Ensuring Rapid Mixing and Low Bias for Asynchronous Gibbs Sampling

Christopher De Sa Kunle Olukotun Christopher Ré

{cdesa,kunle,chrismre}@stanford.edu

Stanford

Overview

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

DeepDive YAHOO! …etc.

Zhang et al, *PVLDB* 2014

Smola et al, *PVLDB* 2010

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

"Folklore" says that asynchronous Gibbs sampling basically works whenever standard (sequential) Gibbs sampling does

…but there's **no theoretical guarantee**.

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

"Folklore" says that asynchronous Gibbs sampling basically works whenever standard (sequential) Gibbs sampling does

…but there's **no theoretical guarantee**.

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

"Folklore" says that asynchronous Gibbs sampling basically works whenever standard (sequential) Gibbs sampling does

…but there's **no theoretical guarantee**.

Our contributions

1. The **"folklore"** is not necesarily true.
2. ...but it works under **reasonable conditions**.

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

~~**“Folklore”** says that asynchronous Gibbs sampling basic[illegible] standard (sequential) Gibbs sampling does~~

…but there's **no theoretical guarantee**.

Our contributions

1. The **“folklore”** is not necesarily true.
2. ...but it works under **reasonable conditions**.

Asynchronous Gibbs sampling is a popular algorithm that's used in practical ML systems.

Question: **when and why** does it work?

~~**“Folklore”** says that asynchronous Gibbs sampling basic[illegible] standard (sequential) Gibbs sampling does~~

…but there's **no theoretical guarantee**.

Our contributions

1. The **“folklore”** is not necesarily true.
2. ...but it works under **reasonable conditions**.

Problem: given a **probability distribution**, produce **samples** from it.

- e.g. to do **inference** in a graphical model

Problem: given a **probability distribution**, produce **samples** from it.

- e.g. to do **inference** in a graphical model

Algorithm: **Gibbs sampling**

- de facto Markov chain Monte Carlo (**MCMC**) method for inference
- produces a series of **approximate** samples that **approach** the target distribution

What is Gibbs Sampling?

What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \leq i \leq n$, and distribution π.

loop

Choose s by sampling uniformly from $\{1, \ldots, n\}$.

Re-sample x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\}\setminus\{s\}})$.

output x

end loop

What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \leq i \leq n$, and distribution π.

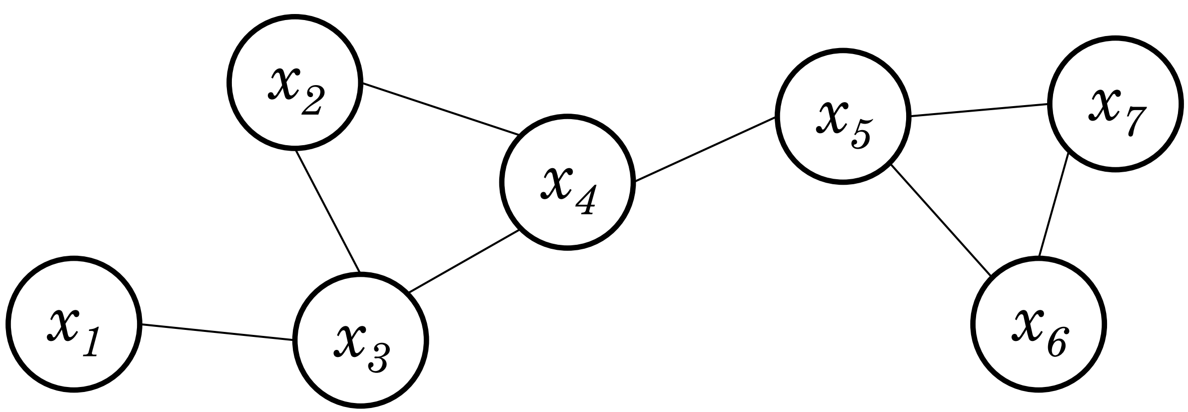
loop

Choose s by sampling uniformly from $\{1, \ldots, n\}$.

Re-sample x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\} \setminus \{s\}})$.

output x

end loop



What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Choose a variable to update at random.

Require: Variables x_i for $1 \leq$ … ution π.

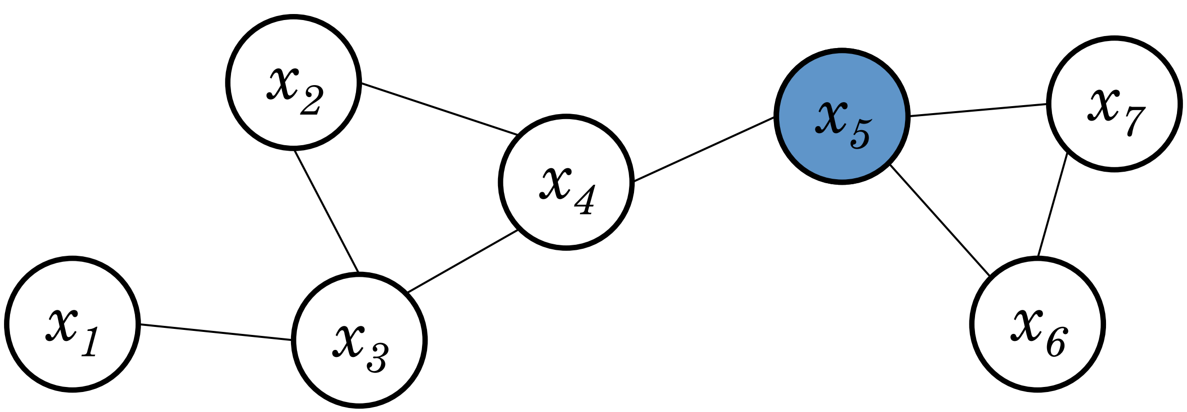
  loop

    Choose s by sampling uniformly from $\{1, \ldots, n\}$.

