Accelerated Gibbs Sampling for the Indian Buffet Process

Finale Doshi-Velez, Cambridge/MIT
Zoubin Ghahramani, Cambridge

UNIVERSITY OF CAMBRIDGE

Motivation

Bilinear models of the form

$$X = UV + E$$

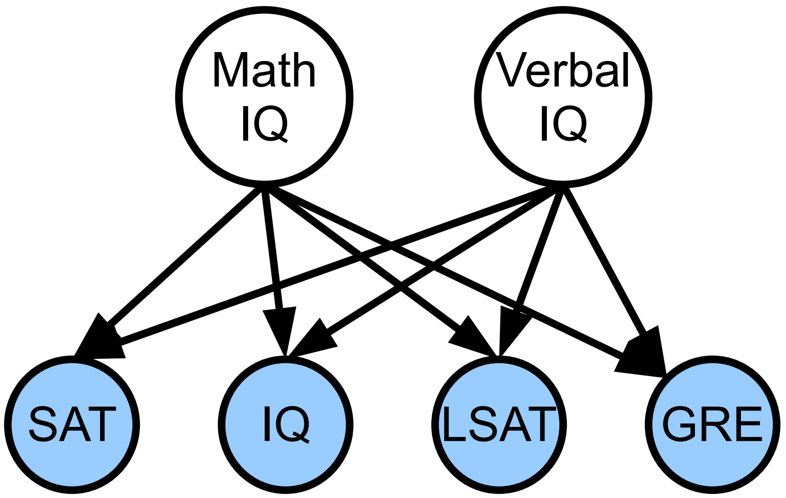
data = matrix product + error

are very common in machine learning.

Examples

Factor Analysis

$Y = LX + E$



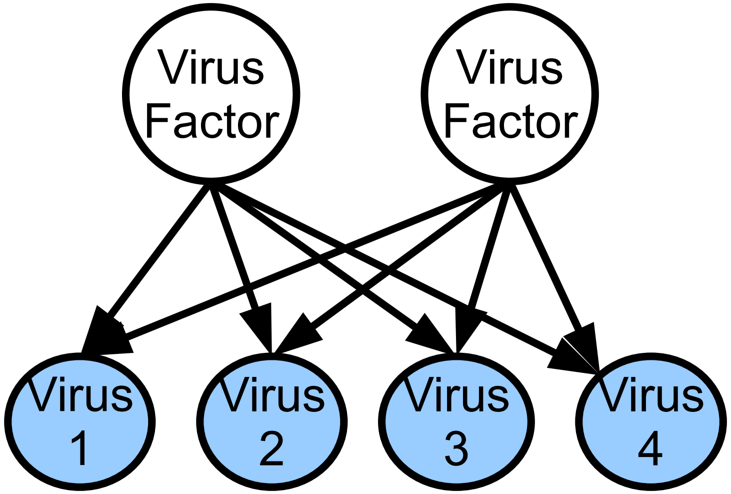
Examples

Factor Analysis

$Y = LX + E$

Probabilistic PCA

$T = WX + E$



Examples

Factor Analysis

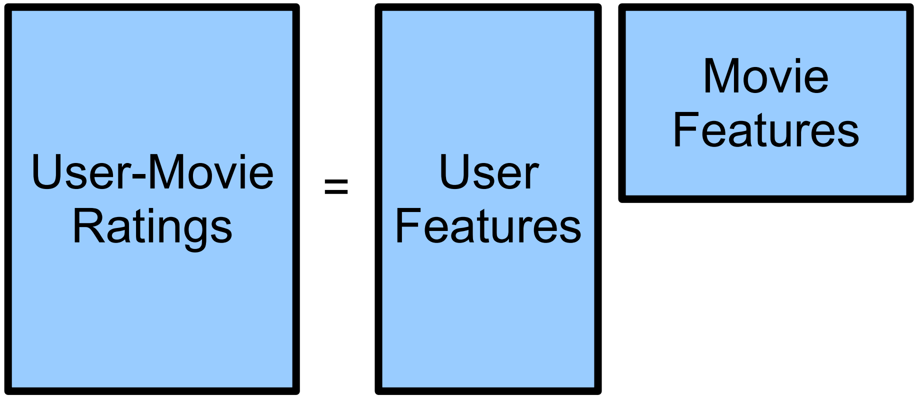
$$Y = LX + E$$

Probabilistic PCA

$$T = WX + E$$

Probabilistic Matrix Factorization

$$X = UV + E$$



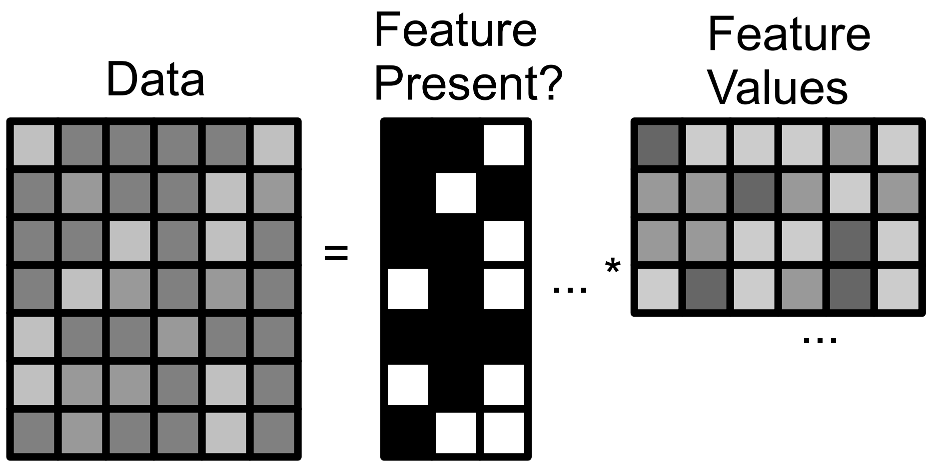
Examples

Factor Analysis

$Y = LX + E$

Probabilistic PCA

$T = WX + E$



Probabilistic Matrix Factorization

$X = UV + E$

Indian Buffet Process with a linear likelihood

$X = ZA + E$

Motivation

- We are interested in doing large-scale Bayesian inference in these models (focus on the IBP for now):

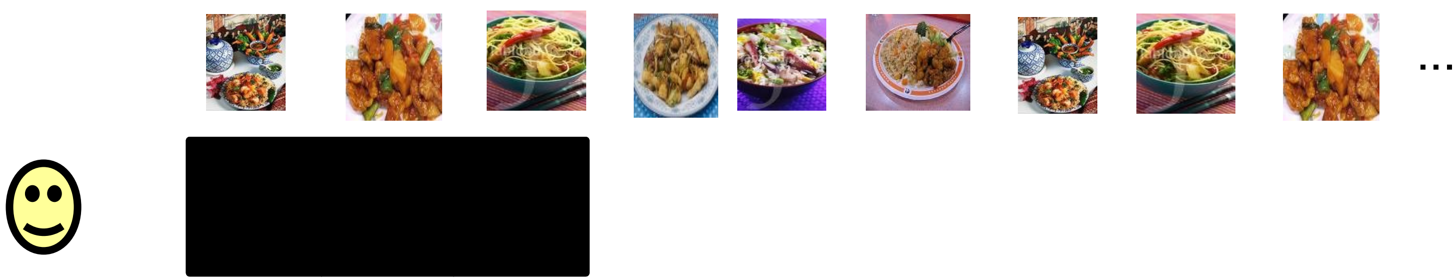
$$X = ZA + E$$

- Suppose
  - We can compute P(X|Z) , but it's expensive
  - We can compute P(A|X,Z)
  - We cannot compute P(Z,A|X)
- We develop a fast sampler for inference in these models.

