Approximate Inference Algorithms

Jie Tang

Department of Computer Science and Technology

Tsinghua University

March 17, 2020

Contents

1. Basic Monte Carlo Sampling Methods
2. Markov Chain Monte Carlo (MCMC)
3. Mean-field Variational Inference
4. Conclusion

Contents

1 Basic Monte Carlo Sampling Methods

2 Markov Chain Monte Carlo (MCMC)

3 Mean-field Variational Inference

4 Conclusion

Basci Monte Carlo Sampling Methods

- What is a sample from a know distribution?
- A simple algorithm.
- Motivation
- Ancestral Sampling
- Transformation Method
- Rejection Sampling
- Importance Sampling

What is a Sample from a Given Distribution?

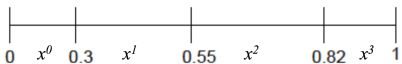
- Given a set of random variables $z = \{z_1, \cdots, z_L\}$.
- A sample is an instantiation of an assignment to all variables:

$$z^t = \{z_1^t, \cdots, z_L^t\}$$

- Each variable in a sample takes one possible allowable value in its domain according to a probability distribution defined over x.

Algorithm to Generate Univariate Samples.

- Assume the domain of x is $\{x^1, x^2, x^3, x^4\}$, the distribution is $p(x = x^1) = 0.3$,$p(x = x^2) = 0.25$,$p(x = x^3) = 0.27$,$p(x = x^4) = 0.18$.
- We divide a real line $[0, 1]$ into K intervals such that width of interval is proportional to $p(x = x^j)$:



- Draw a random number $u \in [0, 1]$ uniformly;
- Determine region j in which u lies;
- Output x^j.

Motivation of Sampling Method

- When exact inference is intractable, we need some form of approximation
- Most situations will require evaluating expectations of unobserved variables, e.g., to make predictions, rather than the posterior distribution.

Sampling-based Estimation

- L samples $x^{(1)}, \cdots x^{(L)}$ are drawn iid from p.
- If $p(x = 1) = p$, estimator for p:

$$\hat{p} = \frac{1}{L} \sum_{l=1}^{L} x^{(l)}$$

- More generally, for any distribution p, function f,

$$\mathbb{E}[f] \approx \hat{f} = \frac{1}{L} \sum_{k=1}^{L} f(x^{(l)})$$

Sampling Bound

- Hoeffding Bound:

$$p(\hat{p} \notin [p - \epsilon, p + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

- For additive bound ϵ on error with probability $> 1 - \delta$:

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

- Chernoff Bound:

$$p(\hat{p} \notin [p(1 - \epsilon), p(1 + \epsilon)]) \leq 2e^{-Mp\epsilon^2/3}$$

- For multiplicative bound ϵ on error with probability $> 1 - \delta$:

$$M \geq 2\frac{\ln(2/\delta)}{p\epsilon^2}$$

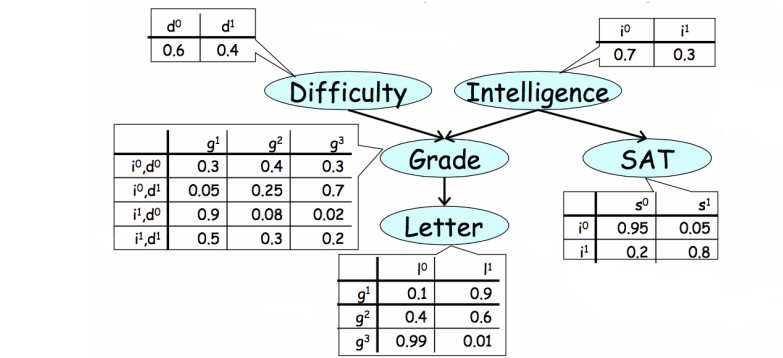
Ancestral Sampling

- If joint distribution is represented by a BN, then the distribution is specified by :

$$p(z) = \prod_{i=1}^{M} p(z_i|pa_i)$$

- To obtain sample from the joint distribution, we start with lowest numbered node (Assume each node has a higher numbered node than its parents), make one pass through set of variable sin order $z_1, \ldots, z_M$, sampling from conditional distribution $p(z|pa_i)$
- After one pass through the graph, we obtain one sample.
- To obtain sampling from marginal distribution, we can sample from full distribution and discard unnecessary values.

Ancestral Sampling - Example



- CPDs: $p(z|pa_i)$
- Joint distribution:
  $p(D, I, G, S, L) = p(D)p(I)p(G|D, I)p(S|I)p(L|G)$
- Generate samples in order of D,I, G, S,L

Ancestral Sampling with Evidence

- If the directed graph has some nodes instantiated with observed values, the method is called *Logic sampling*.
- Use ancestral sampling, except that when sample is obtained from an observed value:
  - if they agree then sample value is retained and proceed to next variable
  - if they don't agree, discard the whole sample.
- Logic sampling corresponds to sampling from joint distribution of hidden and data variables.

Ancestral Sampling - Limitations

- (ϵ, δ)-bounds exist, but usefulness is limited:
  - Additive bounds: useless for low Probability events.
  - Multiplicative bounds: the number of samples grows as $\frac{1}{p(y)}$
- With evidence, the probability of accepting sample decreases as the number of variables increase and the number of stats that variables can take increases.
- Ancestral sampling generally infeasible for MNs.

Transformation Method

- We first consider how to generate random numbers from simple nonuniform distributions, assuming that we have a source of uniformly distributed random numbers

$$f(z) = \begin{cases} \frac{1}{b-a} & a \leq z < b \\ 0 & \text{otherwise.} \end{cases}$$

- Suppose that x is uniformly distributed over the interval $(0, 1)$, thus, $p(z) = 1$ within $(0, 1)$.
- Then we transform the values of z using some function $f(\cdot)$ so that $y = f(z)$.
- The distribution of y will be governed by

$$p(y) = \frac{dh}{dy} = \frac{dh}{dz} \cdot \frac{dz}{dy} = p(z)\frac{dz}{dy} \tag{1}$$

where $h(\cdot)$ is the accumulated distribution of $p(\cdot)$.

