Learning Low-order Models for Enforcing High-order Statistics

Patrick Pletscher, Pushmeet Kohli

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Microsoft Research

Higher-order Statistics

- Standard CRF models usually trained using simple *low-order* losses.
- In real-world often more complex *higher-order* losses used for evaluation.
- Goal here: Train classifier directly with this higher-order loss.
- Our work introduces a higher-order loss for which we can train structured SVMs *exactly*.

Model

Train a predictor of the form

$$\mathbf{f}_{\mathbf{w}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{Y}} E(\mathbf{y}, \mathbf{x}, \mathbf{w}).$$

$$E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle = \sum_{i\in\mathcal{V}} \psi_i(y_i, \mathbf{x}; \mathbf{w}^u) + \sum_{(i,j)\in\mathcal{E}} \psi_{ij}(y_i, y_j, \mathbf{x}; \mathbf{w}^p).$$

Max-margin Learning

The structured SVM considers the following quadratic program:

$$\begin{aligned}
\min_{\mathbf{w},\xi} \quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{n=1}^{N} \xi^n \\
\text{s.t.} \quad & \max_{\mathbf{y}\in\mathcal{Y}} [\langle \mathbf{w}, \phi(\mathbf{x}^n, \mathbf{y})\rangle + \Delta_{\mathbf{y}^n}(\mathbf{y})] - \langle \mathbf{w}, \phi(\mathbf{x}^n, \mathbf{y}^n)\rangle \geq \xi^n \quad \forall n \\
& \xi^n \geq 0.
\end{aligned}$$

- Optimizes convex upper bound on misclassification error.
- Loss $\Delta_{\mathbf{y}^n}(\mathbf{y})$ measures how bad it is to predict $\mathbf{y}$ instead of $\mathbf{y}^n$.
- Solved by the cutting planes algorithm.
- Line 5: Loss augmented inference.

Require: $(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^N, \mathbf{y}^N), \lambda, \epsilon, \Delta_{\mathbf{y}^*}(\cdot)$.

1: $S^n \leftarrow \emptyset$ for $n = 1, \ldots, N$.
2: **repeat**
3: **for** $n = 1, \ldots, N$ **do**
4: $H(\mathbf{y}) := \Delta_{\mathbf{y}^n}(\mathbf{y}) + \langle \mathbf{w}, \phi(\mathbf{x}^n, \mathbf{y}) - \phi(\mathbf{x}^n, \mathbf{y}^n)\rangle$
5: compute $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} H(\mathbf{y})$
6: compute $\xi^n = \max\{0, \max_{\mathbf{y}\in S^n} H(\mathbf{y})\}$
7: **if** $H(\hat{\mathbf{y}}) > \xi^n + \epsilon$ **then**
8: $S^n \leftarrow S^n \cup \{\hat{\mathbf{y}}\}$
9: $\mathbf{w} \leftarrow$ optimize primal over $\bigcup_n S^n$
10: **end if**
11: **end for**
12: **until** no S^n has changed during iteration

Loss Augmented Inference

- Need to efficiently solve the problem:

$$\min_{\mathbf{y}} E(\mathbf{y}, \mathbf{x}, \mathbf{w}) - \Delta_{\mathbf{y}^*}(\mathbf{y}).$$

- Notice the negative sign!
- We assume y_i is binary and $E(\mathbf{y}, \mathbf{x}, \mathbf{w})$ is submodular. Therefore: energy minimization in the original model is exactly solvable.