Indian Buffet Process

Customers enter an "infinite buffet" one at a time and

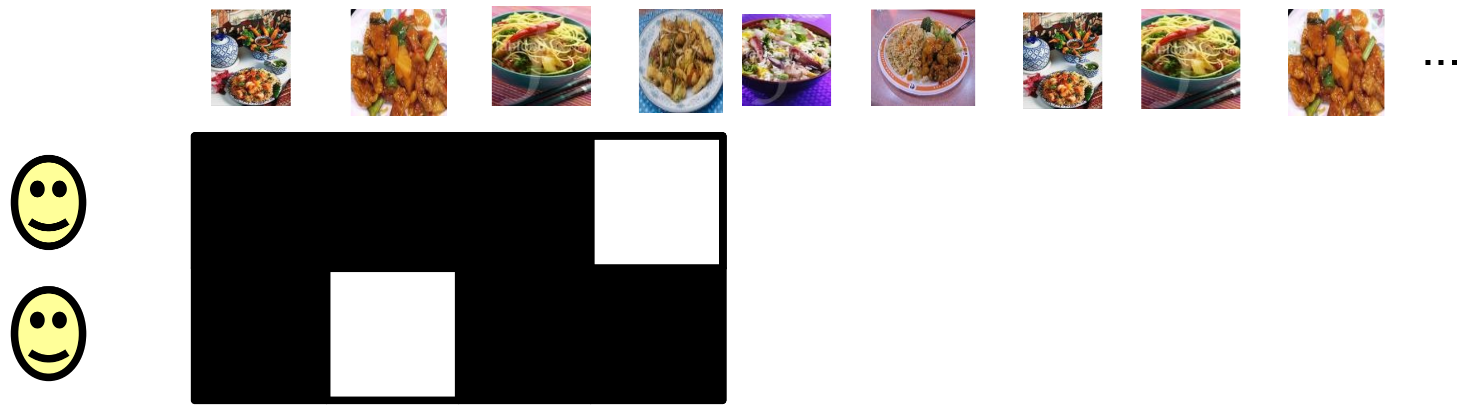
- Sample a previously sampled dish based on its popularity.
- Sample Poisson(alpha / n) new dishes.



Indian Buffet Process

Customers enter an "infinite buffet" one at a time and

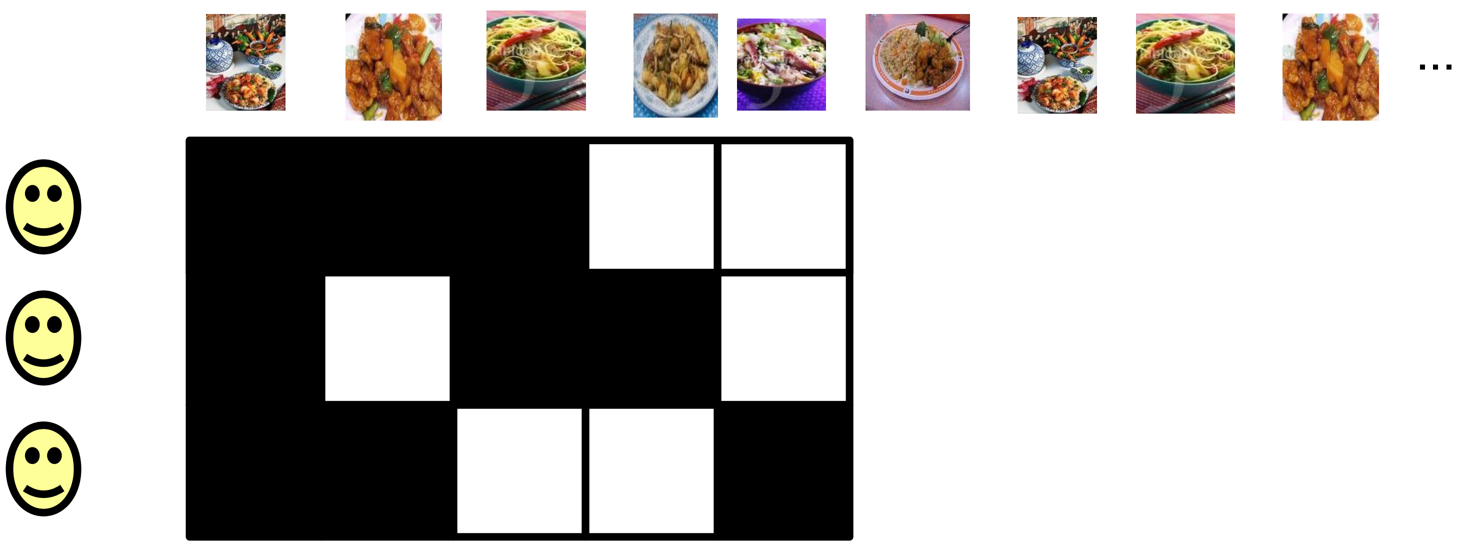
- Sample a previously sampled dish based on its popularity.
- Sample Poisson(alpha / n) new dishes.



