Approximate Inference in GMs
Sampling

Le Song

Machine Learning II: Advanced Topics
CSE 8803ML, Spring 2012
What we have learned so far

- Variable elimination and junction tree
  - Exact inference
  - Exponential in tree-width

- Mean field, loopy belief propagation
  - Approximate inference for marginals/conditionals
  - Fast, but can get inaccurate estimates

- Sample-based inference
  - Approximate inference for marginals/conditionals
  - With “enough” samples, will converge to the right answer
Variational Inference

- What is the approximating structure?

- How to measure the goodness of the approximation of $Q(X_1, \ldots, X_n)$ to the original $P(X_1, \ldots, X_n)$?
  - Reverse KL-divergence $KL(Q||P)$

- How to compute the new parameters?
  - Optimization $Q^* = \text{arg\,min}_Q KL(Q||P)$

New Parameters

mean field
Deriving mean field fixed point condition I

Approximate $P(X) \propto \prod_j \Psi(D_j)$ with $Q(X) = \prod_i Q_i(X_i)$

- $s.t. \forall X_i, \ Q_i(X_i) > 0, \int Q_i(X_i)dX_i = 1$
- $D(Q||P) = -\int Q(X) \ln P(X)dX + \int Q(X) \ln Q(X)dX + \text{const.}$
- $= -\int \prod_i Q_i(X_i)(\Sigma_j \ln \Psi(D_j))dX + \int \prod_i Q_i(X_i)(\Sigma_i \ln Q_i(X_i))dX$
- $= -\Sigma_j \int \prod_{i \in D_j} Q_i(X_i) \ln \Psi(D_j)dX_{D_j} + \Sigma_i \int Q_i(X_i) \ln Q_i(X_i)dX_i$

E.g., $\int \prod_i Q_i(X_i) \ln \Psi(X_1, X_5) \ dX = \int Q_1(X_1)Q_5(X_5) \ln \Psi(X_1, X_5)dX_1dX_5$
Let $L$ be the Lagrangian function ($\forall Q_i(X_i) > 0$?)

$$L = -\sum_j \int \prod_{i \in D_j} Q_i(X_i) \ln \Psi(D_j) dX_{D_j} + \sum_i \int Q_i(X_i) \ln Q_i(X_i) dX_i - \sum_i \beta_i (1 - \int Q_i(X_i)) dX_i$$

For KL divergence, the nonnegativity constraint comes for free.

If $Q$ is a maximum for the original constrained problem, then there exists $\lambda_i, \beta_i$ such that $(Q, \lambda_i, \beta_i)$ is a stationary point for the Lagrange function (derivative equal 0).
Deriving mean field fixed point condition III

\[ L = - \sum_j \int \prod_{i \in D_j} Q_i(X_i) \ln \Psi(D_j) dX_{D_j} + \sum_i \int Q_i(X_i) \ln Q_i(X_i) dX_i - \sum_i \beta_i (1 - \int Q_i(X_i)) dX_i \]

\[ \frac{\partial L}{\partial Q_i(X_i)} = -\sum_{j: i \in D_j} (\prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j)) + \ln Q_i(X_i) - \beta_i = 0 \]

\[ Q_i(X_i) = \exp \left( \sum_{j: i \in D_j} (\prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j)) + \beta_i \right) = \frac{1}{Z_i} \exp \left( \sum_{j: i \in D_j} (\prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j)) \right) \]

\[ Q_i(X_i) = \frac{1}{Z_i} \exp \left( \sum_{D_j: X_i \in D_j} E_Q \left[ \ln \Psi(D_j) \right] \right) \]
Mean Field Algorithm

- Initialize $Q(X_1, ..., X_n) = \prod_i Q(X_i)$ (eg., randomly or smartly)
- Set all variables to unprocessed
- Pick an unprocessed variable $X_i$
  - Update $Q_i$:
    $$Q_i(X_i) = \frac{1}{Z_i} \exp \left( \sum_{D_j: X_i \in D_j} E_Q[\ln \Psi(D_j)] \right)$$
  - Set variable $X_i$ as processed
  - If $Q_i$ changed
    - Set neighbors of $X_i$ to unprocessed

