

Structured Output Prediction: Setting

- Predict $\mathbf{y} \in \mathcal{Y}$ for a given input variable $\mathbf{x} \in \mathcal{X}$.
- Dependencies between y_i specified by parameterized graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
- Parameters are denoted by \mathbf{w} .
- The score (negative energy) of (\mathbf{x}, \mathbf{y}) is given by $\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle$.
- $\phi(\mathbf{x}, \mathbf{y})$ the sufficient statistics follow from the graphical model and its parameterization.

Learning

Conditional Random Field (CRF) models the posterior distribution:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \exp(\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle) \quad Z(\mathbf{x}, \mathbf{w}) = \sum_{\mathbf{y} \in \mathcal{Y}} \exp(\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle)$$

Regularized Maximum Likelihood Learning:

$$\min_{\mathbf{w}} \frac{1}{N} \left[\sum_{n=1}^N -\langle \mathbf{w}, \phi(\mathbf{x}^n, \mathbf{y}^n) \rangle + \log Z(\mathbf{x}^n, \mathbf{w}) \right] + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

⇒ Need to compute the partition sum $Z(\mathbf{x}, \mathbf{w})!$

Prediction

Two approaches for given $P(\mathbf{y}|\mathbf{x})$. Correspond to different loss functions in a minimum Bayes risk framework.

- MAP prediction. Well-studied setting (graph-cut, max-product, ...)

$$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle \Leftrightarrow \text{zero-one loss}$$

- max-marginal (MPM). More challenging (often done by Gibbs sampling)

$$y_i^* = \operatorname{argmax}_{y_i} P(y_i|\mathbf{x}) \Leftrightarrow \text{Hamming loss}$$

Lower Bounding the Structured Output Loss

Given a partition of the variables \mathcal{V} into two sets \mathcal{A} and \mathcal{B} . Trivial lower bound by summing only over a subset $\underline{\mathcal{Y}}_{\mathcal{B}} \subseteq \mathcal{Y}_{\mathcal{B}}$:

$$Z(\mathbf{x}, \mathbf{w}) \geq \sum_{\mathbf{y}_{\mathcal{B}} \in \underline{\mathcal{Y}}_{\mathcal{B}}} \sum_{\mathbf{y}_{\mathcal{A}} \in \mathcal{Y}_{\mathcal{A}}} \exp(\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle) =: Z(\mathbf{x}, \mathbf{w}, \mathcal{B}, \underline{\mathcal{Y}}_{\mathcal{B}})$$

Can do this for several different partitions $\mathcal{D} = \{(\mathcal{A}_1, \mathcal{B}_1), \dots, (\mathcal{A}_M, \mathcal{B}_M)\}$ and corresponding states $\mathcal{Z} = \{\underline{\mathcal{Y}}_{\mathcal{B}_1}, \dots, \underline{\mathcal{Y}}_{\mathcal{B}_M}\}$. Let $Z^m := Z(\mathbf{x}, \mathbf{w}, \mathcal{B}_m, \underline{\mathcal{Y}}_{\mathcal{B}_m})$.

Combining the bounds to get new lower bounds:

- Arithmetic mean:

$$Z^{a, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w}) := \frac{1}{M} \sum_{m=1}^M Z^m$$

- Geometric mean:

$$Z^{g, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w}) := \left(\prod_{m=1}^M Z^m \right)^{1/M}$$

- Maximum (not differentiable w.r.t. \mathbf{w}):

$$Z^{m, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w}) := \max_m Z^m$$

Relation between the three bounds:

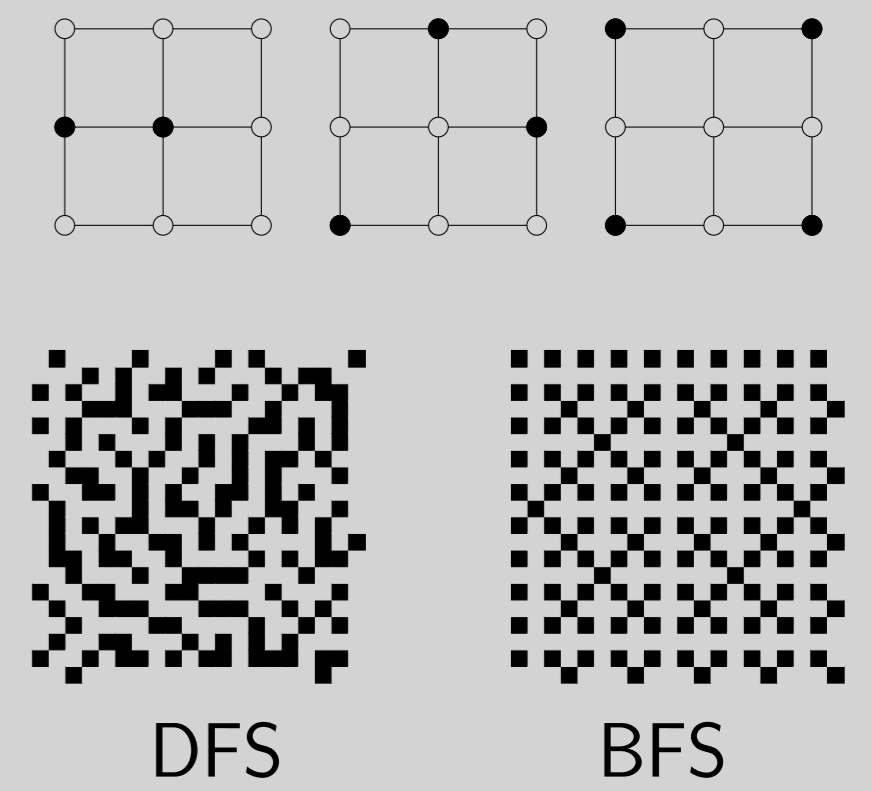
$$Z^{m, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w}) \geq Z^{a, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w}) \geq Z^{g, \mathcal{D}, \mathcal{Z}}(\mathbf{x}, \mathbf{w})$$

