Notes of Sampling

Xiangli Chen

Sampling

Monte Carlo techniques: approximate inference methods based on numerical sampling

- **Sampling from high-dimensional, complicated distributions**
- **Bayesian inference and learning**
  - *Marginalization* $p(x) = \int_Z p(x, z) dz$
  - *Normalization* $p(x|y) = \frac{p(y|x)p(x)}{\int_X p(y|x)p(x)dx}$
  - *Expectation* $\mathbb{E}_{p(x)}(f(x)) = \int_X f(x)p(x)\, dx$
- **Global optimization** $\arg\max_x f(x)$

The Monte Carlo Principle

Estimate: $I(f) = \int_X f(x) p(x) dx$

Idea: Draw i.i.d set of samples $\{x_i\}_{i=1}^N$ from $x \sim p(x)$

Estimator: $I_N(f) = \frac{1}{N} \sum_{i=1}^N f(x_i)$

Theorems:

* $I_N(f) \xrightarrow{P} I(f)$ by the law of large number
* $\mathbb{E}[I_N(f)] = I(f)$ unbiased estimation
* $Var[I_N(f)] = \frac{\sigma_f^2}{N} = O(\frac{1}{N})$ Independent of d
* $\sqrt{N}[I_N - I(f)] \xrightarrow{D} N(0, \sigma_f^2)$ Central Limit Theory

Sampling Using Uniform

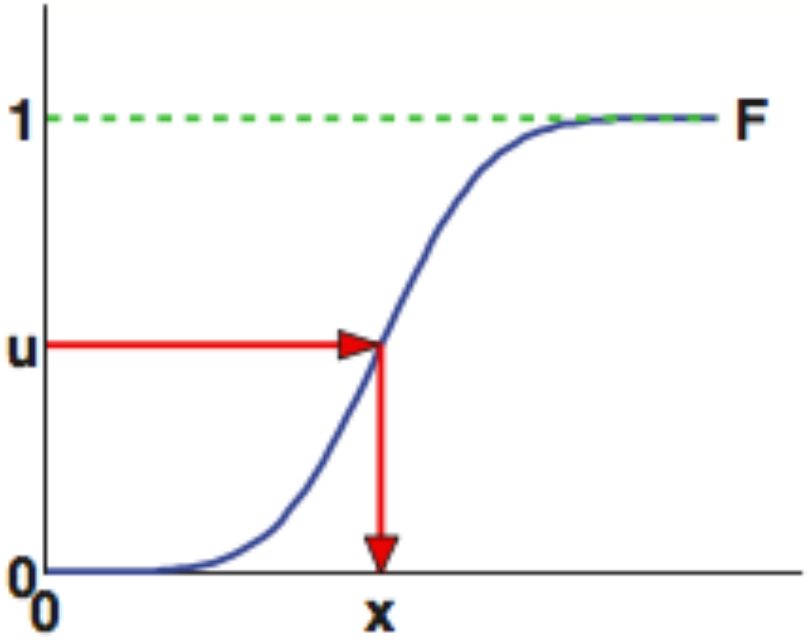
Theorem

Suppose $U \sim \text{Uniform}(0,1)$ let F be a cdf,
then $Z = F^{-1}(U) \sim F_X(x)$

Proof: $P[Z \le x] = P[F^{-1}(U) \le x]$

Note that $F(x)$ is monotone increasing

$$P[F^{-1}(U) \le x] = P[U \in F_X(x)] = F_X(x)$$



Example

Assume $X \sim e(\lambda)$ that $f(x) = \lambda e^{-\lambda x} \quad 0 < x < \infty$

Then $F(x) = 1 - e^{-\lambda x}, \quad x > 0$

$F^{-1}(u) = -(1/\lambda)\log(1-u), \quad 0 < u < 1$

so if $u \sim \text{uniform}(0,1)$, $F^{-1}(U) \sim e(\lambda)$

So we can sample x by $F^{-1}(U)$ using $U \sim \text{uniform}(0,1)$

Rejection Sampling

What if the inverse cdf of $p(x)$ can't be in closed form?

Suppose $p(x) = \frac{1}{Z_p}\tilde{p}(x)$, $\tilde{p}(x)$ can be evaluated, Z_p is unknown

suppose it is easy to sample from $q(x)$ and $p(x) \leq Mq(x)$, $M < \infty$

$q(x)$ is known as proposed distribution

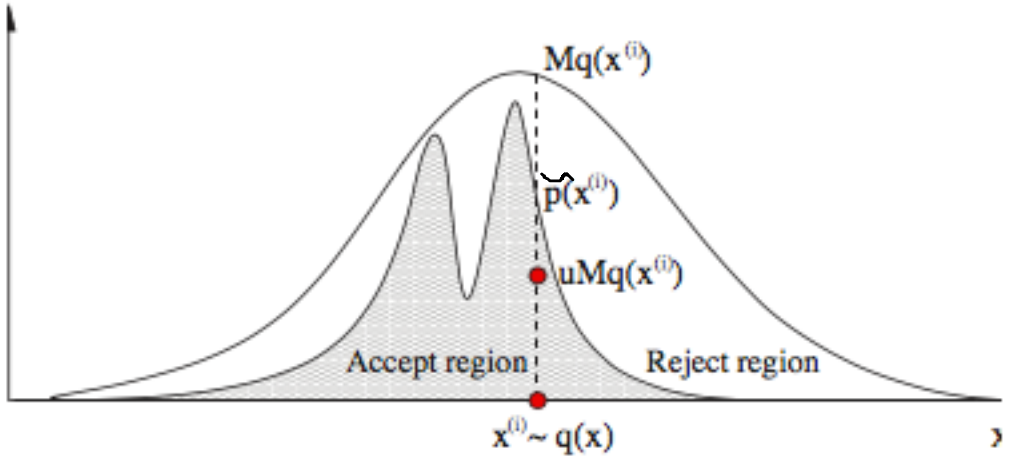
Procedure

Set $i = 1$

Repeat until $i = N$

1. Sample $x_i \sim q(x)$ and $u \sim U(0,1)$
2. If $u Mq(x) < \tilde{p}(x)$ then accept x_i and increase i
   otherwise reject x_i

