Particle-based Approximate Inference

Inference

Professor Ajoodha

Lecture 6

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

ExplainableAI Lab

— MODELLING. DECISION MAKING. CAUSALITY —

Particle-based Methods

- In the previous lecture we looked at one class of approximate inference: Cluster-based message passing.
- We will now look at a completely different approach to approximate inference: Particle-based methods.
- We approximate the joint distribution **as a set of instantiations** (particles) to all or some of the variables in the network.
- We are mainly concerned about two issues:
  1. The notion of particles (i.e. full or collapsed?)
  2. The generation of particles (deterministic or distribution?)

Problem Statement

Problem

Given $P(\mathcal{X})$ *we want to estimate* $\mathbf{Y} = \mathbf{y}$ *relative to* P *for some* $\mathbf{Y} \subseteq \mathcal{X}$ *and* $\mathbf{y} \in Val(\mathbf{Y})$.

More generally we may want to compute the expectation of some $f(\mathcal{X})$.

We can choose $f(\xi) = \mathbb{1}\{\xi\langle\mathbf{Y}\rangle = \mathbf{y}\}$, where $\langle\mathbf{Y}\rangle$ is the assignment in ξ to the variables in $\mathbf{Y}$.

We can approximate the distribution by:

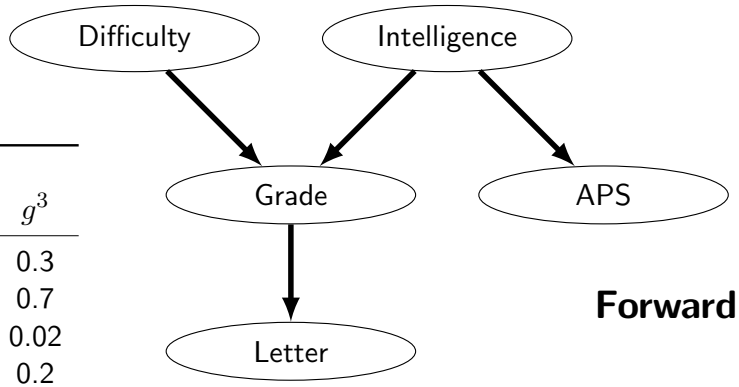
1. Generating a set of M particles
2. Estimating the value of the function (expectation relative to each of the generated particles);
3. Then aggregating the result.

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

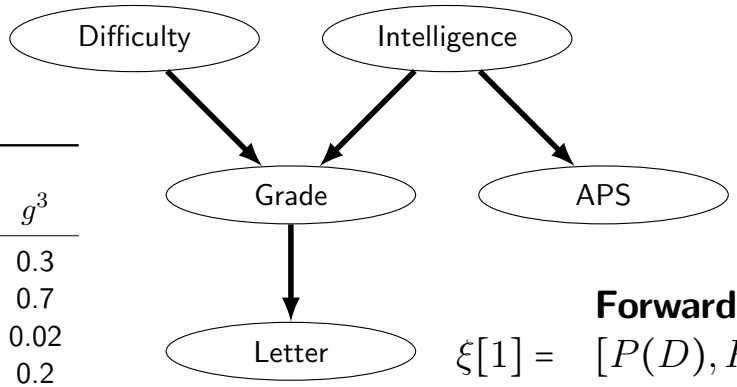
Forward Sampling example:

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

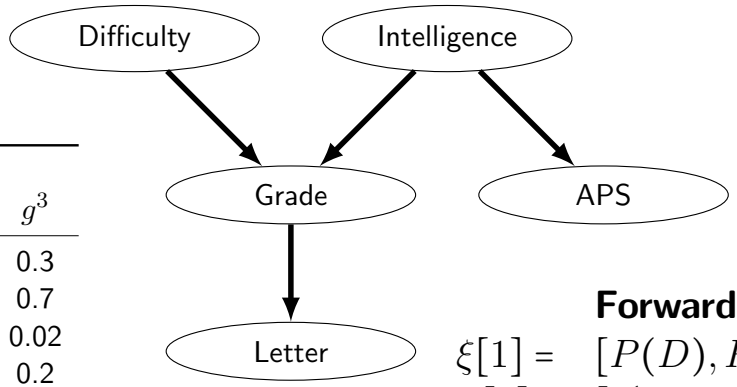
$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

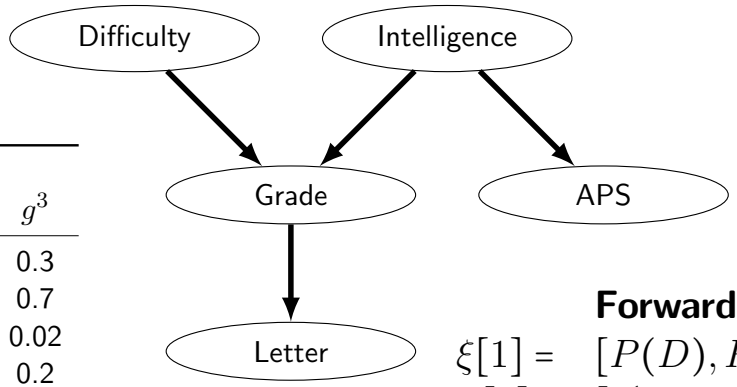
$\xi[1] = [d^1, \ , \ , \ , \ ]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

$\xi[1] = [d^1, \ , \ , \ , \ ]$

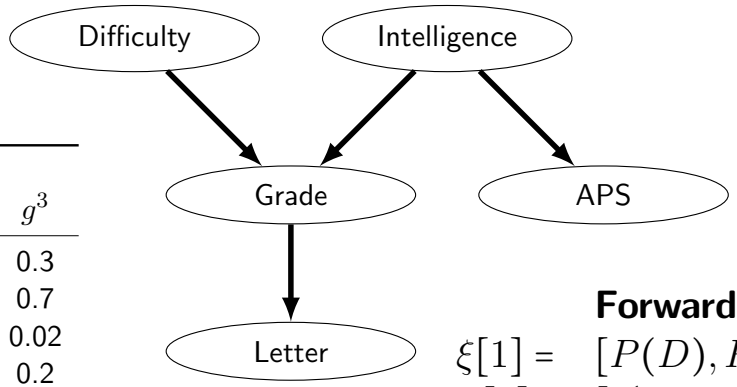
$\xi[1] = [d^1, i^0, \ , \ , \ ]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

$\xi[1] = [d^1, \ , \ , \ , \ ]$

$\xi[1] = [d^1, i^0, \ , \ , \ ]$

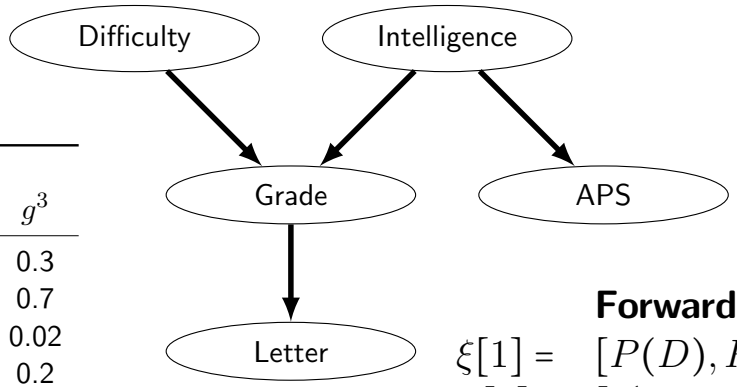
$\xi[1] = [d^1, i^0, g^2, \ , \ ]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

$\xi[1] = [d^1, \ , \ , \ , \ ]$

$\xi[1] = [d^1, i^0, \ , \ , \ ]$

$\xi[1] = [d^1, i^0, g^2, \ , \ ]$

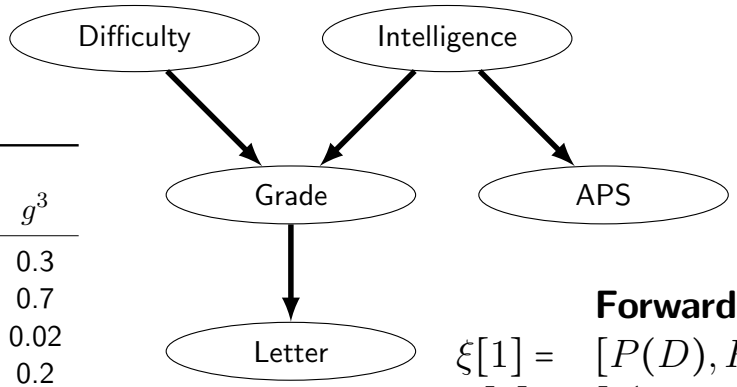
$\xi[1] = [d^1, i^0, g^2, a^0, \ ]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Forward Sampling example:

$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$

$\xi[1] = [d^1, \ , \ , \ , \ ]$

$\xi[1] = [d^1, i^0, \ , \ , \ ]$

$\xi[1] = [d^1, i^0, g^2, \ , \ ]$

$\xi[1] = [d^1, i^0, g^2, a^0, \ ]$

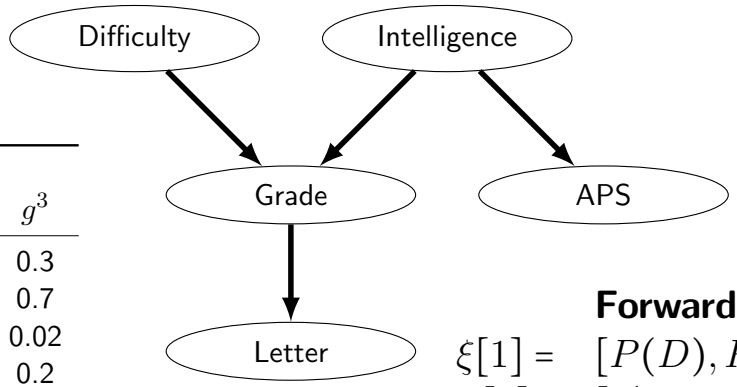
$\xi[1] = [d^1, i^0, g^2, a^0, l^1]$

The Student Example

Forward Sampling is a simple approach to generate random particles $(\xi[1], \ldots, \xi[M])$ from a distribution $P(\mathcal{X})$.

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

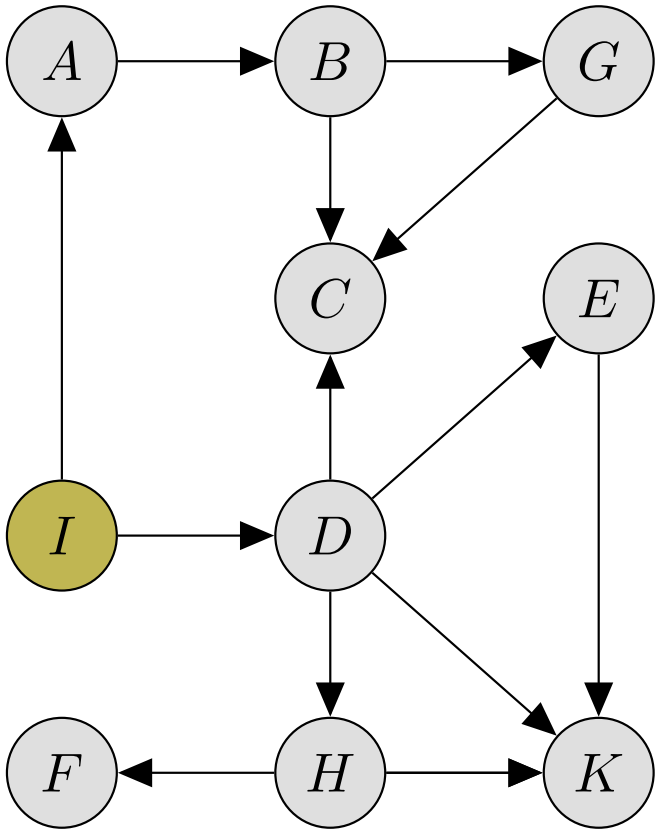
Forward Sampling example:

$$\xi[1] = [P(D), P(I), P(G \mid D, I), P(A \mid I), P(L \mid G)]$$
$$\xi[1] = [d^1, \ , \ , \ , \ ]$$
$$\xi[1] = [d^1, i^0, \ , \ , \ ]$$
$$\xi[1] = [d^1, i^0, g^2, \ , \ ]$$
$$\xi[1] = [d^1, i^0, g^2, a^0, \ ]$$
$$\xi[1] = [d^1, i^0, g^2, a^0, l^1]$$

$$\xi[1] = [d^1, i^0, g^2, a^0, l^1]$$

Topological Ordering

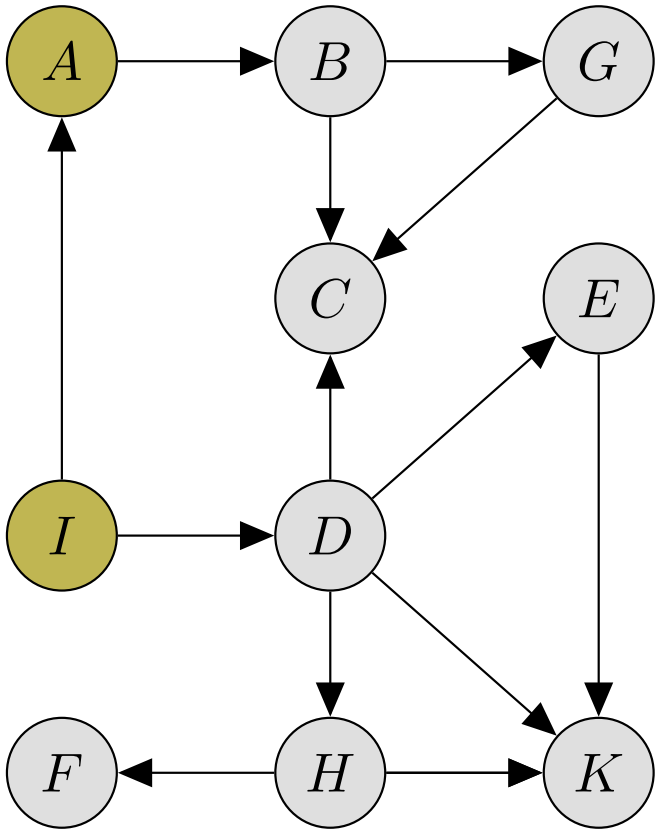
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \rightarrow X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: I

