

Class 14

Markov Chains II

Announcements

- HW6 due Friday!
- HW7 out now!
 - Due a week from Friday.
- Midterms have been graded!
 - Median: 73
 - Mean: 72
 - Std. Dev: 14
 - Maximum: 100
- Great job everyone!! This was a tough exam.
 - We will curve up when computing letter grades at the end of the quarter as needed 😊

Recap: More Markov Chains!

- **Definitions!**

- A chain is **irreducible** if you can get to anywhere from anywhere else.
- A state is **recurrent** if you'll return to it eventually with probability 1.
- Otherwise it is **transient**.

- A chain is **periodic** if there's a state that you can only reach on multiples of c , for some integer $c > 1$.
- Otherwise it is **aperiodic**.
 - Useful fact: any irreducible chain with a self-loop is aperiodic!

Recap: Fundamental Theorem of Markov Chains

- Any irreducible and aperiodic Markov chain over a finite state space has a unique **stationary distribution** π .
- As t gets big, $X_t \rightarrow \pi$
- $\pi P = \pi$, aka if $X_t \sim \pi$, then $X_{t+1} \sim \pi$
- If you start in state i , the expected amount of time to return is $\frac{1}{\pi_i}$

Tie-in to last time...

- **Proposition:** If a Markov chain has symmetric transitions, and is aperiodic and irreducible, then the stationary distribution is uniform.

Symmetric

Always true

$$\begin{array}{c}
 \boxed{P} \begin{array}{|c|} \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline \end{array} = \begin{array}{|c|} \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline \end{array} \Rightarrow \begin{array}{c} \boxed{1\ 1\ 1\ 1\ 1\ 1} \boxed{P} = \boxed{1\ 1\ 1\ 1\ 1\ 1} \\
 \end{array}
 \end{array}$$

uniform dist is stationary!

Irreducible

P has only one eigenvalue that's equal to 1.

Aperiodic

P only has one eigenvalue with **magnitude 1**

Then, the MC converges to a uniform distribution!

$$\boxed{P^t} = \frac{1}{n} \begin{array}{|c|} \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline \end{array} \boxed{V} \begin{array}{|c|} \hline 1 \\ \hline \text{tiny} \\ \hline \text{tiny} \\ \hline \text{tiny} \\ \hline \end{array} \begin{array}{|c|} \hline 1\ 1\ 1\ 1\ 1\ 1 \\ \hline \boxed{V^*} \\ \hline \end{array} \rightarrow \frac{1}{n} \begin{array}{|c|} \hline 1\ 1\ 1\ 1\ 1 \\ \hline 1\ 1\ 1\ 1\ 1 \\ \hline 1\ 1\ 1\ 1\ 1 \\ \hline \end{array}$$

Metropolis algorithm and MCMC

- **Markov Chain Monte Carlo:**

- Set up a Markov Chain with a particular desired stationary distribution.
- To sample from that distribution, run the chain for a while!

- **Metropolis Algorithm:**

- A particular way to set up such a chain.

$$P_{i,j} = \begin{cases} 0 & \text{if } i, j \text{ not neighbors} \\ \frac{1}{d} \min(1, \frac{\pi(j)}{\pi(i)}) & \text{if } i \neq j \text{ and they are neighbors} \\ 1 - \sum_{\ell \neq i} P_{i,\ell} & \text{if } i = j. \end{cases}$$

Questions?

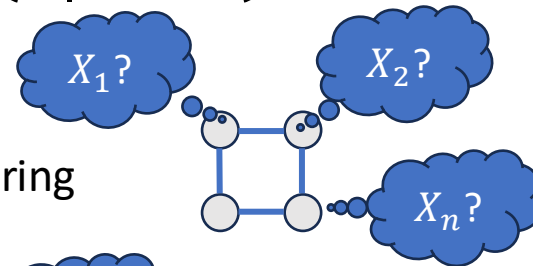
Fundamental Theorem, Metropolis Alg? Quiz?

Today: Gibbs Sampling!

- An MCMC algorithm for multivariate distributions.
- Set-up:
 - π is a joint distribution on random variables X, Y
 - More generally X_1, X_2, \dots, X_m
 - It's hard to sample from π
 - But it's easy to sample from $\pi(X | Y = y)$ or $\pi(Y | X = x)$ for any fixed x, y .

Example: It's not obvious how to sample a uniformly random proper coloring

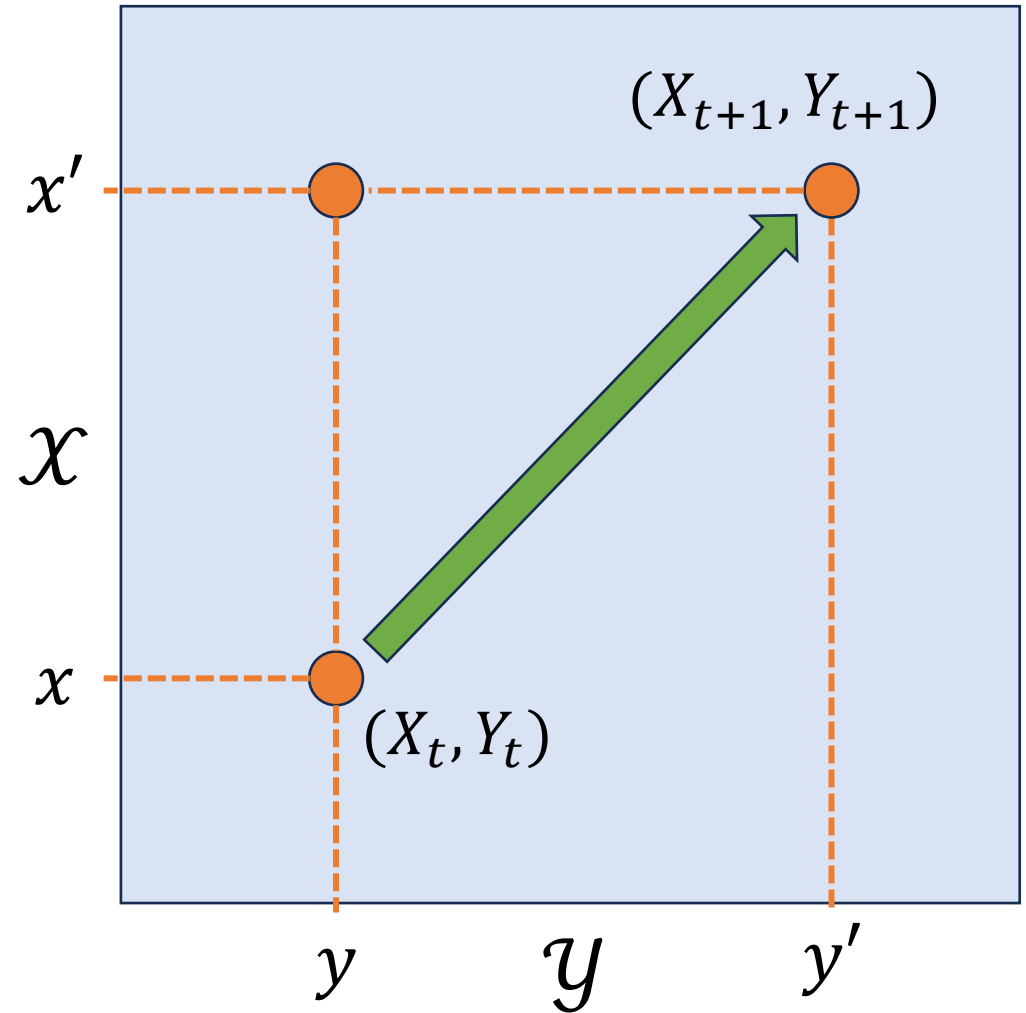
Conditional on all but one of the vertices, it's easy!