    Re-sample x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\} \setminus \{s\}})$.

    output x

  end loop



What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \leq i$ ... π.

loop

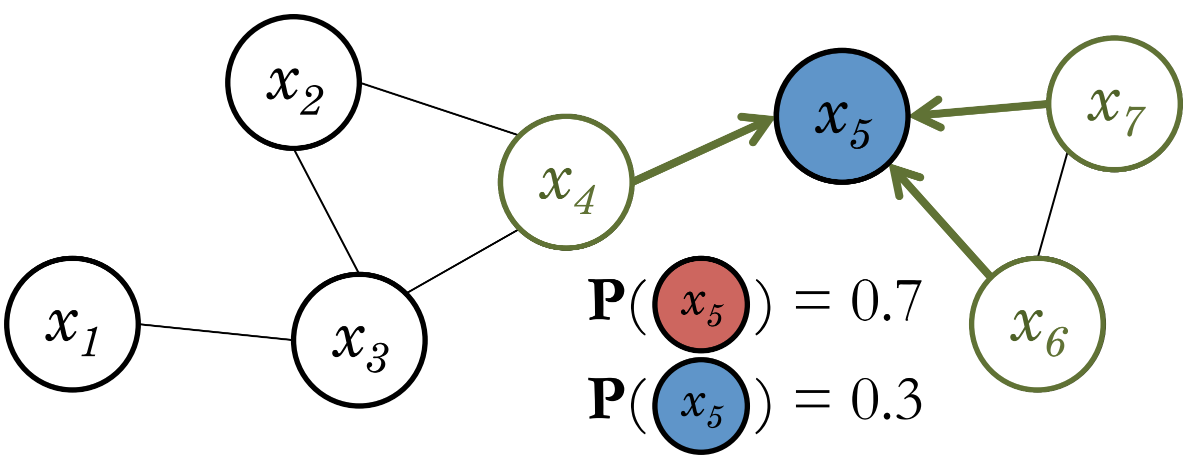
Choose s by sampling uniformly ...n $\{1, \ldots, n\}$.

Re-sample x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\} \setminus \{s\}})$.

output x

end loop

Compute its conditional distribution given the other variables.



What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \leq i$ ... π.

loop

Choose s by sampling uniformly ... n $\{1, \ldots, n\}$.

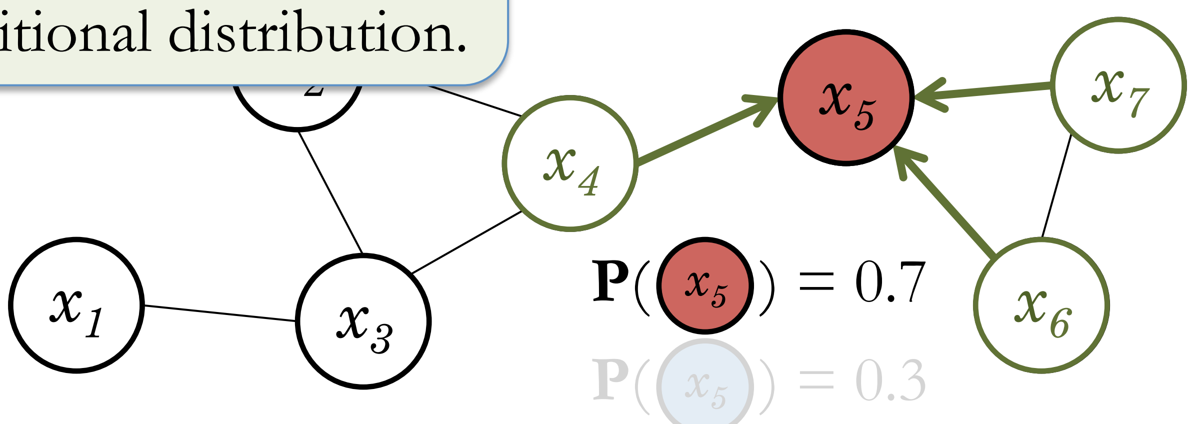
Re-sample x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\}\setminus\{s\}})$.

out... x

e...

Compute its conditional distribution given the other variables.

Update the variable by sampling from its conditional distribution.



What is Gibbs Sampling?

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \le i \le n$, and distribution π.

l

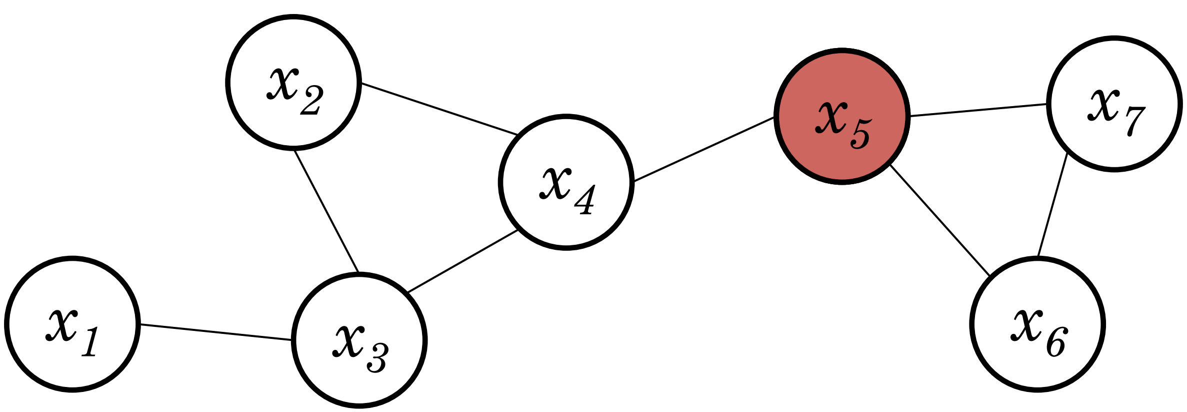
Output the current state as a sample.

…pling uniformly from $\{1, \ldots, n\}$.

Re…mple x_s uniformly from $\mathbf{P}_\pi(x_s | x_{\{1,\ldots,n\}\setminus\{s\}})$.

output x

end loop



Gibbs Sampling: A Practical Perspective

Gibbs Sampling: A Practical Perspective

- **Pros** of Gibbs sampling
  - Easy to implement
  - Updates are sparse → **fast on modern CPUs**

- **Cons** of Gibbs sampling
  - sequential algorithm → **can't naively parallelize**

Gibbs Sampling: A Practical Perspective

- **Pros** of Gibbs sampling
  - Easy to implement
  - Updates are sparse → **fast on modern CPUs**
- **Cons** of Gibbs sampling
  - sequential algorithm → **can't naively parallelize**

e.g.

64 core

No parallelism → Leave up to 98% of performance on the table!

Asynchronous Gibbs Sampling

Asynchronous Gibbs Sampling

- Run multiple threads in parallel **without locks**
  - also known as **HOGWILD!**
  - adapted from a popular technique for stochastic gradient descent (SGD)
- When we read a variable, it could be **stale**
  - while we re-sample a variable, its adjacent variables can be overwritten by other threads
  - semantics **not equivalent** to standard (sequential) Gibbs sampling

Asynchronous Gibbs Sampling

- Run multiple threads in parallel **without locks**
  - also known as **HOGWILD!**
  - adapted from a popular technique for stochastic gradient descent (SGD)
- When we read a variable, it could be **stale**
  - while we re-sample a variable, its adjacent variables can be overwritten by other threads
  - semantics **not equivalent** to standard (sequential) Gibbs sampling

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Two desiderata

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Two desiderata

want to get

accurate estimates

↓

bound the

bias

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Two desiderata

want to get

accurate estimates

↓

bound the

bias

want to be independent of initial conditions

quickly

↓

bound the

mixing time

Previous Work

Previous Work

- **"Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent"** — Niu et al, NIPS 2011.

  follow-up work: Liu and Wright SCIOPS 2015, Liu et al JMLR 2015, De Sa et al NIPS 2015, Mania et al arxiv 2015

- **"Analyzing Hogwild Parallel Gaussian Gibbs Sampling"** — Johnson et al, NIPS 2013.

