## Lect4: Exact Sampling Techniques and MCMC Convergence Analysis

1. Exact sampling
2. Convergence analysis of MCMC
3. First-hit time analysis for MCMC--ways to analyze the proposals.



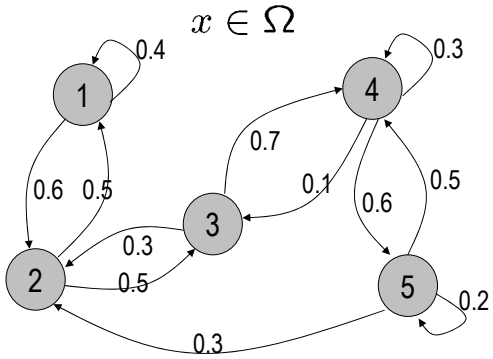
## Outline of the Module

- Definitions and terminologies.
- Exact sampling techniques
- Convergence rate and bounds using eigen-based analysis.
- First hitting time analysis: ways to analyze the proposals.

# A Toy Example

$$(\Omega, K, p_0)$$

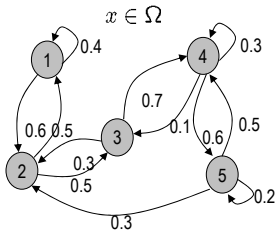
1. State space. 2. Transition kernel. 3. Initial status.




$$K = \begin{pmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.0 & 0.0 & 0.1 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.0 & 0.5 & 0.2 \end{pmatrix}$$



# Target Distribution




$$\lim_{n \to \infty} p_o K^n \to \pi$$

| year | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 2 | 0.4 | 0.6 | 0.0 | 0.0 | 0.0 | | 0.0 | 0.3 | 0.0 | 0.7 | 0.0 |
| 3 | 0.46 | 0.24 | 0.30 | 0.0 | 0.0 | | 0.15 | 0.0 | 0.22 | 0.21 | 0.42 |
| 4 | … | | | | | | … | | | | |
| 5 | | | | | | | | | | | |
| 6 | 0.23 | 0.21 | 0.16 | 0.21 | 0.17 | | 0.17 | 0.16 | 0.16 | 0.26 | 0.25 |
| | 0.17 | 0.20 | 0.13 | 0.28 | 0.21 | | 0.17 | 0.20 | 0.13 | 0.28 | 0.21 |

# Communication Class

A state j is said to be *accessible* from state i if there exists M such $K_{ij}(M)>0$

$$i \longrightarrow j \qquad K_{ij}(M) = \sum_{i_1,\ldots,i_{M-1}} K_{ii_1}\ldots K_{i_{M-1}j} \quad K_{ij}(M) > 0$$

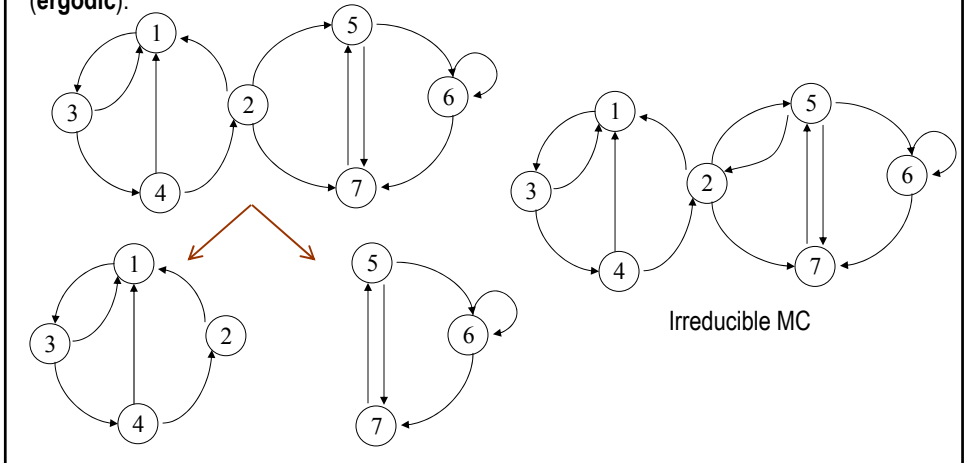
$$i \longleftrightarrow j \qquad i \text{ and } j \text{ are accessible to each other}$$

Communication relation $\longleftrightarrow$ generates a partition of the sate space into disjoint equivalence classes called ***communication classes***.