Composite Likelihood & Non-local Contrastive Divergence

- Inspired by composite likelihood and pseudolikelihood.
- Geometric average and $\underline{\mathcal{Y}}_{\mathcal{B}} = \{\mathbf{y}^n\}$ recovers composite likelihood.
- Pseudolikelihood recovered by particular decomposition.
- Similar to non-local contrastive divergence, but more efficient due to the partition.

Part: Minimum Feedback Vertex Set

- Choose forest-shaped partition $(\mathcal{A}, \mathcal{B})$.
- Like this summation over $\mathcal{Y}_{\mathcal{A}}$ feasible.
- Assumption: all nodes equally important.
- Choose a *minimum feedback vertex set*.
- Greedy randomized growing of forests.
- Breadth-first vs. depth-first variant.
- BFS close to optimal for 4-connected grid.



Clamp: Marginal MAP

- Goal: find state to include in $\underline{\mathcal{Y}}_{\mathcal{B}}$.
- Greedy approach: include state which increases the lower bound the most:

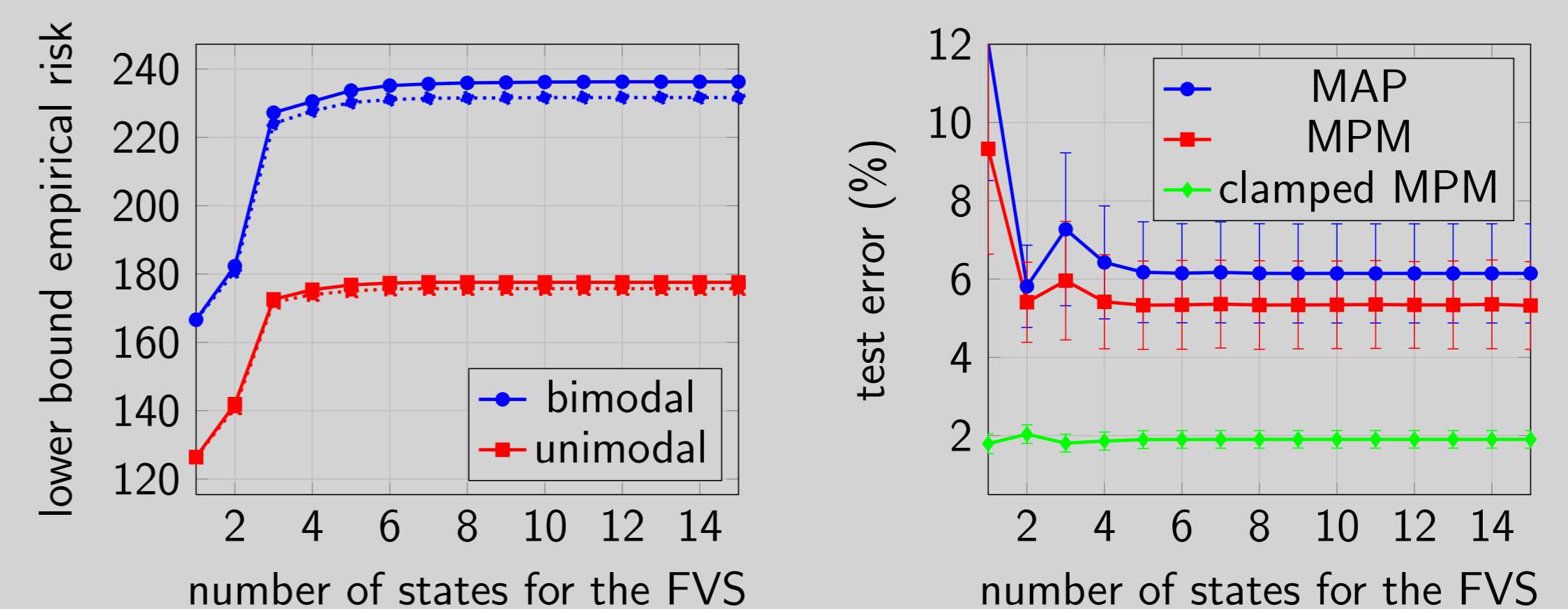
$$\mathbf{y}_{\mathcal{B}}^* = \operatorname{argmax}_{\mathbf{y}_{\mathcal{B}} \in \underline{\mathcal{Y}}_{\mathcal{B}}} \sum_{\mathbf{y}_{\mathcal{A}} \in \mathcal{Y}_{\mathcal{A}}} \exp(\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle)$$