Forward Sampling: Topological Ordering

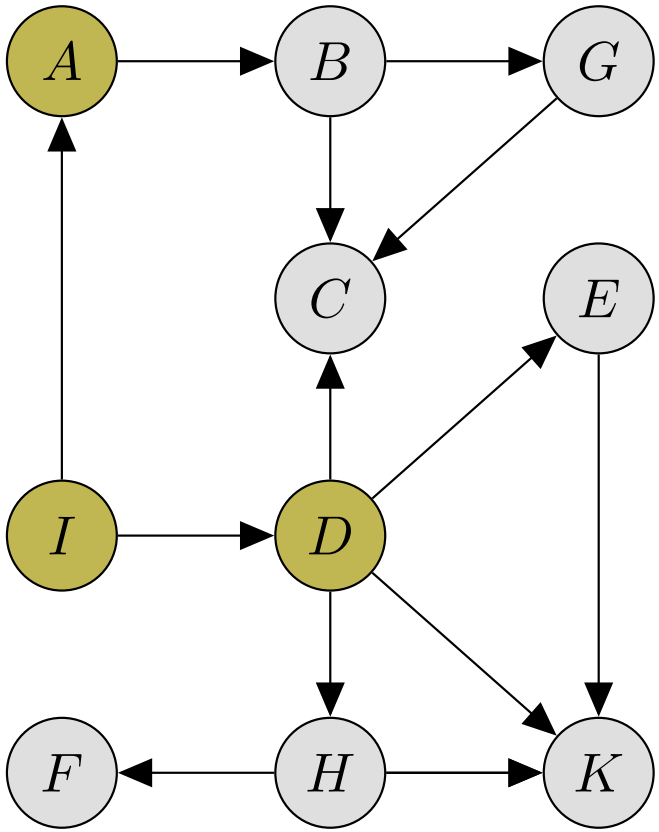
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A**

Forward Sampling: Topological Ordering

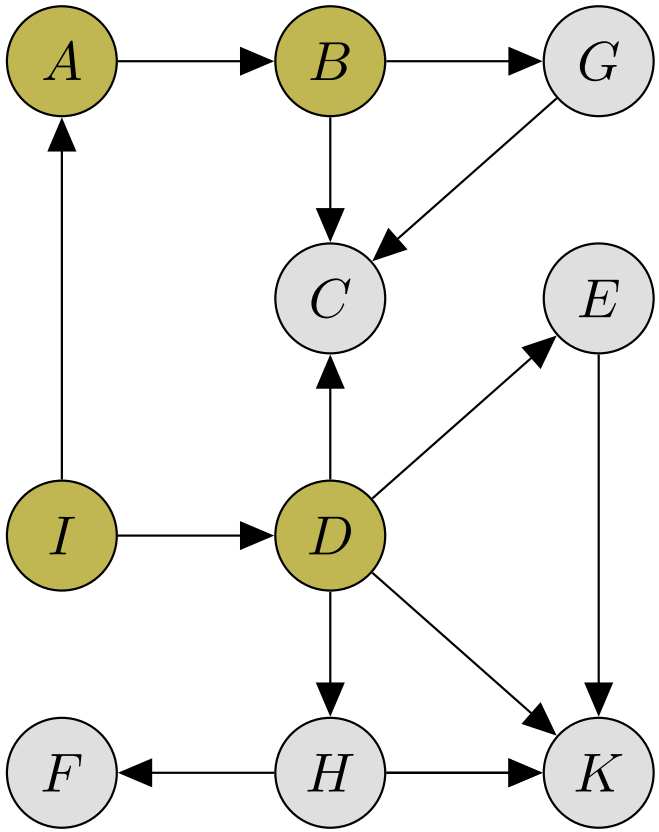
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \rightarrow X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D**

Forward Sampling: Topological Ordering

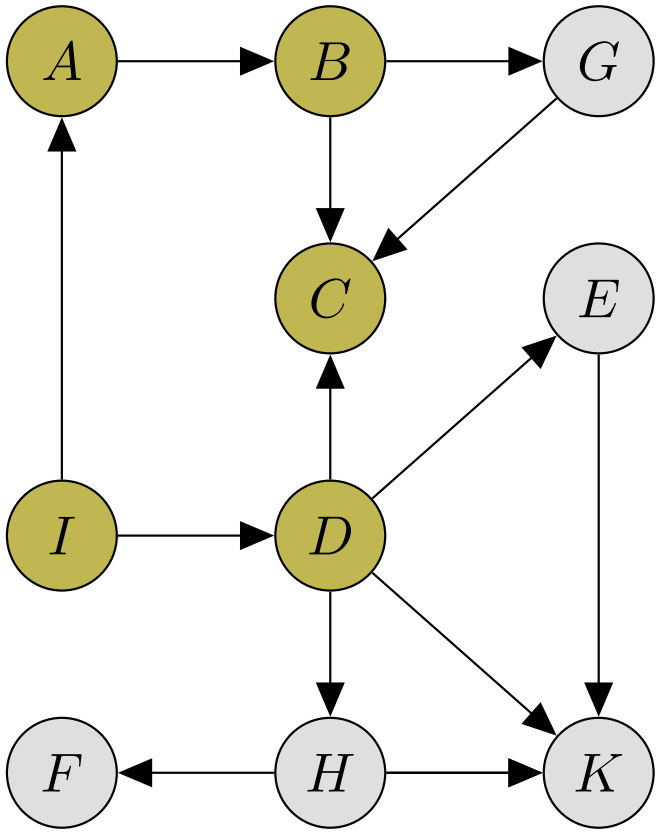
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B**

Forward Sampling: Topological Ordering

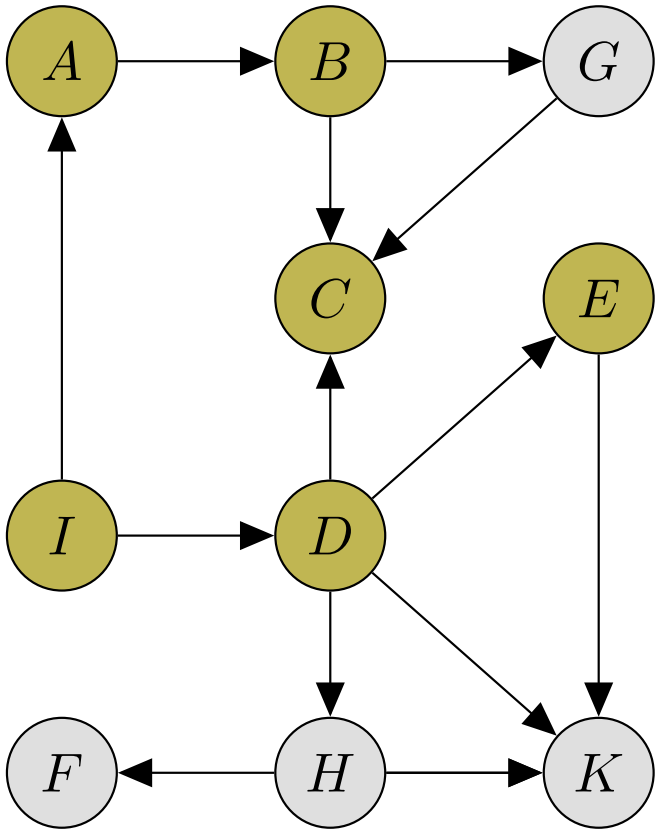
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C**

Forward Sampling: Topological Ordering

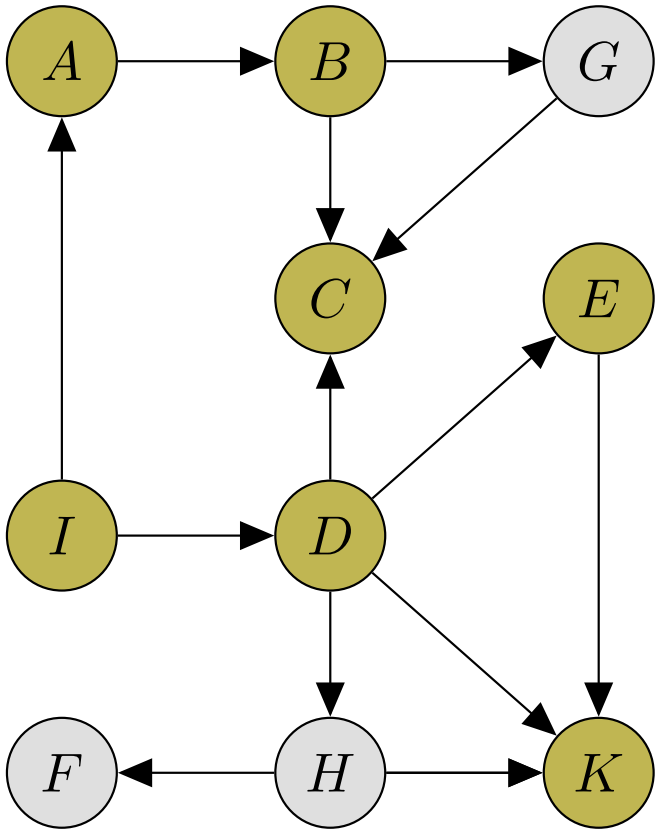
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C E**

Forward Sampling: Topological Ordering

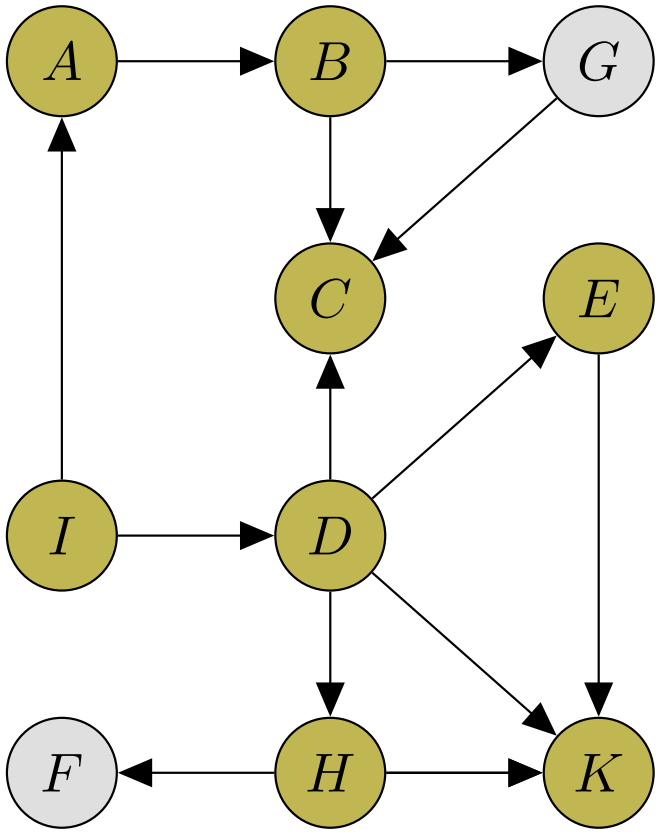
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \rightarrow X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C E L**

Forward Sampling: Topological Ordering

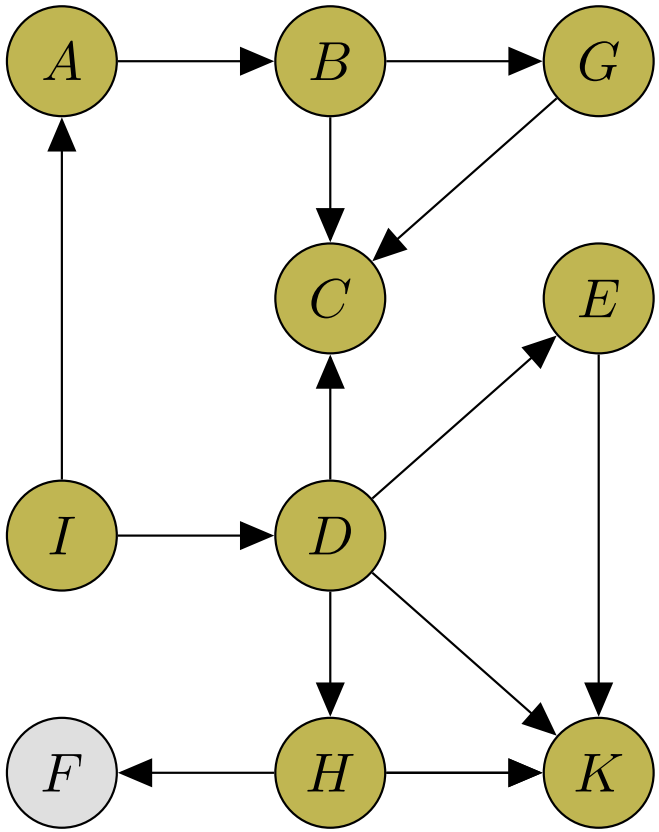
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C E L K**

Forward Sampling: Topological Ordering

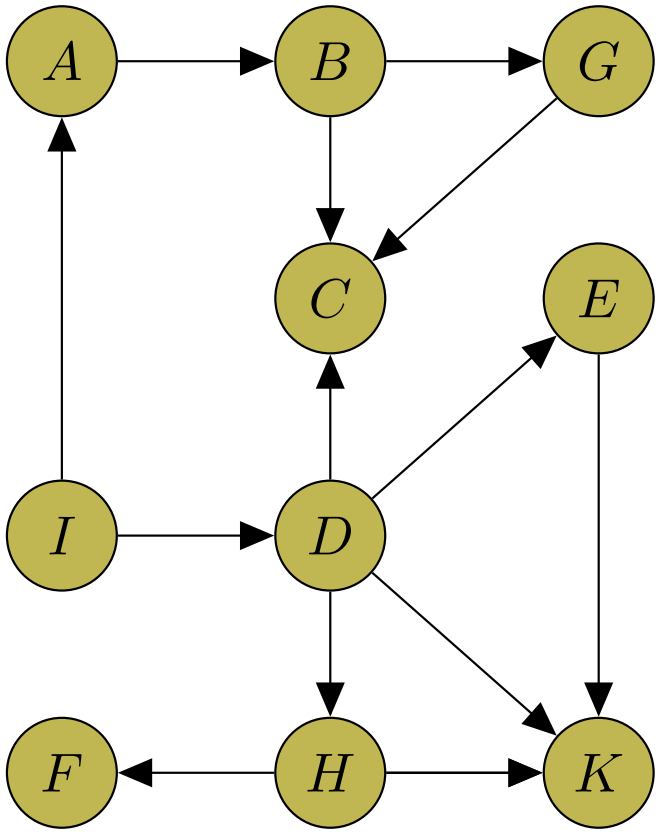
An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \rightarrow X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C E L K G**

Forward Sampling: Topological Ordering

An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \rightarrow X_j) \in \mathcal{E}$, then $i < j$.



