Hoff Chs. 6 and 10: Sampling

Outline

I. Motivation

II. Markov chain Monte Carlo

III. Gibbs sampling (Hoff Ch. 6)

IV. Metropolis-Hastings (Hoff Ch. 10)

V. MCMC diagnostics (Hoff Ch. 6.6)

Keep in mind throughout:

Gibbs sampling is a special case of Metropolis-Hastings which is a special case of Markov chain Monte Carlo

I. Motivation

Bayesian stats in a nutshell.

1. Come up with model
    - Prior, likelihood
2. Compute posterior, $p(\theta \mid y)$
3. Draw samples $\theta_1, \ldots, \theta_S$ from $p(\theta \mid y)$
4. Make inferences via Monte Carlo.

$$\frac{1}{S} \sum_{i}^{S} f(\theta_i) \approx \mathbb{E}[f(\theta) \mid y]$$

- A significant portion of the rest of this course will focus on Step 3.
- When $p(\theta|y)$ belongs to a known family of distributions, e.g., beta, gamma, normal, we can draw iid samples θ_i from $p(\theta|y)$ using standard software packages.
- For many models, drawing iid θ_i is infeasible

Example: Normal model

In the last chapter we studied the model

- $Y_1, \ldots, Y_n \mid \theta, \sigma^2 \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$. (Likelihood)
- $\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$ (Prior on σ^2)
- $\theta \mid \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$ (Prior on θ, given σ^2)

which gave the posterior

$$p(\theta, \sigma^2 \mid \vec{y}_n) = p(\theta \mid \sigma^2, \vec{y}_n)\, p(\sigma^2 \mid \vec{y}_n)$$

$$= N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right) IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right).$$

We could sample from this posterior via

for $1 \leq i \leq S$:

- $\sigma_i^2 \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right)$
- $\theta_i \sim N\left(\mu_n, \frac{\sigma_i^2}{\kappa_n}\right)$

$(*)$

$\Rightarrow \{(\theta_i, \sigma_i^2)\}_{i=1}^S$ iid samples from $p(\theta, \sigma^2 | \vec{y}_n)$

$\{\theta_i\}_{i=1}^S$ iid samples from $p(\theta | \vec{y}_n)$

<u>However</u>: The above model supposes that uncertainty about θ depends on σ^2.

- Suppose instead we want independent priors, e.g.,
  - $\theta \sim N(\mu_0, \tau_0^2)$
  - $\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$

Unlike the previous model where

$$\sigma^2 \mid \vec{y}_n \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right),$$

in this model, $p(\sigma^2 | \vec{y}_n)$ is <u>not</u> an inverse gamma distribution, nor any other distribution we can easily sample from.

$\Longrightarrow$ the sampling scheme (*) is infeasible.

So how to sample from $p(\theta, \sigma^2 | \vec{y}_n)$?

I'll give the brief answer here, as a preview of what's to come. Then, we will step back and look at the bigger picture.

The brief answer.

Instead of finding $p(\theta|\sigma^2, \vec{y}_n)$ and $p(\sigma^2|\vec{y}_n)$, we will compute the full conditionals

$$p(\theta|\sigma^2, \vec{y}_n) \quad \text{and} \quad p(\sigma^2|\theta, \vec{y}_n)$$

We will then generate samples via Gibbs sampling:

Set starting values θ_0, σ_0^2

for $1 \leq i \leq S$:

- $\theta_i \sim p(\theta \mid \sigma_{i-1}^2, \vec{y}_n)$
- $\sigma_i^2 \sim p(\sigma^2 \mid \theta_i, \vec{y}_n)$

Return $\{(\theta_i, \sigma_i^2)\}_{i=0}^{S}$.

Important: These samples are not independent.

However, they still satisfy

$$\lim_{S\to\infty} \frac{1}{S}\sum f(\theta_s, \sigma_s^2) = \mathbb{E}\left[f(\theta, \sigma^2) \,|\, \vec{y}_n \right]$$

$$= \int f(\theta, \sigma^2)\, p(\theta, \sigma^2 \,|\, \vec{y}_n)\, d\theta\, d\sigma^2 .$$

Gibbs sampling is a specific case of

II. Markov chain Monte Carlo (MCMC)

Overview

- Suppose we want to sample from a density p_*, e.g., $p_* = p(\theta | y)$, in order to evaluate

$$\int f(\theta)\, p_*(\theta)\, d\theta$$

but we can't sample from p_* directly.

- Instead, maybe we can construct a sequence

$$\theta^{(0)} \sim p_0, \quad \theta^{(1)} \sim p_1, \quad \theta^{(2)} \sim p_2, \ldots$$

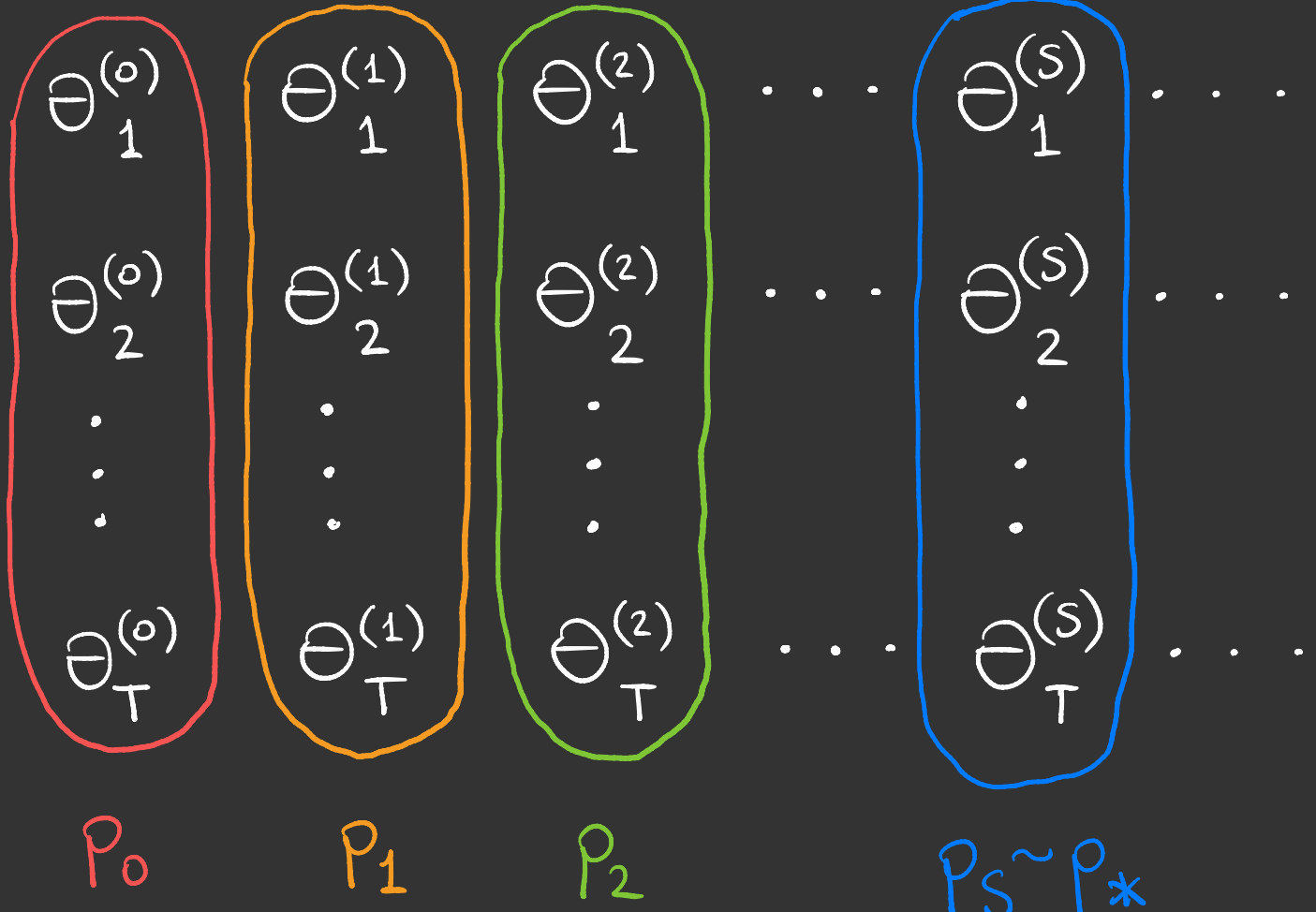
such that the densities p_S converge to p_* as $S \to \infty$. Then (loosely speaking):

