STAT 237
Approximate Inference via Sampling

March 23-28, 2022

Colin Reimer Dawson

Outline

Outline

The (Prior) Predictive Distribution

- We used the product rule and marginalization to show that

$$p(x) = \int_{\text{Range}(\theta)} p(\theta)p(x \mid \theta)\ d\theta$$

  which is a **weighted average** of $p(x \mid \theta)$ values for each mu, weighted by the **prior**, $p(\theta)$, on θ

The (Prior) Predictive Distribution

- We used the product rule and marginalization to show that

$$p(x) = \int_{\text{Range}(\theta)} p(\theta)p(x \mid \theta)\ d\theta$$

  which is a **weighted average** of $p(x \mid \theta)$ values for each mu, weighted by the **prior**, $p(\theta)$, on θ
- In other words, we can write

$$p(x) = \mathbb{E}\left[p(x \mid \theta)\right]$$

  where the expectation is taken with respect to the prior

The (Prior) Predictive Distribution

- This works when we have multiple observations as well

$$\begin{aligned} p(x_1, \ldots, x_N) &= \int_{\text{Range}(\theta)} p(\theta) p(x_1, \ldots, x_N \mid \theta) \, d\theta \\ &= \mathbb{E}\left[p(x_1, \ldots, x_N \mid \theta)\right] \end{aligned}$$

  where the expectation is taken with respect to the prior

The (Prior) Predictive Distribution

- This works when we have multiple observations as well

$$\begin{aligned} p(x_1, \ldots, x_N) &= \int_{\text{Range}(\theta)} p(\theta) p(x_1, \ldots, x_N \mid \theta) \, d\theta \\ &= \mathbb{E}\left[p(x_1, \ldots, x_N \mid \theta)\right] \end{aligned}$$

  where the expectation is taken with respect to the prior
- Note that both of these are instances of taking the expected value of a **function** of θ, because in $p(x \mid \theta)$, x acts as a constant

The (Prior) Predictive Distribution

- The example we looked at in lab involved a conditional Bernoulli PMF

$$\begin{aligned} p(x) &= \int_0^1 p(\mu) p(x \mid \mu) \, d\mu \\ &= \mathbb{E}\left[p(x \mid \mu)\right] \end{aligned}$$

with $p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$

The (Prior) Predictive Distribution

- The example we looked at in lab involved a conditional Bernoulli PMF

$$\begin{aligned} p(x) &= \int_0^1 p(\mu) p(x \mid \mu) \, d\mu \\ &= \mathbb{E}\left[p(x \mid \mu)\right] \end{aligned}$$

  with $p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$

- That is, if $x = 1$, we are asking for $\mathbb{E}\left[\mu\right]$, and if $x = 0$ we are asking for $\mathbb{E}\left[1-\mu\right]$

The (Prior) Predictive Distribution

- The example we looked at in lab involved a conditional Bernoulli PMF

$$p(x) = \int_0^1 p(\mu)p(x \mid \mu)\, d\mu$$
$$= \mathbb{E}\left[p(x \mid \mu)\right]$$

  with $p(x \mid \mu) = \mu^x(1-\mu)^{1-x}$
- That is, if $x = 1$, we are asking for $\mathbb{E}\left[\mu\right]$, and if $x = 0$ we are asking for $\mathbb{E}\left[1-\mu\right]$
- If $p(\mu)$ is a Beta(a, b) distribution, then

$$\mathbb{E}\left[\mu\right] = \frac{a}{a+b} \qquad \mathbb{E}\left[1-\mu\right] = \frac{b}{a+b}$$

The (Posterior) Predictive Distribution

- By the same logic

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_{\text{Range}(\theta)} p(\theta \mid x_{\text{old}}) p(x_{\text{new}} \mid \theta, x_{\text{old}}) \, d\theta$$

The (Posterior) Predictive Distribution

- By the same logic

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_{\text{Range}(\theta)} p(\theta \mid x_{\text{old}}) p(x_{\text{new}} \mid \theta, x_{\text{old}}) \; d\theta$$

- If x_{old} and x_{new} are **conditionally independent** given θ, then this is

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_{\text{Range}(\theta)} p(\theta \mid x_{\text{old}}) p(x_{\text{new}} \mid \theta) \; d\theta$$

  which is a **weighted average** of $p(x_{\text{new}} \mid \theta)$ values for each μ, weighted by the **posterior**, $p(\theta \mid x_{\text{old}})$

The (Posterior) Predictive Distribution

- By the same logic

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_{\text{Range}(\theta)} p(\theta \mid x_{\text{old}}) p(x_{\text{new}} \mid \theta, x_{\text{old}}) \, d\theta$$

- If x_{old} and x_{new} are **conditionally independent** given θ, then this is

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_{\text{Range}(\theta)} p(\theta \mid x_{\text{old}}) p(x_{\text{new}} \mid \theta) \, d\theta$$

  which is a **weighted average** of $p(x_{\text{new}} \mid \theta)$ values for each μ, weighted by the **posterior**, $p(\theta \mid x_{\text{old}})$
- In other words, we can write

$$p(x_{\text{new}} \mid x_{\text{old}}) = \mathbb{E}\left[p(x_{\text{new}} \mid \theta)\right]$$

  where the expectation is taken with respect to the **posterior**

The (Posterior) Predictive Distribution

- The example we looked at in lab involved a conditional Bernoulli PMF

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_0^1 p(\mu \mid x_{\text{old}}) p(x_{\text{new}} \mid \mu) \, d\mu$$
$$= \mathbb{E}\left[p(x_{\text{new}} \mid \mu)\right]$$

with $p(x_{\text{new}} \mid \mu) = \mu^{x_{\text{new}}}(1-\mu)^{1-x_{\text{new}}}$

The (Posterior) Predictive Distribution

- The example we looked at in lab involved a conditional Bernoulli PMF

$$p(x_{\text{new}} \mid x_{\text{old}}) = \int_0^1 p(\mu \mid x_{\text{old}}) p(x_{\text{new}} \mid \mu) \, d\mu$$
$$= \mathbb{E}\left[p(x_{\text{new}} \mid \mu)\right]$$

  with $p(x_{\text{new}} \mid \mu) = \mu^{x_{\text{new}}}(1-\mu)^{1-x_{\text{new}}}$

- If $p(\mu \mid x_{\text{old}})$ is a Beta(a, b) distribution, then

$$\mathbb{E}\left[\mu \mid x_{\text{old}}\right] = \frac{a}{a+b} \qquad \mathbb{E}\left[1-\mu \mid x_{\text{old}}\right] = \frac{b}{a+b}$$

Outline

Beyond Means and Conjugate Priors

- When we have a Beta posterior and the quantity we want the expected value of is something simple like μ itself, we can get these values analytically

Beyond Means and Conjugate Priors

- When we have a Beta posterior and the quantity we want the expected value of is something simple like μ itself, we can get these values analytically
- What do we do if the function of μ we care about is more complicated, and/or our posterior distribution isn't part of a recognizable family?

