

# Markov Chain Monte Carlo

**Mark van der Wilk**

Department of Computing  
Imperial College London

@markvanderwilk  
m.vdwilk@imperial.ac.uk

February 24, 2023

# Goal

We want to create Monte Carlo estimators of integrals:

$$I = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{S} \sum_{s=1}^S f(\mathbf{x}^{[s]}) = \hat{I} \quad \text{with } \mathbf{x}^{[s]} \sim p(\mathbf{x})$$

# Goal

We want to create Monte Carlo estimators of integrals:

$$I = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{S} \sum_{s=1}^S f(\mathbf{x}^{[s]}) = \hat{I} \quad \text{with } \mathbf{x}^{[s]} \sim p(\mathbf{x})$$

Last lecture we saw

- ▶ rejection sampling
- ▶ importance sampling

# Goal

We want to create Monte Carlo estimators of integrals:

$$I = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{S} \sum_{s=1}^S f(\mathbf{x}^{[s]}) = \hat{I} \quad \text{with } \mathbf{x}^{[s]} \sim p(\mathbf{x})$$

Last lecture we saw

- ▶ rejection sampling — High rejection rate in high dim
- ▶ importance sampling — High variance in high dim

# Goal

We want to create Monte Carlo estimators of integrals:

$$I = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{S} \sum_{s=1}^S f(\mathbf{x}^{[s]}) = \hat{I} \quad \text{with } \mathbf{x}^{[s]} \sim p(\mathbf{x})$$

Last lecture we saw

- ▶ rejection sampling — High rejection rate in high dim
- ▶ importance sampling — High variance in high dim

Today: Markov Chain methods for sampling from  $p(\mathbf{x})$

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a proposal density  $q$  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a proposal density  $q$  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a proposal density  $q$  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea**: For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$



# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a proposal density  $q$  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea:** For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$
- ▶ Simplify joint with **Markov property**:  
 $q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}) = q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)})$  only depends on the previous setting/state of the chain

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a proposal density  $q$  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea**: For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$
- ▶ Simplify joint with **Markov property**:  
 $q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}) = q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)})$  only depends on the previous setting/state of the chain
- ▶  $T$  is called a **transition operator**

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a **proposal density  $q$**  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea**: For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$
- ▶ Simplify joint with **Markov property**:  
 $q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}) = q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)})$  only depends on the previous setting/state of the chain
- ▶  $T$  is called a **transition operator**
- ▶ Example:  $T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \sigma^2 \mathbf{I})$

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a **proposal density  $q$**  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

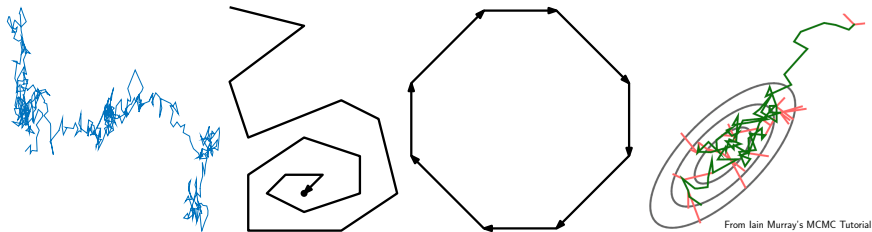
- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea**: For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$
- ▶ Simplify joint with **Markov property**:  
 $q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}) = q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)})$  only depends on the previous setting/state of the chain
- ▶  $T$  is called a **transition operator**
- ▶ Example:  $T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \sigma^2 I)$
- ▶ Samples  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$  form a **Markov chain**

# Markov Chains

Instead of generating independent samples  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ , use a **proposal density  $q$**  that depends on the previous sample (state)  $\mathbf{x}^{(t)}$

- ▶ This generates a **sequence** with a joint  $q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$
- ▶ **Key idea**: For the marginal at  $T$  we want  $q_{X^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$
- ▶ Simplify joint with **Markov property**:  
 $q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}) = q(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)})$  only depends on the previous setting/state of the chain
- ▶  $T$  is called a **transition operator**
- ▶ Example:  $T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \sigma^2 I)$
- ▶ Samples  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$  form a **Markov chain**
- ▶ Samples  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$  are **no longer independent**

