Thesis Presentation

Markov Chain Monte Carlo (MCMC) Algorithms and Convergence Results

Student: Yufan Li

Supervisor: Jeffery Rosenthal

Before the presentation...

- I will skip some slides that are included for completeness. You can ask me questions about these during Q&A.
- Only outline/general ideas of my proof will be presented; we will not discuss technical details

My Thesis

- I focused on studying theoretical properties (convergence) of MCMC algorithms; there was no coding or experiments involved. But I have carefully studied a non-trivial problem concerning asymptotic behavior of Adaptive Gibbs Sampler and have done extensive expository studies in MCMC literature
- Expository studies:
  - Discrete-time, general-state space Markov Chain following Meyn and Tweedie (1997) and Rosenthal's review papers on Markov chain and MCMC
  - Adaptive MCMC: convergence conditions for adaptive MCMC, adversarial adaptive MCMC, adaptive Gibbs sampler, optimal scaling and other papers
  - Quantitative convergence rates: general framework and proofs (Rosenthal 1995) and application on James Estimator (Rosenthal 1998)
  - Weak convergence of Random Walk Metropolis to Langevin diffusion; required theories regarding weak convergence of stochastic processes following Ethier and Kurtz (2009) and other sources; Complexity bounds via diffusion limit

Cont. My Thesis

- Original Studies:
  - Strengthened results regarding “stairway to heaven” example in Latuszynki and Rosenthal (2007).
    - Proof by construction of countably many phases—more intuitive
    - Proof by auxiliary process, similar to the original paper but with more careful bounds
  - Tried to apply Rosenthal (1995)’s quantitative rates results to an MH example. Only had results on a trivial case.

Presentation Objective and Schedule

- A non-technical, intuitive, concise presentation of my major original work:
  - An alternative proof of “stairway to heaven problem/conjecture” proposed in the following paper: *Adaptive Gibbs samplers and related MCMC methods. The Annals of Applied Probability, 23(1), pp.66-98.*
- An overview of the topics I have studied:
  - What is Markov Chain Monte Carlo?
  - Why ”convergence speed” matters?
  - What are the existing approach to measure convergence speed?

What is Markov Chain Monte Carlo (MCMC)? Why we need it?

- It is algorithms, meaning that it can produce sample points given a (potentially highly complicated) probability distribution.
- “But we can always use excel to sample from a distribution right?” e.g. NOMINV(Rand(), mean, Std). Why MCMC?
- From a practical point of view**, MCMC is necessary because often times an explicit expression for inverse of the cdf. Such as (NOMINV) is not easy to find**. E.g. high dimensional integral
- So MCMC can be used to estimate complicated integral as well

Example: Parameter estimation for Hierarchical Bayesian models

$$Y_i|\theta_i \sim N(\theta_i, V), 1 \leq i \leq K$$

$$\theta_i|\mu, A \sim N(\mu, A), 1 \leq i \leq K$$

$$\mu \sim \text{flat prior on } R$$

$$A \sim IG(a, b)$$

*K can be very large. And the model can be more complicated.

Parameters distribution (posterior, conditional) for previous model—goal is to sample these!

$$\theta_i^{(k)} \sim L(\theta_i | A = A^{(k)}, \mu = \mu^{(k-1)}, Y_i) = N(\frac{\mu^{(k)} V + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)} V}{V + A^{(k)}})$$

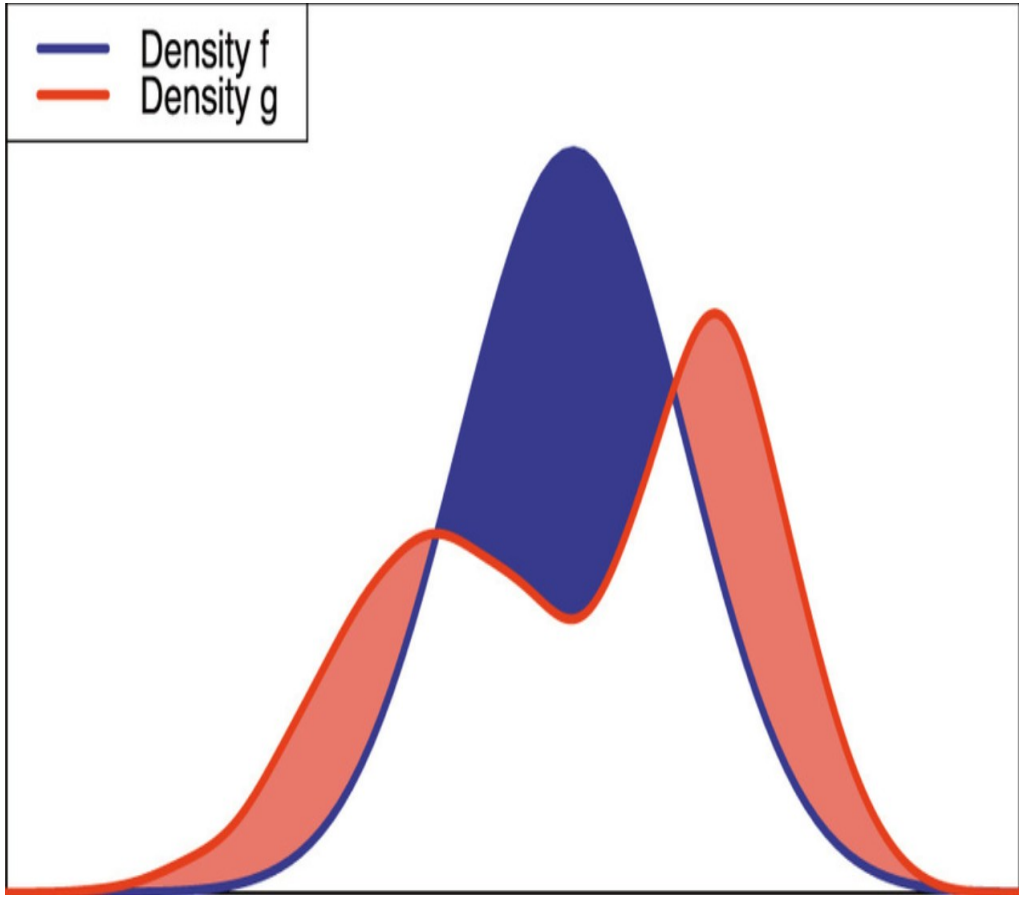
$$A^{(k)} \sim L(A | \theta_i = \theta_i^{(k-1)}, Y_i) = IG((a + \frac{K-1}{2}), b + \frac{1}{2} \sum (\theta_i^{(k-1)} - \bar{\theta}^{(k-1)})^2);$$