Gibbs Sampling

(for two variables)

- Say $(X_t, Y_t) = (x, y)$
- Draw $x' \sim \pi(X|Y = y)$
- Draw $y' \sim \pi(Y|X = x')$
- Set $(X_{t+1}, Y_{t+1}) = (x', y')$



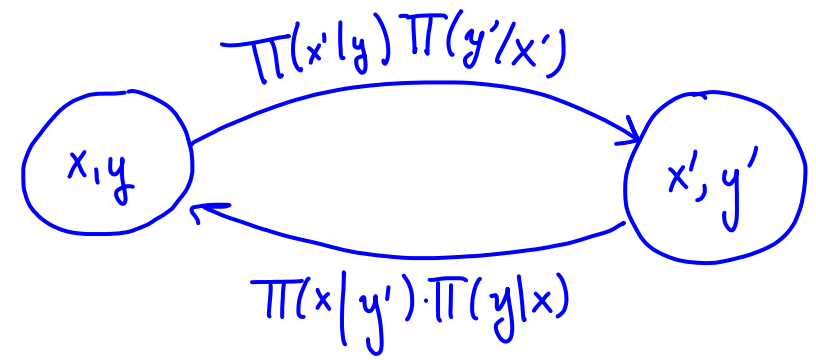
Group Work!

As usual, hints and more info on handout!

1. Show that π is a stationary distribution.
2. Under what conditions on π does FToMC hold?
3. What is the take-away in the context of MCMC?
4. How would you use Gibbs sampling to sample random colorings?
5. [Open-ended] Any other applications of Gibbs sampling/MCMC that you've encountered or can think of?

- Say $(X_t, Y_t) = (x, y)$
- Draw $x' \sim \pi(X|Y = y)$
- Draw $y' \sim \pi(Y|X = x')$
- Set $(X_{t+1}, Y_{t+1}) = (x', y')$

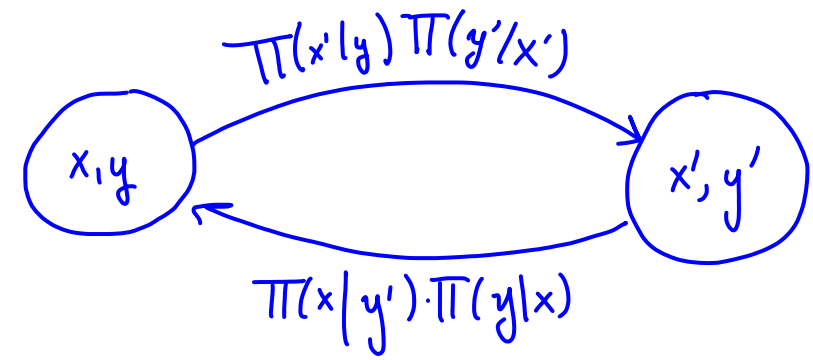
1. Stationary dist is π



- Show $\pi(x, y) = \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)]$

$$\begin{aligned} \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)] &= \sum_{x', y'} \pi(x', y') \pi(x|y') \pi(y|x) \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \sum_{x'} \pi(x', y') \end{aligned}$$

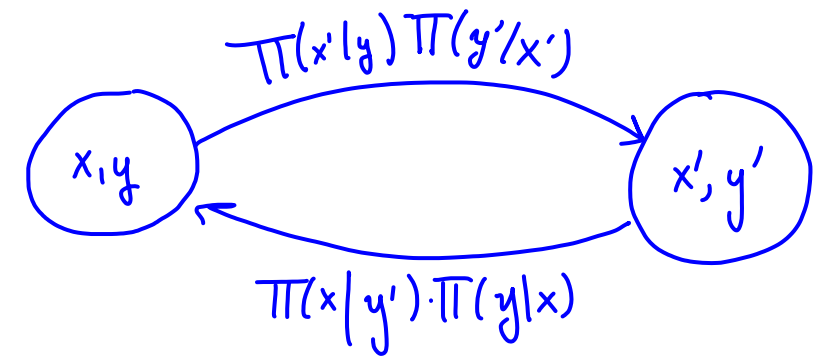
1. Stationary dist is π



- Show $\pi(x, y) = \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)]$

$$\begin{aligned} \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)] &= \sum_{x', y'} \pi(x', y') \pi(x|y') \pi(y|x) \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \sum_{x'} \pi(x', y') \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \pi(y') \end{aligned}$$