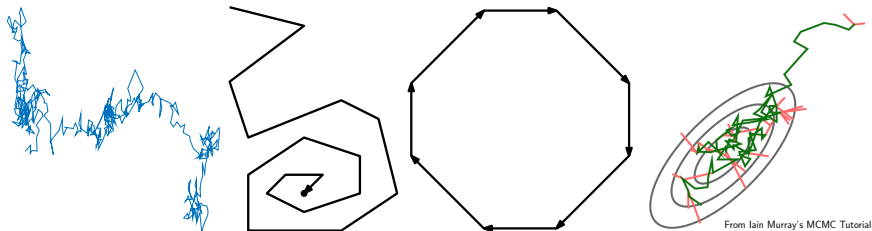
# Behaviour of Markov Chains



Four different behaviors of Markov chains:

- Diverge (e.g., random walk diffusion where  $\mathbf{x}^{(t+1)} \sim \mathcal{N}(\mathbf{x}^{(t)}, \mathbf{I})$ )

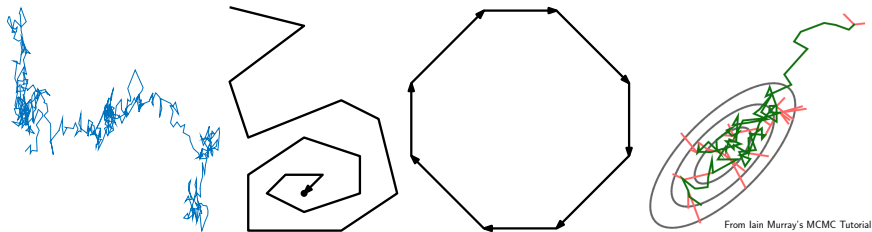
# Behaviour of Markov Chains



Four different behaviors of Markov chains:

- ▶ Diverge (e.g., random walk diffusion where  $\mathbf{x}^{(t+1)} \sim \mathcal{N}(\mathbf{x}^{(t)}, \mathbf{I})$ )
- ▶ Converge to an absorbing state

# Behaviour of Markov Chains

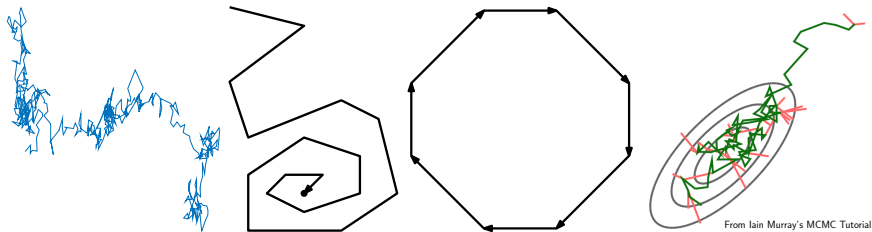


Four different behaviors of Markov chains:

- ▶ Diverge (e.g., random walk diffusion where  $\mathbf{x}^{(t+1)} \sim \mathcal{N}(\mathbf{x}^{(t)}, \mathbf{I})$ )
- ▶ Converge to an absorbing state
- ▶ Converge to a (deterministic) limit cycle



# Behaviour of Markov Chains

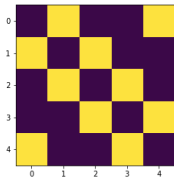


Four different behaviors of Markov chains:

- ▶ Diverge (e.g., random walk diffusion where  $\mathbf{x}^{(t+1)} \sim \mathcal{N}(\mathbf{x}^{(t)}, \mathbf{I})$ )
- ▶ Converge to an absorbing state
- ▶ Converge to a (deterministic) limit cycle
- ▶ Converge to an equilibrium distribution  $p^*$ : Markov chain remains in a region, bouncing around in a random way

## Example: Sampling from a uniform distribution

```
T = np.array([[0.0, 0.5, 0.0, 0.0, 0.5],  
              [0.5, 0.0, 0.5, 0.0, 0.0],  
              [0.0, 0.5, 0.0, 0.5, 0.0],  
              [0.0, 0.0, 0.5, 0.0, 0.5],  
              [0.5, 0.0, 0.0, 0.5, 0.0]])
```



### Procedure:

1. Initialise state at  $t = 1$  by sampling from initial distribution  $p(\mathbf{x}^{(1)})$ . Can be a delta function.
2. Repeat: Sample from  $T(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)})$

# What distribution are we sampling from?

We should ask:

At time  $t$ , what distribution are we sampling from?

# What distribution are we sampling from?

We should ask:

At time  $t$ , what distribution are we sampling from?

Apply sum rule:

$$\begin{aligned} q(x^{(t)}) &= \sum_{x=1}^5 T(x^{(t)} | x^{(t-1)} = x) q(x^{(t-1)} = x) \\ &= \mathbf{T} \mathbf{q}^{(t-1)} \end{aligned}$$

# What distribution are we sampling from?

We should ask:

At time  $t$ , what distribution are we sampling from?