Beyond Means and Conjugate Priors

- When we have a Beta posterior and the quantity we want the expected value of is something simple like μ itself, we can get these values analytically
- What do we do if the function of μ we care about is more complicated, and/or our posterior distribution isn't part of a recognizable family?
- For example, we might have reason to use a non-conjugate prior, or we might have more than one parameter, such that the joint posterior over the parameters is difficult to work with

The Law of Large Numbers

If we generate S independent observations according to the distribution of θ, and apply the function f to each observation, then **the "sample" mean of the $f(\theta)$s** "approaches" $\mathbb{E}[f(\theta)]$ as S increases:

$$\frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)}) \to \mathbb{E}[f(\theta)] \text{ as } S \to \infty$$

Notation Note:

It's conventional to use the superscript with parentheses to denote a **simulated** value — this isn't an exponent.

$$\theta^{(s)} := \text{ the } s^{\text{th}} \text{ simulated value of } \theta$$

Simulation Demonstration of the LLN

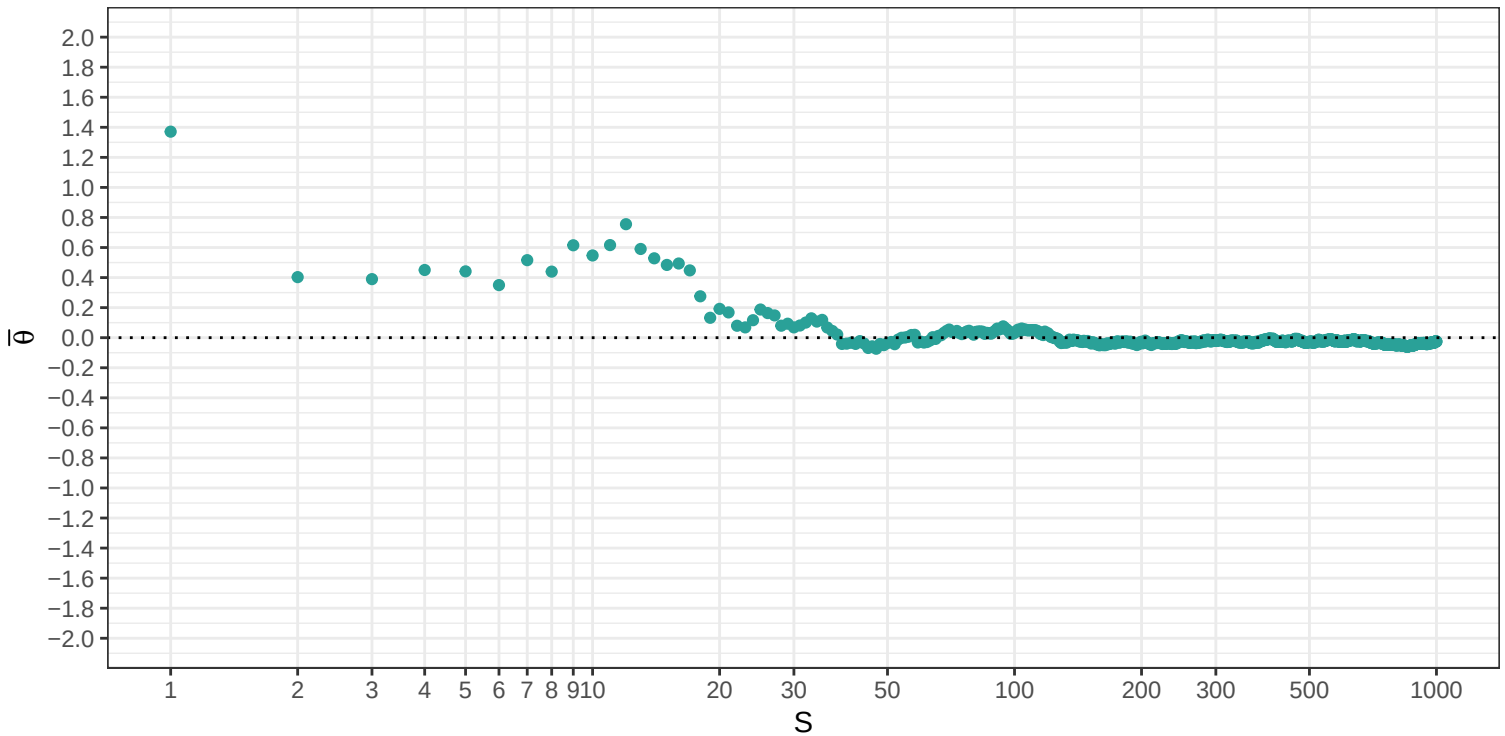


Figure: Simulation of $\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$, for various S, where $\theta^{(s)} \ldots \theta^{(1000)}$ are sampled independently from a $\mathcal{N}(0,1)$ distribution

Sampling from the Posterior

- Therefore, provided we can **sample** from our posterior distribution, $p(\theta \mid x_1, \ldots, x_N)$, we can estimate the expected value of various functions of θ
- For example, if we want $\mathbb{E}\left[p(x_{\text{new}} \mid \mu)\right]$, we could approximate it via

$$\mathbb{E}\left[p(x_{\text{new}} \mid \mu)\right] \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}\left[p(x_{\text{new}} \mid \mu^{(s)})\right]$$

Outline

Approaches to sample/integrate against a given distribution

- How can we get our samples $\theta^{(1)}, \ldots, \theta^{(S)}$?