Transformation Method

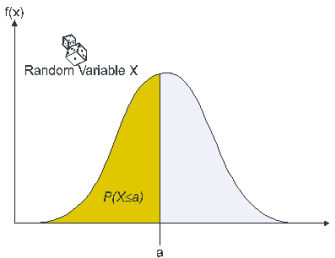
- Our goal is to choose the function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$.
- Integrating Eq. (1) we obtain

$$\begin{aligned} h(y) &= \int_{-\infty}^{y} p(\hat{y}) d\hat{y} \\ &= \int_{-\infty}^{y} \frac{dz}{dy} d\hat{y} \\ &= \int_{-\infty}^{y} \frac{dg(y)}{dy} d\hat{y} \\ &= g(y) = z \end{aligned}$$

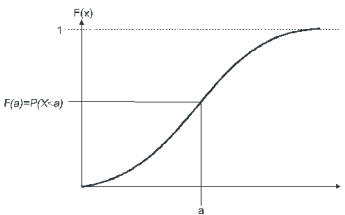
- Thus, $y = h^{-1}(z)$
- So we have to transform the uniformly distributed integral of the desired distribution.

Transformation Method - Example

- Eg. $p(y) = \lambda \exp(-\lambda y)$ where $0 \leq y < \infty$.
  Solution:
  We get $h(y) = 1 - \exp(-\lambda y)$,
  If we transform our uniformly distributed variable z using $y = -\lambda^{-1} \ln(1 - z)$, then y will have an exponential distribution.



(a) PDF



(b) CDF

Transformation Method - Limitation

- Need to first calculate and then invert indefinite integral of the desired distribution
- Feasible only for a small number of distributions

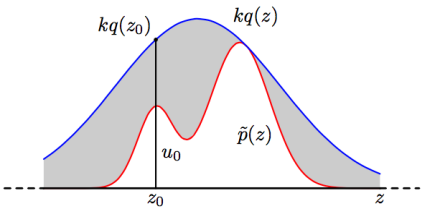
Rejection Sampling

- Distribution $p(z)$ is not a simple standard distribution, and that sampling from $p(z)$ is difficult.
- Need a simpler distribution $q(z)$, called *proposal distribution*, to draw samples.
- Introduce a constant k, such that $kq(z) \leq \tilde{p}(z)$.
- Suppose it is easy to evaluate $p(z)$ for any value of z, up to some normalizing constant Z, so that:

$$p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

where $\tilde{p}(z)$ can readily be evaluated, but Z_p is unknown.

Rejection Sampling - Steps



1. Generate a number z_0 from the distribution $q(z)$.
2. Generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$
   If $u_0 > \tilde{p}(z_0)$ then the sample is rejected, otherwise u_0 is retained.

Rejection Sampling - Acceptance Probablity

- The original values of z are generated from $q(z)$.
- These samples are then accepted with probability $\tilde{p}(z)/kq(z)$.
- So the probability that a sample will be accepted is given by

$$\begin{aligned} p(\text{accept}) &= \int \{\tilde{p}(z)/kq(z)\} q(z) dz \\ &= \frac{1}{k} \int \tilde{p}(z) dz \end{aligned}$$

- Thus the fraction of points that are rejected by this method depends on the ratio of the area under the unnormalized distribution $\tilde{p}(z)$ to the area under the curve $kq(z)$.
- So the constant k should be as small as possible subject to the limitation that $kq(z)$ must be nowhere less than $\tilde{p}(z)$.

Rejection Sampling - Example

- E.g., $p(x) = \frac{e}{e-1} e^{-x} e^{-e^{-x}}$, $0 < x < \infty$, how can we determine $q(x)$?
- Solution:
  If $x > 0$, then $e^{-e^{-x}} < e^0 = 1$
  Then, $p(x) = \frac{e}{e-1} e^{-x} e^{-e^{-x}} \leq \frac{e}{e-1} e^{-x}$
  Then, set $q(x) = e^{-x}$, $c = \frac{e}{e-1}$
  Thus $p(x) \leq c \cdot q(x)$ with $q(x)$ easy to draw samples.
- The principle is to keep the familiar one, replace the unfamiliar one with the upper bound.

Importance Sampling

- Provides a framework for approximating **expectations** directly but does not itself provide a mechanism for drawing samples from distribution $p(z)$.
- It also uses a proposal distribution $q(z)$ to easy draw samples.
- We can express the expectation in the form of a finite sum over samples $\{z^{(l)}\}$ drawn from $q(z)$:

$$\begin{aligned}\mathbb{E} &= \int f(z)p(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})\end{aligned}$$

- The quantities $r_l = \frac{p(z^{(l)})}{q(z^{(l)})}$ are known as importance weights, and they correct the bias introduced by sampling from the wrong distribution.

Importance Sampling - Unnormalized Case

- It will often be the case that the distribution $p(z)$ can only be evaluated up to a normalization constant, so that $p(z) = \tilde{p}(z)/Z_p$ where $\tilde{p}(z)$ can only be evaluated easily, whereas Z_p is unknown.
- Similarly, we may wish to use an importance sampling distribution $q(z) = \tilde{q}(z)/Z_q$. We then have

$$\begin{aligned}
\mathbb{E} &= \int f(z)p(z)dz \\
&= \frac{Z_q}{Z_p}\int f(z)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)da \\
&\approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)})
\end{aligned}$$

where $\tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$.

Importance Sampling - Unnormalized Case (Cont.)