The accepted x_i's $\sim p(x)$



Proof of Rejection Sampling

$$P(X \le x_0 \mid X \text{ accepted}) = P(X \le x_0, X \text{ accepted}) / P(X \text{ accepted})$$

$$= P(X \le x_0, U < \underbrace{\tilde{p}(x) / (M q(x))}_{u(x)})$$

Note that the joint pdf of X, U is

$$f(x,u) = q(x).$$

$$\text{So } P(X \le x_0, X \text{ accepted}) = \int_{-\infty}^{x_0} \int_0^{u(x)} q(x)\, du\, dx = \int_{-\infty}^{x_0} \tilde{p}(x)/M\, dx$$

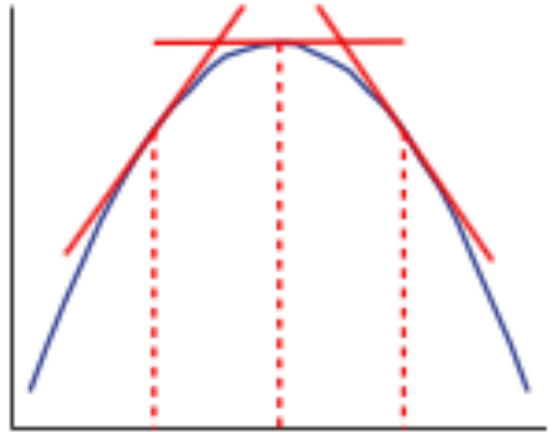
$$\text{So } P(X \le x_0 \mid X \text{ accepted}) = \frac{\int_{-\infty}^{x_0} \tilde{p}(x)\, dx}{\int_{-\infty}^{\infty} \tilde{p}(x)\, dx} = \int_{-\infty}^{x_0} p(x)\, dx = F_X(x_0)$$

Adaptive Rejection Sampling

It is difficult to determine a suitable envelope distribution.

Adaptive rejection sampling: automatically come up with a tight envelope $q(x)$.

As for **Adaptive rejection Metropolis sampling**, log concave is not required.



To be continued

Limitation

Probability of acceptance:

$$p(x \text{ accepted}) = \int_{-\infty}^{\infty} \int_{0}^{Mq(x)} q(x)\,dx = \int_{-\infty}^{\infty} \left(\tilde{p}(x)/Mq(x)\right) q(x)\,dx = \frac{1}{M}\int \tilde{p}(x)\,dx$$

Smaller M is preferred while $Mq(x) \geq \tilde{p}(x)$

Limitation:
If $p(x)$ is multimodal and sharply peaked, it will be extremely difficult to find $q(x)$ and bound $p(x)$ by $Mq(x)$.

The exponential decrease of acceptance rate with dimensionality is a generic feature of rejection sampling.

Importance Sampling - Uniform

Importance sampling provides a framework for approximating expectation directly.

Suppose: simple directly from $p(x)$ is impractical but can evaluate $p(x)$.

Simple strategy using uniform grid method

$$I_N[f] = \sum_{i=1}^{N} p(x_i) f(x_i)$$

Issue: Very low efficiency especially in high dimension —
unable to choose the sample points where $p(x)f(x)$ is large

Use a proposal distribution $q(x)$ where it is easy to draw samples

Importance Sampling - Proposal

Note that

$$I(f) = \int f(x) p(x)\, dx = \int f(x) \frac{p(x)}{q(x)} q(x)\, dx = \int f(x)\, w(x)\, q(x)\, dx$$

$w(x) = p(x)/q(x)$ are the importance weight

Consequently draw N i.i.d x_i from $q(x)$

$$I_N(f) = \frac{1}{N} \sum_{i=1}^{N} f(x_i) w(x_i)$$

Properties

$I_N(f)$ is an unbiased estimator

under weak assumptions, the strong law of large number applies

$$I_N(f) \xrightarrow{P} I(f) = \int f(x) p(x)\, dx$$

Normalizing Constant

Suppose $p(x) = \tilde{p}(x)/Z_p$, $q(x) = \tilde{q}(x)/Z_q$

Then, $E[f] = \frac{Z_q}{Z_p} \int f(x) \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x) dx \approx \frac{Z_q}{Z_p} \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(x_i) f(x_i)$

where $\tilde{w}(x) = \tilde{p}(x)/\tilde{q}(x)$ and $x_i \sim \tilde{q}(x)$

$$Z_p/Z_q = \frac{1}{Z_q} \int \tilde{p}(x) dx = \int \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x) dx \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(x_i)$$

So $\tilde{I}_N(f) = \sum_{i=1}^{N} w(x_i) f(x_i)$ where $w(x_i) \triangleq \frac{\tilde{w}(x_i)}{\sum_j \tilde{w}(x_j)}$

- $\tilde{I}_N(f)$ is biased (ratio of two estimator)
- $\tilde{I}_N(f) \xrightarrow{P} I(f)$ by law of large number

Minimizing Variance

The performance of importance sampling depends on how well $p(x)$ sample from $q(x)$falls into the regions where $p(x)f(x)$is larger.

Some $q(x)$ is much preferable to others.