Topological Order: **I A D B C E L K G F**

Convergence Bound

From this process we can generate infinitely many samples $\mathcal{D} = \{\xi[1], \ldots, \xi[M]\}$

Now we can estimate the empirical expectation of any function f as:

$$\hat{\mathbb{E}}_{\mathcal{D}}(f) = \frac{1}{M} \sum_{m=1}^{M} f(\xi[m]).$$

For example if you are trying to compute $P(\mathbf{y})$:

$$\hat{P}_{\mathcal{D}}(\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{\mathbf{y}[m] = \mathbf{y}\},$$

where $\mathbf{y}[m]$ denotes $\xi[m]\langle\mathbf{Y}\rangle$ - the assignment to $\mathbf{Y}$ in the particle $\xi[m]$.

The quality of the estimate depends heavily on the number of particles generated.

Sampling-based estimation

So then how many particles is enough to get a performance guarantee?

Hoeffding Bound:

$$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y}) - \epsilon, P(\mathbf{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

1. $\hat{P}(\mathbf{y}) \notin [P(\mathbf{y}) - \epsilon, P(\mathbf{y}) + \epsilon]$: The estimated empirical distribution is not within ϵ-distance of the target distribution.
2. $P_{\mathcal{D}}$: Thus a measurement of the probability of a bad dataset.
3. $2e^{-2M\epsilon^2}$: A lower-bound for a bad dataset.

Intuition: The bound grows shrinks exponentially with:

1. the increase in the number of samples.
2. the increase in tolerance (ϵ).

Hoeffding and Chernoff Bound

If we define the probability that we get a bad sample set, as δ, then:

$$2e^{-2M\epsilon^2} \leq \delta$$

then with some algebra we get a required sample size to get an estimator with (ϵ, δ) reliability:

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}.$$

Similarly we can define the **Chernoff Bound:**

$$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y})(\pm\epsilon)]) \leq 2e^{-MP(\mathbf{y})\epsilon^2/3}$$

which gives us:

$$M \geq 3\frac{\ln(2/\delta)}{P(\mathbf{y})\epsilon^2}.$$

Sampling-based estimation

The Chernoff bound considers the chance of getting an estimate that is more than ϵ away from the target value ($P(\mathbf{y})$).

Sampling based estimation is used when $P(\mathbf{y})$ is not too small

Since this directly affects how many samples we require.

Even if we use the Chernoff bound we will still need to calculate $P(\mathbf{y})$ which is a difficult inference task for small probability values.

Additionally, if we know $P(\mathbf{y})$, then why would be want to estimate it.

Conditional Probability Queries

You now know how to calculate $P(\mathbf{Y} = y)$ using particle-based inference.

What about $P(\mathbf{Y} = y \mid \mathbf{E} = e)$?

Strategy:

1. Generate many samples with forward sampling
2. Discard the one which are inconsistent with $P(\mathbf{Y} = y \mid \mathbf{E} = e)$ (**rejection sampling**)

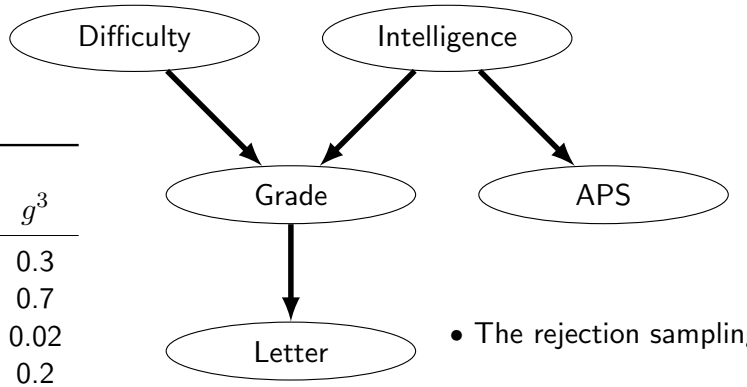
What do you think is the major issue with this approach?

1. Many samples will be rejected.
2. If $P(\mathbf{e}) = 0.001$. Then from 10 000 samples we will only get **10 samples**.
3. In other words, to get M^* samples we need on average $M = \frac{M^*}{P(\mathbf{e})}$ samples.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

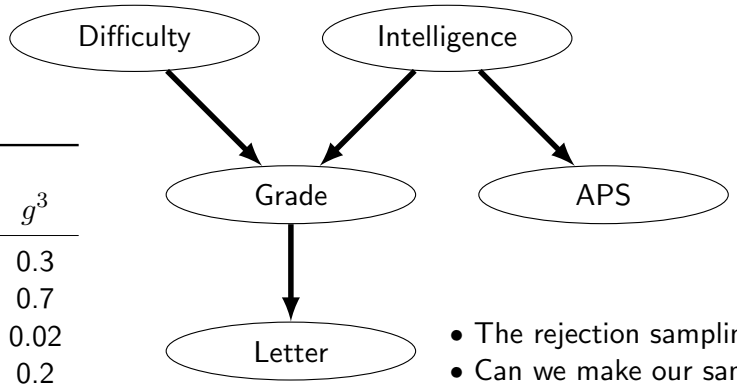
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

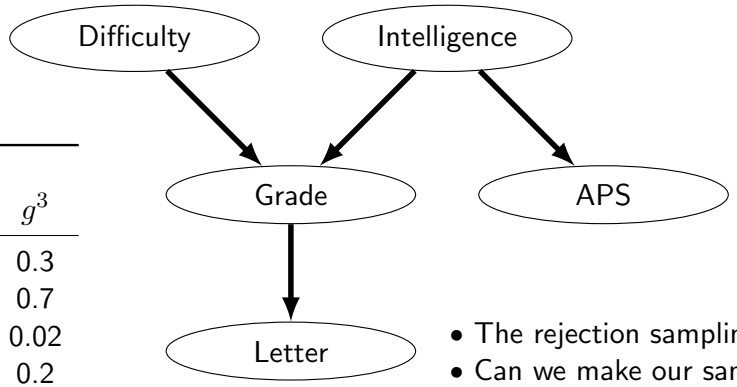
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

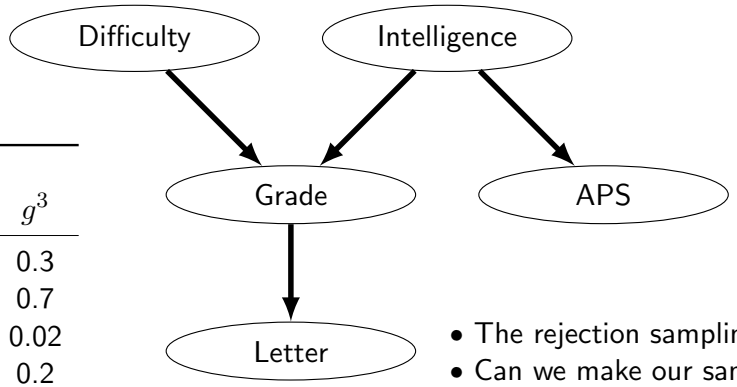
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

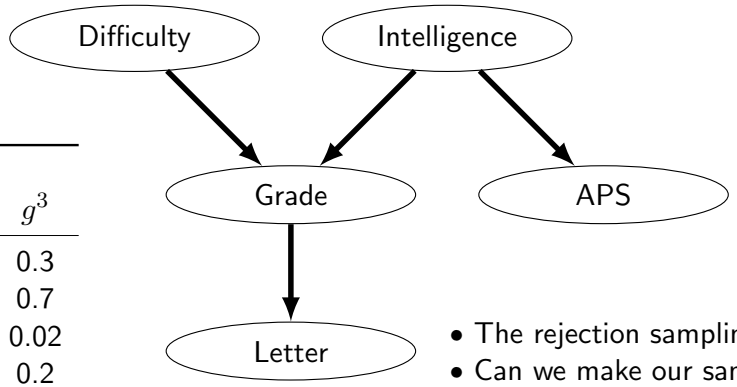
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

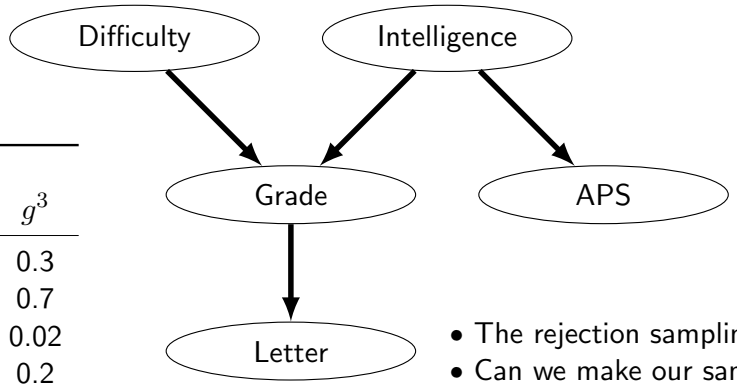
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

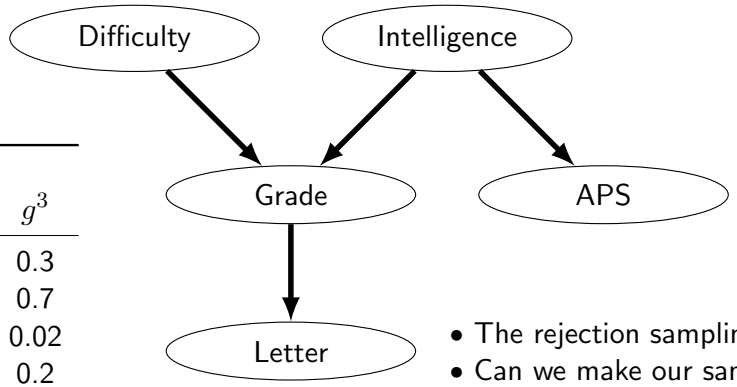
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.
- BUT we should expect i^1 to be more common if a^1 is frequent.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

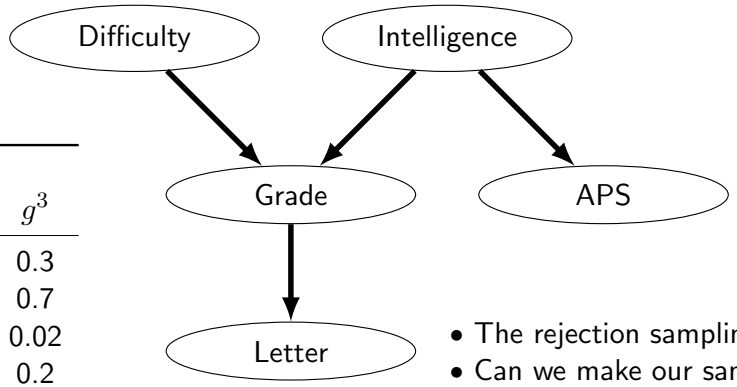
- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.
- BUT we should expect i^1 to be more common if a^1 is frequent.

Therefore, this approach is incorrect!

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.
- BUT we should expect i^1 to be more common if a^1 is frequent.

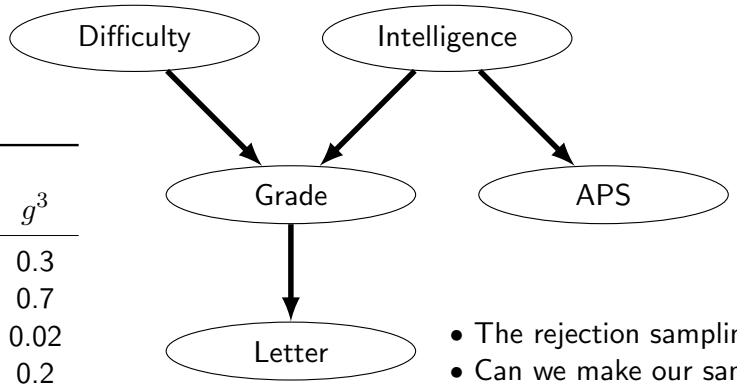
Therefore, this approach is incorrect!

