Monte Carlo Markov Chain (MCMC) method is conceptually related to BP. The bounds on mixing time and the bounds on BP convergence both depend on the amount of influence one node has on the rest. It is hence worth our while to spend one class on this method.

1 Monte Carlo Markov Chain method

Given a joint distribution
\[
\mu(x) = \frac{1}{\mathcal{Z}} \prod_{a \in \mathcal{F}} \Phi(x_a)
\]
where \( x \in X \), we want to compute \( \mu(x) \) by sampling \( x^{(1)}, \ldots, x^{(N)} \) approximately iid from \( \mu(\cdot) \) (the Monte Carlo part). In order to obtain iid samples, we introduce an irreducible and aperiodic Markov Chain that has \( \mu \) as the equilibrium measure.

We’ll use the following example throughout this lecture:
Given \( G = (V, E) \), sample independent sets with probability
\[
\mu(x) = \frac{1}{\mathcal{Z}} \prod_{(i,j) \in E} I_{\{x_i, x_j \neq (1,1)\}} \prod_{i \in V} \lambda^{x_i}
\]
where \( x_i \in \{0,1\} \) and \( i \in V \).

Next, we introduce a method, Metropolis Dynamics, to construct a markov chain with \( \mu \) as its equilibrium measure. The method specifies a transition probability from one configuration to the next, such that
\[
\mu(x)P(x \to \hat{x}) = \mu(\hat{x})P(\hat{x} \to x)
\]
For our example, Metropolis Dynamics works as follows:

- Given a current configuration \( x \), first choose \( i \in G \) uniformly at random.
- Set \( \hat{x}_j = x_j \forall j \neq i \), and choose \( \hat{x} \in \{0,1\} \) as the “proposal”.
- The proposal is accepted with probability
  \[
  \pi = \min \{1, \lambda^{\hat{x}_i - x_i}\}
  \]
  if \( \hat{x}_i = 0 \), or \( \hat{x}_i = 1 \) and all the neighbors of \( i \) are empty. Otherwise, \( \pi = 0 \). The transition rule can also be specified as

  If \( x_i = 1 \),
  \[
  \hat{x}_i = \begin{cases} 
  0 & \text{with prob } \frac{1}{2} \min(1, \lambda^{-1}) \\
  x_i & \text{otherwise}
  \end{cases}
  \]
  If \( x_i = 0 \),
  \[
  \hat{x}_i = \begin{cases} 
  1 & \text{if } x_j = 0 \forall j \sim i \text{ and with prob } \frac{1}{2} \min(1, \lambda) \\
  x_i & \text{otherwise}
  \end{cases}
  \]

Notice that with probability half, \( x_i \) stays the same. Since the chain is irreducible, it can go from one independent set to any other independent set.
2 Mixing Time

After defining a Markov chain with $\mu$ as its equilibrium measure, we start from an arbitrary configuration $x^{(0)}$, e.g. a null independent set, and run the transitions for $t$ steps. We “pretend” that the configuration $x^{(t)}$ is distributed according to $\mu(\cdot)$ to compute the desired marginal. In other words, we would like

$$\mu^{(t)}(x) \approx \mu(x)$$

In order to make this approximation precise, we define

**Total variation distance**

$$\|\mu^{(t)} - \mu\|_{TV} \triangleq \frac{1}{2} \sum_x |\mu^{(t)}(x) - \mu(x)|$$

and

**Mixing time**

$$\tau_{mix}(\epsilon) = \sup_{x_0} \inf_{t} \{t : \|\mu^{(t)} - \mu\|_{TV} \leq \epsilon \forall t \geq \tau\}$$

Notice that

$$\|\mu^{(t)} - \mu\|_{TV} \leq \epsilon \Rightarrow \left| \sum_x f(x) \mu^{(t)}(x) - \sum_x f(x) \mu(x) \right| \leq 2\epsilon \sup_x |f(x)|$$

In particular

$$|\mu^{(t)}(x_i) - \mu(x_i)| \leq \epsilon$$

3 Bound Mixing Time using Coupling

The next question is: How can we upper bound the mixing time?

One method that is easy to apply is Path Coupling, one class of coupling method developed by Bubley and Dyer. There are many other techniques that are more powerful, but Path Coupling is the easiest.

**Path Coupling Method**

Given two rv. $X, Y$ on different probability spaces, a coupling is a rv. $(\tilde{X}, \tilde{Y})$, such that $\tilde{X}$ is distributed as $X$ and $\tilde{Y}$ as $Y$.

We prove a lemma that serves as the foundation of path coupling method.

**Lemma 1.** Given $X_1 \sim \mu_1$, $X_2 \sim \mu_2$ and coupling $(X_1, X_2)$, we have

$$\|\mu_1 - \mu_2\|_{TV} \leq P(X_1 \neq X_2)$$

**Proof**

$$P(X_1 \neq X_2) = \sum_x (P(X_1 = x) - P(X_1 = x, X_2 = x))$$

$$\geq \sum_x (P(X_1 = x) - \min [P(X_1 = x), P(X_2 = x)])$$

$$= \sum_x \max (P(X_1 = x) - P(X_2 = x), 0)$$

$$= \frac{1}{2} \sum_x |P(X_1 = x) - P(X_2 = x)|$$
Corollary 2. Let $x^{(1,t)}, x^{(2,t)}$ be two realizations of the Markov Chain st $x^{(1,0)} = x^{(0)}, x^{(2,t)} \sim \mu$. Then, for any coupling of $\{x^{(1,t)}\}, \{x^{(2,t)}\}$,

$$\|\mu^{(1)} - \mu\|_{TV} \leq P(x^{(1,t)} \neq x^{(2,t)})$$

In order to see how we can apply this corollary, we will go back to our example on independent sets.