Previous Work

- **"Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent"** — Niu et al, NIPS 2011.

  follow-up work: Liu and Wright SCIOPS 2015, Liu et al JMLR 2015, De Sa et al NIPS 2015, Mania et al arxiv 2015

- **"Analyzing Hogwild Parallel Gaussian Gibbs Sampling"** — Johnson et al, NIPS 2013.

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Two desiderata

want to get
accurate estimates
↓
bound the
bias

want to be independent
of initial conditions
quickly
↓
bound the
mixing time

Bias

Bias

- **How close** are samples to target distribution?
  - standard measurement: **total variation distance**

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|$$

- For sequential Gibbs, **no asymptotic bias**:

Bias

- **How close** are samples to target distribution?
  - standard measurement: **total variation distance**

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|$$

- For sequential Gibbs, **no asymptotic bias**:

$$\forall \mu_0,\ \lim_{t \to \infty} \|P^{(t)} \mu_0 - \pi\|_{\mathrm{TV}} = 0$$

Bias

- **How close** are samples to target distribution?
  - standard measurement: **total variation distance**

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|$$

- For sequential Gibbs, **no asymptotic bias**:

$$\forall \mu_0, \ \lim_{t \to \infty} \|P^{(t)} \mu_0 - \pi\|_{\mathrm{TV}} = 0$$

"Folklore": asynchronous Gibbs is also unbiased.

…but this is **not necessarily true**!

