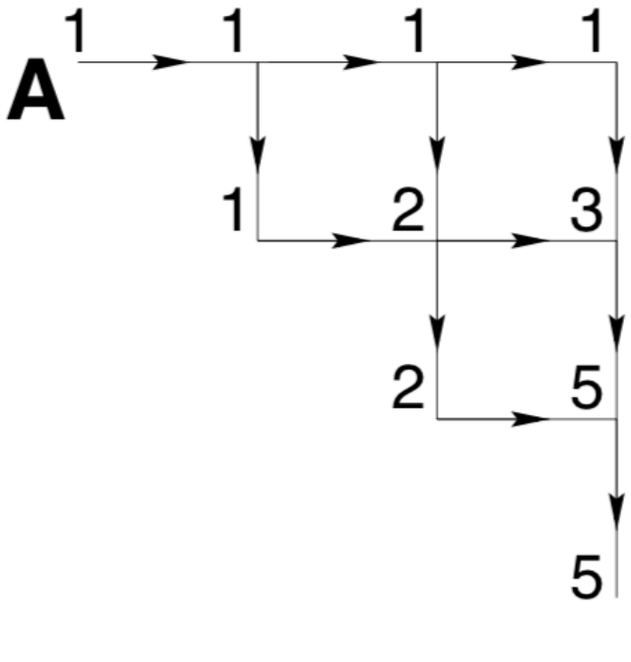
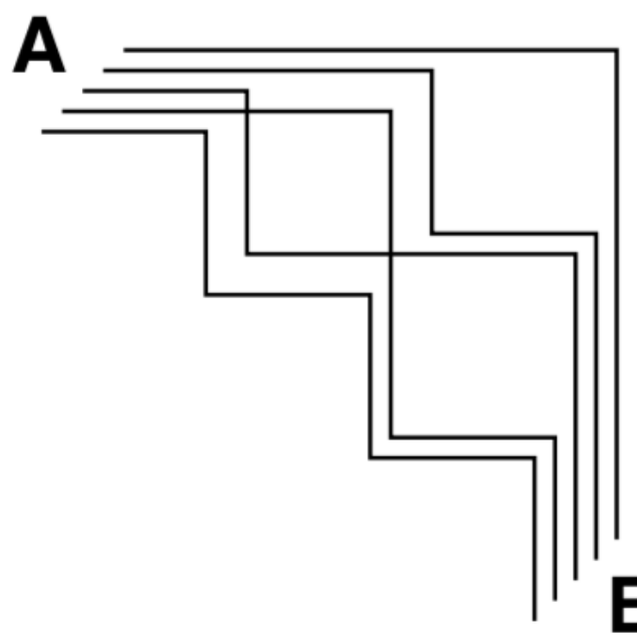
Northeastern

CS 7140:
ADVANCED MACHINE LEARNING

Recap: Message Passing





- Path Counting: pick a point P in the grid and sum over number of paths from A to each of those neighbors.
- Some global functions have a separability property
- Can be computed efficiently with **message passing**

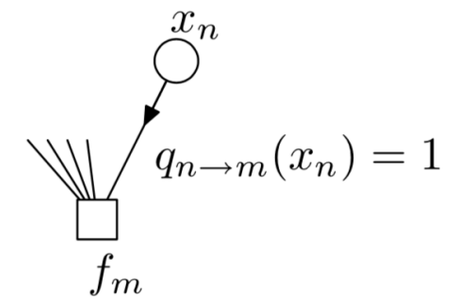
Recap:Sum-Product Algorithm

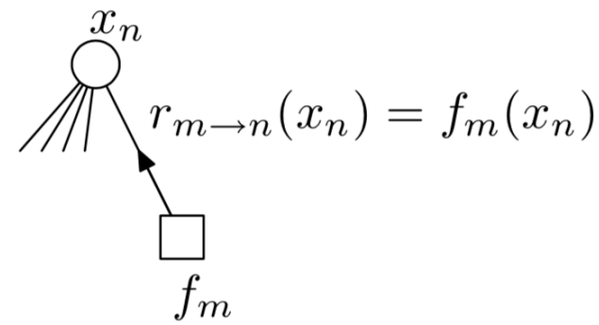
From variable to factor:

$$q_{n\to m}(x_n) = \prod_{m'\in\mathcal{M}(n)\setminus m} r_{m'\to n}(x_n). \qquad (26.11)$$

From factor to variable:

$$r_{m\to n}(x_n) = \sum_{\mathbf{x}_m\setminus n}\left(f_m(\mathbf{x}_m)\prod_{n'\in\mathcal{N}(m)\setminus n} q_{n'\to m}(x_{n'})\right). \qquad (26.12)$$





Two types passing along the edges in the factor graph:

- messages $q_{n\to m}$ from variable nodes to factor nodes
- messages $r_{m\to n}$ from factor nodes to variable nodes.

