# On Markov Chain Monte Carlo

Yevgeniy Kovchegov
Oregon State University

## Metropolis-Hastings algorithm.

**Goal:** simulating an $\Omega$-valued random variable distributed according to a given probability distribution $\pi(z)$, given a complex nature of large discrete space $\Omega$.

**MCMC:** generating a Markov chain $\{X_t\}$ over $\Omega$, with distribution $\mu_t(z) = P(X_t = z)$ converging rapidly to its unique stationary distribution, $\pi(z)$.

**Metropolis-Hastings algorithm:** Consider a connected neighborhood network with points in $\Omega$. Suppose we know the ratios of $\frac{\pi(z')}{\pi(z)}$ for any two neighbor points $z$ and $z'$ on the network.

Let for $z$ and $z'$ connected by an edge of the network, the transition probability be set to

$$p(z, z') = \frac{1}{M} \min\left\{1, \frac{\pi(z')}{\pi(z)}\right\} \qquad \text{for } M \text{ large enough.}$$

## Metropolis-Hastings algorithm.

Consider a connected neighborhood network with points in $\Omega$.

Suppose we know the ratios of $\frac{\pi(z')}{\pi(z)}$ for any two neighbor points $z$ and $z'$ on the network.

Let for $z$ and $z'$ connected by an edge of the network, the transition probability be set to

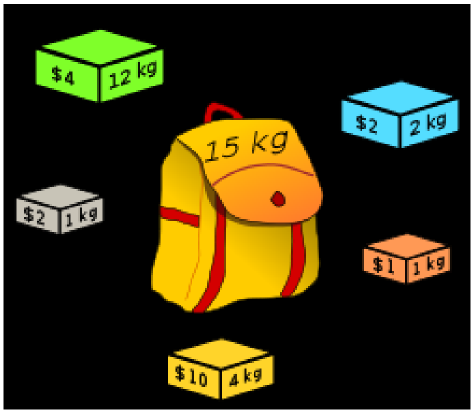
$$p(z,z') = \frac{1}{M} \ \min\left\{1, \ \frac{\pi(z')}{\pi(z)}\right\} \qquad \text{for } M \text{ large enough.}$$

Specifically, $M$ can be any number greater than the maximal degree in the neighborhood network.

Let $p(z,z)$ absorb the rest of the probabilities, i.e.

$$p(z,z) = 1 - \sum_{z': \ z \sim z'} p(z,z')$$

**Knapsack problem.** The **knapsack problem** is a problem in combinatorial optimization: Given a set of items, each with a mass and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. Knapsack problem is NP complete.



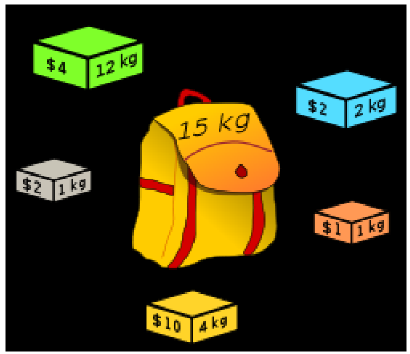

Source: Wikipedia.org

**Knapsack problem.** Given $m$ items of various weights $w_j$ and value $v_j$, and a knapsack with a weight limit $R$. Assuming the volume and shape do not matter, find the most valuable subset of items that can be carried in the knapsack.

Mathematically: we need $z = (z_1, \ldots, z_m)$ in

$$\Omega = \Big\{z \in \{0,1\}^m \ : \ \sum_{j=1}^{m} w_j z_j \le R\Big\}$$

maximizing $U(z) = \sum_{j=1}^{m} v_j z_j$.




Source: Wikipedia.org

**Knapsack problem.** Find $z = (z_1, \ldots, z_m)$ in

$$\Omega = \Big\{ z \in \{0,1\}^m \ : \ \sum_{j=1}^{m} w_j z_j \le R \Big\} \text{ maximizing } \ U(z) = \sum_{j=1}^{m} v_j z_j .$$

