Quantifying Uncertainty

Sai Ravela

M. I. T

Last Updated: Spring 2013

Markov Chain Monte Carlo

- Monte Carlo sampling made for large scale problems via Markov Chains
  - Monte Carlo Sampling
  - Rejection Sampling
  - Importance Sampling
  - Metropolis Hastings
  - Gibbs
- Useful for MAP and MLE problems

MONTE CARLO

Example:

$$P(x) \sim 0.5 \frac{1}{\sqrt{2\pi}} \left\{ e^{-x^2/2} + e^{-(x-2)^2/2} \right\}$$

Calculate $\int (x^2 + cosh(x)) P(x) dx$ May be difficult!

$$\underbrace{\int f(x)P(x)dx}_{\text{When this becomes intractable}} \cong \underbrace{\frac{1}{S}\sum_{s=1}^{S} f(x_s)}_{\substack{Monte-Carlo \\ \text{Sampling may still be feasible}}} \qquad x_s \sim P(x)$$

Properties of Estimator

$$\hat{I}_s = \frac{1}{S}\sum_{s=1}^{S} f(x_s), \qquad x_s \sim P(x)$$

$$I = \int f(x)P(x)dx$$

$$\lim_{S\to\infty} \hat{I}_s = I \qquad \leftarrow \textit{unbiased}$$

$$\sigma_{\hat{I}} = \frac{\sigma}{\sqrt{S}}$$

From Introduction Class.

What's good about this?

The good

* Quick and “dirty” estimate (sometimes, it's the only way out)
* Sampling is useful per se

What's not good?

* Quick and Dirty!
* Rao-Blackwell
  - $\Rightarrow$ Sample based estimator generally worse

Methods

Basics

Via CDF (random and stratified)

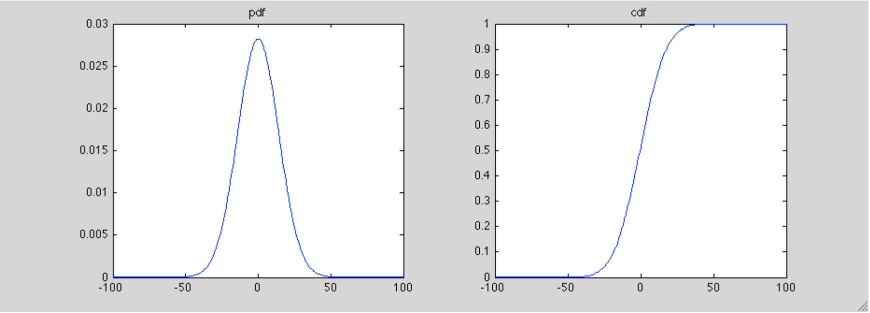
Intermediate

- Importance Sampling
- Rejection Sampling

Objective

- Metropolis
- Metropolis-Hasting
- Gibbs

Sampling from a CDF -Random Sampling



Latin Hypercube Sampling

Stratified Sampling -e.g. Latin hypercube, Orthogonal samplling.

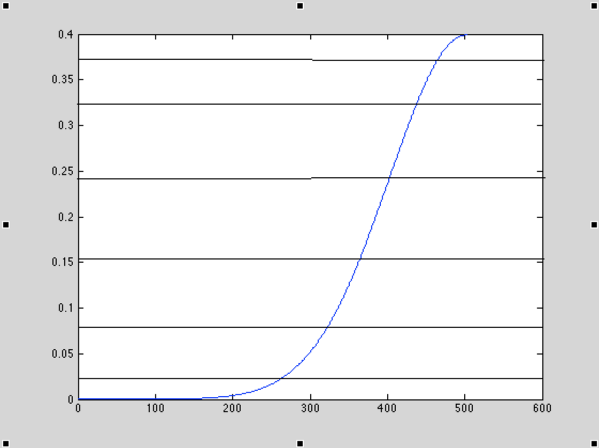
Latin hypercube sampling, motivated by latin squares, the hypercube is in N-D.

- Each row and column have unique selection
- A way to “cover” the square uniformly.

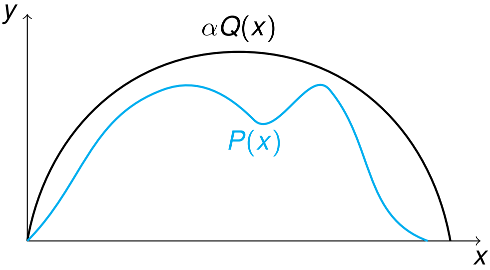
LS example

Photo Credit: Wikipedia

Orthogonal/Stratified Sampling Example



Rejection Sampling



$$\alpha Q(x) \geq P(x)$$

$$x_i \sim Q(x), \quad y_i \sim U[0, \alpha Q(x_i)]$$

If $y_i \leq P(x_i)$ accept
else reject

+ Generates Samples
- Can be very wasteful
- Needs to be upper bound

How to avoid waste?