Apply sum rule:

$$\begin{aligned} q(x^{(t)}) &= \sum_{x=1}^5 T(x^{(t)} | x^{(t-1)} = x) q(x^{(t-1)} = x) \\ &= \mathbf{T} \mathbf{q}^{(t-1)} \end{aligned}$$

Why does it converge?

$$\mathbf{q}^{(t)} = \mathbf{T} \mathbf{q}^{(t-1)} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^{-1} \mathbf{q}^{(t-1)}$$

For this simple-to-analyse case:

- ▶ Only one eigenvector with  $\lambda = 1$ , which is  $\mathbf{p}$ .
- ▶ All other eigenvectors have  $\lambda < 1$ .

## Using Markov Chain samples: Independent chains

If after  $T$  steps, we converge to  $q_{\mathbf{x}^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$ .

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}_s), \quad \mathbf{x}_s \sim q(\mathbf{x}^{(T)}). \quad (1)$$

Where  $q(\mathbf{x}_T)$  is generated from the  $T$ th step of a Markov Chain. Time for a sample to be “good enough” is called **burn-in time**.

# Using Markov Chain samples: Independent chains

If after  $T$  steps, we converge to  $q_{\mathbf{x}^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$ .

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}_s), \quad \mathbf{x}_s \sim q(\mathbf{x}^{(T)}). \quad (1)$$

Where  $q(\mathbf{x}_T)$  is generated from the  $T$ th step of a Markov Chain. Time for a sample to be “good enough” is called **burn-in time**.

- ▶ We run  $S$  separate Markov Chains for  $T$  steps. Samples are **independent**, because the Markov Chains are independent.
- ▶ Samples are approximate. May contain bias from  $T$  not being large enough for the distribution to converge.

# Using Markov Chain samples: Single long chain

Alternative: After  $T$  steps, average all samples

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}^{(T+s)}), \quad \mathbf{x}^{(T+1)}, \dots, \mathbf{x}^{(T+S)} \sim q(\mathbf{x}_{T+1}, \dots, \mathbf{x}_{T+S}). \quad (2)$$

$$q(\mathbf{x}^{(T+1)}, \dots, \mathbf{x}^{(T+S)}) = q(\mathbf{x}^{(T)}) \prod_{s=1}^{S-1} q(\mathbf{x}^{(T+s)} \mid \mathbf{x}^{(T+s-1)}) \quad (3)$$



# Using Markov Chain samples: Single long chain

Alternative: After  $T$  steps, average all samples

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}^{(T+s)}), \quad \mathbf{x}^{(T+1)}, \dots, \mathbf{x}^{(T+S)} \sim q(\mathbf{x}_{T+1}, \dots, \mathbf{x}_{T+S}). \quad (2)$$

$$q(\mathbf{x}^{(T+1)}, \dots, \mathbf{x}^{(T+S)}) = q(\mathbf{x}^{(T)}) \prod_{s=1}^{S-1} q(\mathbf{x}^{(T+s)} | \mathbf{x}^{(T+s-1)}) \quad (3)$$

- ▶ Remember, we choose  $T$  such that  $q_{\mathbf{x}^{(T)}}(\mathbf{x}) \approx p(\mathbf{x})$ .
- ▶ Only requires  $T$  steps for burn-in time **once**.
- ▶ Then can get a single sample per step. However, samples are **correlated**.

Usually more efficient to generate **many correlated samples**, than few independent ones.

# Markov Chain Monte Carlo

Markov Chain Monte Carlo estimates an integral using correlated samples from a Markov Chain. If the chain has converged, the estimate is **unbiased**.

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}^{(s)}) \quad (4)$$

with  $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$  from Markov Chain.

$$\mathbb{E}_{q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots)}[\hat{I}] = \frac{1}{S} \sum_{s=1}^S \mathbb{E}_{q(\mathbf{x}^{(s)})}[g(\mathbf{x}^{(s)})] = I \quad (5)$$

# Markov Chain Monte Carlo

Markov Chain Monte Carlo estimates an integral using correlated samples from a Markov Chain. If the chain has converged, the estimate is **unbiased**.

$$\hat{I} \approx \frac{1}{S} \sum_{s=1}^S g(\mathbf{x}^{(s)}) \quad (4)$$

with  $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$  from Markov Chain.