$$\left| \int f(\theta) p_S(\theta) d\theta - \int f(\theta) p_*(\theta) d\theta \right|$$

$$= \left| \int f(\theta) [p_S(\theta) - p_*(\theta)] d\theta \right|$$

$$\leq \int |f(\theta)| \, |p_S(\theta) - p_*(\theta)| \, d\theta \longrightarrow 0$$

as $S \to \infty$


$\theta_1^{(0)}$
$\theta_2^{(0)}$
$\theta_T^{(0)}$
$\theta_1^{(1)}$
$\theta_2^{(1)}$
$\theta_T^{(1)}$
$\theta_1^{(2)}$
$\theta_2^{(2)}$
$\theta_T^{(2)}$
$\theta_1^{(S)}$
$\theta_2^{(S)}$
$\theta_T^{(S)}$
P_0
P_1
P_2
$P_S \sim P_*$

MCMC: The basic idea

Setup. We have a Bayesian model

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$.

Goal. Draw (approximate) samples from $p(\theta|y)$ s.t.

$$\frac{1}{S}\sum_{i=1}^{S} f(\theta^{(i)}) \approx \int_{\mathbb{R}^d} f(\theta)\, p(\theta|y)\, d\theta. \quad (*)$$

LLN. If independent samples from $p(\theta|y)$ are available, we can just apply the Law of Large Numbers to get (*). If we can't draw independent samples, we can instead try:

MCMC. Construct a Markov chain

$$\vec{\theta}_S = \left(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\right)$$

that satisfies (*).

Examples of MCMC algorithms.

- Gibbs sampling
- Metropolis-Hastings
- Hamiltonian Monte Carlo (HMC)

Markov chains

A sequence of random variables $X_1, X_2, \ldots$ is a **Markov chain** if the probability of the next state depends only on the current one:

$$\mathbb{P}(X_{n+1} = x \mid X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x \mid X_n = x_n).$$

"The future is independent of the past given the present"

Example: Gambling

- We have a coin with probability of heads = θ.
- If the coin lands on heads, you get \$1; if it lands on tails, you lose \$1.
- Let X_i be how much you earn on the i^{th} flip.

Then your total earnings after n flips is

$$S_n = \sum_{i=0}^{n} X_i.$$

- The sequence $(S_0, S_1, S_2, \ldots)$ is a Markov chain: total earnings after $n+1$ flips, S_{n+1}, depends only on the outcome of the $n+1^{th}$ flip and the value of S_n.
- It's helpful to think about real-world stochastic processes that are and are not Markov chains.

Transition kernels

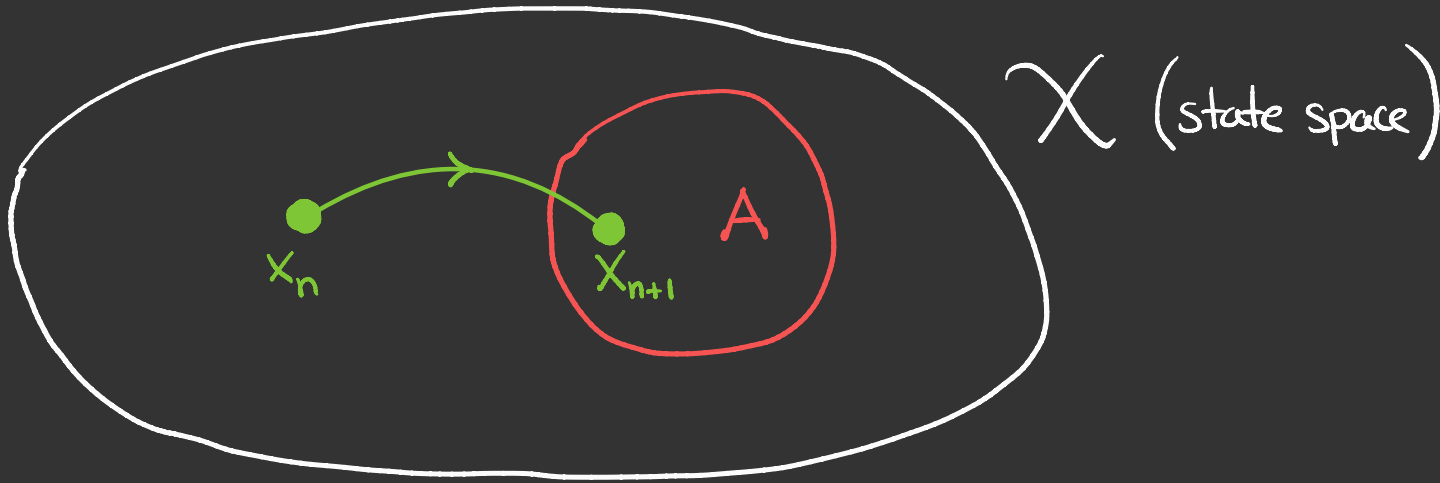
The transition kernel of a Markov chain is a function $K_n : \mathcal{X} \times \mathcal{A} \longrightarrow [0,1]$ given by

$$K_n(x, A) = \mathbb{P}(X_n \in A \mid X_{n-1} = x)$$

where $\mathcal{X}$ is the state space, and $\mathcal{A}$ the collection of all events. The Markov chain is time-homogeneous if $K_n = K_m \quad \forall n \neq m$.