$$\mu^{(k)} \sim L(\mu | A = A^{(k)}, \theta_i = \theta_i^{(k-1)}, Y_i) = N(\bar{\theta}^{(k-1)}, \frac{A^{(k)}}{K})$$

*Intuition: Fix everything else but one parameter, then sample via conditional distribution above (Each parameter here is a dimension). So we “take a step” on one direction at a time. This is Gibbs Sampler!

How and Why MCMC work?

- MCMC is an approximation to direct sampling
- The foundation of MCMC is convergence: unconditional distribution of MCMC at step N where N is large should be “close” to target distribution;
- We make sure it is indeed the case. Then run MCMC long enough (how long?)



Convergence Is Important/Necessary

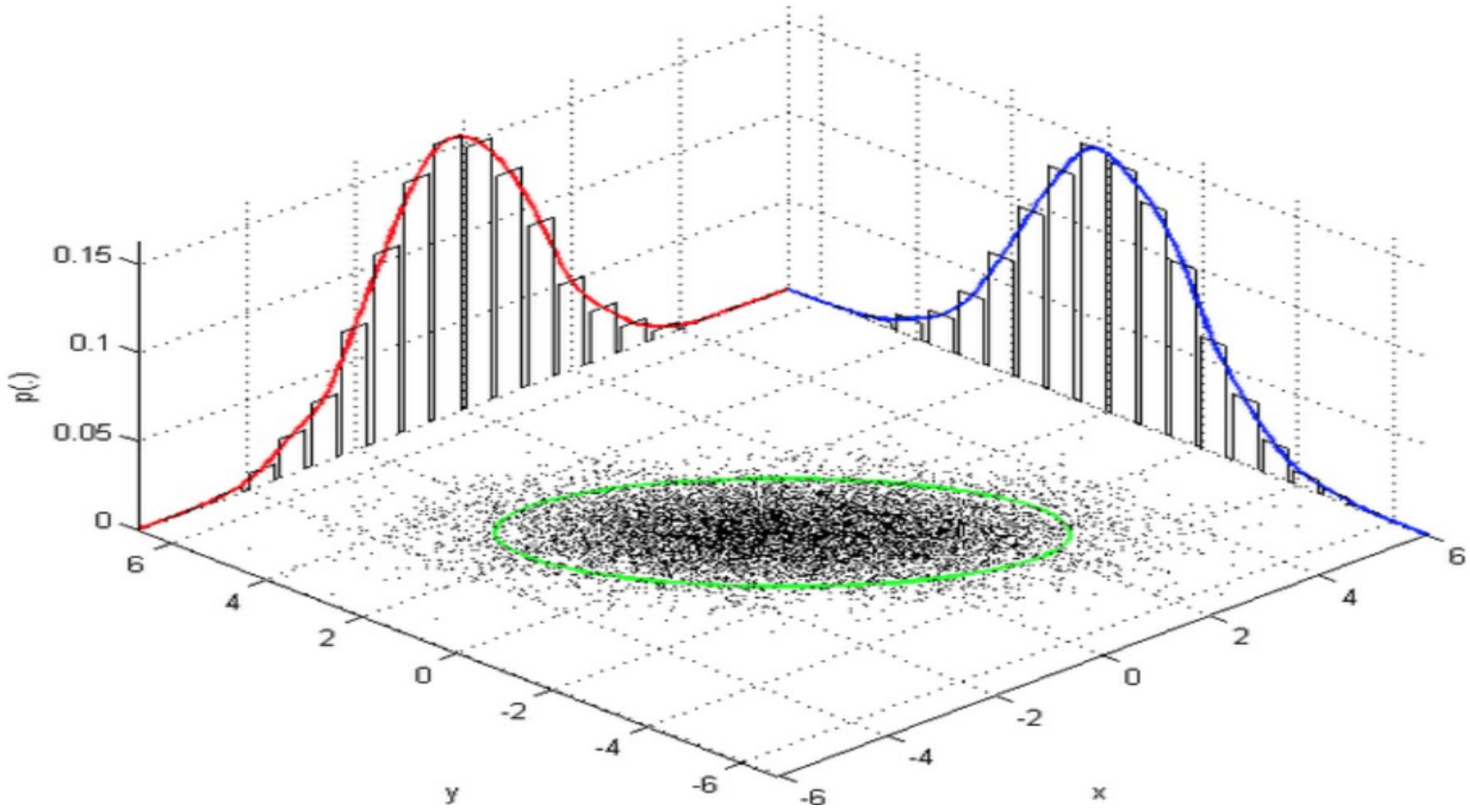
- The sample acquired from MCMC is “correct/accurate” only if
  - we run the algorithm long enough, i.e. collect sufficient amount of sample points (Law of large numbers for MCMC)
  - MCMC used converges to target distribution
- In slightly technical terms, if we measure “difference” between target distribution and the distribution from MCMC (dist at s *ance distance*