Minimizing Variance

Find $q(x)$ that minimizes the variance of the estimator

$$Var_{q(x)}[\hat{I}_w(H)] = \frac{1}{N}\sum_{i=1} Var_{q(x)}[f(x_i)w(x_i)] = E_{q(x)}[f^2(x)w^2(x)] - I^2(f)$$

Only need to minimize the first term, which according to Jensen's inequality

$$E_{q(x)}[f^2(x)w^2(x)] \geq (E_{q(x)}[H(x)|w(x)])^2 = (\int |f(x)|p(x)dx)^2$$

Minimizing Variance

The variance is minimal that the lower bound is attained

$$q^*(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx} = |f(x)|p(x)/C$$

- Not very useful since it is hard to sample from $|f(x)|p(x)$
- High sample efficiency if sample from $p(x)$in the region where $|f(x)|p(x)$ is relatively large, hence the name **importance sampling**
- Super efficient – given $f(x)$, there is possible $q(x)$that yields an estimate with a lower variance than using $q(x) = p(x)$
- In high dimension, it is not efficient either

Adaptive Importance Sampling

Motivation: as the dimension of x increases, it becomes hard to Obtain a $q(x)$.

Adaptive importance sampling: adopt a parameterized $q(x, \theta)$

$$\theta^* = \arg\min \mathbb{E}_{q(x,\theta)}[f^2(x) w^2(x,\theta)]$$

Take the derivative with respect to θ: $\partial z / \partial \theta = \int f^2(x) - p(x) \frac{\partial w}{\partial \theta} dx$

Direct compute the Expectation or sample from $p(x)$ is impractical

We rewrite the Expectation using $q(x, \theta)$

$$\partial z / \partial \theta = \mathbb{E}_{q(x,\theta)}[f^2(x) w(x,\theta) \frac{\partial w}{\partial \theta}] dx$$

Updating θ using empirical expectation

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{N} \sum_{i=1}^{N} f^2(x_i) w(x_i, \theta_t) \frac{\partial w(\theta_t, x_i)}{\partial \theta_t}$$ α: learning rate $x_i \sim q(x, \theta)$

Even with adaption, it is often impossible to obtain proposal distributions that are easy to sample and good approximation simultaneously.

Sampling-Importance-Resampling

Recall: rejection sampling requires a constant $M < \infty$ such that $p(x) \le M q(x)$

Sampling-Importance-Resampling: make use of $q(x)$ but avoids specifying M

Procedure

1. N samples x_i's from $q(x)$
2. Construct N importance weight w_i

   Recall $w_i = \frac{\tilde{w}_i}{\sum \tilde{w}_i}$ where $\tilde{w}_i = \tilde{p}(x_i) / \tilde{q}(x_i)$
3. Draw N samples from a discrete distribution with $(w_1, \dots, w_N)$ and values $(x_1, \dots, x_N)$

   This is $\hat{p}(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$

Claim: $\hat{p}(x) \xrightarrow{D} p(x)$

Prof: $\hat{p}(X \le x_0) = \sum_i \mathbb{I}(x_i \le x_0) w_i = \sum_i \mathbb{I}(x_i \le x_0) \tilde{w}_i / \sum_i \tilde{w}_i$

$\xrightarrow{L \to \infty} \int \mathbb{I}(x \le x_0) \tilde{w}(x) q(x) dx / \int \tilde{w}(x) q(x) dx = \int \mathbb{I}(x \le x_0) \tilde{p}(x) dx / \int \tilde{p}(x) dx = \int \mathbb{I}(x \le x_0) p(x) dx = p(X \le x_0)$

Monte Carlo EM Algorithm

E-step of EM algorithm

$$Q(\theta, \theta^{old}) = \int p(z|x, \theta^{old}) \log p(z, x|\theta) \, dz$$

Draw samples z_i's from the current estimate for $p(z|x, \theta^{old})$

$$Q(\theta, \theta^{old}) \simeq \frac{1}{N} \sum_{i=1}^{N} \log p(z_i, x|\theta)$$

A particular instance

Stochastic EM: Consider a finite mixture model, and draw one sample at each E step

Markov Chain Monte Carlo

Rejection sampling and importance sampling suffer from severe limitation particularly in space of high dimensionality

Markov chain Monte Carlo (MCMC)

- allow sampling from a large class of distributions
- scale well with the dimensionality of the sample space

Markov Chain

A first order Markov chain

$$P(X_{t+1} | X_t, \cdots X_1) = P(X_{t+1} | X_t)$$

Homogenous Markov chain

$P(X_{t+1} | X_t)$ is the same for all t

state transition matrix $T_{ij} = P(X_{t+1} = j | X_t = i)$

A regular Markov chain if $T_{ij}^n > 0$ for some n and $\forall i, j$

K step state matrix: $Q_{ij} = P(X_{t+k} = j | X_t = i)$

Lemma $P(X_{t+k} = j | X_t = i) = [T^k]_{ij}$, $\forall (k, i, j)$

Time Reversible Markov Chain

A markov chain T is time reversible if there exists π such that

$\pi_i P_{ij} = \pi_j P_{ji}$ $\forall (i,j)$ known as detailed balanced equations

that is $P(X_t = i) T(X_{t+1} = j \mid X_t = i) = P(X_t = j) T(X_{t+1} = i \mid X_t = j)$

Theorem: If a Markov Chain is regular and satisfies detailed balance with respect to π, then π is a stationary distribution of the chain.