- We can use the same sample set to evaluate the ratio $\frac{Z_p}{Z_q}$ with the result

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(z)dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l\end{aligned}$$

and hence

$$\mathbb{E} \approx \sum_{l=1}^{L}\omega_l f(z^{(l)})$$

where we have defined

$$\omega_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(l)})/\tilde{q}(z^{(l)})}{\sum_m \tilde{p}(z^{(m)})/\tilde{q}(z^{(m)})}$$

Importance Sampling - Example

- Eg., $p(x) \propto e^{-\frac{x^4}{4}}$, $-\infty < x < \infty$, estimate $\mathbb{E}(x)$
  Solution:
  We set $q(x) = \frac{1}{\sqrt{2\phi}} e^{-\frac{x^2}{2}}$,
  Then draw $x^{(1)} \dots x^{(L)}$ iid from $q(x)$,
  The weight $\omega_l = \frac{e^{-\frac{(x^{(l)})^4}{4}}}{e^{-\frac{(x^{(l)})^2}{2}}}$,
  Estimate $\mathbb{E}(x) = \frac{\sum_{l=1}^{L} \omega_l x^{(l)}}{\sum_{l=1}^{L} \omega_l}$
- The principle is to choose $q(x)$ as close as $f(x)p(x)$

Rejection vs. Importance Sampling

- Rejection sampling
  - Select sample from $p(x)$ - give you samples from $p(x)$
- Importance sampling
  - Reweight samples from $p(x)$ - do not give you samples from $p(x)$

Sampling-Importance Resampling (SIR)

- Rejection sampling depends on suitable value of k where $kq(z) > p(z)$
  - It is impractical to determine a suitable k in that any value that is sufficiently large to guarantee a bound on the desired distribution will lead to impractically small acceptance rates.
- SIR is the approach makes use of a sampling distribution $q(z)$ but avoids having to determine the constant k.

SIR - Steps

1. L samples $z^{(1)}, \cdots, z^{(L)}$ are drawn from $q(z)$.
2. Weights $\omega_1, \cdots, \omega_L$ are constructed as in importance sampling.
3. S second set of L samples are drawn from the discrete distribution $\{z^{(1)}, \cdots, z^{(L)}\}$ with probabilities given by $\{\omega_1, \cdots, \omega_L\}$.

The resulting L samples are only approximately distributed according to $p(z)$, but the distribution becomes correct i the limit $L \to \infty$.

SIR - Proof

- Consider univariate case, and note that the cumulative distribution of the resampled values if given by

$$\begin{aligned} p(z \leq a) &= \sum_{l: z^{(l)} \leq a} \omega_l \\ &= \frac{\sum_l I(z^{(l)} \leq a) \tilde{p}(z^{(l)})/q(z^{(l)})}{\sum_l \tilde{p}(z^{(l)})/q(z^{(l)})} \end{aligned}$$

- Taking the limit $L \to \infty$, we replace the sums by integrals

$$\begin{aligned} p(z \leq a) &= \frac{\int I(z \leq a)\{\tilde{p}(z)/q(z)\}q(z)dz}{\int \{\tilde{p}(z)/q(z)\}q(z)dz} \\ &= \frac{\int I(z \leq a)\tilde{p}(z)dz}{\int \tilde{p}(z)dz} \\ &= \int I(z \leq a)p(z)dz \end{aligned}$$

Contents

1 Basic Monte Carlo Sampling Methods

2 Markov Chain Monte Carlo (MCMC)

3 Mean-field Variational Inference

4 Conclusion

Markov Chain Monte Carlo (MCMC)

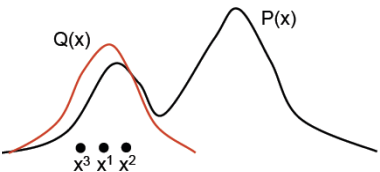
- Essentially, MCMC is widely applied to solve integration and optimization problems in large dimensional space.
  - **Bayesian inference and learning.** Given some unknown variables $y \in \mathcal{Y}$ and data $x \in \mathcal{X}$,
    - *Normalization.*
    $$p(y|x) = \frac{p(x|y)p(y)}{\int_{\mathcal{Y}} p(x|y')p(y')dy'}$$
    - *Marginalization.*
    $$p(y|x) = \int_z p(y, z|x)dz$$
    - *Expectation.*
    $$\mathbb{E}_{p(y|x)}(f(y)) = \int_{\mathcal{Y}} f(y)p(y|x)dy$$
  - **Optimization.** To obtain the solution that minimizes some objective function from a large set of feasible solutions.
  - **Model selection.** To obtain models that are of the most interest. For example first apply maximum likelihood (ML) estimates for each model separately, and then apply MDL, BIC, or AIC to select the best models. The initial space might be very large.

Limitations of Monte Carlo

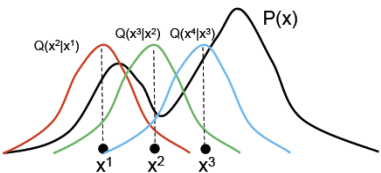
- Direct (unconditional) sampling
  - Hard to get rare events in high-dimensional spaces
  - Infeasible for MRFs, unless we know the normalizer Z
- Rejection sampling, Importance sampling
  - Do not work well if the proposal Q(x) is very different from P(x)
  - Yet constructing a Q(x) similar to P(x) can be difficult
    - Making a good proposal usually requires knowledge of the analytic form of $P(x)$ – but if we had that, we wouldn't even need to sample!
- Intuition: instead of a fixed proposal $Q(x)$, what if we could use an adaptive proposal?