$$\lim_{n\to\infty} ||P^n(x,\cdot) - \pi(\cdot)|| = 0$$

Simpler, more standard MCMC

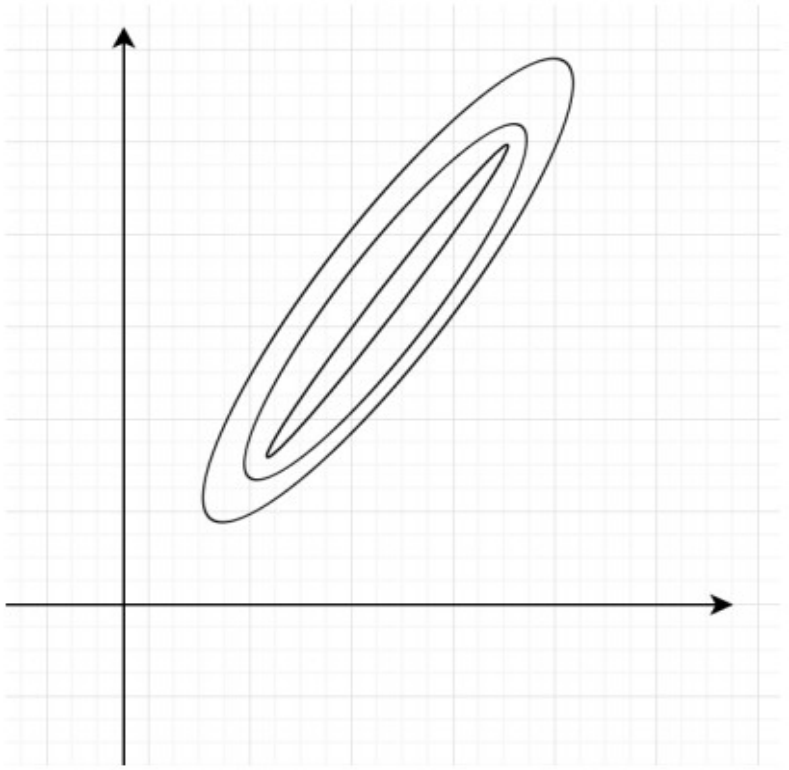
- **Gibbs sampler (we will focus on this)**
  - Used when conditional distributions are simple e.g. Bayesian models
  - Sequential sampling
  - Randomly choose a dimension
- Metropolis Hastings (MH):
  - Propose via a known distribution (e.g. Random Walk with certain step size)
  - Reject or Accept (depending on target distribution)
- They are designed to be convergent. You don’t have to know any math or check anything – just follow the instructions/recipes

Gibbs sampler



So why we are still studying MCMC: Standard MCMC is not always efficient

- Because standard method such as MH and Gibbs samplers are not always optimal to use– can be very slow: think of the following (slanted, elongated distribution). How would a Gibbs sampler behave
- What “step size” should you choose for MH?



