Vignettes in Learning Theory II: Graphical Models

Ankur Moitra (MIT)

Columbia Data Science Institute, May 2nd

Let me tell you a story about learning and a story about sampling….

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- **Ising Models and Their Applications**
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

ISING MODELS

Rich model for defining high-dimensional distributions based on pairwise interactions

ISING MODELS

Rich model for defining high-dimensional distributions based on pairwise interactions

Definition: An **Ising model** is a distribution on $\{\pm 1\}^n$ with

$$\mathbb{P}[X = x] = \frac{1}{Z} \exp\Big(\sum_{i,j} J_{i,j} x_i x_j + \sum_i h_i x_i \Big)$$

interaction matrix **external field**

where Z is the partition function

ISING MODELS

Rich model for defining high-dimensional distributions based on pairwise interactions

Definition: An **Ising model** is a distribution on $\{\pm 1\}^n$ with

$$\mathbb{P}[X = x] = \frac{1}{Z} \exp\Big(\sum_{i,j} J_{i,j} x_i x_j + \sum_i h_i x_i \Big)$$

interaction matrix **external field**

where Z is the partition function

Wide-ranging applications in science and engineering

CONDITIONAL INDEPENDENCE

Often helpful to look at their graph structure:

$G = (\{X_1, \cdots, X_n\}, E)$ with $E = \{(X_i, X_j) \text{ s.t. } J_{i,j} \neq 0\}$

CONDITIONAL INDEPENDENCE

Often helpful to look at their graph structure:

$G = (\{X_1, \cdots, X_n\}, E)$ with $E = \{(X_i, X_j) \text{ s.t. } J_{i,j} \neq 0\}$

Markov Property: Two nodes are independent when conditioned on a separator

CONDITIONAL INDEPENDENCE

Often helpful to look at their graph structure:

$G = (\{X_1, \cdots, X_n\}, E)$ with $E = \{(X_i, X_j) \text{ s.t. } J_{i,j} \neq 0\}$

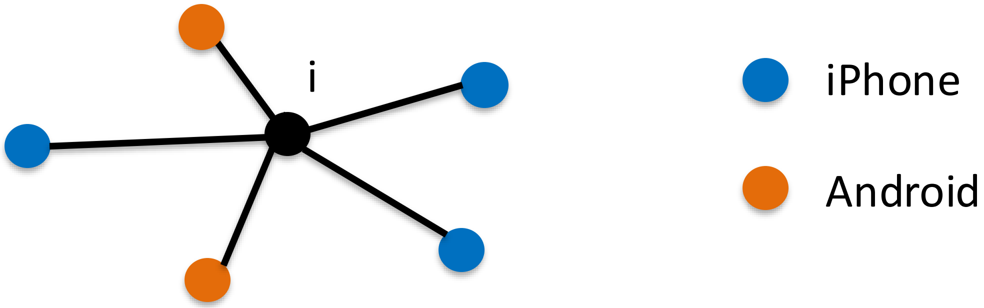
Markov Property: Two nodes are independent when conditioned on a separator – i.e.

$$X_i \perp X_j | X_U$$

provided that all paths from X_i to X_j pass through X_U

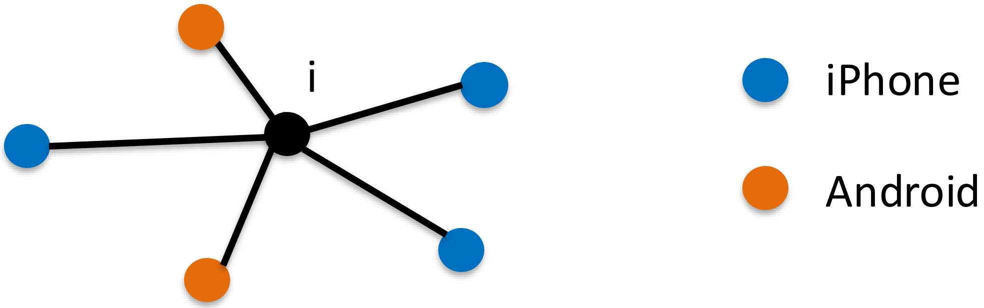
RUNNING EXAMPLE

Modeling the adoption of technology in a social network



RUNNING EXAMPLE

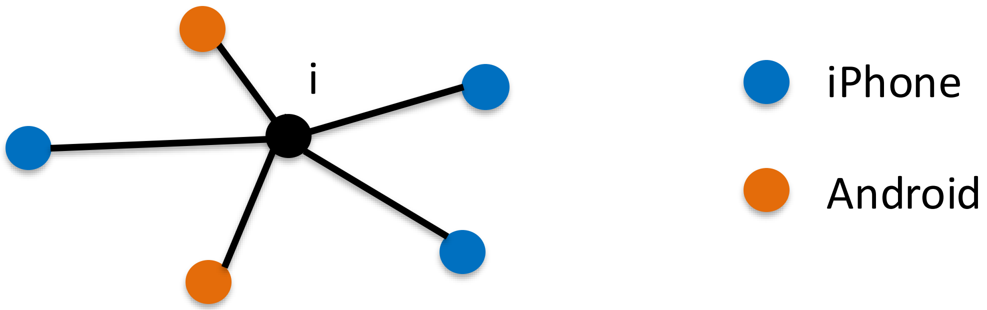
Modeling the adoption of technology in a social network



Many works formalize as an Ising model

RUNNING EXAMPLE

Modeling the adoption of technology in a social network



Many works formalize as an Ising model

Each person's decision is independent of everyone else, conditioned on his neighbors

HISTORY

What’s known about learning Ising models from random samples?

[Chow, Liu ‘68]: Polynomial time algorithm on trees

HISTORY

What’s known about learning Ising models from random samples?

[Chow, Liu ‘68]: Polynomial time algorithm on trees

[Karger, Srebro ‘01]: Polynomial time algorithm on graphs of bounded treewidth

HISTORY

What's known about learning Ising models from random samples?

[Chow, Liu '68]: Polynomial time algorithm on trees

[Karger, Srebro '01]: Polynomial time algorithm on graphs of bounded treewidth

[Bresler '15]: Polynomial time algorithm on graphs of bded degree (doubly-exponential dependence on max degree)

HISTORY

What's known about learning Ising models from random samples?