APPROXIMATE INFERENCE

Monte Carlo Method

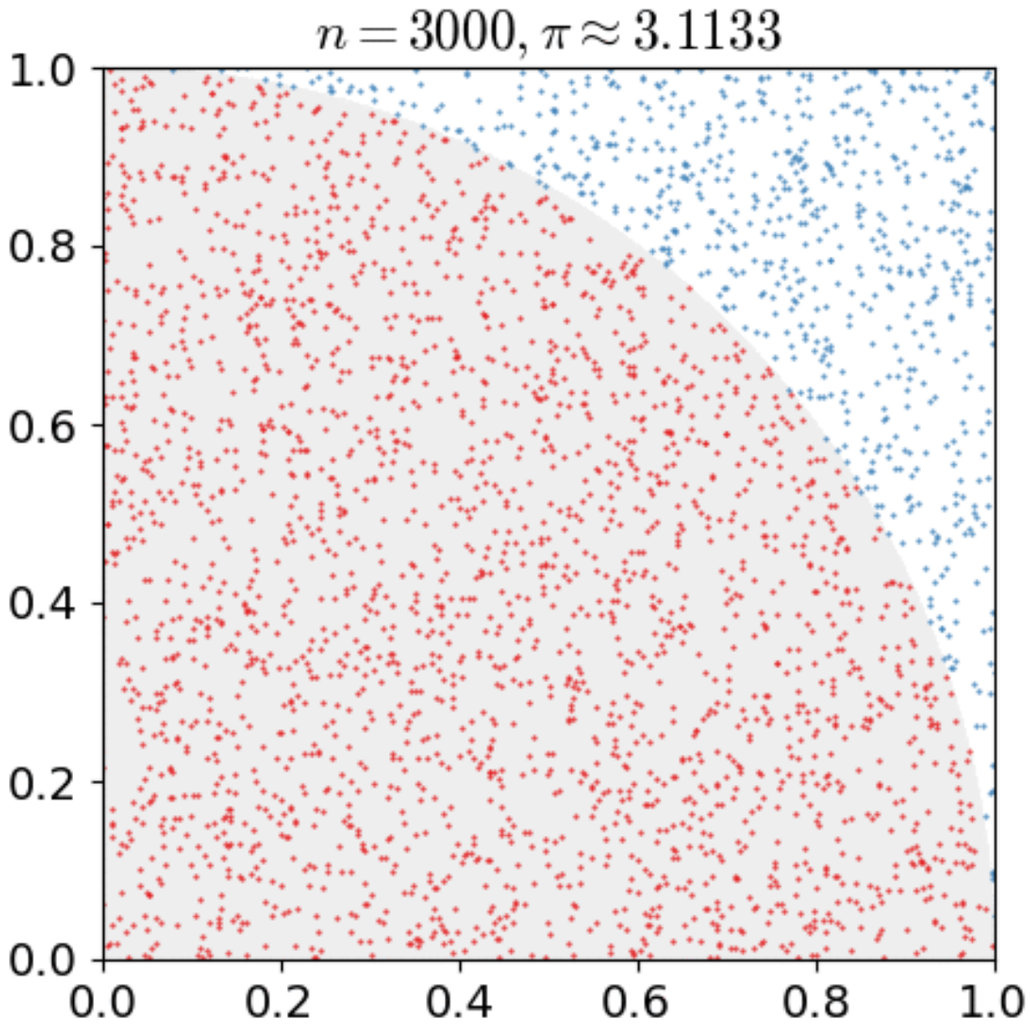
Stanislaw Ulam

John von Neumann

- Computational techniques that make use of random numbers
- Solve problems that might be deterministic with randomness
- Central to the simulations required for the Manhattan project

Monte Carlo Method

Estimating the value of π



- The expected value for finite n is equal to the correct value
- For very large n, the error converges “almost surely” to 0

General Problems

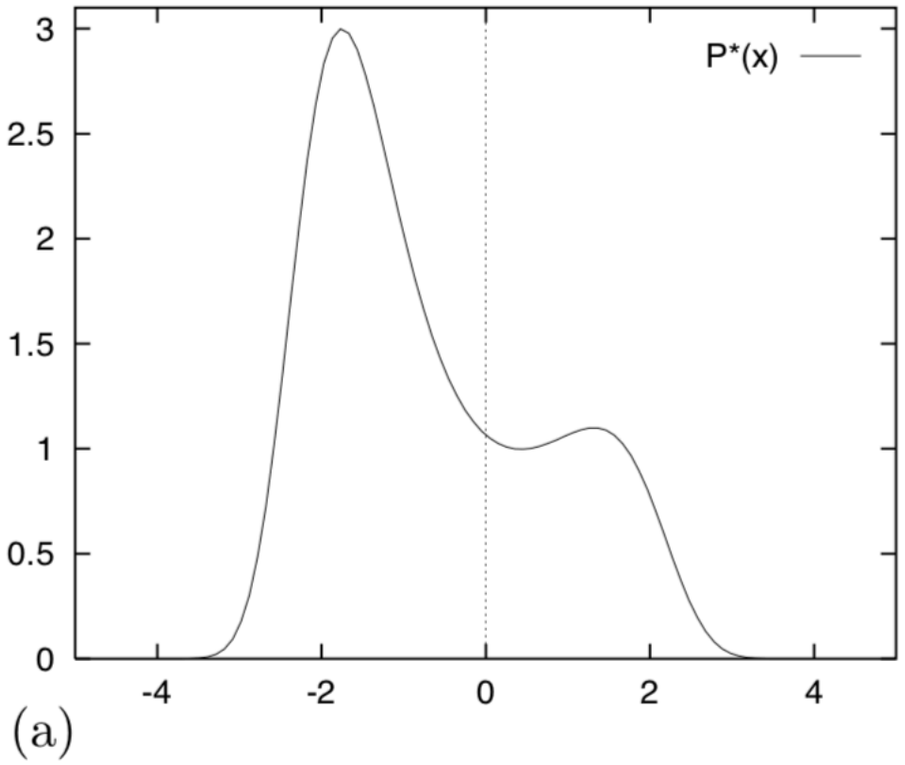
- **Problem 1**: Generate sample from $\{\mathbf{x}^{(r)}\}_{r=1}^{R}$ from a given probability distribution $P(\mathbf{x})$