Guaranteed to converge
Understanding mean field algorithm

\[
Q_1(X_1) = \frac{1}{Z_1} \exp \left( E_{Q(X_2)} \left[ \ln \Psi(X_1, X_2) \right] + E_{Q(X_3)} \left[ \ln \Psi(X_1, X_3) \right] + E_{Q(X_4)} \left[ \ln \Psi(X_1, X_4) \right] + E_{Q(X_5)} \left[ \ln \Psi(X_1, X_5) \right] \right)
\]

\[
Q_2(X_2) = \frac{1}{Z_1} \exp \left( E_{Q(X_1)} \left[ \ln \Psi(X_1, X_2) \right] + E_{Q(X_6)} \left[ \ln \Psi(X_2, X_6) \right] + E_{Q(X_7)} \left[ \ln \Psi(X_2, X_7) \right] \right)
\]
Example of Mean Field

- Mean field equation
  \[ Q_i(X_i) = \frac{1}{Z_i} \exp \left( \sum_{D_j : X_i \in D_j} E_Q[\ln \Psi(D_j)] \right) \]

- Pairwise Markov random field
  \[
P(X_1, \ldots, X_n) \propto \exp \left( \sum_{ij} \theta_{ij}X_iX_j + \sum_i \theta_iX_i \right) = \prod_{ij} \exp(\theta_{ij}X_iX_j) \prod_i \exp(\theta_iX_i)
\]

- The mean field equation has simple form:
  \[
  Q_i(X_i) = \frac{1}{Z_i} \exp \left( \sum_{j \in N(i)} \sum_{X_j} \theta_{ij}X_iX_j Q_j(X_j) + \theta_iX_i \right) = \frac{1}{Z_i} \exp \left( \sum_{j \in N(i)} \theta_{ij}X_i < X_j >_{Q_j} + \theta_iX_i \right)
  \]
Example of generalized mean field

\[ Q(C_1) \propto \exp(\sum_{i \in C_1} \theta_i X_i + \sum_{(ij) \in E, i \in C_1, j \in C_1} \theta_{ij} X_i X_j + \sum_k \sum_{i \in C_1, j \in C_k, j \in MB(C_1)} \theta_{ij} X_i < X_j >_{Q(c_k)}) \]

- Node potential within \( C_1 \)
- Edge potential within \( C_1 \)
- Mean of variables in Markov blanket
- Edge potential across \( C_1 \) and \( C_k \)
Why Sampling

- Previous inference tasks focus on obtaining the entire posterior distribution \( P(X_i|e) \)

- Often we want to take expectations
  - Mean \( \mu_{X_i|e} = E[X_i|e] = \int X_i P(X_i|e) dX_i \)
  - Variance \( \sigma^2_{X_i|e} = E[(X_i - \mu_{X_i|e})^2|e] = \int (X_i - \mu_{X_i|e})^2 P(X_i|e) dX_i \)
  - More general \( E[f] = \int f(X) P(X|e) dX \), can be difficult to do it analytically

- **Key idea**: approximate expectation by sample average
  \[
  E[f] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)
  \]
  where \( x_1, ..., x_N \sim P(X|e) \) independently and identically
Sampling

- Samples: points from the domain of a distribution $P(X)$

- The higher the $P(x)$, the more likely we see $x$ in the sample

- Mean $\hat{\mu} = \frac{1}{N} \sum_i x_i$

- Variance $\hat{\sigma}^2 = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2$
Sampling Example

For $n$ independent Gaussian variable with 0 mean and unit variance, use $x \sim \text{randn}(n, 1)$ in Matlab

For multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
- $y \sim \text{randn}(n, 1)$
- Transform the sample: $x = \mu + \Sigma^{1/2}y$

$X$ takes finite number of values, e.g., $\{1, 2, 3\}$, $P(X = 1) = 0.7, P(X = 2) = 0.2, P(X = 3) = 0.1$
- Draw a sample from uniform distribution $U[0,1]$, $u \sim \text{rand}$
  - If $u$ falls into interval $[0, 0.7)$, output sample $x = 1$
  - If $u$ falls into interval $[0.7, 0.9)$, output sample $x = 2$
  - If $u$ falls into interval $[0.9, 1]$, output sample $x = 3$
Generate Samples from Bayesian Networks