Adaptive MCMC

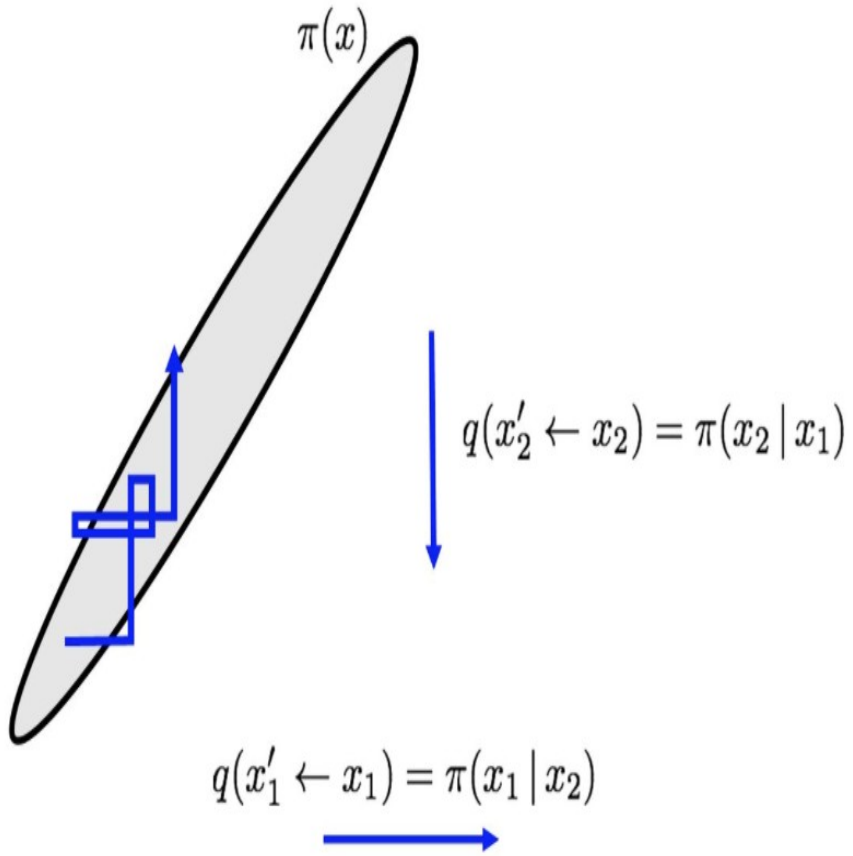
- **MCMC that can learn from sample path so far to figure out optimal parameters**, including orientation of coordinate systems, preference for coordinate to update, step size—can be anything
- Examples of Adaptive models
  - Directional sampling: rotating axis
  - Adaptive Gibbs: adapt dimension selection probability
  - Adaptive MH: scale the “step size” such that acceptance rate is 0.23

How do we know if Adaptive MCMC converges?

- There is no “recipe” anymore! Designers have to figure out themselves.
- They may not be Markovian (next step does not entirely depend on current location but may be entire history of the chain). Typical Markov chain theory cannot be applied directly.
- Rosenthal and Roberts (2001) gave **two general conditions (proved via coupling)**:
  - Containment Condition: all your adaptation kernels must converge to target distribution in a “somewhat” uniform fashion
  - Diminishing Adaptation: you must adapt “less and less”

Downside of Adaptive Algorithms: Need to show Convergence

- Consider a random scan Gibbs sampler where we adapt on the probability of selecting each coordinate to update.
- Apply the two general conditions. Latuszynki and Rosenthal (2007) manages to show the following convergence conditions:



Adaptive Random Scan Gibbs sampler

Theorem 4.1. *Let the selection probabilities $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ *in probability for fixed starting values* $x_0 \in \mathcal{X}$ *and* $\alpha_0 \in \mathcal{Y}$.

(b) there exists $\beta \in \mathcal{Y}$ *s.t.* ***RSG***(β) *is uniformly ergodic.*

Then ***AdapRSG*** *is ergodic, i.e.*

$$T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{6}$$

Moreover, if

(a') $\sup_{x_0,\alpha_0} |\alpha_n - \alpha_{n-1}| \to 0$ *in probability,*

then convergence of ***AdapRSG*** *is also uniform over all* x_0, α_0, *i.e.*

$$\sup_{x_0,\alpha_0} T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{7}$$

Simpler conditions? They are wrong and we want to refute them

Levine and Casella (2006) proposed the following conditions for adaptive random scan Gibbs sampler, which does not require condition (b) in previous slide . Latuszynki and Rosenthal (2007) is able to prove it false through a counter example (“stairway to heaven”):

(i) $\alpha_n \to \alpha$ a.s. for some fixed $\alpha \in (0,1)^d$; and

(ii) The random scan Gibbs sampler with fixed selection probabilities α induces an ergodic Markov chain with stationary distribution π.

The counter example

Example 3.1. Let $\mathbb{N} = \{1, 2, \dots\}$, and let the state space $\mathcal{X} = \{(i, j) \in \mathbb{N} \times \mathbb{N} : i = j \text{ or } i = j + 1\}$, with target distribution given by $\pi(i, j) \propto j^{-2}$. On $\mathcal{X}$, consider a class of adaptive random scan Gibbs samplers for π, as defined by Algorithm 2.2, with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i, j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j + 1, \end{cases} \tag{4}$$

