Monte Carlo Methods

Leon Gu

CSD, CMU

Approximate Inference

EM: y-observed variables; x-hidden variables; θ-parameters;

$$\begin{aligned} \text{E-step:} \quad & q(x) = p(x|y, \theta^{t-1}) \\ \text{M-step:} \quad & \theta^t = \arg\max_{\theta} E_{q(x)} \left[\log p(y, x|\theta)\right] \end{aligned}$$

Monte Carlo EM: if the expectation cannot be derived analytically, we can approximate it by use of sampling methods: first draw a set of samples $x_i \overset{\text{i.i.d}}{\sim} p(x|y, \theta^{t-1}), i = 1 \dots N$, then approximate the integral by a finite sum

$$E_{q^t(x)} \left[\log p(y, x|\theta)\right] \approx \frac{1}{N} \sum_{i=1}^{N} \log p(y, x_i|\theta)$$

In M step we maximize this approximated Q-function. This technique is called Monte Carlo EM.

Monte Carlo methods are a set of computational techniques for 1) generating samples from a target distribution; 2) approximating the expectations of some random quantities under this distribution.

Sampling

Sampling is Difficult

If we can write a density function $P(X)$ in an explicit parametric form and if we can evaluate its value at any point $X = x$, why is sampling $P(X)$ still a hard problem?

1. Correct samples from $P(X)$ by definition tend to come from places where $P(X)$ is big;
2. How can we identify those places where $P(X)$ is big *without* evaluating $P(X)$ everywhere?

A Weaker Assumption

We can evaluate $P^*(x)$, a function that is proportional to the density function $P(x)$ up to a normalizing constant, i.e., $P(x) = P^*(x)/Z$).

Uniform Sampling

Suppose that our goal is to compute the expectation $E[R(x)] = \int R(x)dP(x)$ under the distribution $x \sim P(x)$.

1. Sample $\{x_n : i = 1 \dots N\}$ points i.i.d from Uni(X).
2. Evaluate $P^*(x)$ at those points.
3. Estimate the normalizing constant $Z_N = \sum_{n=1}^{N} P^*(x_n)$.
4. Approximate the expectation by,

$$E_P(R(x)) \approx \sum_{n=1}^{N} R(x_n) \frac{P^*(x_n)}{Z_N}$$

Typical set $\mathcal{T}$: a high dimensional distribution is often concentrated in a few small regions in its state space called "typical set". Its volume can be roughly estimated by $\|\mathcal{T}\| \approx 2^{H(X)}$, where $H(X)$ is the entropy of X.

Consider a distribution $P(X)$ in a binary, D dimensional state space. A uniform sample has a chance of $2^H/2^D$ to hit the typical set, therefore we need roughly $O(2^{D-H})$ samples for an accurate estimate of E[R(x)].

Importance Sampling

Importance sampling is a direct generalization of Uniform sampling. Suppose there is a "proposal" density $Q(X)$ from which we can easily generate samples.

1. Sample $\{x_n : i = 1 \dots N\}$ i.i.d from $Q(X)$.
2. Compute $P^*(x) = ZP(X)$ for each sample and evaluate their *importance* $w_n = P^*(x_n)/Q(x_n)$.
3. Compute the normalizing constant $Z_N = \sum_{n=1}^{N} w_n$.
4. Approximate the expectation by

$$E_P(R(x)) \approx \sum_{n=1}^{N} R(x_n) \frac{w_n}{Z_N}$$

Importance sampling is efficient if $Q(X)$ is close enough to $P(X)$.