Irreducibility, Aperiodicity and Ergodicity

- **Irreducibility:** For each pair of states (i, j), there is a positive probability, starting in i, that the process will ever enter state j. Equivalent to the transition matrix cannot be reduced to separate smaller matrices. Equivalent to transition graph is connected
- **Aperiodicity:** The chain cannot get trapped in cycles. In other words, a state i is aperiodic if there exists n such that for all $\mathrm{m} \geq n$,

$$\Pr(x_m = i | x_0 = i) > 0$$

  A Markov chain is aperiodic if every state is aperiodic.
- **Ergodicity:** A state is ergodic if it is aperiodic ,recurrent and non-null. A chain is ergodic if all its states are ergodic.

Limiting Distribution

A chain has a limiting distribution if

$$\lim_{t\to\infty} T_{ij}^{t} = \pi_j \text{ exists and is independent of } i, \text{ for all } j$$

If this hold

$$\lim_{t\to\infty} \pi_0 T^{t} = \pi \text{ for all } \pi_0 \text{ independent of the starting state}$$

A limiting distribution has to be stationary

Proof: suppose not

$$\delta = \min_{\pi_i} (|\pi - \pi T|)$$

Let $\varepsilon \ll \delta$ and $\varepsilon > \max_{\pi_i}(|\pi - \pi T^{N})$ for N large enough. $\pi T^{N} \cdot T \simeq \pi T$

But by the property of limit $\varepsilon > \max_{\pi_i}(|\pi - \pi T^{N} \cdot T|)$ leads to contradiction

Limiting Distribution

Theorem1: Every irreducible, aperiodic finite state Markov Chain, has a limiting distribution, which is its unique stationary distribution π.

Theorem2: Every irreducible, ergodic Markov Chain, has a limiting distribution, which is its unique stationary distribution π.
For irreducible finite-state chains, all states are recurrent and non-null

Markov Chain Monte Carlo

MCMC for sampling $p(x)$ is motivated to find a Markov Chain such that $p(x)$ is the limiting distribution.
Given significant larger N, we can draw samples from distribution close to or exactly $p(x)$.

The Hastings-Metropolis Algorithm

Suppose we can sample j from a proposal distribution $q(j|i)$ given i.

Let $X_0, X_1, \ldots, X_n, \ldots$ Markov Chain defined as follows:

1. from state i goto state j with prob. $q(j|i)$
2. $\begin{cases} \text{with prob. } 1-\alpha(i,j) & \text{go back to state } i, \\ \text{with prob. } \alpha(i,j) & \text{stay in state } j \end{cases}$

Thm $P(X_{n+1}=j \mid X_n=i) = q(j|i)\alpha(i,j) \quad \forall j \neq i$

$$P(X_{n+1}=i \mid X_n=i) = q(i|i) + \sum_{k \neq i} q(i,k)(1-\alpha(i,k))$$

The Hastings-Metropolis Algorithm

Thm

If $\alpha(i,j)=\min\left(\frac{\pi_j q(i|j)}{\pi_i q(j|i)},1\right)=\min\left(\frac{\tilde{\pi}_j q(i|j)}{\tilde{\pi}_i q(j|i)},1\right)$, $\pi_i=\frac{\tilde{\pi}_i}{\sum_i \tilde{\pi}_i}$

The markov chain is time reversible that

π satisfies the detailed balanced equations

$\forall j\neq i \quad \pi_i q(j|i)\alpha(i,j)=\pi_j q(i|j)\alpha(j,i)$

Prof: If $\alpha(i,j)=\frac{\pi_j q(i|j)}{\pi_i q(j|i)} \Leftrightarrow \alpha(j,i)=1$

So if T is regular, π is the stationary distribution

if T is regular, irreducible, ergodic,

π is the unique stationary distribution that it is the limiting distribution

The Hastings-Metropolis Algorithm

1. Initialize $x_0 = k$

2. For $i = 0$ to $N-1$

   sample $u \sim U(0,1)$

   sample $x^* \sim q(x^* | x_i)$

   If $u < \alpha(x_i, x^*)$ $x_{n+1} = x^*$

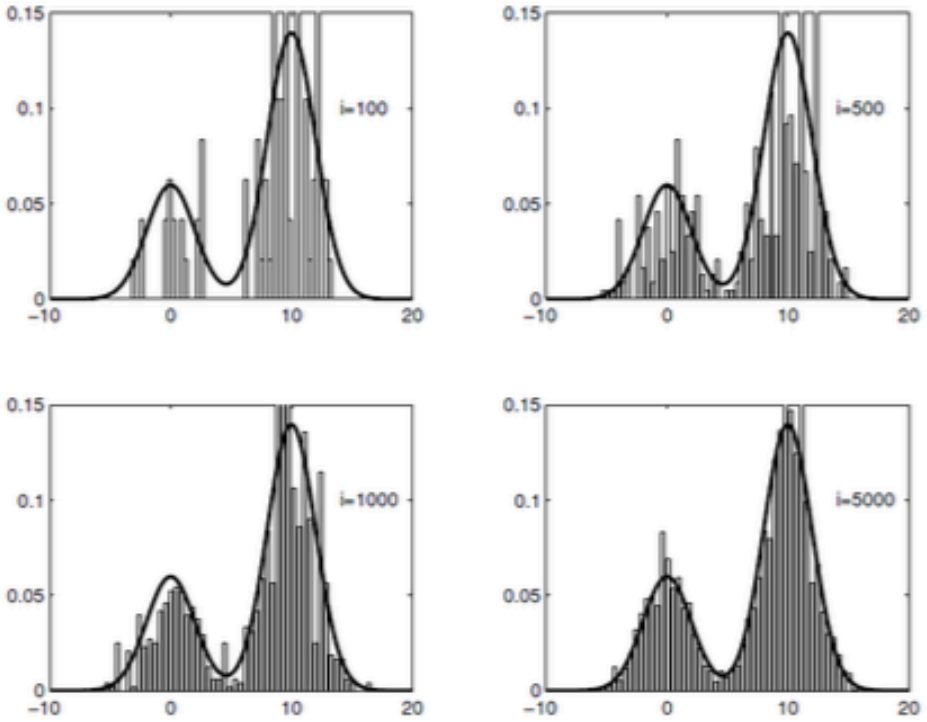
   else $x_{n+1} = x_n$

we use all samples $x_0, x_1, \cdots, x_n, \cdots,$

Proposal Distribution

For continuous state space, a Gaussian distribution is commonly used.