Importance Sampling

$$\int f(x)P(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx$$
$$\cong \frac{1}{S}\sum_{s=1}^{S} f(x_s)\frac{P(x_s)}{Q(x_s)}, \quad x_S \sim Q(x)$$

$$\frac{P(x_s)}{Q(x_s)} \equiv \text{Importance of sample} \doteq \omega_s$$

$$\hat{I}_S = \frac{1}{S}\sum_{s=1}^{S} f(x_s)\omega_s$$

Unbiased

Works with Potentials

$$I = \int f(x)p(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx$$

Let's write $Z_p = \int \acute{P}(x)dx$ & $Z_q = \int \acute{Q}(x)dx$
and define

$$P(x) = \frac{\acute{P}(x)}{Z_p}$$

$$Q(x) = \frac{\acute{Q}(x)}{Z_q}$$

Here $\acute{P}(x)$ is just un-normalized, i.e. a potential as opposed to a probability we have access to.

Q is still a proposal distribution we constructed.

Contd.

Then,

$$\begin{aligned} I &= \frac{Z_q}{Z_p}\int f(x)\frac{\acute{P}(x)}{\acute{Q}(x)}Q(x)dx \\ &= \frac{Z_q}{Z_p}\int f(x)\acute{\omega}(x)Q(x)dx \\ &\cong \frac{Z_q}{Z_p}\cdot\frac{1}{S}\sum_{s=1}^{S} f(x_s)\acute{\omega}_s; \quad x_s \sim Q(x) \\ &= \frac{Z_q}{Z_p}\cdot\frac{1}{S}\sum_{s=1}^{S} f(x_s)\acute{\omega}_s \end{aligned}$$

we still don't know what to do with Z_q/Z_p!

A simple normalization works

Turns out

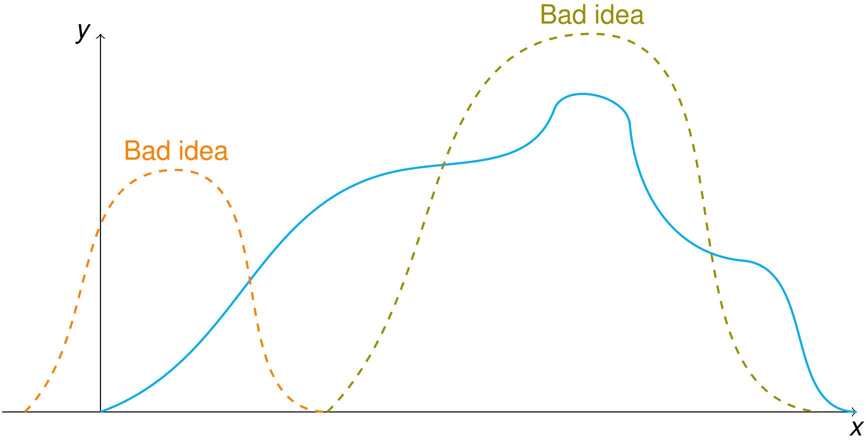
$$\frac{Z_q}{Z_p} = \frac{1}{S}\sum_{s=1}^{S} \acute{\omega}_s$$

So,

$$\hat{I} = \frac{\sum_s f(x_s)\acute{\omega}_s}{\sum_{\acute{s}} \acute{\omega}_{\acute{s}}}$$

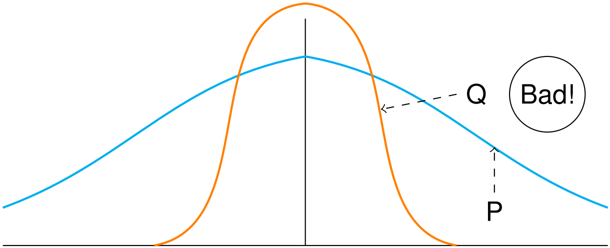
A weighted normalization.
$\rightarrow$Biased

How to select Q ?