Bias-Variance Analysis

Importance sampling estimator is *unbiased*, i.e. the convergence to $E_P(R(x))$ is guaranteed. First, note that Z_N is an approximation of Z,

$$\begin{aligned} Z_N &= \sum_{n=1}^{N} w_n = \sum_{n=1}^{N} P^*(x_n)/Q(x_n) \\ &\approx N \int (P^*(x)/Q(x))Q(X)dx \\ &= N \int P^*(x)dx = N \int P(x)Z dx = NZ \end{aligned}$$

$$\begin{aligned} E_P[R(x)] &= \int R(x)P(x)dx = \int \frac{R(x)P(x)}{Q(x)}Q(x)dx \\ &= E_Q[R(x)\frac{P(x)}{Q(x)}] = E_Q[R(x)\frac{P^*(x)}{ZQ(x)}] \\ &\approx \frac{1}{N}\sum_n R(x_n)\frac{P^*(x_n)}{ZQ(x_n)} \\ &\approx \sum_n R(x_n)\frac{P^*(x_n)}{Z_N Q(x_n)} \\ &= \sum_{n=1}^{N} R(x_n)\frac{w_n}{Z_N} \end{aligned}$$

However, the variance could be *substantially large* if $Q(X)$ is *not* close to $P(X)$.

Markov Chain Monte Carlo

Direct Monte Carlo Sampling (Uniform/Importance/Rejection) methods approximate the target density $P(X)$ or the expectation $E_P[R(X)]$ using the samples drawn from some proposal density $Q(X)$. So we face a paradox here: how can we find a simple $Q(X)$ that is close to the complex $P(X)$? Although several remedies are available, in general these methods will fail if the dimension is high or if $P(X)$ is too complex.

Markov Chain Monte Carlo Sampling methods are based on a different strategy: building a sequence of random variables X_t (a Markov chain) whose distribution converges to $P(X)$ as $t \to \infty$.

- Starting point, transition, convergence to $P(X)$
- Convergence rate

Markov Chain

A *Markov chain* is a sequence of discrete random variables $X_0, X_1, \ldots$ (also called a *stochastic process*) which satisfy *Markov property* $\{X_t; t < T\} \perp \{X_s; s > T\} | X_T$, i.e., given the present state, the past and future states are independent.

The behavior of a Markov chain is decided by its *initial distribution* $\mu_0 = p(X_0)$, and its *transition probability matrix* P^t. The ij-th element $p_{ij}^t = P(X_{t+1} = S_i | X_t = S_j)$ defines the probability of X being in state S_i in $t+1$ given that its previous state is S_j.

A Markov chain is called *homogeneous* or *stationary* if the transition probability $P(X_{t+1}|X_t)$ is independent with t.

Irreducible and Aperiodic

We are interesting in studying such chains:

1. they satisfy some *specific properties* that lead to useful results;
2. and these properties must be as *general* as possible to be holden in most of real-world applications.

A Markov chain is called *irreducible* if for any two states S_i and S_j, starting from any one of them, the other state is accessible by the chain within finite many steps. S_j is *accessible* from S_i (written as $S_i \to S_j$) means $\exists n < \infty, P(X_n = S_j | X_0 = S_i) > 0$. We say the state S_i is *self-accessible* if $\exists n < \infty, P(X_n = S_i | X_0 = S_i) > 0$, and n is called *return time*.

We define the *period* of a state S_i as the greatest common divisor (gcd) of its all possible return times $d(S_i) = \gcd\{n \geq 1 : (P^n)_{i,i} > 0\}$. If $d(S_i) = 1$ then we say the state is *aperiodic*. A Markov chain is called *aperiodic* if all its states are aperiodic.

Convergence of Markov Chain

An irreducible and aperiodic Markov chain converges to an unique distribution $P(X_t) \rightarrow \pi(X)$, as $t \rightarrow \infty$. This distribution $\pi(X)$ is called *stationary distribution*, because $\pi(X_{t+1}) = \sum_{X_t} P(X_t, X_{t+1})\pi(X_t)$, or equally, $\pi = P\pi$.