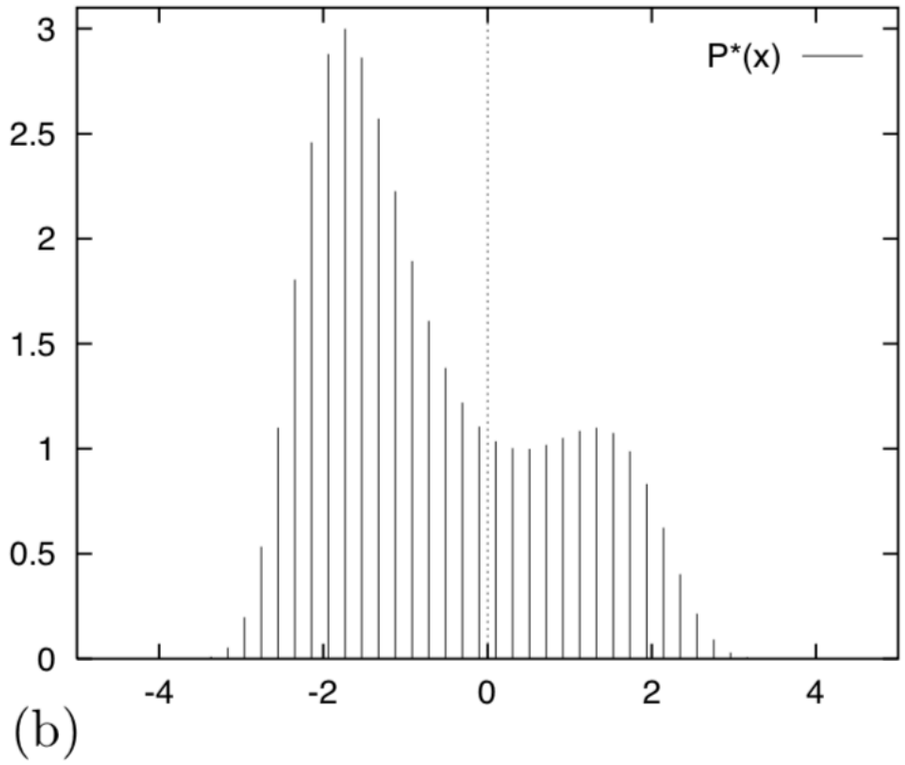
- **Problem 2**: Estimate expectations of functions under the distribution

$$\Phi = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\phi(\mathbf{x})] \equiv \int P(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}$$

$$\hat{\Phi} \equiv \frac{1}{R}\sum_{r} \phi(\mathbf{x}^{(r)})$$

1D Example





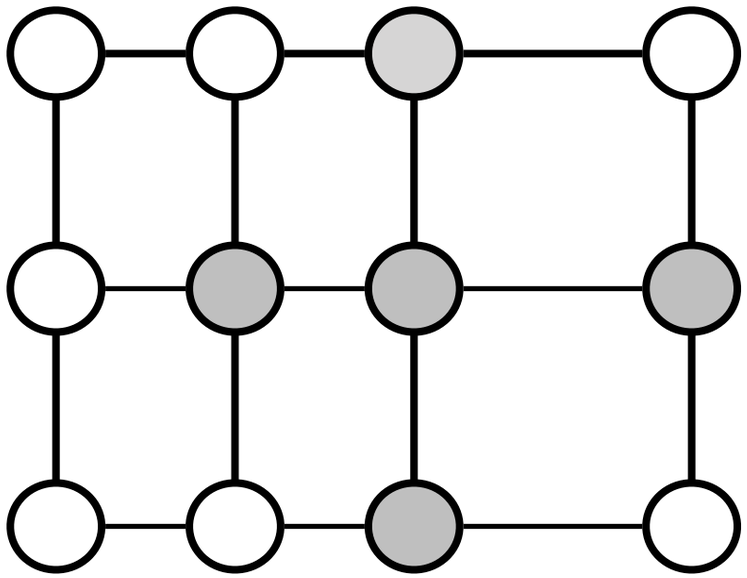
$$P(\mathbf{x}) = P^{\star}(\mathbf{x})/Z$$

$$Z = \int_{\mathbf{x}} P^{\star}(\mathbf{x}) d(\mathbf{x})$$

Sample from $P^{\star}(x) = \exp[0.4(x-0.4)^2 - 0.08x^4], \quad x \in (-\infty, \infty)$

- Discretize x
- Ask for samples from a uniformly spaces points $\{x_i\}$
- Evaluate $p_i^{\star} = P^{\star}(x_i)$ and $Z = \sum_i p_i^{\star}$
- Sample from $p_i = p_i^{\star}/Z$

High-Dimensional Distribution



Ising Model

$$p(x_1, \cdots, x_T) = \frac{1}{Z} \prod_{i,j} \phi(x_i, x_j)$$

- Binary variable, 2^T states
- Evaluate $p_i^\star = P^\star(x_i)$ and $Z = \sum_i p_i^\star$ total 2^T number of times

Draw samples from a high-dimensional distribution is difficult!

Importance Sampling

- A method for Problem 2, not Problem 1

Problem 2: Estimate expectations of functions under the distribution

$$\Phi = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\phi(\mathbf{x})] \equiv \int P(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}$$