Indian Buffet Process

Customers enter an "infinite buffet" one at a time and

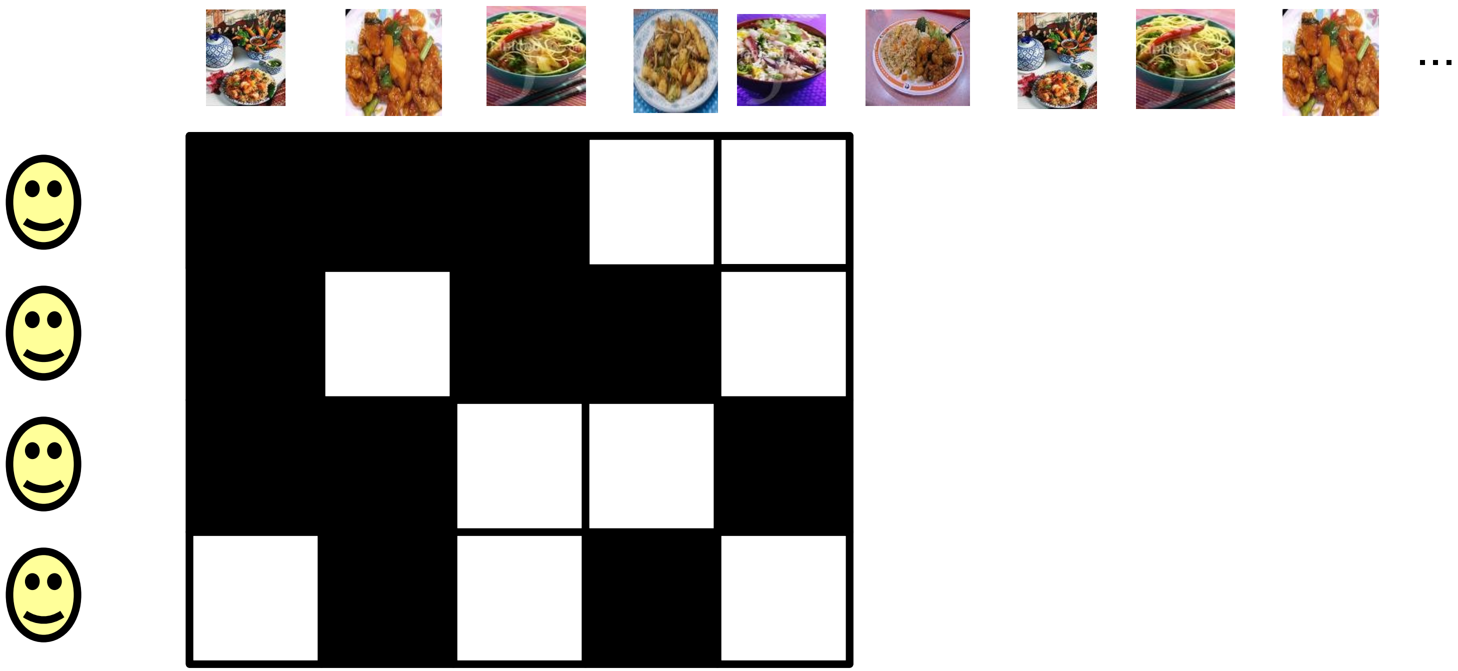
- Sample a previously sampled dish based on its popularity.
- Sample Poisson(alpha / n) new dishes.



Indian Buffet Process

Customers enter an “infinite buffet” one at a time and

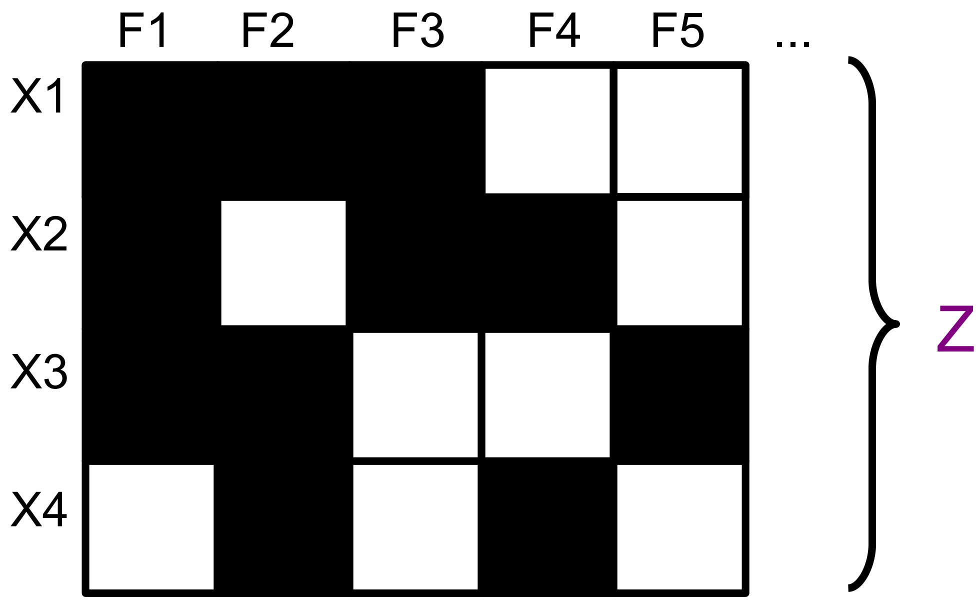
- Sample a previously sampled dish based on its popularity.
- Sample Poisson(alpha / n) new dishes.



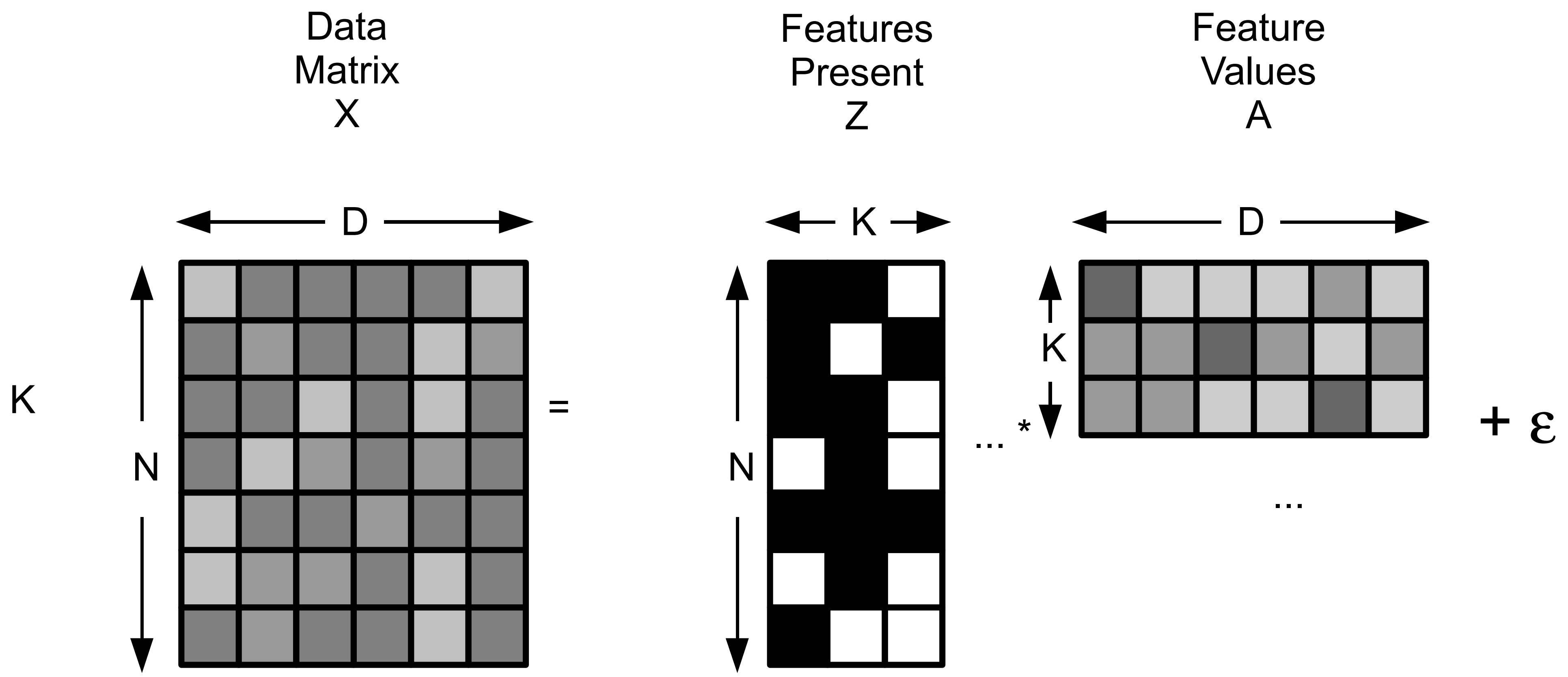
Indian Buffet Process

Result is a non-parametric prior on feature assignments—a general tool for many latent feature models—with some nice properties:

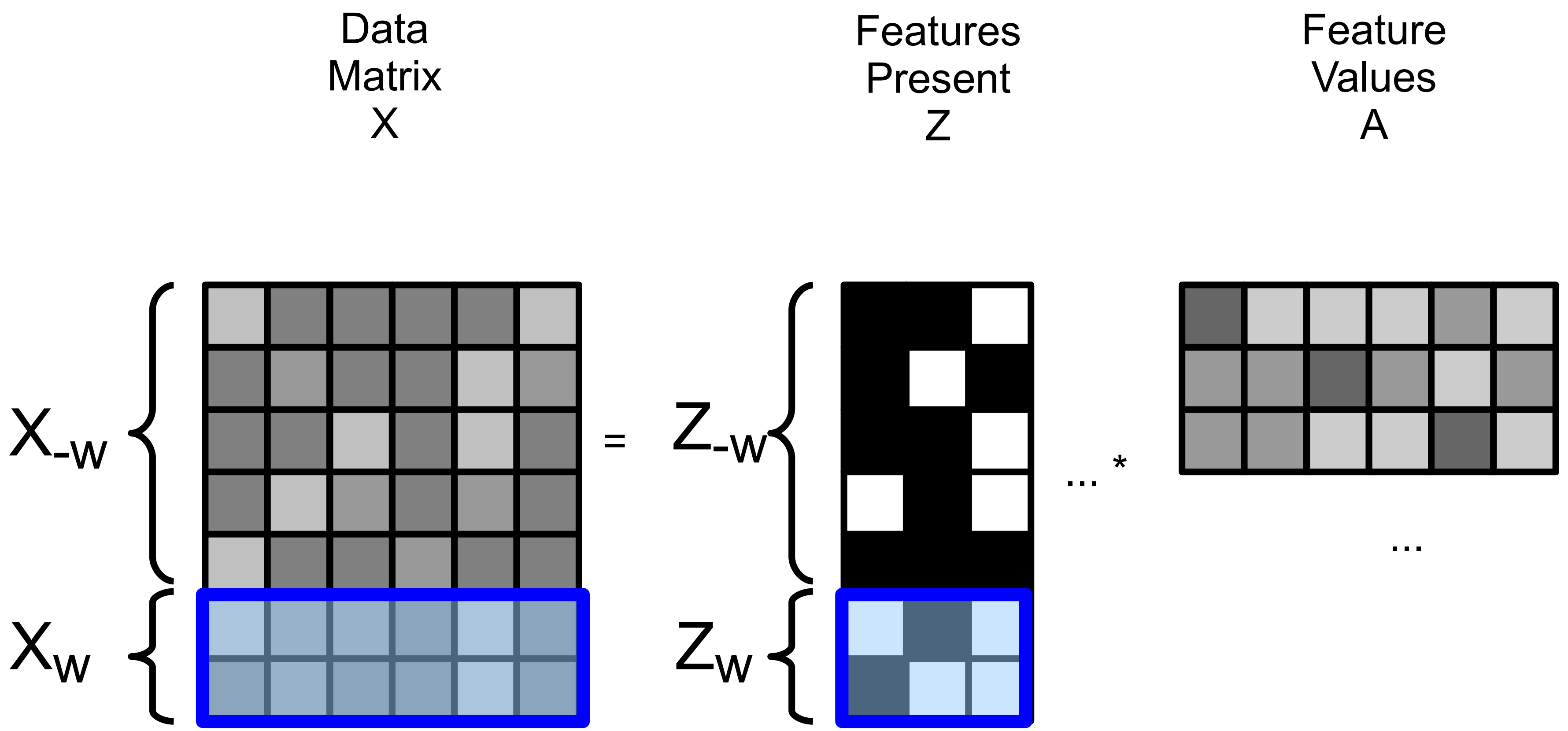
- Observations are exchangeable.
- Infinite features, but finite datasets contain a finite number of features.



Full Model

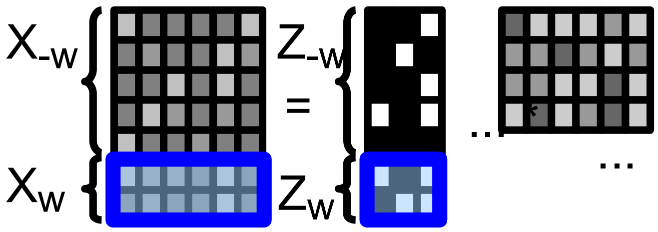


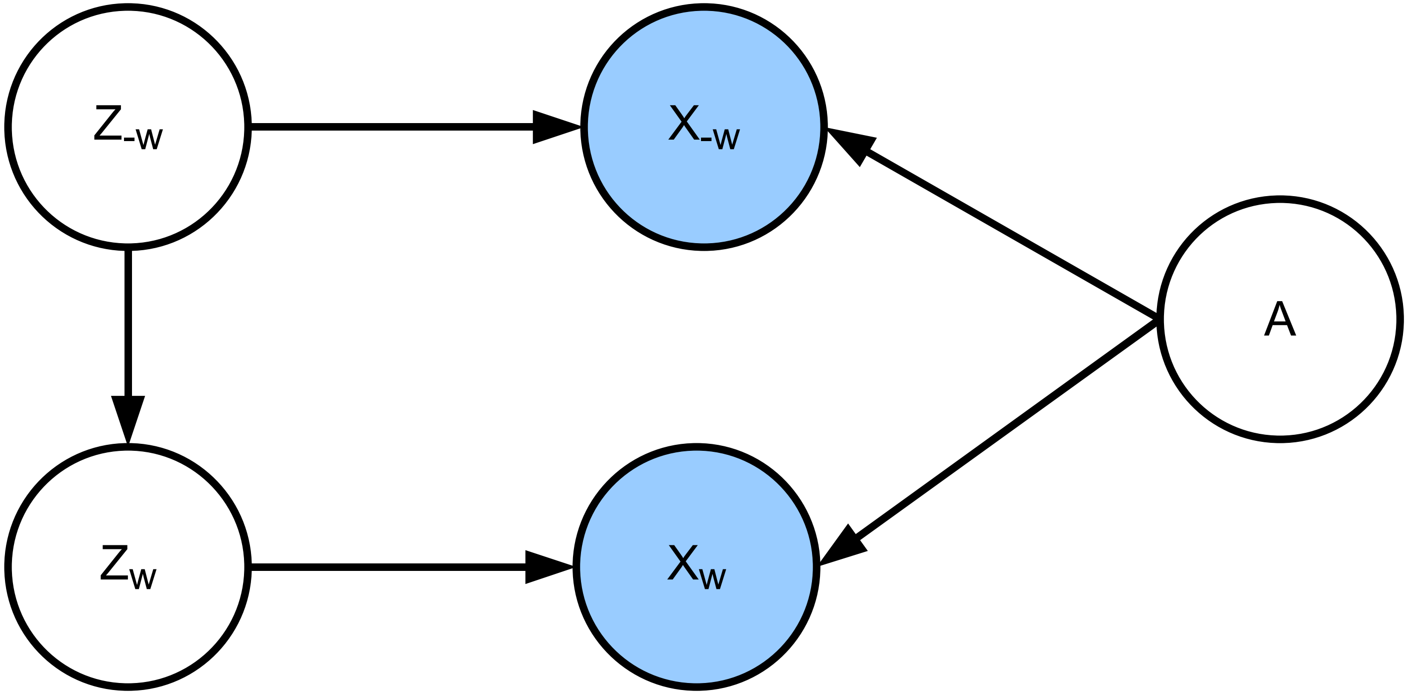
Full Model



Note: this is not Blocked Gibbs Sampling!