$$\mathbb{E}_{q(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots)}[\hat{I}] = \frac{1}{S} \sum_{s=1}^S \mathbb{E}_{q(\mathbf{x}^{(s)})}[g(\mathbf{x}^{(s)})] = I \quad (5)$$

Variance decreases depending on **covariance**

$$\begin{aligned} \mathbb{V}_{q(\{\mathbf{x}^{(s)}\})}[\hat{I}] &= \frac{1}{S^2} \left[ \sum_{s=1}^S \mathbb{V}_{q(\mathbf{x}^{(s)})}[g(\mathbf{x}^{(s)})] + \sum_t \sum_{t' \neq t} \mathbb{C}_{q(\mathbf{x}^{(t)}, \mathbf{x}^{(t')})}[g(\mathbf{x}^{(t)}), g(\mathbf{x}^{(t')})] \right] \\ &= \frac{1}{S} \mathbb{V}_{p(\mathbf{x})}[g(\mathbf{x})] + \left[ \sum_t \sum_{t' \neq t} \mathbb{C}_{q(\mathbf{x}^{(t)}, \mathbf{x}^{(t')})}[g(\mathbf{x}^{(t)}), g(\mathbf{x}^{(t')})] \right] \end{aligned}$$

# Correlation vs steps trade-off

Independent chains:

- ▶ Require  $T \cdot S$  transitions for  $S$  samples
- ▶ Generate independent samples, so don't need too many  $S$ .

Single chain:

- ▶ Require  $T + S$  transitions for  $S$  samples
- ▶ Generates dependent samples so may need more  $S$ .

# Converging to an Equilibrium Distribution

To get a Markov Chain that converges to a desired distribution  $p(\mathbf{x})$ , we need two properties:

1. Transition leaves  $p(\mathbf{x})$  **invariant**:

$$p(\mathbf{x}) = \int T(\mathbf{x}|\mathbf{x}')p(\mathbf{x}')d\mathbf{x}' \quad (6)$$

i.e. if we start with a sample from  $p(\mathbf{x})$ , the marginal distribution after the transition is unchanged.

# Converging to an Equilibrium Distribution

To get a Markov Chain that converges to a desired distribution  $p(\mathbf{x})$ , we need two properties:

1. Transition leaves  $p(\mathbf{x})$  **invariant**:

$$p(\mathbf{x}) = \int T(\mathbf{x}|\mathbf{x}')p(\mathbf{x}')d\mathbf{x}' \quad (6)$$

i.e. if we start with a sample from  $p(\mathbf{x})$ , the marginal distribution after the transition is unchanged.

2. Transition is **ergodic**. Definition is technical, but it is needed to ensure that  $\pi(\mathbf{x}^{(t)}) \rightarrow p(\mathbf{x})$  as  $t \rightarrow \infty$ .

Ergodic chains only have one equilibrium distribution.

# Invariance and Detailed Balance

- **Invariance:** Each step leaves the distribution  $p$  invariant (we stay in  $p$ ):

$$p(\mathbf{x}') = \sum_x T(\mathbf{x}'|\mathbf{x})p(\mathbf{x}) \qquad p(\mathbf{x}') = \int T(\mathbf{x}'|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

# Invariance and Detailed Balance

- **Invariance:** Each step leaves the distribution  $p$  invariant (we stay in  $p$ ):

$$p(\mathbf{x}') = \sum_x T(\mathbf{x}'|\mathbf{x})p(\mathbf{x}) \qquad p(\mathbf{x}') = \int T(\mathbf{x}'|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

Once we sample from  $p$ , the transition operator will not change this, i.e., we do not fall back to some funny distribution  $\pi \neq p$



# Invariance and Detailed Balance

- **Invariance:** Each step leaves the distribution  $p$  invariant (we stay in  $p$ ):

$$p(\mathbf{x}') = \sum_x T(\mathbf{x}'|\mathbf{x})p(\mathbf{x}) \qquad p(\mathbf{x}') = \int T(\mathbf{x}'|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

Once we sample from  $p$ , the transition operator will not change this, i.e., we do not fall back to some funny distribution  $\pi \neq p$

- **Sufficient condition** for  $p$  being invariant:

**Detailed balance:**

$$p(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) = p(\mathbf{x}')T(\mathbf{x}|\mathbf{x}')$$

# Why is invariance not enough?

- ▶ Invariance only says something about the transitions once we have **reached** the stationary distribution.
- ▶ Invariance doesn't say anything about how the chain converges.

