Metropolis Hastings Algorithm

1. Choose any proposal distribution for transferring from x to x' $T^Q(x \to x')$
2. Use T^Q to propose a transition from x to x'. We accept the proposal with probability $A(x \to x')$ and transition, or stay in x.

$$T(x \to x') = T^Q(x \to x') A(x \to x') \quad x \neq x'$$
$$T(x \to x) = T^Q(x \to x) + \sum_{x' \neq x} T^Q(x \to x')(1 - A(x \to x'))$$

How to design A?

Reversible Chains

Definition: A finite state Markov chain T is reversible if $\exists$ a unique π such that $\forall x, x' \in \chi$

$\pi(x')T(x' \to x) = \pi(x)T(x \to x')$
Above is called the Detailed balance Equation (DBE)

Theorem

If $\pi(x)$ satisfies above then $\pi(x)$ is a stationary distribution of T.

Proof-

$\sum_{x'} \pi(x')T(x' \to x) = \pi(x) \sum_{x'} T(x \to x') = \pi(x)$

Example: Reversibility check for Gibbs

To show that: $P(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) = P(\mathbf{x}')T(\mathbf{x}|\mathbf{x}'), \quad \mathbf{x} \neq \mathbf{x}'$
Proof: $\mathbf{x}$ and $\mathbf{x}'$ can only differ in one position. Let that be i. Then : $\mathbf{x}' = x_i', \mathbf{x}_{-i}$

$$
\begin{aligned}
P(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) &= \frac{1}{n} P(\mathbf{x}) P(x_i'|\mathbf{x}_{-i}) \\
&= \frac{1}{n} P(\mathbf{x}) \frac{P(x_i', \mathbf{x}_{-i})}{\sum_{x_i'} P(x_i', \mathbf{x}_{-i})} \\
&= \frac{1}{n} P(\mathbf{x}) \frac{P(\mathbf{x}')}{\sum_{x_i'} P(x_i', \mathbf{x}_{-i})} \\
&= P(\mathbf{x}')T(\mathbf{x}|\mathbf{x}')
\end{aligned}
$$

This gives us an alternative easier proof that $P(\mathbf{x})$ is a stationary distribution under Gibbs sampling.

Choosing A

Design A to satisfy detailed balance equation for $x \neq x'$

$$\pi(x) T^Q(x \to x') A(x \to x') = \pi(x') T^Q(x' \to x) A(x' \to x)$$

$$A(x \to x') = \min\left[1, \frac{\pi(x') T^Q(x' \to x)}{\pi(x) T^Q(x \to x')}\right] \text{ satisfies this}$$

[Proof easy: Either numerator $\pi(\mathbf{x}') T^Q(\mathbf{x}' \to \mathbf{x})$ is lower or the denominator. Assume numerator. Then $A(x' \to x) = 1$. Plug in these values and DBE will be satisfied.]

Given a desired stationary distribution $P(\mathbf{x})$, designing the $A(\cdot)$ just requires the (user provided) T^Q and the ratio of probabilities $\frac{P(\mathbf{x}')}{P(\mathbf{x})}$.

Example from book

Let us we desire a stationary distribution:
$\pi = [\pi_1, \pi_2, \pi_3] = [2/Z, 3/Z, 2/Z]$
Earlier we had started with a T that gave this π. Not we choose an arbitrary T^Q and compute A.

1. Example $T(x \rightarrow x') = 1/3$
2. Compute $A(1 \rightarrow 2) = \min[1, \frac{\pi(2)T^Q(1|2)}{\pi(1)T^Q(2|1)}] = \min(1, 3/2) = 1$
3. Compute $A(2 \rightarrow 3) = 2/3$

Langevin Monte-Carlo

(https://en.wikipedia.org/wiki/Metropolis-adjusted_Langevin_algorithm)
Sampling from an arbitrary differentiable (in x) function eg. Neural network representing P($\mathbf{x}$) [eg: audio,images]

$P(\mathbf{x}) \propto \exp(-E_\theta(\mathbf{x}))$ $\quad P(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z}$ where $E_\theta(\mathbf{x}) \mapsto R$ is an arbitrary differentiable function in $\mathbf{x}$ and Z_θ is intractable to compute
But given two $\mathbf{x}$ and $\mathbf{x}'$ easy to compute the ratio of their probabilities as $\frac{P(\mathbf{x})}{P(\mathbf{x}')} = \frac{\exp(-E_\theta(\mathbf{x})}{\exp(-E_\theta(\mathbf{x}')}$.

How to design $T^Q(\mathbf{x}'|\mathbf{x})$?
Intuition: transition from any $\mathbf{x}$ to another $\mathbf{x}'$ along directions of maximum increase in $\log P(\mathbf{x})$ by using gradients.

$$\mathbf{x}' = \mathbf{x} + \tau \nabla_{\mathbf{x}} \log P(\mathbf{x}) = \mathbf{x} - \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x})$$

Langevin Monte-Carlo: proposal distribution

Making the transitions probabilistic by adding a small Guassian noise.

$$\mathbf{x}' = \mathbf{x} - \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x}) + \sqrt{2\tau}\xi, \qquad \xi \sim N\left(0, I_d\right)$$

$$T^Q\left(\mathbf{x} \to \mathbf{x}'\right) = \frac{1}{\sqrt{2pi}2\tau} \exp\left(-\frac{1}{4\tau}\left\|\mathbf{x}' - \mathbf{x} + \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x})\right\|^2\right)$$

$$A\left(\mathbf{x} \to \mathbf{x}'\right) = \min\left\{1, \frac{e^{-E_\theta(\mathbf{x}')} T^Q\left(\mathbf{x}' \to \mathbf{x}\right)}{e^{-E_\theta(\mathbf{x})} T^Q\left(\mathbf{x} \to \mathbf{x}'\right)}\right\}$$

Algorithm for Langevin Monte-Carlo

Input: $E_\theta(\mathbf{x})$, Initial sample $\mathbf{x}^1$.

1. For t $= 1 \ldots M$
   $\mathbf{x} = \mathbf{x}^t$
   Compute gradient $\nabla_{\mathbf{x}} E_\theta(\mathbf{x})$
   Sample a $\xi \sim N(0, I)$
   Compute $\mathbf{x}' = \mathbf{x} - \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x}) + \sqrt{2\tau}\xi$
   Compute $T^Q(\mathbf{x} \rightarrow \mathbf{x}') \propto \exp\left(-\frac{1}{4\tau}\left\|\mathbf{x}' - \mathbf{x} + \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x})\right\|^2\right)$
   Compute gradient $\nabla_{\mathbf{x}} E_\theta(\mathbf{x}')$
   Compute $T^Q(\mathbf{x}' \rightarrow \mathbf{x}) \propto \exp\left(-\frac{1}{4\tau}\left\|\mathbf{x} - \mathbf{x}' + \tau \nabla_{\mathbf{x}} E_\theta(\mathbf{x}')\right\|^2\right)$
   $A = \min\left\{1, \frac{e^{-E_\theta(\mathbf{x}')} T^Q(\mathbf{x}' \rightarrow \mathbf{x})}{e^{-E_\theta(\mathbf{x})} T^Q(\mathbf{x} \rightarrow \mathbf{x}')}\right\}$
   Sample a $u \sim U(0, 1)$
   $\mathbf{x}^{t+1} = \mathbf{x}^t \quad \text{if} \quad u \geq A, \quad \text{else} \quad \mathbf{x}'$
2. Return $x^1, \ldots, x^M$ as samples

Applications of Langevin Monte Carlo

Training better generators for high-dimensional objects using deep networks via energy-based networks.