1. Stationary dist is π



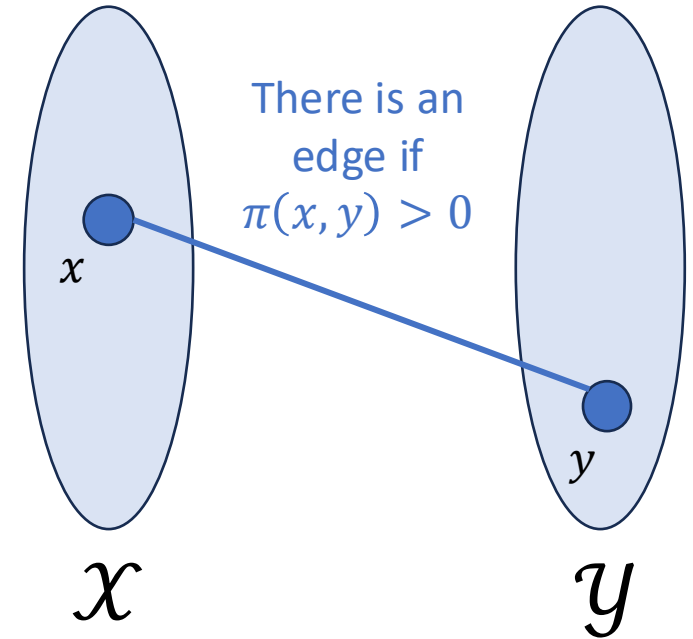
- Show $\pi(x, y) = \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)]$

$$\begin{aligned} \sum_{x', y'} \pi(x', y') \Pr[(x', y') \rightarrow (x, y)] &= \sum_{x', y'} \pi(x', y') \pi(x|y') \pi(y|x) \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \sum_{x'} \pi(x', y') \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \pi(y') \\ &= \pi(y|x) \pi(x) \\ &= \pi(x, y) \end{aligned}$$

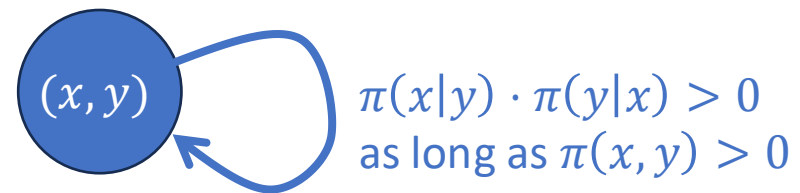
2. Do the conditions hold?

- Irreducible: Not without more assumptions!

We can view our states as edges in this graph:
We need this graph to be connected.



- Aperiodic: Assuming irreducible, yes!



- Finite: As long as π has finite support.

3. Why is this useful?

Suppose we can sample from the marginals...

- If we can easily sample from $\pi(X | Y = y)$ or $\pi(Y | X = x)$, then we can sample (X_t, Y_t) .
- The FToMC says that $(X_t, Y_t) \rightarrow \pi$ as $t \rightarrow \infty$
- So eventually we can sample from $\pi!$
 - But how long does it take to converge???? More next time!

4. Graph Coloring

- Say $(X_t, Y_t) = (x, y)$
- Draw $x' \sim \pi(X|Y = y)$
- Draw $y' \sim \pi(Y|X = x')$
- Set $(X_{t+1}, Y_{t+1}) = (x', y')$

- What is the “right” multivariate generalization?

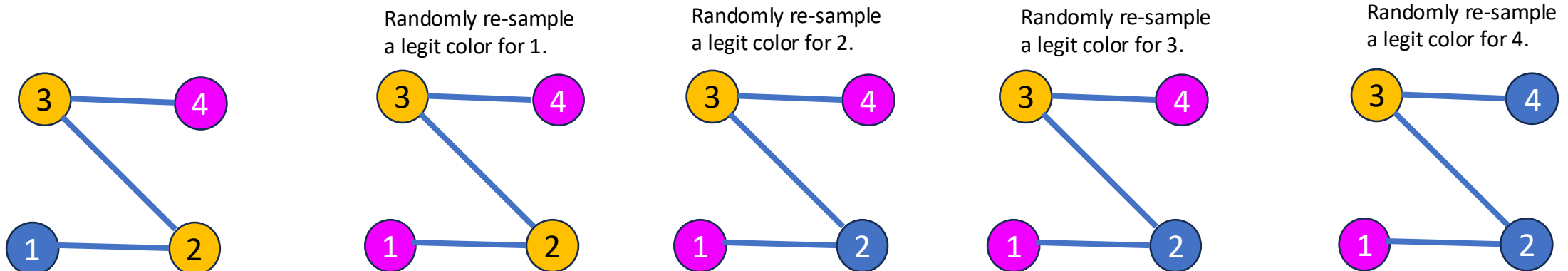
Let's just do it for $n = 3$ variables to save on subscripts...

Say $X_t = (x, y, z)$
Sample $x' \sim P(X | y, z)$
Sample $y' \sim P(Y | x', z)$
Sample $z' \sim P(Z | x', y')$
Let $X_{t+1} = (x', y', z')$

Note: other reasonable generalizations (also called Gibbs sampling) include:

- Update X_t with each intermediate step
- Choose random coordinates to update.

- How to apply to sampling a uniformly random proper coloring?



What is Gibbs sampling useful for?

- Imputing missing data (e.g., survey data, medical records)
- Image denoising
- Clustering
- Recommendation Systems
- Uncertainty Quantification
- ...

Everything from here on is super handwavey and you are not responsible for it for HW or the final exam!

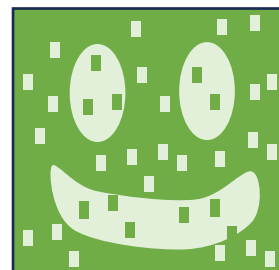
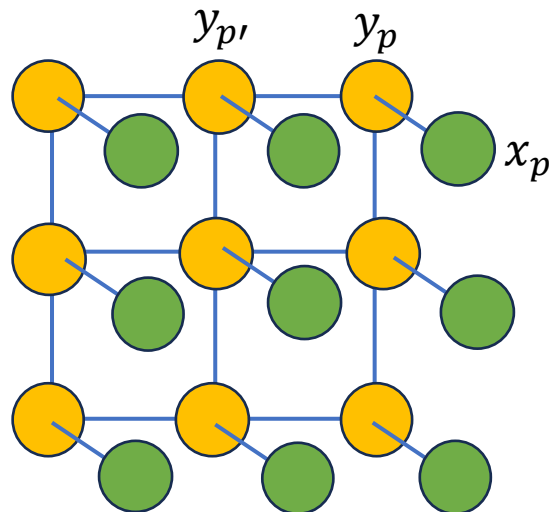
Another Example: Image Denoising

- Say you get a noisy (black and white, say) image $X = (x_1, \dots, x_N)$.
 - Each pixel x_p is ± 1
- Sample an “un-noisy” version $Y = (y_1, \dots, y_N)$, so that the probability of Y is proportional to:

$$\exp(\eta \sum_p x_p y_p + \beta \sum_{p \sim p'} y_p y_{p'})$$

This distribution is hard to sample from directly, but the marginals are easy!

