CS839: Probabilistic Graphical Models

Lecture 13: Variational Inference and Mean Field (continued)

Theo Rekatsinas
Variational Principle

• The dual function takes the form

\[ A^*(\mu) = \begin{cases} 
-H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \notin \mathcal{M}.
\end{cases} \]

• The log partition function has the variational form

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu - A^*(\mu) \right\} \]

• For all \( \theta \) the above optimization problem is attained uniquely at \( \mu(\theta) \) that satisfies

\[ \mu(\theta) = \mathbb{E}_\theta[\phi(X)] \]
Example: Two-node Ising Model

• The distribution

\[ p(x; \theta) \propto \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_{12}\} \]

\[ \phi(x) = \{x_1, x_2, x_{12}\} \]

• Sufficient statistics

• The marginal polytope is characterized by

\[
\begin{align*}
\mu_1 & \geq \mu_{12} \\
\mu_2 & \geq \mu_{12} \\
\mu_{12} & \geq 0 \\
1 + \mu_{12} & \geq \mu_1 + \mu_2
\end{align*}
\]

• The dual has an explicit form

\[
A^*(\mu) = \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log(\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log(\mu_2 - \mu_{12}) + (1 + \mu_{12} - \mu_1 - \mu_2) \log(1 + \mu_{12} - \mu_1 - \mu_2)
\]

• The variational problem is

\[
A(\theta) = \max_{\{\mu_1, \mu_2, \mu_{12}\} \in \mathcal{M}} \{\theta_1 \mu_1 + \theta_2 \mu_2 + \theta_{12} \mu_{12} - A^*(\mu)\}
\]

• The optimum is attained at

\[
\mu_1(\theta) = \frac{\exp\{\theta_1\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}{1 + \exp\{\theta_1\} + \exp\{\theta_2\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}
\]
Variational Principle

• Exact variational formulation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \theta^T \mu - A^*(\mu) \} \]

- \( \mathcal{M} \): the marginal polytope, difficult to characterize
- \( A^* \): the negative entropy function, no explicit form

• Mean field method: non-convex inner bound and exact form of entropy

• Bethe approximation and Loopy BP: polyhedral outer bound and non-convex Bethe approximation
Belief Propagation Algorithm

• Message passing rule:

\[ M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in N(t)/s} M_{ut}(x'_t) \right\} \]

• Marginals

\[ \mu_s(x_s) = \kappa \psi_s(x_s) \prod_{t \in N(s)} M^*_t(x_s) \]

• Exact for trees but approximate for loopy graphs
• How does this relate to the variational principle? For trees/generic graphs?
Tree Graphical Models

• Discrete variables $X_s \in \{0, 1, \ldots, m_s - 1\}$ on a tree $T = (V,E)$

\[ \mathbb{I}_j(x_s) \quad \text{for} \ s = 1, \ldots n, \ j \in \mathcal{X}_s \]
\[ \mathbb{I}_{jk}(x_s, x_t) \quad \text{for} \ (s, t) \in E, \ (j, k) \in \mathcal{X}_s \times \mathcal{X}_t \]

• Sufficient statistics

• Exponential representation of distribution?

• Mean parameters are marginal probabilities:

\[ \mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in \mathcal{X}_s, \quad \mu_s(x_s) = \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s) = \mathbb{P}(X_s = x_s) \]
\[ \mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j, k) \in \mathcal{X}_s \times \mathcal{X}_t. \]
\[ \mu_{st}(x_s, x_t) = \sum_{(j, k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t) = \mathbb{P}(X_s = x_s, X_t = x_t) \]
Marginal Polytope for Trees

• Marginal polytope for general graphs

\[ \mathcal{M}(G) = \{ \mu \in \mathbb{R}^d \mid \exists p \text{ with marginals } \mu_{s;j}, \mu_{st;jk} \} \]

• By junction tree we have:

\[ \mathcal{M}(T) = \left\{ \mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \right\} \]

• If \( \mu \in \mathcal{M}(T) \) then

\[ p_\mu(x) := \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)}. \]
Decomposition of Entropy for Trees

• For trees the entropy decomposes as (this is also our dual!):

\[
H(p(x; \mu)) = - \sum_x p(x; \mu) \log p(x; \mu)
\]

\[
= \sum_{s \in V} \left( - \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s) \right) - H_s(\mu_s)
\]

\[
- \sum_{(s,t) \in E} \left( \sum_{x_s,x_t} \mu_{st}(x_s,x_t) \log \frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)} \right) I_{st}(\mu_{st}), \text{KL-Divergence}
\]

\[
= \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})
\]

\[
A^*(\mu) = -H(p(x; \mu))
\]
Exact Variational Principle for Trees

• Variational formulation

\[
A(\theta) = \max_{\mu \in \mathcal{M}(T)} \left\{ \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) \right\}
\]

