{Y}|$ dual variables are required. The primal and dual variables are related by

$$\mathbf{w} = \frac{1}{C}\sum_{n=1}^{N}\sum_{\mathbf{y}\in\mathcal{Y}} u_{n,\mathbf{y}}\mathbf{g}_{n,\mathbf{y}}.$$

Insights from the dual

Three terms: data, margin, entropy

$$\frac{1}{2C}\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u} - \mathbf{b}^\mathsf{T}\mathbf{u} + \frac{1}{\beta}\sum_{n=1}^{N}\sum_{\mathbf{y}\in\mathcal{Y}} u_{n,\mathbf{y}} \log u_{n,\mathbf{y}}$$

CRF vs. SVM

- $\beta \to \infty$: max-margin dual. C for regularization.
- β and C small: margin term vanishes $\to$ CRF dual. β/C for regularization.
- No additional scaling of margin term required.

Algorithmics

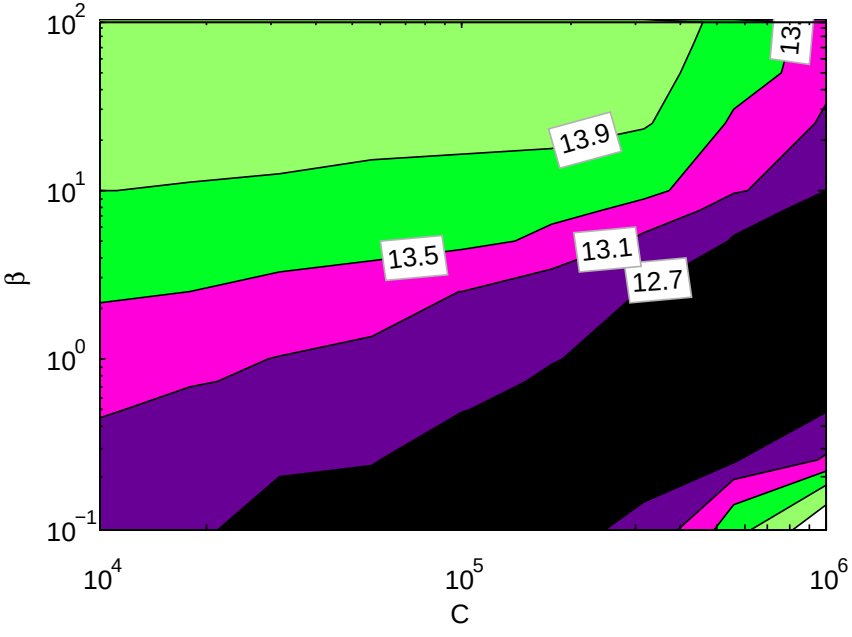
Probabilistic Inference

- Compute log-partition sum for loss.
- Compute marginals $P_\beta(y'_i|\mathbf{x}, \mathbf{w}, \mathbf{y})$ for gradient.
- Here: only focus on cases where this is efficiently possible.

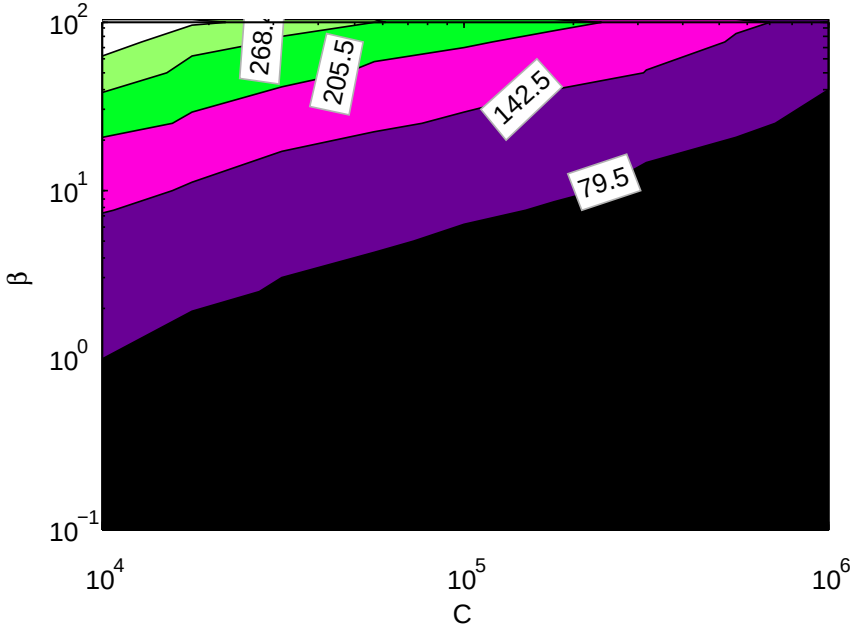
$\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y})$ is differentiable everywhere

- Can use standard quasi Newton methods like e.g., L-BFGS.
- Increasing $\beta \rightarrow$ problem becomes "almost" piece-wise linear.
- Thus: second-order gradient information will not help in optimization.