• **MCMC approach:** Assign weights $\pi(z) = \frac{1}{Z_\beta} \exp\big\{\beta\, U(z)\big\}$ to each $z \in \Omega$ with $\beta = \frac{1}{T}$, where

$$Z_\beta = \sum_{z \in \Omega} \exp\big\{\beta \ U(z)\big\}$$

is called partition function. Next, for each $z \in \Omega$ consider a **clique** $\mathcal{C}_z$ of neighbor points in $\Omega$. Consider a Markov chain over $\Omega$ that jumps from $z$ to a neighbor $z' \in \mathcal{C}_z$ with probability

$$p(z, z') = \frac{1}{M} \ \min\left\{ 1, \ \frac{\pi(z')}{\pi(z)} \right\} \qquad \text{for } M \text{ large enough.}$$

**Knapsack problem.** Assign weights $\pi(z) = \frac{1}{Z_\beta} \exp\left\{\beta\ U(z)\right\}$ to each $z \in \Omega$ with $\beta = \frac{1}{T}$, where

$$Z_\beta = \sum_{z \in \Omega} \exp\left\{\beta\ U(z)\right\}$$

is called partition function. Next, for each $z \in \Omega$ consider a **clique** $\mathcal{C}_z$ of neighbor points in $\Omega$. Consider a Markov chain over $\Omega$ that jumps from $z$ to a neighbor $z' \in \mathcal{C}_z$ with probability

$$p(z, z') = \frac{1}{M}\ \min\left\{1,\ \frac{\pi(z')}{\pi(z)}\right\} \qquad \text{for } M \text{ large enough.}$$

Observe that

$$\frac{\pi(z')}{\pi(z)} = \exp\left\{\beta\ \big(U(z') - U(z)\big) = \exp\left\{\beta\ \big(v \cdot (z' - z)\big)\right\},\right.$$

where $v = (v_1, \ldots, v_m)$ is the values vector.

## Knapsack and other optimization problems.

- **Issues:**

**(i)** Running time?

Analyzing mixing time is challenging in MCMC for real-life optimization problems such as knapsack problem. With few exceptions – no firm foundation exists, and no performance guaranteed.

**(ii)** Optimal $T$?

$T$ is usually chosen using empirical observations, trial and error, or certain heuristic.

Often, simulated annealing approach is used.

## Simulated annealing.

Usually, we let $\pi(z) = \frac{1}{Z_\beta} \exp\left\{\beta\ U(z)\right\}$ to each $z \in \Omega$ with $\beta = \frac{1}{T}$, and $p(z, z') = \frac{1}{M}\ \min\left\{1,\ \frac{\pi(z')}{\pi(z)}\right\}$.

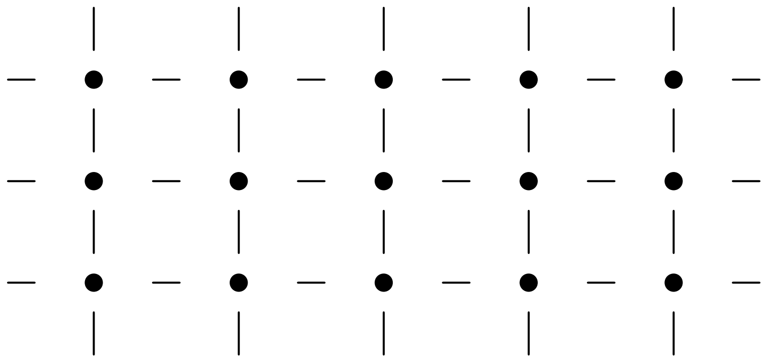
- **Idea:** What if we let temperature $T$ change with time $t$, i.e. $T = T(t)$? When $T$ is large, the Markov chain is more diffusive; as $T$ gets smaller, the value $X_t$ stabilizes around the maxima.

The method was independently devised by S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi in 1983, and by V. Černý in 1985.

Name comes from *annealing in metallurgy*, a technique involving heating and controlled cooling.