# Why is invariance not enough?

- ▶ Invariance only says something about the transitions once we have **reached** the stationary distribution.
- ▶ Invariance doesn't say anything about how the chain converges.

Trivial solutions leave  $p(\mathbf{x})$  invariant, e.g.  $T(\mathbf{x}_{t+1} | \mathbf{x}_t) = \delta(\mathbf{x}_{t+1} - \mathbf{x}_t)$ :

$$\int T(\mathbf{x}_{t+1} = \mathbf{x} | \mathbf{x}_t = \mathbf{x}') p(\mathbf{x}') d\mathbf{x}' = p(\mathbf{x}) \quad (7)$$

# Why is invariance not enough?

- ▶ Invariance only says something about the transitions once we have **reached** the stationary distribution.
- ▶ Invariance doesn't say anything about how the chain converges.

Trivial solutions leave  $p(\mathbf{x})$  invariant, e.g.  $T(\mathbf{x}_{t+1} | \mathbf{x}_t) = \delta(\mathbf{x}_{t+1} - \mathbf{x}_t)$ :

$$\int T(\mathbf{x}_{t+1} = \mathbf{x} | \mathbf{x}_t = \mathbf{x}') p(\mathbf{x}') d\mathbf{x}' = p(\mathbf{x}) \quad (7)$$

Ergodicity has a rather technical definition, but thankfully it is easy to guarantee!

# Ergodicity and communication

A Markov Chain is ergodic if there is some probability for any state to reach any state in bounded steps. If this is true, all states are said to **communicate**.

When designing Markov Chains, the easiest way to guarantee this is to have transitions that satisfy:

$$T(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}) > 0 \quad \forall \mathbf{x}^{(t+1)}, \mathbf{x}^{(t)} \quad (8)$$

Then, all states will communicate in 1 step.

# Metropolis-Hastings

- ▶ Assume that  $\tilde{p} = Zp$  can be evaluated easily
- ▶ **Proposal density**  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})$  depends on last sample  $\mathbf{x}^{(t)}$ .  
Example: Gaussian with mean  $\mathbf{x}^{(t)}$ :  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t)}, \Sigma)$

# Metropolis-Hastings

- ▶ Assume that  $\tilde{p} = Zp$  can be evaluated easily
- ▶ **Proposal density**  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})$  depends on last sample  $\mathbf{x}^{(t)}$ .  
Example: Gaussian with mean  $\mathbf{x}^{(t)}$ :  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t)}, \Sigma)$

## Metropolis-Hastings Algorithm

1. Generate proposal  $\mathbf{x}' \sim \hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})$
2. If

$$\frac{\hat{T}(\mathbf{x}^{(t)}|\mathbf{x}')\tilde{p}(\mathbf{x}')}{\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})\tilde{p}(\mathbf{x}^{(t)})} \geq u, \quad u \sim U[0, 1]$$

accept the sample  $\mathbf{x}^{(t+1)} = \mathbf{x}'$ . Otherwise set  $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$ .

# Metropolis-Hastings

- ▶ Assume that  $\tilde{p} = Zp$  can be evaluated easily
- ▶ **Proposal density**  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})$  depends on last sample  $\mathbf{x}^{(t)}$ .  
Example: Gaussian with mean  $\mathbf{x}^{(t)}$ :  $\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t)}, \Sigma)$

## Metropolis-Hastings Algorithm

1. Generate proposal  $\mathbf{x}' \sim \hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})$
2. If

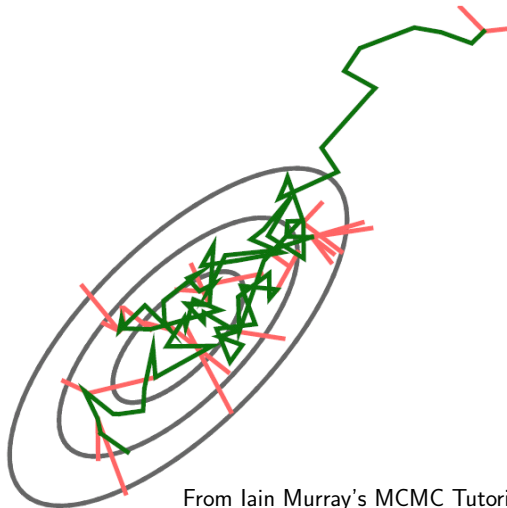
$$\frac{\hat{T}(\mathbf{x}^{(t)}|\mathbf{x}')\tilde{p}(\mathbf{x}')}{\hat{T}(\mathbf{x}'|\mathbf{x}^{(t)})\tilde{p}(\mathbf{x}^{(t)})} \geq u, \quad u \sim U[0, 1]$$

accept the sample  $\mathbf{x}^{(t+1)} = \mathbf{x}'$ . Otherwise set  $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$ .