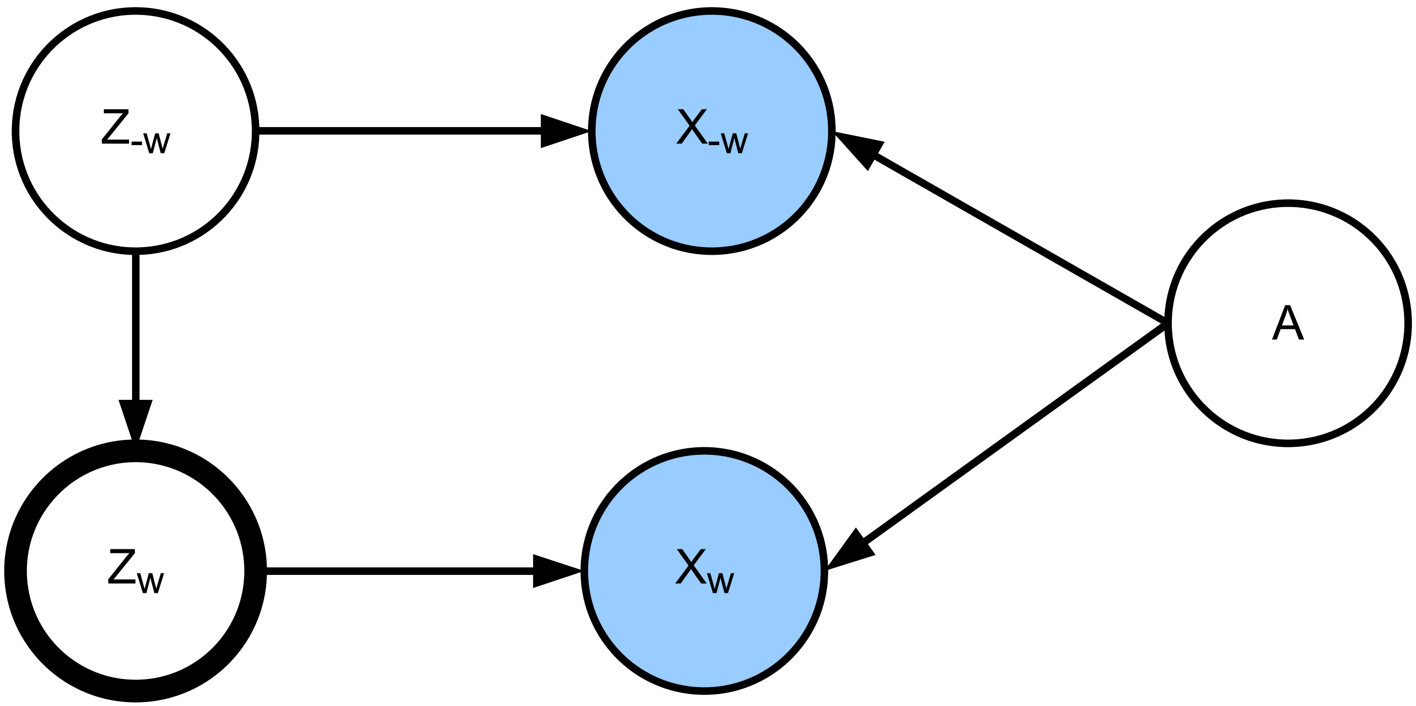
The Graphical Model





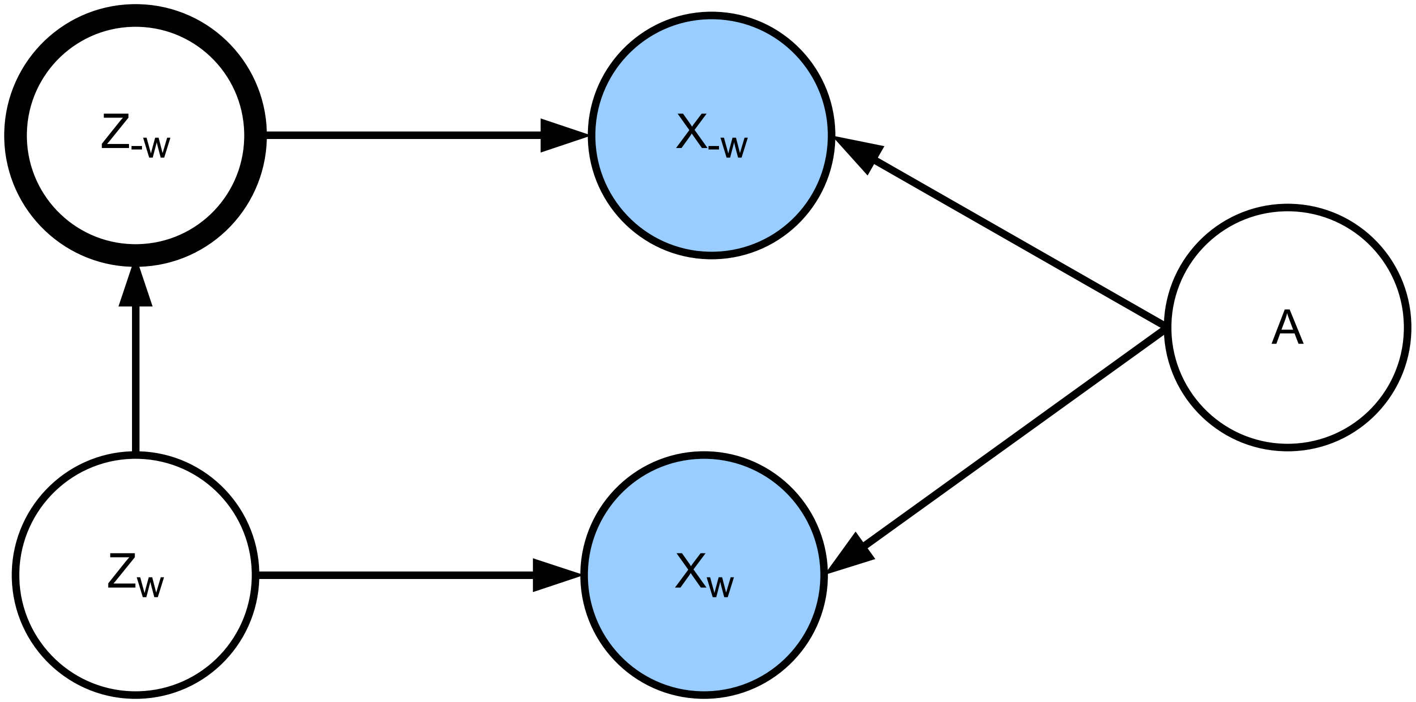
Basic Sampling

First sample $Z_w | X, A, Z_{-w}$



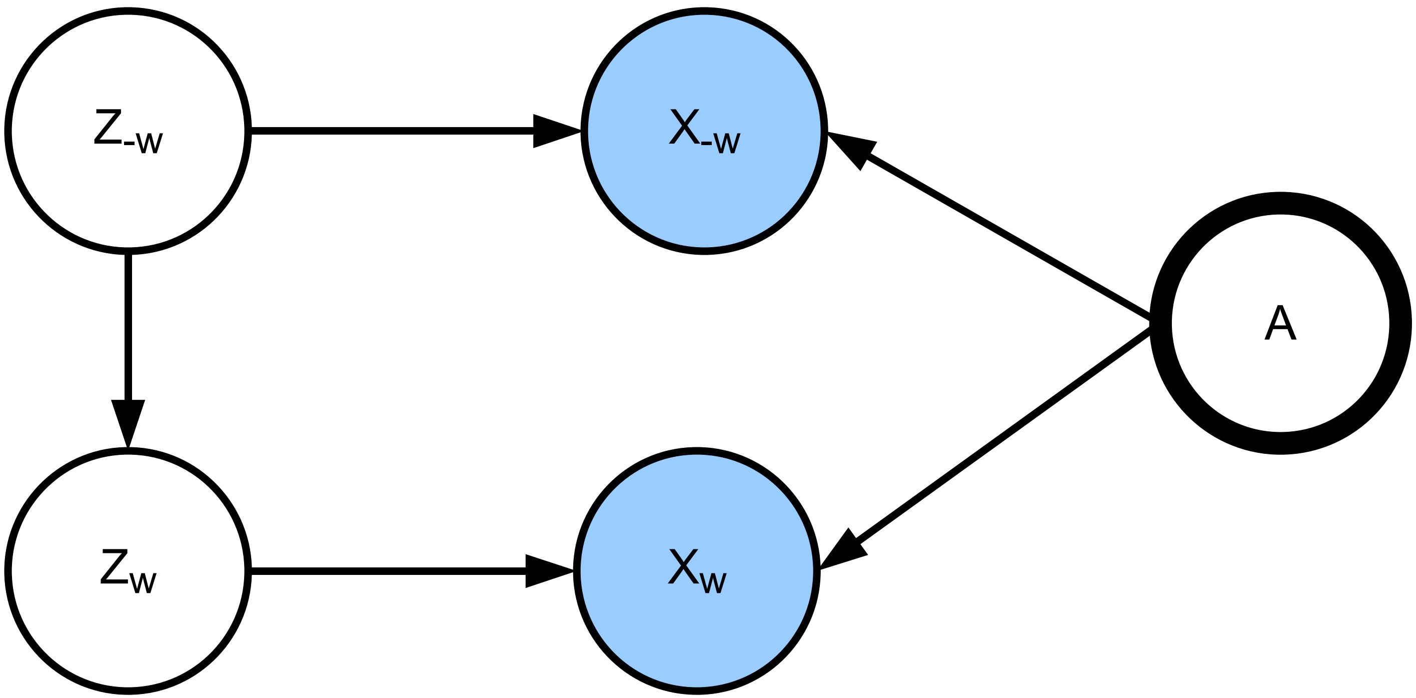