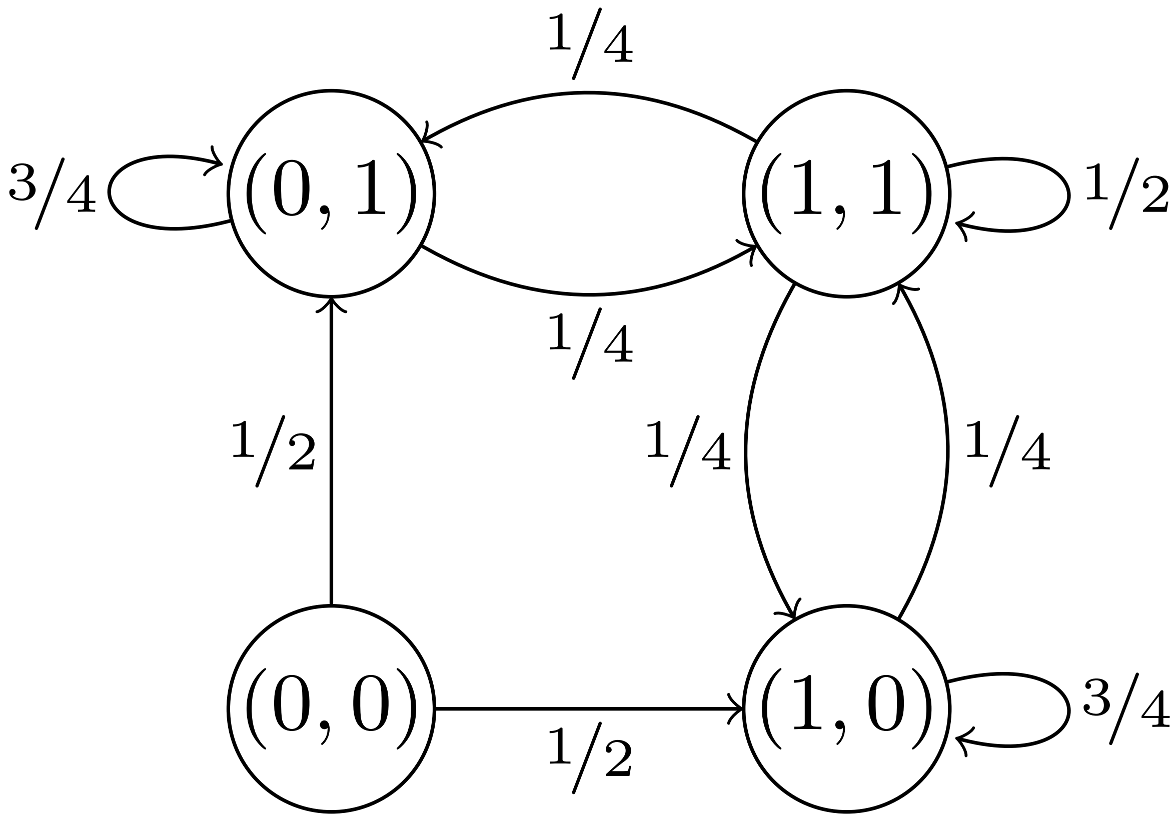
Simple Bias Example

Simple Bias Example

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$

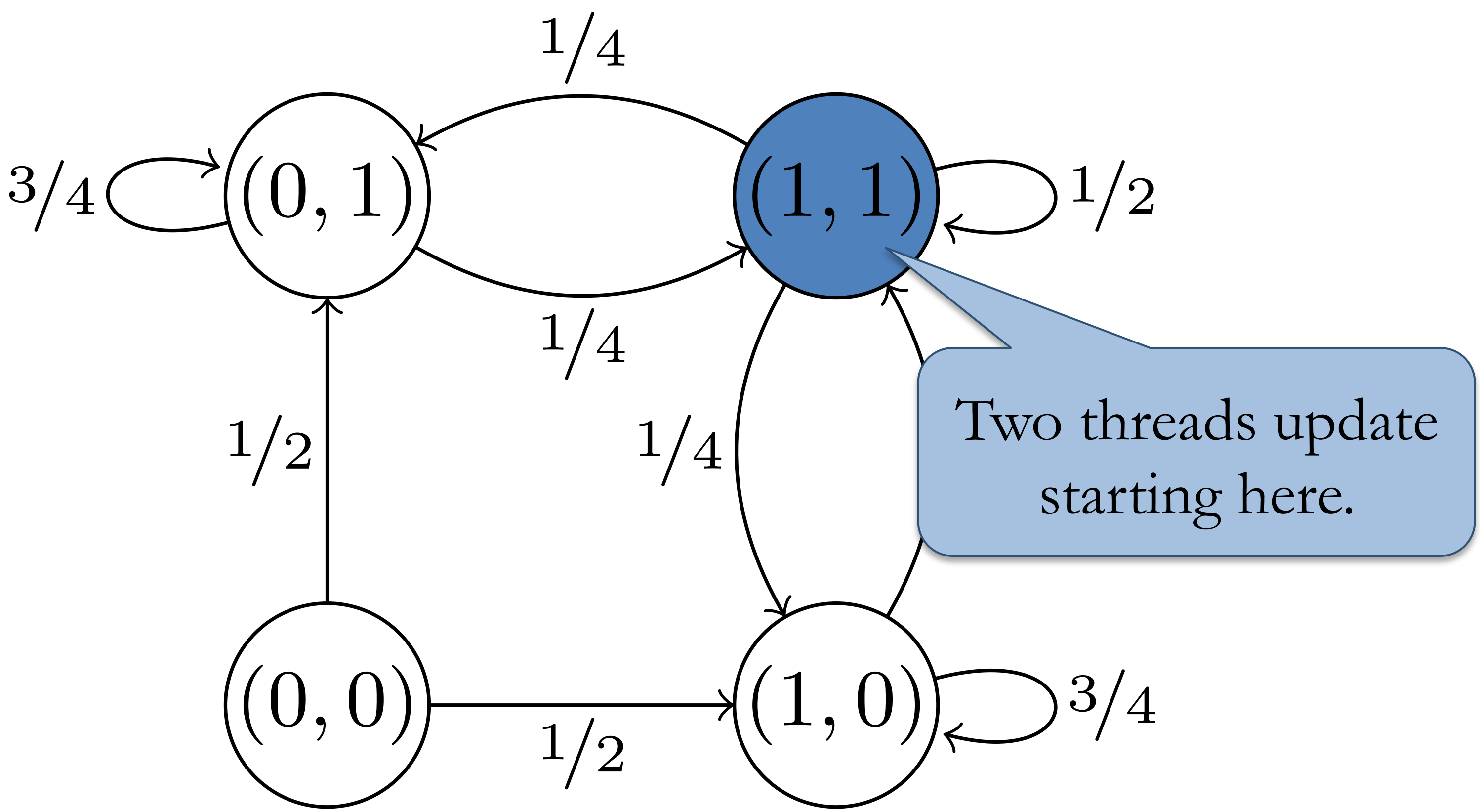
Simple Bias Example

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$



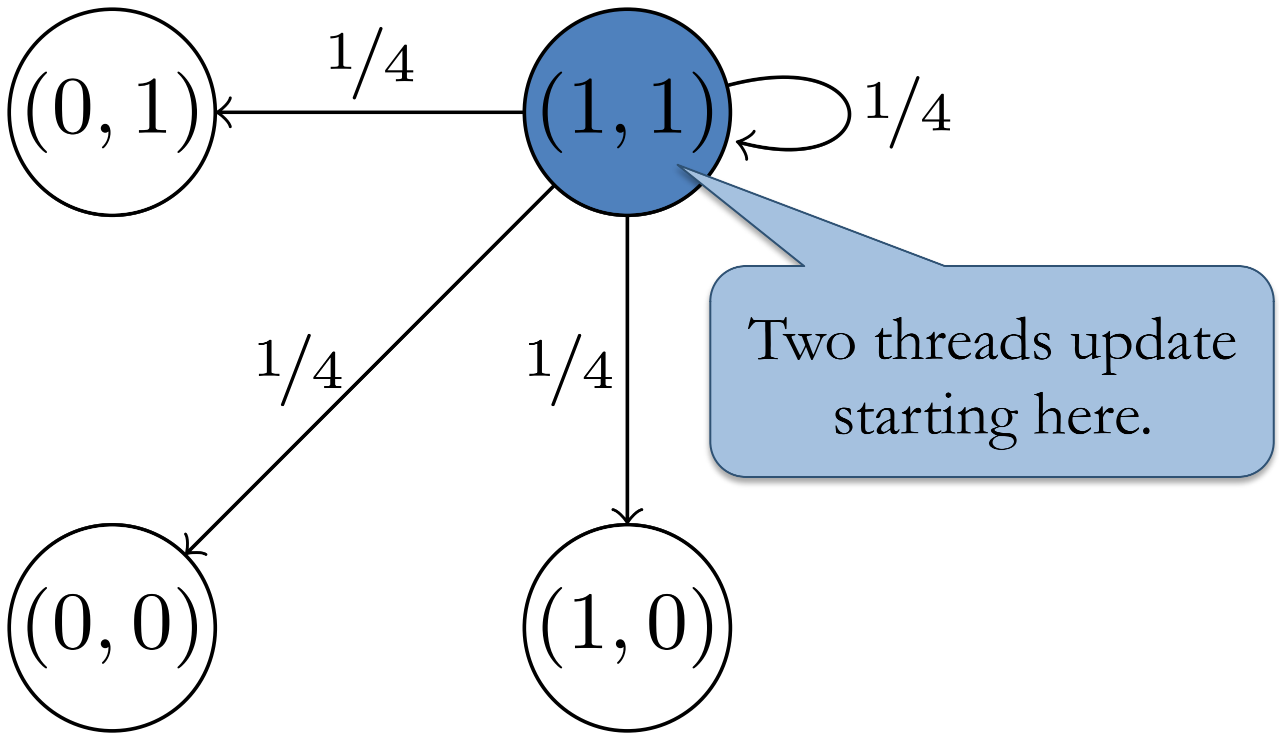
Simple Bias Example