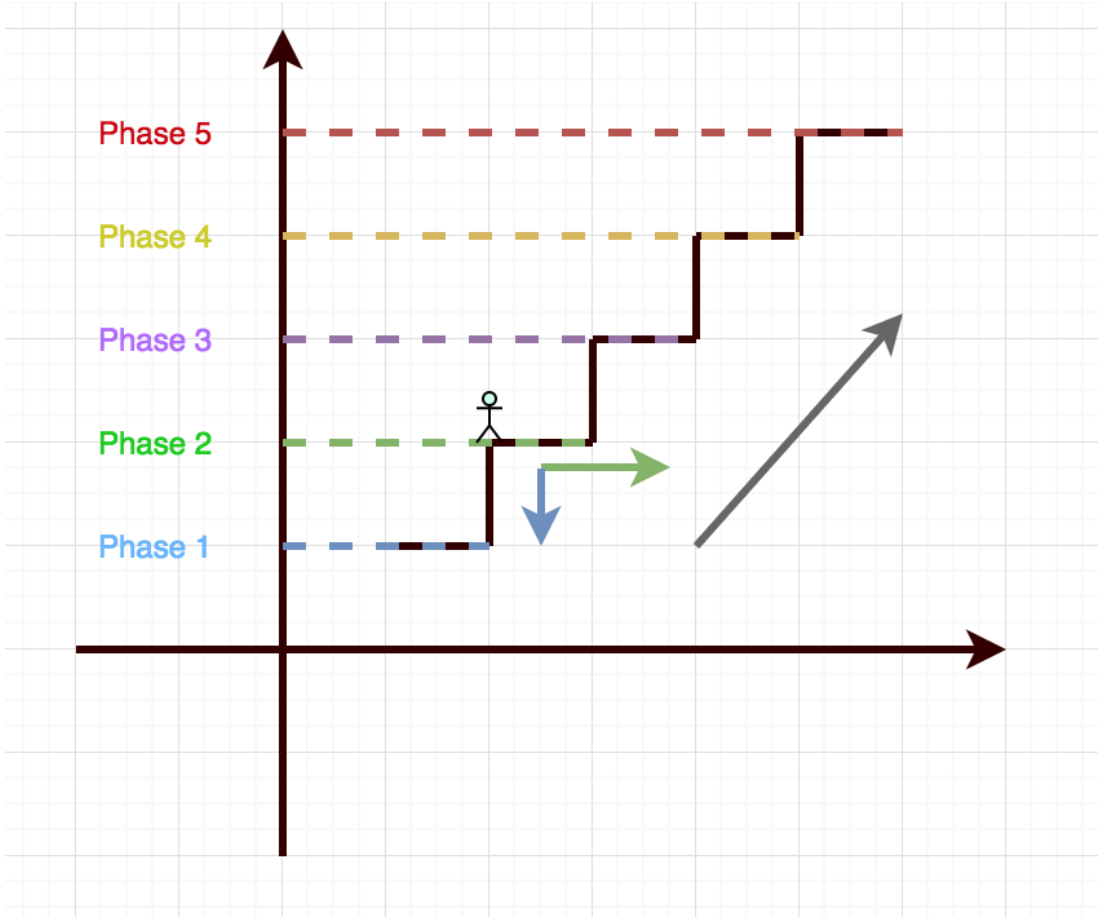
where $\{a_n\}_{n=0,1,2,\dots}$ is some sequence satisfying $a_n >= 8$

Note that this process satisfies (i) and (ii) given by Levine and Casella (2006). We only need to prove that this tends to infinite with positive probability to prove it not converging to target distribution π.

Graphically, the state space looks like a staircase:

Step 1. You will select either x or y direction by Rn.

Step 2. You will choose to stay where you are or move one step by conditional distribution.



How we move on the staircase?

After some simplification, this is equivalent to determine
(i) Move one step forward, or
(ii) Move one step backward, or
(iii) Stay
by tossing a three-way coin with the following distribution where xn is current x coordinate