- We'll generally assume time-homogeneity for simplicity

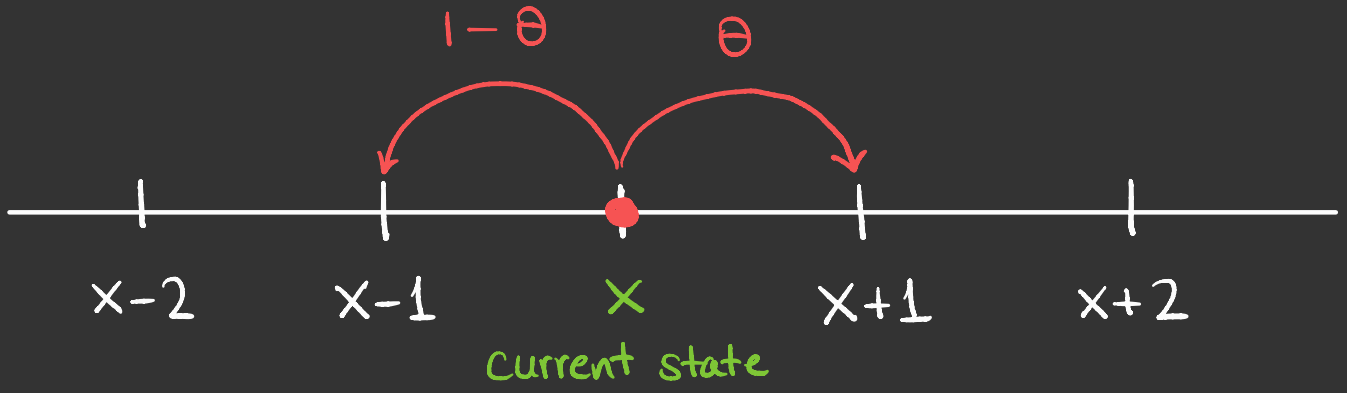
- Intuitively, $K(x, A)$ is the probability that you end up in $A \subseteq \mathcal{X}$ in the next step, given that your current state is x.



- **Example (Gambling continued)**

In the gambling example above,

$$K(x, \{y\}) = \begin{cases} \theta & \text{if } y = x+1 \\ 1-\theta & \text{if } y = x-1 \\ 0 & \text{otherwise} \end{cases}$$



- Transition kernels "act" on probability measures. Specifically, let
  - π be a probability measure on $(\mathcal{X}, \mathcal{A})$ where $\mathcal{X}$ is the state space and $\mathcal{A}$ is the collection of all events. So the probability of event A is $\pi(A)$.
  - $K : \mathcal{X} \times \mathcal{A} \longrightarrow [0, 1]$ be a transition kernel

  Then $K\pi$ is a new probability measure on

$(\mathcal{X}, \mathcal{A})$ given by

$$K\pi(A) = \int_{\mathcal{X}} K(x, A)\, \pi(dx).$$

if π has a pdf p, i.e., $\pi(dx) = p(x)dx$

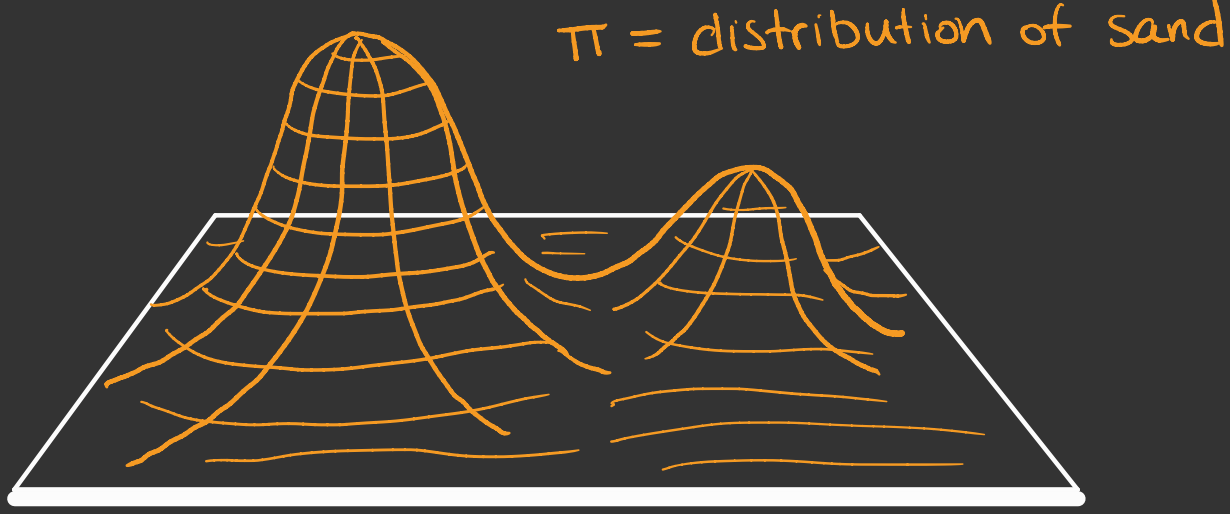
$$= \int_{\mathcal{X}} K(x, A)\, p(x)\, dx.$$

Observe that this is in fact a probability measure, e.g.,

$$K\pi(\mathcal{X}) = \int_{\mathcal{X}} K(x, \mathcal{X})\pi(dx) = \int_{\mathcal{X}} \pi(dx) = 1.$$