$q(x^*|x_i) = N(x^*|x_i, \varepsilon)$ known as a random walk Metropolis algorithm



Bimodal target distribution: $p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

$q(x \mid x^{(i)}) = N(x^{(i)}, 100)$, 5000 iterations

Proposal Distribution

Set variance of proposal distribution is important

- Variance is too low, only explore one of the modes
- Variance is too large, reject rate can be vary high, result in high correlation

Benefits of Gibbs sampling:
Doesn’t need to choose the proposed distribution. The acceptance rate is 100%

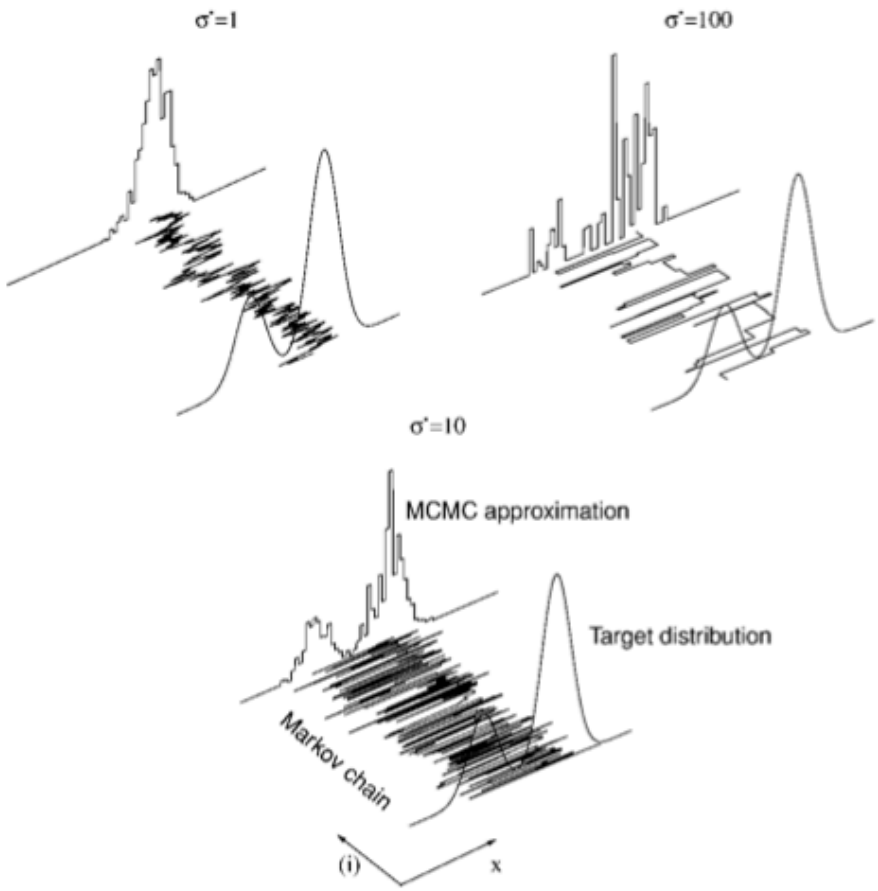


Figure 7. Approximations obtained using the MH algorithm with three Gaussian proposal distributions of different variances.

Convergence of MCMC

Ignore samples before the claim has reached its stationary distribution

Mixing time: amount of time taking for convergence

Mixing time from state x_0, for any constant $\varepsilon > 0$,

$$T_\varepsilon(x_0) \triangleq \min\{t : \|\delta_{x_0}(x) T^t - \pi\|_1 \le \varepsilon\}$$

$\delta_{x_0}(x)$: distribution with all its mass in state x_0

Mixing time of the chain $T_\varepsilon \triangleq \max_{x_0} T_\varepsilon(x_0)$

Theorem $\pi T = \pi$, π is left eigenvector of T with eigenvalue 1

Perron–Frobenius theorem shows $1 = \lambda_1 > \lambda_2 \ge \lambda_3 \cdots$

In particular, one can show $T_\varepsilon \le O\left(\frac{1}{1-\lambda_2} \log \frac{n}{\varepsilon}\right)$

So we want λ_2 as small as possible

Practical Diagnostics of Convergence

Trace plot: assess when the method has converged running multiple chains from very different over dispersed starting point.
If the chain has mixed, the trace plots should converge to the same distribution, thus overlap with each other.