Basic Sampling

First sample $Z_w|X,A,Z_{-w}$ and then $Z_{-w}|X,A,Z_w$



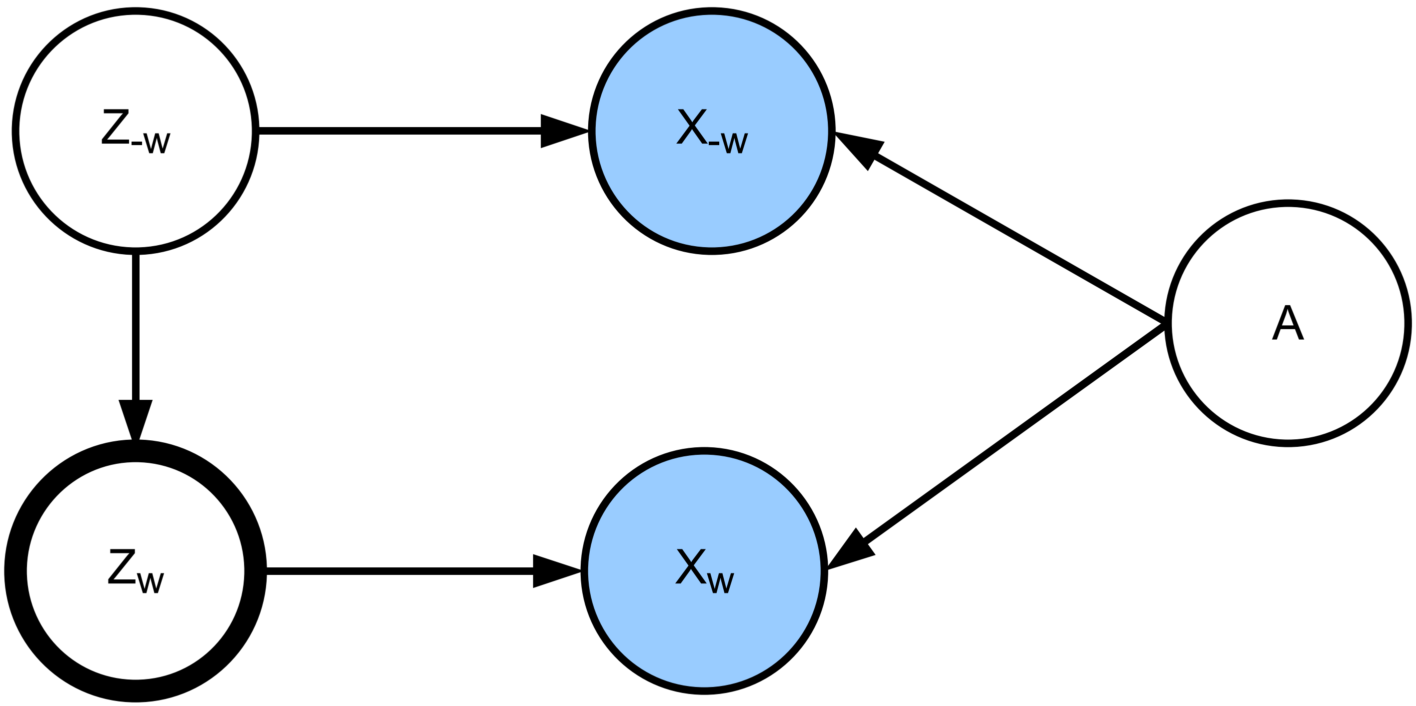
Basic Sampling

First sample $Z_w|X,A,Z_{-w}$ and then $Z_{-w}|X,A,Z_w$ and then $A|Z,X$ ...