Test error and running time vs. β and C



test error



running time

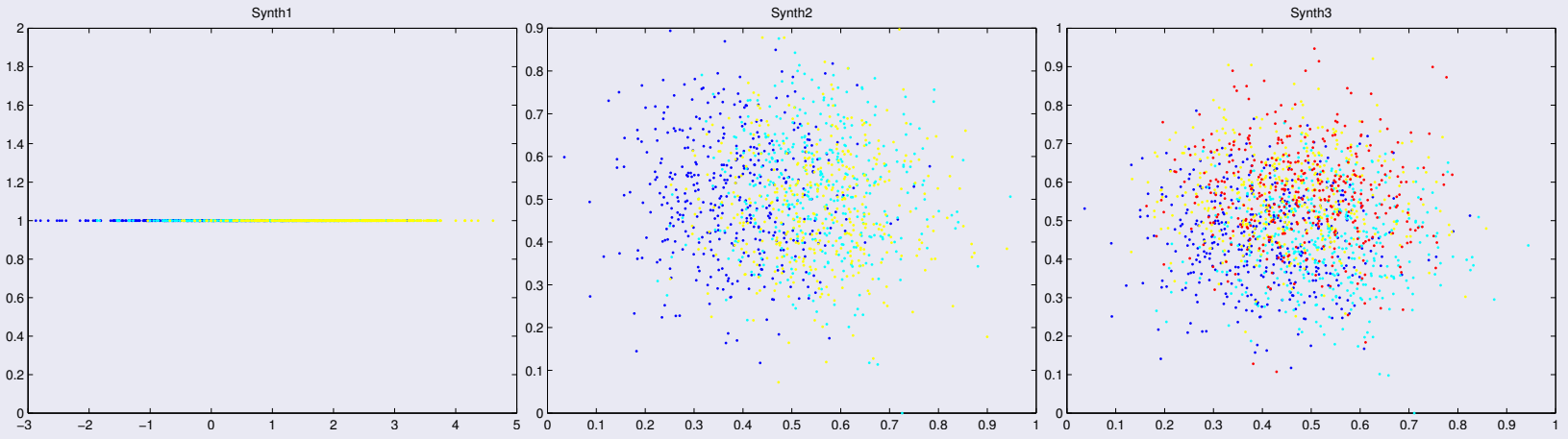
Synthetic multiclass data

Setup

3 configurations

- Synth1: Large overlap. Consistency should help.
- Synth2: okay to confuse classes. Margin term should help.
- Synth3: combination of Synth1 and Synth2.