[Chow, Liu '68]: Polynomial time algorithm on trees

[Karger, Srebro '01]: Polynomial time algorithm on graphs of bounded treewidth

[Bresler '15]: Polynomial time algorithm on graphs of bded degree (doubly-exponential dependence on max degree)

Improved to singly-exponential in **[Vuffray et al. '16]** and **[Klivans, Meka '17]**, which is tight

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- **Markov Random Fields and Lower Bounds**

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

MARKOV RANDOM FIELDS

Even more expressive model allows higher-order interactions:

MARKOV RANDOM FIELDS

Even more expressive model allows higher-order interactions:

Definition: A **Markov random field of order k** is a distribution on $\{\pm 1\}^n$ with

$$\mathbb{P}[X = x] = \frac{1}{Z} \exp\Big(f(x) \Big)$$

degree k polynomial

where, again, Z is the partition function

MARKOV RANDOM FIELDS

Even more expressive model allows higher-order interactions:

Definition: A **Markov random field of order k** is a distribution on $\{\pm 1\}^n$ with

$$\mathbb{P}[X = x] = \frac{1}{Z} \exp\Big(\underset{\uparrow}{f(x)} \Big)$$

degree k polynomial

where, again, Z is the partition function

How do higher-order interactions affect learning?

HISTORY, CONTINUED

Algorithms and hardness for higher order dependencies

[Klivans, Meka '17], **[Hamilton et al. '17]**: There are $n^{O(k)}$ time algorithms for learning order k MRFs on n variables with bounded degree

HISTORY, CONTINUED

Algorithms and hardness for higher order dependencies

[Klivans, Meka '17], **[Hamilton et al. '17]**: There are $n^{O(k)}$ time algorithms for learning order k MRFs on n variables with bounded degree

[Bresler et al. '14], **[Klivans, Meka '17]**: Under standard hardness assumptions, learning an order k MRF on n variables takes $n^{\Omega(k)}$ time

learning a k-sparse parity w/ noise takes time $n^{\Omega(k)}$

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

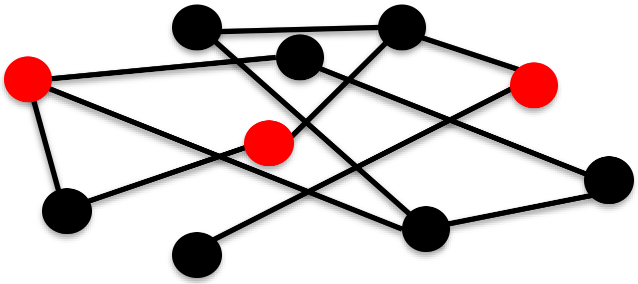
There is a natural dynamics that eventually reaches the target distribution

GLAUBER DYNAMICS

Create a Markov chain on the state space with local updates

GLAUBER DYNAMICS

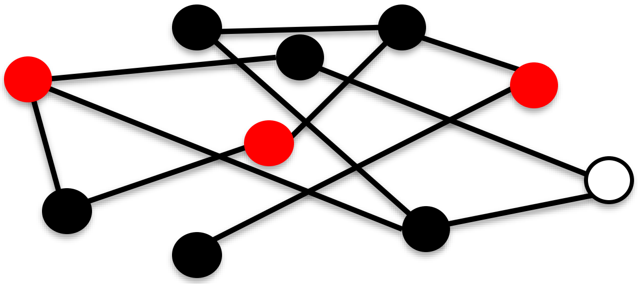
Create a Markov chain on the state space with local updates



GLAUBER DYNAMICS

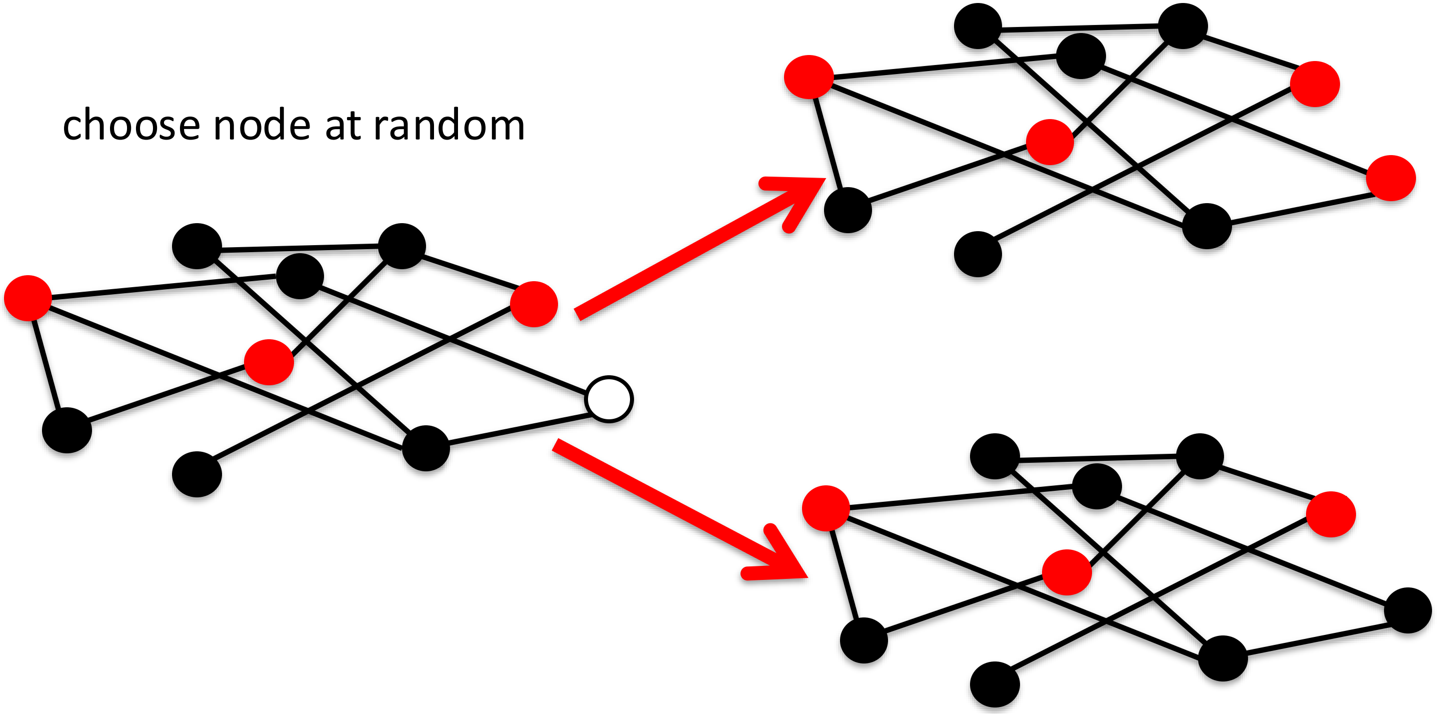
Create a Markov chain on the state space with local updates

