Learning Deep Latent-variable MRFs with Amortized Bethe Free Energy Minimization

Sam Wiseman

Motivating Questions

- How good are popular approximate inference methods at learning deep structured models with discrete latent variables?
- Are there learning objectives for such models that don't require sampling-based gradient estimators?

Main Idea

- Use a learning objective based on the Bethe free energy (BFE) approximation to the partition function.
- The BFE approximation can be computed *exactly* for many models of interest.
- This is only an advantageous approximation for undirected models (i.e., MRFs).

Bethe Approximations

- **Notation:**
  - Denote a factor graph by $\mathcal{G} = (\mathcal{V} \cup \mathcal{F}, \mathcal{E})$.
  - $\boldsymbol{x}$: observed variables in $\mathcal{V}$
  - $\boldsymbol{z}$: latent variables in $\mathcal{V}$
  - $\boldsymbol{x}_\alpha$: subvector of $\boldsymbol{x}$ associated with factor α
  - Ψ_α: potential func. associated with factor α
  - $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}'} \sum_{\boldsymbol{z}'} \Pi_\alpha \Psi_\alpha(\boldsymbol{x}'_\alpha, \boldsymbol{z}'_\alpha; \boldsymbol{\theta})$
  - $Z(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}'} \Pi_\alpha \Psi_\alpha(\boldsymbol{x}_\alpha, \boldsymbol{z}'_\alpha; \boldsymbol{\theta})$
- **BFE** (Bethe, 1935; Yedidia et al., 2001):

$$F(\boldsymbol{\tau}) = \mathrm{KL}[q_{\boldsymbol{\tau}}(\boldsymbol{x}, \boldsymbol{z}) || P(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})] - \log Z(\boldsymbol{\theta})$$

$$= \sum_\alpha \sum_{\boldsymbol{x}'_\alpha, \boldsymbol{z}'_\alpha} \boldsymbol{\tau}_\alpha(\boldsymbol{x}'_\alpha, \boldsymbol{z}'_\alpha) \log \frac{\boldsymbol{\tau}_\alpha(\boldsymbol{x}'_\alpha, \boldsymbol{z}'_\alpha)}{\Psi_\alpha(\boldsymbol{x}'_\alpha, \boldsymbol{z}'_\alpha)}$$

$$- \sum_{v \in \mathcal{V}} (\mathrm{ne}(x_v) - 1) \sum_{x'_v} \boldsymbol{\tau}(x'_v) \log \boldsymbol{\tau}(x'_v)$$

  - $\boldsymbol{\tau}_\alpha(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$: (pseudo) marginal associated with factor α
- **Partition Function Approximation**:
  - Let $\mathcal{C}$ contain all locally consistent, (pseudo) marginals.
  - For a tree, $\min_{\boldsymbol{\tau} \in \mathcal{C}} F(\boldsymbol{\tau}) = -\log Z(\boldsymbol{\theta})$.
  - Otherwise, $\min_{\boldsymbol{\tau} \in \mathcal{C}} F(\boldsymbol{\tau}) \approx -\log Z(\boldsymbol{\theta})$.
  - Loopy BP finds stationary points of $\min_{\boldsymbol{\tau} \in \mathcal{C}} F(\boldsymbol{\tau})$ (Yedidia et al., 2001).

Why the BFE is Attractive

- Only linear in the number of factors!
- But, having a large number of low-degree factors is only interesting for MRFs (c.f., products of experts (Hinton, 2002) and Figure 1).