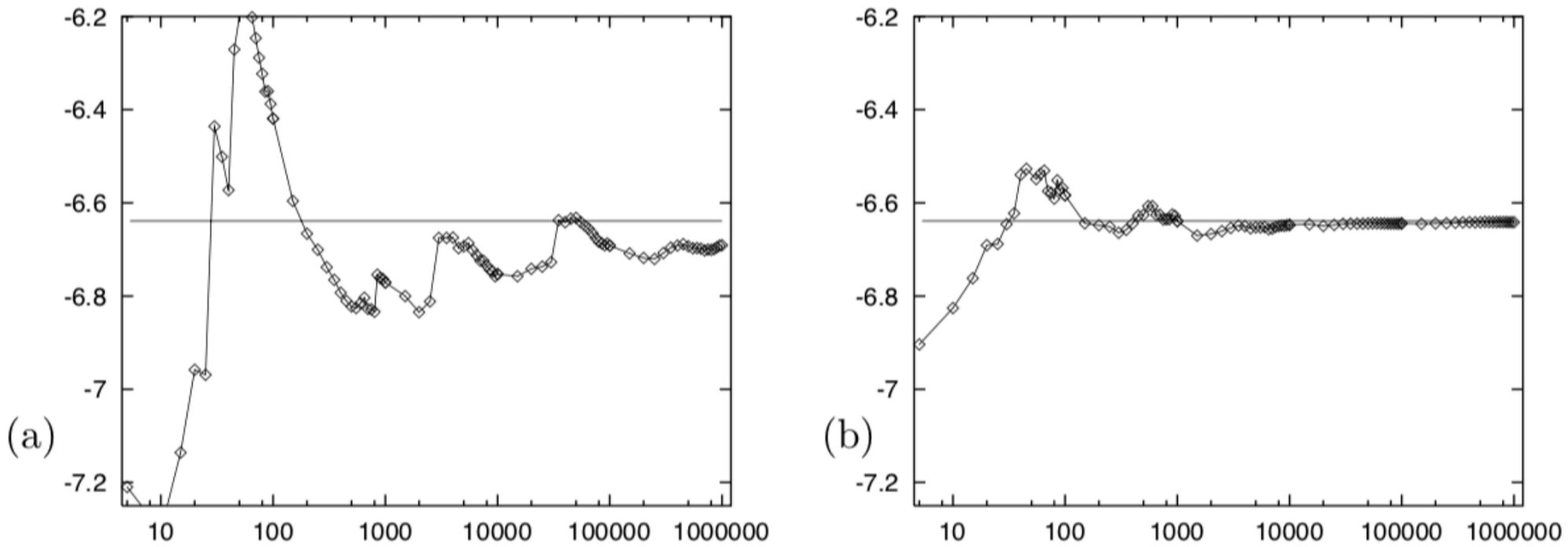
- Sample from $P(x)$ is **hard**, sample from $Q(x)$ is **simple**

Importance: $w_r \equiv \dfrac{P^\star(x^{(r)})}{Q^\star(x^{(r)})}$ target / proposal

$$\hat{\Phi} \equiv \frac{\sum_r w_r \phi(x^{(r)})}{\sum_r w_r}$$

Proposal Distribution

- Proposal density $Q(x)$ is small in a region where $|\phi(x)P^{\star}(x)|$ is large
- After many samples, none of them have fallen in that region
- Estimation would be drastically wrong



Gaussian Cauchy

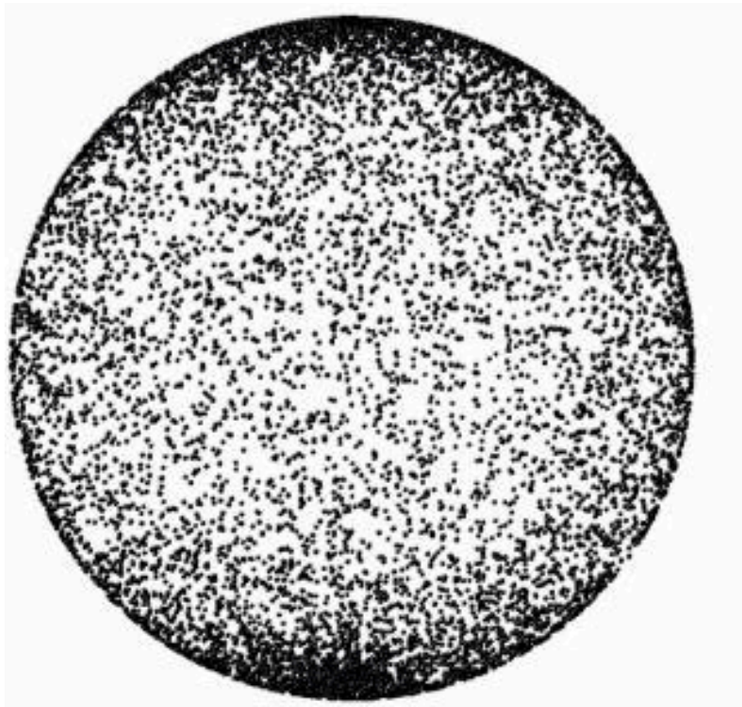
An importance sampler should have **heavy tail**.

IS in High-Dimensions

- $P^{\star}(x) = \begin{cases} 1 & 0 \le \rho(x) \le R \\ 0 & \rho(x) > R \end{cases}$, $\rho(x) \equiv (\sum_i x_i^2)^{1/2}$

  $Q(x) = \prod_i N(x_i; 0, \sigma^2)$

- What is the range of weights w_r ?

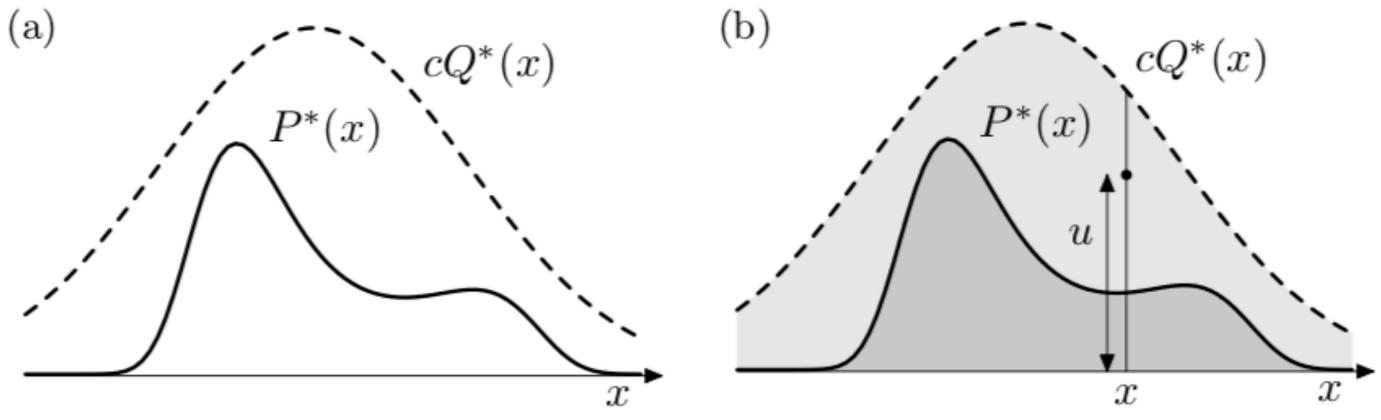


- We know $\rho^2 \sim N\sigma^2 \pm \sqrt{2N}\sigma^2$
- The weights $w_r \sim (2\pi\sigma^2)^{N/2} \exp\left(\frac{N}{2} \pm \frac{2N}{2}\right)$
- The ratio of largest and median weight $\frac{w_r^{max}}{w_r^{med}} = \exp(\sqrt{2N})$
- For 1000 dimension

