CS281A/Stat241A Lecture 24

Variational Methods

Peter Bartlett

Announcements

- Poster sessions will be on Tue Dec 1 (Stat241A) and Thu Dec 3 (CS281A), here, 11-12:30. Please attend both sessions.
- Project reports are due at 5pm on Friday December 4. In the box outside 723 SD Hall. This deadline is firm.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Variational Methods

- Represent quantity of interest as solution to (or value of) an optimization problem.
- Then approximate the optimization problem:
  - Approximate the constraint set.
  - Approximate the criterion.

Variational Approach: Ingredients

1. Exponential family representation of graphical model.
2. Mean parameters μ correspond to desired marginal (conditional) clique probabilities.
3. Realizable mean parameter set $\mathcal{M}$ (marginal polytope).
4. Inference as optimization problem via conjugate dual representation of log normalization.

Variational Approach: Ingredients

Exponential family:

$$p(x) = h(x)\exp\left(\langle\theta,\phi(x)\rangle - A(\theta)\right).$$

Example: pairwise MRF $(x_v \in \{0,1,\ldots,r-1\})$.

$$p(x) = \exp\Bigg(\sum_{v\in V}\sum_i \theta_{v,i}\mathbf{1}[x_v = i] + \sum_{\{u,v\}\in E}\sum_{i,j}\theta_{u,i;v,j}\mathbf{1}[x_u = i]\mathbf{1}[x_v = j]\Bigg),$$

for $\theta \in \Omega = \{\theta : A(\theta) < \infty\} = \mathbb{R}^{r|V|+r^2|E|}$.

Variational Approach: Ingredients

1. Exponential family representation of graphical model.
2. Mean parameters μ correspond to desired marginal (conditional) clique probabilities.
3. Realizable mean parameter set $\mathcal{M}$ (marginal polytope).
4. Inference as optimization problem via conjugate dual representation of log normalization.

Variational Approach: Ingredients

- Define the set $\mathcal{M}$ of *realizable mean parameters* (marginal polytope) as

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \forall \alpha,\ \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha \right\}$$

if $\mathcal{X}$ is finite:

$$= \text{co}\{\phi(x) : x \in \mathcal{X}\},$$

where co represents the convex hull.

- Example: pairwise MRF $(x_v \in \{0, 1, \ldots, r-1\})$.

$$\mu_v = \mathbb{E}_p \mathbf{1}[X_v = i] = \Pr(X_v = i)$$

$$\mu_{u,v} = \mathbb{E}_p \mathbf{1}[x_u = i] \mathbf{1}[x_v = j] = \Pr(X_u = i,\ X_v = j).$$

Variational Approach: Ingredients

1. Exponential family representation of graphical model.
2. Mean parameters μ correspond to desired marginal (conditional) clique probabilities.
3. Realizable mean parameter set $\mathcal{M}$ (marginal polytope).
4. Inference as optimization problem via conjugate dual representation of log normalization.

Variational Approach: Ingredients

The conjugate dual of the log normalization A is

$$A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \mu, \theta \rangle - A(\theta) \right) = -H(p_{\theta(\mu)}),$$

where $\mu \in \mathbb{R}^d$ for $\Omega \subseteq \mathbb{R}^d$ and $H(p)$ is the entropy.
For $\theta \in \Omega$,

$$\begin{aligned} A(\theta) &= \sup_{\mu \in \mathcal{M}} \left(\langle \theta, \mu \rangle - A^*(\mu) \right) \\ &= \sup_{\mu \in \mathcal{M}} \big(\langle \theta, \mu \rangle + H(p_{\theta(\mu)}) \big) . \end{aligned}$$

Variational Approach: Ingredients

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left(\langle \theta, \mu \rangle + H(p_{\theta(\mu)}) \right).$$

- Solving this optimization problem gives the value $A(\theta)$ and the mean parameters $\mu = \mathbb{E}_\theta[\phi(X)]$.
- These correspond to the expectation of the sufficient statistics. (conditional expectation, if evidence has been incorporated).
- For example, for discrete pairwise MRFs, they give the marginal singleton and pairwise distributions.