Approaches to sample/integrate against a given distribution

- How can we get our samples $\theta^{(1)}, \ldots, \theta^{(S)}$?
- Three approaches:

Approaches to sample/integrate against a given distribution

- How can we get our samples $\theta^{(1)}, \ldots, \theta^{(S)}$?
- Three approaches:
  1. Inverse CDF Method

Approaches to sample/integrate against a given distribution

- How can we get our samples $\theta^{(1)}, \ldots, \theta^{(S)}$?
- Three approaches:
  1. Inverse CDF Method
  2. Rejection Sampling

Approaches to sample/integrate against a given distribution

- How can we get our samples $\theta^{(1)}, \ldots, \theta^{(S)}$?
- Three approaches:
  1. Inverse CDF Method
  2. Rejection Sampling
  3. Markov Chain Monte Carlo

Outline

Inverse CDF Method

- Recall, the CDF, $F(x)$, gives us $P(X \leq x)$ for every x.

Inverse CDF Method

- Recall, the CDF, $F(x)$, gives us $P(X \leq x)$ for every x.
- The inverse, $F^{-1}(p)$ returns the value of x such that $P(X \leq x) = p$. That is, it returns the pth quantile of the distribution of X.

Inverse CDF Method

- Recall, the CDF, $F(x)$, gives us $P(X \leq x)$ for every x.
- The inverse, $F^{-1}(p)$ returns the value of x such that $P(X \leq x) = p$. That is, it returns the pth quantile of the distribution of X.
- Since the quantiles uniquely define a distribution, if we can generate a variable with the same quantiles as X, it will have the same distribution as X.

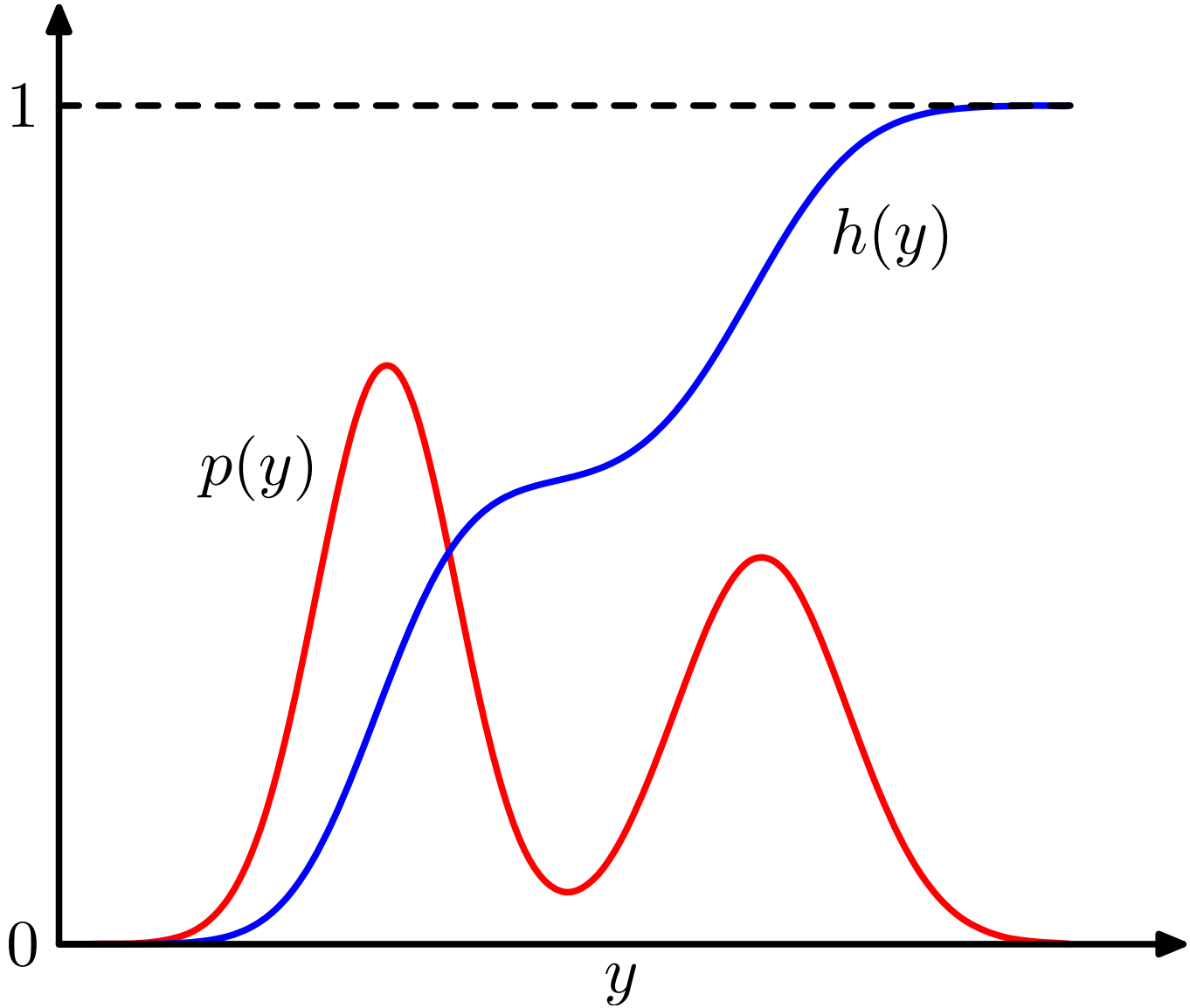
Inverse CDF Method

- Recall, the CDF, $F(x)$, gives us $P(X \leq x)$ for every x.
- The inverse, $F^{-1}(p)$ returns the value of x such that $P(X \leq x) = p$. That is, it returns the pth quantile of the distribution of X.
- Since the quantiles uniquely define a distribution, if we can generate a variable with the same quantiles as X, it will have the same distribution as X.
- Solution: Generate $U \sim \text{Unif}(0, 1)$, and return $F^{-1}(U)$

Inverse CDF Method

- Recall, the CDF, $F(x)$, gives us $P(X \leq x)$ for every x.
- The inverse, $F^{-1}(p)$ returns the value of x such that $P(X \leq x) = p$. That is, it returns the pth quantile of the distribution of X.
- Since the quantiles uniquely define a distribution, if we can generate a variable with the same quantiles as X, it will have the same distribution as X.
- Solution: Generate $U \sim \text{Unif}(0, 1)$, and return $F^{-1}(U)$
- As long as F^{-1} is defined, this will produce samples distributed as X