$$\text{If } X_n = (x_n, x_n),$$

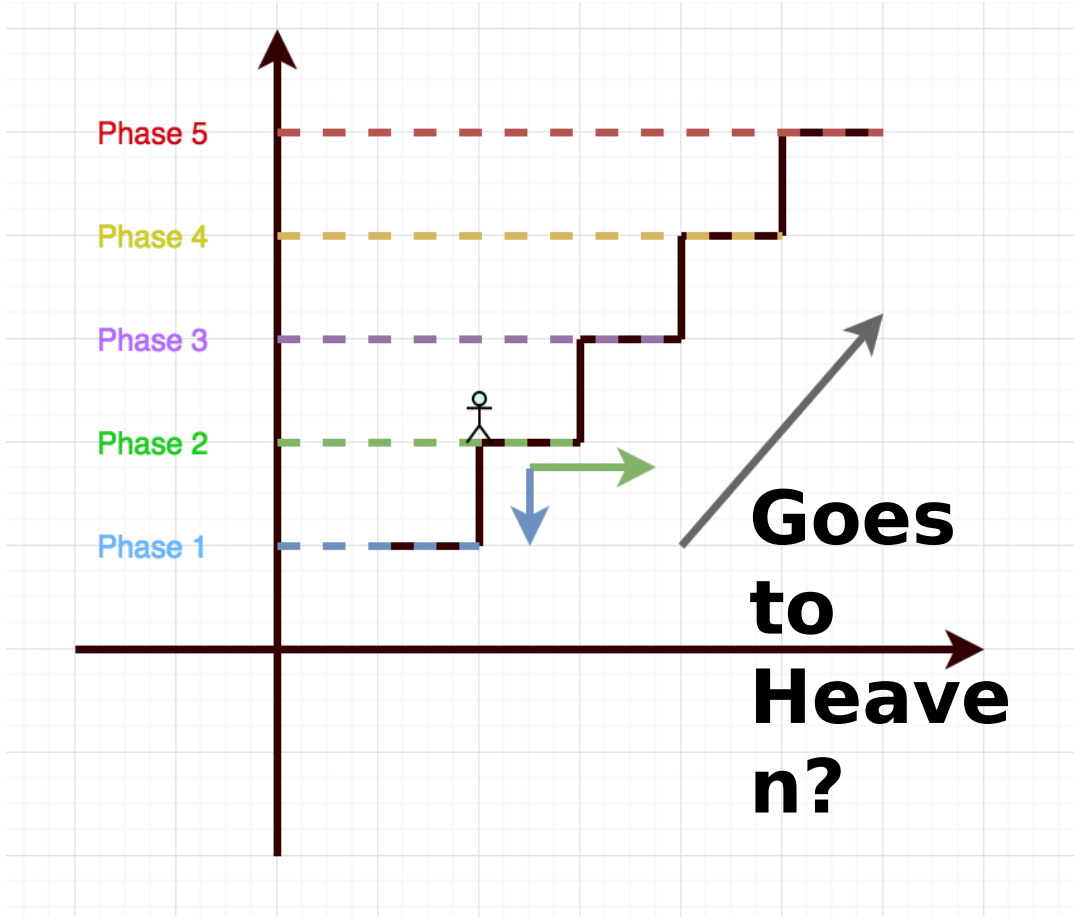
$$\left((\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n - 1)^2}, 1 - (\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n - 1)^2} - (\frac{1}{4} + \frac{2}{a_n}), \frac{1}{4} + \frac{2}{a_n} \right) \quad (59)$$

$$\text{If } X_n = (x_n, x_n - 1),$$

$$\left(\frac{1}{4} - \frac{2}{a_n}, 1 - (\frac{1}{4} - \frac{2}{a_n}) - (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n - 1)^2}{x_n^2 + (x_n - 1)^2}, (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n - 1)^2}{x_n^2 + (x_n - 1)^2} \right) \quad (60)$$

Conjecture: “Stairway to Heaven” -- yet to be proved

There exists {an} such that {an>=8} converges to infinity for which the process tends to infinity with probability 1.



Partial results

Original result: Can find a sequence {an} that tends to infinity such that the process tends to infinity with probability
larger than 0. Sufficient to refuse false claim.

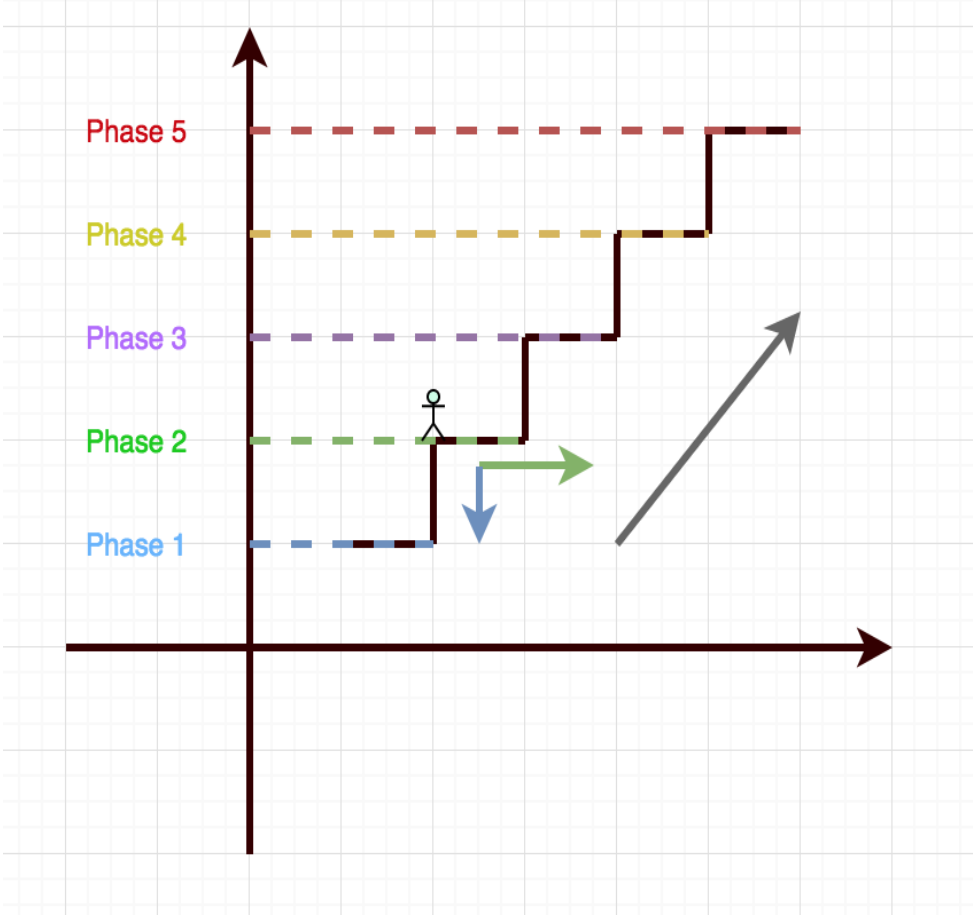
My result: Can find a sequence {an} that tends to infinity such that the process tends to infinity with any probability (fixed) less than 1. Simpler proof

Theorem 3.9. For any fixed $\sigma \in [0, 1)$, there exists a choice of $\tilde{a}_n$ such that $P(\tilde{X}_n \to \infty) > \sigma$.