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$



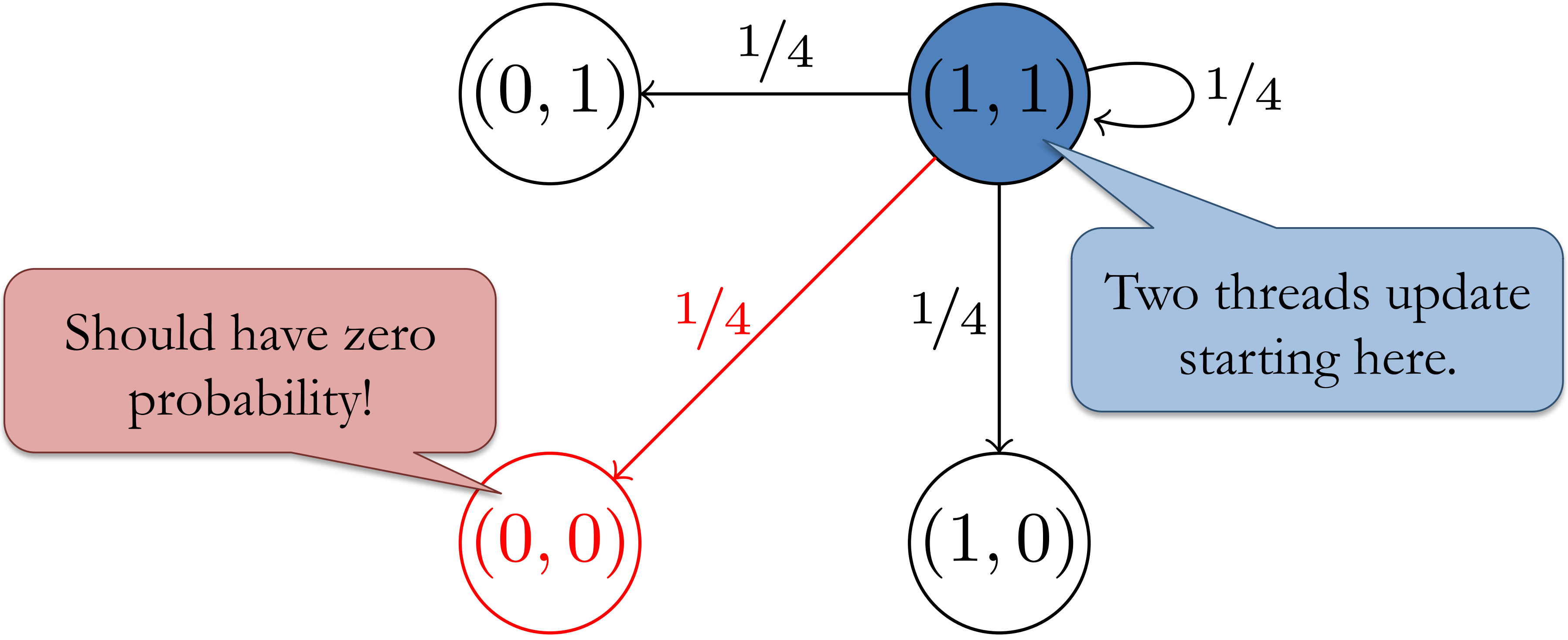
Simple Bias Example

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$

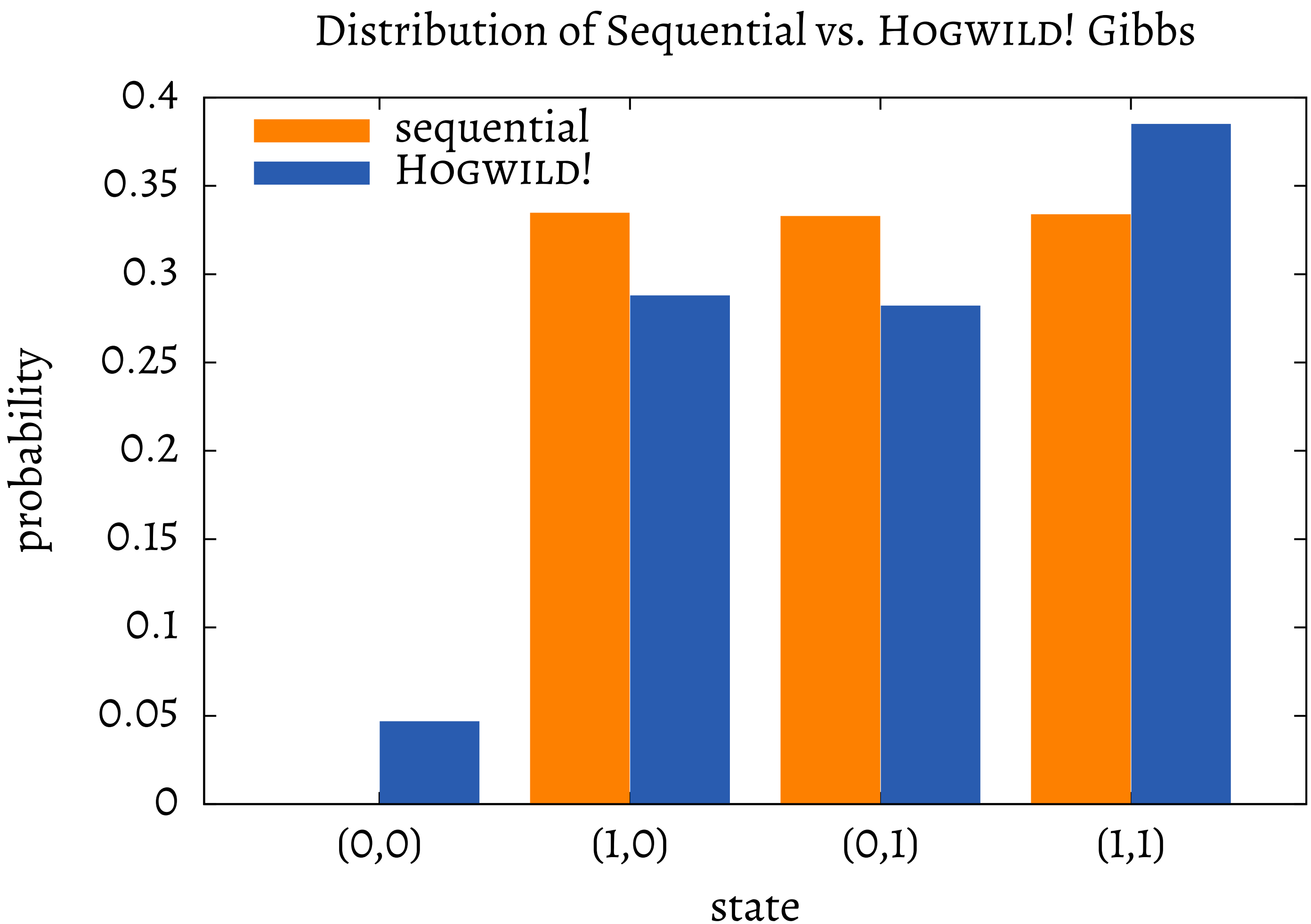


Simple Bias Example

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$



Nonzero Asymptotic Bias



Measured

Bias

(total variation distance)