Inverse CDF Method



Example: Exponential Distribution

A positive real-valued random variable X has an **exponential distribution** with parameter λ if its PDF is

$$p(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

Example: Exponential Distribution

A positive real-valued random variable X has an **exponential distribution** with parameter λ if its PDF is

$$p(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

The CDF is

$$F(x_0) = \int_0^{x_0} f(x)dx = 1 - e^{-\lambda x_0}$$

Example: Exponential Distribution

A positive real-valued random variable X has an **exponential distribution** with parameter λ if its PDF is

$$p(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

The CDF is

$$F(x_0) = \int_0^{x_0} f(x)dx = 1 - e^{-\lambda x_0}$$

The inverse CDF is

$$x = F^{-1}(p) = -\frac{\log(1-p)}{\lambda}$$

Example: Exponential Distribution

A positive real-valued random variable X has an **exponential distribution** with parameter λ if its PDF is

$$p(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

The CDF is

$$F(x_0) = \int_0^{x_0} f(x)dx = 1 - e^{-\lambda x_0}$$

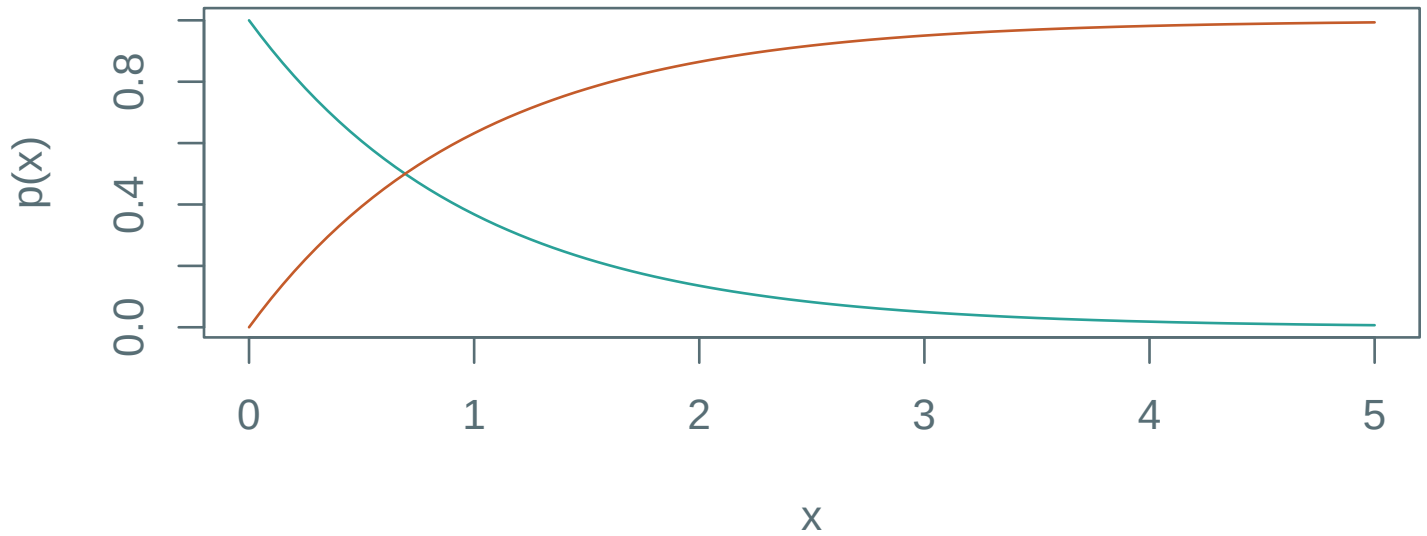
The inverse CDF is

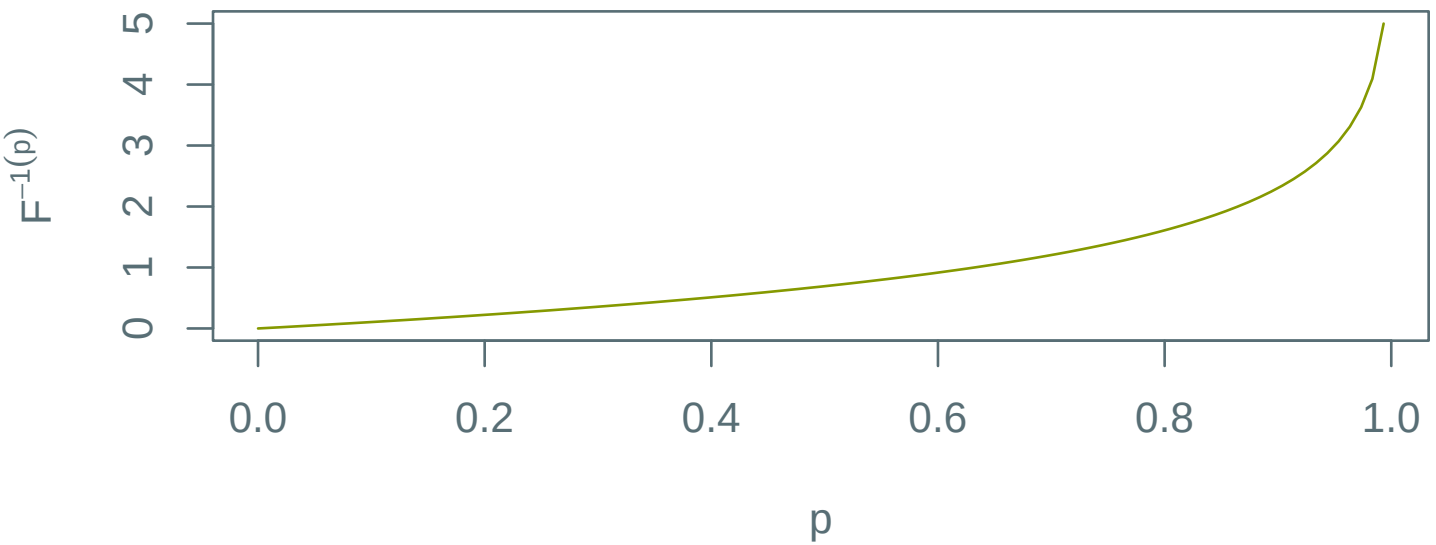
$$x = F^{-1}(p) = -\frac{\log(1-p)}{\lambda}$$

Hence we can sample X values by generating $U \sim \text{Unif}(0,1)$ and returning

$$x = -\frac{\log(1-U)}{\lambda}$$

Example: Exponential Distribution





Discrete Case

- When the variable is discrete (e.g., Poisson, Binomial, etc.), the CDF has flat regions and gaps; so the inverse is not well-defined.