Fix any natural number $K \geq 1$.

My proof by constructing phases (strategy)

- The strategy is to have countably infinitely many phases.
- Each phase is comprised of finitely many steps.
- Infinite sequence {an} is constant for each phase but increases as we enter the next phase (so that {an} still tends to infinity).
- Goal: “the little person here moves up certain amount during each phase”



Constructions: {an}

- The sequence $\{a_n\}$ is constant during each phase
- But $\{a_n\}$ converges to infinity as phase $i \to \infty$

The sequence {an} is constant during each phase
But {an} converges to infinity as phase

Definition 3.7. Define a_n as follows:

$$\begin{cases} a_n = 8, \text{ if } 0 \leq n < N_1 \\ a_n = 8\frac{2i^2+1-2i}{2i-1+0.1} - 0.001, \text{ if } N_{i-1} \leq n < N_i, \forall i \geq 2 \end{cases} \tag{48}$$

Construction: phases

Intuitively,

- M is a large number that satisfies certain technical conditions;
- Phase 1 has M steps
- Phase 2 has M-2 steps
- Phase 3 has M steps
- Phase 4 has M+2 steps
- Phase 5 has M+4 steps
- Phase 6 has M+6 steps
- Phase 7 has M+8 steps
- …

Definition 3.5. Fix any $M \in \mathbb{N}$ such that

$$0.01 \cdot (M-2) - 2\sqrt{K \cdot (M-2)\ln(M-2)} \geq 4$$

and

$$M > \max(M_0(\sigma), M_1(\sigma))$$

where $M_0(\sigma)$ and $M_1(\sigma)$ are finite numbers that depend only on the fixed σ. We will define $M_0(\sigma)$ and $M_1(\sigma)$ later in the proof: this is mostly for more concise presentation–$M_0(\sigma)$ and $M_1(\sigma)$ can be defined right away given σ as we will see later. M apparently exists for any fixed natural number K.

Definition 3.6. Define sequence N_i such that $N_0 = 0$, $N_1 = M_0(\sigma)$ and $N_i - Ni - 1 = M - 2 + 2(i-2), i = 2, 3, ...$

Proof based on monotone coupling technique (also called stochastic domination)

As I will show in the next slide, couple the current process to a simpler process will simplify the proof:

Theorem 3.10. The real random variable X is stochastically larger than Y if and only if there is a coupling between X, Y such that

$$P(X \geq Y) = 1$$

A technical lemma: when can we couple?

Here I proved that when “the little person” is above the i-th staircase during phase i, there is a ”coupling” to a biased i.d.d. random walk {Zn}. {Zn} is biased to move forward each step in a “homogeneous” fashion. This is important. **No matter what happens, the process should not fall below i-th stairway or everything breaks.**

Proposition 3.8. For each $i \geq 2$, if $x_n \geq i$ and $N_{i-1} \leq n < N_i$, there exists a sequence of i.d.d random variable $\{Z_i\}$ such that Z_i is stochastically smaller than $S_{n+1} - S_n$ for each $n \in [N_{i-1}, N_i)$ and $Z_{i,n}$ take value $\{-1, 0, 1\}$ and $\mathbb{E}(Z_i) \geq 0.01$

Sequence of “Omega Events”: Moving up certain amount during each phase

Omega here denotes events that are associated with each phase (we will refer to these as Omega events), that after phase i the process is above certain level $(M+4(i-1))$, **these are events we want to happen and events that will not to happen with large probability (how large, we will see) – because if all Omega events happen, the process tends to infinity.**

And we shall prove via induction that if events for all phases before i occurs, event i will occur with a sufficiently large probability.

Let's adopt the following notation:

$$\Omega_1 = \{S_{M_0} > M\}$$

$$\Omega_i = \{S_{N_i} \geq M + 4(i-1)\}$$

Proof by induction: First Inductive Lemma

If Omega events for all phases prior to i occurs, then we are guaranteed that "the little person" will not fall below i-th staircase during phase i.

Recall from previous technical lemma, this means we **can** couple to the biased random walk during phase i!

The proof is by construction of {an} and phase steps.

Lemma 3.3. For each $i \geq 2$, under event $\cap_{j=1}^{i-1} \Omega_i$, $x_n \geq i$ for all $n \in [N_{i-1}, N_i)$.