choose node at random



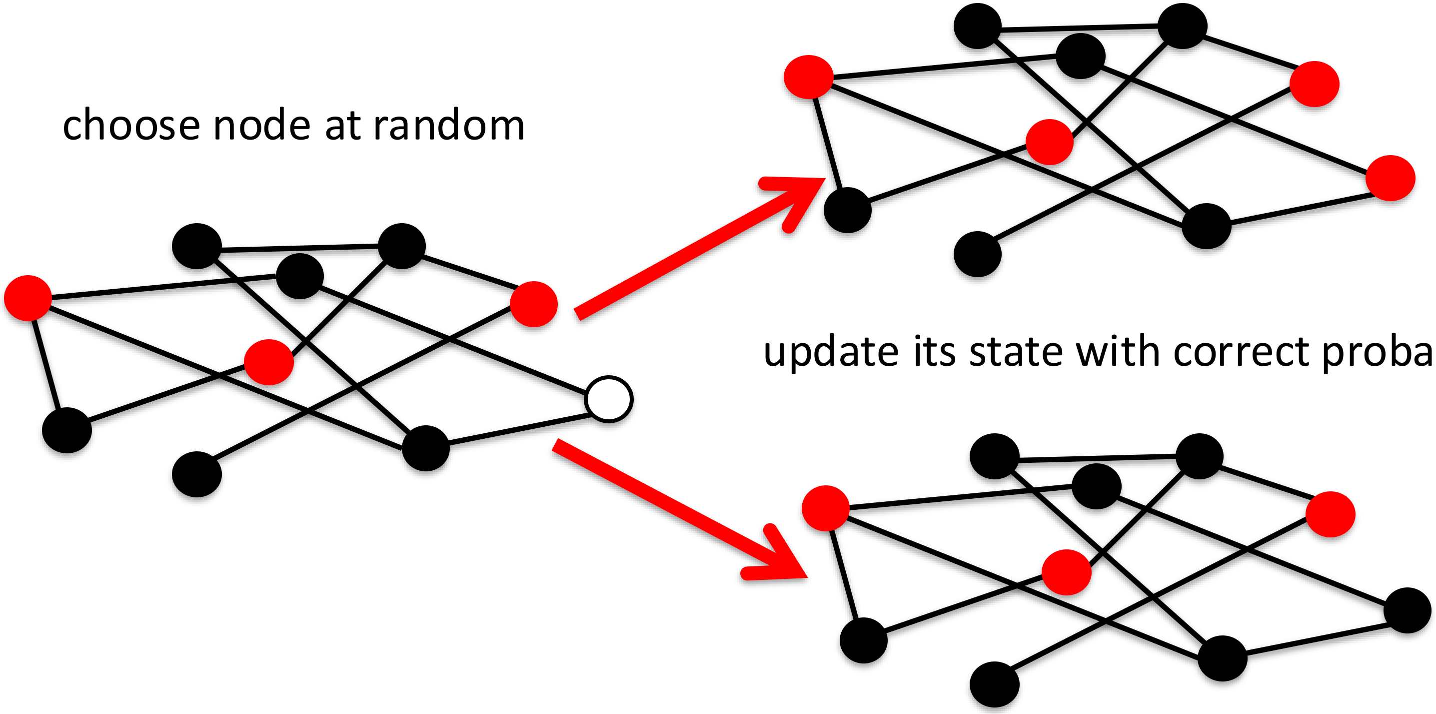
GLAUBER DYNAMICS

Create a Markov chain on the state space with local updates



GLAUBER DYNAMICS

Create a Markov chain on the state space with local updates

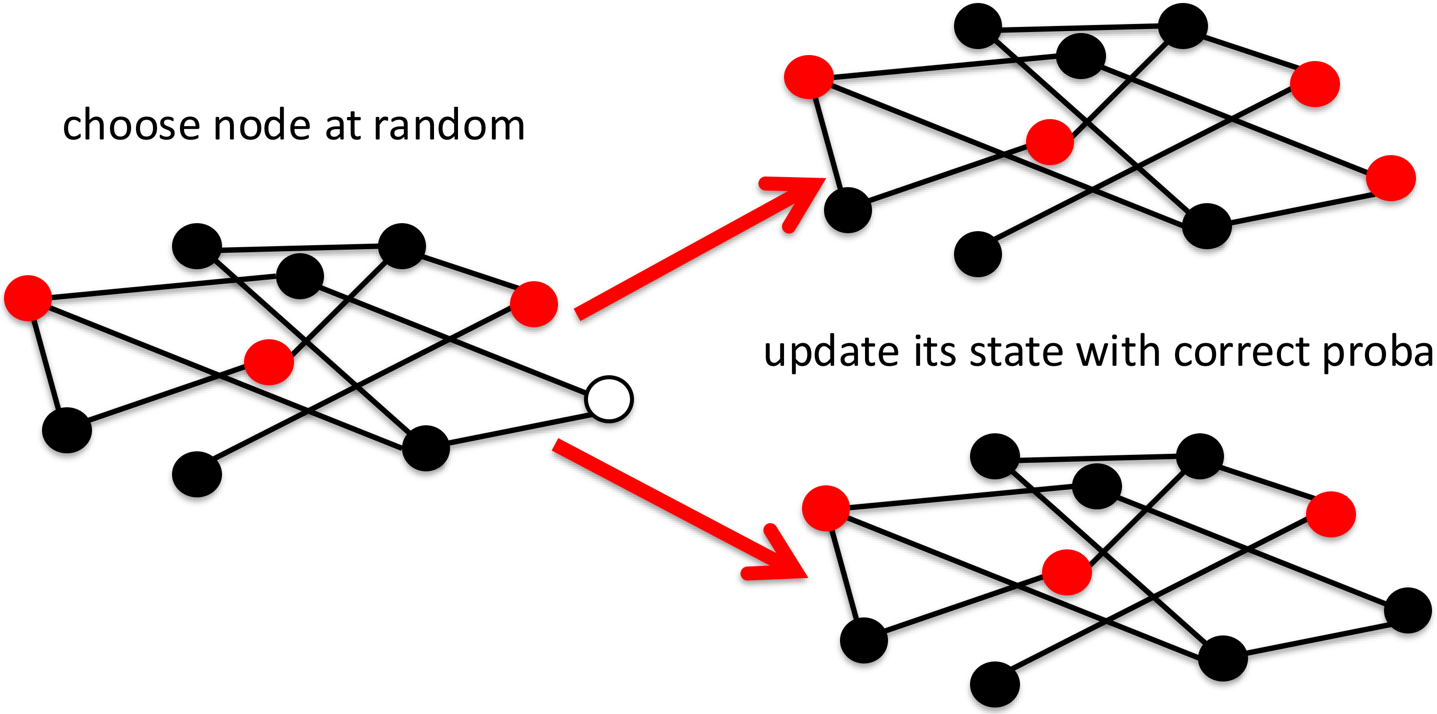


choose node at random

update its state with correct probability

GLAUBER DYNAMICS

Create a Markov chain on the state space with local updates



choose node at random

update its state with correct probability

When does the Markov chain mix rapidly?