A distribution $\pi(X), X \in S = \{S_k | k = 1...N\}$ is said to be *reversible* for a Markov chain $\{X_t, \mu_0, P\}$, if for any states $S_i, S_j \in S$, the probability mass "flowing" from S_i to S_j is same as that of the inverse. That is called the *detailed balance condition*,

$$\pi(S_i)P_{ij} = \pi(S_j)P_{ji}$$

If there exists a reversible distribution for a Markov chain, then it is also the *unique* stationary distribution of the chain. This is true because

$$\begin{aligned} \sum_{X_t} \pi(X_t)P(X_t, X_{t+1}) &= \sum_{X_t} \pi(X_{t+1})P(X_{t+1}, X_t) \\ &= \pi(X_{t+1}) \sum_{X_t} P(X_{t+1}, X_t) = \pi(X_{t+1}) \end{aligned}$$

Convergence Rate

If a markov chain converges, the absolute values of all eigenvalues of the transition probability matrix P must be less than or equal to one.

1. $\lambda = 1$: the subspace spanned by these eigenvectors contains all stationary distributions. When the chain is aperiodic and irreducible, there is only one stationary distribution.
2. $\|\lambda\| = 1$: this type of eigenvectors exists only if the Markov chain is periodic.
3. $\|\lambda\| < 1$: all other eigenvectors.

Note the stationary distribution π is an eigenvector of P with eigenvalue $= 1$, since $P\pi = \pi$.

The convergence rate of a Markov chain is controlled by *the second largest eigenvalue* of its transition probability matrix. To see that, we can expand the initial distribution p_0 in the eigen-space of the transition matrix:

$$p_0 = \pi + \lambda_2 v_2 + \lambda_3 v_3 + \dots$$

where the eigenvalues are ordered as "$1 > \|\lambda_2\| \geq \|\lambda_3\| \geq \dots$". After n steps, the distribution becomes:

$$p_n = P^n p_0 = \pi + \lambda_2^n v_2 + \lambda_3^n v_3 + \dots$$

As n increases, p_n will converge to π with the rate determined by $\|\lambda_2\|$.

Gibbs Sampling

Our goal is to sample $P(X)$, where $X = (x_1, \ldots, x_d)^t$. Suppose that $P(x)$ is complex, but the conditionals $P(x_i|\{x_j\}_{j\neq i})$ are tractable and easy to sample. *Gibbs Sampling* is a MCMC method that constructers a Markov Chain $X^0, X^1, \ldots$ with $P(x_i|\{x_j\}_{j\neq i})$ as its one-dimensional transition prabability.

1. Initialize $X^0 = (x_1^0, \ldots, x_d^0)$.
2. Randomly choose one coordinate $i \in [1..d]$.
3. Draw a sample x_i^{t+1} from the conditional $P(x_i|\{x_j\}_{j\neq i})$ and keep the value of all other coordinates ($x_j^{t+1} = x_j^t, \forall j \neq i$). Repeat steps $2 \sim 3$.

$P(X)$ is the reversible distribution of the Markov chain $X^0, X^1, \ldots$ constructed by Gibbs sampler. Suppose we select x_i at time t. Let $\diamond$ denote all other dimensions of X except x_i:

$$\begin{aligned} & P(x_i = S_1, \diamond)P(x_i = S_2, \diamond) = P(x_i = S_2, \diamond)P(x_i = S_1, \diamond) \\ \Leftrightarrow \quad & P(x_i = S_1, \diamond)P(x_i = S_2|\diamond) = P(x_i = S_2, \diamond)P(x_i = S_1|\diamond) \\ \Leftrightarrow \quad & \pi(S_1, \diamond)P_{S_1,\diamond \to S_2,\diamond} = \pi(S_2, \diamond)P_{S_2,\diamond \to S_1,\diamond} \end{aligned}$$

Therefore $P(X)$ is also the stationary distribution of the chain. In other words, $P(X^t)$ converges to $P(X)$ as $t \to \infty$. Gibbs sampler is easy to implement and there is *no tuning parameters*(such as step size in other MCMC methods like Metroplis Hastings sampling).

Gibbs sampler is useful for *sampling graphical models*, because the conditionals are simply specified by the distribution of a node given its Markov blanket.