Loss Functions

- Should reflect scoring used for evaluation.
- But at the same time loss augmented inference should also be tractable!
- In practice for many segmentation problems Hamming loss is used:

$$\Delta^{hamming}_{\mathbf{y}^*}(\mathbf{y}) = \sum_{i\in\mathcal{V}} y_i \neq y_i^*.$$

  Loss augmented inference has same complexity as inference for original model.
- Only modifies the unaries.
- A low-order loss. What about higher-order losses?
- Here we study the label-count loss:

$$\Delta^{count}_{\mathbf{y}^*}(\mathbf{y}) = \left|\sum_{i\in\mathcal{V}} y_i - \sum_{i\in\mathcal{V}} y_i^*\right|.$$

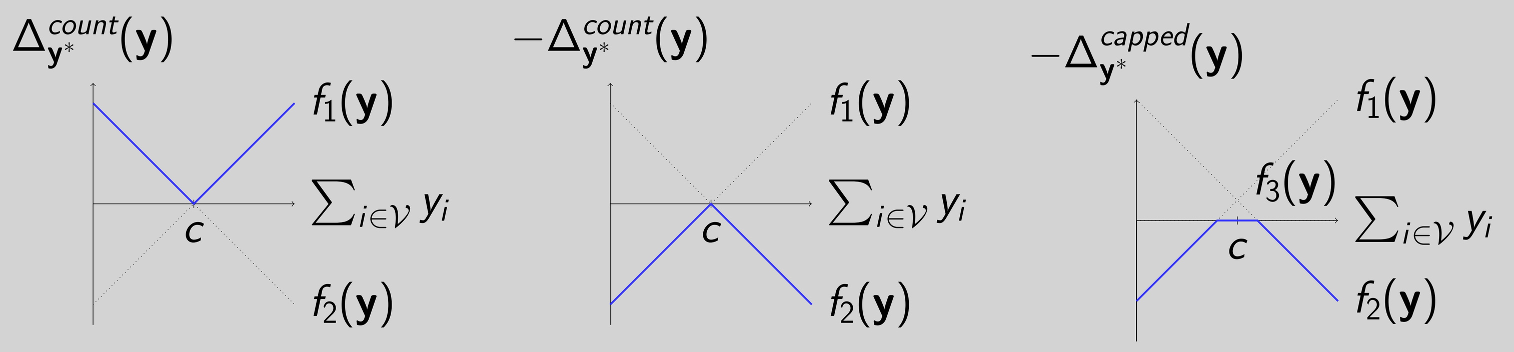
- Useful if we are only interested in predicting the number of foreground pixels, but not their location.
- Unfortunately label-count loss no longer factorizes!

Lower and Upper Envelopes

- Many higher order functions can be represented as:

$$f^h(\mathbf{y}) = \otimes_{q\in\mathcal{Q}} f^q(\mathbf{y})$$

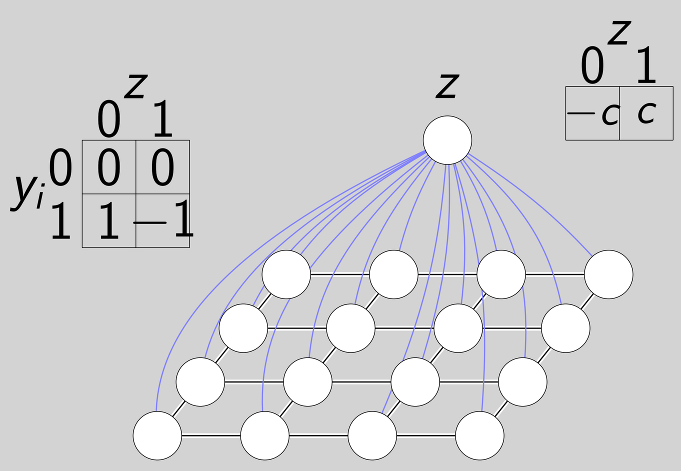
  where $\otimes = \{\max, \min\}$, and $\mathcal{Q}$ indexes a set of linear functions.
- min: lower envelope, max: upper envelope.
- Inference for upper envelope substantially more difficult (min-max).
- Label-count is upper envelope representable.
- Fortunately, negative sign makes loss lower envelope representable:



Label-count Loss Augmented Inference

Obtain the pairwise minimization problem:

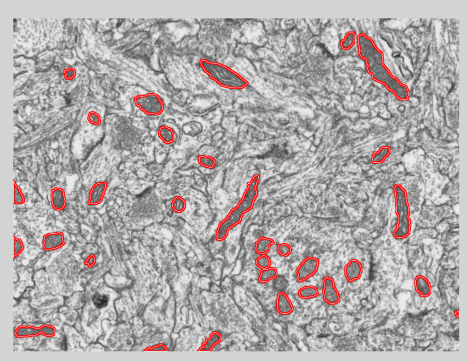
$$\min_{\mathbf{y}, z\in\{0,1\}} E(\mathbf{y}, \mathbf{x}, \mathbf{w}) + 2z\left(\sum_{i\in\mathcal{V}} y_i^* - \sum_{i\in\mathcal{V}} y_i\right) + \sum_{i\in\mathcal{V}} y_i - \sum_{i\in\mathcal{V}} y_i^*.$$

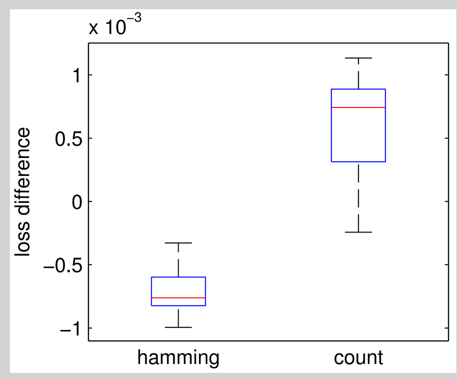


- Can be solved by standard graph-cut with an auxiliary variable.
- Or alternatively by two standard graph-cut calls with modified unaries.
- Former approach also works if label-count loss for several parts is desired.
- Iterative breadth-first search graph-cut leads to better performance.

Cell Segmentation

- Goal: Counting number of mitochondria cell pixels in an electroscopic image.
- Right: Hamming loss trained model minus count-loss trained model.





Background-Foreground Segmentation

H (c: 0.077, h: 0.077) C (c: 0.037, h: 0.040) Ground-truth

H (c: 0.069, h: 0.069) C (c: 0.012, h: 0.124) Ground-truth

Eval \ Train		Hamming better (%)	Count better (%)
4/S	Hamming	52.1 ± 7.0	47.9 ± 7.0
	Count	33.8 ± 8.3	66.2 ± 8.3
4/D	Hamming	39.4 ± 6.1	60.6 ± 6.1
	Count	29.6 ± 8.3	70.4 ± 8.3
8/S	Hamming	48.2 ± 11.9	51.8 ± 11.9
	Count	32.0 ± 13.1	68.0 ± 13.1
8/D	Hamming	50.0 ± 9.2	50.0 ± 9.2
	Count	40.5 ± 14.3	59.5 ± 14.3

Conclusions

- Max-margin learning with the label-count loss can be done exactly.
- Leads to better results if only interested in the number of foreground pixels.
- Also see Danny Tarlow's poster here at AISTATS.

References

- I. Tsochantaridis et al. (2005). "Large Margin Methods for Structured and Interdependent Output Variables". In: *Journal of Machine Learning Research* 6, pp. 1453–1484
- B. Taskar, C. Guestrin, and D. Koller (2003). "Max-Margin Markov Networks". In: *Advances in Neural Information Processing Systems (NIPS)*
- P. Kohli and M. P. Kumar (2010). "Energy Minimization for linear Envelope MRFs". In: *CVPR*
- A. Blake et al. (2004). "Interactive Image Segmentation Using an Adaptive GMMRF Model". In: *ECCV*, pp. 428–441
- A. V. Goldberg et al. (2011). "Maximum Flows by Incremental Breadth-First Search". In: *ESA*, pp. 457–468
- V. Lempitsky and A. Zisserman (2010). "Learning To Count Objects in Images". In: *NIPS*
- D. Tarlow and R. Zemel (2011). "Big and Tall: Large Margin Learning with High Order Losses". In: *CVPR 2011 Workshop on Inference in Graphical Models with Structured Potentials*
- D. Tarlow and R. Zemel (2012). "Structured Output Learning with High Order Loss Functions". In: *AISTATS*