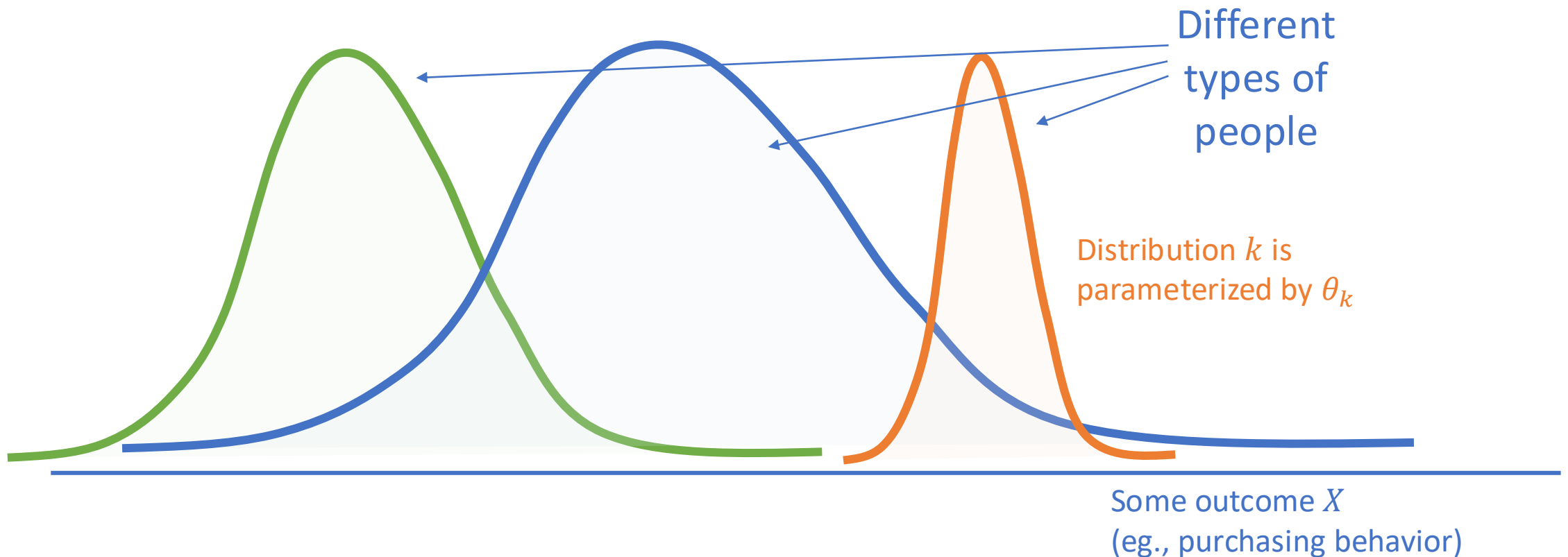
Intuition: every node “wants” to be the same as its neighbors



This is a really high-probability y under this distribution

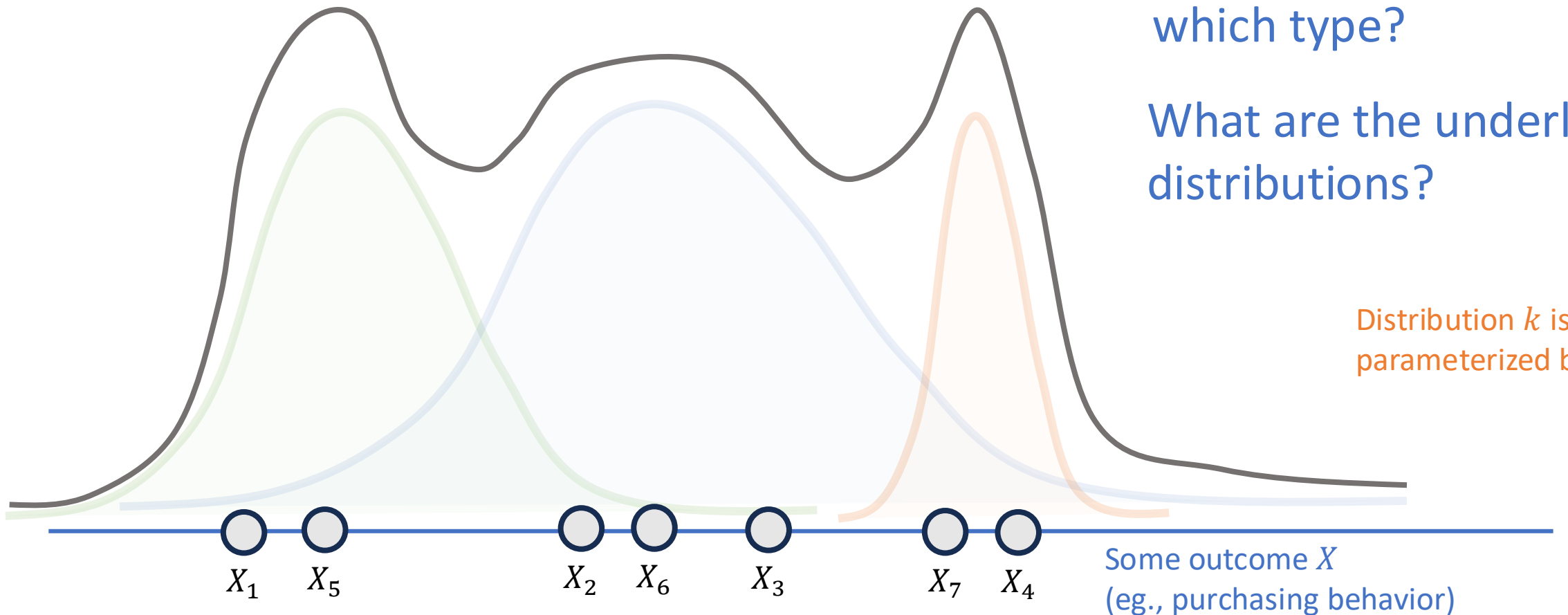
Another Example: Clustering

- Say we think that the “true” model of the world is:



Another Example: Clustering

- What we see: Draws from the mixture:



Which person is of which type?

What are the underlying distributions?

