Probabilistic Graphical Models

Lecture 15 – Inference as Optimization

CS/CNS/EE 155
Andreas Krause
Announcements

- Homework 3 due next Monday (Nov 23)
- Project poster session on Friday December 4 (tentative)
- Final writeup (8 pages NIPS format) due Dec 9
Approximate inference

- Three major classes of general-purpose approaches

- Message passing
  - E.g.: Loopy Belief Propagation

- Inference as optimization
  - Approximate posterior distribution by simple distribution
  - Mean field / structured mean field

- Sampling based inference
  - Importance sampling, particle filtering
  - Gibbs sampling, MCMC

- Many other alternatives (often for special cases)
What if we apply BP to a graph with loops?
- Apply BP and hope for the best..

$$\delta_{i \rightarrow j}(X_j) = \sum_{x_i} \pi_i(x_i) \pi_{i,j}(x_i, X_j) \prod_{s \in N(i) \setminus \{j\}} \delta_{s \rightarrow i}(x_i)$$

- Will not generally converge.. 😞
- If it converges, will not necessarily get correct marginals 😞

However, in practice, answers often still useful!
Approximate inference

- Three major classes of general-purpose approaches

- **Message passing**
  - E.g.: Loopy Belief Propagation (today!)

- **Inference as optimization**
  - Approximate posterior distribution by simple distribution
  - Mean field / structured mean field
  - Assumed density filtering / expectation propagation

- **Sampling based inference**
  - Importance sampling, particle filtering
  - Gibbs sampling, MCMC

- Many other alternatives (often for special cases)
Variational approximation

- Graphical model with intractable (high-treewidth) joint distribution $P(X_1,\ldots,X_n)$
- Want to compute posterior distributions

$$P(X_3, X_4 | X_1 = x_1, X_2 = x_2)$$

$$P(X_3, X_4 | X_1 = x_1, X_2 = x_2) \propto \sum \sum \cdots \sum P(x_1, x_2, \ldots, x_n)$$

- Computing posterior exactly is intractable
- **Key idea**: Approximate posterior with *simpler* distribution that’s *as close* to $P$ as possible
Why should we hope that we can find a simple approximation?

- Prior distribution is complicated
  - Need to describe all possible states of the world (and relationships between variables)

- Posterior distribution is often simple:
  - Have made many observations \( \Rightarrow \) less uncertainty
  - Variables can become “almost independent”

For now: Represent posterior as undirected model (and instantiate observations)

\[
P(X_1, \ldots, X_n \mid obs) = \frac{1}{Z} \prod_j \Psi_j(C_j)
\]
Variational approximation

**Key idea:** Approximate posterior with simpler distribution that’s as close as possible to $P$
- What is a “simple” distribution?
- What does “as close as possible” mean?

**Simple** = efficient inference
- Typically: factorized (fully independent, chain, tree, ...)
- Gaussian approximation

**As close as possible** = KL divergence (typically)
- Other distance measures can be used too, but more challenging to compute
Kullback-Leibler (KL) divergence

- Distance between distributions

\[ D(P \| Q) = \int P(x) \log \frac{P(x)}{Q(x)} \, dx \]

- Properties:
  - \( D(P \| Q) \geq 0 \)
  - \( P(x) = Q(x) \) almost everywhere \( \iff \) \( D(P \| Q) = 0 \)
  - In general, \( D(P \| Q) \neq D(Q \| P) \)
    - \( P \) determines when difference is important
      \[ P(x) = 0 \; \land \; Q(x) \neq 0 \; \implies \; 0 \cdot \log \frac{Q}{0} = 0 \]
      \[ P(x) = \varepsilon \; \land \; Q(x) > 0 \; \implies \; \varepsilon \cdot \log \frac{\varepsilon}{0} = \infty \]
Finding simple approximate distributions

- KL divergence not symmetric; need to choose directions
- P: true distribution; Q: our approximation

\[ D(P \mid \mid Q) \]
- The “right” way
- Q chosen to “support” P
- Often intractable to compute

\[ D(Q \mid \mid P) \]
- The “reverse” way
- Underestimates support (overconfident)
- Will be tractable to compute

Both special cases of \( \alpha \)-divergence
“Simple” distributions

Simplest distribution: $Q$ fully factorized
- $Q(X_1,\ldots,X_n) = \prod_i Q_i(X_i)$