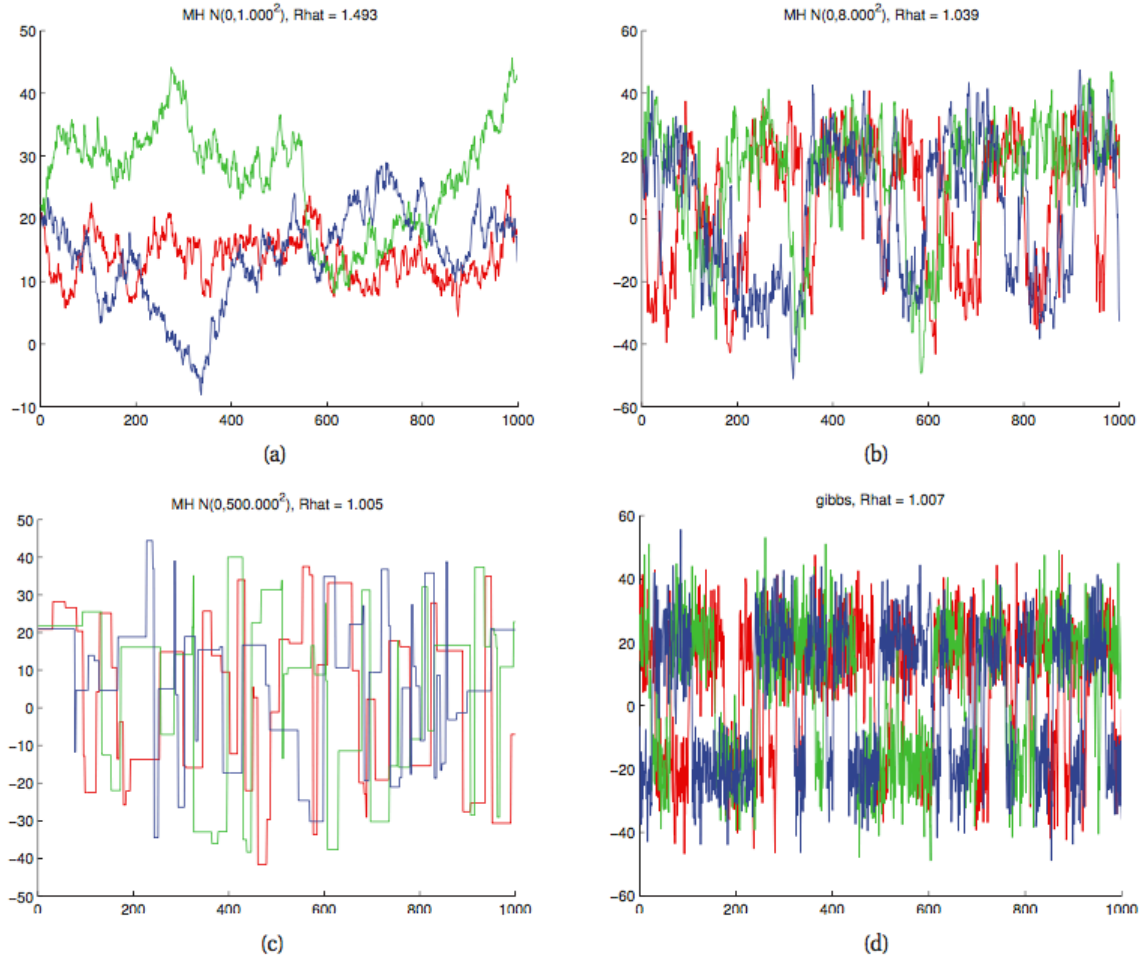


Figure 24.12 Traceplots for MCMC samplers. Each color represents the samples from a different starting point. (a-c) MH with proposal $\mathcal{N}(x'|x, \sigma^2)$ for $\sigma^2 \in \{1, 8, 500\}$, corresponding to Figure 24.7. (d) Gibbs sampling. Figure generated by `mcmcGmmDemo`.

Accuracy of MCMC

The samples produced by MCMC are auto-correlated instead of i.i.d.

Suppose $X \sim p(x)$

$f^* \triangleq \mathbb{E}[f(X)]$ $\bar{f} = \frac{1}{N}\sum_{n=1}^{N} f_n$

$$\mathrm{Var}_{\mathrm{MCMC}}[\bar{f}] = \mathbb{E}[(\bar{f} - f^*)^2] = \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N}(f_n - f^*)^2\right] + \frac{1}{N^2}\sum_{n \neq m}\mathbb{E}[(f_n - f^*)(f_n - f^*)]$$

$$= \mathrm{Var}_{\mathrm{MC}}(\bar{f}) + \underbrace{\frac{1}{N^2}\sum_{n \neq m}\mathbb{E}[(f_n - f^*)(f_n - f^*)]}_{\text{This is zero if samples are i.i.d}}$$

Reduce the auto-correlation using thinning, in which we keep every $n'th$ sample.

In practice, it is common to run a medium number of chains of medium length, and take samples from each after discarding the first half of the samples.

Gibbs Sampling

Suppose we want to sample $p(x) = p(x_1, \cdots, x_2)$

Gibbs sampling procedure involves replacing the value of the variables by a value draw from the distribution of that variable conditioned on the values of the remaining variables.

1. Initialise $x_{0,1:n}$.
2. For $i = 0$ to $N - 1$
   - Sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
   - Sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

     $\vdots$
   - Sample $x_j^{(i+1)} \sim p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

     $\vdots$
   - Sample $x_n^{(i+1)} \sim p(x_n | x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

The algorithm above generates $N \times n$ samples.