Dominated by a few samples with huge weights

Rejection Sampling

- A method for Problem 1
  - Assume again sampling from $P(x)$ is **hard**, sampling from $Q(x)$ is **simple**
  - Assume we know c, such that $cQ^\star(x) > P^\star(x)$

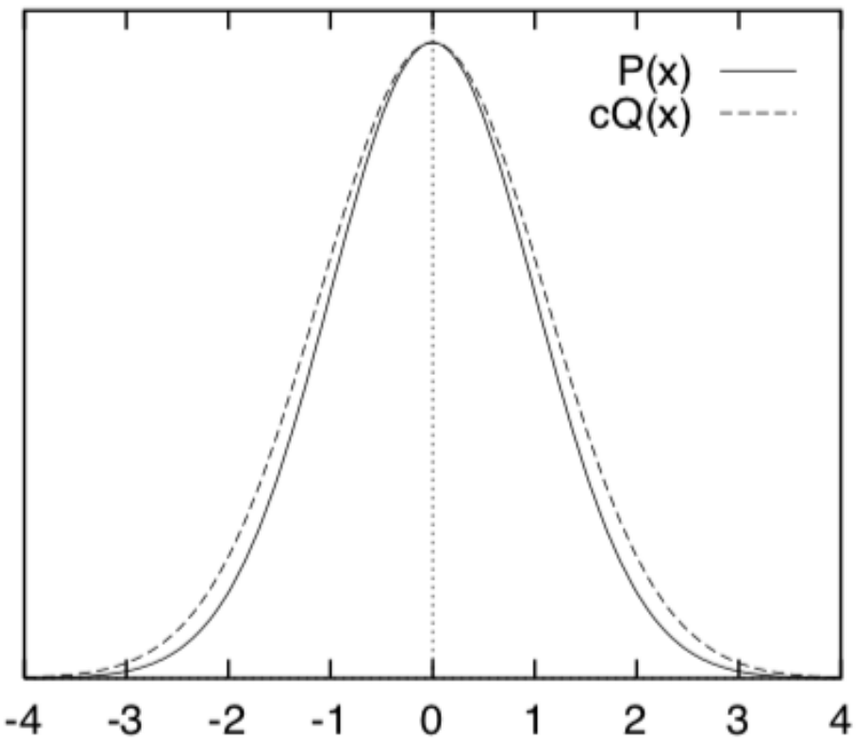


1. Draw a sample x from $Q(x)$
2. Draw a point u from uniform $[0, cQ^\star(x)]$
3. Reject x is $u > P^\star(x)$, accept otherwise

RS in High Dimensions

- Generating samples from one Gaussian with σ_P using the other Gaussian with σ_Q

- With dimension $N = 1000$ and $\frac{\sigma_Q}{\sigma_P} = 1.01$, what is the value of c?



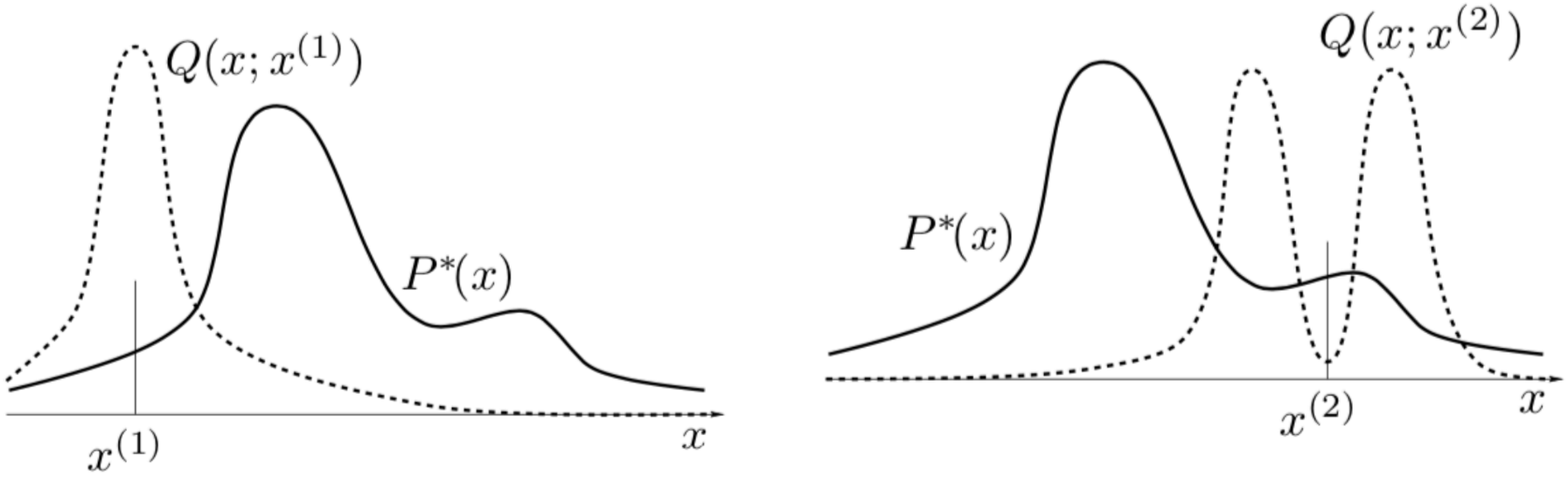
the mean of Q is $1/(2\pi\sigma_Q^2)^{N/2}$

$$c = \frac{(2\pi\sigma_Q^2)^{N/2}}{(2\pi\sigma_P^2)^{N/2}} = \exp\left(N \ln \frac{\sigma_Q}{\sigma_P}\right)$$