# Irreducible

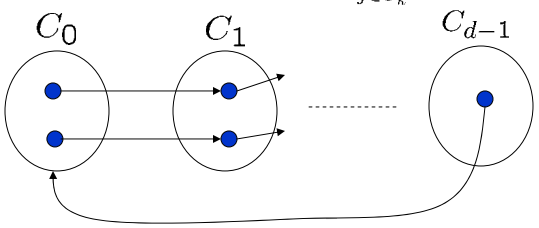
If there exists only one communication class then we call its transition graph to be irreducible (**ergodic**).



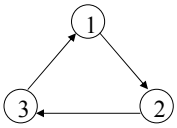
# Periodic Markov Chain

For any irreducible Markov chain, one can find a unique partition of graph G into d classes:

$$C_0, ..., C_{d-1}, \quad i \in C_k \qquad \sum_{j \in C_k} K_{ij} = 1$$




An example:




$$K = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

The Markov Chain has period 3 and it alternates at three distributions:

$$(1\ 0\ 0) \rightarrow (0\ 1\ 0) \rightarrow (0\ 0\ 1)$$



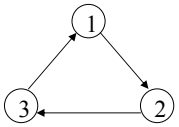
# Stationary Distribution

$$\pi = \pi K$$

There maybe many stationary distributions w.r.t K.

Even there is a stationary distribution, Markov chain may not always converge to it.

$$\pi = (\tfrac{1}{3}\ \tfrac{1}{3}\ \tfrac{1}{3}) \qquad K = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \qquad (\tfrac{1}{3}\ \tfrac{1}{3}\ \tfrac{1}{3}) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = (\tfrac{1}{3}\ \tfrac{1}{3}\ \tfrac{1}{3})$$




$$(1\ 0\ 0) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = (0\ 1\ 0)$$

## Burn-in Time

$$(\Omega, K, p_0)$$

**Burn-in time**:

The initial convergence time is called the "burn-in" time.

$$x \sim \pi(x)$$

This measures how quickly a Markov chain is not biased by the starting point $x_0$. $p_0 = (0, ..., \underset{x_0}{1}, ..., 0)$

**Mixing rate**:

It measures how fast Markov chain convergences.



## MCMC Design

In general, for a given target distribution $\pi$, we want to design a **irreducible**, **aperiodic** Markov chain which has low **burn-in period** and **mixes fast**.

$$p_0 K^n \rightarrow \pi$$

$$x \sim \pi(x)$$

Ideally, ***x*** should be as i.i.d as possible.

## Outline

- Definitions and terminologies.
- Exact sampling techniques.
- Convergence rate and bounds using eign-based analysis.
- First hitting time analysis: ways to analyze the proposals.



## Exact Sampling

A natural and general question we want to ask is:

When do we want to stop a MC?

But how long is long enough?

# Exact Sampling (literature)

**Exact (perfect) sampling is a new technique.**

J. Propp and D. Wilson, 1996, "**Exact sampling** with coupled Markov chains and applications to statistical mechanics", *Random Structures and Algorithms*, 9:223-252.

W. Kendall, 1998, "Perfect simulation for the area-interaction point process", Probability Towards 2000, pp.218~234.

J. Fill, 1998, "An interruptible algorithm for exact sampling via Markov chains", *Ann. Applied Prob.*, 8:131-162.

Casella et al. 1999, "Perfect slice samplers for mixtures of distributions", *Technical Report BU-1453-M*, Dept. of Biometrics, Cornell University.

L. Breyer and G. Roberts, "Catalytic perfect simulation", Technical Report, Dept. of Statistics, Univ. of Lancaster.

...

**Introduction web:** http://dbwilson.com/exact/



# Coupling

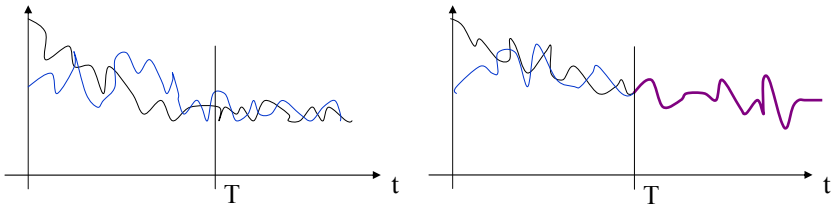
**Definition:** Coupling

We say that two chains are coupled if they use the same sequence of random numbers from the transitions.