- The marginal MAP problem.
- Recent message-passing algorithms for marginal MAP which include max-product and sum-product updates.
- Alternative: simply use MAP algorithm. Advantage: More efficient!

Part & Clamp Learning Algorithms

- Efficient computations: two passes through the tree for each clamping state in $\underline{\mathcal{Y}}_{\mathcal{B}}$ and decomposition.
- Batch Learning (cutting planes like):
 1. Full parameter learning using L-BFGS for current bound.
 2. Tighten bound for each example with current parameters.
 3. Repeat.
- Online Learning (stochastic gradient descent):
 1. Sample an example.
 2. Tighten bound for this particular example with current parameters.
 3. SGD step.
 4. Repeat.
- Budget version: keep size $\underline{\mathcal{Y}}_{\mathcal{B}}$ within a budget.

Experiment: Binary Image Denoising



| Prediction | Train | Pseudo-likelihood | Composite likelihood | Contrastive divergence | Part & Clamp batch | Part & Clamp online |
|-------------|--------------|-------------------|----------------------|------------------------|--------------------|---------------------|
| bimodal | | | | | | |
| MAP | 15.58 ± 4.11 | 12.02 ± 3.50 | 7.01 ± 1.71 | 6.14 ± 1.27 | 5.16 ± 0.77 | |
| MPM | 11.86 ± 3.40 | 9.33 ± 2.69 | 6.72 ± 1.67 | 5.32 ± 1.12 | 5.20 ± 0.80 | |
| clamped MPM | 1.77 ± 0.25 | 1.80 ± 0.26 | 1.96 ± 0.22 | 1.90 ± 0.22 | 2.23 ± 0.25 | |
| unimodal | | | | | | |
| MAP | 5.28 ± 1.47 | 4.43 ± 1.26 | 2.39 ± 0.47 | 2.40 ± 0.50 | 2.40 ± 0.46 | |
| MPM | 4.13 ± 1.18 | 3.66 ± 0.96 | 2.37 ± 0.45 | 2.40 ± 0.42 | 2.42 ± 0.43 | |
| clamped MPM | 0.98 ± 0.22 | 1.01 ± 0.21 | 1.05 ± 0.21 | 1.03 ± 0.22 | 1.17 ± 0.23 | |

Conclusions

- Simple lower bound that leads to good parameter estimates in practice.
- Generalizes pseudolikelihood and composite likelihood.
- Efficient if $|\underline{\mathcal{Y}}_{\mathcal{B}}|$ small, go through graph twice for each state.
- Would not expect this to work well in settings where posteriori has large entropy.

References

- J. Lafferty, A. McCallum, and F. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *ICML*, pp. 282–289
- B. Lindsay (1988). "Composite Likelihood Methods". In: *Contemporary Mathematics* 80
- J. Besag (1975). "Statistical Analysis of Non-Lattice Data". In: *The Statistician* 24.3, pp. 179–195
- D. Vickrey, C. Lin, and D. Koller (2010). "Non-Local Contrastive Objectives". In: *ICML*
- G. Hinton (2000). "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Computation* 14.8, pp. 1771–1800
- E. Horvitz, J. Suermondt, and G. Cooper (1989). "Bounded Conditioning: Flexible Inference for Decisions Under Scarce Resources". In: *Proceedings of Conference on Uncertainty in Artificial Intelligence*
- Jiarong Jiang, Piyush Rai, and Hal Daumé III (2011). "Message-Passing for Approximate MAP Inference with Latent Variables". In: *NIPS*
- Q. Liu and A. Ihler (2011). "Variational Algorithms for Marginal MAP". In: *UAI*
- S. Kumar and M. Hebert (2006). "Discriminative Random Fields". In: *IJCV* 68.2, pp. 179–201
- P. Pletscher, C. Ong, and J. Buhmann (2010). "Entropy and Margin Maximization for Structured Output Learning". In: *ECML*