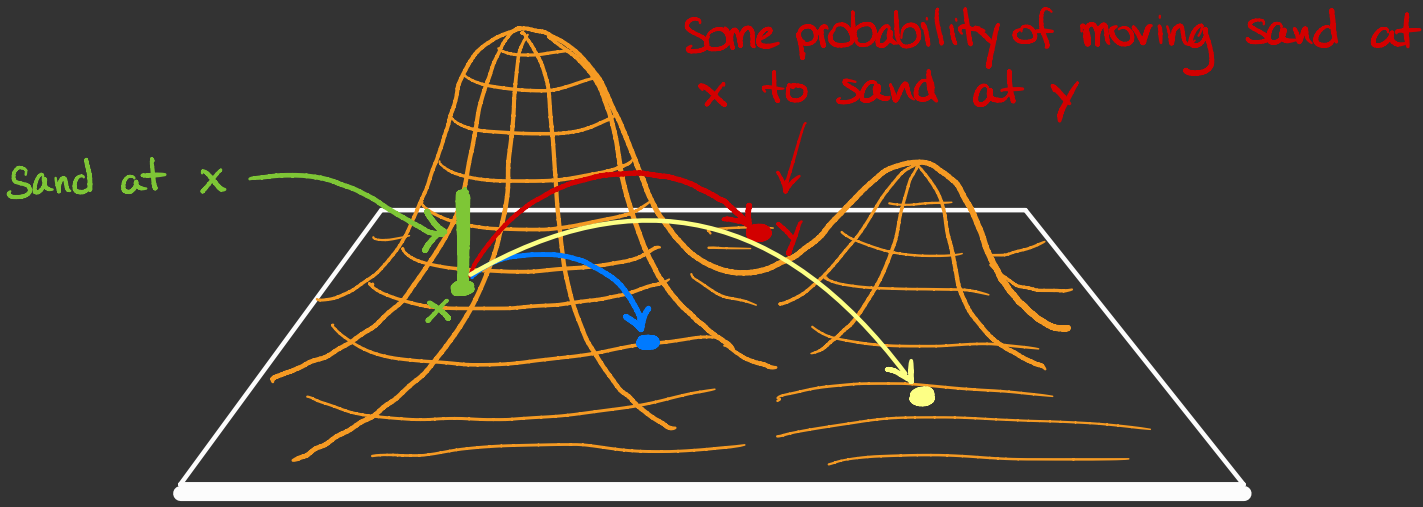
Intuition.

Imagine π as the distribution of a unit mass of sand at a construction site:



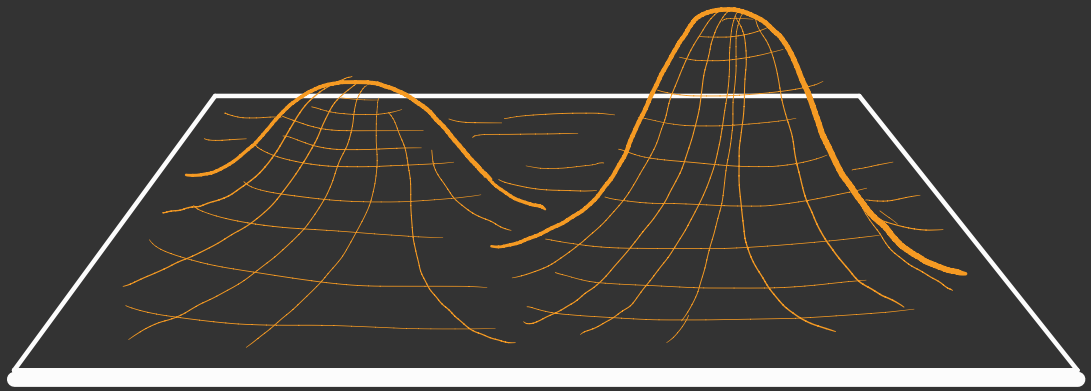
2-dimensional construction site X

The transition kernel $K(x, A)$ gives the probability of moving sand from each point x into each set A.



$K\pi$ is the distribution of the sand after one step of the Markov chain, i.e., after moving sand from each point x to its next location according to K.

$K\pi$ = distribution of sand after one step of K



2-dimensional construction site X

- Because $K\pi$ is itself a probability distribution, we can again apply K to it:

$$K^2\pi(A) = K(K\pi)(A)$$

$$= \int_{\mathcal{X}} K(x_1, A)\, K\pi(dx_1)$$

$$= \int_{\mathcal{X}} \int_{\mathcal{X}} K(x_1, A)\, K(x_0, dx_1)\, \pi(dx_0)$$

$K^m\pi$ is the distribution of the sand after m steps.

Invariant measures

- A probability measure π is <u>invariant</u> for K if $K\pi = \pi$ (and hence $K^m\pi = \pi \ \forall m$).

- Let π_* be the unique invariant measure for K. Under certain assumptions,

$$K^m \pi_0 \longrightarrow \pi_* \qquad (*)$$

as $m \longrightarrow \infty$ for any initial distribution π_0.

- Let's recall at this point our primary goal:

Construct a Markov chain $\theta^{(0)}, \theta^{(1)}, \ldots$ such that the $\theta^{(i)}$ are (approximate) samples from the posterior $p(\theta|y)$ satisfying

$$\lim_{S\to\infty} \frac{1}{S} \sum_{i}^{S} f(\theta^{(i)}) = \int f(\theta) p(\theta|y)\, d\theta.$$

- Thus, we must construct our Markov chain in a manner consistent with $p(\theta|y)$ (not any old Markov chain will do). Specifically, we aim to construct a Markov chain with <u>invariant measure</u> $p(\theta|y)$.

$\Rightarrow$ If $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ is a Markov chain w/ kernel K and K-invariant measure $p(\theta|y)$, then by (*), $\theta^{(m)} \sim K^m \pi_0$ is an approximate sample from $p(\theta|y)$ for sufficiently large m.