Basic Sampling

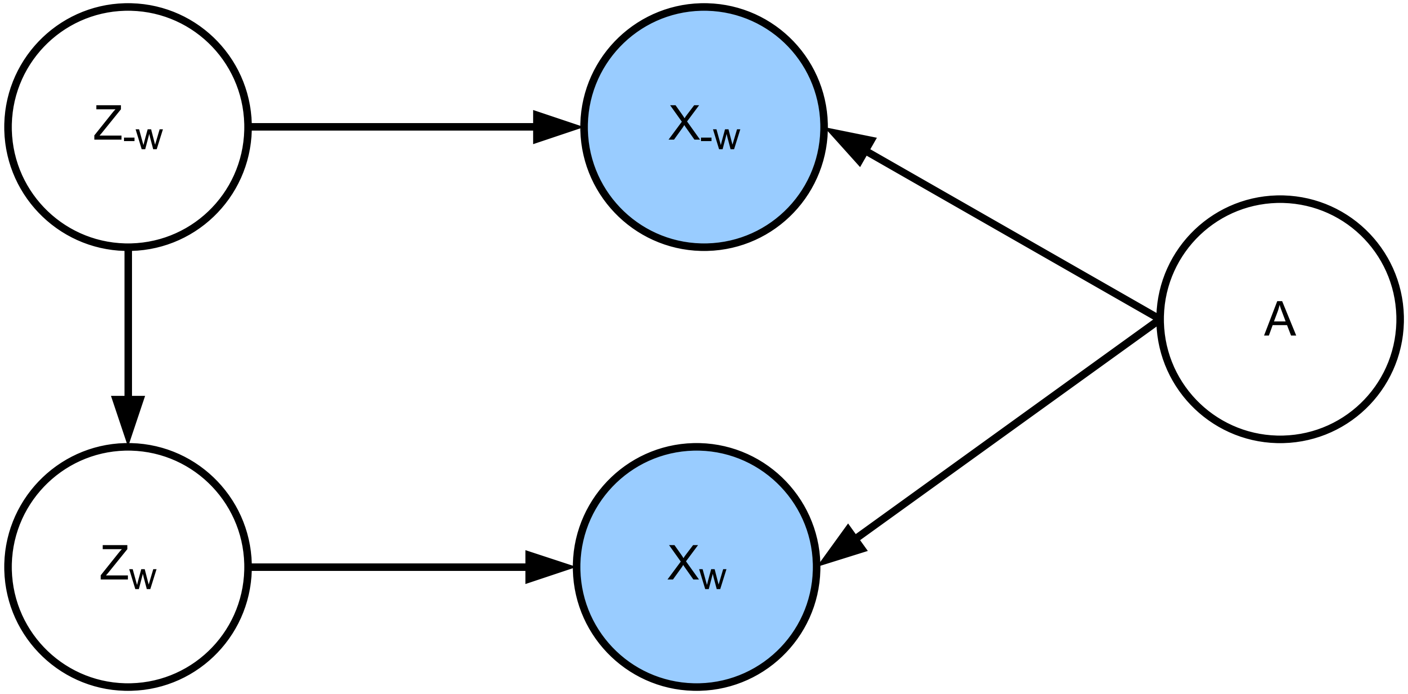
First sample $Z_w|X,A,Z_{-w}$ and then $Z_{-w}|X,A,Z_w$ and then $A|Z,X$ and then $Z_w|X,A,Z_{-w}$ ...



Basic Sampling

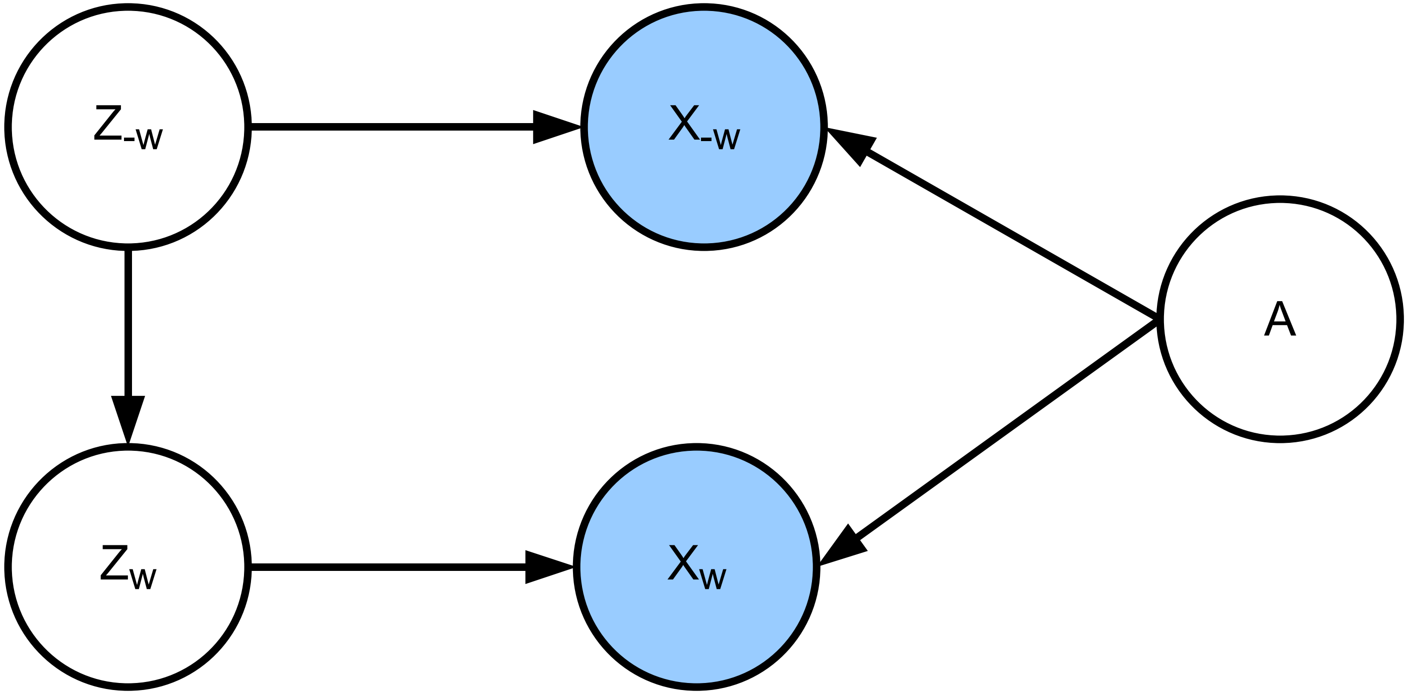
Advantage: Each iteration is fast to compute.

Disadvantage: Often slow to mix.