- This approach is wrong because variable A is more likely to be a^1 if and only if I is i^1.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.
- BUT we should expect i^1 to be more common if a^1 is frequent.

Therefore, this approach is incorrect!

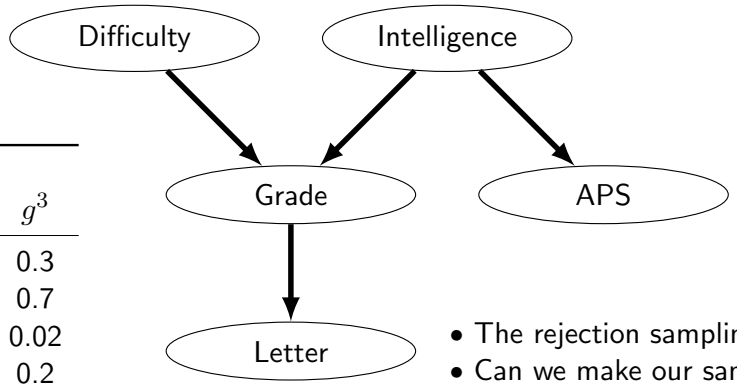
- This approach is wrong because variable A is more likely to be a^1 if and only if I is i^1.

If only we had a way to associate a sample with its likelihood.

Likelihood Weightings

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The rejection sampling method seems very wasteful.
- Can we make our samples more relevant to our evidence?
- Suppose that have the evidence a^1. (High APS)
- If we just force the samples to take on a^1.
- Since we will get 70% and 30% of samples with i^0 and i^1 respectively.
- BUT we should expect i^1 to be more common if a^1 is frequent.

Therefore, this approach is incorrect!

- This approach is wrong because variable A is more likely to be a^1 if and only if I is i^1.

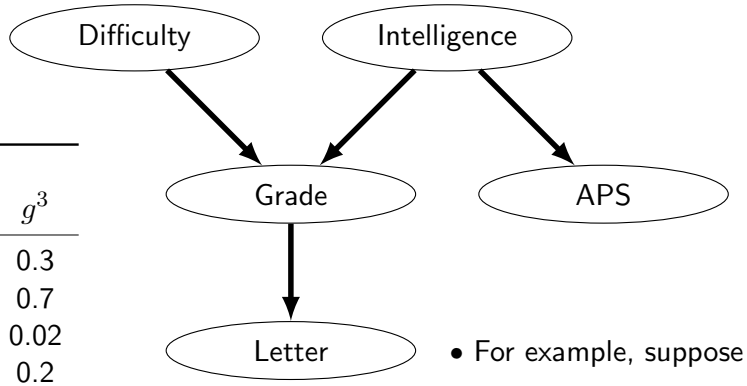
If only we had a way to associate a sample with its likelihood.

- That is i^1 which sets a^1 should occur 80% of the time (weight) and i^0 which sets a^1 is weighted as 5%

Likelihood Weighting Example

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

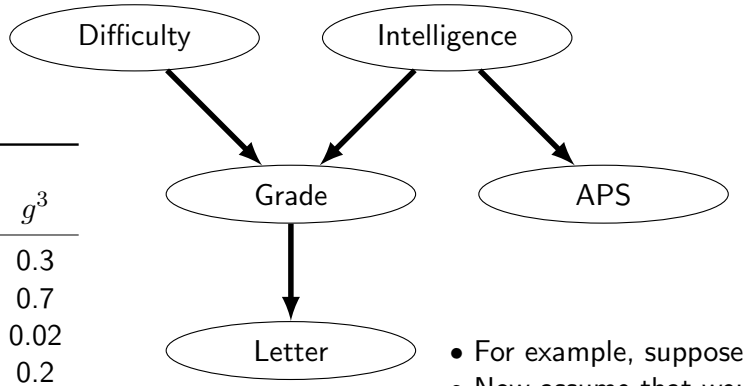
	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- For example, suppose that we have the evidence l^0, a^1.

Likelihood Weighting Example

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

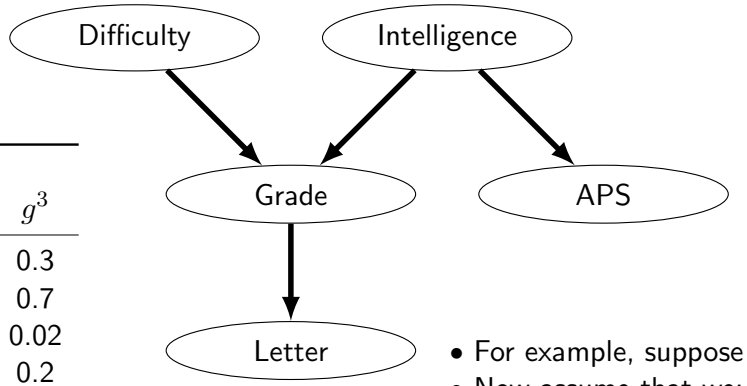
- For example, suppose that we have the evidence l^0, a^1.
- Now assume that we:

sample $D = d^1$,
sample $I = i^0$,
set $A = a^1$,
sample $G = g^2$, and
set $L = l^0$.

Likelihood Weighting Example

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- For example, suppose that we have the evidence l^0, a^1.
- Now assume that we:

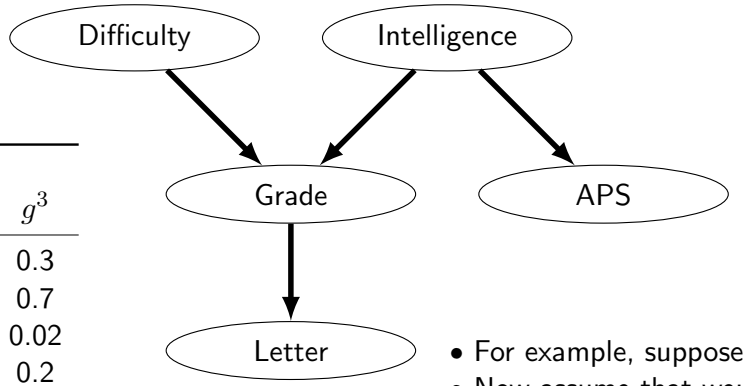
sample $D = d^1$,
sample $I = i^0$,
set $A = a^1$,
sample $G = g^2$, and
set $L = l^0$.

- The weight required for this sample to compensate for setting the evidence is $0.05 \times 0.4 = 0.02$.

Likelihood Weighting Example

Difficulty	
d^0	d^1
0.6	0.4

Intelligence	
i^0	i^1
0.7	0.3



	Grade		
i,d	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	APS	
i	a^0	a^1
i^0	0.95	0.05
i^1	0.2	0.8

	Letter	
g	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- For example, suppose that we have the evidence l^0, a^1.
- Now assume that we:

sample $D = d^1$,
sample $I = i^0$,
set $A = a^1$,
sample $G = g^2$, and
set $L = l^0$.

- The weight required for this sample to compensate for setting the evidence is $0.05 \times 0.4 = 0.02$.
- The weights of different samples are derived from the **likelihood of the evidence accumulated** throughout the sampling process.

Likelihood Weighting

This process provides a weighted particle from which we can calculate $P(\mathbf{y} \mid \mathbf{e})$.

We simple use likelihood weighting to generate M particles in $\mathcal{D} = \{\langle \xi[1], w[1] \rangle, \ldots, \langle \xi[M], w[M] \rangle\}$.

Then calculate:

$$\hat{P}_{\mathcal{D}}(\mathbf{y} \mid \mathbf{e}) = \frac{\sum_{m=1}^{M} w[m] \mathbb{1}\{\mathbf{y}[m] = \mathbf{y}\}}{\sum_{m=1}^{M} w[m]}.$$

(Notice that the equation does not include the evidence since every sample is generated w.r.t. the evidence anyway.)

Notice that samples here and in forward sampling is generated IID.

Likelihood Weighting Limitations

1. The evidence node influences its descendants, while non-descendants are accounted for by the weights.
2. If evidence is only at the leaves, we're essentially sampling from the prior, not desired **posterior distribution**.
3. Forward sampling and Likelihood Weightings can only be used for directed models.
4. We now look at methods that generate a **sequence of samples**, called Markov Chains (for directed and undirected).
5. This sample sequence generally `creeps` into the desired posterior.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Gibbs Sampling: Gibbs Chain Trajectory

Intuition: The Gibbs sampling method attempts to "fix" the sample, by re-sampling some of the variables.

- It generates a sample with arbitrary values (except for the evidence), and then re-samples all variables while updating the evidence in arbitrary order.

  For example, suppose we have evidence a^0, l^1.

  1. Suppose we want to produce the sample:

$P(l^1 \mid G)$	$P(G \mid D, I)$	$P(a^0 \mid I)$	$P(I)$	$P(D)$

  2. Insert evidence:

l^1	$P(G \mid D, I)$	a^0	$P(I)$	$P(D)$

  3. Insert arbitrary initialisation:

l^1	g^3	a^0	i^1	d^0

  4. re-sample $P(G \mid D, I, A, L)$:

l^1	g^2	a^0	i^1	d^0

  5. re-sample $P(I \mid D, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

  6. re-sample $P(D \mid I, G, A, L)$:

l^1	g^2	a^0	i^0	d^0

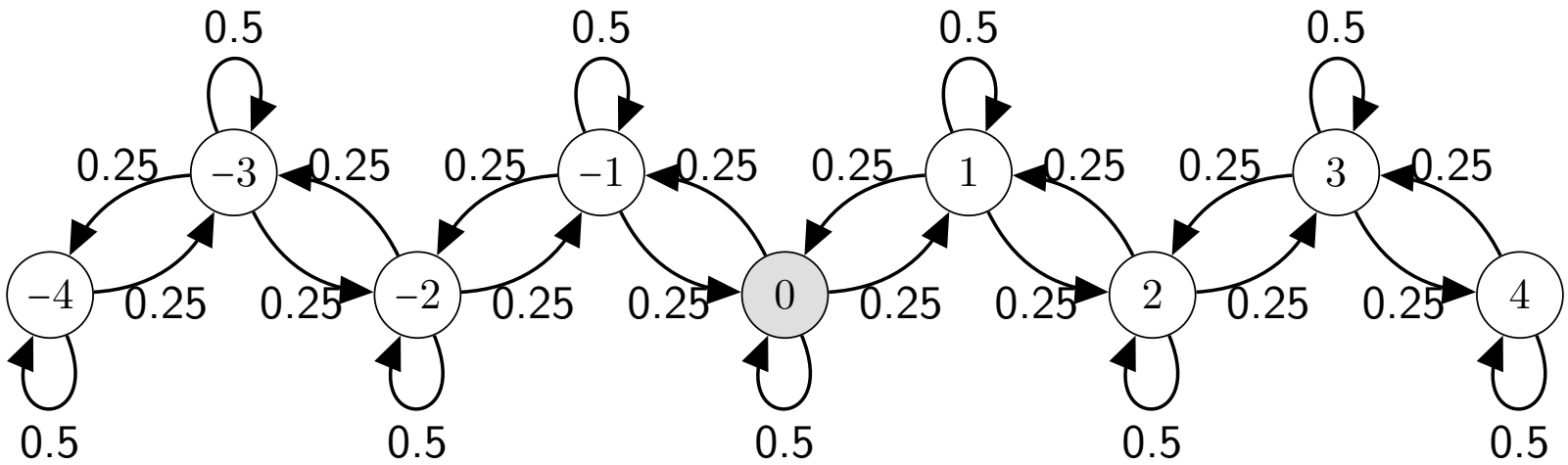
  7. We do this process to generate a sequence of samples, reusing unobserved variables at each step.

Markov Chain Monte Carlo

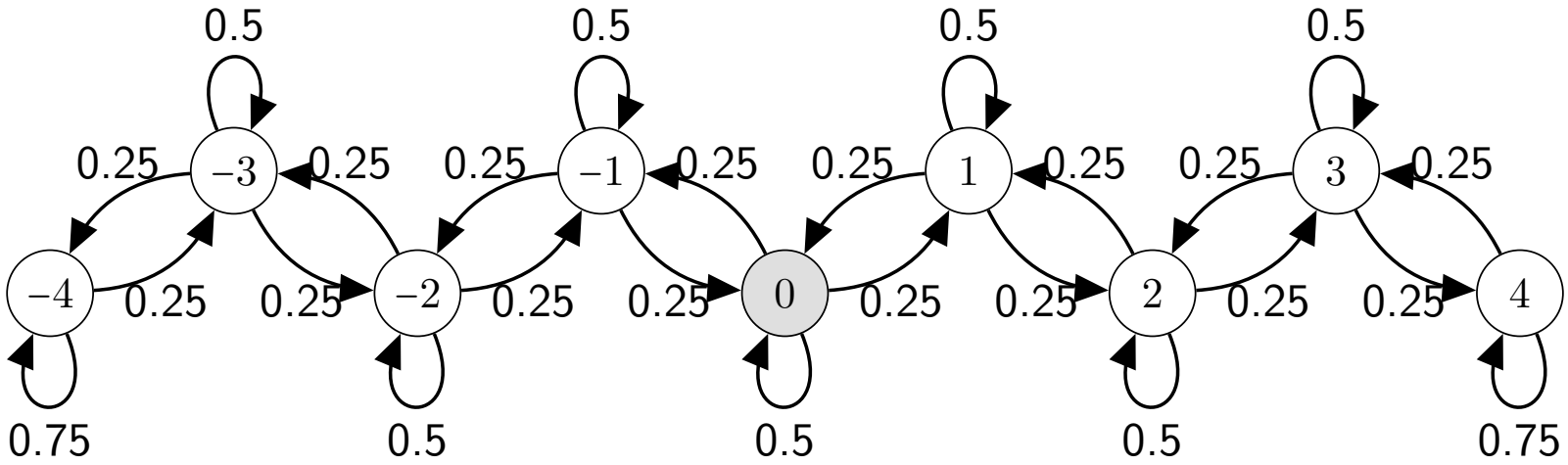
- Unlike forward sampling, the Gibbs sampling technique takes into account the **downstream evidence**.
- As we repeat the sampling process, we get closer and closer to the desired posterior: $P_\Phi(\mathbf{X}) = P(\mathbf{X} \mid \mathbf{e})$.
- Gibbs sampling is an example of a class of methods called Markov Chain Monte Carlo (MCMC).
- MCMC is an iterative process where we move from distribution to distribution getting closer to the desired posterior.
- But how many iterations do we need before we get a "good" sample?