Define **deterministic update function** $x^{t+1} = \phi(x^t, U^{t+1})$ where $U$ is iid from a **fixed** distribution. $x \in \Omega = \{x_1, x_2, ..., x_N\}$

MCs are then coupled.

# Coupling from the Past




The chance of two MC meeting at T is $p(X^T(x_1)) \cdot p(X^T(x_2)))$ if they are not coupled.

$$X^t(t, x_1) = \phi_t \circ ... \circ \phi_1(x_1)$$
$$X^t(t, x_2) = \phi_t \circ ... \circ \phi_1(x_2)$$
$$X^T(T, x_1) = X^T(T, x_2)$$

If the two MC coalesce at any time t, they become identical forever after.



# Coupling from the Past (CFTP)

1. Set the starting value for the time to go back, $T_0 \leftarrow -1$.

2. Generate a random vector $U^{T_0+1}$

3. Start a chain in each state $x_m, m = 1, ..., N \;\; at \;\; T_0$

   and run the chains:

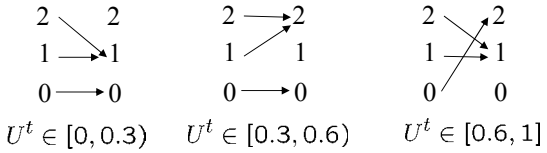
$$X^{t+1}(T_0, x_m) = \phi(x^t, U^{t+1}) \; to \; time \; 0, t = T_0, T_0+1, ..., -1$$

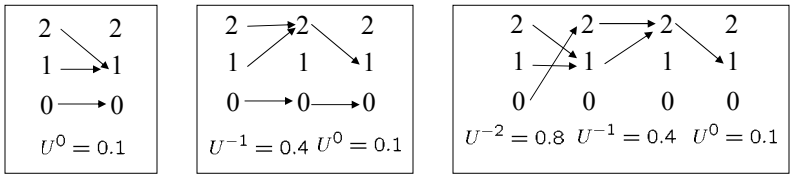
4. Check for coalescence at time 0. If so, common value $X^0$ is returned. Otherwise let $T_0 \leftarrow (T_0 - 1)$ and repeat 2.

# An Example

$$x \in \Omega = \{0, 1, 2\} \qquad x_{t+1} = \phi(x_t, U^t)$$

**Define:**






# Convergence

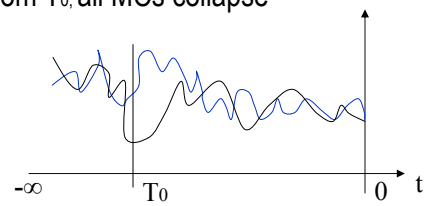
Propp and Wilson's Algorithm:

The algorithm produces a random variable distributed exactly according to the stationary distribution of the Markov chain.

Detailed proof see Propp and Wilson 1996.

To understand: Since for any states $x_i$, from $T_0$, all MCs collapse at time 0.

$$x_{-\infty} \rightarrow x_0, \ then \ \ x_0 \sim \pi$$




Traditional forward MCMC can not guarantee this!

## Computational Issue with CFTP

1. Do we need to check for each $T_0$?

**No!**

$$For\ example\ \ T_i = -2^i.$$

2. What if the state space of x is very big?

**Monotone CFTP**



## Montonicity and Envelopes

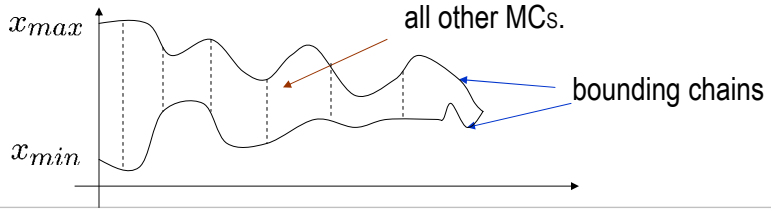
Coupling from the past (CFTP) is a nice theory but it applies only to a finite state space with a manageable number of points.

**Monotonicity structure**:

There exists an ordering structure $\preceq$ on the space $\Omega$:

$$x_1 \preceq x_2,\ \phi_t(x_1) \preceq \phi_t(x_2), for\ \forall\ t.$$

We only need to run Markov chains from $x_{min}\ and\ x_{max}$.