Variational Methods

- Represent quantity of interest as solution to (or value of) an optimization problem:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left(\langle \theta, \mu \rangle + H(p_{\theta(\mu)}) \right) .$$

- Then approximate the optimization problem:
  - Approximate the constraint set, $\mathcal{M}$.
  - Approximate the criterion, $\langle \theta, \mu \rangle + H(p_{\theta(\mu)})$.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Mean Field Algorithm

Consider the Ising model:

$$x_u \in \{0, 1\}.$$

$$\psi_{u,v}(x_u, x_v) = \exp\left(\theta_{u,v} x_u x_v\right),$$

$$\psi_v(x_v) = \exp\left(\theta_v x_v\right).$$

$$p(x) = \exp\left(\sum_{v \in V} \theta_v x_v + \sum_{\{u,v\} \in E} \theta_{u,v} x_u x_v - A(\theta)\right).$$

Mean Field Algorithm

- Consider Gibbs sampling, and replace X_u by its expectation:

$$\mu_v := \frac{1}{1 + \exp\left(-\theta_v - \sum_{u \in N(v)} \theta_{v,u} \mu_u\right)}.$$

- *Naive mean field algorithm* for the Ising model.

Variational Interpretation

- Consider the optimization problem

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left(\langle \theta, \mu \rangle + H(p_{\theta(\mu)}) \right).$$

- If we approximate $\mathcal{M}$ with the smaller set:

$$\hat{\mathcal{M}} = \{ \mu \in \mathcal{M} : \mu_{u,v} = \mu_u \mu_v \}.$$

- Then we have

$$A(\theta) \geq \sup_{\mu \in \hat{\mathcal{M}}} \left(\sum_{v \in V} \theta_v \mu_v + \sum_{\{u,v\} \in E} \theta_{u,v} \mu_u \mu_v \right.$$
$$\left. - \sum_{v \in V} \left(\mu_v \log \mu_v + (1 - \mu_v) \log(1 - \mu_v) \right) \right).$$

Variational Interpretation

- Coordinate ascent in μ_v gives

$$\mu_v = \frac{1}{1 + \exp\left(-\theta_v - \sum_{u \in N(v)} \theta_{u,v}\mu_u\right)},$$

which is the mean field update.

- The criterion is strictly concave in each coordinate μ_v.
- But it is not a concave maximization problem...

Mean Field $\hat{\mathcal{M}}$ is Not Convex

$$\mathcal{M} = \text{co}\{\phi(x) : x \in \mathcal{X}\},$$

$$\hat{\mathcal{M}} = \{\mu \in \mathcal{M} : \mu_{u,v} = \mu_u \mu_v\}.$$

- $\hat{\mathcal{M}} \subseteq \mathcal{M}$.
- $\phi(x) \in \hat{\mathcal{M}}$:
  Place all mass on x. For such a distribution, $\mu_v \in \{0, 1\}$, and so $\mu_{u,v} = \mu_u \mu_v$.
- But $\mathcal{M}$ is the convex hull of these points in $\hat{\mathcal{M}}$.
- So if $\hat{\mathcal{M}}$ is a proper subset of $\mathcal{M}$, it must be nonconvex.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Mean Field and KL-Divergence

For the exponential family

$$p(x) = h(x) \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right),$$

consider two parameters θ^1 and θ^2.
The KL-divergence between the distributions p_{θ^1} and p_{θ^2} (with mean parameters μ^1 and μ^2) is

$$\begin{aligned}
D(\theta^1; \theta^2) &= \mathbb{E}_{\theta^1} \log \frac{p_{\theta^1}(X)}{p_{\theta^2}(X)} \\
&= \langle \mu^1, \theta^1 - \theta^2 \rangle - A(\theta^1) + A(\theta^2) \\
&= A(\theta^2) - \left(A(\theta^1) + \langle \mu^1, \theta^2 - \theta^1 \rangle \right).
\end{aligned}$$

Mean Field and KL-Divergence

$$D(\theta^1;\theta^2) = A(\theta^2) - \left(A(\theta^1) + \langle \mu^1, \theta^2 - \theta^1\rangle\right).$$

Using conjugate duality,

$$\begin{aligned} A^*(\mu^1) &= \sup_{\theta\in\Omega} \left(\langle \mu^1, \theta\rangle - A(\theta)\right) \\ &= \langle \mu^1, \theta^1\rangle - A(\theta^1), \end{aligned}$$

we have

$$D(\theta^1;\theta^2) = A(\theta^2) - \left(\langle \mu^1, \theta^2\rangle - A^*(\mu^1)\right).$$