Gibbs Sampling

A special case of Metropolis Hastings algorithm

The proposal distribution of Gibbs Sampling

For $j=1,\dots,n$, $x_{-j}=(x_1,\cdots x_{j-1},x_{j+1},\cdots,x_n)$

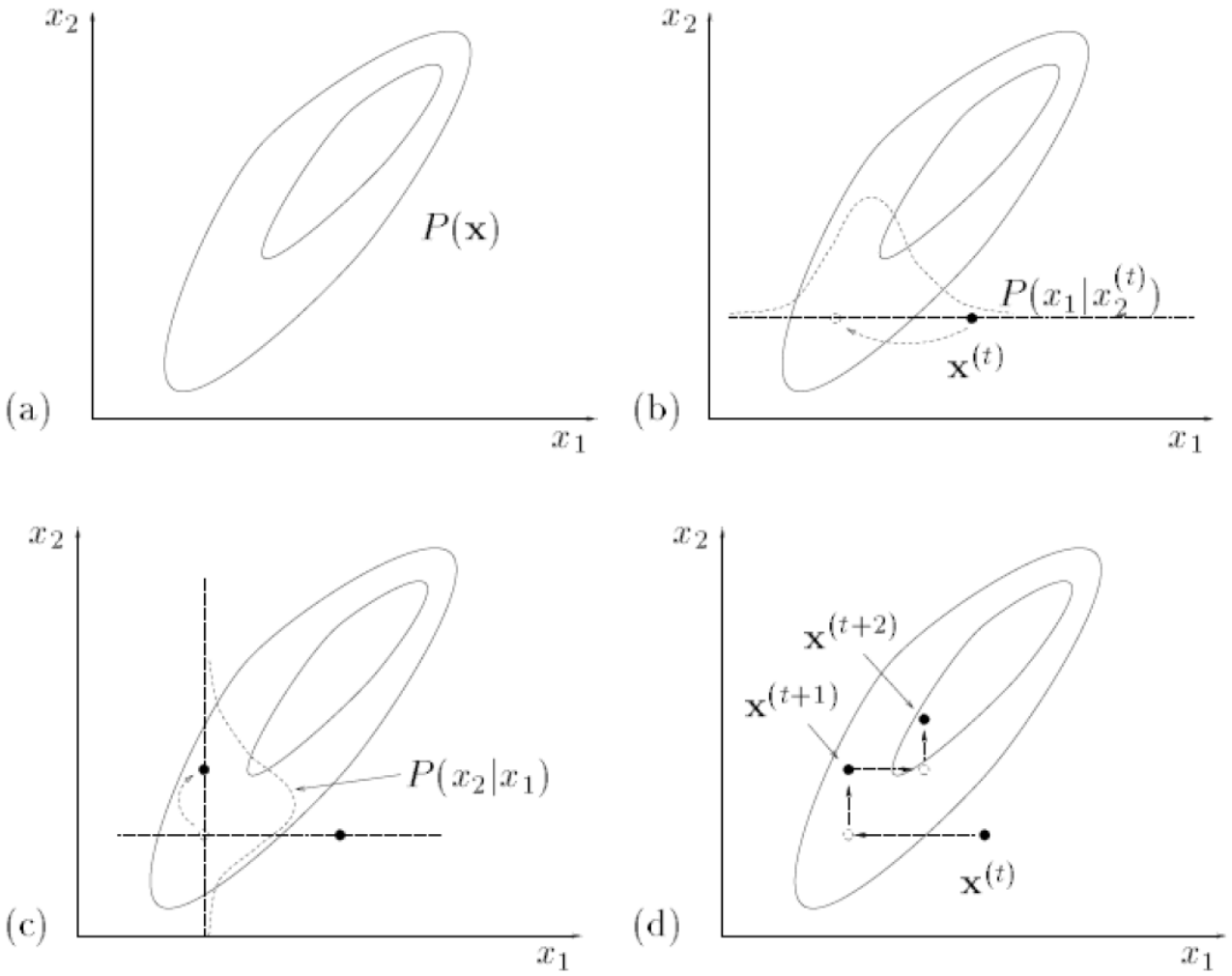
$$q(x^*|x)=\begin{cases}p(x_j^*|x_{-j}) & \text{If } x_{-j}^*=x_{-j}\\ 0 & \text{otherwise}\end{cases}$$

The acceptance probability

$$\alpha(x,x^*)=\min\left(1,\frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}\right)=\min\left(1,\frac{p(x^*)p(x_j|x_{-j}^*)}{p(x)p(x_j^*|x_{-j})}\right)$$

$$=\min\left(1,\frac{p(x^*)p(x_j|x_{-j})}{p(x)p(x_j^*|x_{-j}^*)}\right)=\min\left(1,\frac{p(x_{-j}^*)}{p(x_{-j})}\right)=1$$

Gibbs Sampling

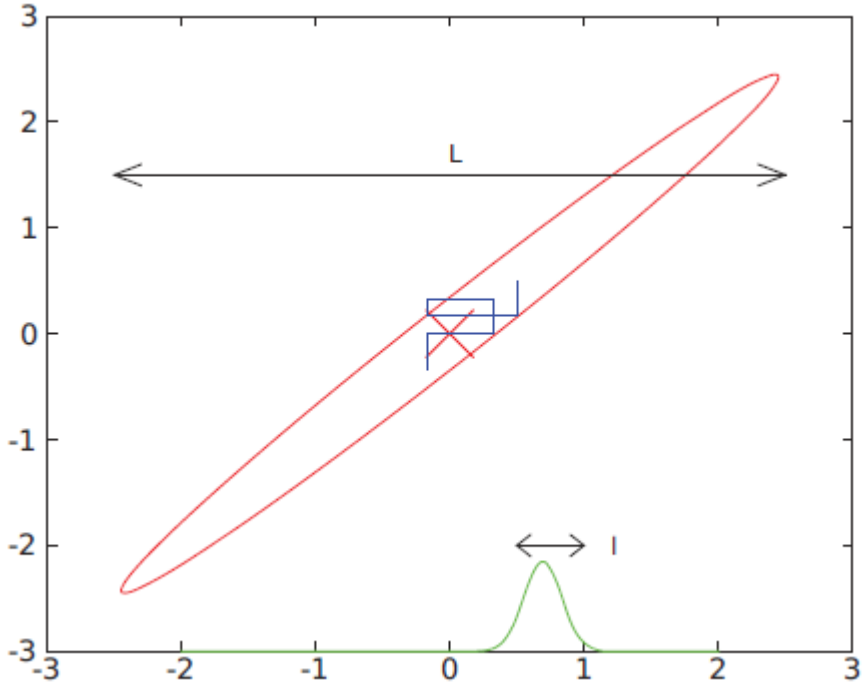


Independence sample of Gibbs Sampling

Correlated Gaussian in two variables
Having conditional distribution of width l
Having marginal distribution of width L
Number of steps needed to obtain independent samples

$$O((L/l)^2)$$

Consider over-relation



Slice Sampling

Hybrid Monte Carlo

Simulated Annealing

Let $\mathcal{A}$ be a huge finite set of vectors.