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

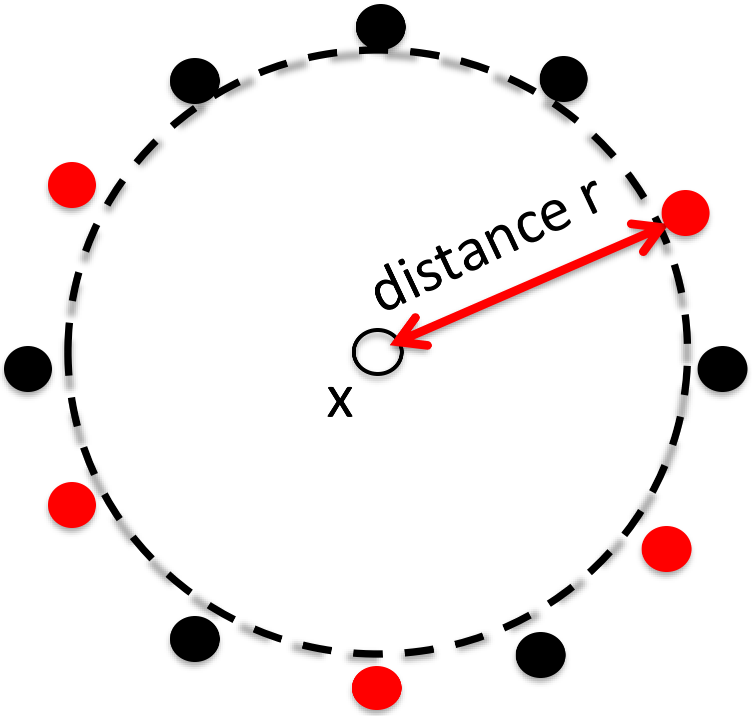
- Glauber Dynamics
- **Phase Transitions and More Lower Bounds**

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

CORRELATION DECAY

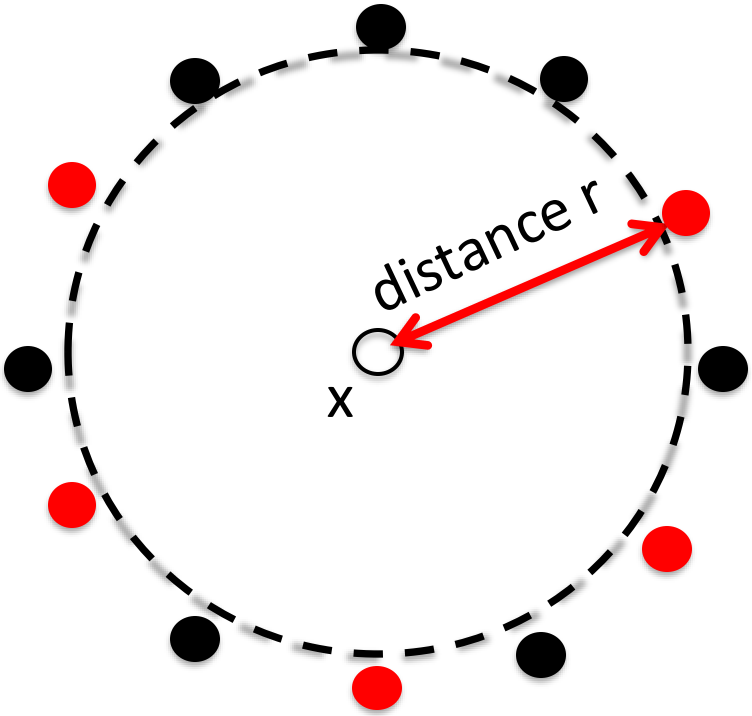
Definition: We say a distribution exhibits **spatial mixing** if



for any boundary at distance r…

CORRELATION DECAY

Definition: We say a distribution exhibits **spatial mixing** if



for any boundary at distance r, effect on marginal distribution at x goes to zero as r increases, i.e. no long-range dependencies

PHASE TRANSITIONS

Definition: A **Gibbs distribution** has the form

$$\mathbb{P}[X = x] = \frac{1}{Z}\exp\Big(- \overset{\textbf{inverse temperature}}{\beta}\, \overset{\textbf{Hamiltonian}}{H}(x)\Big)$$

PHASE TRANSITIONS

Definition: A **Gibbs distribution** has the form

$$\mathbb{P}[X = x] = \frac{1}{Z}\exp\Big(-\overset{\text{inverse temperature}}{\beta}\,\overset{\text{Hamiltonian}}{H}(x)\Big)$$

At lower temperature (larger β) more of the distribution is concentrated on low-energy states

PHASE TRANSITIONS

Definition: A **Gibbs distribution** has the form

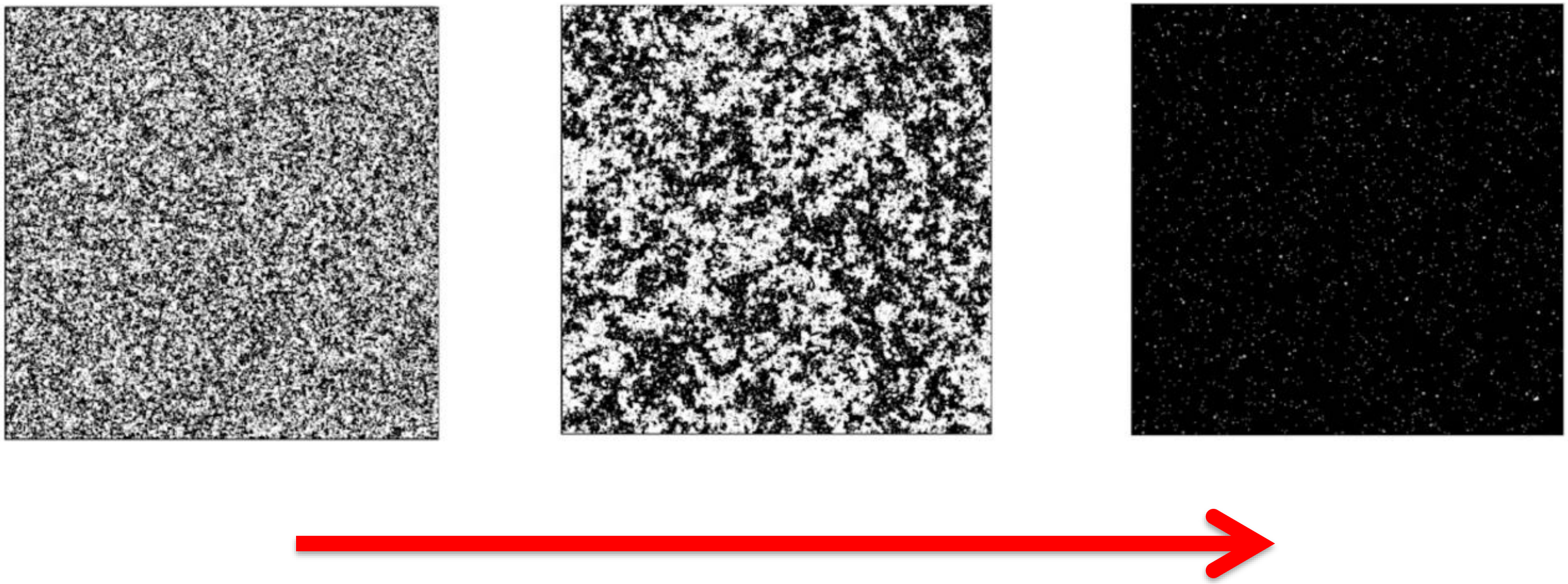
$$\mathbb{P}[X = x] = \frac{1}{Z}\exp\Big(- \overset{\text{inverse temperature}}{\beta}\, \overset{\text{Hamiltonian}}{H}(x)\Big)$$