**Gibbs Sampling: Ising Model.** Every vertex $v$ of $G = (V, E)$ is assigned a spin $\sigma(v) \in \{-1, +1\}$. The probability of a configuration $\sigma \in \{-1, +1\}^V$ is

$$\pi(\sigma) = \frac{e^{-\beta \mathcal{H}(\sigma)}}{Z(\beta)}, \quad \text{where} \quad \beta = \frac{1}{T}$$



**Gibbs Sampling: Ising Model.** Every vertex $v$ of $G = (V, E)$ is assigned a spin $\sigma(v) \in \{-1, +1\}$. The probability of a configuration $\sigma \in \{-1, +1\}^V$ is

$$\pi(\sigma) = \frac{e^{-\beta \mathcal{H}(\sigma)}}{Z(\beta)}, \quad \text{where} \quad \beta = \frac{1}{T}$$

$$\begin{array}{ccccccccccc}
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & -1 & - & +1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & +1 & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | & 
\end{array}$$

**Gibbs Sampling: Ising Model.** $\forall \sigma \in \{-1,+1\}^V$, the Hamiltonian

$$\mathcal{H}(\sigma) = -\frac{1}{2} \sum_{u,v:\ u \sim v} \sigma(u)\sigma(v) = -\sum_{edges\ \ e=[u,v]} \sigma(u)\sigma(v)$$

and probability of a configuration $\sigma \in \{-1,+1\}^V$ is

$$\pi(\sigma) = \frac{e^{-\beta\mathcal{H}(\sigma)}}{Z(\beta)}, \quad \text{where} \quad \beta = \frac{1}{T}$$

$Z(\beta) = \sum_{\sigma \in \{-1,+1\}^V} e^{-\beta\mathcal{H}(\sigma)}$ - normalizing factor.

## Ising Model: local Hamiltonian

$$\mathcal{H}(\sigma) = -\frac{1}{2} \sum_{u,v:\ u \sim v} \sigma(u)\sigma(v) = -\sum_{edges\ e=[u,v]} \sigma(u)\sigma(v)$$

The local Hamiltonian

$$\mathcal{H}_{local}(\sigma, v) = -\sum_{u:\ u \sim v} \sigma(u)\sigma(v)\ .$$

Observe: conditional probability for $\sigma(v)$ is given by $\mathcal{H}_{local}(\sigma, v)$:

$$\mathcal{H}(\sigma) = \mathcal{H}_{local}(\sigma, v) - \sum_{e=[u_1,u_2]:\ u_1,u_2 \neq v} \sigma(u_1)\sigma(u_2)$$

## Gibbs Sampling: Ising Model via Glauber dynamics.

$$
\begin{array}{ccccccccccc}
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & -1 & - & +1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & \sigma(v) & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | &
\end{array}
$$

Observe: conditional probability for $\sigma(v)$ is given by $\mathcal{H}_{local}(\sigma, v)$:

$$\mathcal{H}(\sigma) = \mathcal{H}_{local}(\sigma, v) - \sum_{e=[u_1,u_2]:\ u_1,u_2 \neq v} \sigma(u_1)\sigma(u_2)$$

## Gibbs Sampling: Ising Model via Glauber dynamics.

$$\begin{array}{ccccccccccc}
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & -1 & - & +1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & ? & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & -1 & - & -1 & - & +1 & - & -1 & - \\
 & | & & | & & | & & | & & | &
\end{array}$$

Randomly pick $v \in G$, erase the spin $\sigma(v)$. Choose $\sigma_+$ or $\sigma_-$:

$$\mathbf{Prob}(\sigma \to \sigma_+) = \frac{e^{-\beta \mathcal{H}(\sigma_+)}}{e^{-\beta \mathcal{H}(\sigma_-)} + e^{-\beta \mathcal{H}(\sigma_+)}}$$

$$= \frac{e^{-\beta \mathcal{H}_{local}(\sigma_+, v)}}{e^{-\beta \mathcal{H}_{local}(\sigma_-, v)} + e^{-\beta \mathcal{H}_{local}(\sigma_+, v)}} = \frac{e^{-2\beta}}{e^{-2\beta} + e^{2\beta}} \; .$$

## Glauber dynamics: Rapid mixing.

Glauber dynamics - a random walk on state space $S$ (here $\{-1,+1\}^V$) s.t. needed $\pi$ is stationary w.r.t. Glauber dynamics.

In high temperatures (i.e. $\beta = \frac{1}{T}$ small enough) it takes $O(n \log n)$ iterations to get "$\varepsilon$-close" to $\pi$. Here $|V| = n$.

Need: $\boxed{\max_{v \in V} deg(v) \cdot \tanh(\beta) < 1}$

Thus the Glauber dynamics is a fast way to generate $\pi$. It is an important example of **Gibbs sampling**.