$M = \{Q: Q$ fully factorized\}$
- $= \{Q: Q(X) = \prod_i Q_i(X_i)\}$

$$Q^* = \arg\min_{Q\in M} D(Q\|P)$$

Can also find more structured approximations
- Chains: $Q(X_1,\ldots,X_n) = \prod_i Q_i(X_i \mid X_{i-1})$
- Trees
- Any distributions one can do efficient inference on
Mean field approximation the “right way”

\[
D(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}
\]

\[
= \sum_x P(x) \log P(x) - \sum_x P(x) \log Q(x)
\]

\[
= \sum_x P(x) \log \prod_i Q_i(x_i)
\]

\[
= \sum_x P(x) \sum_i \log Q_i(x_i)
\]

\[
= \sum_{i} \sum_{x_i} P(x_i) \log Q_i(x_i)
\]

\[
= \sum_{i} \left( \sum_{x_i} P(x_i) \log Q_i(x_i) \right) \left( \sum_{x_i} P(x_i) \right)
\]

Need \( P(x_i) \) Intractable!
Mean field approximation the reverse way

\[ D(Q \parallel P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)} \]

\[ = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(x) \]

(i) \[ H(Q) = \sum_x Q(x) \log \frac{1}{\hat{Q}} \cdot \hat{Q}(x) \]

\[ = \sum_x \sum_k Q(x) \log \hat{Q}(x) \]

\[ = \sum_x \sum_k Q(x) \hat{Q}(x) \log \hat{Q}(x) \left( \frac{\sum \hat{Q}(x)}{\sum_k \hat{Q}(x)} \right) \]

\[ = - \sum_k H(Q_k) \]

(ii) \[ = \sum_x Q(x) \log \frac{1}{\hat{Q}} \cdot \hat{Q}(x) \]

\[ = - \sum_x Q(x) \log \hat{Q} + \sum_k Q(x) \sum \log \hat{Q}(x) \]

\[ = - \log \hat{Q} + \log \hat{Q}(C) \]

(iii) \[ \sum_k Q(x) \log \hat{Q}(x) \]

\[ \text{Need marginal for } C_i \]
Reverse KL for fully factorized case

\[
D(Q||P) = -\sum_i \sum_x Q(x) \log \Psi_i(x) - \sum_i H(Q_i) + \ln Z
\]

\[
\ln Z = D(Q||P) + \sum_i H(Q_i) + \sum_i \mathbb{E}_Q[\log \Psi_i]
\]

\[
\Rightarrow \min_{Q} F(Q; P)
\]
Suppose \( P(X_1,\ldots,X_n) = \frac{1}{Z} \prod_i \Psi_i(C_i) \) is Markov Network

**Theorem:** For any \( Q \) (not necessarily fully factorized)

\[
\ln Z = \mathbb{F}[P;Q] + \mathbb{D}(Q\|P)
\]

Hereby, \( F[P;Q] \) is the following energy functional

\[
F[P;Q] = \sum_i \mathbb{E}_Q[\ln \Psi_i] + H(Q)
\]
Reverse KL vs. log-partition function

\[ \ln Z = \uparrow F[P; Q] + \downarrow D(Q\|P) \geq 0 \]

\[ F[P; Q] = \sum_i \mathbb{E}[\ln \Psi_i] + H(Q) \]

Maximizing energy functional \( \iff \) Minimizing reverse KL

**Corollary:**

Energy function is lower bound on log partition function

\[ P(x) = \frac{1}{Z} \prod_i \Psi_i(x_i) \]

\[ P(x) \leq \frac{1}{F(P; Q)} \prod_i \Psi_i(x_i) \]

Implies upper bound on event probabilities!
Optimizing for mean field approximation

- Want to solve $\max_{Q} F[P; Q] = \max_{Q} \sum_{j} \mathbb{E}_{Q}[\ln \Psi_{j}] + \sum_{i} H(Q_{i})$

  \[ \text{s.t. } \sum_{x_{i}} Q_{i}(x_{i}) = 1 \]

- Solved via Lagrange multipliers: There exist $\lambda_{1}, \ldots, \lambda_{n}$ such that optimization of (\ref{eq:mean-field-approximation}) is equivalent to

  $\max_{Q} \sum_{j} \mathbb{E}_{Q}[\ln \Psi_{j}] + \sum_{j} H(Q_{j}) + \sum_{j} \lambda_{j} \left[ \sum_{x_{j}} Q_{j}(x_{j}) - 1 \right]$

  Differentiate and set to 0!