- ▶  $q(\mathbf{x}^{(t)}) \xrightarrow{t \rightarrow \infty} p(\mathbf{x})$  ► Converge to equilibrium distribution
- ▶ If proposal distribution is symmetric: **Metropolis Algorithm** (Metropolis et al., 1953); Otherwise **Metropolis-Hastings**



# Example



From Iain Murray's MCMC Tutorial

# Step-Size Demo

- ▶ Explore  $p(x) = \mathcal{N}(x | 0, 1)$  for different step sizes  $\sigma$ .
- ▶ We can only evaluate  $\log \tilde{p}(x) = -x^2/2$
- ▶ Proposal distribution  $q$ : Gaussian  $\mathcal{N}(x^{(t+1)} | x^{(t)}, \sigma^2)$  centered at the current state for various step sizes  $\sigma$
- ▶ Expect to explore the space between  $-2, 2$  with high probability

# Step-Size Demo: Discussion

- ▶ Acceptance rate depends on the step size of the proposal distribution
  - ▶▶ Exploration parameter

# Step-Size Demo: Discussion

- ▶ Acceptance rate depends on the step size of the proposal distribution
  - ▶▶ Exploration parameter
- ▶ If we do not reject enough, the method does not work.

## Step-Size Demo: Discussion

- ▶ Acceptance rate depends on the step size of the proposal distribution
  - ▶▶ Exploration parameter
- ▶ If we do not reject enough, the method does not work.
- ▶ In rejection sampling we do not like rejections, but in MH rejections tell you where the target distribution is.

# Step-Size Demo: Discussion

- ▶ Acceptance rate depends on the step size of the proposal distribution
  - ▶▶ Exploration parameter
- ▶ If we do not reject enough, the method does not work.
- ▶ In rejection sampling we do not like rejections, but in MH rejections tell you where the target distribution is.
- ▶ Theoretical results: in 1D 44%, in higher dimensions about 25% acceptance rate for good mixing properties

# Step-Size Demo: Discussion

- ▶ Acceptance rate depends on the step size of the proposal distribution
  - ▶▶ Exploration parameter
- ▶ If we do not reject enough, the method does not work.
- ▶ In rejection sampling we do not like rejections, but in MH rejections tell you where the target distribution is.
- ▶ Theoretical results: in 1D 44%, in higher dimensions about 25% acceptance rate for good mixing properties
- ▶ Tune the step size

# Properties

- ▶ Samples are correlated



# Properties

- ▶ Samples are correlated
- ▶ If  $\hat{T} > 0$  everywhere, we will end up in the **equilibrium distribution**:  $\pi(\mathbf{x}^{(t)}) \xrightarrow{t \rightarrow \infty} p^*(\mathbf{x})$

# Properties

- ▶ Samples are correlated
- ▶ If  $\hat{T} > 0$  everywhere, we will end up in the **equilibrium distribution**:  $\pi(\mathbf{x}^{(t)}) \xrightarrow{t \rightarrow \infty} p^*(\mathbf{x})$
- ▶ Explore the state space by random walk
  - ▶▶ May take many steps, if the steps are short compared to the distribution
- ▶ No further catastrophic problems in high dimensions

# MCMC Diagnostics: Trace Plots

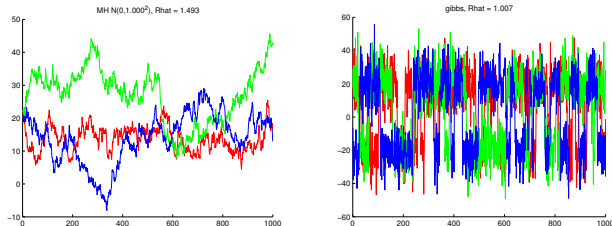


Figure from Murphy (2012)

- ▶ **Mixing time:** Amount of time it takes the Markov chain to converge to the stationary distribution and forget its initial state.
- ▶ **Trace plots:** Run multiple chains from very different starting points, plot the samples of the variables of interest. If the chain has mixed, the trace plots should converge to the same distribution.

# Summary

- ▶ MCMC generates a Markov chain of dependent samples that allow us to generate samples from the target distribution
- ▶ Metropolis Hastings algorithm

# Further Reading

- ▶ MacKay, ch 29
- ▶ Murphy, ch 24

# References I