- BN describe a generative process for observations

- First, sort the nodes in topological order

- Then, generate sample using this order according to the CPTs

- Generate a set of sample for (A, F, S, N, H):
  - Sample $a_i \sim P(A)$
  - Sample $f_i \sim P(F)$
  - Sample $s_i \sim P(S | a_i, f_i)$
  - Sample $n_i \sim P(N | s_i)$
  - Sample $h_i \sim P(H | s_i)$
Challenge in sampling

- How to draw sample from a given distribution?
  - Not all distributions can be trivially sampled

- How to make better use of samples (not all samples are equally useful)?

- How do know we’ve sampled enough?
Sampling Methods

- **Direct Sampling**
  - Simple
  - Works only for easy distributions

- **Rejection Sampling**
  - Create samples like direct sampling
  - Only count samples consistent with given evidence

- **Importance Sampling**
  - Create samples like direct sampling
  - Assign weights to samples

- **Gibbs Sampling**
  - Often used for high-dimensional problem
  - Use variables and its Markov blanket for sampling
Rejection sampling I

- Want to sample from distribution $P(X) = \frac{1}{Z} f(X)$
- $P(X)$ is difficult to sample, but $f(X)$ is easy to evaluate

Key idea: sample from a simpler distribution $Q(X)$, and then select samples

Rejection sampling (choose $M$ s.t. $f(X) \leq M \cdot Q(X)$)

1. $i = 1$
2. Repeat until $i = N$
   - $x \sim Q(X), u \sim U[0,1]$
   - if $u \leq \frac{f(X)}{M \cdot Q(X)}$, accept $x$ and $i = i + 1$; otherwise, reject $x$

3. Reject sample with probability $1 - \frac{f(X)}{M \cdot Q(X)}$
Rejection sampling II

- Sample $x \sim Q(X)$ and reject with probability $1 - \frac{f(x)}{M Q(x)}$

The likelihood of seeing a particular $x$

$$P(x) = \frac{\frac{f(x)}{M Q(x)} Q(x)}{\int \frac{f(x)}{M Q(x)} Q(x) dX} = \frac{1}{Z} f(x)$$
Drawback of rejection sampling

- Can have very small acceptance rate

Adaptive rejection sampling: use envelop function for $Q$

$\mathbf{f}(X)$

$\mathbf{M} \mathbf{Q}(X)$

$\mathbf{M} \mathbf{Q}(x_1)$

$u_1 \sim \mathbf{U}[0,1]$

$x_1 \sim \mathbf{Q}(X)$

Big rejection region

$X$
Importance Sampling I

- Support sampling from $P(X)$ is hard

- We can sample from a **simpler proposal** distribution $Q(X)$

- If $Q$ dominates $P$ (i.e., $Q(X) > 0$ whenever $P(X) > 0$), we can sample from $Q$ and reweight

$$E_P[f(X)] = \int f(X)P(X)dX = \int f(X)\frac{P(X)}{Q(X)}Q(X)dX$$

  - Sample $x_i \sim Q(X)$
  
  - Reweight sample $x_i$ with $w_i = \frac{P(x_i)}{Q(x_i)}$
  
  - approximate expectation with $\frac{1}{N} \sum_i f(x_i)w_i$
Importance Sampling II

- Instead of reject sample, reweight sample instead

\[ x_2 \sim Q(X) \quad w_2 \sim P(x_2)/Q(x_2) \]
\[ x_1 \sim Q(X) \quad w_1 \sim P(x_1)/Q(x_1) \]

- The efficiency of importance sampling depends on how close the proposal Q is to the target P
Suppose we only evaluate $P(X) = \frac{1}{Z} f(X)$ (e.g., grid or MRF, $Z$ is difficult to compute), and don’t know Z.