Mean Field and KL-Divergence

$$D(\theta^1;\theta^2) = A(\theta^2) - \left(\langle \mu^1, \theta^2 \rangle - A^*(\mu^1)\right).$$

So choosing $\mu \in \hat{\mathcal{M}}$ to maximize

$$\langle \mu^1, \theta \rangle - A(\theta)$$

corresponds to choosing the distribution μ from the approximating set $\hat{\mathcal{M}}$ to minimize the KL-divergence

$$D(\mu;\theta) = A(\theta) - (\langle \mu, \theta \rangle - A^*(\mu)).$$

That is, the mean field algorithm aims for the *best approximation* (in terms of KL-divergence) in $\hat{\mathcal{M}}$.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Gaussian Mean Field

- Another mean field example: Gaussian MRF.
- Mean parameters:

$$\mu = \mathbb{E}X \in \mathbb{R}^d,$$

$$\Sigma = \mathbb{E}XX' \in \mathcal{S}_+^d.$$

- Approximate with disconnected graph (empty edge set):

$$\begin{aligned}\hat{\mathcal{M}} = \big\{(\mu, \Sigma) : \Sigma - \mu\mu' &= \text{diag}(\Sigma - \mu\mu') \\ \Sigma - \mu\mu' &\geq 0\big\}\,.\end{aligned}$$

Gaussian Mean Field

- Entropy for a Gaussian is

$$\frac{1}{2}\ln\left((2\pi e)^d \left|\Sigma - \mu\mu'\right|\right).$$

Since covariance matrix is diagonal, we have

$$A^*(\mu,\Sigma) = -\frac{d}{2}\ln(2\pi e) - \frac{1}{2}\sum_{i=1}^{d}\ln\left(\Sigma_{ii} - \mu_i^2\right).$$

- Optimization problem becomes

$$\max_{(\mu,\Sigma)\in\hat{\mathcal{M}}}\left(\langle\theta,\mu\rangle + \langle\Theta,\Sigma\rangle + \frac{1}{2}\sum_{i=1}^{d}\ln\left(\Sigma_{ii} - \mu_i^2\right)\right).$$

Gaussian Mean Field

- Calculus shows that fixed point satisfies, for all $i \in V$,

$$\Theta_{ii} = -\frac{1}{2(\mu_{ii} - \mu_i^2)},$$
$$\frac{\mu_i}{2(\mu_{ii} - \mu_i^2)} = \theta_i + \sum_{j \in N(i)} \theta_{ij}\mu_j.$$

- Iteration

$$\mu_i := -\frac{1}{\Theta_{ii}}\left(\theta_i + \sum_{j \in N(i)} \Theta_{ij}\mu_j\right)$$

solves these fixed point equations (provided $-\Theta$ is diagonally dominant):
corresponds to Gauss-Seidel iteration.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Loopy Belief Propagation

Consider a pairwise MRF:

- Graph $G = (V, E)$.
- $X_v \in \mathcal{X} := \{0, \ldots, r-1\}$ for $v \in V$.
- Sufficient statistics are indicators for singleton and pairwise marginals (nodes and edges):

$$\mathbf{1}[x_v = i] \qquad v \in V,\ i \in \mathcal{X}$$

$$\mathbf{1}[x_u = i, x_v = j] \qquad \{u, v\} \in E,\ i, j \in \mathcal{X}$$

Loopy Belief Propagation

- Exponential representation:

$$
\begin{aligned}
p(x) &= \exp\left(\sum_{v \in V} \sum_i \theta_{v,i} \mathbf{1}[x_v = i] \right. \\
&\qquad \left. + \sum_{\{u,v\} \in E} \sum_{i,j} \theta_{u,i;v,j} \mathbf{1}[x_u = i] \mathbf{1}[x_v = j] \right) \\
&= \exp\left(\sum_{v \in V} \theta_v(x_v) + \sum_{\{u,v\} \in E} \theta_{u,v}(x_u, x_v) \right),
\end{aligned}
$$

where $\theta_v(x_v) = \sum_{i \in \mathcal{X}} \theta_{v,i} \mathbf{1}[x_v = i]$,

$$
\theta_{u,v}(x_u, x_v) = \sum_{i,j \in \mathcal{X}} \theta_{u,i;v,j} \mathbf{1}[x_u = i] \mathbf{1}[x_v = j].
$$