## Perfect Slice Sampling

The slice sampling transition can be coupled in order to respect the natural ordering.

$$u \sim [0, \pi(x)]$$
$$x \sim \{x : \pi(x) \geq u\}$$

$$\pi(x_1) \leq \pi(x_2)$$

$$A_2 = \{x : \pi(x) \geq u\pi(x_2)\} \subset A_1 = \{x : \pi(x) \geq u\pi(x_1)\}$$

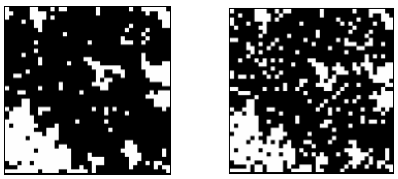
Similar to the Propp and Wilson's algorithm, we can have a **perfect monotone slice sampler**:

There is coalescence when $\pi(x^0_{-t}) \geq u_{-t}\pi(x^1_{-t-1})$



## An Example

Image restoration using CFTP (We know when to stop).

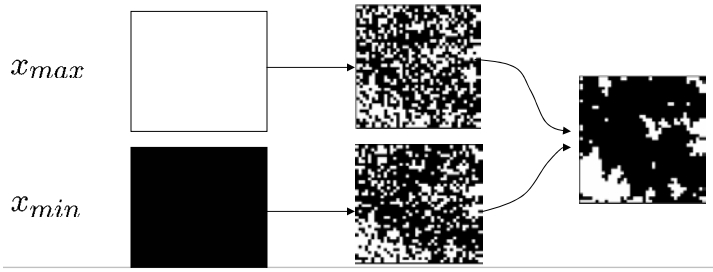


true image



observed image

$$\pi(x|y) \propto exp\{\beta \sum x_i x_j + log((1-e)/e)/2 \sum x_j y_j\}$$



## Some Other Methods

1. Kac's perfect sampling. (Murdoch and Green 1998).
2. Automtic coupling. (Breyer and Roberts 2000).
3. Forward perfect sampling. (Fill 1998).
4. …..

This is a new direction and has many potential promises for MCMC convergence analysis.



## Outline of the Module

- Exact sampling techniques
- Some definitions of MCMC.
- Convergence rate and bounds using eigen-based analysis.
- First hitting time analysis: ways to analyze the proposals.

# MCMC Convergence

CFTP: **I will let you know when to stop once we get there, but I can not tell you how long it will take in advance.**

How long will a MC converge?

A basic MCMC consists of three key components:

$$(\Omega, K, p_0) \quad p_0 K^n \rightarrow \pi$$

How to estimate n?



# Convergence Analysis (literature)

F. Gantmacher, 1995, "Application of the Theory of Matrices", Inter Science, New York, .

M. Jerrum and A. Sinclair, 1989 "Approximating the permanent", SIAM Journal of Computing, pp.1149-1178.

J.A. Fill, 1991, "Eigenvalue bounds on convergence to stationarity for non-reversible Markov chains", The Annals of Applied Porbability.

P. Diaconis and J.A. Fill, 1996, "Strong stationary times via a new form of duality", The Annals of Probability, p. 1483-1522.

P. Bremaud, 1999, "Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues", Springer.

J. Liu, 2000, "Monte Carlo Strategies in Scientific Computing", Springer.

R. Maciuca and S.C. Zhu, "First-hitting-time Analysis of Independence Metropolis Sampler", Journal of Theoretical Probability, 2005.

…

## Perron-Frobenius Theorem

For any primitive stochastic matrix K, K has eigen-values

$$\lambda_1 \geq |\lambda_2| \geq \cdots > |\lambda_r|$$

Each eigen-value has left and right eigen-vectors $(\mu_i, \nu_i)$

$$K^n = \lambda_1^n v_1 u_1^T + O(n^{m_2 - 1}|\lambda_2|^n)$$

m2 is the algebraic multiplicity of $l_2$, i.e. $m_2$ eigen-values that have the same modulus.



## Perron-Frobenius Theorem

$$K^n = \lambda_1^n v_1 u_1^T + O(n^{m_2 - 1}|\lambda_2|^n)$$

If K is irreducible with period *d*>1, then there are exactly d distinct eigen-values of modulus 1, namely the *d*th roots of unity, and all other eigen-values have modulus strictly less than 1.