At lower temperature (larger β) more of the distribution is concentrated on low-energy states

Claim (informal): Usually exhibit spatial mixing iff $\beta \leq \beta_c$

PHASE TRANSITIONS

For a ferromagnet, where neighbors want to agree, in pictures:



Increasing β

COMPLEXITY OF SAMPLING

Theorem (informal): The Glauber dynamics has polynomial mixing time if $\beta < \beta_c$

COMPLEXITY OF SAMPLING

Theorem (informal): The Glauber dynamics has polynomial mixing time if $\beta < \beta_c$

Above the critical temperature Glauber mixes slowly, even worse:

Theorem [Sly '10], [Sly, Sun '12]: In certain classes, sampling from the Gibbs distribution when $\beta > \beta_c$ is NP-hard

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

Often NP-hard

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

Often NP-hard

A quip due to Kamal Jain:

> **“If my laptop can’t find it, then neither can the market”**

SAMPLING ALGORITHMS

Algorithms for learning graphical models assume we are given independent samples

But where do these samples come from?

Often NP-hard

A quip due to Kamal Jain:

"If my laptop can't find it, then ~~neither can the market"~~

it's not a defensible generative model

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

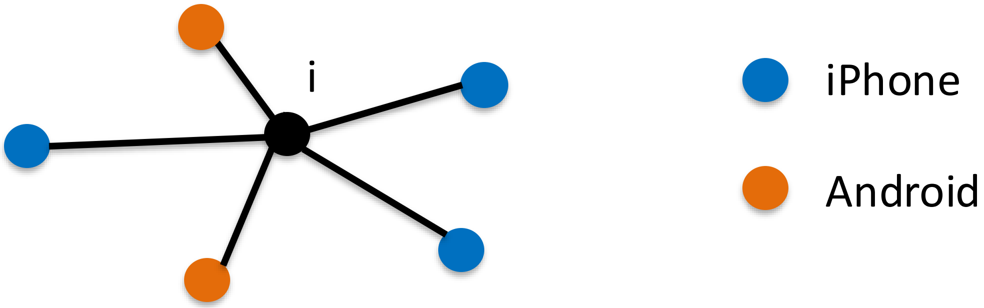
- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

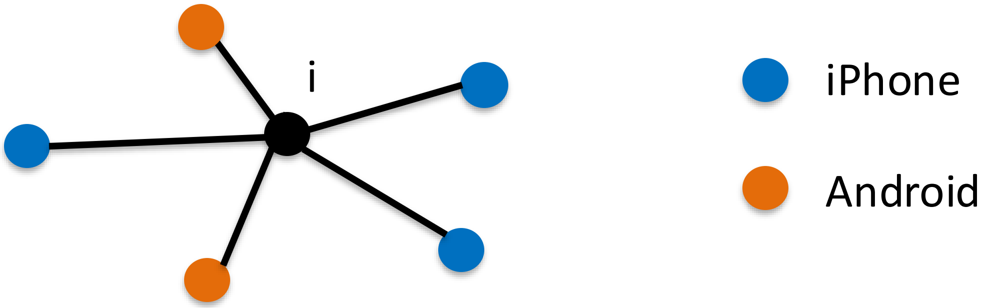
RUNNING EXAMPLE, REVISITED

In fact, for adoption of technology on a social network…



RUNNING EXAMPLE, REVISITED

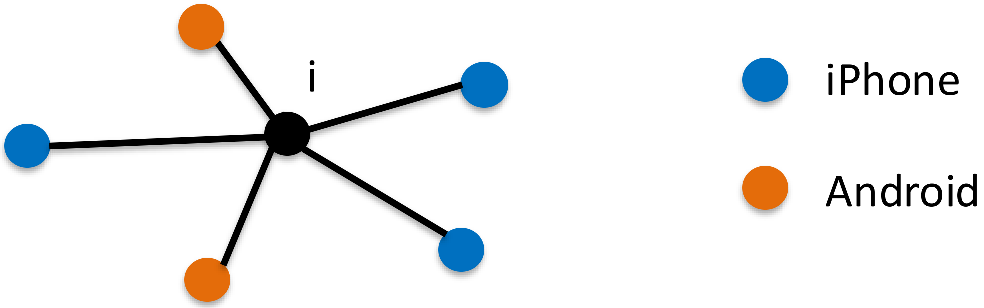
In fact, for adoption of technology on a social network



Many works **[Montanari, Saberi ‘10]** model the time-dependent process as Glauber dynamics

RUNNING EXAMPLE, REVISITED

In fact, for adoption of technology on a social network



Many works **[Montanari, Saberi '10]** model the time-dependent process as Glauber dynamics

Even when Glauber doesn't mix, can we still learn from it?

EVEN MORE HISTORY

Learning from the Glauber dynamics*

EVEN MORE HISTORY

Learning from the Glauber dynamics*

[Bresler, Gamarnik, Shah ‘15]: Polynomial time algorithm for structure learning for Ising models

EVEN MORE HISTORY

Learning from the Glauber dynamics*

[Bresler, Gamarnik, Shah ‘15]: Polynomial time algorithm for structure learning for Ising models

[Gaitonde, Mossel ‘25]: Can do parameter learning too

EVEN MORE HISTORY

Learning from the Glauber dynamics*

[Bresler, Gamarnik, Shah ‘15]: Polynomial time algorithm for structure learning for Ising models

[Gaitonde, Mossel ‘25]: Can do parameter learning too

***Learning from Glauber means seeing not just when nodes switch, but also when they considered switching and didn’t**

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- **An Algorithmic Surprise**
- The Trouble with Higher-Order Interactions

MAIN QUESTION, AND OUR RESULTS

Instead of hoping to get the same results for learning from Glauber as is possible from i.i.d. samples….