More on Q

1. Must generally “cover” the distribution
2. Not lead to undue importance



3. Uniform is OK when P(.) is bounded

What's different

Importance Sampling $\rightarrow$

Does not reject a sample,
just reweights it
May be problematic to carry around
weights during uncertainty propagation

Rejection Sampling $\rightarrow$

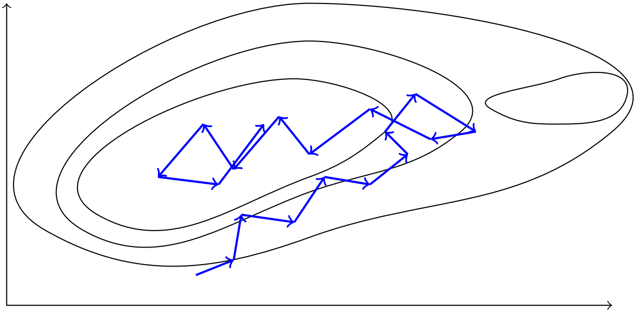
Wastes time (computation)
Produces samples

What's common

- Neither technique scales to high dimension
- Sampling (all Monte Carlo so far) is brute force! (Dumb)

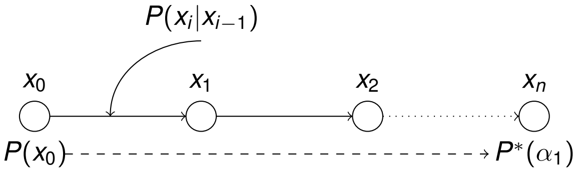
$\rightarrow$ Markov chain Monte Carlo

Markov Chain Monte Carlo



1. A proposal distribution from local moves (not globally, as in RS/IS).
   1.1 Local moves could be in some subspace of state space.
2. Move is conditioned on most recent sample

Primer



Forward Problem: Given Transition end up where?
MCMC: Given target, how to transition?

Transitions, Invariance and Equilibrium

Contruct a transition

$$x_t \sim \underbrace{P_T(x_{t-1}) \to x_t}_{\text{Markov chain}}$$

such that the equilibrium distribution π^* of P_T, defined as:

$$\pi^* \leftarrow P_T^N \pi_0$$

is the invariant distribution, i.e.

$$\pi^* = P_T \pi^*$$

Which implies Condition 1: General balance.

$$\sum_{x'} P_T(x' \to x)\pi^*(x') = \pi^*(x)$$

And, π^* is the target distribution to sample from.

Regularity and Ergodicity

Condition #2 (The whole state space is reachable)

$$P_T^N(x' \rightarrow x) > o \quad \forall x : \pi^*(x) > 0$$

$\Rightarrow$ Ergodicity

Condition 2 says that all states are reachable, i.e. the chain is irreducible. When the states are aperiodic, i.e. transitions don't deterministically return to state i in integer multiples of a period, then chain is ergodic.

Detailed Balance

Condition #3: Detailed Balance

$$P_T(x' \to x)\pi^*(x') = P_T(x \to x')\pi^*(x)$$

$$\Rightarrow \sum_{x'} P_T(x' \to x)\pi^*(x') = \pi^*(x) \underbrace{\sum_{x'} P(x \to x')}_{=1} \quad (\textit{Invariance})$$

- Detailed balance implies general balance but easier to check for.
- Detailed balance implies convergence to a stationary distribution
- If π^* is in detailed balance with P_T, then irrespective of π_0, there is some N for which $\pi_0 \to \pi_N$.
- Detailed balance implies reversibility.

Metropolis Hastings

Draw $x' \sim Q(x'; x)$, the proposal distribution

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

Accept x' with prob. a, else retain x.

- $\Rightarrow$ No need to have pmf in $Q(x'; x)$
- $\Rightarrow$ Satisfies detailed balance
- $\Rightarrow$ Equilibrium distribution is target distribution

Note: $P_T(x \to x') = aQ(x'; x)$

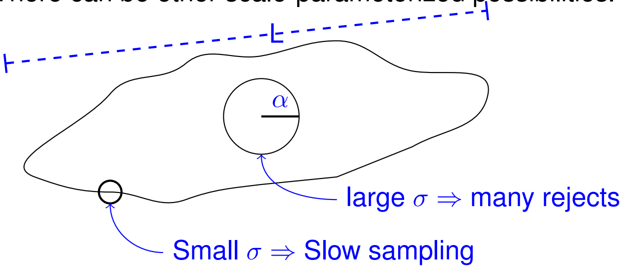
MH Satisfied detailed balance

Proof is easy

$$
\begin{aligned}
P_T(x \to x')\pi^*(x) &= Q(x'; x) \min\left(1, \frac{\pi^*(x')Q(x; x')}{\pi^*(x)Q(x'; x)}\right) \pi^*(x) \\
&= \min\left(\pi^*(x)Q(x'; x), \pi^*(x')Q(x; x')\right) \\
&= Q(x; x') \min\left(\frac{\pi^*(x)Q(x'; x)}{\pi^*(x')Q(x; x')}, 1\right) \pi^*(x') \\
&= P_T(x' \to x)\pi^*(x')
\end{aligned}
$$

Limitations of MH

The transition distribution $N(x, \sigma^2) \Rightarrow$ A local kernel.
There can be other scale-parameterized possibilities.