Markov Chain Monte Carlo

- MCMC algorithms feature adaptive proposals
  - Instead of $Q(x)$, they use $Q(x'|x)$ where x' is the new state being sampled, and x is the previous sample
  - As x changes, $Q(x'|x)$ can also change (as a function of x')



(c) Importance sampling with $Q(x')$



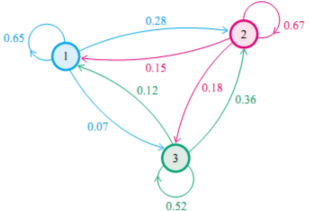
(d) MCMC with $Q(x'|x)$

Markov Chain

- A Markov Chain is a sequence of random variables $x^1, x^2, \cdots, x^n$ with the Markov property

$$P(x^n = x|x^1, \cdots, x^{n-1}) = P(x^n = x|x^{n-1})$$

  - where $P(x^n = x|x^{n-1})$ is called transition kernel
  - x^n only depends on x^{n-1}. Recall HMM!
  - To simplify, we use T to denote the transition matrix and $T(x'|x)$ the transition kernel from state x to x'



$$T = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$$

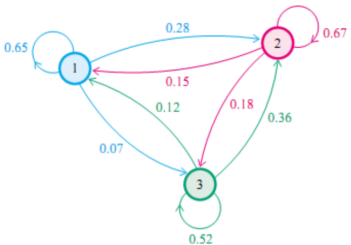
Invariant Distribution

- One major concern of Markov Chain theory is to determine
  - conditions under which there exists an invariant distribution $P(x^i)$
  - conditions under which iterations of the transition kernel converge to the invariant distribution
- The invariant distribution satisfies

$$P(x^i) = \sum_{x^{i-1}} P(x^{i-1}) T(x^i | x^{i-1})$$

- If $P(x^n)$ converge to the invariant distribution as $n \to \infty$ under the transition kernel, then we also call $P(x^n)$ as a stationary distribution.

Invariant Distribution: An Example



$$T = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$$

- Initial distribution

  $\pi_0 = [0.21, 0.68, 0.11]$

iteration	1	2	3
0	0.210	0.680	0.110
1	0.252	0.554	0.194
2	0.270	0.512	0.218
3	0.278	0.497	0.225
4	0.282	0.490	0.225
5	0.286	0.489	0.225
6	0.286	0.489	0.225
7	0.286	0.489	0.225
...	...	...	...

- Invariant distribution

  $\pi = [0.286, 0.489, 0.225]$

Invariant Distribution

- X_i is random variable in i-th iteration.

$$
\begin{aligned}
X_0 &\sim \pi_0(x) \\
X_1 &\sim \pi_1(x) \\
X_2 &\sim \pi_2(x) \\
&\cdots \\
X_n &\sim \pi_n(x) = \pi(x) \\
X_{n+1} &\sim \pi(x) \\
X_{n+2} &\sim \pi(x) \\
&\cdots
\end{aligned}
$$

- $x_n, x_{n+1}, x_{n+2}, \cdots$ are samples of given distribution $P(x)$ if invariant distribution $\pi(x) = P(x)$.

MC Concepts and Reversibility Condition

- MCMC methods turn the Markov Chain theory around:
  - $P(.)$ is known, but $T(x'|x)$ is unknown.
  - To generate samples from $P(.)$, the methods find transition kernels $T(x'|x)$ whose n^{th} iterate converges to $P(.)$ for large n
- The problem then is to find an appropriate kernel $T(x'|x)$.
- Reversibility condition (detailed balance, microscopic reversibility, and time reversibility):

$$P(x')T(x|x') = P(x)T(x'|x)$$

- then $P(x)$ is a stationary distribution of the MC! Proof:

$$\begin{aligned}
P(x')T(x|x') &= P(x)T(x'|x) \\
\sum_x P(x')T(x|x') &= \sum_x P(x)T(x'|x) \\
P(x')\sum_x T(x|x') &= \sum_x P(x)T(x'|x) \\
P(x') &= \sum_x P(x)T(x'|x)
\end{aligned}$$

Metropolis-Hastings

- Metropolis-Hastings is one of the best known MCMC methods
  - Proposed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and generalized by Hastings (1970)
  - Refer to Chib and Greenberg (1995) for a detailed introduction
  - Gibbs sampler is a special case of MH
- Metropolis-Hastings (MH)
  - Draws a sample x' from $Q(x'|x)$
  - Estimate acceptance probability

$$A(x'|x) = \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$$

  - Accept the new sample x' according to the probability $A(x'|x)$. $A(x'|x)$ is like a ratio of importance sampling weights
  - If accept, then $x = x'$

The MH Algorithm

- The algorithm process
  - Initialize staring state x^t, $t = 0$
  - Burn-in:
    - set $x = x^t$
    - $t = t + 1$
    - sample $x' \sim Q(x'|x)$
    - calculate $A(x'|x)$

$$A(x'|x) = \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$$

    - sample $\mu \sim Uniform(0, 1)$
    - if $A(x'|x) > \mu$ then
    - $\quad x^t = x'$
    - else
    - $\quad x^t = x$
  - Take samples from $P(x)$
    - $x^{t+1} \leftarrow$ Draw samples $x^{(t)}$

Why does MH work?

- Recall that we draw a sample x' according to $Q(x'|x)$, and then accept/reject (A-R) according to $A(x'|x)$
- Generally speaking, the $Q(x'|x)$ distribution corresponding to the A-R process does not satisfy the reversibility condition, e.g.,

$$P(x)Q(x'|x) > P(x')Q(x|x')$$

  - This basically means that the process moves from x to x' too often and from x' to x too rarely.
  - To balance this, we introduce a probability of movement $\alpha(x, x') < 1$, thus define the transition kernel as

$$T(x'|x) = Q(x'|x)\alpha(x, x')$$

  - On the other hand, we should define $\alpha(x', x)$ as large as possible. However, since it is a probability, its upper limit is 1. We then have

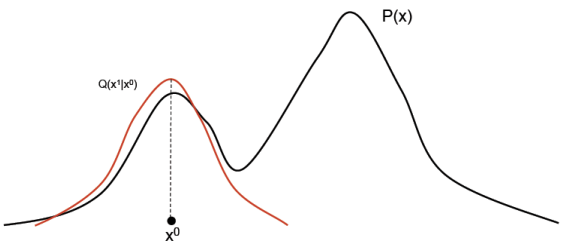
$$P(x)Q(x'|x)\alpha(x, x') = P(x')Q(x'|x)$$

Why does MH work? (cont.)