MAIN QUESTION, AND OUR RESULTS

Instead of hoping to get the same results for learning from Glauber as is possible from i.i.d. samples….

Is learning from Glauber dynamics *easier*?

MAIN QUESTION, AND OUR RESULTS

Instead of hoping to get the same results for learning from Glauber as is possible from i.i.d. samples….

Is learning from Glauber dynamics *easier*?

Theorem [Gaitonde, Moitra, Mossel]: There is an algorithm that learns Markov random fields of order k in $O(n^2 \log n)$ time and from $O(n \log n)$ length trajectories of Glauber dynamics

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- The Trouble with Higher-Order Interactions

OUTLINE

Part I: Learning Graphical Models from Samples

- Ising Models and Their Applications
- Markov Random Fields and Lower Bounds

Part II: Sampling from Graphical Models

- Glauber Dynamics
- Phase Transitions and More Lower Bounds

Part III: Learning from Trajectories

- An Algorithmic Surprise
- **The Trouble with Higher-Order Interactions**

First let’s review **[Bresler, Gamarnik, Shah]**…

PROOF SKETCH (ISING CASE)

How can we decide if (X_u, X_v) is an edge?

PROOF SKETCH (ISING CASE)

How can we decide if (X_u, X_v) is an edge?

Definition: A **good sequence** is a contiguous sequence where

(1) X_u attempts to update (might or might not flip)

(2) Then X_v attempts to update, and flips

(3) Then X_u attempts to update (might or might not flip)

PROOF SKETCH (ISING CASE)

How can we decide if (X_u, X_v) is an edge?

Definition: A **good sequence** is a contiguous sequence where

(1) X_u attempts to update (might or might not flip)

(2) Then X_v attempts to update, and flips

(3) Then X_u attempts to update (might or might not flip)

Intuition: We get one sample from the conditional distribution on X_u when $X_v = +1$ and one sample when $X_v = -1$ but everything else stays the same

PROOF SKETCH (ISING CASE)

Good sequences give us empirical estimates of the quantity

$$p - p'$$

probability X_u = +1 after update, if X_v = +1

same, but with X_v = -1

PROOF SKETCH (ISING CASE)

Good sequences give us empirical estimates of the quantity

$$p - p'$$

probability $X_u = +1$ after update, if $X_v = +1$

same, but with $X_v = -1$

The probabilities p and p' depend on the rest of the configuration, but have a consistent sign depending on $J_{u,v}$

PROOF SKETCH (ISING CASE)

Good sequences give us empirical estimates of the quantity

$$p - p'$$

probability $X_u = +1$ after update, if $X_v = +1$

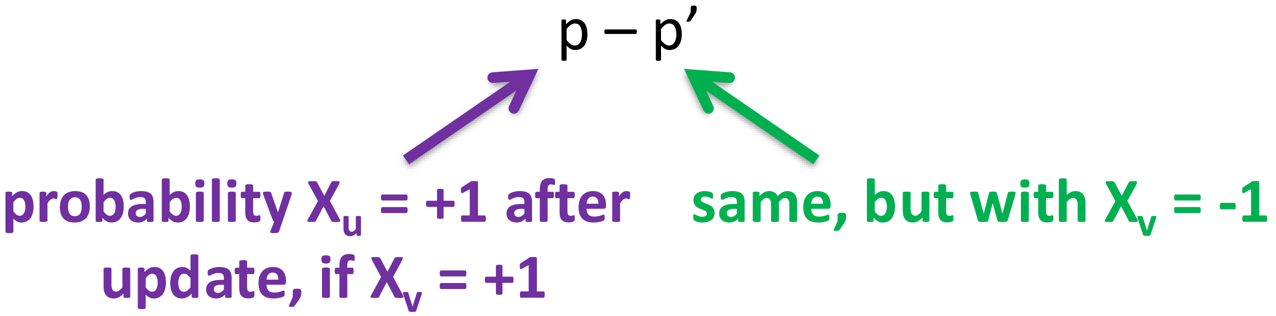
same, but with $X_v = -1$

The probabilities p and p' depend on the rest of the configuration, but have a consistent sign depending on $J_{u,v}$

Claim: With high probability, edges correspond to pairs where the statistic averaged over good sequences is far enough from zero

PROOF SKETCH (ISING CASE)

Good sequences give us empirical estimates of the quantity



$$p - p'$$

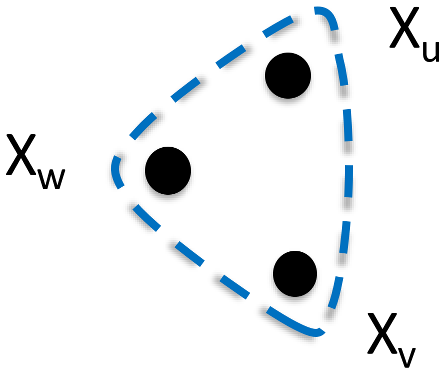
probability X_u = +1 after update, if X_v = +1

same, but with X_v = -1

⚠ **Caution:** There is a minor bug because conditioning on X_v flipping changes things, can be fixed

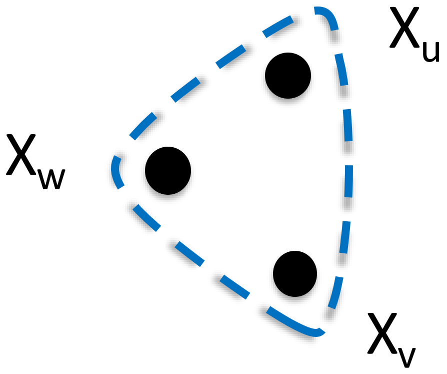
A STUMBLING BLOCK

Why does this strategy not work with higher-order interactions?



A STUMBLING BLOCK

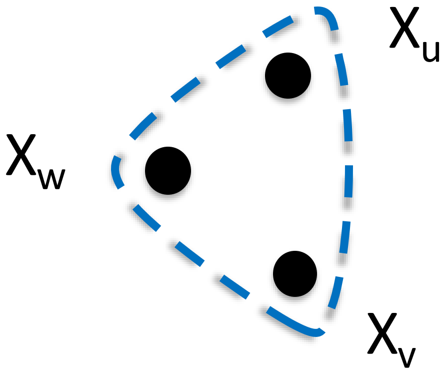
Why does this strategy not work with higher-order interactions?