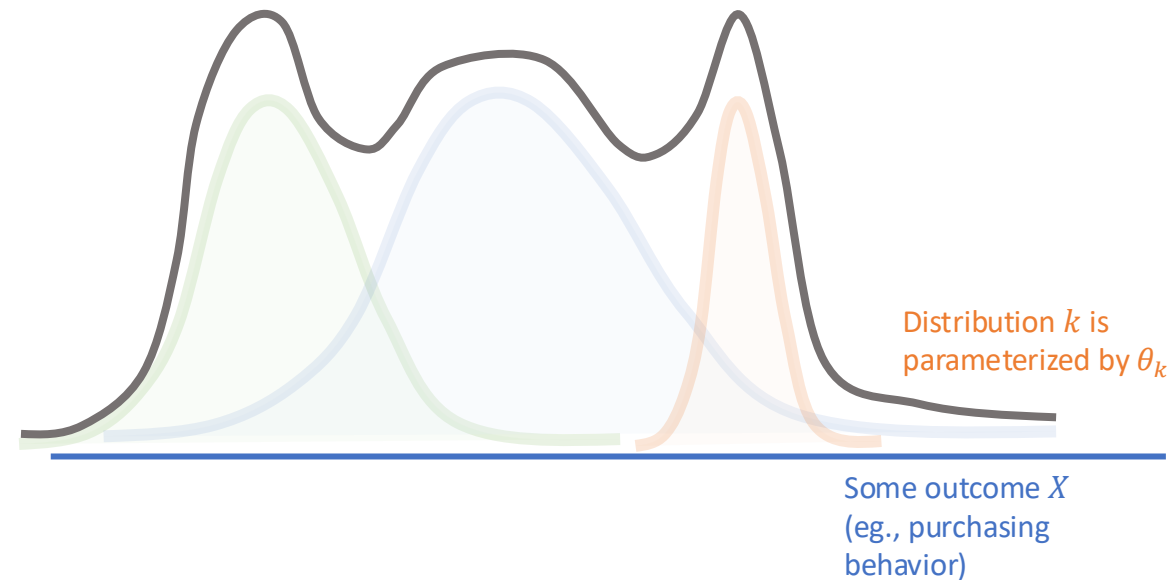
Distribution k is parameterized by θ_k

Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

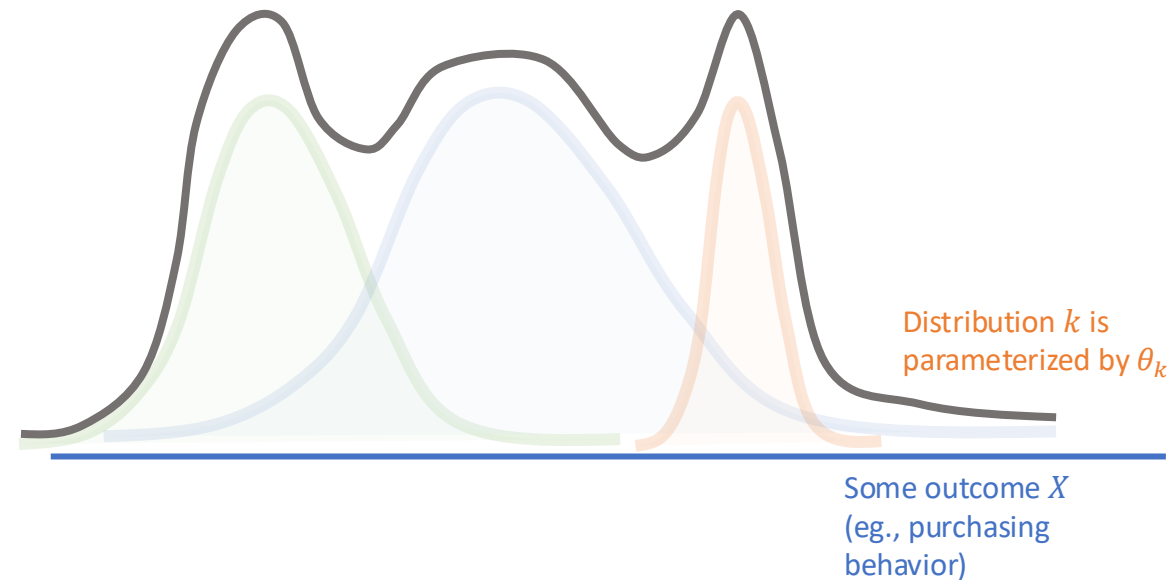


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

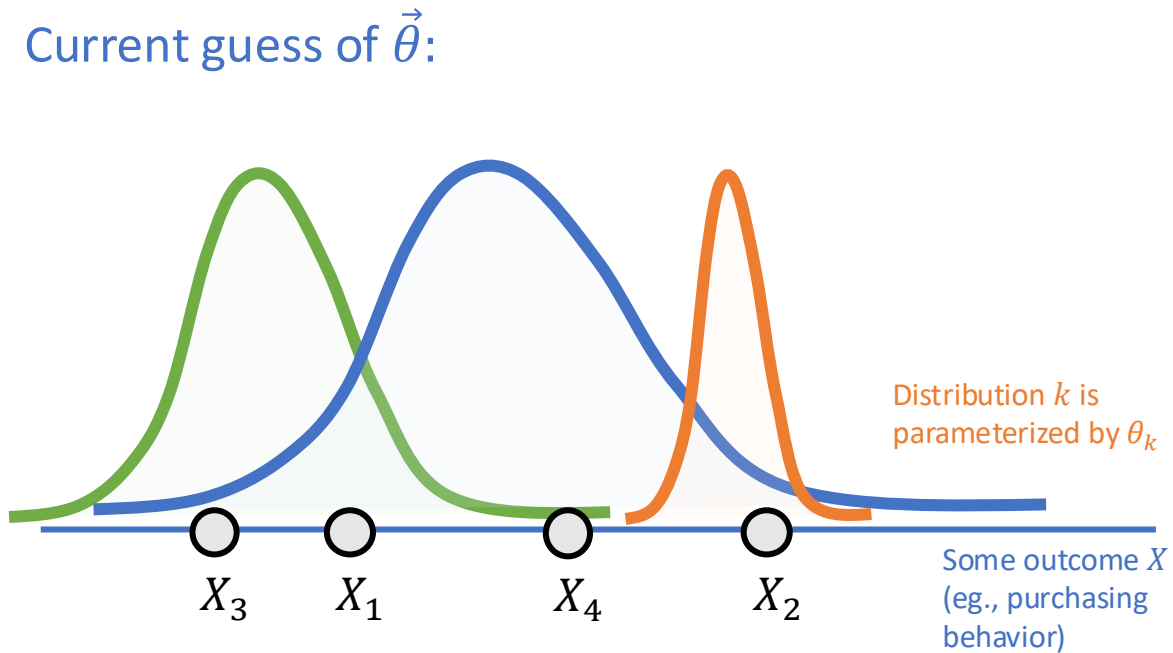


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

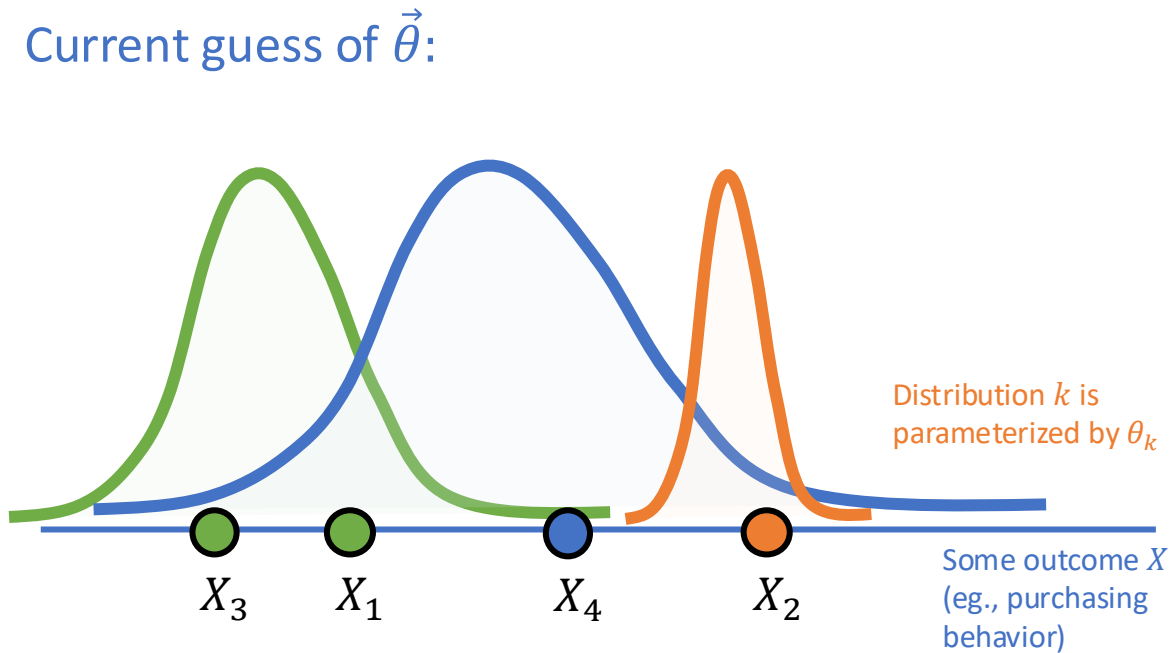