• Assign a Lagrange multiplier for the normalization constraint and each marginalization constraint

\[
\mathcal{L}(\mu, \lambda) = \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) + \sum_{s \in V} \lambda_{ss} C_{ss}(\mu) \\
+ \sum_{(s,t) \in E} \left[ \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) \right]
\]
Lagrangian Derivation

• Taking the derivatives of the Lagrangian wrt to $\mu_s \mu_{st}$

\[
\frac{\partial L}{\partial \mu_s(x_s)} = \theta_s(x_s) - \log \mu_s(x_s) + \sum_{t \in \mathcal{N}(s)} \lambda_{ts}(x_s) + C
\]

\[
\frac{\partial L}{\partial \mu_{st}(x_s, x_t)} = \theta_{st}(x_s, x_t) - \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'
\]

• Setting them to zeros yields

\[
\mu_s(x_s) \propto \exp\{\theta_s(x_s)\} \prod_{t \in \mathcal{N}(s)} \exp\{\lambda_{ts}(x_s)\} \exp\{M_{ts}(x_s)\}
\]

\[
\mu_s(x_s, x_t) \propto \exp \left\{ \theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t) \right\} \times 
\prod_{u \in \mathcal{N}(s) \setminus t} \exp \left\{ \lambda_{us}(x_s) \right\} \prod_{v \in \mathcal{N}(t) \setminus s} \exp \left\{ \lambda_{vt}(x_t) \right\}
\]
BP on Arbitrary Graphs

• Two main difficulties of the variation formulation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \theta^T \mu - A^*(\mu) \} \]

• The marginal polytope is hard to characterize, so let’s use the tree-based outer bound

\[ \mathcal{L}(G) = \left\{ \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \right\} \]

• Exact entropy lacks explicit form, so let’s approximate it using the exact expression for trees

\[ -A^*(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}). \]
Bethe Variational Problem

• Combining the two gives us the Bethe variational problem

\[
\max_{\tau \in \mathbb{L}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}.
\]

• What is happening?
• Tree-based outer bound

\[ \mathcal{M}(G) \subseteq \mathbb{L}(G) \]
Mean Field Approximation
Tractable Subgraphs

• For an exponential family with sufficient statistics $\phi$ defined on graph $G$ the set of realizable mean parameter set is

$$\mathcal{M}(G; \phi) := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

• Idea: restrict $p$ to a subset of distributions associated with a tractable subgraph

$$\Omega := \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$
Mean Field Methods

• For a given tractable subgraph $F$, a subset of canonical parameters is

$$\mathcal{M}(F; \phi) := \{\tau \in \mathbb{R}^d \mid \tau = \mathbb{E}_\theta[\phi(X)] \text{ for some } \theta \in \Omega(F)\}$$

• Inner approximation

$$\mathcal{M}(F; \phi)^o \subseteq \mathcal{M}(G; \phi)^o$$

• Mean field solves the relaxed problem

$$\max_{\tau \in \mathcal{M}_F(G)} \{\langle \tau, \theta \rangle - A^*_F(\tau)\}$$

• $A^*_F = A^*|_{\mathcal{M}_F(G)}$ is the exact dual function restricted to $\mathcal{M}_F(G)$
Example: Naïve Mean Field for Ising Model

- Ising model in \{0,1\} representation

\[
p(x) \propto \exp \left\{ \sum_{s \in V} x_s \theta_s + \sum_{(s,t) \in E} x_s x_t \theta_{st} \right\}
\]

- Mean parameters

\[
\mu_s = E_p[X_s] = P[X_s = 1] \quad \mu_{st} = E_p[X_s X_t] = P[(X_s, X_t) = (1, 1)]
\]

- For fully disconnected graph \( F \)

\[
\mathcal{M}_F(G) := \{ \tau \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \tau_s \leq 1, \forall s \in V, \tau_{st} = \tau_s \tau_t, \forall (s, t) \in E \}
\]

- The dual decomposes into sum, one for each node

\[
A^*_F(\tau) = \sum_{s \in V} [\tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s)]
\]
Example: Naïve Mean Field for Ising Model

• Mean field problem

\[ A(\theta) \geq \max_{(\tau_1, \ldots, \tau_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \tau_s + \sum_{(s,t) \in E} \theta_{st} \tau_s \tau_t - A_F^*(\tau) \right\} \]

• The same objective function as in free energy based approach

• The naïve mean field update equations

\[ \tau_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \tau_t \right) \]