- Now we have
$$\alpha(x, x') = \frac{P(x)Q(x'|x)}{P(x')Q(x|x')}$$
- To ensure that the transition kernel $T(x'|x)$ satisfies the reversibility condition, the probability of move $\alpha(x, x')$ must be set to
$$\begin{aligned} \alpha(x, x') = A(x'|x) &= \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right), && \text{if } P(x)Q(x'|x) > 0 \\ &= 1, && \text{otherwise.} \end{aligned}$$
- Finally we have (this is exactly the detailed balance condition)
$$P(x)T(x'|x) = P(x')T(x|x')$$
- Now, we prove that MH eventually converges to the stationary distribution $P(x)$.

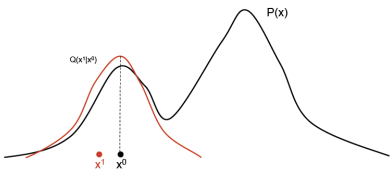
The MH Algorithm

- Example
  - Let $Q(x'|x)$ be a Gaussian centered on x
  - $P(x)$ itself is a biomodal distribution.
- Initialize x^0

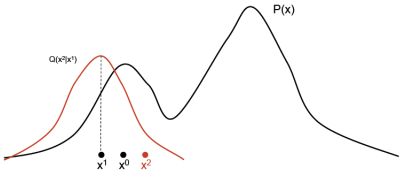


The MH Algorithm

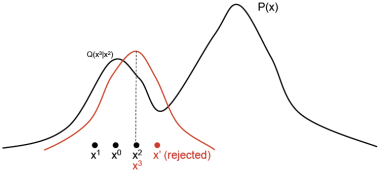
$$A(x'|x) = \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$$



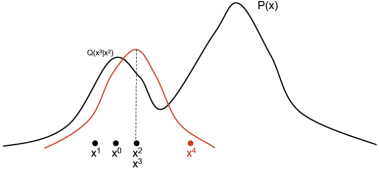
(a) Draw x^1, accept



(b) Draw x^2, accept



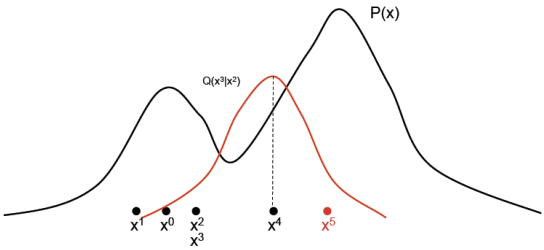
(c) Draw x^3, reject, because $\frac{P(x^3)}{P(x^2)}$ is very small and thus $A(x^3|x^2)$ is small



(d) Draw x^4, accept

The MH Algorithm (cont.)

$$A(x'|x) = \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$$



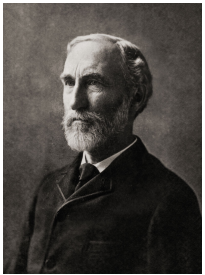
(e) Draw x^5, accept.

The adaptive proposal $Q(x'|x)$ allows us to sample both modes of $P(x)$!

Gibbs Sampling

- Gibbs Sampling is a special case of the MH algorithm that samples each random variable of a graphical model
- Consider distribution $P(\mathbf{x}) = P(x_1, \cdots, x_n)$ from which we wish to sample
- GS algorithms
  - Are fairly easy to derive for many graphical models (e.g. mixture models, Latent Dirichlet allocation)
  - Have reasonable computation and memory requirements, because they sample one r.v. at a time

Josiah Willard Gibbs



- 1839-1903, Born New Haven Connecticut, PhD in Engineering in US
- Invented so many ...
  - Gibbs entropy
  - Gibbs free energy
  - Gibbs paradox
  - Vector calculus
  - Gibbs algorithm

Gibbs Sampling

- Suppose we have a $n-$dimensional vector $\{x_1, \cdots, x_n\}$ and the expressions for the full conditionals $P(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$
- Initialize $x_1, \cdots, x_n$
- For $t = 0, \cdots, T$
  - Sample $x_1^{(t+1)}$ according to $P(x_1^{(t+1)}|x_2^{(t)}, \cdots, x_n^{(t)})$;
  - Sample $x_2^{(t+1)}$ according to $P(x_2^{(t+1)}|x_1^{(t+1)}, x_3^{(t)}, \cdots, x_n^{(t)})$;
  - $\cdots\cdots$
  - Sample $x_n^{(t+1)}$ according to $P(x_n^{(t+1)}|x_1^{(t+1)}, \cdots, x_{(n-1)}^{(t+1)})$;
- When we update x_i, we immediately use its new value for sampling other variables x_j

Gibbs Sampling is a special case of MH

- The GS transition kernel is (Recall MH)

$$T(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i}) = Q(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i})A(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i})$$

- The GS proposal distribution is

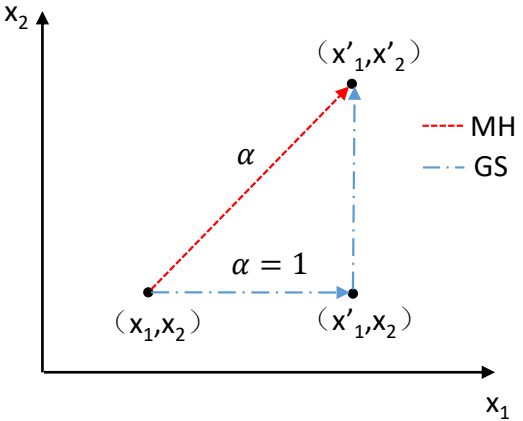
$$Q(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i}) = P(x_i'|\mathbf{x}_{-i})$$