Loopy Belief Propagation

An alternative protocol for *belief propagation in trees*:

1. $m_{v,u}^{(0)}(x_u) = 1$ for all $\{u, v\} \in \mathcal{E}$.
2. At iteration $t = 1, 2, \ldots$,

$$m_{v,u}^{(t)}(x_u) = \sum_{x_v} \exp\left(\theta_v(x_v) + \theta_{u,v}(x_u, x_v)\right) \prod_{w \in N(v) \setminus \{u\}} m_{w,v}^{(t-1)}(x_v)$$

- This protocol makes sense for arbitrary graphs: pretend that the graph is a tree.
- If there are a few long cycles, we might expect this to work well.

Variational Interpretation

If we

1. Approximate the marginal polytope $\mathcal{M}$ with a tree-based outer bound $\hat{\mathcal{M}}$,
2. Approximate the entropy $-A^*(\mu)$ with something tractable (the *Bethe* approximation),
3. Iteratively update variables to find stationary points of the Lagrangian,

then we arrive at loopy belief propagation.

Mean Parameters

$$\mu_v(x_v) := \sum_{i \in \mathcal{X}} \mu_{v;i} \mathbf{1}[x_v = i],$$

$$\mu_{u,v}(x_u, x_v) := \sum_{i,j \in \mathcal{X}} \mu_{u,i;v,j} \mathbf{1}[x_u = i] \mathbf{1}[x_v = j].$$

$$\mathcal{M} = \Bigg\{ \mu : \mu_v(x_v) = \sum_{x_u, u \neq v} p(x),$$

$$\mu_{u,v}(x_u, x_v) = \sum_{x_w, w \neq u,v} p(x) \Bigg\}.$$

Tree-Based Outer Bound on $\mathcal{M}$

$$\hat{\mathcal{M}} = \left\{ \tau : \tau \geq 0, \sum_{x_v} \tau_v(x_v) = 1, \sum_{x_u} \tau_{u,v}(x_u, x_v) = \tau_v(x_v) \right\}.$$

- For any G,
  $\mathcal{M} \subseteq \hat{\mathcal{M}}$.
- If G is a tree, there is a junction tree, so local consistency implies global consistency:
  $\hat{\mathcal{M}} = \mathcal{M}$.

Variational Interpretation

1. Approximate the marginal polytope $\mathcal{M}$ with a tree-based outer bound $\hat{\mathcal{M}}$,
2. Approximate the entropy $-A^*(\mu)$ with something tractable (the *Bethe* approximation),
3. Iteratively update variables to find stationary points of the Lagrangian.

Bethe Entropy Approximation

$$H_{\text{Bethe}}(\mu) = \sum_{v \in V} H_v(\mu_v) - \sum_{\{u,v\} \in E} I_{u,v}(\mu_{u,v}),$$

where H_v is the single node entropy,

$$H_v(\mu_v) = -\sum_{x_v} \mu_v(x_v) \log \mu_v(x_v),$$

and $I_{u,v}$ is the mutual information between X_u and X_v,

$$\begin{aligned} I_{u,v}(\mu_{u,v}) &= D(\mu_{u,v}; \mu_u \mu_v) \\ &= -\sum_{x_u, x_v} \mu_{u,v}(x_u, x_v) \log \frac{\mu_{u,v}(x_u, x_v)}{\mu_u(x_u)\mu_v(x_v)}. \end{aligned}$$

Bethe Entropy Approximation

Recall that, if an undirected graph G has a junction tree, then the joint distribution can be expressed as

$$p(x) = \frac{\prod_{c \in C} p(x_C)}{\prod_{s \in S} p(x_s)},$$

where C is the set of cliques and S the set of separators. This implies that if G is a tree, we can write

$$p(x) = \prod_{v \in V} \mu_v(x_v) \prod_{\{u,v\} \in E} \frac{\mu_{u,v}(x_u, x_v)}{\mu_u(x_u)\mu_v(x_v)}.$$

Bethe Entropy Approximation

If G is a tree,

$$p(x) = \prod_{v \in V} \mu_v(x_v) \prod_{\{u,v\} \in E} \frac{\mu_{u,v}(x_u, x_v)}{\mu_u(x_u)\mu_v(x_v)}.$$