For d=1: $\lambda_1 = 1, \nu_1 = 1, \mu_1 = \pi$

Rate of convergence is decided by $\lambda_{SLEM} = |\lambda_2|$

## Markov Design

Given a target distribution π, we want to design an irreducible and aperiodic K

$$\pi K = \pi \quad \text{and } K \text{ has small } \lambda_{SLEM}$$

The easiest would be: $K = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix}$ then any $pK = \pi \quad \lambda_{SLEM}(K) = 0$

But in general *x* is in a big space and we don't know the landscape of π, though we can compute each π(x).



## Necessary and Sufficient Conditions for Convergence

**Irreducible (ergodic):**

$$\forall i \leftrightarrow j, K_{ij}(M) > 0 \;\; and \;\; K_{ji}(M) > 0$$

**Detailed Balance:** $\pi(i)K_{ij} = \pi(j)K_{ji}$

Detailed balance implies stationarity:

$$\pi K = \sum_i \pi(i)K_i = \sum_i \pi(i)(K_{i1}, ..., K_{in})$$

$$= \sum_i (\pi(j)K_{1i}, ..., \pi(n)K_{ni}) = \pi$$

## Choice of K

**Markov Chain Design**:

(1) K is irreducible (egordic).

(2) K is aperiodic (with only one eigen-value to be 1).

(3) $p(i)K_{ij} = p(j)K_{ji}$

There are almost infinite number of ways to construct K given a $\boldsymbol{\pi}$.

r equations with unknowns r x r unknowns

Different Ks have different performances.



## Convergence Rate

$The\ \frac{1}{\pi}\ Bound$

For any initial distribution p:

$$||pK^n - \pi||_{\frac{1}{\pi}} \leq \lambda_{SLEM}^n ||p - \pi||_{\frac{1}{\pi}}$$

In particular, if we start from a specific state

$$||K^n(i_0, .) - \pi||_{TV} \leq \sqrt{\frac{1-\pi(i_0)}{\pi(i_0)}} \lambda_{SLEM}^n$$

The convergence rate depends on :

(1) The second largest eigen-value modulus.

(2) The initial state.

# Bounds of Second Largest Eigen-value Modulus

We see that the convergence of MCMC is mostly decided by the second largest eigen-value modulus from a couple of theorems.

How do we connect the second largest eigen-value modulus to our algorithm design?

Jerrum and Sinclair's theorem:

$$1 - 2\Psi(K) \leq \lambda_2 \leq 1 - \frac{\Psi(K)^2}{2}$$

$\Psi(K)$ is the conductance of the transition matrix K.



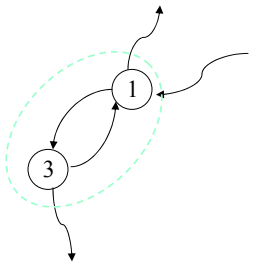
# Conductance

For a nonempty set $B \subset E = \{1, ..., r\}$, the **capacity** of B

$$\pi(B) = \sum_{i \in B} \pi(i)$$

The **ergodic flow** out of B,

$$F(B) = \sum_{i \in B, j \in \bar{B}} \pi(i) k_{ij}$$




Define:

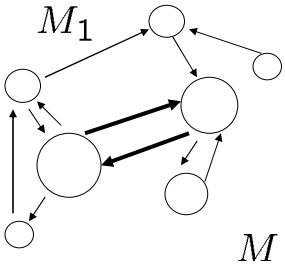
$$\Psi(B, K) = \frac{F(B)}{\pi(B)}$$

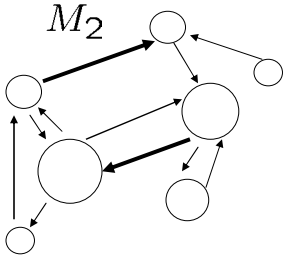
The **conductance** of (K, p):

$$\Psi(K) = inf(\Psi(B, K); 0 < |B| < r, \pi(B) \leq \tfrac{1}{2})$$

**It is the bottleneck of the transition graph!**

## Intuition




$$M = K \cdot K$$

$$\Psi(B, M) = \frac{F(B)}{p(B)} \quad \Psi(M) = inf(\Psi(B, M); 0 < |B| < r, p(B) \leq \tfrac{1}{2})$$

$$1 - 2\Psi(M) \leq \lambda_{SLEM} \leq 1 - \frac{\Psi(M)^2}{2}$$

**This is analogous to traffic network design. To put major resources on big populations.**

**Problem: (1) We still do not know what is an optimal design strategy.**

**(2) Any small probability mass matters.**



## Outline of the Module

- Definitions and terminologies.
- Exact sampling techniques.
- Convergence rate and bounds using on eigen-based analysis.
- First hitting time analysis: ways to analyze the proposals.