Sometimes X_u and X_v want to agree, sometimes they want to disagree, depending on the sign of X_w

A STUMBLING BLOCK

Why does this strategy not work with higher-order interactions?



Sometimes X_u and X_v want to agree, sometimes they want to disagree, depending on the sign of X_w

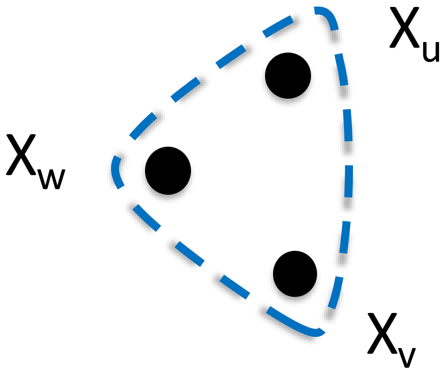
But now the quantity p – p'…

probability $X_u = +1$ after update, if $X_v = +1$

same, but with $X_v = -1$

A STUMBLING BLOCK

Why does this strategy not work with higher-order interactions?



Sometimes X_u and X_v want to agree, sometimes they want to disagree, depending on the sign of X_w

But now the quantity $p - p'$

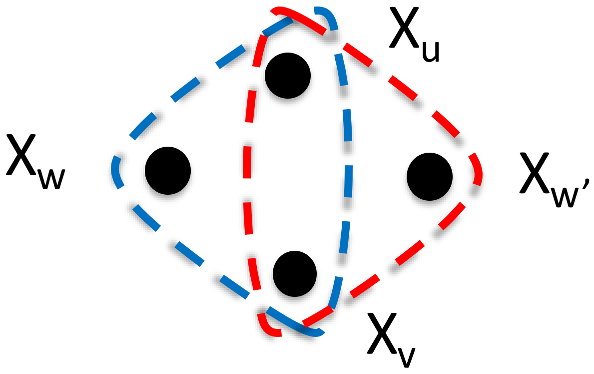
probability X_u = +1 after update, if X_v = +1

same, but with X_v = -1

doesn't necessarily have a consistent sign

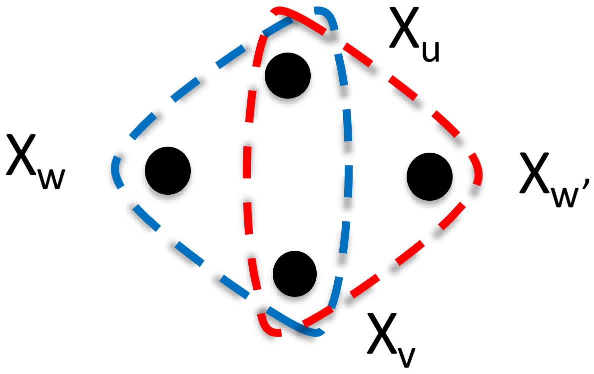
A STUMBLING BLOCK

Could also have situations like



A STUMBLING BLOCK

Could also have situations like



Where X_v affects the distribution of X_u through two paths, but they cancel out

OUTLINE

Bonus Part: Population of Means

- Taming Higher-Order Interactions

OUTLINE

Bonus Part: Population of Means

- Taming Higher-Order Interactions

POPULATION OF MEANS

Classic problem in statistics:

Setup: Suppose $\theta \sim \mu$ and $X_1, \cdots, X_k \sim \mathrm{Bern}(\theta)$

random student

their score on k exam questions

POPULATION OF MEANS

Classic problem in statistics:

Setup: Suppose $\theta \sim \mu$ and $X_1, \cdots, X_k \sim \mathrm{Bern}(\theta)$

random student

their score on k exam questions

Are students just randomly guessing?

POPULATION OF MEANS

Classic problem in statistics:

Setup: Suppose $\theta \sim \mu$ and $X_1, \cdots, X_k \sim \mathrm{Bern}(\theta)$

random student

their score on k exam questions

Are students just randomly guessing?

There's no hope of answering this question if k =1, but…

POPULATION OF MEANS

Classic problem in statistics:

Setup: Suppose $\theta \sim \mu$ and $X_1, \cdots, X_k \sim \mathrm{Bern}(\theta)$

random student

their score on k exam questions

Fact (shrinkage): We can’t estimate the θ’s, but can estimate the first k moments of μ from batches

POPULATION OF MEANS

Classic problem in statistics:

Setup: Suppose $\theta \sim \mu$ and $X_1, \cdots, X_k \sim \mathrm{Bern}(\theta)$

random student

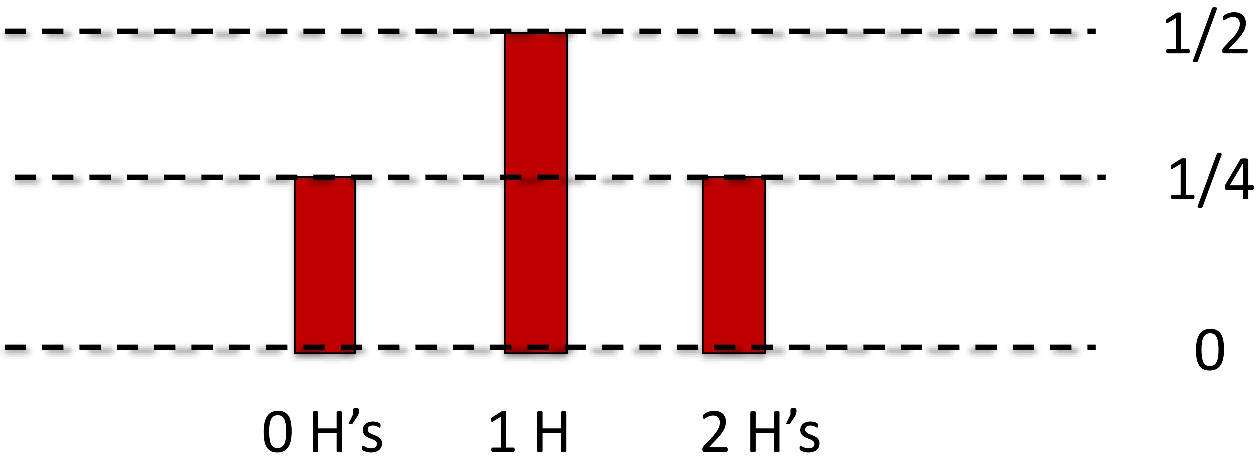
their score on k exam questions

Fact (shrinkage): We can’t estimate the θ’s, but can estimate the first k moments of μ from batches