Discrete Case

- When the variable is discrete (e.g., Poisson, Binomial, etc.), the CDF has flat regions and gaps; so the inverse is not well-defined.
- Big jumps occur at values that have large probability; that is, the "left-hand edge" of a flat region is the value that "owns" the gap just below its CDF value.

Discrete Case

- When the variable is discrete (e.g., Poisson, Binomial, etc.), the CDF has flat regions and gaps; so the inverse is not well-defined.
- Big jumps occur at values that have large probability; that is, the "left-hand edge" of a flat region is the value that "owns" the gap just below its CDF value.
- So we can modify our algorithm by having our "pseudoinverse" map "gap values" to the value just above the gap.

Discrete Case

- When the variable is discrete (e.g., Poisson, Binomial, etc.), the CDF has flat regions and gaps; so the inverse is not well-defined.
- Big jumps occur at values that have large probability; that is, the "left-hand edge" of a flat region is the value that "owns" the gap just below its CDF value.
- So we can modify our algorithm by having our "pseudoinverse" map "gap values" to the value just above the gap.
- Sample $U \sim \text{Unif}(0, 1)$ and return

$$\min\{x : F(x) \geq U\}$$

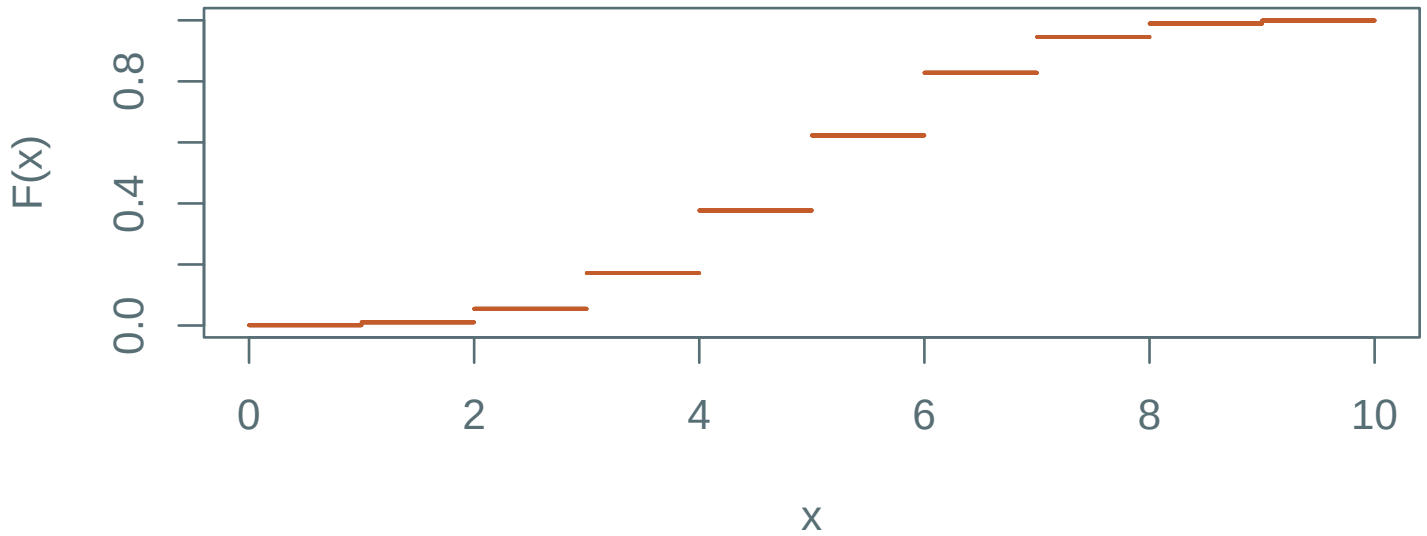
Discrete Case

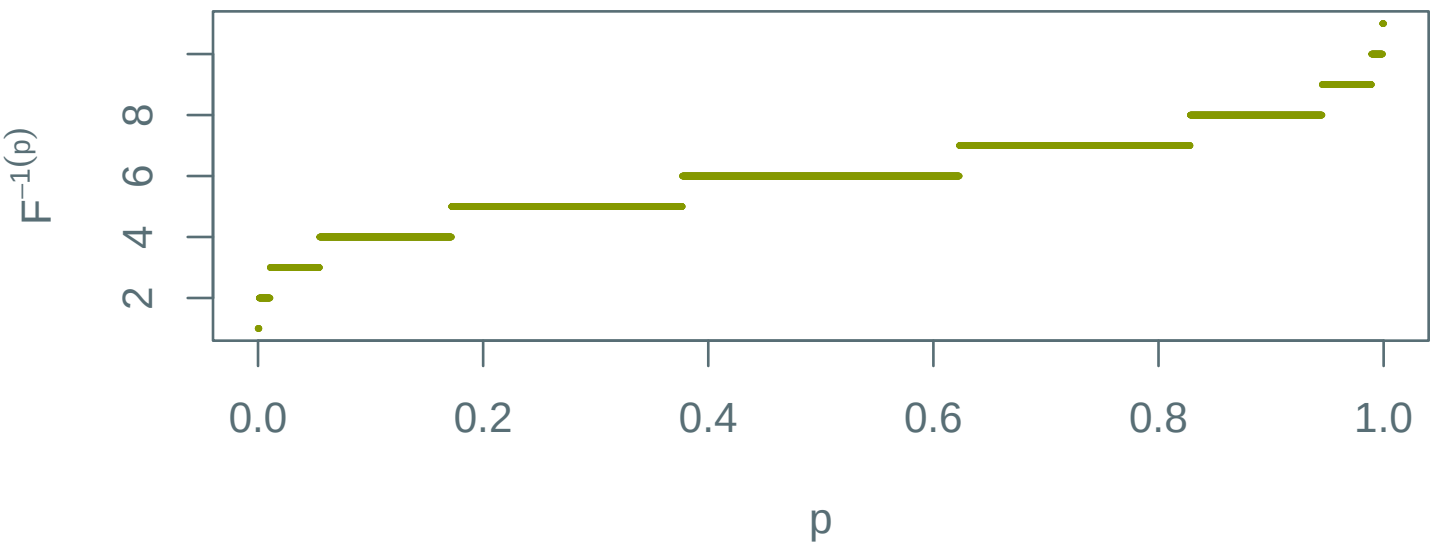
- When the variable is discrete (e.g., Poisson, Binomial, etc.), the CDF has flat regions and gaps; so the inverse is not well-defined.
- Big jumps occur at values that have large probability; that is, the "left-hand edge" of a flat region is the value that "owns" the gap just below its CDF value.
- So we can modify our algorithm by having our "pseudoinverse" map "gap values" to the value just above the gap.
- Sample $U \sim \text{Unif}(0, 1)$ and return

$$\min\{x : F(x) \geq U\}$$

- I.e., find the two x values on either side of the "gap" enclosing U, and choose the upper one.