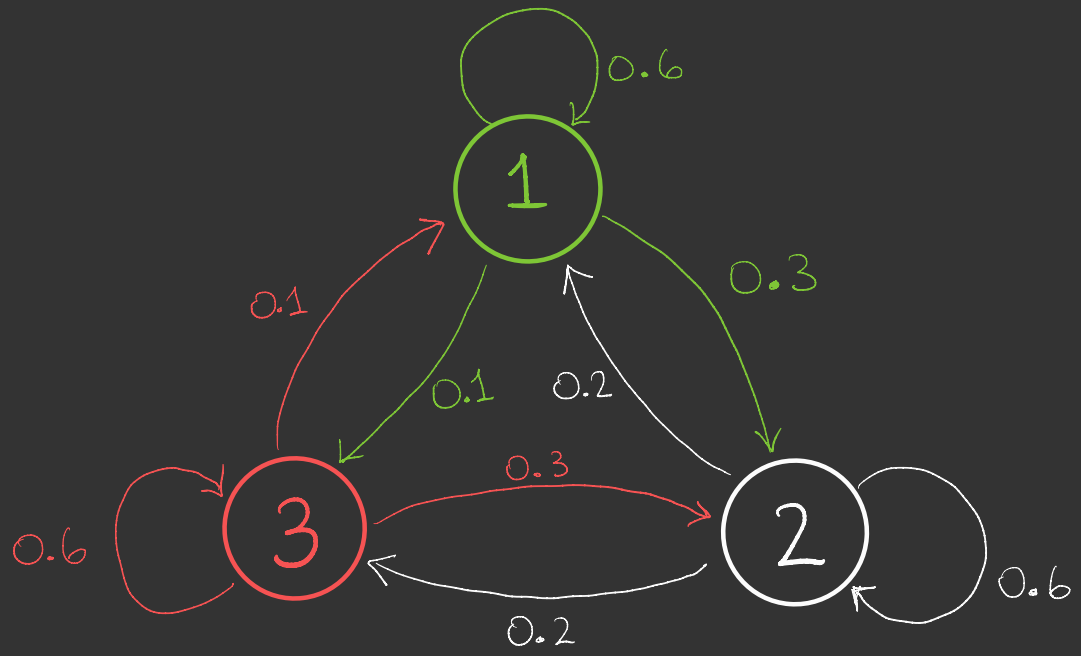
One more example: Finite state space

- For finite state spaces:
  - Probability measures are vector
  - Transition kernels are matrices
- Let's consider the state space $\mathcal{X} = \{1, 2, 3\}$.
  - A probability distribution on $\mathcal{X}$ is a 3-dimensional vector of nonnegative values whose sum is 1:

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} \quad \text{w/} \quad \pi_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{3} \pi_i = 1.$$

π_i = probability of being in state i

- A transition kernel on $\mathcal{X}$ gives probability of going from one state to another, e.g.,



- This can be written in matrix form:

$$K = \begin{pmatrix} 0.6 & 0.2 & 0.1 \\ 0.3 & 0.6 & 0.3 \\ 0.1 & 0.2 & 0.6 \end{pmatrix}$$

$$K_{ij} = \mathbb{P}\left(X_{n+1} = i \mid X_n = j \right)$$

- $\pi \in \mathbb{R}^3$ is an invariant measure for K if

$$K\pi = \pi$$

i.e. if π is an eigenvector of K w/ eigenvalue 1.

MCMC summary

- Construct Markov chain w/ invariant measure $p(\theta|y)$ starting from any <u>initial state</u> $\theta^{(0)} \sim \pi_0$.
- For sufficiently large m, the m^{th} sample $\theta^{(m)} \sim K^m \pi_0$ will be an approximate sample from $p(\theta|y)$ and for $S \gg m$,

$$\frac{1}{S-m+1} \sum_{i=m}^{S} f(\theta^{(i)}) \approx \int f(\theta) p(\theta|y) \, d\theta.$$

Ⅲ. Gibbs sampling

- Gibbs sampling is an MCMC algorithm that can be used when we know the full conditionals
- Specifically, let $\vec{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ be a d-dimensional parameter. The full conditional of θ_j is

$$p(\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d, \vec{y}_n) = p(\theta_j \mid \vec{\theta}_{-j}, \vec{y}_n)$$

where $\vec{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d)$.

Gibbs sampling algorithm

- Initialize $\vec{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.
- for $1 \leq i \leq S$:
  - $\theta_1^{(i)} \sim p(\theta_1 | \theta_2^{(i-1)}, \ldots, \theta_d^{(i-1)}, \vec{y}_n)$
  - $\theta_2^{(i)} \sim p(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)}, \vec{y}_n)$
  - $\theta_3^{(i)} \sim p(\theta_3 | \theta_1^{(i)}, \theta_2^{(i)}, \theta_4^{(i-1)}, \ldots, \theta_d^{(i-1)}, \vec{y}_n)$
  - $\vdots$
  - $\theta_d^{(i)} \sim p(\theta_d | \theta_1^{(i)}, \ldots, \theta_{d-1}^{(i)}, \vec{y}_n)$

- Return $\{\vec{\theta}^{(i)} : 1 \le i \le S\}$.

- Under certain assumptions that hold for the models considered in this course, $\{\vec{\theta}^{(i)} : 1 \le i \le S\}$ is a Markov chain satisfying

$$\lim_{S \to \infty} \frac{1}{S} \sum_{i}^{S} f(\vec{\theta}^{(i)}) = \int f(\vec{\theta}) p(\vec{\theta} | \vec{y}_n) d\vec{\theta}$$

Example: Normal model w/ independent priors

- $Y_1, \dots, Y_n \mid \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$ ← Likelihood
- $\theta \sim N(\mu_0, \tau_0^2)$ ← prior for θ
- $\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$ ← prior for σ^2
- In Ch 5 we saw that

Note that both of these depend on σ^2

$$p(\theta \mid \sigma^2, \vec{Y}_n) = N\left(\mu_n(\sigma^2), \tau_n^2(\sigma^2)\right)$$

(see previous notes for values of μ_n and τ_n^2)

- After some mathematical manipulation (see Section 6.3 in Hoff) one finds that

$$p(\sigma^2 | \theta, \vec{y}_n) \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2(\theta)\right)$$

where $\nu_n = \nu_0 + n$ and

$$\sigma_n^2(\theta) = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + n s_n^2(\theta)\right]$$

where $s_n^2(\theta) = \frac{1}{n}\sum_i (y_i - \theta)^2$.

- The Gibbs sampler for $p(\theta, \sigma^2 | \vec{y}_n)$ is therefore
  - Initialize $\theta^{(0)}, \sigma^{2(0)}$
  - for $1 \leq i \leq S$:
    - $\theta^{(i)} \sim N\left(\mu_n(\sigma^{2(i-1)}), \tau_n^2(\sigma^{2(i-1)})\right)$
    - $\sigma^{2(i)} \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2(\theta^{(i)})\right)$

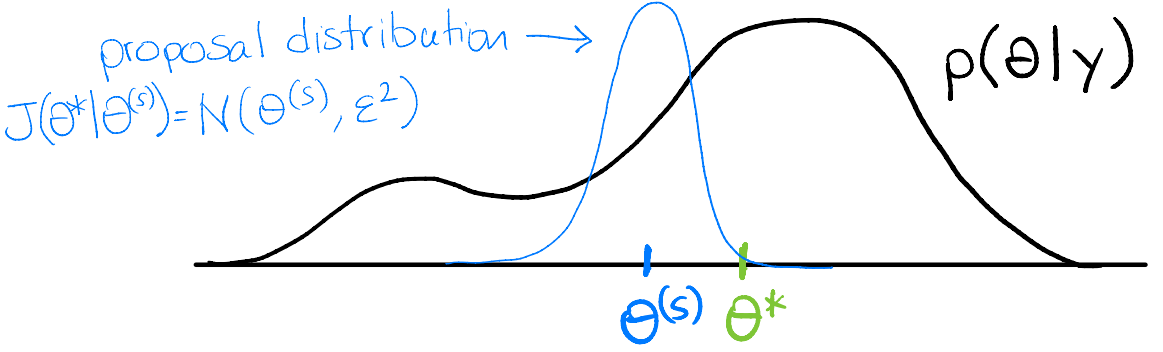
$\Rightarrow \{(\theta^{(i)}, \sigma^{2(i)}) : 1 \leq i \leq S\}$ approximate samples from $p(\theta, \sigma^2 | \vec{y}_n)$.