## Metropolis-Hastings Algorithm

Detailed balance: $\pi(i)K(i,j) = \pi(j)K(j,i)$

Metropolis-Hastings:

$$\underbrace{K(i,j)}_{transition\ probability} = \underbrace{Q(i,j)}_{proposal} \cdot \underbrace{\alpha(i,j)}_{acceptance\ rate}$$

$$\alpha(i,j) = min\Big(1, \underbrace{\frac{Q(j,i)}{Q(i,j)}}_{proposal} \underbrace{\frac{\pi(j)}{\pi(i)}}_{verification}\Big)$$

The previous convergence analysis in terms of **K** still applies.

But we want to know its behavior w.r.t. **Q**.



## MCMC Convergence w.r.t. KL Divergence

Suppose we are not limited by a fixed kernel K (inhomogeneous),

$$\pi(i)K_\alpha(i,j) = \pi(j)K_\alpha(i,j)$$

The Markov chain is monotonically approaching to the target distribution.

$$KL(\pi||p) = \sum_i \pi(i) \log \frac{\pi(i)}{p(i)}$$

Let $p_t$ be the distribution at t, and $p_{t+1} = p_t K_\alpha$

$$KL(\pi||p_t) - KL(\pi||p_{t+1}) \geq 0$$

## Why is it working?

Detailed balance is satisfied (easy to check!). Therefore, $\pi$ is the stationary distribution of *K*.

The unspecified part of Metropolis-Hastings algorithm is *Q*, the choice of which determines, if the Markov chain is ergodic.

The choice of *Q* is problem specific.



## Independent Metropolis Sampler (IMS):

$$Q(i,j) = q(j)$$

This implies that each move does not depend on the current state. This is probably the simplest case in MCMC.

$\lambda_{SLEM}$ can be computed analytically.

$$\lambda_i = \sum_{k \geq i} (q(k) - \pi(k)w(i)), 1 \leq i \leq N-1, \lambda_0 = 1$$

where w(i) = q(i)/π(i), are sorted increasingly, w(1)≤w(2)≤…≤w(N).

$$\lambda_{SLEM} = 1 - w(1)$$

## Convergence of IMS

$$\lambda_{SLEM} = 1 - w(1), \quad w(1) = \frac{q(1)}{\pi(1)}$$

The convergence depends on the smallest value of q(i)/π(i).

This is consistent with the previous conductance analysis (bottleneck).

But it doesn't sound right. What if π(1) is extremely small and negligible.

The problem is due to the worst case analysis!



## What is the Alternative?

**Worst Case v.s. Average Case**

- Assume we are interested in a particular state (the mode of some distribution for instance) → search problems.
- One can ask, how fast will the algorithm hit x*, in average → *average case analysis*.
- This can be much quicker than the total convergence time → *worst case scenario*!

## First Hitting Times

- Let $\Omega=\{1,2,\ldots,N\}$ the state space of a finite Markov chain $\{X_n\}_{n\geq 0}$
- The first hitting time (f.h.t) of $i \in \Omega$ is defined to be the number of steps for reaching $i$ for the first time :

$$\tau(i)= \min\{n\geq 0 \mid X_n=i\}$$

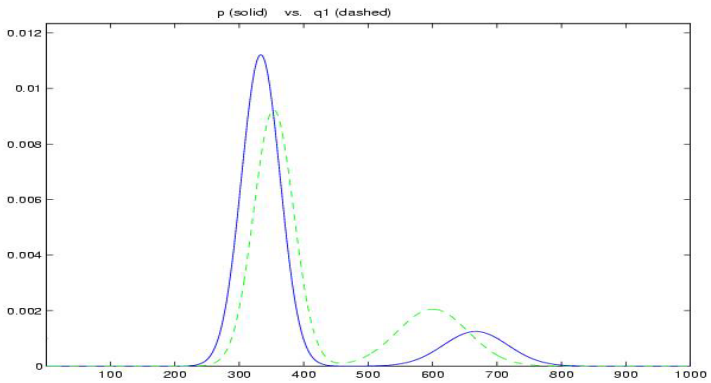
E[τ(i)]- often more relevant than the time to converge to equilibrium (mixing time).