Example: Binomial Distribution





Outline

Rejection Sampling

- The inverse CDF method is great when we can find an inverse CDF. But (a) this only has any hope for 1D distributions, and (b) even then, it is often intractable to invert the CDF analytically.

Rejection Sampling

- The inverse CDF method is great when we can find an inverse CDF. But (a) this only has any hope for 1D distributions, and (b) even then, it is often intractable to invert the CDF analytically.
- An alternative approach: **rejection sampling**

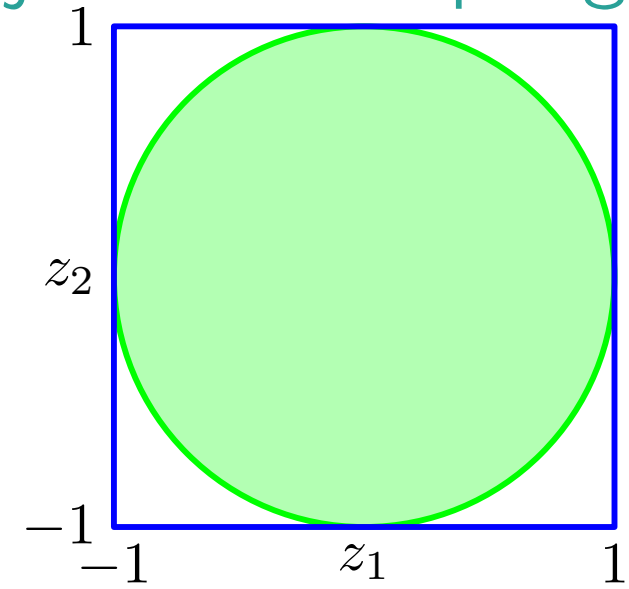
Rejection Sampling

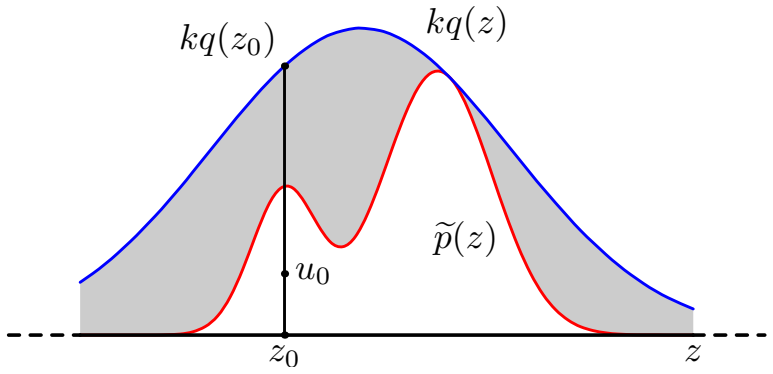
- The inverse CDF method is great when we can find an inverse CDF. But (a) this only has any hope for 1D distributions, and (b) even then, it is often intractable to invert the CDF analytically.
- An alternative approach: **rejection sampling**
- Goal: sample from some density, $p(x)$.

Rejection Sampling

- The inverse CDF method is great when we can find an inverse CDF. But (a) this only has any hope for 1D distributions, and (b) even then, it is often intractable to invert the CDF analytically.
- An alternative approach: **rejection sampling**
- Goal: sample from some density, $p(x)$.
- Idea: find a similar distribution, $q(x)$ to sample from, and "filter" the results using a "p-shaped" filter.

Rejection Sampling





Rejection Sampling Algorithm

1. Choose a "proposal density" q "similar" to target p.
2. Find a scaling constant k so that $k \cdot q(x)$ is at or above $p(x)$ for all x.
3. Sample x^* from q
4. Accept x^* with probability $\frac{p(x)}{k \cdot q(x)}$; otherwise, reject, and try again until acceptance.

Example: Truncation

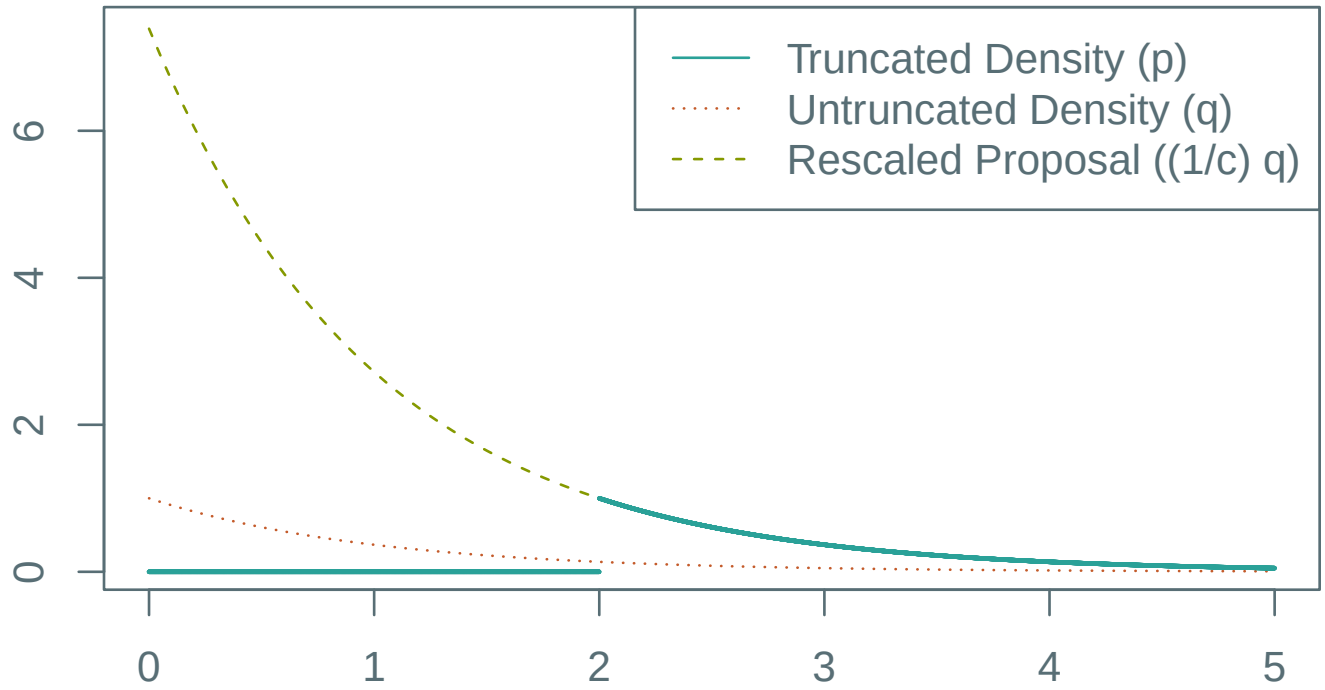
- Sometimes we want to sample from a distribution which looks like a known distribution, except it has a restricted range.