IV. Metropolis-Hastings

- The Metropolis (1953) and Metropolis-Hastings (1970) algorithms provide ways to sample from essentially any distribution
- As we'll see below, Gibbs sampling is a special case of MH when the full conditionals are known

Intuition

- Goal: Construct Markov chain with invariant measure $p(\theta|y)$.
- Suppose the current state in our chain is $\theta^{(s)}$.
- We want to propose a new value/state θ^*
  - To satisfy Markov property, θ^* should depend only on $\theta^{(s)}$
  - Example: $\theta^* \sim N(\theta^{(s)}, \varepsilon^2)$

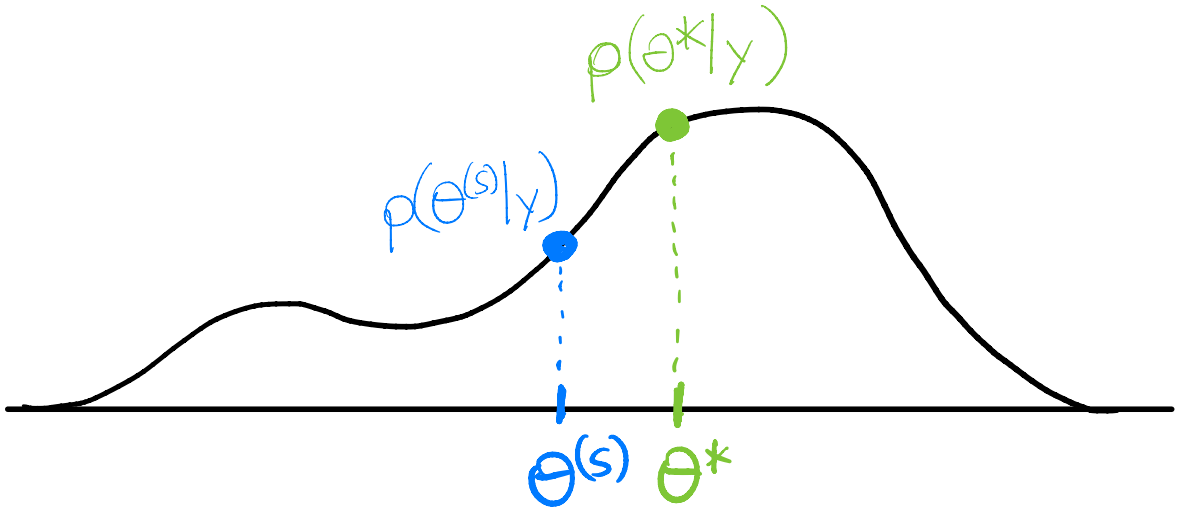


- Is θ^* a good fit for $p(\theta|y)$?
  - If θ^* is a "good fit", set $\theta^{(s+1)} = \theta^*$
  - Otherwise, "accept" θ^* w/ some probability r.
- Intuitively, θ^* is a good fit if

$$p(\theta^*|y) > p(\theta^{(s)}|y)$$

or, equivalently, if

$$\frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} > 1.$$



- By Bayes rule,

$$\frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{\frac{p(y|\theta^*)p(\theta^*)}{p(y)}}{\frac{p(y|\theta^{(s)})p(\theta^{(s)})}{p(y)}} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}$$

Metropolis-Hastings algorithm

- Specify a proposal or "jumping" distribution $J(\theta^* | \theta)$
- Initialize $\theta^{(0)}$
- for $1 \leq i \leq S$:
  - Sample $\theta^* \sim J(\theta^* | \theta^{(i-1)})$
  - Compute $r = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})} \times \frac{J(\theta^{(i-1)} | \theta^*)}{J(\theta^* | \theta^{(i-1)})}$
  - Set $\theta^{(i)} = \begin{cases} \theta^* & \text{w/ probability } \min\{1, r\} \\ \theta^{(i-1)} & \text{otherwise} \end{cases}$
- Return samples $\{\theta^{(i)} : i = 0, \ldots, S\}$.

- If the proposal distribution is <u>symmetric</u>, i.e., if

$$J(\theta^*|\theta) = J(\theta|\theta^*) \quad \forall \theta, \theta^*, \text{ then}$$

$$r = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})} \times \frac{J(\theta^{(i-1)}|\theta^*)}{J(\theta^*|\theta^{(i-1)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})}.$$

This is the <u>Metropolis algorithm</u>.

- Examples of symmetric distributions
  - Normal: $J(\theta^*|\theta) = N(\theta, \varepsilon^2)$

$$J(\theta^*|\theta) = \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp\left[-\left(\frac{\theta^*-\theta}{\varepsilon}\right)^2\right] = J(\theta|\theta^*)$$

  - Uniform: $J(\theta^*|\theta) = \text{Uniform}(\theta-\varepsilon, \theta+\varepsilon)$, $\varepsilon > 0$.

Example: Normal w/ known variance

- $Y_1, \ldots, Y_n \mid \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$
- $p(\theta) = N(\mu_0, \tau_0^2)$
- In Chapter 5, we computed the posterior $p(\theta \mid \sigma^2, \vec{y}_n)$ in closed form. Since it was a normal, we could sample from it directly.
- Suppose, however, we want to run Metropolis with a normal proposal, i.e., $J(\theta^* \mid \theta) = N(\theta, \varepsilon^2)$.