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

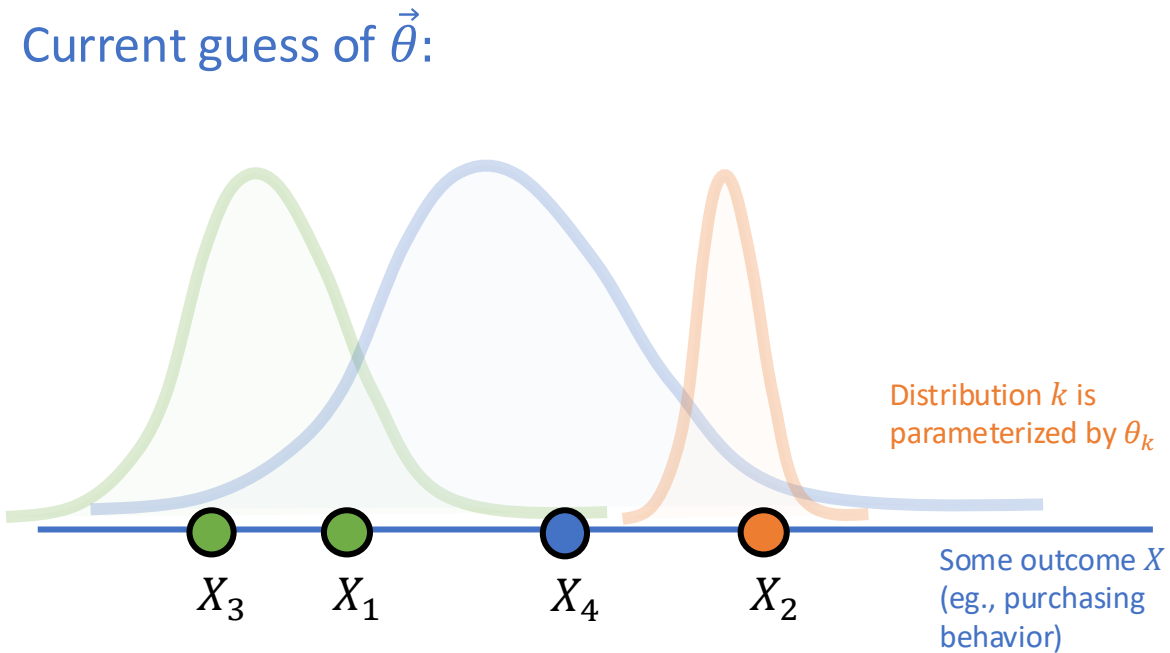


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

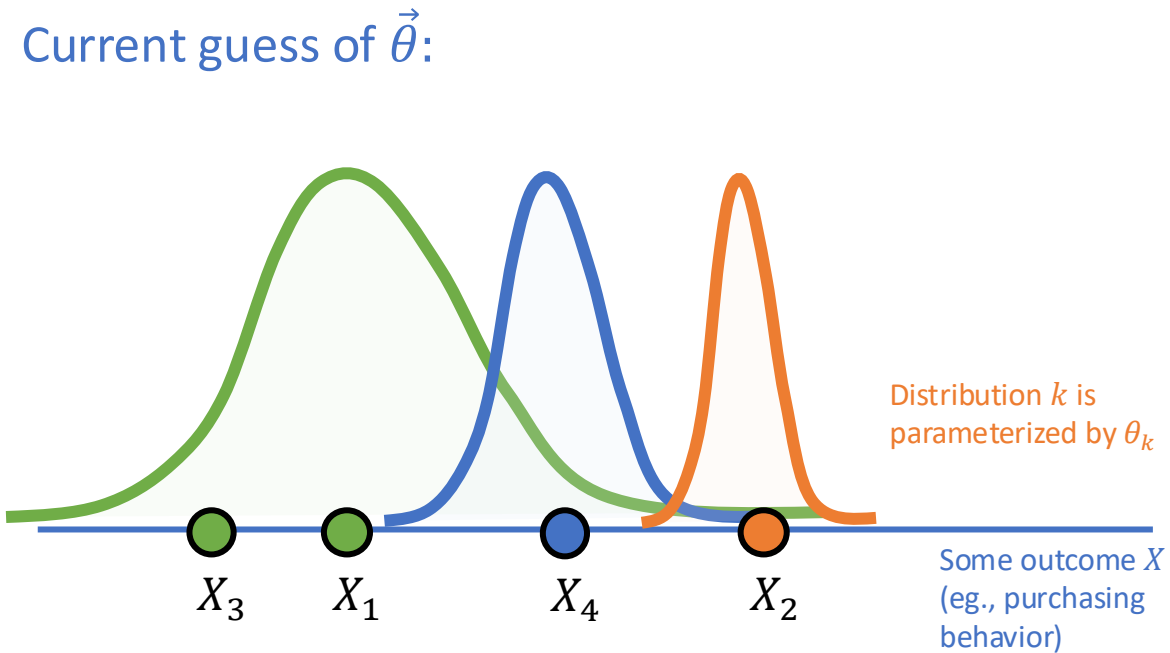


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$

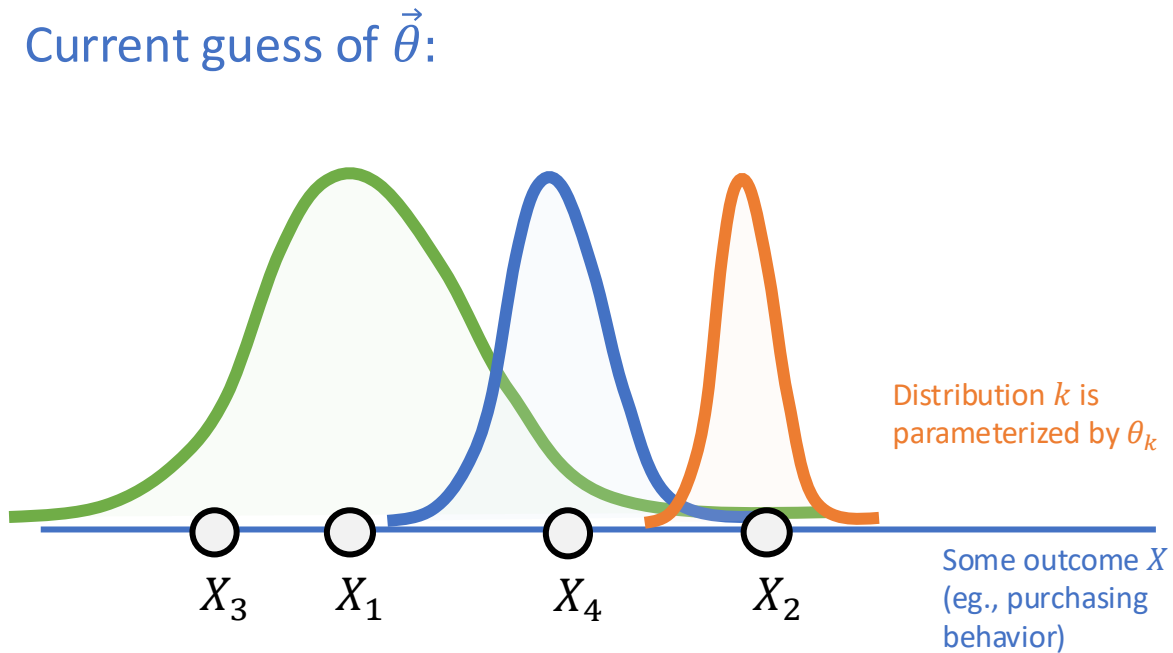


Another Example: Clustering

- We'd like to sample:
 - $\theta_1, \theta_2, \dots, \theta_k$ (parameters for each distribution)