Basic Definition

- A Markov chain is defined as a graph of states over which the sampling algorithm takes a random walk.
  - The nodes of this graph are **possible assignments** of variables $\mathbf{x} \in Val(\mathbf{X})$.
  - The edges is a transition model, $\mathcal{T}$, which specifies for each pair of states **x** and **x'** the probability $\mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$ of going from state **x** to **x'**.



Example of Markov Chain



- In this Markov chain the sampler starts at (0).
- $\mathcal{T}(i \to i) = 0.5$
- $\mathcal{T}(i \to i+1) = 0.25$
- $\mathcal{T}(i \to i-1) = 0.25$
- $\mathcal{T}(4 \to 4) = \mathcal{T}(-4 \to -4) = 0.75$

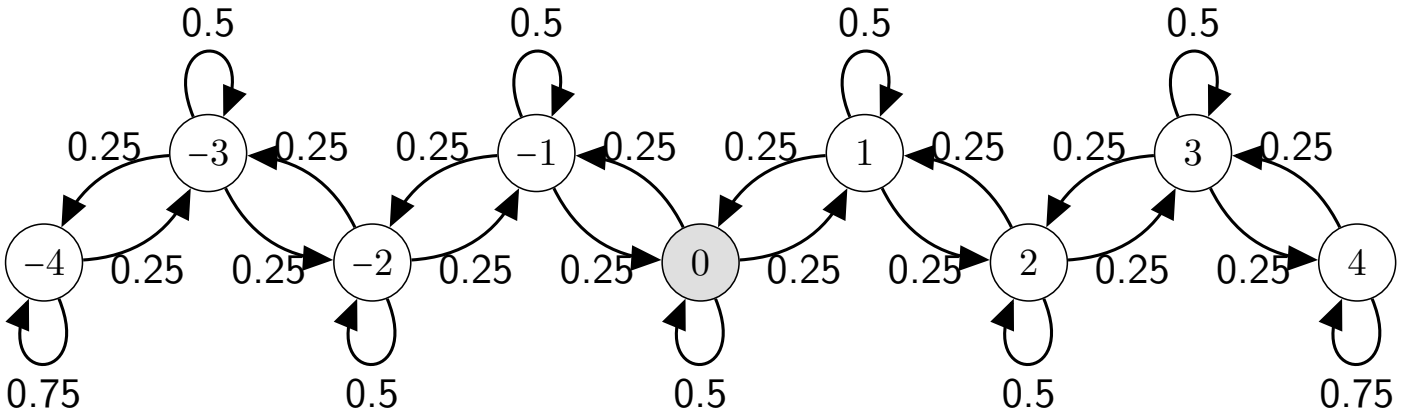
Sampling with MCMC

- By moving around states we can sample: $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \ldots$
- It is useful to imagine a state as a random variable, $\mathbf{X}^{(t)}$.
- Then we can assume that the initial assignment is generated from $P^{(0)}(\mathbf{X}^{(0)})$.
- Then distributions from subsequent states is just: $P^{(0)}(\mathbf{X}^{(0)}), P^{(1)}(\mathbf{X}^{(1)}), P^{(2)}(\mathbf{X}^{(2)})$.
- More generally:

$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

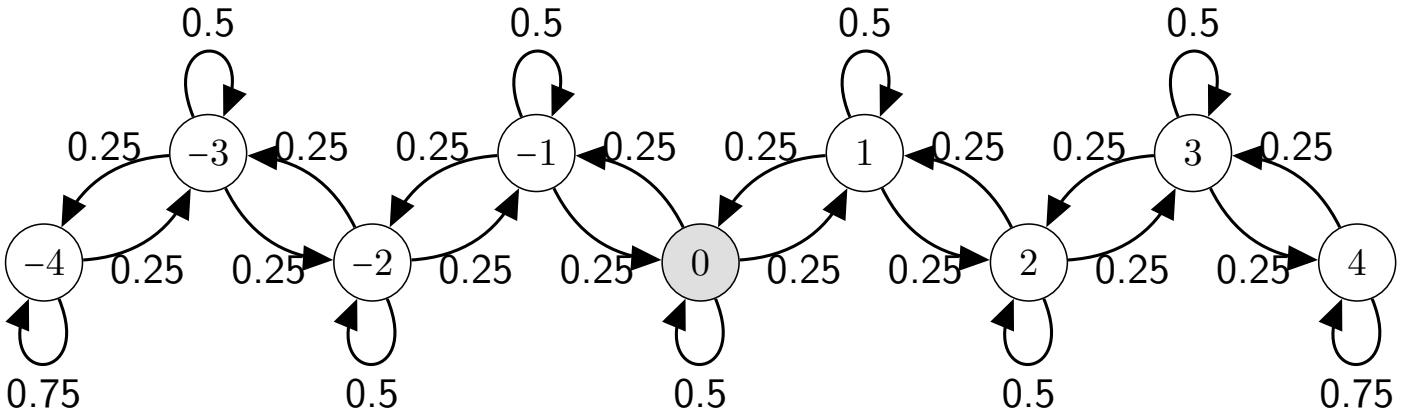
Intuition: The next state $\mathbf{x'}$ is the sum over all previous neighbouring assignments $\mathbf{x}$ of the probability being in state $\mathbf{x}$ multiplied by the transition from $\mathbf{x}$ to $\mathbf{x}'$.

The First 7 iterations of MCMC



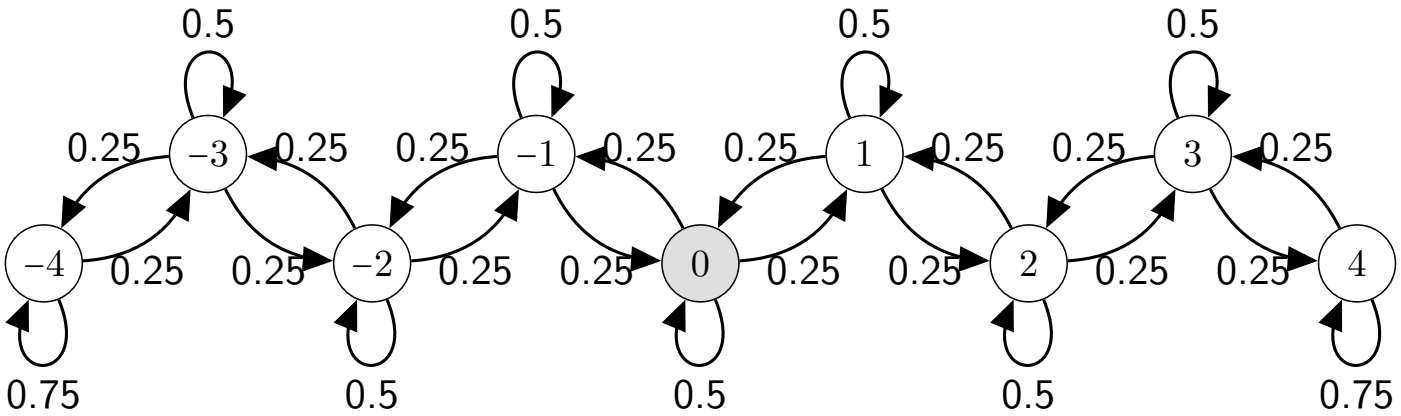
	-4	-3	-2	-1	0	1	2	3	4
$P^{(0)}(\mathbf{X}^{(0)})$	0	0	0	0	1	0	0	0	0
$P^{(1)}(\mathbf{X}^{(1)})$									
$P^{(2)}(\mathbf{X}^{(2)})$									
$P^{(3)}(\mathbf{X}^{(3)})$									
$P^{(4)}(\mathbf{X}^{(4)})$									
$P^{(5)}(\mathbf{X}^{(5)})$									
$P^{(6)}(\mathbf{X}^{(6)})$									
$P^{(7)}(\mathbf{X}^{(7)})$									

The First 7 Iterations of MCMC



	-4	-3	-2	-1	0	1	2	3	4
$P^{(0)}(\mathbf{X}^{(0)})$	0	0	0	0	1	0	0	0	0
$P^{(1)}(\mathbf{X}^{(1)})$	0	0	0	0.25	0.5	0.25	0	0	0
$P^{(2)}(\mathbf{X}^{(2)})$									
$P^{(3)}(\mathbf{X}^{(3)})$									
$P^{(4)}(\mathbf{X}^{(4)})$									
$P^{(5)}(\mathbf{X}^{(5)})$									
$P^{(6)}(\mathbf{X}^{(6)})$									
$P^{(7)}(\mathbf{X}^{(7)})$									

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) = \mathbf{0.375}$

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) = \mathbf{0.25}$

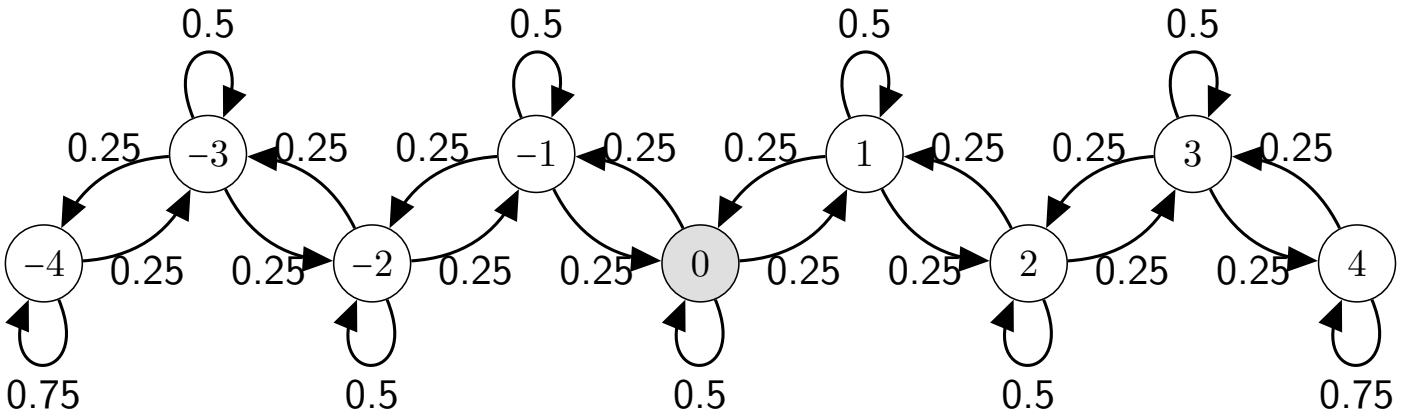
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) = \mathbf{0.25}$

For State -2: $(0.25 \times 0.25) = \mathbf{0.063}$

For State 2: $(0.25 \times 0.25) = \mathbf{0.063}$

For State -3, 3, -4, 4: $= \mathbf{0}$

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

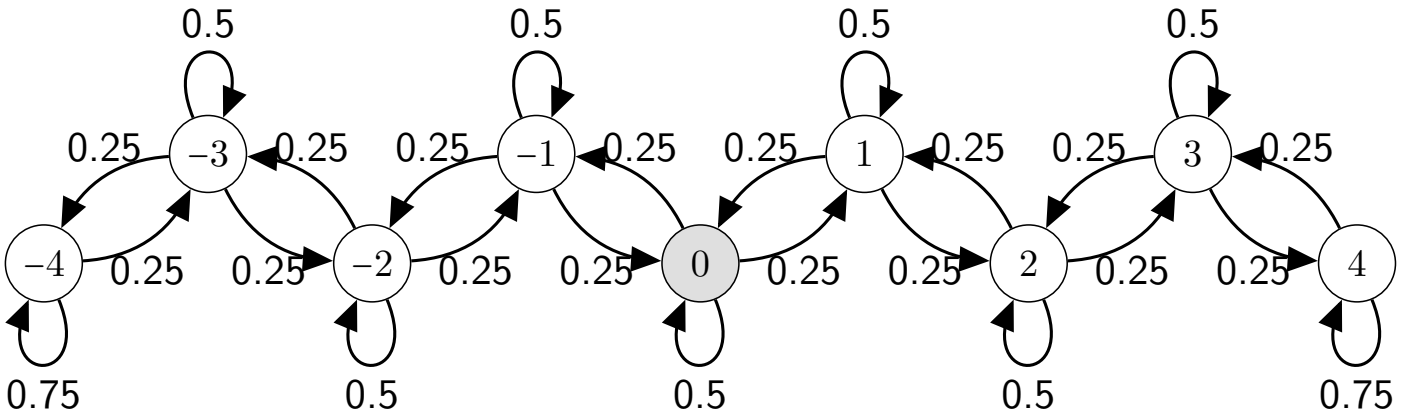
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