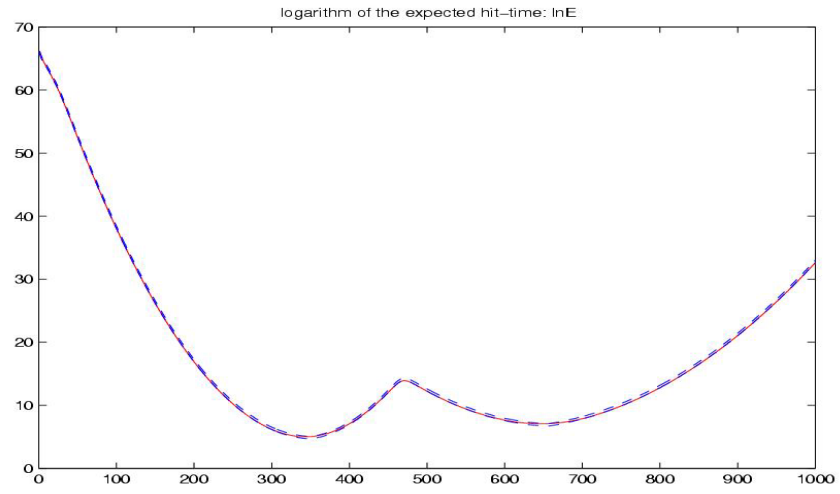
## Bounds

$$max\{\frac{1}{\pi(i)}, \frac{1}{q(i)}\} \leq E[\tau(i)] \leq max\{\frac{1}{\pi(i)}, \frac{1}{q(i)}\}\frac{1}{1-||\pi - q||_{TV}}$$

Example: π, q are mixtures of gaussians with N=1000 states.

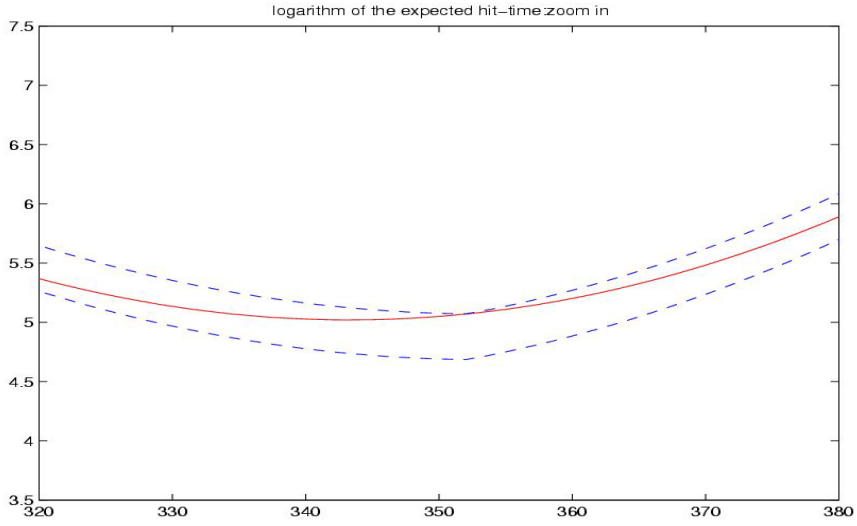


# Plot of the Expectation with Bounds






# Zoom in around the mode



## Ideally, $q=\pi$

Three types of states:

(1) i is said to be over-informed if q(i)>π(i).

(2) i is said to be under-informed if q(i)<π(i).

(3) i is said to be exactly-informed if q(i)=π(i).



## Equality Cases

(1) $E[\tau(i)] = \frac{1}{\pi(i)}$, at the most informed state $i_{max}$

(2) $E[\tau(i)] = \frac{1}{q(i)}$, at the least informed state $i_{min}$

(3) $E[\tau(i)] = \frac{1}{\pi(i)||\pi - q||_{TV}}$,

at the exactly informed states.

## Take-home Messages of MCMC Convergence

1. MCMC in general converges to the target distribution $\pi$.
2. Exact sampling is a technique telling us when it converges. (But we don't know how to measure it.)
3. Eigen-based analysis gives us bounds on the convergence. (But it is based on the worst-case scenario.)
4. First-hitting time analysis on IMS gives us intuitive ideas about algorithm design. (We still need to remove the independence assumptions).
5. ????