sequential

< 0.1%

unbiased

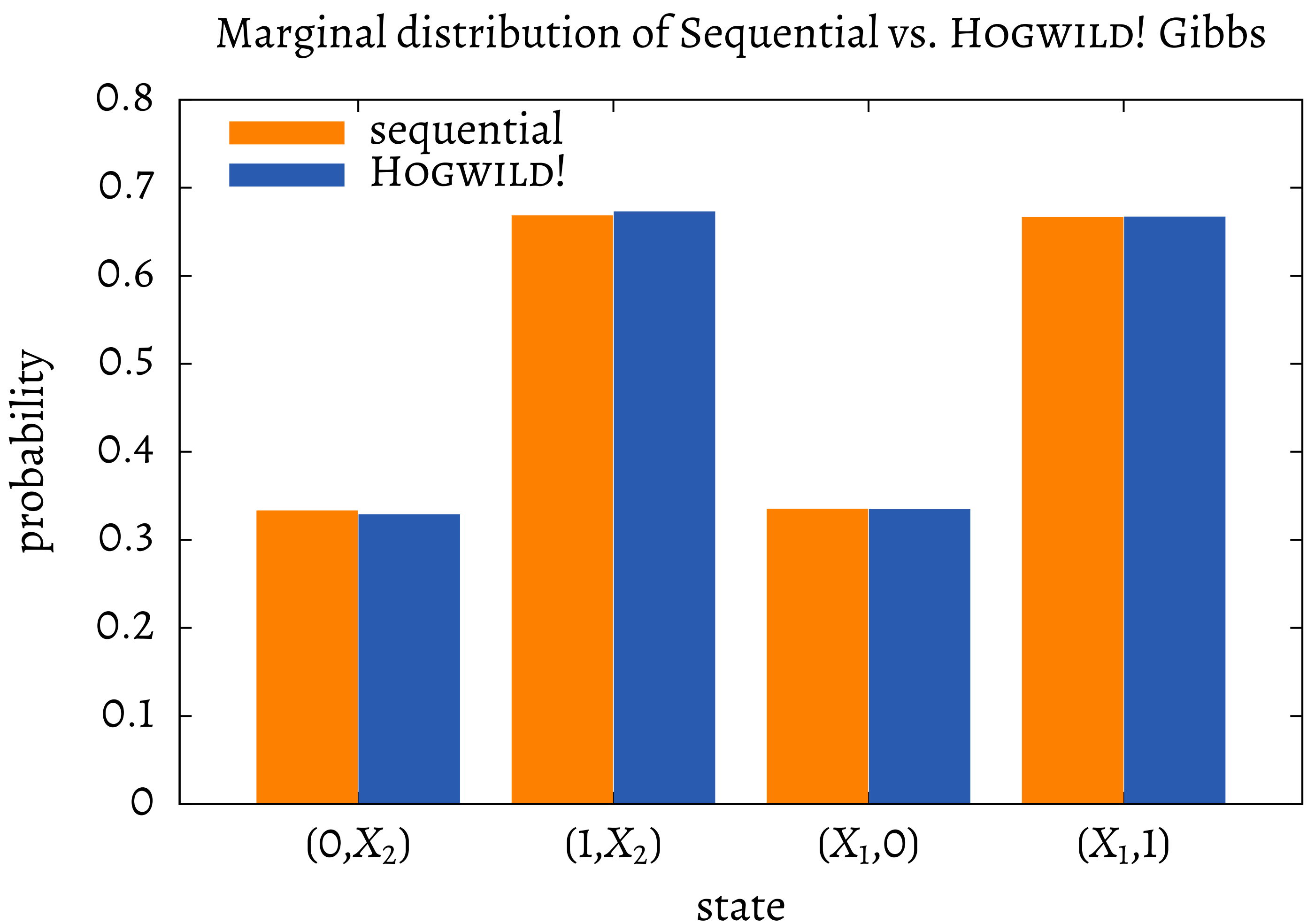
asynchronous

9.8%

biased

Bias introduced by HOGWILD!-Gibbs (10^6 samples).

Nonzero Asymptotic Bias



Measured

Bias

(total variation distance)

sequential

< 0.1%

unbiased

asynchronous

9.8%

biased

Bias introduced by HOGWILD!-Gibbs (10^6 samples).

Are we using the right metric?

Are we using the right metric?

- No, total variation distance is **too conservative**
  - depends on events that don't matter for inference
  - usually only care about small number of variables

- New metric: **sparse variation distance**

where $|A|$ is the number of variables on which event A depends

Are we using the right metric?

- No, total variation distance is **too conservative**
  - depends on events that don't matter for inference
  - usually only care about small number of variables

- New metric: **sparse variation distance**

$$\|\mu - \nu\|_{\mathrm{SV}(\omega)} = \max_{|A| \leq \omega} |\mu(A) - \nu(A)|$$

where |A| is the number of variables on which event A depends

Are we using the right metric?

- No, total variation distance is **too conservative**
  - depends on events that don't matter for inference
  - usually only care about small number of variables
- New metric: **sparse variation distance**

$$\|\mu - \nu\|_{\mathrm{SV}(\omega)} = \max_{|A| \le \omega} |\mu(A) - \nu(A)|$$

where $|A|$ is the number of variables on which event A depends

Simple Example: Bias of Asynchronous Gibbs

Total variation: 9.8% **Sparse Variation ($\omega = 1$): 0.4%**

Total Influence Parameter

Total Influence Parameter

- Old condition that was used to study **mixing times** of spin statistics systems

$$\alpha = \max_{i \in I} \sum_{j \in I} \max_{(X,Y) \in B_j} \left\| \pi_i(\cdot | X_{I \setminus \{i\}}) - \pi_i(\cdot | Y_{I \setminus \{i\}}) \right\|_{\mathrm{TV}}$$

  – $(X, Y) \in B_j$ means X and Y equal except variable j.

  – $\pi_i(\cdot | X_{I \setminus \{i\}})$ is conditional distribution of variable i given the values of all the other variables in state X.

  – **Dobrushin's condition** holds when

Total Influence Parameter

- Old condition that was used to study **mixing times** of spin statistics systems

$$\alpha = \max_{i \in I} \sum_{j \in I} \max_{(X,Y) \in B_j} \left\| \pi_i(\cdot | X_{I \setminus \{i\}}) - \pi_i(\cdot | Y_{I \setminus \{i\}}) \right\|_{\mathrm{TV}}$$

  - $(X, Y) \in B_j$ means X and Y equal except variable j.
  - $\pi_i(\cdot | X_{I \setminus \{i\}})$ is conditional distribution of variable i given the values of all the other variables in state X.
  - **Dobrushin's condition** holds when $\alpha < 1$.

Asymptotic Result

- For any class of distributions with **bounded total influence** $\alpha = O(1)$.
  - big-O notation is over number of variables n.
- If $O(n)$ timesteps of sequential Gibbs suffice to achieve arbitrarily small bias
  - measured by ω-sparse variation distance, for fixed ω
- …then asynchronous Gibbs **requires only $O(1)$ additional timesteps** to achieve **the same bias**!

Asymptotic Result

- For any class of distributions with **bounded total influence** $\alpha = O(1)$.
  - big-O notation is over number of variables n.
- If $O(n)$ timesteps of sequential Gibbs suffice to achieve arbitrarily small bias
  - measured by ω-sparse variation distance, for fixed ω
- …then asynchronous Gibbs **requires only** $O(1)$ **additional timesteps** to achieve **the same bias**!

more details, explicit bounds, et cetera in the paper

Question

Does **asynchronous Gibbs sampling** work?

…and **what does it mean** for it to work?

Two desiderata

want to get

accurate estimates

↓

bound the

bias

want to be independent of initial conditions

quickly

↓

bound the

mixing time

Mixing Time

Mixing Time

- **How long** do we need to run until the samples are **independent of initial conditions**?