For State -3, 3, -4, 4: = **0**

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

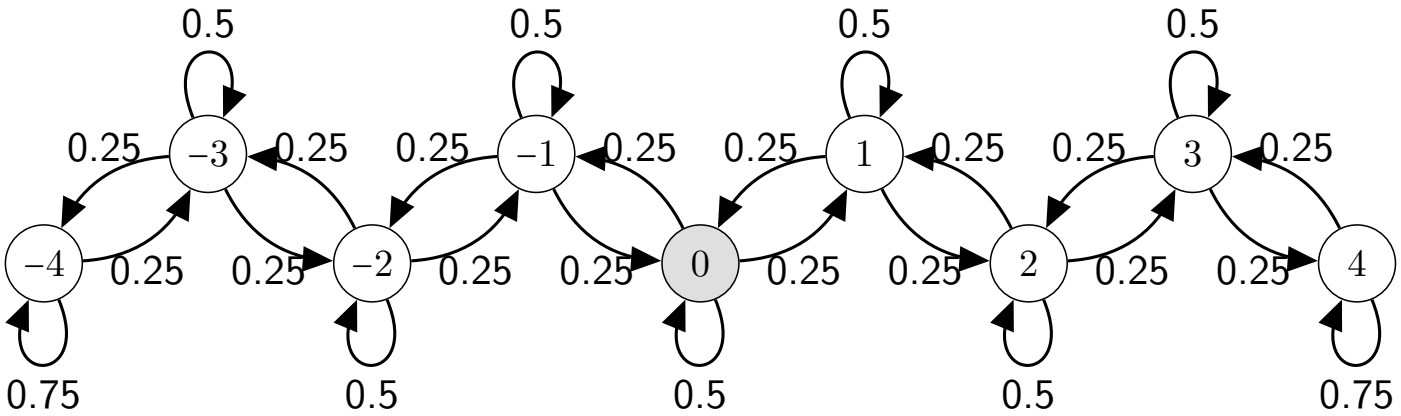
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

For State -3, 3, -4, 4: = **0**

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

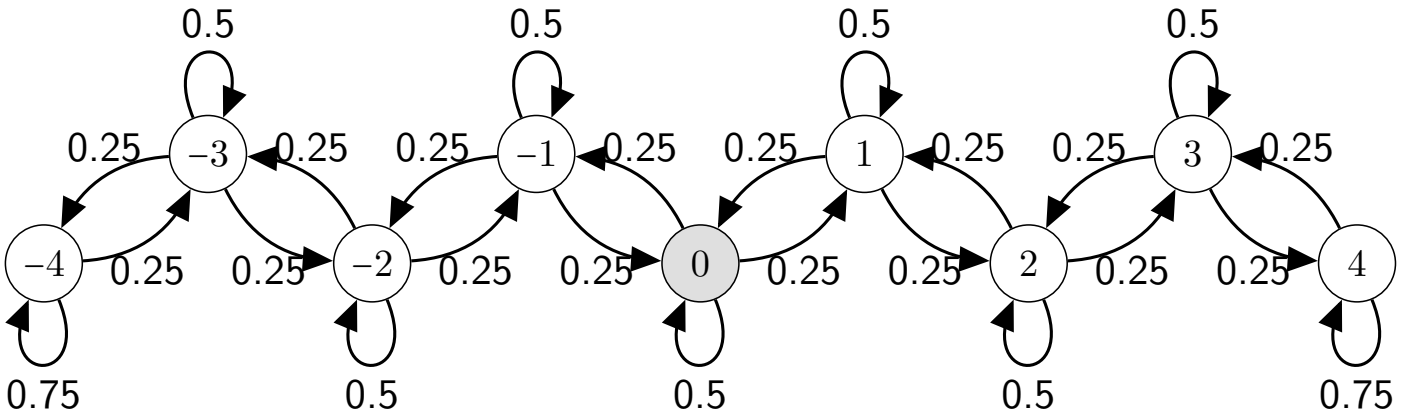
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

For State -3, 3, -4, 4: = **0**

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

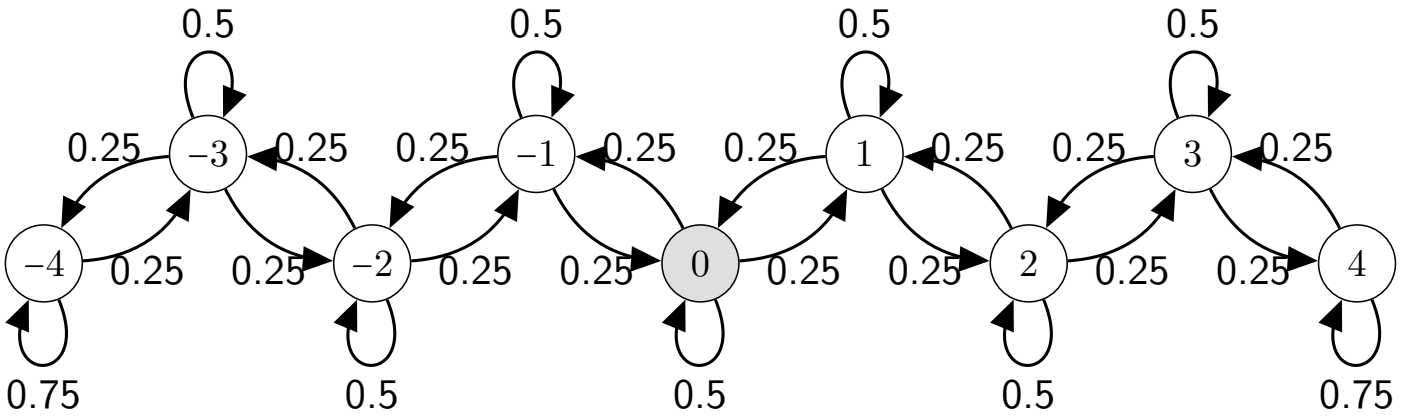
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

For State -3, 3, -4, 4: $= 0$

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

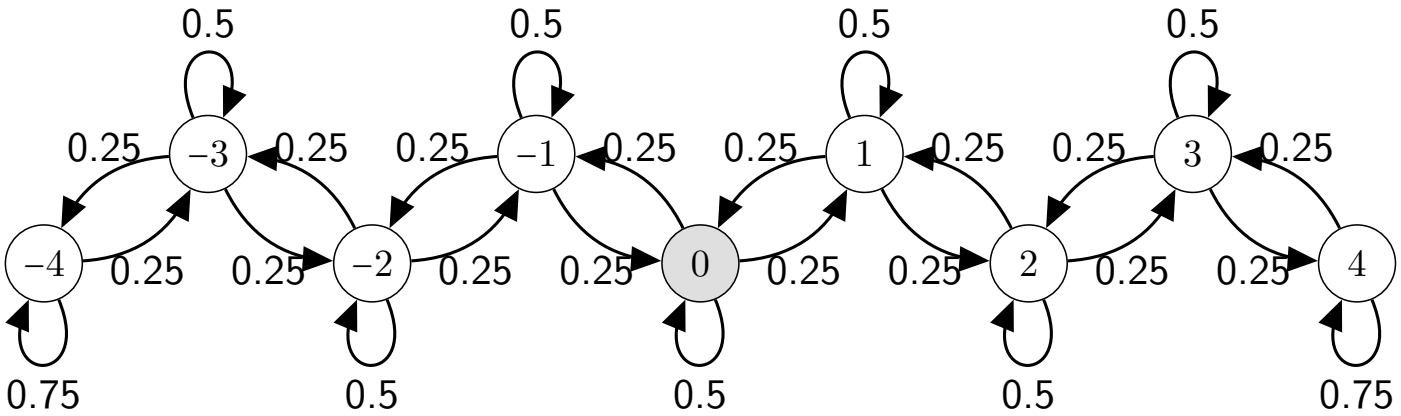
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

For State -3, 3, -4, 4: = 0

The First 3 Iterations of MCMC



$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

For State 0: $(0.5 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) =$ **0.375**

For State -1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

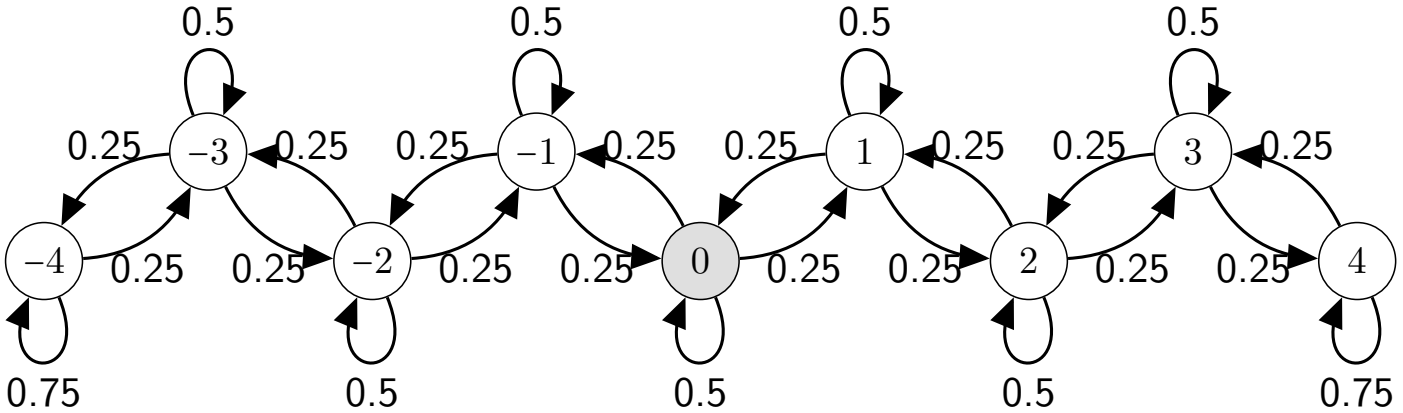
For State 1: $(0.25 \times 0.5) + (0.5 \times 0.25) =$ **0.25**

For State -2: $(0.25 \times 0.25) =$ **0.063**

For State 2: $(0.25 \times 0.25) =$ **0.063**

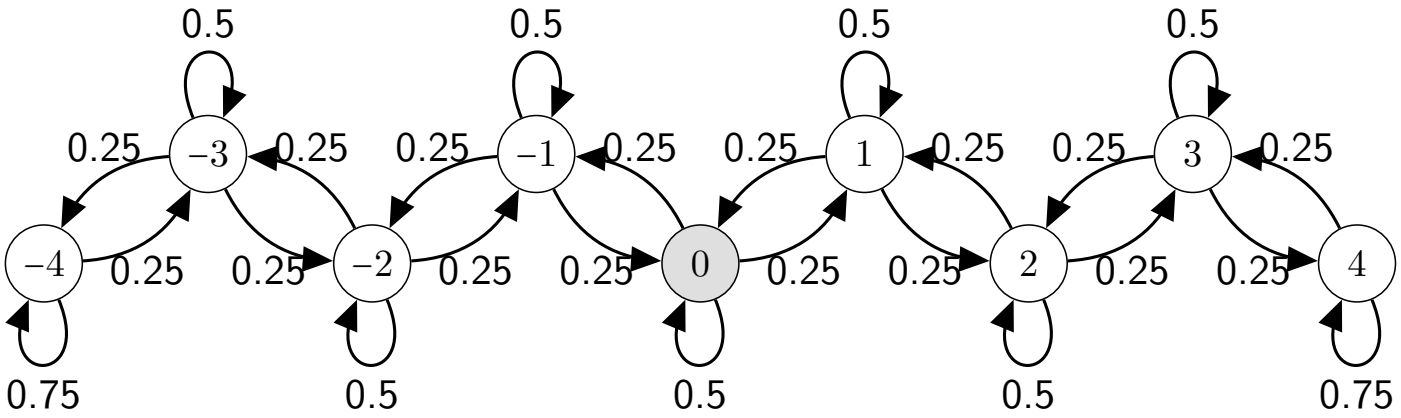
For State -3, 3, -4, 4: $=$ **0**

The First 3 Iterations of MCMC



	-4	-3	-2	-1	0	1	2	3	4
$P^{(0)}(\mathbf{X}^{(0)})$	0	0	0	0	1	0	0	0	0
$P^{(1)}(\mathbf{X}^{(1)})$	0	0	0	0.25	0.5	0.25	0	0	0
$P^{(2)}(\mathbf{X}^{(2)})$	0	0	0.063	0.25	0.375	0.25	0.063	0	0
$P^{(3)}(\mathbf{X}^{(3)})$									
$P^{(4)}(\mathbf{X}^{(4)})$									
$P^{(5)}(\mathbf{X}^{(5)})$									
$P^{(6)}(\mathbf{X}^{(6)})$									
$P^{(7)}(\mathbf{X}^{(7)})$									

The First 3 Iterations of MCMC



	-4	-3	-2	-1	0	1	2	3	4
$P^{(0)}(\mathbf{X}^{(0)})$	0	0	0	0	1	0	0	0	0
$P^{(1)}(\mathbf{X}^{(1)})$	0	0	0	0.25	0.5	0.25	0	0	0
$P^{(2)}(\mathbf{X}^{(2)})$	0	0	0.063	0.25	0.375	0.25	0.063	0	0
$P^{(3)}(\mathbf{X}^{(3)})$	0	0.016	0.094	0.235*	0.313**	0.235	0.094	0.016	0
$P^{(4)}(\mathbf{X}^{(4)})$									
$P^{(5)}(\mathbf{X}^{(5)})$									
$P^{(6)}(\mathbf{X}^{(6)})$									
$P^{(7)}(\mathbf{X}^{(7)})$									