Theorem: $Q$ stationary point iff for each $i$ and $x_{i}$:

\[ Q_{i}(x_{i}) = \frac{1}{Z_{i}} \exp\left( \sum_{j} \mathbb{E}[\ln \Psi_{j} | x_{i}] \right) \]
Fixed point iteration for MF

- Initialize factors $Q^{(0)}_i$ arbitrarily; $t=0$
- Until converged, do
  - $t \leftarrow t+1$
  - For each variable $i$ and each assignment $x_i$ do

  $$Q_{i}(x_{i})^{(t+1)} = \frac{1}{Z_i} \exp \left( \sum_{j} \mathbb{E}_{Q^{(t)}} [\ln \Psi_{j} | x_{i}] \right)$$

  $$Z_i = \sum_{x_i} Q_i(x_i)$$

- Guaranteed to converge! ☺
- Gives both approx. distribution $Q$ and lower bound on $\ln Z$
- Can get stuck in local optimum 😞
Computing updates

Need to compute

\[ Q_i(x_i)^{(t+1)} = \frac{1}{Z_i} \exp \left( \sum_j \mathbb{E}_{Q^{(t)}}[\ln \Psi_j | x_i] \right) \]

Must compute expected log potentials:

\[ (\Psi) = \mathbb{E}_Q[\ln \Psi_j | x_i] \]

\[
\Psi_j(C_i) = \sum_{x \sim C_i} Q^{(t)}(x | x_i) \ln \Psi_j(C_i) = \sum_{x \sim C_i} Q^{(t)}(x | x_i) \ln \Psi_j(C_i)
\]

\[
= \Psi_j(C_i) = \prod_{k \in C_j \setminus C_i} Q_k(x_k)
\]
Example iteration

\[ \mathbb{E}_Q \left[ \ln \psi_{DG} \mid X_D = 1 \right] \]

\[ = \sum_{X_G} Q^*_G(X_G \mid X_D = 1) \ln \psi_{DG}(X_D = 1, X_G) \]

\[ = Q^*_G(X_G) \]

\[ Q^*_D(X_D = 1) = \frac{1}{Z_D} \exp \left( \mathbb{E}_Q \left[ \ln \psi_{DG} \mid X_D = 1 \right] + \mathbb{E}_Q \left[ \ln \psi_{CD} \mid X_D = 1 \right] \right) \]
Structured mean field

Goal of variational inference:
Approximate complex distribution by simple distribution

True dist.  Fully-factorized mean field  Structured mean field

\[ p(x) \propto \prod_c \phi_c(x_c) \]
\[ q(x) \propto \prod_i q_i(x_i) \]
\[ q(x) \propto q_A(x_A) q_B(x_B) \]
Structured mean-field approximations

- Can get better approximations using **structured approximations**:

\[
\max_{Q \in \mathcal{M}} F[P; Q] = \max_{Q \in \mathcal{M}} \sum_j \mathbb{E}_Q[\ln \Psi_j] + H(Q)
\]

- Only need to be able to compute energy functional
- Can do whenever we can perform efficient inference in Q (e.g., chains, trees, low-treewidth models)
  - Update equations look similar as for fully-factorized case (see reading)
Example: Factorial HMM

- Simultaneous tracking and camera registration
- State space decomposed into object location and camera parameters

Mei and Porikli ‘08
Variational approximations for FHMMs

\[ \max_{Q \in \mathcal{M}} F[P; Q] = \max_{Q \in \mathcal{M}} \sum_j \mathbb{E}_Q[\ln \Psi_j] + H(Q) \]

- Approximate posterior by independent chains

\[ \mathcal{M} = \left\{ Q : Q(X) = \prod_c \prod_t Q_{c,t}(X_{c,t} | X_{c,t-1}) \right\} \]
Summary: Variational inference

- Approximate complex (intractable) distribution by simpler distribution that is “as close as possible”
- **Simple** = tractable (efficient inference)
- **Closeness** = Reverse KL (efficient to compute)
- Interpretation: Optimize **lower bound** on the log-partition function
  - Implies upper bound on event probabilities
- Efficient algorithm that’s guaranteed to converge (in contrast to Loopy BP..), but possibly to local optimum
Approximate inference

- Three major classes of general-purpose approaches

- **Message passing**
  - E.g.: Loopy Belief Propagation (today!)