Acceptance rate (1/c) is exponentially small in N

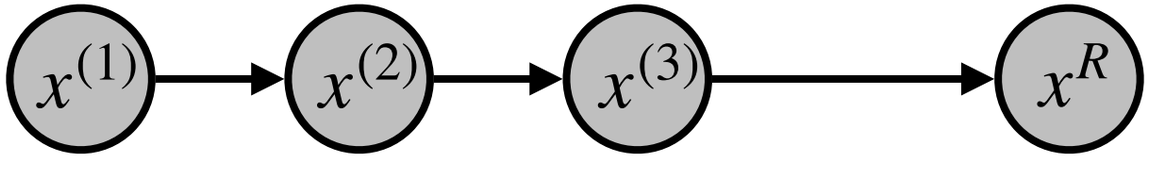
Metropolis-Hastings Method

- IS and RS work well when Q is similar to P
- Difficult to find for large and complex problem
- MHS: Q depends on the current state $x^{(t)}$



Metropolis-Hastings Method

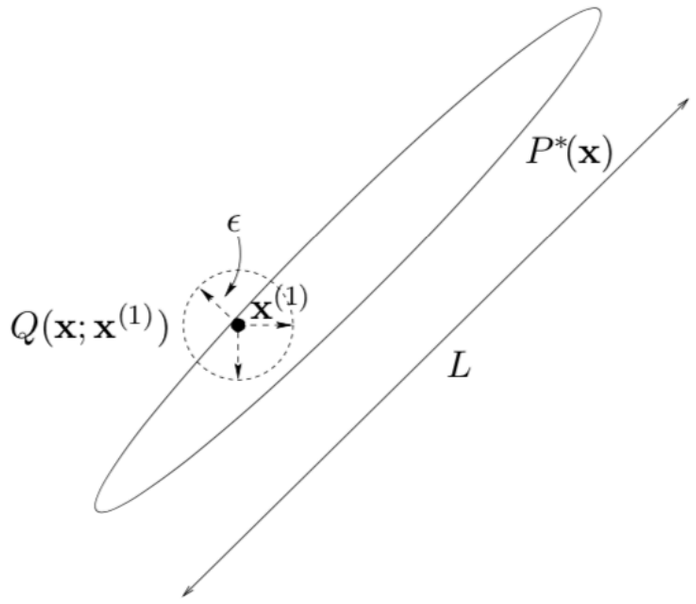
- Draw a sample x from $Q(x; x^{(t)})$
- Evaluate $a = \dfrac{P^{\star}(x)Q(x^{(t)}; x)}{P^{\star}(x^{(t)})Q(x; x^{(t)})}$
- If $a \geq 1$,accept, set $x^{(t+1)} = x$
- Otherwise, reject, set $x^{(t+1)} = x^{(t)}$



Proposal Distribution

MHS is an example of **Markov chain Monte Carlo** (MCMC)

- For any positive Q , the probability distribution of $x^{(t)}$ tends to $P(x) = P^\star(x)/Z$
- Employ a proposal distribution with a small length scale ϵ



Rule of thumb: lower bound on number of iterations of a Metropolis method. If the largest length scale of the space of probable states is L, a Metropolis method whose proposal distribution generates a random walk with step size ϵ must be run for at least

$$T \simeq (L/\epsilon)^2 \qquad (29.32)$$

iterations to obtain an independent sample.