*$(0.25 \times 0.5) + (0.063 \times 0.25) + (0.375 \times 0.25) = 0.2345$

**$(0.375 \times 0.5) + (0.25 \times 0.25) + (0.25 \times 0.25) = 0.3125$

Iterations 1-20

	-4	-3	-2	-1	0	1	2	3	4
$P^{(0)}(\mathbf{X}^{(0)})$	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00
$P^{(1)}(\mathbf{X}^{(1)})$	0,00	0,00	0,00	0,25	0,50	0,25	0,00	0,00	0,00
$P^{(2)}(\mathbf{X}^{(2)})$	0,00	0,00	0,06	0,25	0,38	0,25	0,06	0,00	0,00
$P^{(3)}(\mathbf{X}^{(3)})$	0,00	0,02	0,09	0,23	0,31	0,23	0,09	0,02	0,00
$P^{(4)}(\mathbf{X}^{(4)})$	0,00	0,03	0,11	0,22	0,27	0,22	0,11	0,03	0,00
$P^{(5)}(\mathbf{X}^{(5)})$	0,01	0,04	0,12	0,21	0,25	0,21	0,12	0,04	0,01
$P^{(6)}(\mathbf{X}^{(6)})$	0,02	0,05	0,12	0,19	0,23	0,19	0,12	0,05	0,02
$P^{(7)}(\mathbf{X}^{(7)})$	0,03	0,06	0,12	0,18	0,21	0,18	0,12	0,06	0,03
$P^{(8)}(\mathbf{X}^{(8)})$	0,04	0,07	0,12	0,17	0,20	0,17	0,12	0,07	0,04
$P^{(9)}(\mathbf{X}^{(9)})$	0,04	0,07	0,12	0,17	0,19	0,17	0,12	0,07	0,04
$P^{(10)}(\mathbf{X}^{(10)})$	0,05	0,08	0,12	0,16	0,18	0,16	0,12	0,08	0,05
$P^{(11)}(\mathbf{X}^{(11)})$	0,06	0,08	0,12	0,15	0,17	0,15	0,12	0,08	0,06
$P^{(12)}(\mathbf{X}^{(12)})$	0,06	0,09	0,12	0,15	0,16	0,15	0,12	0,09	0,06
$P^{(13)}(\mathbf{X}^{(13)})$	0,07	0,09	0,12	0,14	0,16	0,14	0,12	0,09	0,07
$P^{(14)}(\mathbf{X}^{(14)})$	0,07	0,09	0,12	0,14	0,15	0,14	0,12	0,09	0,07
$P^{(15)}(\mathbf{X}^{(15)})$	0,08	0,09	0,12	0,14	0,15	0,14	0,12	0,09	0,08
$P^{(16)}(\mathbf{X}^{(16)})$	0,08	0,10	0,12	0,13	0,14	0,13	0,12	0,10	0,08
$P^{(17)}(\mathbf{X}^{(17)})$	0,09	0,10	0,12	0,13	0,14	0,13	0,12	0,10	0,09
$P^{(18)}(\mathbf{X}^{(18)})$	0,09	0,10	0,12	0,13	0,13	0,13	0,12	0,10	0,09
$P^{(19)}(\mathbf{X}^{(19)})$	0,09	0,10	0,11	0,13	0,13	0,13	0,11	0,10	0,09
$P^{(20)}(\mathbf{X}^{(20)})$	0,09	0,10	0,11	0,13	0,13	0,13	0,11	0,10	0,09

Iterations 20-40

	-4	-3	-2	-1	0	1	2	3	4
$P^{(20)}(\mathbf{X}^{(20)})$	0,09	0,10	0,11	0,13	0,13	0,13	0,11	0,10	0,09
$P^{(21)}(\mathbf{X}^{(21)})$	0,10	0,10	0,11	0,12	0,13	0,12	0,11	0,10	0,10
$P^{(22)}(\mathbf{X}^{(22)})$	0,10	0,10	0,11	0,12	0,13	0,12	0,11	0,10	0,10
$P^{(23)}(\mathbf{X}^{(23)})$	0,10	0,10	0,11	0,12	0,12	0,12	0,11	0,10	0,10
$P^{(24)}(\mathbf{X}^{(24)})$	0,10	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,10
$P^{(25)}(\mathbf{X}^{(25)})$	0,10	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,10
$P^{(26)}(\mathbf{X}^{(26)})$	0,10	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,10
$P^{(27)}(\mathbf{X}^{(27)})$	0,10	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,10
$P^{(28)}(\mathbf{X}^{(28)})$	0,10	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,10
$P^{(29)}(\mathbf{X}^{(29)})$	0,11	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,11
$P^{(30)}(\mathbf{X}^{(30)})$	0,11	0,11	0,11	0,12	0,12	0,12	0,11	0,11	0,11
$P^{(31)}(\mathbf{X}^{(31)})$	0,11	0,11	0,11	0,11	0,12	0,11	0,11	0,11	0,11
$P^{(32)}(\mathbf{X}^{(32)})$	0,11	0,11	0,11	0,11	0,12	0,11	0,11	0,11	0,11
$P^{(33)}(\mathbf{X}^{(33)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(34)}(\mathbf{X}^{(34)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(35)}(\mathbf{X}^{(35)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(36)}(\mathbf{X}^{(36)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(37)}(\mathbf{X}^{(37)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(38)}(\mathbf{X}^{(38)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(39)}(\mathbf{X}^{(39)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
$P^{(40)}(\mathbf{X}^{(40)})$	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11

Asymptotic Behaviour

- At t $=$ 1-4, the probabilities get spread out over more and more of the states.
- At t $=$ 40, the distribution is almost uniform. Each state has probability 0.11 (Stationary Distribution).
- As the process converges we would expect $P^{(t+1)}$ to be close to $P^{(t)}$.

$$P^{(t)}(\mathbf{x'}) \approx P^{(t+1)}(\mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

- In other words we say a distribution is stationary if:

$$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

Asymptotic Behaviour

- At t $=$ 1-4, the probabilities get spread out over more and more of the states.
- At t $=$ 40, the distribution is almost uniform. Each state has probability 0.11 (Stationary Distribution).
- As the process converges we would expect $P^{(t+1)}$ to be close to $P^{(t)}$.

$$P^{(t)}(\mathbf{x'}) \approx P^{(t+1)}(\mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

- In other words we say a distribution is stationary if:

$$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

Asymptotic Behaviour

- At t $=$ 1-4, the probabilities get spread out over more and more of the states.
- At t $=$ 40, the distribution is almost uniform. Each state has probability 0.11 (Stationary Distribution).
- As the process converges we would expect $P^{(t+1)}$ to be close to $P^{(t)}$.

$$P^{(t)}(\mathbf{x'}) \approx P^{(t+1)}(\mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

- In other words we say a distribution is stationary if:

$$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

Asymptotic Behaviour

- At t $=$ 1-4, the probabilities get spread out over more and more of the states.
- At t $=$ 40, the distribution is almost uniform. Each state has probability 0.11 (Stationary Distribution).
- As the process converges we would expect $P^{(t+1)}$ to be close to $P^{(t)}$.

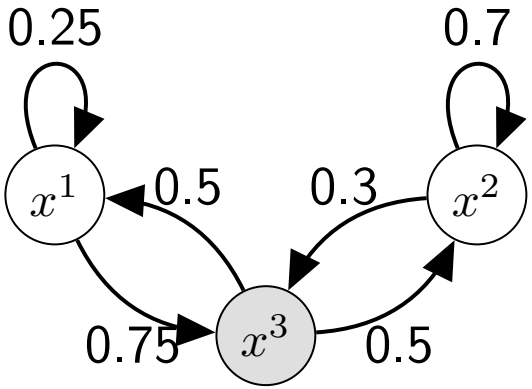
$$P^{(t)}(\mathbf{x'}) \approx P^{(t+1)}(\mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

- In other words we say a distribution is stationary if:

$$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \rightarrow \mathbf{x'})$$

Stationary Distributions

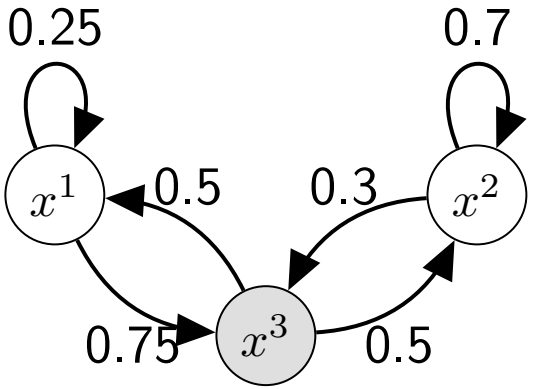
- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:
  $\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)
  $\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)
  $\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)
  $\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)
- You can solve this linear system using whatever method you like.

Stationary Distributions

- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:

$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)

$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)

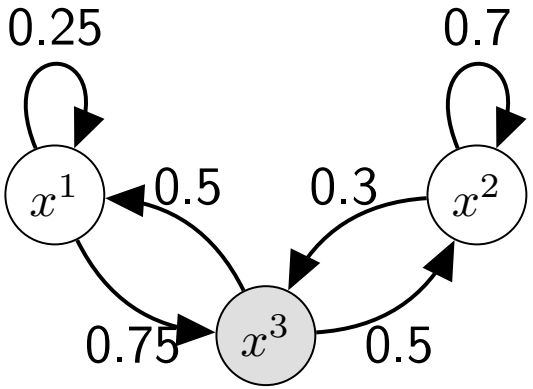
$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)

$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)

- You can solve this linear system using whatever method you like.

Stationary Distributions

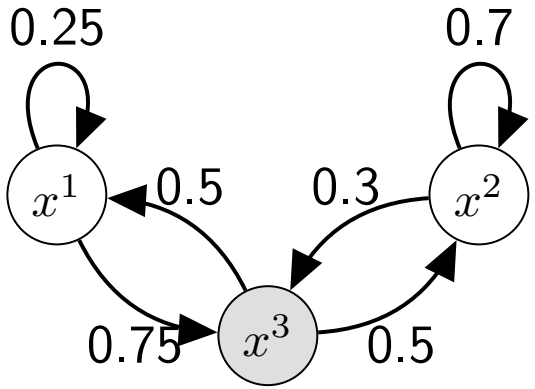
- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:
  $\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)
  $\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)
  $\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)
  $\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)
- You can solve this linear system using whatever method you like.

Stationary Distributions

- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:

$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)

$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)

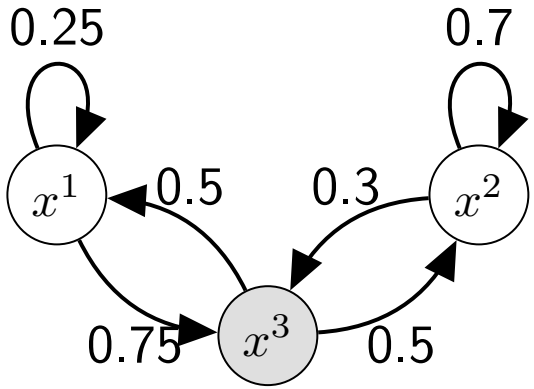
$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)

$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)

- You can solve this linear system using whatever method you like.

Stationary Distributions

- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:

$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)

$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)

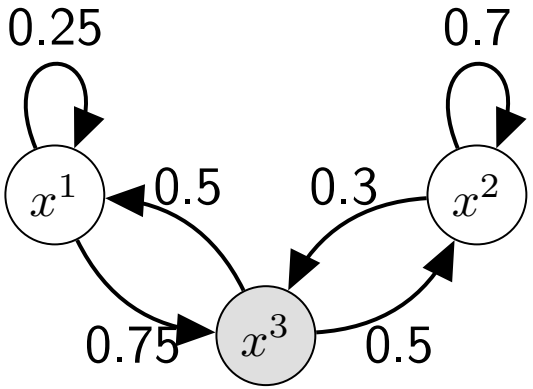
$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)

$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)

- You can solve this linear system using whatever method you like.

Stationary Distributions

- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:

$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)

$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)

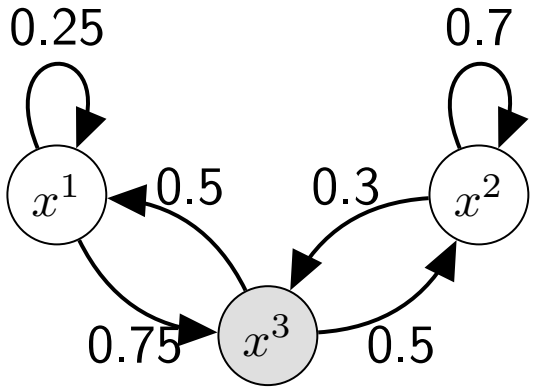
$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)

$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)

- You can solve this linear system using whatever method you like.

Stationary Distributions

- Lets look at another example, consider the following Markov chain:



- Then the stationary distribution must satisfy:

  $\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$ (ways of getting to x^1)

  $\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$ (ways of getting to x^2)

  $\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$ (ways of getting to x^3)

  $\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$ (legal probability distribution)
- You can solve this linear system using whatever method you like.