We can get around the normalization $Z$ by normalize the importance weight

$$r(X) = \frac{f(X)}{Q(X)} \Rightarrow E_Q[r(X)] = \int \frac{f(X)}{Q(X)} Q(X) dX = \int f(X) dX = Z$$

$$E_P[g(X)] = \int g(X)P(X) dX = \frac{1}{Z} \int g(X) \frac{f(X)}{Q(X)} Q(X) dX = \frac{\int g(X)r(X)Q(X) dX}{\int r(X)Q(X) dX} \approx \frac{\sum g(x_i)r(x_i)}{\sum r(x_i)}$$

define $w_i = \frac{r(x_i)}{\sum r(x_i)}$ as the new importance weight.
Difference between normalized and unnormalized

- Unnormalized importance sampling is unbiased

- Normalized importance sampling is biased, eg. for $N=1$
  \[ E_Q \left[ \frac{f(x_1)w_1}{w_1} \right] = E_Q [f(x_1)] \neq E_P [f(x_1)] \]

- However, the normalized importance sampler usually has lower variance

- But most importantly, unnormalized importance sampler work for unnormalized distributions

- For unnormalized sampler, one can also do resampling based on $w_i$ (multinomial distribution with probability $w_i$ for $x_i$)
Gibbs Sampling

- Both rejection sampling and importance sampling do not scale well to high dimensions

- Markov Chain Monte Carlo (MCMC) is an alternative

- **Key idea**: Construct a Markov chain whose stationary distribution is the target distribution $P(X)$

- Sampling process: random walk in the Markov chain

- Gibbs sampling is a very special and simple MCMC method.
Markov Chain Monte Carlo

- Wan to sample from $P(X)$, start with a random initial vector $X$
  - $X^t$: $X$ at time step $t$
  - $X^t$ transition to $X^{t+1}$ with probability
    - $Q(X^{t+1} | X^t, \ldots, X^1) = T(X^{t+1} | X^t)$
  - The stationary distribution of $T(X^{t+1} | X^t)$ is our $P(X)$

- Run for an initial $M$ samples (burn-in time) until the chain converges/mixes/reaches the stationary distribution

- Then collect $N$ (correlated) sample as $x_i$

- **Key issues:** Designing the transition kernel, and diagnose convergence
Gibbs Sampling

- A very special transition kernel, works nicely with Markov blanket in GMs.

- The procedure
  - We have variables set $X = \{X_1, \ldots, X_K\}$ variables in a GM.
  - At each step, one variable $X_i$ is selected (at random or some fixed sequence), denote the remaining variables as $X_{-i}$, and its current value as $x_{-i}^t$
  - Compute the conditional distribution $P(X_i \mid x_{-i}^t)$
  - A value $x_i^t$ is sampled from this distribution
  - This sample $x_i^t$ replaces the previous sampled value of $X_i$ in $X$
Gibbs Sampling in formula

- **Gibbs sampling**
  - \( X = x^0 \)
  - For \( t = 1 \) to \( N \)
    - \( x_1^t = P(X_1 \mid x_2^{t-1}, \ldots, x_K^{t-1}) \)
    - \( x_2^t = P(X_2 \mid x_1^t, \ldots, x_K^{t-1}) \)
    - ... 
    - \( x_K^t = P(X_K \mid x_1^t, \ldots, x_{K-1}^t) \)

- **Variants:**
  - Randomly pick variable to sample
  - sample block by block

*For graphical models*
Only need to condition on the Variables in the Markov blanket
Gibbs Sampling: Examples

- Pairwise Markov random field
  \[ P(X_1, ..., X_k) \propto \exp(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \theta_i X_i) \]

- In the Gibbs sampling step, we need conditional \( P(X_1 | X_2, ..., X_k) \)
  \[ P(X_1 | X_2, ..., X_k) \propto \exp(\theta_{12} X_1 X_2 + \theta_{13} X_1 X_3 + \theta_{14} X_1 X_4 + \theta_{15} X_1 X_5 + \theta_1 X_1) \]
Diagnose convergence

- Good chain

Sampled Value

Iteration number
Diagnose convergence

- Bad chain

Sampled Value

Iteration number
Sampling Methods

- Direct Sampling
  - Simple
  - Works only for easy distributions
- Rejection Sampling
  - Create samples like direct sampling
  - Only count samples consistent with given evidence
- Importance Sampling
  - Create samples like direct sampling
  - Assign weights to samples
- Gibbs Sampling
  - Often used for high-dimensional problem
  - Use variables and its Markov blanket for sampling