- The corresponding acceptance probability is:

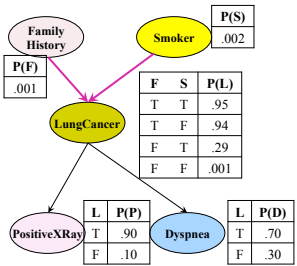
$$\begin{aligned} A(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i}) &= \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})Q(x_i, \mathbf{x}_{-i}|x_i', \mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})Q(x_i', \mathbf{x}_{-i}|x_i, \mathbf{x}_{-i})}\right) \\ &= \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})P(x_i|\mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})P(x_i'|\mathbf{x}_{-i})}\right) \\ &= \min\left(1, \frac{P(x_i'|\mathbf{x}_{-i})P(\mathbf{x}_{-i})P(x_i|\mathbf{x}_{-i})}{P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})P(x_i'|\mathbf{x}_{-i})}\right) \\ &= \min(1, 1) = 1 \end{aligned}$$

- Thus, GS is a MH with a proposal that is always accepted!

Gibbs sampling is a special case of MH



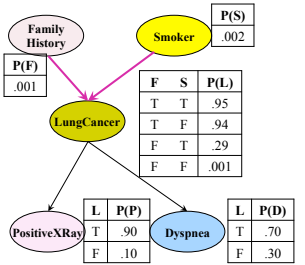
Gibbs Sampling: An Example



t	F	S	L	P	D
0	F	F	F	F	F
1					
2					
3					
4					

- Consider the long cancer network
  - Assume we sample variables in the order F, S, L, P, D
  - Initialize all variables at t = 0 to False

Gibbs Sampling: An Example



t	F	S	L	P	D
0	F	F	F	F	F
1	**F**				
2					
3					
4					

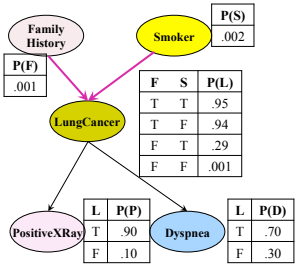
- Sampling $P(F|S, L)$ at $t = 1$, using Bayes Rule,

$$P(F|S, L) \propto P(L|F, S)P(F)$$

- As $S = F$ and $L = F$, so we have

$$P(F = T|S = F, L = F) \propto 0.06 \times 0.01 = 0.0006$$
$$P(F = F|S = F, L = F) \propto 0.999 \times 0.999 = 0.998$$

Gibbs Sampling: An Example



t	F	S	L	P	D
0	F	F	F	F	F
1	F	**T**			
2					
3					
4					

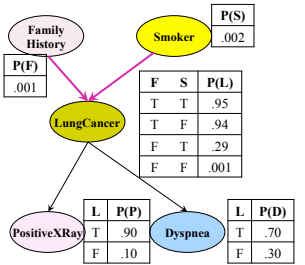
- Sampling $P(S|F, L)$ at $t = 1$,

$$P(S|F, L) \propto P(L|F, S)P(S)$$

- As $F = F$ and $L = F$, so we have

$$P(S = T|F = F, L = F) \propto 0.71 \times 0.02 = 0.0142$$
$$P(S = F|F = F, L = F) \propto 0.999 \times 0.998 = 0.997$$

Gibbs Sampling: An Example



t	F	S	L	P	D
0	F	F	F	F	F
1	F	T	F		
2					
3					
4					

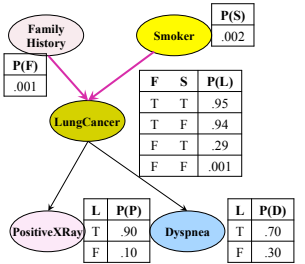
- Sampling $P(L|F,S,P,D)$ at $t=1$,

$$P(L|F,S,P,D) \propto P(L|F,S)P(P|L)P(D|L)$$

- As $(F,S,P,D) = (F,T,F,F)$, so we have

$$P(L=T|F=F,S=T,P=F,D=F) \propto 0.1 \times 0.3 \times 0.29 = 0.0087$$
$$P(L=F|F=F,S=T,P=F,D=F) \propto 0.95 \times 0.99 \times 0.71 = 0.6678$$

Gibbs Sampling: An Example

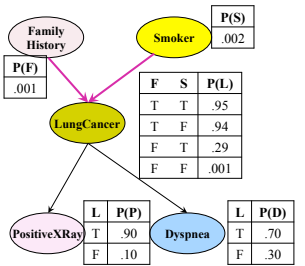


t	F	S	L	P	D
0	F	F	F	F	F
1	F	T	F	**T**	
2					
3					
4					

- Directly sampling $P(P|L)$ at $t = 1$,
- As $L = F$, so we have

$$P(P = T|L = F) \propto 0.10$$
$$P(P = F|L = F) \propto 0.90$$

Gibbs Sampling: An Example

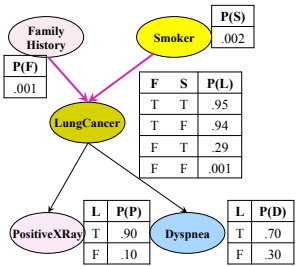


t	F	S	L	P	D
0	F	F	F	F	F
1	F	T	F	T	**F**
2					
3					
4					

- Directly sampling $P(D|L)$ at $t = 1$,
- As $L = F$, so we have

$$P(D = T|L = F) \propto 0.30$$
$$P(D = F|L = F) \propto 0.70$$

Gibbs Sampling: An Example



t	F	S	L	P	D
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3	T	F	T	F	T
4	T	F	T	F	F