$\Longrightarrow$ **Can decide if μ is a point mass even when k = 2**

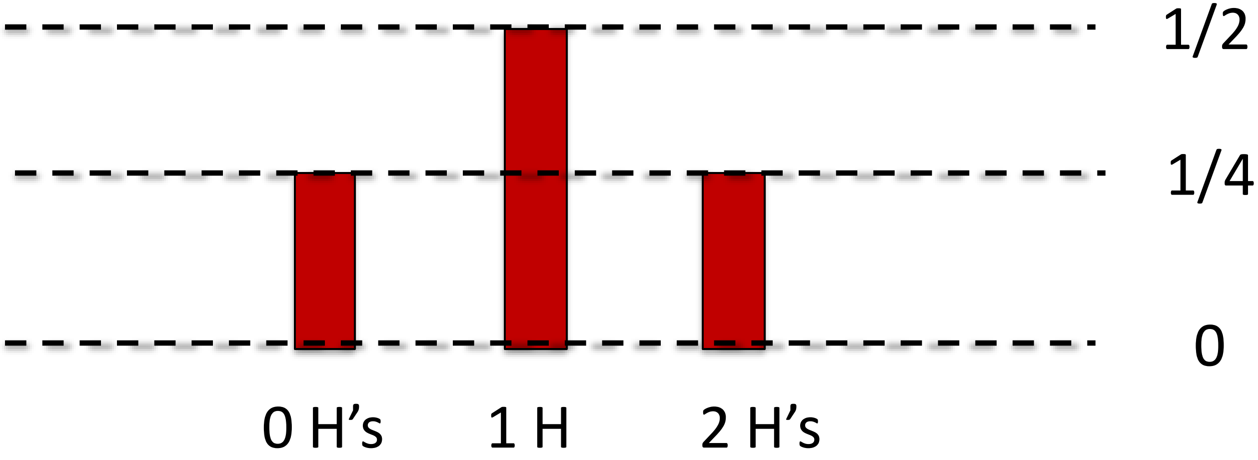
POPULATION OF MEANS

When k = 2 we can’t estimate each student’s θ well, but the point is the only way we can get



POPULATION OF MEANS

When k = 2 we can't estimate each student's θ well, but the point is the only way we can get



is if μ is a point mass on 1/2, i.e. each student is randomly guessing

OUTLINE

Bonus Part: Population of Means

- Taming Higher-Order Interactions

OUTLINE

Bonus Part: Population of Means

- **Taming Higher-Order Interactions**

KEY IDEA

Can we instead estimate the following quantity?

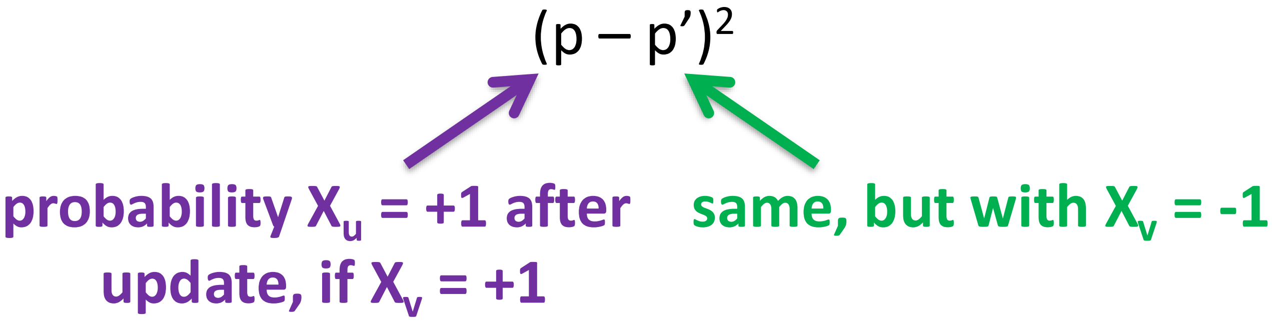
$$(p - p')^2$$

probability X_u = +1 after update, if X_v = +1

same, but with X_v = -1

KEY IDEA

Can we instead estimate the following quantity?



$$(p - p')^2$$

probability $X_u = +1$ after update, if $X_v = +1$

same, but with $X_v = -1$

All we need is a more complicated definition of a good sequence

KEY IDEA

Definition': A **better sequence** is a contiguous sequence where

1. (1) X_u attempts to update **twice** (might or might not flip)
2. (2) Then X_v attempts to update, and flips
3. (3) Then X_u attempts to update **twice** (might or might not flip)

KEY IDEA

Definition': A **better sequence** is a contiguous sequence where

1. X_u attempts to update $\overset{\textbf{twice}}{\wedge}$ (might or might not flip)
2. Then X_v attempts to update, and flips
3. Then X_u attempts to update $\overset{\textbf{twice}}{\wedge}$ (might or might not flip)

Intuition: We get **two** samples from the conditional distribution on X_u when X_v = +1 and **two** samples when X_v = -1 but everything else stays the same

KEY IDEA

Definition': A **better sequence** is a contiguous sequence where

(1) X_u attempts to update $\overset{\textbf{twice}}{\wedge}$ (might or might not flip)

(2) Then X_v attempts to update, and flips

(3) Then X_u attempts to update $\overset{\textbf{twice}}{\wedge}$ (might or might not flip)

Let $Y_1 Y_2$ ___ $Y'_1 Y'_2$ be the indicators that $X_u = +1$ at corresponding locations

KEY IDEA

Definition': A **better sequence** is a contiguous sequence where

(1) X_u attempts to update twice (might or might not flip)

(2) Then X_v attempts to update, and flips

(3) Then X_u attempts to update twice (might or might not flip)

Let $Y_1 Y_2$ ___ $Y'_1 Y'_2$ be the indicators that $X_u = +1$ at corresponding locations

Now we can estimate $(p-p')^2$ using the statistic

$$Z = Y_1 Y_2 - 2Y_1 Y'_1 + Y'_1 Y'_2$$

EPILOGUE

This strategy would need very long trajectories…

EPILOGUE

This strategy would need very long trajectories…

Idea (informal): Even if the attempted updates to X_u and X_v are not contiguous, if they're close enough we don't expect any updates to their neighbors in between

EPILOGUE

This strategy would need very long trajectories…

Idea (informal): Even if the attempted updates to X_u and X_v are not contiguous, if they're close enough we don't expect any updates to their neighbors in between

Also can learn from just the updates:

Theorem [Gaitonde, Moitra, Mossel]: Can learn Ising models from just the timestamps of updates in continuous time

Summary:

- Learning Markov Random Fields from samples is computationally hard
- But you can’t generate samples anyways and given the trajectories there are much better algorithms
- **What other learning problems become easier from dynamics?**

Summary:

- Learning Markov Random Fields from samples is computationally hard
- But you can’t generate samples anyways and given the trajectories there are much better algorithms
- **What other learning problems become easier from dynamics?**

Thanks! Any Questions?