Example: Truncation

- Sometimes we want to sample from a distribution which looks like a known distribution, except it has a restricted range.
- E.g., detection limit θ with a uniform likelihood and a Gamma prior:

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \qquad \theta > 0$$
$$p(y \mid \theta) = \theta^{-1} \qquad \theta \geq y$$
$$p(\theta \mid y) \propto \frac{b^a}{\Gamma(a)}\theta^{a-1-1}e^{-b\theta}, \qquad \theta \geq y$$

Example: Truncation



1. $p(x) = c \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1-1} e^{-b\theta}, \qquad \theta \geq y$
2. Choose $q(x) = \frac{b^a}{\Gamma(a)} \theta^{a-1-1} e^{-b\theta}, \qquad \theta \geq 0$
3. Then $p(x)/cq(x) = 1$ for all $\theta \geq 0$, and 0 otherwise.
4. So, generate x from $q(x)$, and accept if $\theta \geq y$; reject otherwise.

Drawbacks of Rejection Sampling

1. For high-dimensional distributions, it's very hard to find a good proposal.
2. It can be quite difficult to find a valid rescaling constant, without making the rejection probability unacceptably high.

Outline

Non-Independent Samples

- In practice, generating independent samples is often intractable

Non-Independent Samples

- In practice, generating independent samples is often intractable
- We can often generate **correlated** samples, however

Non-Independent Samples

- In practice, generating independent samples is often intractable
- We can often generate **correlated** samples, however
- Idea: Use the current value to "seed" the next one

Sequential vs Independent Samples

- With **independent** sampling methods, the parameter value $\theta^{(s)}$ generated at each iteration comes **directly** from the target distribution (such as the posterior)

$$\theta^{(s)} \sim p(\theta)$$

Sequential vs Independent Samples

- With **independent** sampling methods, the parameter value $\theta^{(s)}$ generated at each iteration comes **directly** from the target distribution (such as the posterior)

$$\theta^{(s)} \sim p(\theta)$$

- With **sequential** methods, we sample $\theta^{(s)}$ from a distribution that depends on $\theta^{(s-1)}$

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

Sequential vs Independent Samples

- With **independent** sampling methods, the parameter value $\theta^{(s)}$ generated at each iteration comes **directly** from the target distribution (such as the posterior)

$$\theta^{(s)} \sim p(\theta)$$

- With **sequential** methods, we sample $\theta^{(s)}$ from a distribution that depends on $\theta^{(s-1)}$

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

- **Goal**: Choose p^* so that the **marginal** distribution of $\theta^{(s)}$ is the target, $p(\theta)$

"Stationary" Distributions

- **Procedure:** Sample

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

"Stationary" Distributions

- **Procedure:** Sample

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

- **Goal:** Choose p^* so that the **marginal** distribution $p^*(\theta^{(s)})$ is the same as the target, $p(\theta)$

"Stationary" Distributions

- **Procedure:** Sample

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

- **Goal:** Choose p^* so that the **marginal** distribution $p^*(\theta^{(s)})$ is the same as the target, $p(\theta)$
- For this to work, we must have

$$p(\theta^{(s)}) = \int_\Theta p^*(\theta^{(s-1)}, \theta^{(s)}) \, d\theta^{(s-1)}$$

"Stationary" Distributions

- **Procedure:** Sample

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

- **Goal:** Choose p^* so that the **marginal** distribution $p^*(\theta^{(s)})$ is the same as the target, $p(\theta)$
- For this to work, we must have

$$p(\theta^{(s)}) = \int_\Theta p^*(\theta^{(s-1)}, \theta^{(s)}) \, d\theta^{(s-1)}$$

- Assuming it works at $s-1$, this becomes (product rule)

$$p(\theta^{(s)}) = \int_\Theta p^*(\theta^{(s)} \mid \theta^{(s-1)}) p(\theta^{(s-1)}) \, d\theta^{(s-1)}$$

"Stationary" Distributions

- **Procedure:** Sample

$$\theta^{(s)} \sim p^*(\theta^{(s)} \mid \theta^{(s-1)})$$

- **Goal:** Choose p^* so that the **marginal** distribution $p^*(\theta^{(s)})$ is the same as the target, $p(\theta)$
- For this to work, we must have

$$p(\theta^{(s)}) = \int_\Theta p^*(\theta^{(s-1)}, \theta^{(s)}) \, d\theta^{(s-1)}$$

- Assuming it works at $s-1$, this becomes (product rule)

$$p(\theta^{(s)}) = \int_\Theta p^*(\theta^{(s)} \mid \theta^{(s-1)}) p(\theta^{(s-1)}) \, d\theta^{(s-1)}$$

- In other words, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$ **preserves** $p(\theta)$ once it "finds" it

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:
     (i) Flip the coin

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:
     (i) Flip the coin
     (ii) If heads, set $\theta^{(s)} = \theta^{(s-1)} + 1$ (wrapping around so that 6 goes to 1)

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:
     (i) Flip the coin
     (ii) If heads, set $\theta^{(s)} = \theta^{(s-1)} + 1$ (wrapping around so that 6 goes to 1)
     (iii) If tails, set $\theta^{(s)} = \theta^{(s-1)} - 1$ (wrapping around so that 1 goes to 6)

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:
     (i) Flip the coin
     (ii) If heads, set $\theta^{(s)} = \theta^{(s-1)} + 1$ (wrapping around so that 6 goes to 1)
     (iii) If tails, set $\theta^{(s)} = \theta^{(s-1)} - 1$ (wrapping around so that 1 goes to 6)
- What is the **conditional** distribution, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$?