Let $V : \mathcal{A} \to \mathbb{R}_+$.

V can have lots of local maximum points.

Goal: Find $V^* = \max_{x \in \mathcal{A}} V(x)$

Let $\mathcal{M} = \{x \in \mathcal{A} : V(\mathbf{x}) = V^*\}$, set of maximum points

Limit Distribution

Let $\lambda > 0$ $\quad P_\lambda(\mathbf{x}) = \dfrac{\exp(\lambda V(x))}{\sum_{\mathbf{x}\in\mathcal{A}} \exp(\lambda V(x))}$

Theorem: $P_\lambda(x) \xrightarrow{\lambda\to\infty} \dfrac{\delta(x,\mathcal{M})}{|\mathcal{M}|}$

where $\delta(x,\mathcal{M}) = 1$, if $x \in \mathcal{M}$, and 0 otherwise.

Proof: $P_\lambda(x) = \dfrac{\exp(\lambda(V(x)-V^*))}{|\mathcal{M}| + \sum_{\substack{x\in\mathcal{A}\\ x\notin\mathcal{M}}} \exp(\lambda(V(x)-V^*))}$

$(V(x)-V^*) \le 0,\ \forall x$

If $(V(x)-V^*) < 0$, then $\exp(\lambda(V(x)-V^*)) \xrightarrow{\lambda\to\infty} = 0$

Main idea

- Let λ be big.
- Generate a Markov chain with limit distribution $P_\lambda(x)$.
- In long run, the Markov chain will jump among the maximum points of $P_\lambda(x)$.

Hasting – Metropolis Sampling

Introduce the relationship of **neighboring vectors:**

For example, let $\mathbf{x} \in \mathcal{A}$, and $\mathbf{y} \in \mathcal{A}$ be neighbors, if they only differ in one coordinate.

Let $N(\mathbf{x})$ be the set of neighbors of $\mathbf{x}$

Let $q(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x}) = \frac{1}{|N(\mathbf{x})|}$ Uniform distribution

We want $\pi(\mathbf{x}) = \exp(\lambda V(\mathbf{x}))$ limit distribution.

Use the Hastings- Metropolis sampling:

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1\right)$$

$$= \min\left(1, \frac{\exp(\lambda V(\mathbf{y}))|N(\mathbf{y})|}{\exp(\lambda V(\mathbf{x}))|N(\mathbf{x})|}\right)$$

Hasting – Metropolis Sampling

1) At iteration t we are in $\mathbf{x}(t) \in \mathcal{A}$
2) Let $\mathbf{z} \in \mathcal{A}$ be a neighbor of $\mathbf{x}(t)$.

   drawn from uniform distribution $(\frac{1}{|N(x(t))|})$.

3) Let $u \sim U_{[0,1]}$
4) If $u < \alpha(\mathbf{x}(t), \mathbf{z}) \Rightarrow \mathbf{x}(t+1) = \mathbf{z}$

   With prob. α accept the new state

   If $u \geq \alpha(\mathbf{x}(t), \mathbf{z}) \Rightarrow \mathbf{x}(t+1) = \mathbf{x}(t)$

   with prob. (1-α) don't accept and stay

5) Back to 2

Hasting – Metropolis Sampling

$$\text{If } |N(\mathbf{z})| = |N(\mathbf{x})| \ \forall \mathbf{x}, \mathbf{z} \in \mathcal{A}$$
$$\Rightarrow \alpha(\mathbf{x}, \mathbf{z}) = \min\left\{1, \exp(\lambda(V(\mathbf{z}) - V(\mathbf{x}))\right\}$$

In this special case:

4)

$$\text{If } V(\mathbf{z}) \geq V(\mathbf{x}(t)) \Rightarrow \mathbf{x}(t+1) = \mathbf{z}$$

With prob. α=1 accept the new state since we increased V

$$\text{If } V(\mathbf{z}) < V(\mathbf{x}(t)) \Rightarrow \begin{cases} \exp\{\lambda[V(\mathbf{z}) - V(\mathbf{x}(t))]\} = \alpha < 1 \\ \text{with prob. } 1-\alpha\text{: } \mathbf{x}(t+1) = \mathbf{x}(t) \\ \text{with prob. } \alpha\text{: accept } \mathbf{z}, \ \mathbf{x}(t+1) = \mathbf{z} \end{cases}$$

Problems

- If $V(\mathbf{z}) < V(\mathbf{x}(t))$, then the probability to move to $\mathbf{z}$ is exp small.
- If λ is big and $\mathbf{x}(t)$ is a local maximum, then it might take a looong time to get to a new $\mathbf{z}$ from $\mathbf{x}(t)$.
- Nonetheless, we need big λ to find the set $\mathcal{M}$.
- Solution: Increase λ slowly, e.g. $\lambda_t = c\log(1+t)$, $c > 0$