First we define the distance of two configurations $x, \hat{x}$.

Definition 3.

$$D(x, \hat{x}) = \left[ \text{minimal number of "allowed" moves to go from } x \text{ to } \hat{x} \right]$$

where “allowed” moves are defined as the change of one of the $x_i$s in such a way that the resulting configuration is still an independent set. Notice that $D(x, \hat{x}) \leq 2n$.

Suppose we are able to prove

$$E[D(x^{(1,t+1)}, x^{(2,t+1)}) | x^{(1,t)}, x^{(2,t)}] \leq \beta D(x^{(1,t)}, x^{(2,t)})$$

for some $\beta < 1$, then

$$E[D(x^{(1,t)}, x^{(2,t)})] \leq \beta^t \cdot 2N$$

$$\Rightarrow P(x^{(1,t)} \neq x^{(2,t)}) \leq 2N/\beta^t$$

$$\Rightarrow P(x^{(1,t)} \neq x^{(2,t)}) < \epsilon \text{ for } t \geq \left( \frac{\log 2N}{\log 1/\beta} \right)$$

$$\Rightarrow \tau_{mix}(\epsilon) \leq \frac{\log 2N}{\log 1/\beta}$$

Refer to figure (1) for a picture of coupling.

4 Proof of Inequality (1)

Idea: Consider the paths between $x$ and $y$. If we prove that each step in the path decreases in expectation by a factor $\beta$, we get inequality (1). Hence we have to prove that if $D(x, y) = 1$, then

$$E[D(\hat{x}, \hat{y}) | x, y] \leq \beta$$

Figure (2) illustrates this idea.

Assume $y$ is obtained from $x$ by flipping the variable at $i$. We define the coupling in detail as follows:

- Pick a vertex $j$ same for the two system
- Pick $x' \in \{0, 1\}$ same for the two system
- Let $\pi(x)$ and $\pi(y)$ as defined before and draw $W \in [0, 1]$.
- Set $\hat{x}_j = x'$ if $W \leq \pi(x)$ and $\hat{y}_j = x'$ if $W \leq \pi(y)$.

Let us assume for the sake of analysis that $G$ has uniform degree $k$, and compute $E[D(\hat{x}, \hat{y})]$. We consider three cases:
If \( j \neq i, j \sim i \) (with probability \( 1 - \frac{k+1}{n} \)) then \( D(\hat{x}, \hat{y}) = D(x, y) = 1 \).

If \( j \neq i, j \sim i \) (with probability \( \frac{k}{n} \)) then
\[
D(\hat{x}, \hat{y}) = 2 \text{ with probability } \alpha = |\pi_j(\hat{x}) - \pi_j(\hat{y})|,
\]
\( D(\hat{x}, \hat{y}) = 1 \) otherwise.

If \( j = i \) (with probability \( \frac{1}{n} \)) then
\[
D(\hat{x}, \hat{y}) = 1 \text{ with probability } \gamma = |\pi_j(\hat{x}) - \pi_j(\hat{y})|,
\]
\( D(\hat{x}, \hat{y}) = 0 \) otherwise.

We now compute a worst-case upper bound for \( \alpha \). A filled circle in the pictures indicates \( x_i = 1 \) and an unfilled circle indicates \( x_i = 0 \).

In figure 3, \( x_j \) has to be 0 since \( y_i = 0 \).

\[
\pi_j(\hat{x}) = \min(1, \lambda^{1^{-0}}) = \lambda, \text{ assuming } \lambda < 1
\]
Figure 3: The picture for computing $\alpha$

$\alpha \leq \lambda$

Hence $\alpha \leq \lambda$.

Similarly, we compute $\gamma$.

Figure 4: The picture for computing $\gamma$

- Case 1: $x'_i = 1$.
  $\pi_j(x) = \min(1, \lambda^{1-0}) = \lambda$, assuming $\lambda < 1$.
  $\pi_j(y) = 1$.

- Case 2: $x'_i = 0$.
  $\pi_j(x) = 1$.
  $\pi_j(y) = \min(1, \lambda^{0-1}) = 1$, assuming $\lambda < 1$.

Hence

$\gamma \leq \frac{1}{2}(1 - \lambda) + \frac{1}{2} \cdot 0 \leq \frac{1}{2}$
Together,

\[ \mathbb{E} D(\hat{x}, \hat{y}) \leq 1 \cdot \left( 1 - \frac{k + 1}{n} \right) + \frac{k}{n} (1 - \lambda + 2\lambda) + \frac{1}{n} \cdot \frac{1}{2} \]

\[ = 1 - \frac{1}{2n} + \frac{k\lambda}{n} \]

Hence

\[ \beta \leq 1 - \frac{1}{2n} (1 - 2k\lambda) \]

Which implies, for \( \lambda < \frac{1}{2k} \), the chain is rapidly mixing.

Note that we were very lousy in computing \( \alpha \), we could have got \( \lambda < \frac{1}{k} \).