## Close enough distribution and mixing time.

What is "$\varepsilon$-close" to $\pi$? Start with $\sigma_0$:

$$\begin{array}{ccccccccccc}
 & | & & | & & | & & | & & | & \\
- & +1 & - & +1 & - & +1 & - & +1 & - & +1 & - \\
 & | & & | & & | & & | & & | & \\
- & +1 & - & +1 & - & +1 & - & -1 & - & -1 & - \\
 & | & & | & & | & & | & & | & \\
- & -1 & - & -1 & - & -1 & - & -1 & - & -1 & - \\
 & | & & | & & | & & | & & | &
\end{array}$$

If $P_t(\sigma)$ is the probability distribution after $t$ iterations, the total variation distance

$$\|P_t - \pi\|_{TV} = \frac{1}{2} \sum_{\sigma \in \{-1,+1\}^V} |P_t(\sigma) - \pi(\sigma)| \le \varepsilon \; .$$

## Close enough distribution and mixing time.

**Total variation distance:**

$$\|\mu - \nu\|_{TV} := \frac{1}{2}\sum_{x\in S} |\mu(x) - \nu(x)| = \sup_{A\subset S} |\mu(A) - \nu(A)|$$

**Mixing time:**

$$t_{mix}(\varepsilon) := \inf\left\{t \ : \ \|P_t - \pi\|_{TV} \leq \varepsilon, \quad \text{all } \sigma_0\right\} .$$

In high temperature, $t_{mix}(\varepsilon) = O(n \log n)$.

## Coupling Method.

$S$ - sample space

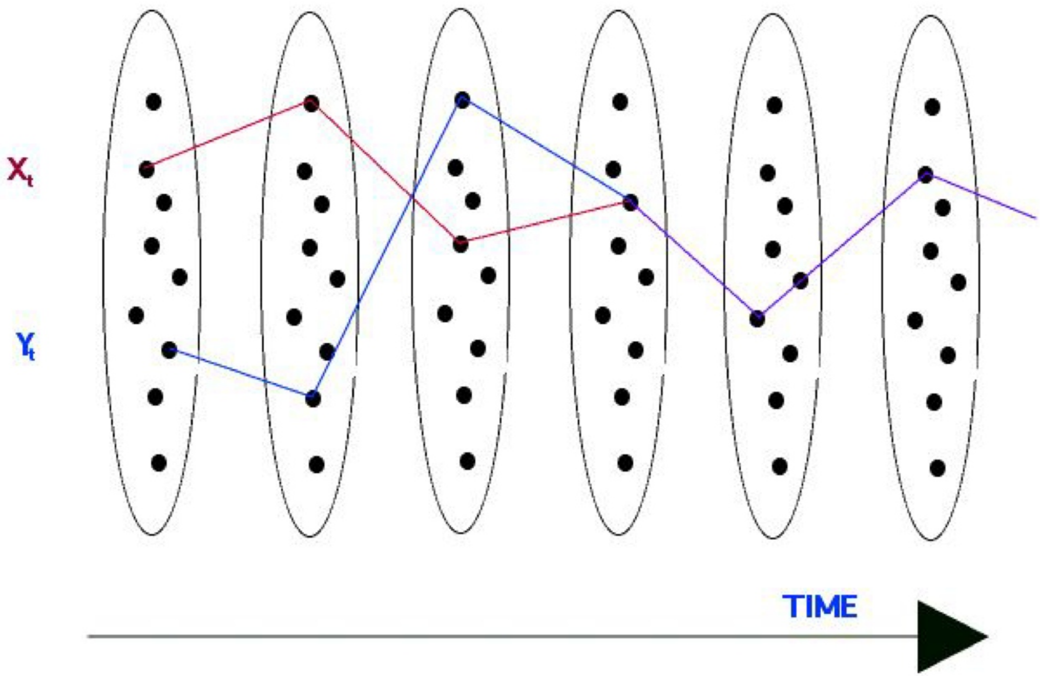
$\{p(i,j)\}_{i,j \in S}$ - transition probabilities

Construct process $\begin{pmatrix} X_t \\ Y_t \end{pmatrix}$ on $S \times S$ such that