So for a tree, we can write the entropy as

$$\begin{aligned} H(\mu) &= -\sum_x p(x) \log p(x) \\ &= \sum_{v \in V} H_v(\mu_v) - \sum_{\{u,v\} \in E} I_{u,v}(\mu_{u,v}) \\ &= H_{\text{Bethe}}(\mu). \end{aligned}$$

Bethe Variational Problem

1. Approximate the marginal polytope $\mathcal{M}$ with a tree-based outer bound $\hat{\mathcal{M}}$,
2. Approximate the entropy $-A^*(\mu)$ with something tractable (the *Bethe* approximation).

$$\max_{\tau \in \hat{\mathcal{M}}} \left(\langle \theta, \tau \rangle + \sum_{v \in V} H_v(\mu_v) - \sum_{\{u,v\} \in E} I_{u,v}(\tau_{u,v}) \right).$$

Variational Interpretation

1. Approximate the marginal polytope $\mathcal{M}$ with a tree-based outer bound $\hat{\mathcal{M}}$,
2. Approximate the entropy $-A^*(\mu)$ with something tractable (the *Bethe* approximation),
3. Iteratively update variables to find stationary points of the Lagrangian.

Lagrangian Formulation

Marginalization constraints:

$$C_{u,v}(x_v) := \tau_v(x_v) - \sum_{x_u} \tau_{u,v}(x_u, x_v).$$

Lagrangian:

$$\mathcal{L}(\tau;\lambda) = \langle \theta, \tau \rangle + \sum_{v \in V} H_v(\mu_v) - \sum_{\{u,v\} \in E} I_{u,v}(\tau_{u,v})$$

$$+ \sum_{\{u,v\} \in E} \left(\sum_{x_v} \lambda_{u,v}(x_v) C_{u,v}(x_v) + \sum_{x_u} \lambda_{v,u}(x_u) C_{v,u}(x_u) \right)$$

Lagrangian Formulation

Taking partial derivatives w.r.t. τ_v and $\tau_{u,v}$, and setting to 0 gives

$$\tau_v(x_v) \propto \exp(\theta_v(x_v)) \prod_{u \in N(v)} \exp(\lambda_{u,v}(x_v))$$

$$\tau_{u,v}(x_u, x_v) \propto \exp\left(\theta_u(x_u) + \theta_v(x_v) + \theta_{u,v}(x_u, x_v)\right)$$
$$\times \prod_{w \in N(u) \setminus \{v\}} \exp(\lambda_{w,u}(x_u)) \prod_{z \in N(v) \setminus \{u\}} \exp(\lambda_{z,v}(x_v))$$

Consider the *messages* $m_{v,u}(x_u) = \exp(\lambda_{v,u}(x_u))$, set $C_{v,u}(x_u) = 0$, and solve to obtain the loopy belief propagation update rule.

Lagrangian Formulation

Messages $m_{v,u}(x_u) = \exp(\lambda_{v,u}(x_u))$ are updated via

$$m_{v,u}(x_u) := \sum_{x_v} \exp\left(\theta_v(x_v) + \theta_{u,v}(x_u, x_v)\right) \prod_{w \in N(v) \setminus \{u\}} m_{w,v}(x_v).$$

This is loopy belief propagation.

Key ideas of this lecture

- Variational approach: Inference as optimization.
- Mean field algorithm.
  - Approximate $\mathcal{M}$ with smaller set $\hat{\mathcal{M}}$.
  - Coordinate ascent is mean field algorithm.
  - $\hat{\mathcal{M}}$ is not convex.
  - Equivalent to finding closest (KL) μ in $\hat{\mathcal{M}}$.
  - Example: Gaussian mean field.
- Loopy belief propagation.
  - Approximate $\mathcal{M}$ with larger tree-based $\hat{\mathcal{M}}$.
  - Approximate $H(\mu)$ with $H_{\text{Bethe}}(\mu)$.
  - Updates to find stationary points of Lagrangian: Loopy belief propagation.

Announcements

- Poster sessions will be on Tue Dec 1 (Stat241A) and Thu Dec 3 (CS281A), here, 11-12:30. Please attend both sessions.
- Project reports are due at 5pm on Friday December 4. In the box outside 723 SD Hall. This deadline is firm.