• Lower bound on log partition function
Geometry of Mean Field

• Mean field optimization is always non-convex for any exponential family in which the state space is finite

\[ \mathcal{M}(G) = \text{conv}\{\phi(e); e \in \mathcal{X}^m\} \]

• Marginal polytope is a convex hull

• \( \mathcal{M}_F(G) \) contains all the extreme points (if it is a strict subset then it must be non-convex)

• Example: two-node Ising

\[ \mathcal{M}_F(G) = \{0 \leq \tau_1 \leq 1, 0 \leq \tau_2 \leq 1, \tau_{12} = \tau_1 \tau_2\} \]

• Parabolic cross section along \( \tau_1 = \tau_2 \)
Variational (Bayesian) Inference

• Variational Bayes is most often used to infer the conditional distribution over the latent variables given the observations (and parameters).

• This is also known as the posterior distribution over the latent variables

\[ p(z|x, \alpha) = \frac{p(z, x|\alpha)}{\int_z p(z, x|\alpha)} \]
Motivation

• Why variational Bayes?

• We cannot directly compute the posterior distribution for many interested models.
  • Posterior density is in an intractable form
Variational Bayes

• The main idea behind variational Bayes:

  • Choose a family of distributions over the latent variables $z$ with its own set of variational parameters $\nu$ $q(z_1:m|\nu)$

  • Find parameters $\nu$ that make approximation $q$ to be closest to the posterior distribution
    • Optimization!

  • Then use $q$ with the fitted parameters in place of the posterior
    • E.g., for predictions or to find models etc.
Kullback-Leibler Divergence

• We measure the closeness of the two distributions with KL divergence

\[
KL(q||p) = \int q(z) \log \frac{q(z)}{p(z|x)} = \mathbb{E}_q \left[ \log \frac{q(z)}{p(z|x)} \right]
\]

• Important cases:
  • q high and p is high (low KL)
  • q high and p is low (high KL)
  • q is low then we don’t care

• Why not KL(p || q)?
The Evidence Lower Bound

• To do variational Bayes, we want to minimize KL between our approximation q and our posterior p

• However, we cannot actually minimize this quantity (quantity of interest). Let’s minimize a function that is equal to it up to a constant.

• Evidence lower bound (ELBO)

• What is evidence? (Hint: logZ)
The Evidence Lower Bound

- Jensen’s inequality: if $f$ is concave $f(E[X]) \geq E[f(x)]$
- Take the log (marginal) probability of the observations and get the ELBO by using Jensen’s

$$
\log p(x) = \log \int_z p(x, z) = \log \int_z p(x, z) \frac{q(z)}{q(z)} = \log \left( \mathbb{E}_q \left[ \frac{p(x, z)}{q(z)} \right] \right) \geq \mathbb{E}_q \left[ \log p(x, z) \right] - \mathbb{E}_q \left[ \log q(z) \right]
$$
The Evidence Lower Bound

• The evidence lower bound for a probability model $p(x,z)$ and approximation $q(z)$ to the posterior is:

$$E_q [\log p(x, z)] - E_q [\log q(z)]$$

• This quantity is less than or equal to the evidence (log marginal probability of the observations)

• We optimize this quantity (over densities $(q(z))$ in Variational Bayes to find an “optimal approximation”
Mean Field and ELBO

• We choose a family of variational distributions (i.e. a family of approximations) such that these two expectations can be computed.

• The second expectation is the entropy

• In variational inference, we find settings of the variational parameters $v$ that maximize the ELBO, which is equivalent to minimizing the KL divergence.

\[
KL(q||p) = \mathbb{E}_q \left[ \log \frac{q(z)}{p(z|x)} \right] = \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z|x)] = \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z,x)] + \log p(x) = - (\mathbb{E}_q [\log p(z,x)] - \mathbb{E}_q [\log q(z)]) + \log p(x)
\]
Optimizing the ELBO in Mean Field Variational Inference

• How do we optimize the ELBO in mean field variational inference?

• Typically, use coordinate ascent

• We optimize each latent variable’s variational approximation $q$ in turn while holding the others fixed.
  • At each iteration we get an updated local variational approximation
  • We iterate through each latent variable until convergence
Optimizing the ELBO in Mean Field Variational Inference

• Recall that:

\[ p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j | z_{1:(j-1)}, x_{1:n}) \]

• Latent variables can occur in any order

• We can decompose the entropy term of ELBO as

\[ \mathbb{E}_q [\log q(z_{1:m})] = \sum_{j=1}^{m} \mathbb{E}_{q_j} [\log q(z_j)] \]
Optimizing the ELBO in Mean Field Variational Inference