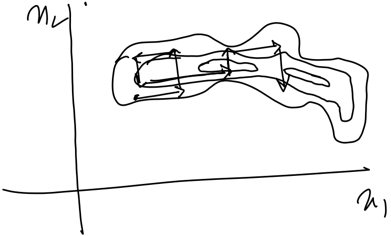
How to select σ adaptively?

On Transitions

$$P_T^N(x_n|x_\gamma) = P_T(x_n|x_{n-1})P_T(x_{n-1}|x_{n-2})\dots P(x_1)$$
$$\text{or } P_T^a(x_n|x_{n-1})P_T^b(x_{n-1}|x_{n-2})\dots$$

- Each transition can be different and individually not be ergodic
- But if P_T^N leaves P^* invariant and is ergodic then OK
- Allows adaptive transitions

Gibbs Sampler: a different transition



Let $\underline{x} = x_1, \cdots, x_n$
(a huge dimensional space) and we want to sample

$$P(\underline{x}) = P(x_1 \cdots x_n)$$
$$P(\underline{x}) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1) \ldots P(x_n|x_{n-1} \ldots x_1)$$

Gibbs:

$$P(x_1) \to P(x_2|x_1) \to P(x_3|x_1, x_2) \to \cdots$$
$$\to P(x_n|x_n - 1 \ldots x_1) \to P(x_1|x_{i\neq 1}) \to P(x_2|x_{i\neq 2}) \ldots$$

Transitions are simple

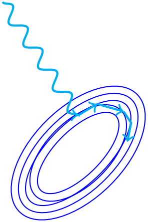
$$P(x_i|x_{j\neq i}) = \frac{P(x_i, x_{j\neq i})}{\sum_{x_i'} P(x_i', x_{j\neq i})}$$

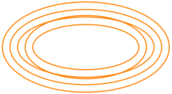
Generally only one dimensional! easy to calculate
Amenable to direct sampling $\rightarrow$ no need for acceptance

Satisfies Detailed Balance

$$
\begin{aligned}
\pi^*(\underline{x})P_T(\underline{x} \to \underline{x}') &= P(x_j', x_{\neq j})P(x_j|x_{\neq j}) \\
&= P(x_j', x_{\neq j})P(x_j|x_{\neq j}) \\
&= P(x_j'|x_{\neq j})P(x_{\neq j})P(x_j|x_{\neq j}) \\
&= P(x_j'|x_{\neq j})P(x_j, x_{\neq j}) \\
&= \pi^*(\underline{x}')P_T(\underline{x} \to \underline{x}')
\end{aligned}
$$

MCMC caveats





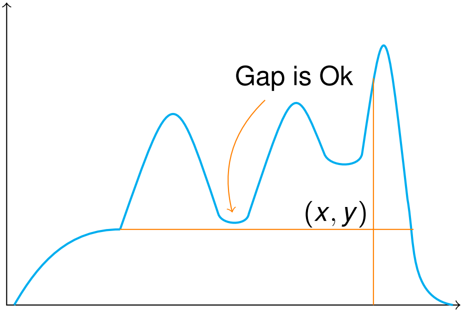
Stuck?

What about burn in?

Stuck in a well?

MCMC typically started from multiple initial starting points, and information is exchanged between chains to better track the underlying probability surface.

Slice Sampler

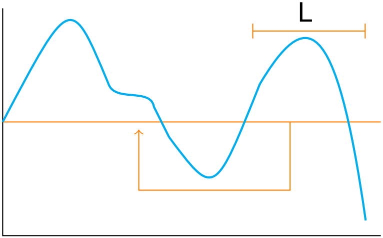


$$P(y|x) = u[0, P(x)] \quad y \sim P(y|x)$$
$$x \sim U[xmin, xmax]$$
$$P(x|y) \propto L(x; y) = \begin{cases} 1 & P(x) \geq y \\ 0 & \text{otherwise} \end{cases}$$

Accept if $L(x; y) = 1$, reject otherwise

Slicing the Slice Sampler

1. No step size like M-H. L/σ iterations vs L^2/σ^2
2. A kind of Gibbs sampler.
3. Bracketing and Rejction can be incorporated.
4. Needs just evaluations of P(x)
5. Scaling in high dimensions?



12.S990 Quantifying Uncertainty
Fall 2012