 - z_1, z_2, \dots, z_N (types for each person)
 - ...so that we are very likely to choose $(\vec{\theta}; \vec{z})$ under which our samples were very likely.

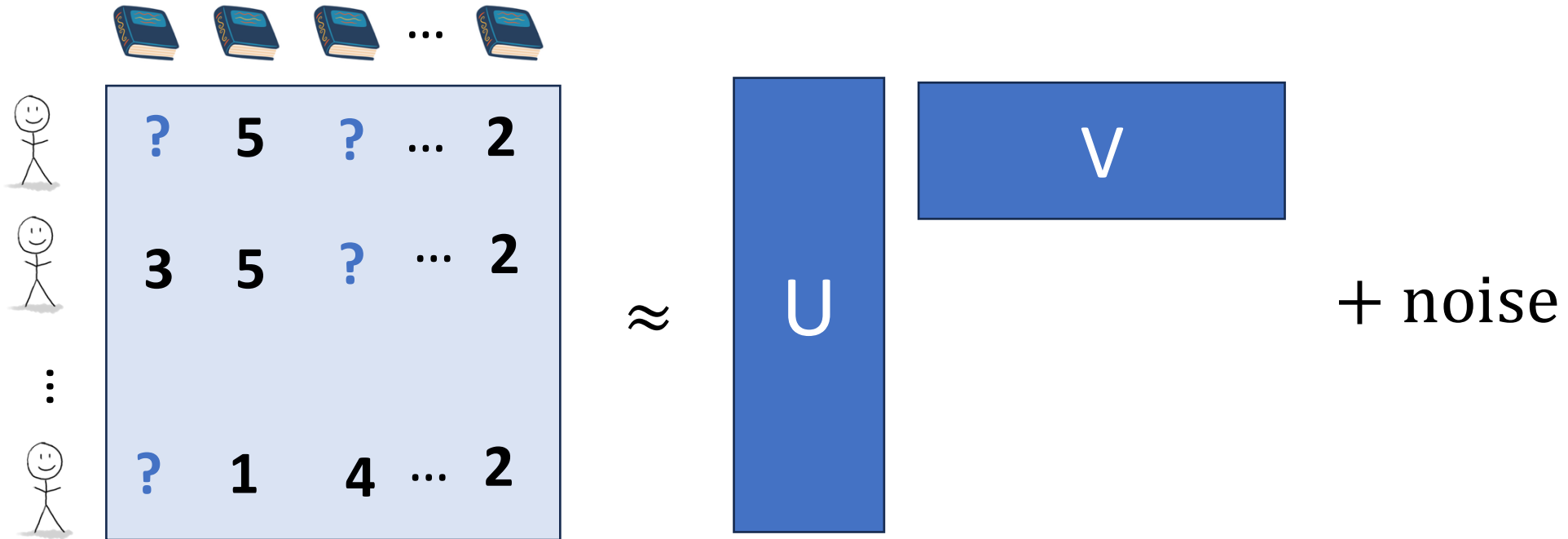
Gibbs sampling:

- For $i = 1, \dots, N$:
 - Resample z_i so that $\Pr[z_i = k] \propto \Pr[X_i | \theta_k]$
- For $j = 1, \dots, k$:
 - Resample θ_k to best fit the X_i with $z_i = k$