- We can now decompose the ELBO loss in a nice form

- ELBO is defined as: $\mathbb{E}_q [\log p(x, z)] - \mathbb{E}_q [\log q(z)]$

- We have that

$$L = \log p(x_{1:n}) + \sum_{j=1}^{m} (\mathbb{E}_q [\log p(z_j | z_{1:(j-1)}, x_{1:n})] - \mathbb{E}_{q_j} [\log q(z_j)])$$
Some terminology

• The conditional for a latent variable \(z_j\) is

\[
p(z_j|z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_m, x) = p(z_j|z_{-j}, x)
\]

• Here \(-j\) denotes all indices other than the j-th

• This is the posterior conditional of \(z_j\) given all other latent variables and observations

• This posterior conditional is very important in mean field variational Bayes, and will be important in sampling algorithms (Gibbs sampling)
Optimizing the ELBO in Mean Field Variational Inference

• We can now decompose the ELBO loss in a nice form

• ELBO is defined as: \( \mathbb{E}_q [\log p(x, z)] - \mathbb{E}_q [\log q(z)] \)

• We have that

\[
\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m} (\mathbb{E}_q [\log p(z_j | z_{1:(j-1)}; x_{1:n})] - \mathbb{E}_{q_j} [\log q(z_j)])
\]

• And we have for each \( q_j \)

\[
\arg\max_{q_j} \mathcal{L} = \arg\max_{q_j} \left( \mathbb{E}_q [\log p(z_j | z_{-j}, x)] - \mathbb{E}_{q_j} [\log q(z_j)] \right)
\]

\[
= \arg\max_{q_j} \left( \int q(z_j) \mathbb{E}_{q_{-j}} [\log p(z_j | z_{-j}, x)] dz_j - \int q(z_j) \log q(z_j) dz_j \right)
\]
Optimizing the ELBO in Mean Field Variational Inference

• To find the argmax we take the derivative with respect to \( q_j \) and use Lagrange multipliers and set the derivative to zero

\[
\frac{d\mathcal{L}_j}{dq(z_j)} = \mathbb{E}_{q_{-j}} [\log p(z_j|z_{-j}, x)] - \log q(z_j) - 1 = 0
\]

• We have that:

\[q^*(z_j) \propto \exp \left\{ \mathbb{E}_{q_{-j}} [\log p(z_j|z_{-j}, x)] \right\}\]

• And since the denominator of the conditional does not depend on \( z_j \) we have:

\[q^*(z_j) \propto \exp \left\{ \mathbb{E}_{q_{-j}} [\log p(z_j, z_{-j}, x)] \right\}\]
Notes

• This coordinate ascent procedure converges to a local maximum
• The coordinate update for \( q(z_j) \) only depends on the other fixed approximations \( q(z_k) \)
• While this determines the optimal \( q(z_j) \) we have not specified the form of \( q \) we aim to use, we only talked about factorization.
• Depending on what form we use, the coordinate update \( q^*(z_j) \) might not be easy to work with.
Mean field approx. to the Gibbs free energy

• Given a disjoint clustering $C_1, C_2, ..., C_n$ of all variables

• Define $q$ as

$$q(X) = \prod_i q_i(X_{C_i})$$

• Mean-field free energy

$$G_{MF} = \sum_i \sum_{x_{C_i}} \prod_i q_i(x_{C_i})E(x_{C_i}) + \sum_i \sum_{x_{C_i}} q_i(x_{C_i}) \ln q_i(x_{C_i})$$

  e.g., $G_{MF} = \sum_{i<j} q(x_i)q(x_j)\phi(x_ix_j) + \sum_i \sum_{x_i} q(x_i)\phi(x_i) + \sum_i \sum_{x_i} q(x_i)\ln q(x_i)$  
  (naïve mean field)

• Will never be equal to the exact Gibbs free energy. It is always a lower bound
Generalized Mean Field

**Theorem**: The optimum GMF approximation to the cluster marginal is isomorphic to the cluster posterior of the original distribution given internal evidence and its generalized mean fields:

\[ q_i^* (X_{H,C_i}) = p(X_{H,C_i} | X_{E,C_i}, \langle X_{H,MB_i} \rangle_{q_{j \neq i}}) \]

**GMF algorithm**: Iterate over each \( q_i \)
Automatic Variational Inference

- For each new model derive the variational update equations and write application-specific code to find the solution.
- Can we have a general-purpose inference engine which autoamates this procedure? Black-box variational inference Ranganath et al.
Summary

• Variation methods in general turn inference into an optimization problem via exponential families and convex duality

• The exact variational principle is intractable to solve; Two approximations:
  • Either inner or outer bound to the marginal polytope
  • Various approximations to the entropy function

• Mean-field: non-convex inner bound and exact form of entropy
• BP: polyhedral outer bound and non-convex Bethe approximation