Large scale: low acceptance Small scale: slow progress

Example of MHS

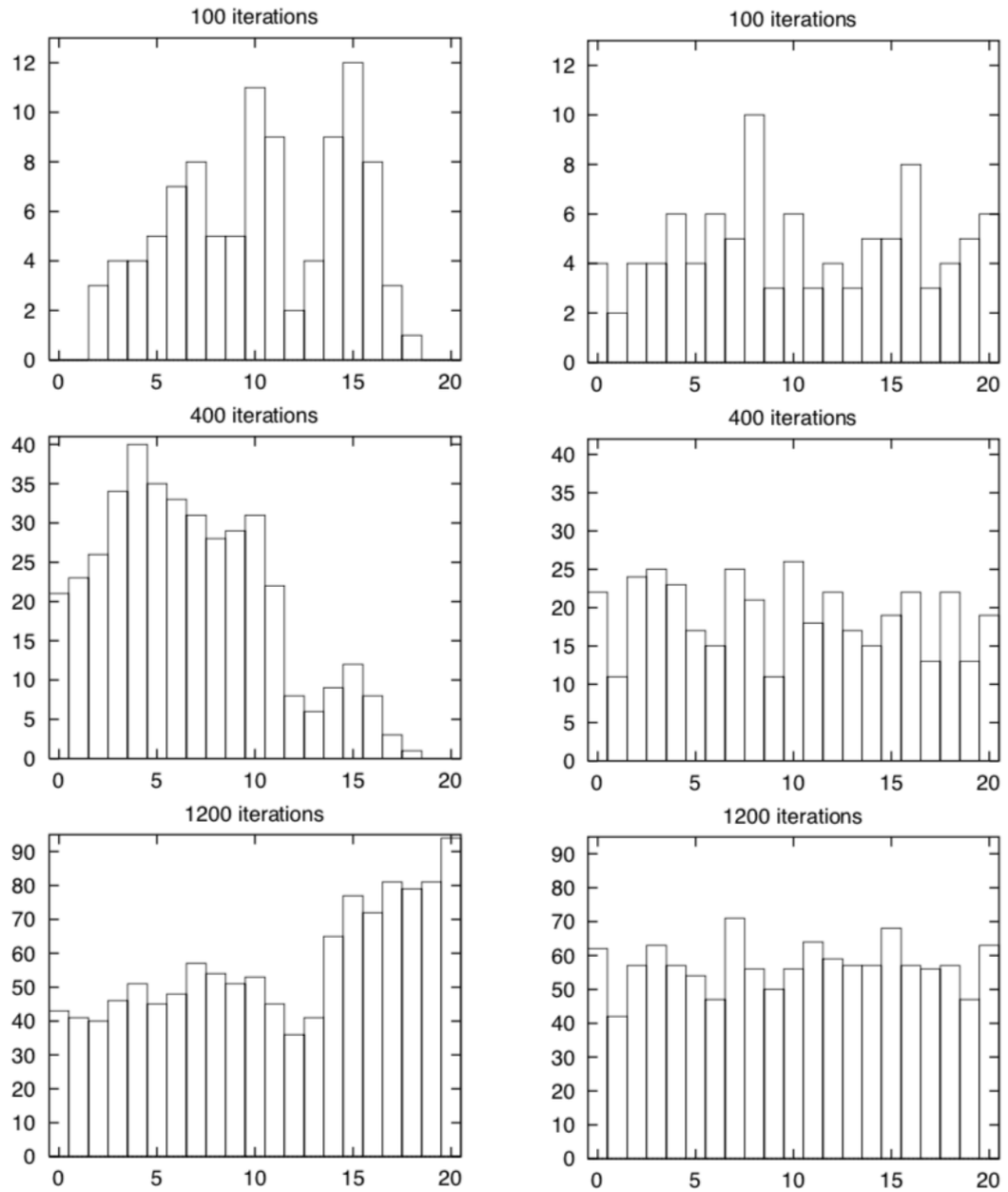
- Sample from

$$P(x) = \begin{cases} 1/21 & x \in (0,1,\ldots,20) \\ 0 & \text{otherwise} \end{cases}$$

- Proposal distribution

$$Q(x', x) = \begin{cases} 1/2 & x' = x \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

Random Walk behavior in Monte Carlo methods



Metropolis Independent

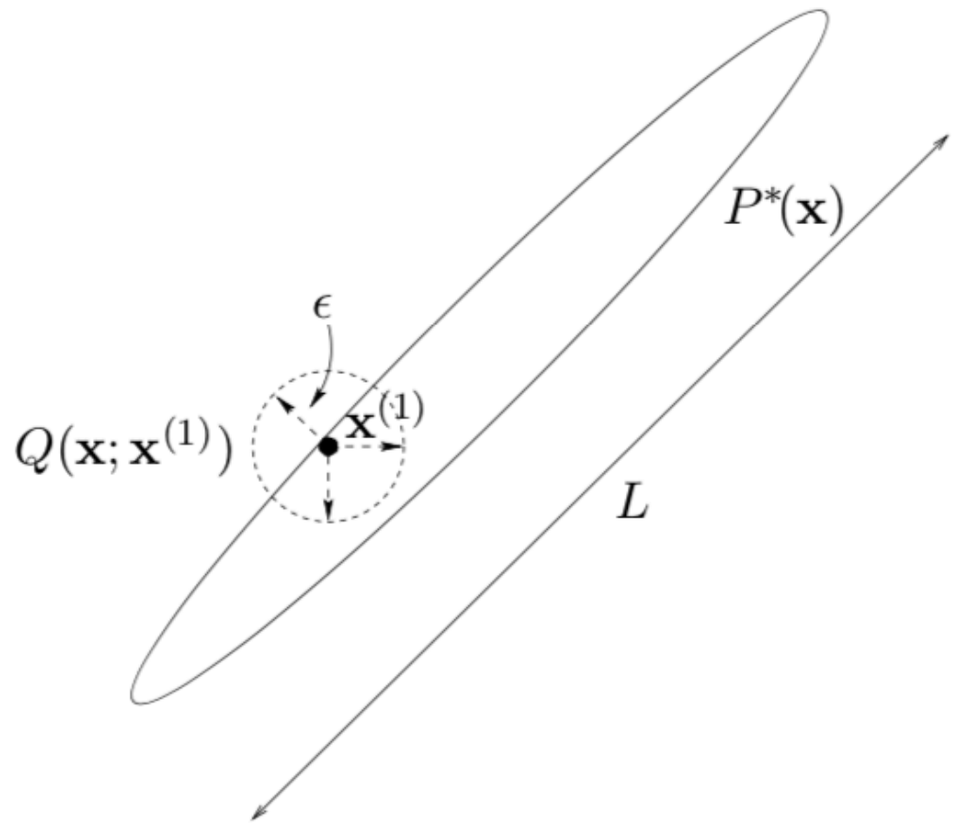
MHS in High Dimensions

- Target distribution P: N-dimensional Gaussian
- Proposal distribution Q: a spherical Gaussian
- P is separable with $\{x_n\}$, σ^{max} and σ^{min} be the standard deviation over dimensions