- **Inference as optimization**
  - Approximate posterior distribution by simple distribution
  - Mean field / structured mean field
  - Assumed density filtering / expectation propagation

- **Sampling based inference**
  - Importance sampling, particle filtering
  - Gibbs sampling, MCMC

- Many other alternatives (often for special cases)
KL-divergence the “right” way:

- Find distribution $Q^* \in \mathcal{M}$:

$$Q^* = \arg\min_{Q \in \mathcal{M}} D(P \| Q)$$

- In some applications, can compute $D(P \| | Q)$
  - Important example: Assumed density filtering in DBNs
Recall: Dynamic Bayesian Networks

- At every timestep have a Bayesian Network

\[ S_t = \{ A_t, B_t, \ldots, E_t \} \]

- Variables at each time step \( t \) called a “slice” \( S_t \)
- “Temporal” edges connecting \( S_{t+1} \) with \( S_t \)
Flow of influence in DBNs

Can we do efficient filtering in BNs?
Approximate inference in DBNs?

Want to find **tractable** approximation to marginals that’s **as close** to true marginals as possible.
Assumed Density Filtering

Assume distribution $P(S_t)$ for slice $t$ factorizes

$P(S_{t+1})$ is fully connected 😞

Want to compute best-approximation $Q^*$ for $P(S_{t+1})$

\[
Q^* = \text{argmin } D(P || Q)
\]
Assumed Density Filtering

\[ \sum_{s_{t+1}} P(s_{t+1}) \log \frac{Q(s_{t+1})}{\tau} = \prod_{i} Q_i(s_{t+1}) \]

\[ = \sum_{s_{t+1}} \sum_{s_{t+1}^{(i)}} P(s_{t+1}^{(i)}) \log Q_i(s_{t+1}^{(i)}) \]

\[ = \sum_{s_{t+1}^{(i)}} P(s_{t+1}^{(i)}) \log Q_i(s_{t+1}^{(i)}) \]

Eq. (1) \( P(A_{t+1}) \)

Can compute expectations efficiently

Get optimal \( Q^* \) by setting \( Q_i^*(s_{t+1}^{(i)}) = P(s_{t+1}^{(i)}) \)
Recall: Bayesian filtering

- Start with $P(X_1)$
- At time $t$
  - Assume we have $P(X_t \mid y_{1\ldots t-1})$
  - Condition: $P(X_t \mid y_{1\ldots t})$

\[
P(X_t \mid y_{1\ldots t}) \propto P(X_t \mid y_{1\ldots t-1}) \frac{P(y_t \mid X_t, Y_{1\ldots t-1})}{P(Y_t \mid X_t)}
\]

- Prediction: $P(X_{t+1}, X_t \mid y_{1\ldots t})$

\[
P(X_{t+1}, X_t \mid y_{1\ldots t}) = P(X_t \mid y_{1\ldots t}) \cdot \frac{P(X_{t+1} \mid X_t, y_{1\ldots t})}{P(X_{t+1} \mid X_t)} = P(X_{t+1} \mid X_t)
\]

- Marginalization: $P(X_{t+1} \mid y_{1\ldots t})$

\[
P(X_{t+1} \mid y_{1\ldots t}) = \sum_{X_t} P(X_{t+1}, X_t \mid y_{1\ldots t})
\]
Assumed Density Filtering

- Start with $P(S_1)$
- At every time step $t$: tractable approximation $Q_t$
  $Q_t(S_t) \approx P(S_t \mid O_{1:t-1})$
- **Condition** on observation $O_t \subseteq S_t$: $Q_t(S_t \mid O_t)$
- **Predict**: multiply transition model to get $Q_t(S_{t+1}, S_t \mid O_t)$
  
  $$Q_t(S_{t+1}, S_t \mid O_t) = Q_t(S_t \mid O_t) \ P(S_{t+1} \mid S_t)$$
- **Marginalize** $S_t$
  - This is intractable (connects all variables in $S_{t+1}$)
  - Approximate $Q_t(S_{t+1} \mid O_t)$ by $Q^*$ s.t.
    
    $$Q^* = \text{argmin}_Q \ D(Q_t(S_{t+1}) \mid \mid Q(S_{t+1}))$$
  - This is done by matching moments:
    for discrete models, ensure that $Q_{t+1}(s_{t+1}) = Q_t(s_{t+1} \mid o_t)$
Summary of Assumed Density Filtering

- Variational inference technique for dynamical Bayesian Networks
- Find tractable approximation for each time slice that minimizes KL divergence (in the “right” way)
- Can show that errors don’t add up too much
- Examples:
  - Tractable inference in DBNs
  - Unscented Kalman Filter
Summary: Inference as optimization

- Approximate intractable distribution by a **tractable** one
- **Optimize parameters** of the distribution to make approximation as tight as possible
- Common distance measure: KL-divergence (both ways)
  - Special case of $\alpha$-divergence
- Can get upper bounds on event probabilities, etc.