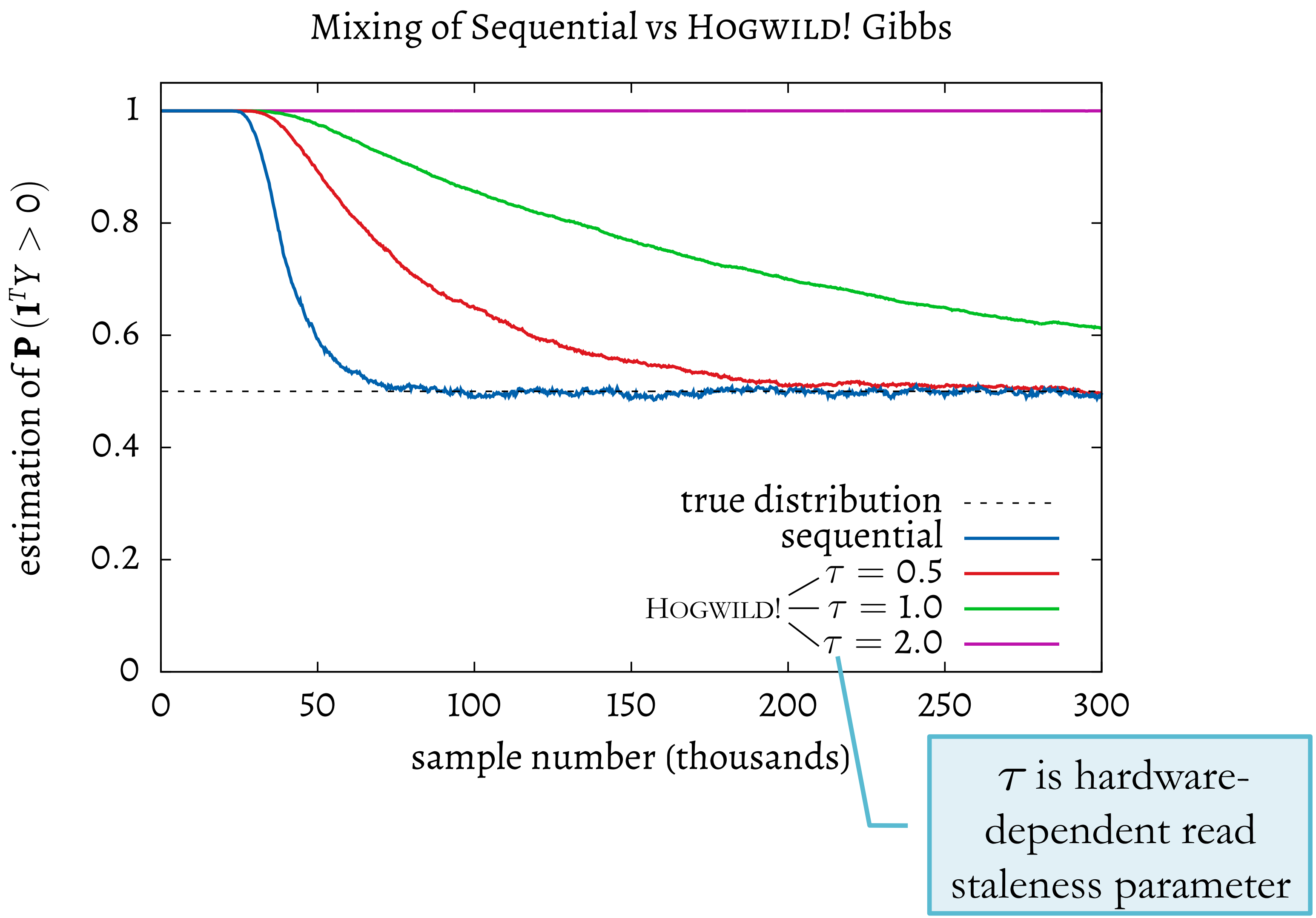
- **Mixing time** of a Markov chain is the first time at which the distribution of the sample is close to the stationary distribution.
  - in terms of total variation distance
  - feasible to run MCMC **if mixing time is small**

Mixing Time

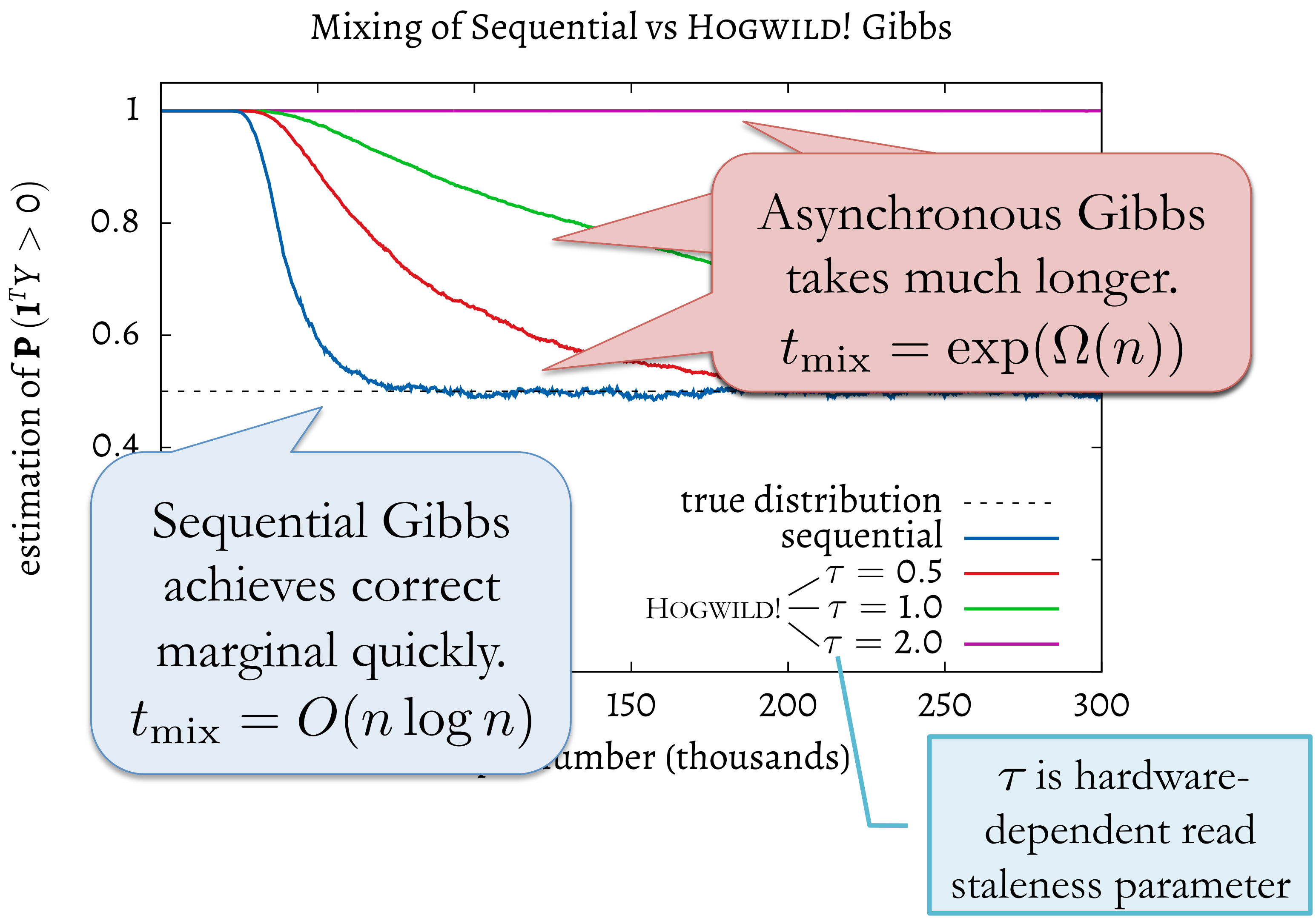
- **How long** do we need to run until the samples are **independent of initial conditions**?
- **Mixing time** of a Markov chain is the first time at which the distribution of the sample is close to the stationary distribution.
  - in terms of total variation distance
  - feasible to run MCMC **if mixing time is small**

"Folklore": asynchronous Gibbs has the same mixing time as sequential Gibbs…also **not necessarily true**!