If we compute a point estimate $\hat{X}_i^t$ that maximizes $P(x_i|\{x_j\}_{j \neq i})$ at each stage, instead of drawing samples, then we have the *Iterated Conditional Modes*(ICM) algorithm.

Random Walk Behaviour

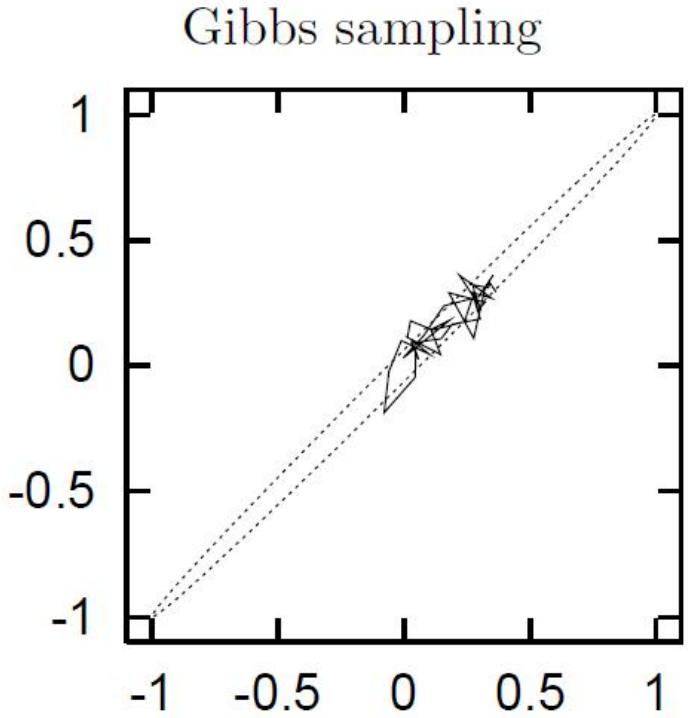
Consider a Markov chain in a integer state space, with initial state $x^0 = 0$ and transition probabilities,

$$\begin{aligned} p(x^{t+1} = x^t) &= 0.5 \\ p(x^{t+1} = x^t + 1) &= 0.25 \\ p(x^{t+1} = x^t - 1) &= 0.25 \end{aligned}$$

By symmetry the expected state at time t will also be zero $E[X^t] = 0$, and $E[(X^t)^2] = Var[X^t] = t/2$. So after t steps, the Markov chain has only travelled a distance that on average is proportional to the square root of t.

This square root dependence is a typical random walk behavior.

Burn-in Time



The successive samples are usually *correlated*. The *burn-in* time is the number of steps to obtain an independent sample as the state evolves. Consider approximating a correlated gaussian of two variables, x_1 and x_2, with a Gibbs sampler. The guassian having marginal distribution of width L and conditional distribution of width l. The typical step size is governed by the conditional distribution so that will be of order l. Since the state evolves according to a random walk, the number of steps needed to obtain independent samples is of order $(L/l)^2$.

An illustrative example

Consider a bivariate Gaussian variable,

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

Recall the formula conditional Gaussian density

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

We can construct a gibbs sampler based on the conditional densities

$$\begin{aligned} x^t|y^t &\sim \mathcal{N}(\rho y^t, (1-\rho^2)) \\ y^{t+1}|x^t &\sim \mathcal{N}(\rho x^t, (1-\rho^2)) \end{aligned}$$

It can be shown

$$\begin{pmatrix} x^t \\ y^t \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \rho^{2t-1}y^0 \\ \rho^{2t}y^0 \end{pmatrix}, \begin{pmatrix} 1-\rho^{4t-2} & \rho-\rho^{4t-1} \\ \rho-\rho^{4t-1} & 1-\rho^{4t} \end{pmatrix} \right\}$$

as $t \to \infty$, $(x^t, y^t)^T$ converges to the target distribution at a rate of ρ^2.

Other Sampling Techniques

- Rejection Sampling
- Metropolis-Hastings Sampling
- Sequential Importance Sampling
- Hybrid Monte Carlo Sampling
- Slice Sampling
- .....