Proof by induction: Second Inductive lemma

Notice that this lemma is exactly what we what, i.e**., if Omega events for all phases before i occurs, event i will occur with a sufficiently large probability: the probability in blue box.**

We will talk about how we get this from first inductive lemma, coupling and Hoeffding inequality bounds in next slide

Lemma 3.4. For each $i \geq 2$, under event $\cap_{j=1}^{i-1}\Omega_i$, for all $n \in [N_{i-1}, N_i)$,

$$P\{S_{N_i} \geq M + 4(i-1)\} \geq 1 - \frac{2}{(M-2+2(i-2))^{2K}}$$

Key to Prove the Second Inductive Lemma

- First Inductive Lemma guarantees that we can couple to the biased random walk sequence {Zn}
- Coupling to {Zn} allows us to apply Hoeffding's inequality whereby by choosing appropriate t, we establish the bounds necessary for second inductive

Define $I_{i,m} := \sum_{j=1}^{m} Z_i$. Since Z_i is strictly bounded by $[-1, 1]$, by Hoeffding's inequality, for any $t > 0$,

$$P(|I_{i,m} - E(I_{i,m})| \geq t) \leq 2\exp(-\frac{t^2}{2m}) \tag{54}$$

Proof of the main theorem via Second Inductive Lemma

- Note again, if all Omega events are true, the process tends to infinity. **Therefore, the probability of all Omega events are true is the probability the process tends to infinity, which is the following:**

As a result,

$$P(S_n \to \infty) \geq (1-\epsilon) \cdot \Pi_{j=0}^{\infty}(1 - \frac{2}{(M-2+2j)^{2K}})$$

$$\geq (1-\sigma/2)^2 > 1-\sigma$$

Overview Expository Studies

- Discrete-time, general-state space Markov Chain following Meyn and Tweedie (1997) and Rosenthal's review papers on Markov chain and MCMC
- Adaptive MCMC: convergence conditions for adaptive MCMC, adversarial adaptive MCMC, adaptive Gibbs sampler, optimal scaling and other papers
- Quantitative convergence rates: general framework and proofs (Rosenthal 1995) and application on James Estimator (Rosenthal 1998)
- Weak convergence of Random Walk Metropolis to Langevin diffusion; required theories regarding weak convergence of stochastic processes following Ethier and Kurtz (2009) and other sources; Complexity bounds via diffusion limit

How fast is the chain converging?

- Quantitative convergence rates (Rosenthal, 1995):

Proposition 1. *Let $P(x,\cdot)$ be the transition probabilities for a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose there exist $\epsilon > 0$, $0 < \lambda < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{1-\lambda}$, a non-negative function $f : \mathcal{X} \to \mathbf{R}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, such that*

$$\mathbf{E}\,(f(X_1)\,|\,X_0 = x) \;\le\; \lambda f(x) + \Lambda\,, \qquad x \in \mathcal{X} \tag{1}$$

and

$$P(x,\cdot) \;\ge\; \epsilon\, Q(\cdot)\,, \qquad x \in f_d \tag{2}$$

where $f_d = \{x \in \mathcal{X} \,|\, f(x) \le d\}$, and where $P(x,\cdot) \;\ge\; \epsilon\, Q(\cdot)$ means $P(x,S) \;\ge\; \epsilon\, Q(S)$ for every measurable $S \subseteq \mathcal{X}$. Then for any $0 < r < 1$, we have

$$\|\mathcal{L}(X_k) - \pi(\cdot)\| \;\le\; (1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}\gamma^{r}\right)^{k} \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}\,(f(X_0))\right),$$

where

$$\alpha^{-1} \;=\; \frac{1 + 2\Lambda + \lambda d}{1+d} \;<\; 1\,; \qquad \gamma \;=\; 1 + 2(\lambda d + \Lambda)\,.$$

Quantitative convergence rates

- How to use this method?
  - Step 1: Pick an appropriate auxiliary function “f”
  - Step 2: Find parameters and verify the Drift and Minorisation conditions in blue box
  - Step 3: Plug in parameters into the red box and claim your quantitative bounds!
- The proof is based on a coupling argument and Minorization condition: basically locate a ”special” set called small set and set up two chains from different initial distribution and analyze how long they will converge.

Optimal Scaling (Optimal step size): a proof of weak convergence of random walk MH to Langevin diffusion

Theorem 2.20. Suppose f is a real-valued function with continuous second derivative and satisfies the conditions outlined in the set-up. Let $\sigma_n^2 = l^2/(n-1)$ and $U_t^n := X_{[nt],1}^n$ where $X_{[nt],1}^n$ denotes the first-dimension component of the sped-up (by $n-$times for setting of n dimension) MH process. Assume that $X_{0,j}^i = X_{0,j}^j \forall i \leq j$. Then, as $n \to \infty$,

$$U^n \Rightarrow U$$

where $\Rightarrow$ denotes weak convergence of processes in the Skorokhod topology and U_0 is distributed according to f and U satisfies the Langevin SDE

$$dU_t = (h(l))^{1/2} dB_t + h(l) \frac{f'(U_t)}{2f(U_t)} dt$$

and

$$h(l) = 2l^2 \Phi(-\frac{l\sqrt{I}}{2})$$

with Φ being the standard normal cumulative cdf and

$$I \equiv E_f[(\frac{f'(X)}{f(X)})^2]$$

Thank you!