- Similarly, for $t = 2, 3, 4$, we repeat the procedure to sample new values of F, S, L, P, D

One More: Monte Carlo EM

- The EM algorithm for ML and MAP. Suppose we have hidden variable $y \in \mathcal{Y}$, observation x and parameters θ. With EM, we can iteratively update the following two steps:
  - E Step.
$$Q(\theta) = \sum_{\mathcal{Y}} \log(P(y, x|\theta)) P(y|x, \theta^{old})$$
  - M Step. Perform the maximization
$$\theta^{new} = \arg\max_{\theta} Q(\theta)$$
- In practice, the expectation might be difficult if the space $\mathcal{Y}$ is huge.
- We consider replace the expectation with a small sum of samples by MCMC

One More: Monte Carlo EM

- Initialize (y^0, θ^0)
- For $i = 0, \cdots$
  - Sample $\{y^j\}_{j=1}^{N_i}$ with an MCMC algorithm, e.g.,

$$A = \min\left(1, \frac{P(x|y', \theta)P(y'|\theta)Q(y^j|y')}{P(x|y^j, \theta)P(y^j|\theta)Q(y'|y^j)}\right)$$

  - E Step.

$$\hat{Q}(\theta) = \frac{1}{N_i}\sum_{j=1}^{N_i}\log(P(y^j, x|\theta))$$

  - M Step. Maximize

$$\theta^{new} = \arg\max_{\theta}\hat{Q}(\theta)$$

- For more, please read stochastic EM (Celeux & Diebolt, 1985), Monte Carlo EM (Wei & Tanner, 1990).

Contents

1 Basic Monte Carlo Sampling Methods

2 Markov Chain Monte Carlo (MCMC)

3 Mean-field Variational Inference

4 Conclusion

Recall Probabilistic Inference

- Let $P(x_1, \cdots, x_n; z_1, \cdots, z_m)$ is the distribution of interest over variables $\{x_1, \cdots, x_n\}$ and $\{z_1, \cdots, z_m\}$.
  - "visible" variables x_v whose marginal distribution $P(x_v)$ we are interested in evaluating and
  - "hidden" variables z_h whose posterior distribution $P(z_h|x_v)$ we want

  evaluating the marginal or posterior involves summing over all configuration of the hidden variables z_h

$$P(x_v) = \sum_{z_h} P(x_v, z_h)$$

- To find a variational solution we need to cast these computations optimization problems

Variational Probabilistic Inference

- We start with a KL-divergence based objective function

$$J(Q) = -KL(Q(z_h) \| P(z_h|x_v)) + \log(P(x_v))$$

where Q is a variational distribution over the hidden variables z_h.

  - If $Q(z_h) = P(z_h|x_v)$, the posterior probability over the hidden variables z_h, we recover the desired log-marginal $\log P(x_v)$
  - $\log P(x_v) \geq J(Q)$ for all Q—J constitutes a lower bound of $\log P(x_v)$
  - The slack in the bound is given by the KL-divergence between Q and the true posterior.

Variational Probabilistic Inference

- The objective function can be rewritten as

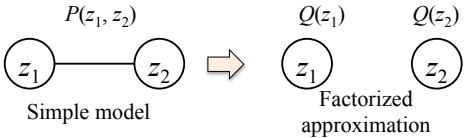
$$J(Q) = -KL(Q(z_h)\|P(z_h|x_v)) + \log(P(x_v))$$
$$= \sum_{z_h} Q(z_h) \log P(z_h, x_v)) + H(Q)$$

  where $H(Q)$ is the (Shannon) entropy of Q.
- The variational formulation $Q(z_h)$ is still intractable and one simple idea is to assume that Q can be factorized, following the mean field distributions

$$Q(z_h) = \prod_{i \in h} Q_i(z_i)$$

  while keeping the remaining marginals fixed.
- Again, we could have an EM algorithm
  - E-step: Find Q to get a good lower bound of the $\log P$, by fixing all the remaining marginals
  - M-step: Update P based on the obtained Q in the E-step.

Simple Example

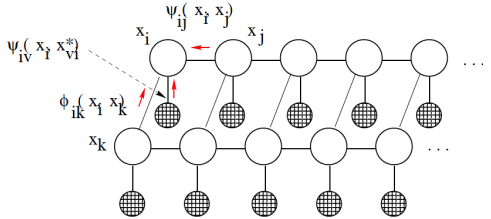


- The updated are made "locally" by geometrically averaging the ignored dependencies with respect to the other component distributions

$$Q_1(z_1) = \frac{1}{Z_1} \exp\left(\sum_{z_2} Q_2(z_2) \log P(z_1, z_2)\right)$$

  where Z_1 normalizes the right hand side across z_1.
- In other words, each z_i only sees the mean effect of the dependencies
- These update equations are the "fixed point equations"

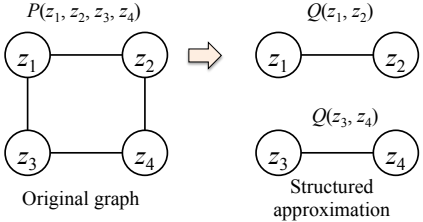
Application to undirected graphs



- Example: coupled HMMs

$$Q_i(x_i) = \frac{1}{Z_i} \times \overbrace{\psi_{iv}(x_i, x_{vi}^*)}^{\text{evidence at } i} \times \overbrace{\exp(\sum_{x_j} Q_j(x_j) \log \psi_{ij}(x_i, x_j))}^{\text{within chain dependencies}} \underbrace{\exp(\sum_{x_k} Q_k(x_k) \log \phi_{ik}(x_i, x_k))}_{\text{between chains dependencies}}$$