Flavors of High-order HMM

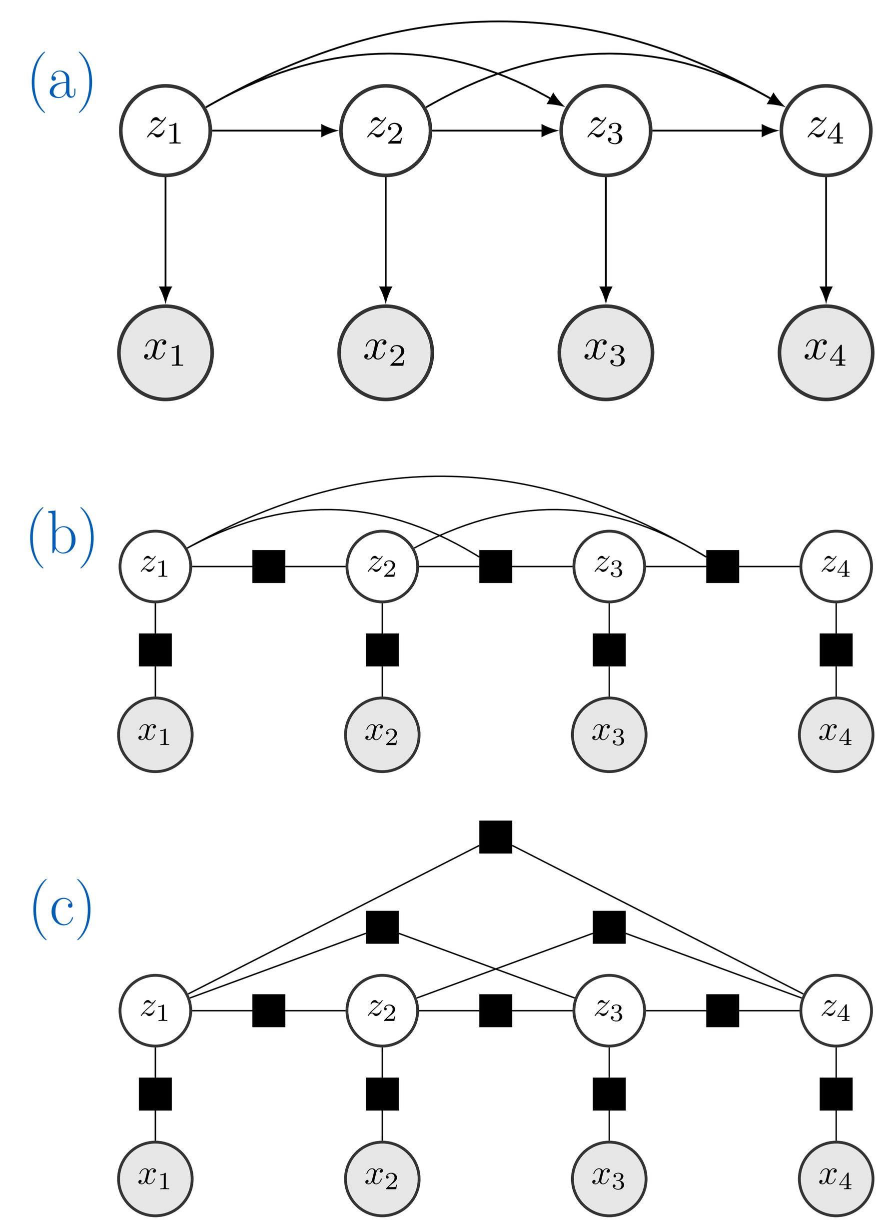

Figure 1

A BFE-based Objective

- Replace clamped and unclamped partition functions in the log marginal with their BFE approximations:

$$-\log p(\boldsymbol{x}; \boldsymbol{\theta}) = -Z(\boldsymbol{x}, \boldsymbol{\theta}) + Z(\boldsymbol{\theta})$$

$$\approx \min_{\boldsymbol{\tau}_{\boldsymbol{x}} \in \mathcal{C}} F_{\boldsymbol{x}}(\boldsymbol{\tau}_{\boldsymbol{x}}) - \min_{\boldsymbol{\tau} \in \mathcal{C}} F(\boldsymbol{\tau})$$

- Gives rise to a saddle-point objective:

$$\min_{\boldsymbol{\theta}} \ell_F(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} [\min_{\boldsymbol{\tau}_{\boldsymbol{x}} \in \mathcal{C}} F_{\boldsymbol{x}}(\boldsymbol{\tau}_{\boldsymbol{x}}) - \min_{\boldsymbol{\tau} \in \mathcal{C}} F(\boldsymbol{\tau})]$$

$$= \min_{\boldsymbol{\theta}, \boldsymbol{\tau}_{\boldsymbol{x}} \in \mathcal{C}} \max_{\boldsymbol{\tau} \in \mathcal{C}} [F_{\boldsymbol{x}}(\boldsymbol{\tau}_{\boldsymbol{x}}) - F(\boldsymbol{\tau})]$$

- We train inference networks $f, f_{\boldsymbol{x}}$ to output approximate minimizers of $F(\boldsymbol{\tau})$ and $F_{\boldsymbol{x}}(\boldsymbol{\tau}_{\boldsymbol{x}})$.

Constraining Optimization

- Local consistency and sum-to-one constraints are linear and can be eliminated:
  - $\boldsymbol{\tau} \leftarrow \boldsymbol{V}\boldsymbol{V}^{+} f(\mathcal{G}; \boldsymbol{\phi}) + \hat{\boldsymbol{\tau}}$, where $\boldsymbol{V}$ is a basis for the null space of the constraint matrix and $\hat{\boldsymbol{\tau}}$ is feasible.
- We impose a linear penalty on negative elements of $\boldsymbol{V}\boldsymbol{V}^{+} f(\mathcal{G}; \boldsymbol{\phi}) + \hat{\boldsymbol{\tau}}$.

Experiments

- Model: 2nd or 3rd order neural HMM, K=20
- Data: Penn Treebank sentences, length ≤ 20
- We compare BFE minimization with VAE variants on true held-out PPL.

Model Parameterizations

Directed/VAE models:

- Neural directed HMM: emission and transition distributions parameterized by residual feed-forward nets.
- Mean-field (MF) inference net: BLSTM over input into linear decoder for each token.
- First-order (FO) inference net: 1st order neural HMM, but also conditions on averaged BLSTM states of input.

MRF/Bethe models:

- Pairwise HMM MRF: transition factors are residual feed-forward function of distance; emissions are the same as directed version.
- Bethe inference net: BLSTM over discrete representation of MRF edges and associated potentials into linear to predict marginals.

Results

	PPL	ELBO/ℓ_F
2nd Order HMM		
MF VAE + BL	348.06	7318.82
MF IWAE, $L=5$	338.07	7224.11
MF IWAE, $L=10$	328.42	7087.25
FO HMM VAE	286.40	298.08
Exact	160.62	N/A
ℓ_F + true marginals	151.78	125.10
$\ell_F + f, f_x$	243.43	308.33
Exact	149.27	N/A
3rd Order HMM		
MF VAE + BL	350.98	7382.30
MF IWAE, $L=5$	346.57	7290.14
MF IWAE, $L=10$	335.24	7273.60
FO HMM VAE	270.25	274.12
Exact	159.78	N/A
ℓ_F + true marginals	170.07	152.88
$\ell_F + f, f_x$	253.35	254.54
Exact	141.71	N/A

ELBO Gradient Variance