Mixing Time Example



Mixing Time Example



Bounding the Mixing Time

$\alpha < 1$

Bounding the Mixing Time

Suppose that our target distribution satisfies **Dobrushin's condition** (total influence $\alpha < 1$).

- Mixing time of sequential Gibbs (known result)

$$t_{\text{mix-seq}}(\epsilon) \leq \frac{n}{1-\alpha} \log\left(\frac{n}{\epsilon}\right).$$

- Mixing time of asynchronous Gibbs is

$$t_{\text{mix-hog}}(\epsilon) \leq \frac{n + \alpha\tau}{1-\alpha} \log\left(\frac{n}{\epsilon}\right).$$

τ is hardware-dependent read staleness parameter

Bounding the Mixing Time

Suppose that our target distribution satisfies **Dobrushin's condition** (total influence $\alpha < 1$).

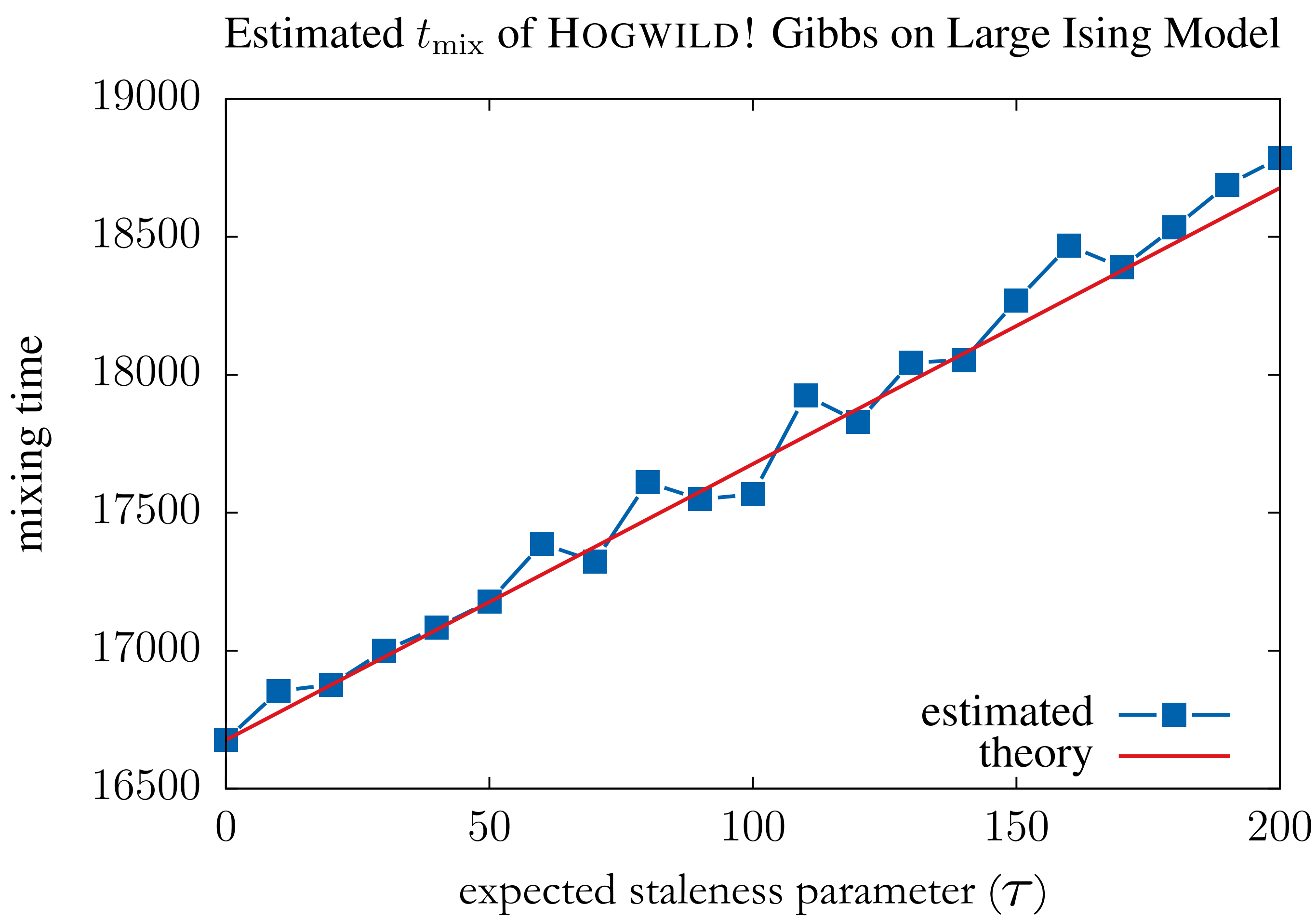
Takeaway message: can compare the two mixing time bounds with

$$t_{\text{mix}-\text{hog}}(\epsilon) \approx \left(1 + \alpha\tau n^{-1}\right) t_{\text{mix}-\text{seq}}(\epsilon)$$

…they differ by a **negligible factor**!

τ is hardware-dependent read staleness parameter

Theory Matches Experiment



Conclusion

- Analyzed and modeled **asynchronous Gibbs sampling**, and identified **two success metrics**
  - sample bias → **how close** to target distribution?
  - mixing time → **how long** do we need to run?
- Showed that asynchronicity can cause problems
- Proved bounds on the effect of asynchronicity
  - using the new **sparse variation distance**, together with
  - the classical condition of **total influence**

Conclusion

Thank you!

cdesa@stanford.edu
stanford.edu/~cdesa

- Analyzed and modeled **asynchronous Gibbs sampling**, and identified **two success metrics**
  - sample bias → **how close** to target distribution?
  - mixing time → **how long** do we need to run?
- Showed that asynchronicity can cause problems
- Proved bounds on the effect of asynchronicity
  - using the new **sparse variation distance**, together with
  - the classical condition of **total influence**