Structured mean field approach



We can approximate a complicated graph model with a tractable one

$$P(z) = \frac{1}{Z}\psi_{12}(z_1, z_2)\psi_{13}(z_1, z_3)\psi_{34}(z_3, z_4)\psi_{42}(z_2, z_4)$$

$$Q(z) = Q(z_1, z_2)Q(z_3, z_4)$$

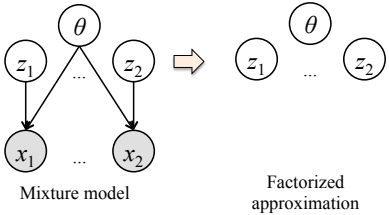
$$= \frac{1}{Z_{12}} \overbrace{\psi_{12}(z_1, z_2)}^{\text{exact potential}} \times \overbrace{\exp(\sum_{z_3} Q_{34}(z_3) \log \psi_{13}(z_1, z_3)}^{\text{mean } 1-3 \text{ dependence}}$$

$$\times \underbrace{\exp(\sum_{z_4} Q_{34}(z_4) \log \psi_{42}(z_4, z_2)}_{\text{mean } 2-4 \text{ dependence}}$$

Extensions

- General thoughts on accuracy
  - weak dependencies $\Rightarrow$ good accuracy
  - accurate even for large systems with weak interactions (law of large numbers)
  - strong dependencies $\Rightarrow$ poor approximation
- Extensions
  - The components are not necessarily to be disjoint
  - The structure of mean field can be extended by structure learning

Application to Bayesian networks



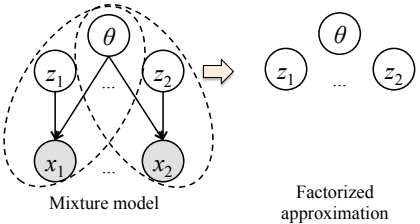
- Mixture model

$$P(x, z, \theta) = P(\theta) \prod_{i=1}^{n} P(z_i) P(x_i | z_i, \theta)$$

- Factorized approximation

$$Q(z, \theta) = Q(z) Q(\theta) = Q(\theta) \prod_{i=1}^{n} Q_i(z_i)$$

Bayesian mixture model: updates



1. Variational posterior over mixing variables

$$Q_i(z_i) \leftarrow \frac{1}{Z_i} \times P(z_i) \times \overbrace{\exp(\int Q(\theta) \log P(x_i|z_i,\theta) d\theta)}^{\text{mean } \theta - z_i \text{ dependencies}}$$

2. Variational posterior over parameters

$$Q(\theta) \leftarrow \frac{1}{Z} \times P(\theta) \times \prod_{i=1}^{n} \overbrace{\exp(\sum_{z_i} Q_i(z_i) \log P(x_i|z_i,\theta))}^{\text{mean } \theta - z_i \text{ dependencies}}$$

Bayesian mixture model: discussion

- Imposing any factorization of the form

$$Q(z, \theta) = Q(z)Q(\theta)$$

  makes the solution liable to sudden deterioration of the component marginals
- Useful factorization "decouples" dependent variables as in

$$P(x_i|x_{pa_i}) \sim \prod_{j \in pa_i} \psi_{ij}(x_i, x_j)$$

- Question: when can we impose such factorization in a principle way?

Imposing additional factorization

Theorem

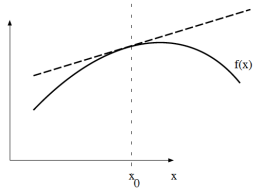
(Factorization Theory.) Assuming x_i and all its parents x_{pa_i} are discrete variables, then we can always find a variational transform of the form

$$P(x_i|x_{pa_i}) = \min_{\epsilon} \left(\prod_{j \in pa_i} \psi_{ij}(x_i, x_j; \epsilon) \right)$$

(analogous formulation holds for "max")

- However, such transform is insufficient; we need transforms that lead to accurate approximations, which is much harder to find

Factorization via convex duality



- Basic idea: for a concave function $f(x)$, any tangent plane serves as a upper bound

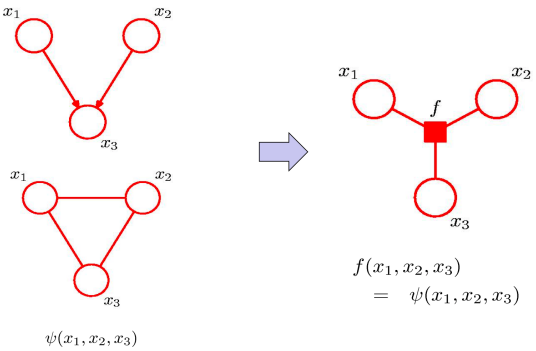
$$f(x) \leq \lambda x - f^*(\lambda)$$

- The bound is exact for one of the tangents

$$f(x) = \min_{\lambda} \lambda x - f^*(\lambda)$$

(the conjugate function $f^*(\lambda)$, also concave, can be similarly expressed in terms of $f(x)$, hence the duality)

Factorization



- transformation from (un)directed graphical models to factor graph model

Contents

1 Basic Monte Carlo Sampling Methods

2 Markov Chain Monte Carlo (MCMC)

3 Mean-field Variational Inference

4 Conclusion

Conclusion

- Basic Monte Carlo sampling methods:
  - Ancestral Sampling
  - Transformation Method
  - Rejection Sampling
  - Importance Sampling
- MCMC sampling methods
  - Metropolis-Hasting
  - Gibbs Sampling
- Mean-field approximation
  - Mean-field variational inference
  - Application to undirected graphs
  - Application to Bayesian networks

Thanks.

HP: http://keg.cs.tsinghua.edu.cn/jietang/
Email: jietang@tsinghua.edu.cn