- $p(\vec{y}_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$

$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\left(\frac{y_i-\theta}{\sigma}\right)^2\right]$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]$$

- $p(\theta) = (2\pi\tau_0^2)^{-\frac{1}{2}} \exp\left[-\left(\frac{\theta-\mu_0}{\tau_0}\right)^2\right]$

$$\Rightarrow r = \frac{\exp\left[-\frac{1}{\sigma^2}\sum(y_i-\theta^*)^2\right]\exp\left[-\left(\frac{\theta^*-\mu_0}{\tau_0}\right)^2\right]}{\exp\left[-\frac{1}{\sigma^2}\sum(y_i-\theta^{(s)})^2\right]\exp\left[-\left(\frac{\theta^{(s)}-\mu_0}{\tau_0}\right)^2\right]}$$

$$= \exp\left[-\frac{1}{\sigma^2}\left(\sum_i (y_i-\theta^*)^2-(y_i-\theta^{(s)})^2\right)-\left(\frac{\theta^*-\mu_0}{\tau_0}\right)^2+\left(\frac{\theta^{(s)}-\mu_0}{\tau_0}\right)^2\right]$$

Take log for numerical stability:

$$\log(r) = -\frac{1}{\sigma^2}\left(\sum_i (y_i-\theta^*)^2-(y_i-\theta^{(s)})^2\right)-\frac{1}{\tau_0^2}\left[(\theta^*-\mu_0)^2-(\theta^{(s)}-\mu_0)^2\right]$$

$$\min\{1, r\} \iff \min\{0, \log(r)\}$$

Algorithm.

- Initialize $\theta^{(0)}$
- for $1 \leq i \leq S$:
  - $\theta^* \sim N(\theta^{(i-1)}, \varepsilon^2)$
  - Compute $\log(r)$ as above
  - Draw $u \sim \text{Uniform}(0,1)$
  - $\theta^{(i)} = \begin{cases} \theta^* & \text{if } \log(u) < \log(r) \\ \theta^{(i-1)} & \text{otherwise} \end{cases}$
- Return $\{\theta^{(0)}, \ldots, \theta^{(S)}\}$.

V. MCMC diagnostics

Summary so far

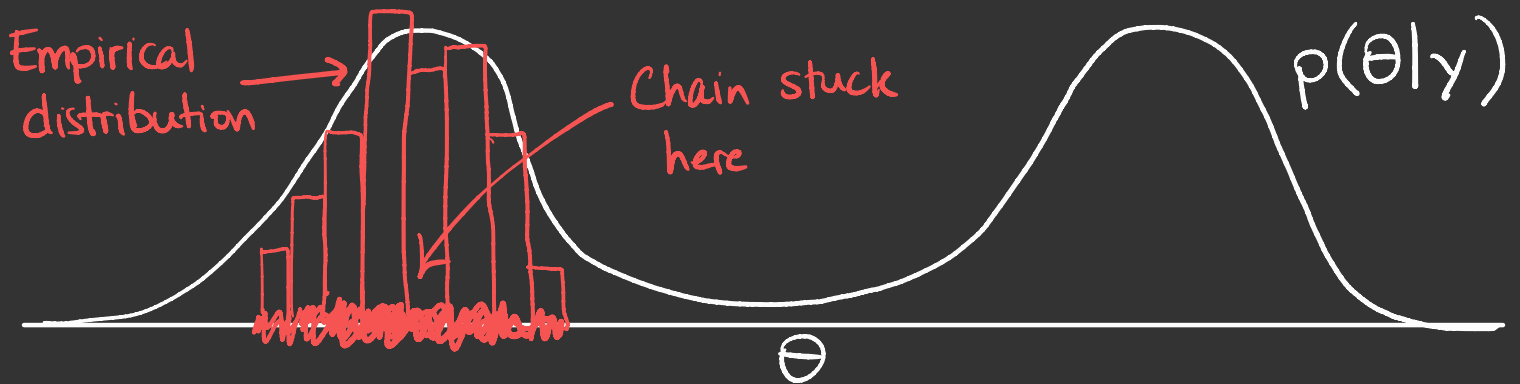
When we cannot draw independent samples from the posterior (or any distribution), we can use Gibbs/Metropolis/Metropolis-Hastings/MCMC to draw correlated samples $\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta|y)$ such that

$$\lim_{S \to \infty} \frac{1}{S} \sum_{i=1}^{S} f(\theta^{(i)}) = \int_{\Theta} f(\theta) p(\theta|y) d\theta.$$

Great in theory. What about in practice?

Intuition

- We want the empirical distribution of the chain $\theta^{(1)}, \ldots, \theta^{(S)}$ (i.e., the histogram of these samples) to resemble the true distribution $p(\theta|y)$.
- This requires the chain to sufficiently explore the state space, e.g., it can't get "stuck"



- Terms/phrases like "the chain has achieved stationarity" or "the chain has converged" or "good mixing" refer to an MCMC algorithm performing well

More formally

- Is S large enough that $\frac{1}{S}\sum f(\theta^{(i)})$ is a good approximation of $\int f(\theta)p(\theta|y)\,d\theta$?
- For classical Monte Carlo (independent samples) we saw that the Monte Carlo error is

$$\mathbb{E}\left[\left(\frac{1}{S}\sum f(\theta^{(i)}) - \int f(\theta)p(\theta|y)\,d\theta\right)^2\right] = \mathrm{Var}\left(\frac{1}{S}\sum_i f(\theta^{(i)})\right)$$

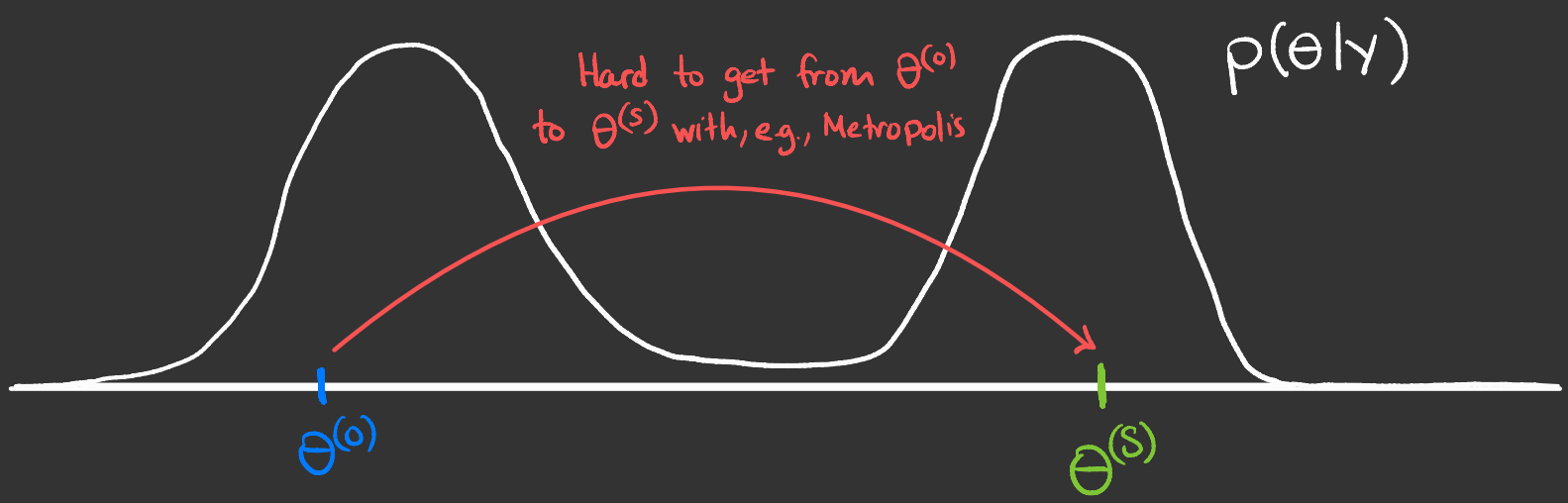
$$= \frac{1}{S^2}\sum_{i=1}^{S}\mathrm{Var}\left(f(\theta^{(i)})\right) = \frac{\mathrm{Var}(f(\theta))}{S}.$$