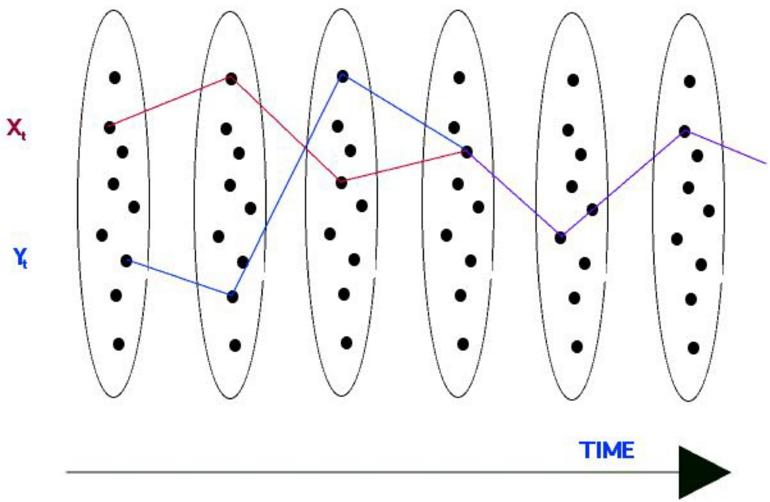
$X_t$ is a $\{p(i,j)\}$-Markov chain
$Y_t$ is a $\{p(i,j)\}$-Markov chain

Once $X_t = Y_t$, let $X_{t+1} = Y_{t+1}$, $X_{t+2} = Y_{t+2}, \ldots$

## Coupling Method.



## Coupling Method.




Coupling time: $T_{coupling} = \min\{t : \ X_t = Y_t\}$

Successful coupling: $\mathbf{Prob}(T_{coupling} < \infty) = 1$

## Mixing times via coupling.

Let $T_{i,j}$ be coupling time for $\begin{pmatrix} X_t \\ Y_t \end{pmatrix}$ given $X_0 = i$ and $Y_0 = j$. Then

$$\|P_{X_t} - P_{Y_t}\|_{TV} \leq P[T_{i,j} > t] \leq \frac{E[T_{i,j}]}{t}$$

Now, if we let $Y_0 \sim \pi$, then for any $X_0 \in S$,

$$\|P_{X_t} - \pi\|_{TV} = \|P_{X_t} - P_{Y_t}\|_{TV} \leq \frac{\max_{i,j \in S} E[T_{i,j}]}{t} \leq \varepsilon$$

whenever $t \geq \frac{\max_{i,j \in S} E[T_{i,j}]}{\varepsilon}$.

## Mixing times via coupling.

$\|P_{X_t} - \pi\|_{TV} \leq \varepsilon$ whenever $t \geq \frac{\max_{i,j\in S} E[T_{i,j}]}{\varepsilon}$.

Thus

$$t_{mix}(\varepsilon) = \inf\left\{t : \ \|P_{X_t} - \pi\|_{TV} \leq \varepsilon\right\} \leq \frac{\max_{i,j\in S} E[T_{i,j}]}{\varepsilon}.$$

So,

$$O(t_{mix}) \leq O(T_{coupling}) \ .$$

Thus constructing a coupled process that minimizes $E[T_{coupling}]$ gives an effective upper bound on mixing time.

## Coupon collector.


$n$ types of coupons: $\boxed{1}$, $\boxed{2}$, $\ldots$, $\boxed{n}$

Collecting coupons: coupon / unit of time, each coupon type is equally likely.

Goal: To collect a coupon of each type.

Question: How much time will it take?

## Coupon collector.

$$\begin{array}{ccccccccc}
\boxed{\text{c1}} & \boxed{\text{c2}} & \boxed{\text{c1}} & \boxed{\text{c3}} & \boxed{\text{c3}} & \boxed{\text{c2}} & \boxed{\text{c4}} & \boxed{\text{c5}} & \ldots \\
\uparrow & \uparrow & & \uparrow & & & \uparrow & \uparrow & \\
\tau_1 = 1 & \tau_2 & & \tau_3 & & & \tau_4 & \tau_5 &
\end{array}$$

Here $\tau_1 = 1$, $E[\tau_2 - \tau_1] = \frac{n}{n-1}$, $E[\tau_3 - \tau_2] = \frac{n}{n-2}, \ldots, E[\tau_n - \tau_{n-1}] = n$.

Hence

$$E[\tau_n] = n\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right) = n \log n + O(n)$$