Random Walks

A simple example:

- We want to simulate rolling a fair six-sided die using only a fair coin
- An algorithm:
  1. Pick some initial $\theta^{(0)}$ (somehow)
  2. For $s = 1, \ldots S$:
     (i) Flip the coin
     (ii) If heads, set $\theta^{(s)} = \theta^{(s-1)} + 1$ (wrapping around so that 6 goes to 1)
     (iii) If tails, set $\theta^{(s)} = \theta^{(s-1)} - 1$ (wrapping around so that 1 goes to 6)
- What is the **conditional** distribution, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$?
- What is the **marginal** distribution, $p^*(\theta^{(s)})$?

Random Walks

- What is the **conditional** distribution, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$?

$$p^*(\theta^{(s)} \mid \theta^{(s-1)}) = \begin{cases} 0.5 & \theta^{(s)} = \theta^{(s-1)} - 1 \\ 0.5 & \theta^{(s)} = \theta^{(s+1)} - 1 \\ 0 & \text{otherwise} \end{cases}$$

Random Walks

- What is the **conditional** distribution, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$?

$$p^*(\theta^{(s)} \mid \theta^{(s-1)}) = \begin{cases} 0.5 & \theta^{(s)} = \theta^{(s-1)} - 1 \\ 0.5 & \theta^{(s)} = \theta^{(s+1)} - 1 \\ 0 & \text{otherwise} \end{cases}$$

- What is the **marginal** distribution, $p^*(\theta^{(s)})$?

Random Walks

- What is the **conditional** distribution, $p^*(\theta^{(s)} \mid \theta^{(s-1)})$?

$$p^*(\theta^{(s)} \mid \theta^{(s-1)}) = \begin{cases} 0.5 & \theta^{(s)} = \theta^{(s-1)} - 1 \\ 0.5 & \theta^{(s)} = \theta^{(s+1)} - 1 \\ 0 & \text{otherwise} \end{cases}$$

- What is the **marginal** distribution, $p^*(\theta^{(s)})$?
- Depends on the marginal distribution of $\theta^{(s-1)}$

Random Walks

- What is the **marginal** distribution, $p^*(\theta^{(s)})$ if $\theta^{(s-1)}$ is distributed uniformly over 1 to 6; i.e., if it has the target distribution?

Random Walks

- What is the **marginal** distribution, $p^*(\theta^{(s)})$ if $\theta^{(s-1)}$ is distributed uniformly over 1 to 6; i.e., if it has the target distribution?
- We can use the product and sum rules to find out. Take $\theta^{(s)} = 4$ for instance:

$$p^*(\theta^{(s)} = 4) = \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)}) p^*(\theta^{(s-1)})$$

$$= \frac{1}{6} \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)})$$

Random Walks

- What is the **marginal** distribution, $p^*(\theta^{(s)})$ if $\theta^{(s-1)}$ is distributed uniformly over 1 to 6; i.e., if it has the target distribution?
- We can use the product and sum rules to find out. Take $\theta^{(s)} = 4$ for instance:

$$p^*(\theta^{(s)} = 4) = \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)}) p^*(\theta^{(s-1)})$$

$$= \frac{1}{6} \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)})$$

- What is $p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)})$ for each $\theta^{(s-1)}$ from 1 to 6?

Random Walks

- What is the **marginal** distribution, $p^*(\theta^{(s)})$ if $\theta^{(s-1)}$ is distributed uniformly over 1 to 6; i.e., if it has the target distribution?
- We can use the product and sum rules to find out. Take $\theta^{(s)} = 4$ for instance:

$$p^*(\theta^{(s)} = 4) = \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)}) p^*(\theta^{(s-1)})$$

$$= \frac{1}{6} \sum_{\theta^{(s-1)}=1}^{6} p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)})$$

- What is $p^*(\theta^{(s)} = 4 \mid \theta^{(s-1)})$ for each $\theta^{(s-1)}$ from 1 to 6?
- **Answer:** 0, except if $\theta^{(s-1)} = 3$ or 5. In that case, $1/2$

$$p^*(\theta^{(s)} = 4) = \frac{1}{6}(0 + 0 + \frac{1}{2} + 0 + \frac{1}{2} + 0)$$

$$= \frac{1}{6}$$

Random Walks

- The same logic holds for other values

Random Walks

- The same logic holds for other values
- So, once $\theta^{(s-1)}$ has the right distribution, our algorithm **preserves** it

Random Walks

- The same logic holds for other values
- So, once $\theta^{(s-1)}$ has the right distribution, our algorithm **preserves** it
- But if we could set $\theta^{(0)}$ to the right distribution, we wouldn't be doing this...

Random Walks

- The same logic holds for other values
- So, once $\theta^{(s-1)}$ has the right distribution, our algorithm **preserves** it
- But if we could set $\theta^{(0)}$ to the right distribution, we wouldn't be doing this...
- What happens if $\theta^{(0)} = 1$ with probability 1?

Random Walks

- The distribution of $\theta^{(1)}$ is then

$$p^*(\theta^{(s)} \mid \theta^{(s-1)}) = \begin{cases} \frac{1}{2} & \theta^{(s)} = 2 \\ \frac{1}{2} & \theta^{(s)} = 6 \\ 0 & \text{otherwise} \end{cases}$$

Random Walks

- The distribution of $\theta^{(1)}$ is then

$$p^*(\theta^{(s)} \mid \theta^{(s-1)}) = \begin{cases} \frac{1}{2} & \theta^{(s)} = 2 \\ \frac{1}{2} & \theta^{(s)} = 6 \\ 0 & \text{otherwise} \end{cases}$$

- The distribution of $\theta^{(2)}$ is

$$p^*(\theta^{(2)}) = \begin{cases} \frac{1}{2} \cdot \frac{1}{2} & \theta^{(s)} = 5 \\ \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} & \theta^{(s)} = 1 \\ \frac{1}{2} \cdot \frac{1}{2} & \theta^{(s)} = 3 \end{cases}$$

Random Walks

- The distribution of $\theta^{(3)}$ is then

$$p^*(\theta^{(3)}) = \begin{cases} \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} & \theta^{(s)} = 4 \\ \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} & \theta^{(s)} = 6 \\ \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} & \theta^{(s)} = 2 \end{cases}$$