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

1. Extract Coefficients:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ -0.75 & -0.3 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
2. $R_3 = R_3 + R_1$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & -0.3 & 0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
3. $R_3 = R_2 + R_3$.
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
4. Remove R_3
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right)$$
5. $R_3 = (0.75 \times R_3) - R_1$:
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0.75 & 1.25 & 0.75 \end{array}\right)$$
6. $R_3 = (2.5 \times R_2) - R_3$
$$\left(\begin{array}{ccc|c} 0.75 & 0 & -0.5 & 0 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

Gauss-Jordan Elimination

7. $R_3 = (0.2 \times R_3) + R_1$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & -0.5 & 0 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

8. $R_3 = (0.2 \times R_3) + R_2$.

$$\left(\begin{array}{ccc|c} 0.75 & 0 & 0 & 0.15 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

9. Row echelon form: $R_1 = \frac{4}{3} \times R_1$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.15 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

10. $R_2 = \frac{10}{3} \times R_2$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 2.5 & 0.75 \end{array}\right)$$

11. $R_3 = \frac{2}{5} \times R_3$.

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.3 \end{array}\right)$$

Solution:

$\pi(x^1) = 0.20$

$\pi(x^2) = 0.50$

$\pi(x^3) = 0.30$

What does this all mean?

$\pi(x^1) = 0.20$

	(x^1)	(x^3)	(x^2)
1	**1**	0	0
2	0,25	0,75	0,00
3	0,44	0,19	0,38
4	0,20	0,44	0,36
5	0,27	0,26	0,47
6	0,20	0,34	0,46
7	0,22	0,29	0,49
8	0,20	0,31	0,49
9	0,21	0,29	0,50
10	**0,20**	**0,30**	**0,50**

$\pi(x^2) = 0.50$

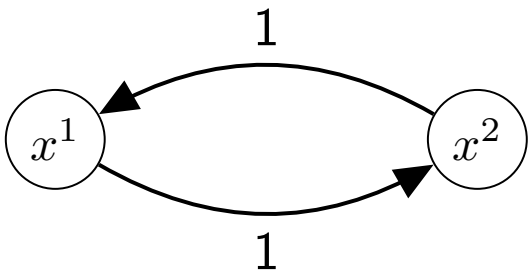
	(x^1)	(x^3)	(x^2)
1	0	**1**	0
2	0,50	0,00	0,50
3	0,13	0,53	0,35
4	0,29	0,20	0,51
5	0,17	0,37	0,45
6	0,23	0,27	0,50
7	0,19	0,32	0,49
8	0,21	0,29	0,50
9	0,20	0,31	0,50
10	**0,20**	**0,30**	**0,50**

$\pi(x^3) = 0.30$

	(x^1)	(x^3)	(x^2)
1	0	0	**1**
2	0,00	0,30	0,70
3	0,15	0,21	0,64
4	0,14	0,30	0,55
5	0,19	0,27	0,54
6	0,18	0,30	0,51
7	0,20	0,29	0,51
8	0,20	0,30	0,50
9	0,20	0,30	0,50
10	**0,20**	**0,30**	**0,50**

Stationary Distribution

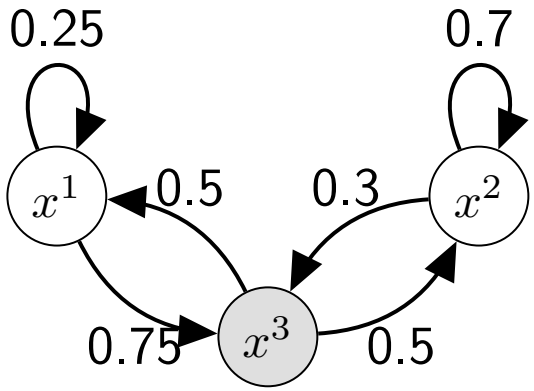
- There is **no guarantee** that our MCMC sampling process will converge to a stationary distribution.
- For example, this Markov chain is periodic.



- There is also **no guarantee** that the stationary distribution is unique.
- Sometimes, the stationary distribution reached also depends on the starting point, $P^{(0)}$.
- This situations occur when several distinct regions are not reachable in the Markov chain (reducible Markov chains).

Ergodic Markov Chains

- It is useful to restrict our attention to Markov chains which have a unique stationary distribution which can be reached from any starting point.
- These type of Markov chains are called **ergodic Markov chains** and must satisfy the following:
  A Markov chain is said to be regular if there exists some number k such that for every pair $x, x' \in Val(\mathbf{X})$, the probability of getting from x to x' in exactly k steps is > 0.



For example, we can get between any pair in 2 steps in the above Markov chain.

Multiple Transition Models

- We can think of each state in a Markov chain being an assignment to several variables.
- However, it is also useful to think of updating only a single component of the state vector at a time.
- In other-words, updating only a single value for a single variable in one transition.
- In this setting we can have multiple transitions models (kernels) $\mathcal{T}_1, \ldots, \mathcal{T}_k$. Each $\mathcal{T}_i$ is associated with a variable $X_i \in \mathbf{X}$.
- $\mathcal{T}_i$ takes us from state (x_{-i}, x_i) to (x_{-i}, x_i') (Gibbs chain)

$$(x_{-i}, x_i) \xrightarrow{\mathcal{T}_i} (x_{-i}, x_i') \quad = P(x_i' \mid x_{-i})$$

- This process is actually the **Gibbs sampling method** we discussed earlier in this lecture.

A Quick Recap

- Using MCMC methods relies on a Markov chain that:
  1. is regular/Ergodic,
  2. and has a stationary distributions which is the desired posterior.
- We described the Gibbs chain which is a general purpose method for sampling in PGMs.
- Only works when we can calculate $P(X_i \mid x_{-i})$ which is easy in discrete models.
- Unfortunately, this is not true for continuous PGMs since it yields an intractable equation.
- Furthermore, Gibbs mixes very slowly where variables have strong correlation.
- Lets now discuss a framework to construct multiple chains (with a desired stationary distribution) using Detailed Balance and the Metropolis-Hasting Algorithm.

The Detailed Balance Equation

- Firstly we need to define the detailed balance equation:

$$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$$

- For example, the stationary distribution $\pi(x^1) = 0.2, \pi(x^2) = 0.5, \pi(x^3) = 0.3.$ and $\mathcal{T}$:

- The detailed balance between x^2 and x^3 is:

$$\pi(x^2)\mathcal{T}(x^2 \to x^3) = \pi(x^3)\mathcal{T}(x^3 \to x^2) = 0.5 \times 0.3 = 0.3 \times 0.5 = 0.15$$

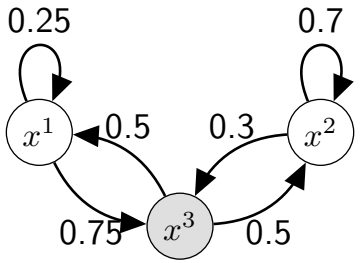
- If $\mathcal{T}$ is regular and it satisfied the detailed balance equation relative to π, then π is the **unique stationary distribution** of $\mathcal{T}$.

The Detailed Balance Equation

- Firstly we need to define the detailed balance equation:

$$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$$

- For example, the stationary distribution $\pi(x^1) = 0.2, \pi(x^2) = 0.5, \pi(x^3) = 0.3.$ and $\mathcal{T}$:



- The detailed balance between x^2 and x^3 is:

$$\pi(x^2)\mathcal{T}(x^2 \to x^3) = \pi(x^3)\mathcal{T}(x^3 \to x^2) = 0.5 \times 0.3 = 0.3 \times 0.5 = 0.15$$

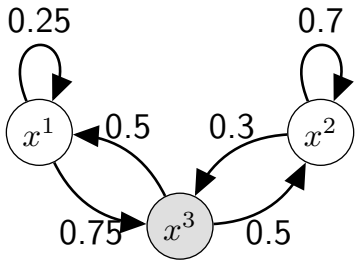
- If $\mathcal{T}$ is regular and it satisfied the detailed balance equation relative to π, then π is the **unique stationary distribution** of $\mathcal{T}$.

The Detailed Balance Equation

- Firstly we need to define the detailed balance equation:

$$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$$

- For example, the stationary distribution $\pi(x^1) = 0.2, \pi(x^2) = 0.5, \pi(x^3) = 0.3.$ and $\mathcal{T}$:



- The detailed balance between x^2 and x^3 is:

$$\pi(x^2)\mathcal{T}(x^2 \to x^3) = \pi(x^3)\mathcal{T}(x^3 \to x^2) = 0.5 \times 0.3 = 0.3 \times 0.5 = 0.15$$

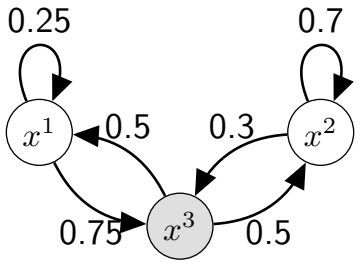
- If $\mathcal{T}$ is regular and it satisfied the detailed balance equation relative to π, then π is the **unique stationary distribution** of $\mathcal{T}$.

The Detailed Balance Equation

- Firstly we need to define the detailed balance equation:

$$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$$

- For example, the stationary distribution $\pi(x^1) = 0.2, \pi(x^2) = 0.5, \pi(x^3) = 0.3.$ and $\mathcal{T}$:



- The detailed balance between x^2 and x^3 is:

$$\pi(x^2)\mathcal{T}(x^2 \to x^3) = \pi(x^3)\mathcal{T}(x^3 \to x^2) = 0.5 \times 0.3 = 0.3 \times 0.5 = 0.15$$

- If $\mathcal{T}$ is regular and it satisfied the detailed balance equation relative to π, then π is the **unique stationary distribution** of $\mathcal{T}$.

The Metropolis-Hastings Algorithm Intuition

- The Metropolis-Hastings Algorithm is a general construction that allows us to build a reversible Markov chain with a particular stationary distribution.
- The intuition is that since we can not sample from the desired posterior, we rather sample from any distribution (called a proposed distribution), but then correct for the resulting error.
- In other words, we randomly choose whether to accept a proposed transition, with a probability that corrects for the discrepancy between the proposal distribution and the target.

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

The Metropolis-Hastings Algorithm Details

1. Suppose we have a proposed distribution $\mathcal{T}^Q$ which defines a transition model.
2. That is, for each state x, $\mathcal{T}^Q$ defines a distribution over possible successor states in $Val(\mathbf{X})$.
3. From these possible assignment states, we randomly select x' as the next possible transition.
4. We can either accept this transition, or reject it (staying at x) using acceptance probability $\mathcal{A}(x \to x')$.
5. The transition model for the Markov chain is then:

$$\mathcal{T}(x \to x') = \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \text{ for } x \neq x'$$

$$\mathcal{T}(x \to x) = \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')).$$

Choosing the Distributions

- For $\mathcal{T}^Q$, you can pick any distribution you like. As long as it induces a regular chain.
- For $\mathcal{A}(x \to x')$, we can to incorporate the detailed balance equation to obtain the desired stationary distribution.
- Incorporating the acceptance probability we get:
  $\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x)$
- Moving things around we get:

$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x' \to x)}{\pi(x)\mathcal{T}^Q(x \to x')}].$$

- For PGMs, Matropolis-Hastings extends naturally:

$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i \to x_{-i}, x_i')}].$$

Choosing the Distributions

- For $\mathcal{T}^Q$, you can pick any distribution you like. As long as it induces a regular chain.
- For $\mathcal{A}(x \to x')$, we can to incorporate the detailed balance equation to obtain the desired stationary distribution.
- Incorporating the acceptance probability we get:
$$\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x)$$
- Moving things around we get:
$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x' \to x)}{\pi(x)\mathcal{T}^Q(x \to x')}].$$
- For PGMs, Matropolis-Hastings extends naturally:
$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i \to x_{-i}, x_i')}].$$

Choosing the Distributions

- For $\mathcal{T}^Q$, you can pick any distribution you like. As long as it induces a regular chain.
- For $\mathcal{A}(x \to x')$, we can to incorporate the detailed balance equation to obtain the desired stationary distribution.
- Incorporating the acceptance probability we get:
  $\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x)$
- Moving things around we get:

$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x' \to x)}{\pi(x)\mathcal{T}^Q(x \to x')}].$$

- For PGMs, Matropolis-Hastings extends naturally:

$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i \to x_{-i}, x_i')}].$$

Choosing the Distributions

- For $\mathcal{T}^Q$, you can pick any distribution you like. As long as it induces a regular chain.
- For $\mathcal{A}(x \to x')$, we can to incorporate the detailed balance equation to obtain the desired stationary distribution.
- Incorporating the acceptance probability we get:
  $\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x)$
- Moving things around we get:

$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x' \to x)}{\pi(x)\mathcal{T}^Q(x \to x')}].$$

- For PGMs, Matropolis-Hastings extends naturally:

$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i \to x_{-i}, x_i')}].$$

Choosing the Distributions

- For $\mathcal{T}^Q$, you can pick any distribution you like. As long as it induces a regular chain.
- For $\mathcal{A}(x \to x')$, we can to incorporate the detailed balance equation to obtain the desired stationary distribution.
- Incorporating the acceptance probability we get:
  $\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x)$
- Moving things around we get:

$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x' \to x)}{\pi(x)\mathcal{T}^Q(x \to x')}].$$

- For PGMs, Matropolis-Hastings extends naturally:

$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^Q(x_{-i}, x_i \to x_{-i}, x_i')}].$$

Mixing Time and Using Samples

- So how long do we need to mix the Markov chain until we can collect a sample?
- Unfortunately, you **can not prove that a Markov chain has mixed**.
- However, we can use some heuristics to show that it hasn't mixed.
- We can compute chain statistics, such as the ...
  - Log probability of a sample, or
  - probability that a state is in a set
- in ...
  - different windows of a single run of the chain.
  - different runs that are initialised differently.
- For using samples you may want to **use every other sample** since they are correlated.