The time taken to generate independent samples

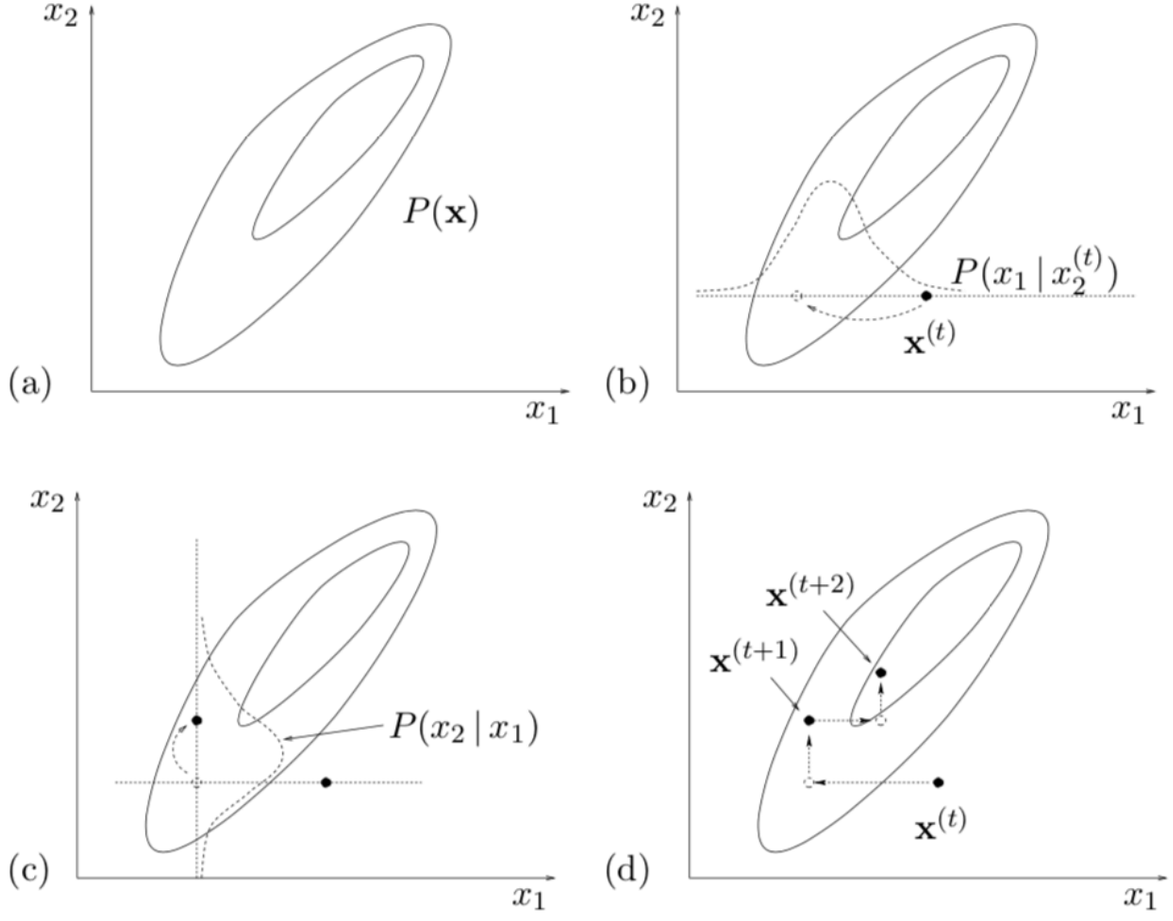
$$T \approx (\sigma^{max}/\epsilon)^2$$

Does NOT dependent on N !



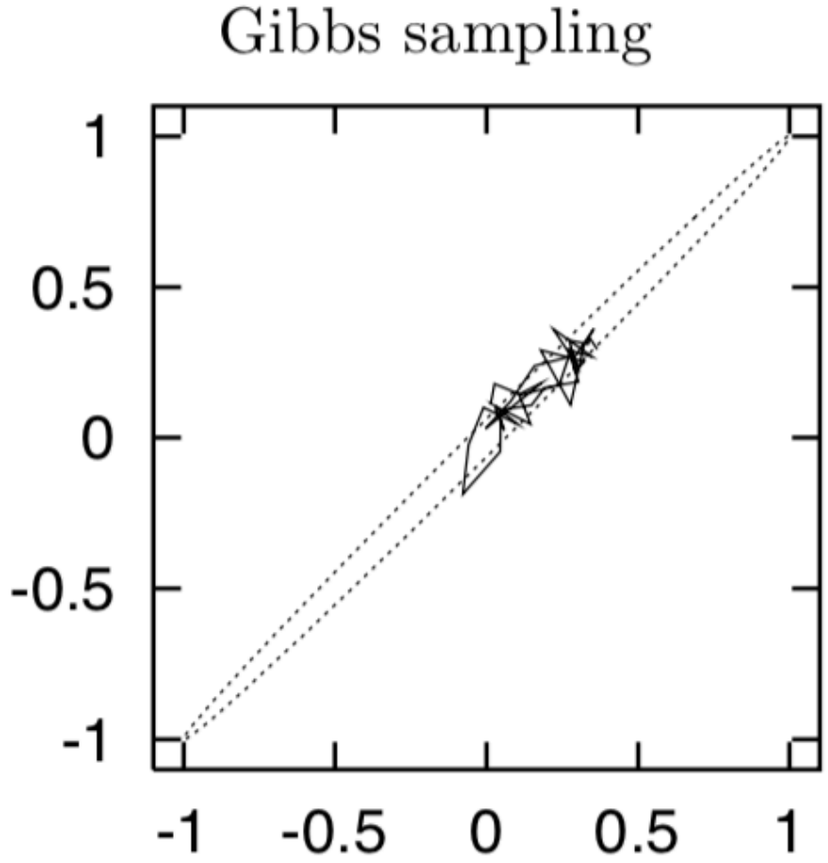
Gibbs Sampling

- A method for Problem 1 with at least two dimensions
- Sample from *joint* distribution $P(\mathbf{x})$ is **hard**, sample from *conditional* distribution $P(x_i \mid \{x\}_{j \neq i})$ is **easy**



GS in High Dimensions

- Same defect as Metropolis method: slow random walk
- Solution: make the probability $P(\mathbf{x})$ separable
  - Hamiltonian Monte Carlo
  - Over-relaxation
  - Simulated Annealing



Monte Carlo Strategies

- How long a MCMC simulation needs to run?
- How to obtain samples with limited computer resources?

1. Make one long run, obtaining all R samples from it.
2. Make a few medium-length runs with different initial conditions
3. Make R short runs, each starting from a different random initial condition, with the only state that is recorded