Another Example: Recommendation Engines

- Say our model of the world is:



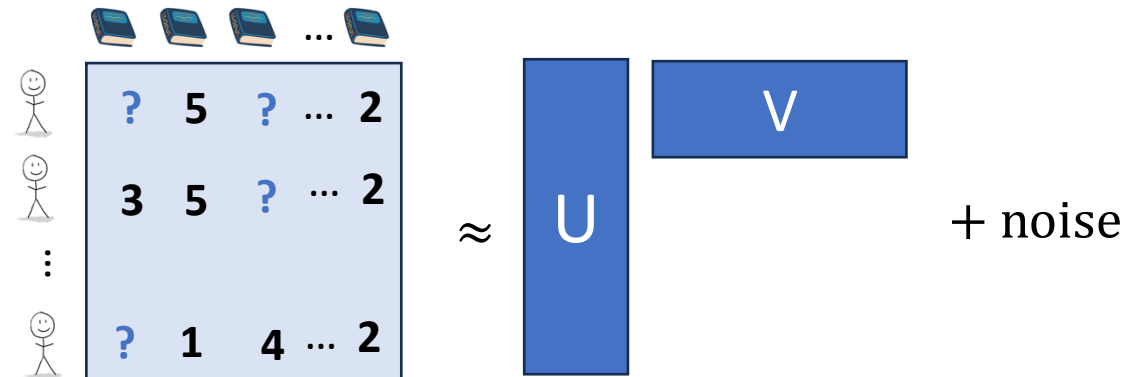
Another Example: Recommendation Engines

- We want to learn U, V , so that we can predict the missing entries.

Gibbs sampling:

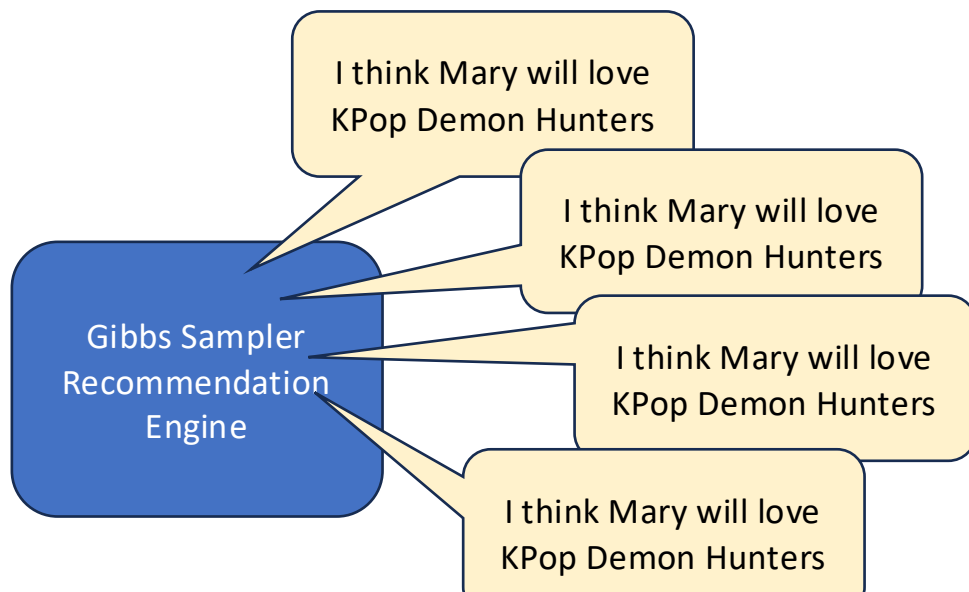
- For $i = 1, \dots, N$:
 - Resample $U_i \propto \Pr[\text{person } i\text{'s ratings} \mid V, U_i]$
- For $i = 1, \dots, N$:
 - Resample $V_i \propto \Pr[\text{person } i\text{'s ratings} \mid V_i, U]$

The point: If U is fixed, it's easy to sample V , and vice versa.

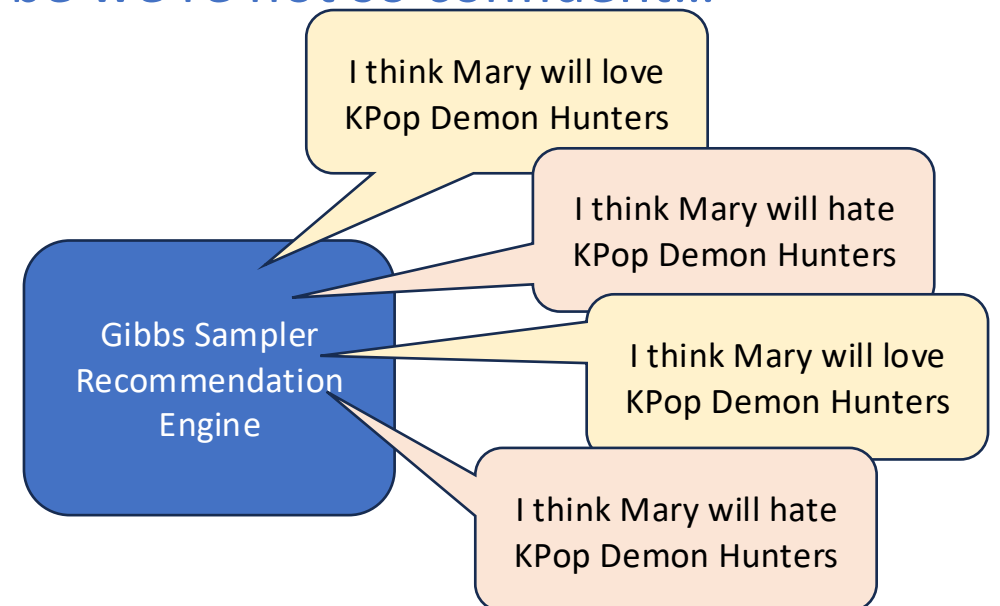


Uncertainty quantification

- We do some task (clustering, recommending, etc)
- We want to know how likely we were to get a pretty good answer
- Gibbs sampling not only gives us a point estimate, it gives us some idea of the distribution we are sampling from.
 - If that distribution is super spread out, maybe we're not so confident...



VS



5. Other examples?

- Y'all come from many different areas – have any of you used Gibbs sampling or any other MCMC method before? For what applications?

Recap

- The fundamental theorem of Markov chains can be useful!
- But it sure would be more useful if we knew how fast we approached the stationary distribution...
 - Next time!