That is, the standard deviation of the error in the approximation

$$\frac{1}{S}\sum f(\theta^{(i)}) \approx \int f(\theta)p(\theta|y)d\theta$$

goes to zero at rate $\frac{1}{\sqrt{S}}$.

- For MCMC, samples $\theta^{(i)}$ are not independent. They can get "stuck" in certain regions of state space, causing MCMC to be very slow.



- Setting $f(\theta) = \theta$ for simplicity (though the following holds for most functions f) and $\bar{\theta}_S = \frac{1}{S}\sum \theta^{(i)}$ and $\theta_* = \int \theta\, p(\theta|y)\, d\theta$, the MCMC error is

$$\mathbb{E}\left[(\bar{\theta}_S - \theta_*)^2\right] = \mathbb{E}\left[\left(\frac{1}{S}\sum(\theta^{(i)} - \theta_*)\right)^2\right]$$

Call this $\mathrm{Var}_{MCMC}(\bar{\theta}_S)$

$$= \frac{1}{S^2}\mathbb{E}\left[\left(\sum_{i=1}^{S}\underbrace{(\theta^{(i)} - \theta_*)}_{a_i}\right)^2\right]$$

$$= \frac{1}{S^2}\mathbb{E}\left[\left(\sum_{i=1}^{S} a_i\right)^2\right]$$

$$= \frac{1}{S^2}\mathbb{E}\left[\sum_{i=1}^{S} a_i^2 + \sum_{i\neq j} a_i a_j\right]$$

$$= \frac{1}{S^2}\mathbb{E}\left[\sum_{i=1}^{S}(\theta^{(i)} - \theta_*)^2\right] + \frac{1}{S^2}\mathbb{E}\left[\sum_{i\neq j}(\theta^{(i)} - \theta_*)(\theta^{(j)} - \theta_*)\right]$$

$$= \frac{\mathrm{Var}(\theta)}{S} + \underbrace{\frac{1}{S^2} \sum_{i \neq j} \mathbb{E}\left[(\theta^{(i)} - \theta_*)(\theta^{(j)} - \theta_*)\right]}$$

Monte Carlo variance/error

Additional error from correlation in Markov chain

- The lag-t sample autocorrelation function is

$$\mathrm{acf}_t(\theta^{(1)}, \ldots, \theta^{(S)}) = \frac{\frac{1}{S-t} \sum_{i=1}^{S-t} (\theta^{(i)} - \bar{\theta}_S)(\theta^{(i+t)} - \bar{\theta}_S)}{\frac{1}{S-1} \sum_{i=1}^{S} (\theta^{(i)} - \bar{\theta}_S)^2}.$$

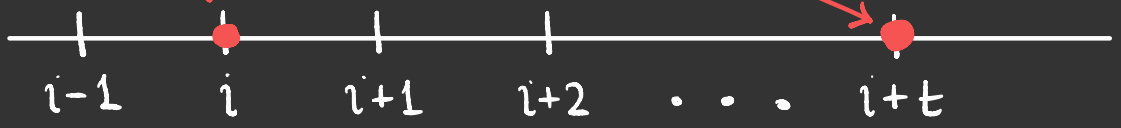
Helpful to write out explicitly. For example, if $S=10$ and $t=4$, then the sum in the numerator becomes

$$\sum_{i=1}^{6}(\theta^{(i)}-\bar{\theta}_{10})(\theta^{(i+t)}-\bar{\theta}_{10}) = (\theta^{(1)}-\bar{\theta}_{10})(\theta^{(5)}-\bar{\theta}_{10}) + (\theta^{(2)}-\bar{\theta}_{10})(\theta^{(6)}-\bar{\theta}_{10})$$

$$+ (\theta^{(3)}-\bar{\theta}_{10})(\theta^{(7)}-\bar{\theta}_{10}) + (\theta^{(4)}-\bar{\theta}_{10})(\theta^{(8)}-\bar{\theta}_{10})$$

$$+ (\theta^{(5)}-\bar{\theta}_{10})(\theta^{(9)}-\bar{\theta}_{10}) + (\theta^{(6)}-\bar{\theta}_{10})(\theta^{(10)}-\bar{\theta}_{10})$$

How correlated are $\theta^{(i)}$ and $\theta^{(i+t)}$?



So acf_t is basically the average correlation in the chain $\theta^{(1)}, \ldots, \theta^{(S)}$ between samples that are t steps apart.

- Another diagnostic metric is <u>effective sample size</u>. Intuitively this measures how many independent samples samples are needed to obtain the same accuracy as the MCMC samples. Formally, the effective sample size (ESS) of a chain $\theta^{(1)}, \ldots, \theta^{(S)}$ is

$$\mathrm{ESS} = \frac{\mathrm{Var}(\theta)}{\mathrm{Var}_{\mathrm{MCMC}}(\bar{\theta}_S)}$$

$$= \frac{\mathrm{Var}(\theta)}{\frac{\mathrm{Var}(\theta)}{S} + \frac{1}{S^2}\sum_{i \neq j} \mathbb{E}\left[(\theta^{(i)} - \theta_*)(\theta^{(j)} - \theta_*)\right]}$$

$$= \left(\frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\theta) + \frac{1}{S}\sum_{i \neq j} \mathbb{E}\left[(\theta^{(i)} - \theta_*)(\theta^{(j)} - \theta_*)\right]}\right) S$$

- Note that ESS is generally some fraction of the total number of samples, S.
- If samples are independent, $ESS = S$.
- Large S can indicate good mixing, but does not guarantee it. See mixture of normal example in code
- There is no definitive way to tell whether an MCMC algorithm is performing well in practice

- So rely on multiple heuristics, such as
  - ESS
  - lag-t autocorrelation
  - mixing of multiple chains
  - trace plots
  - domain knowledge

End Chs. 6 and 10