Visualization of the data sets



Synthetic multiclass data

Results

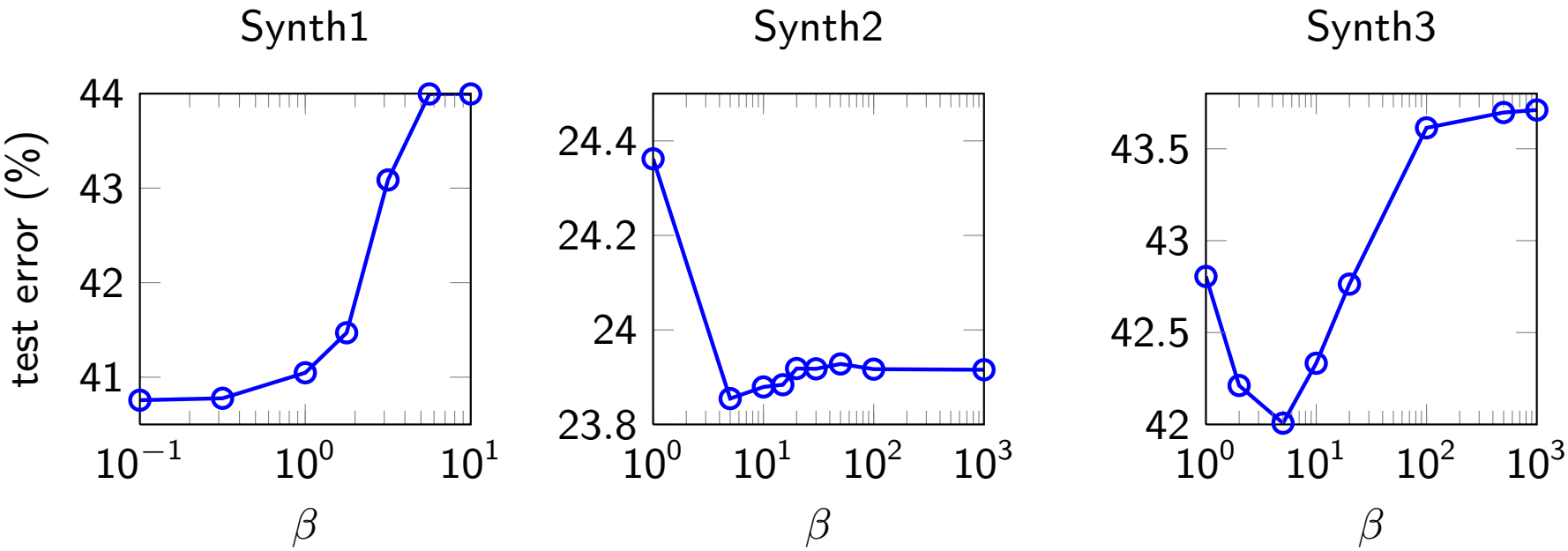
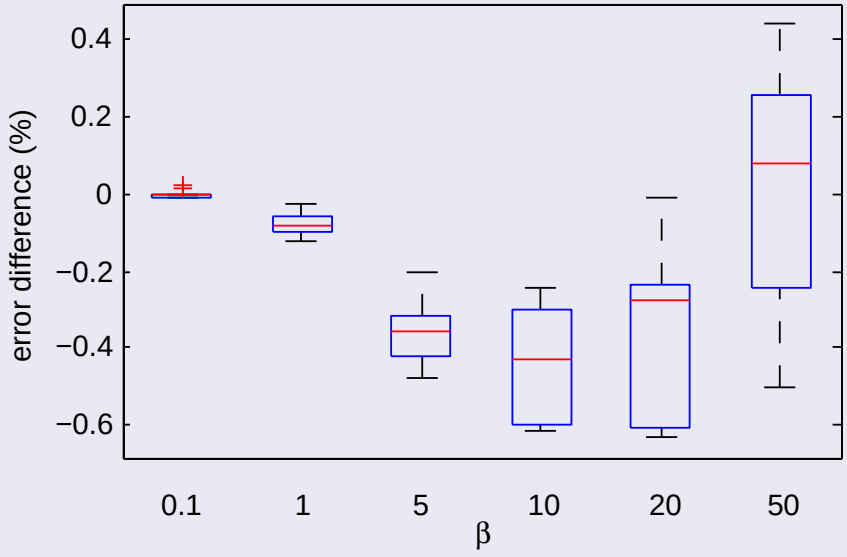


Figure: Results on the different synthetic multiclass datasets. Changing the parameter β leads to different test errors.

OCR data

Dataset from Taskar, Guestrin, and Koller, 2003

Result: around 20% test error



Summary

Take home message

CRF and structured SVM can be derived as special cases of a more general loss.

More details in the paper

- Slack rescaling.
- Hidden variables.

Summary

Take home message

CRF and structured SVM can be derived as special cases of a more general loss.

More details in the paper

- Slack rescaling.
- Hidden variables.

Thank you. Any questions?

References I

Collins, M. et al. (2008). "Exponentiated Gradient Algorithms for Conditional Random Fields and Max-Margin Markov Networks". In: *J. Mach. Learn. Res.* 9, pp. 1775–1822.

Gimpel, K. and N. Smith (2010). "Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions". In: *HLT*, pp. 733–736.

Hazan, Tamir and Raquel Urtasun (2010). "Approximated Structured Prediction for Learning Large Scale Graphical Models". In: *CoRR* abs/1006.2899.

Lafferty, J., A. McCallum, and F. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *ICML*.

Taskar, B., C. Guestrin, and D. Koller (2003). "Max-Margin Markov Networks". In: *NIPS*.

Teo, C. H. et al. (2009). "Bundle Methods for Regularized Risk Minimization". In: *J. Mach. Learn. Res.* submitted.

Tsochantaridis, I. et al. (2004). "Support vector machine learning for interdependent and structured output spaces". In: *ICML*, p. 104.

References II

Zhang, T. (2005). "Class-size Independent Generalization Analysis of Some Discriminative Multi-Category Classification". In: *NIPS*. Cambridge, MA.