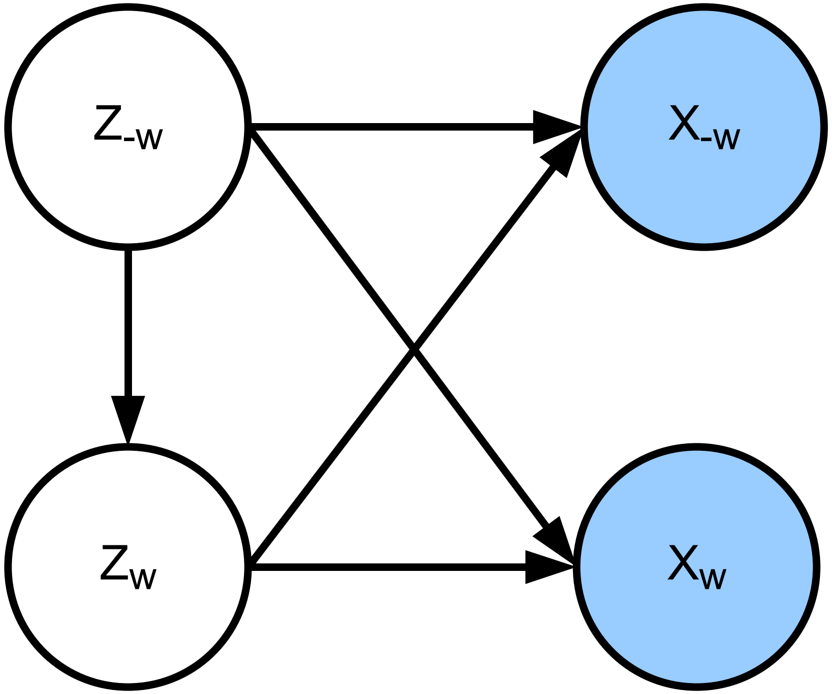
Collapsed Gibbs Sampling

Since we can compute P(X|Z), integrate out A



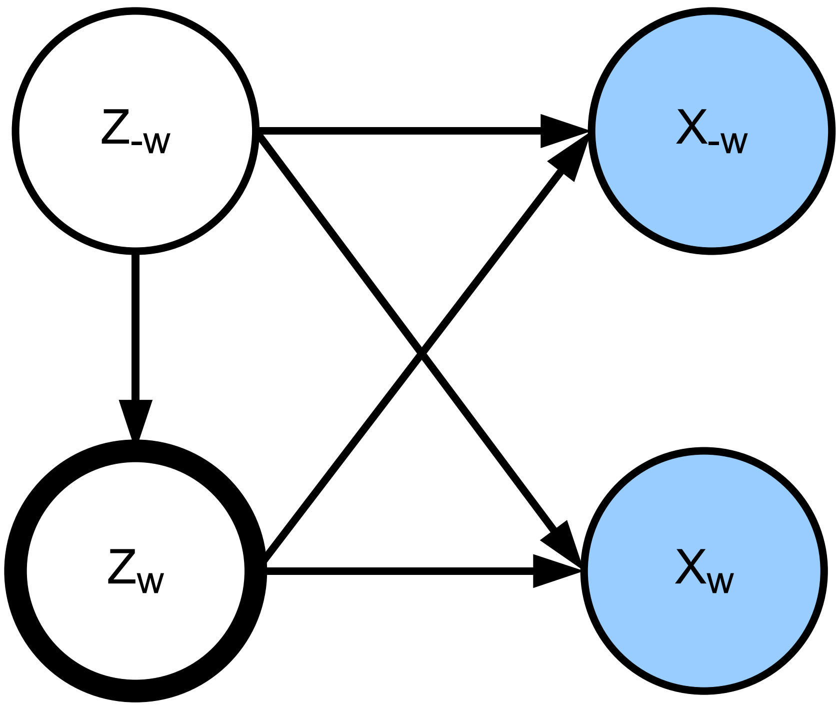
Collapsed Gibbs Sampling

Since we can compute P(X|Z), integrate out A



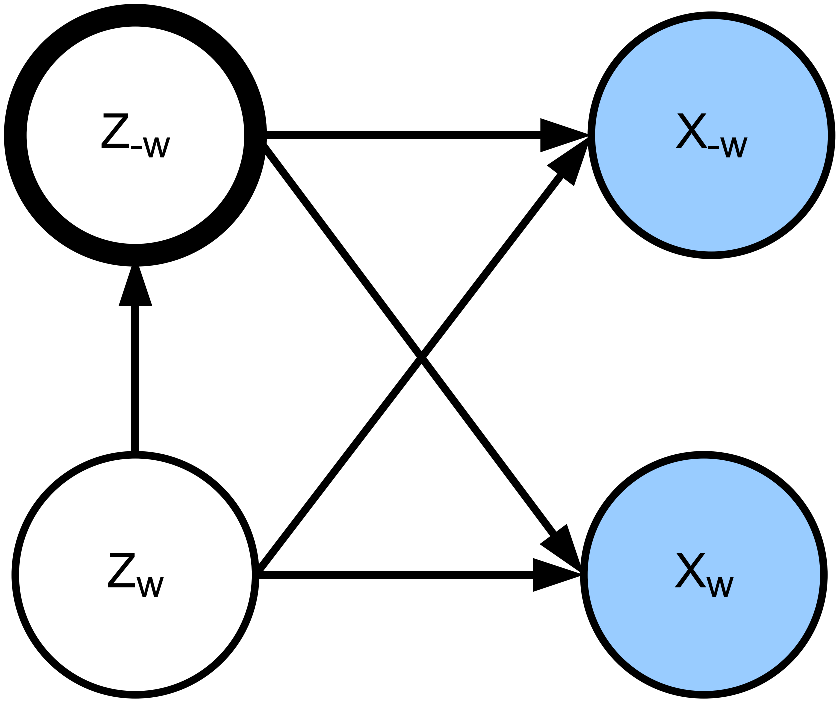
Collapsed Gibbs Sampling

Sample each Z in turn, as before



Collapsed Gibbs Sampling

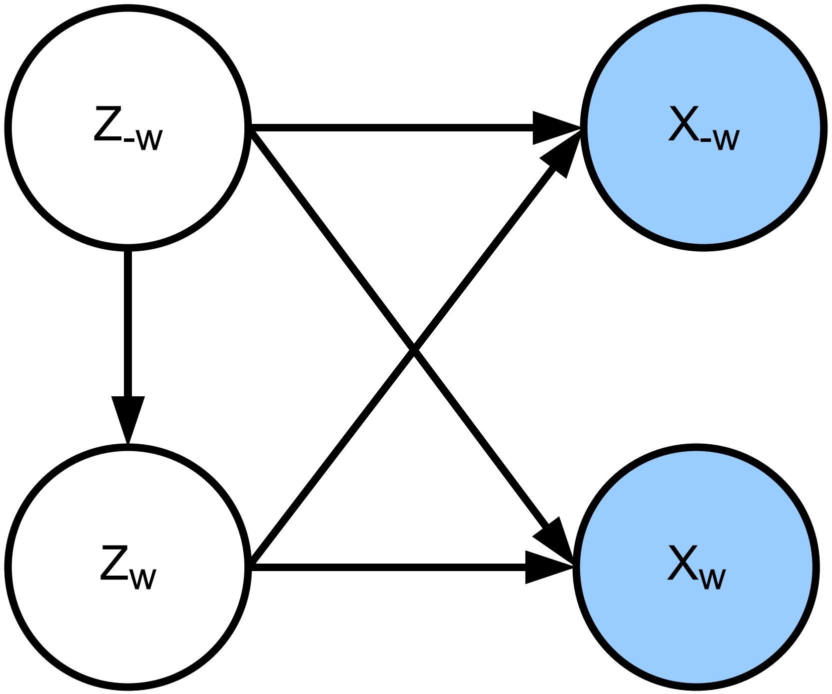
Sample each Z in turn, as before



Collapsed Gibbs Sampling

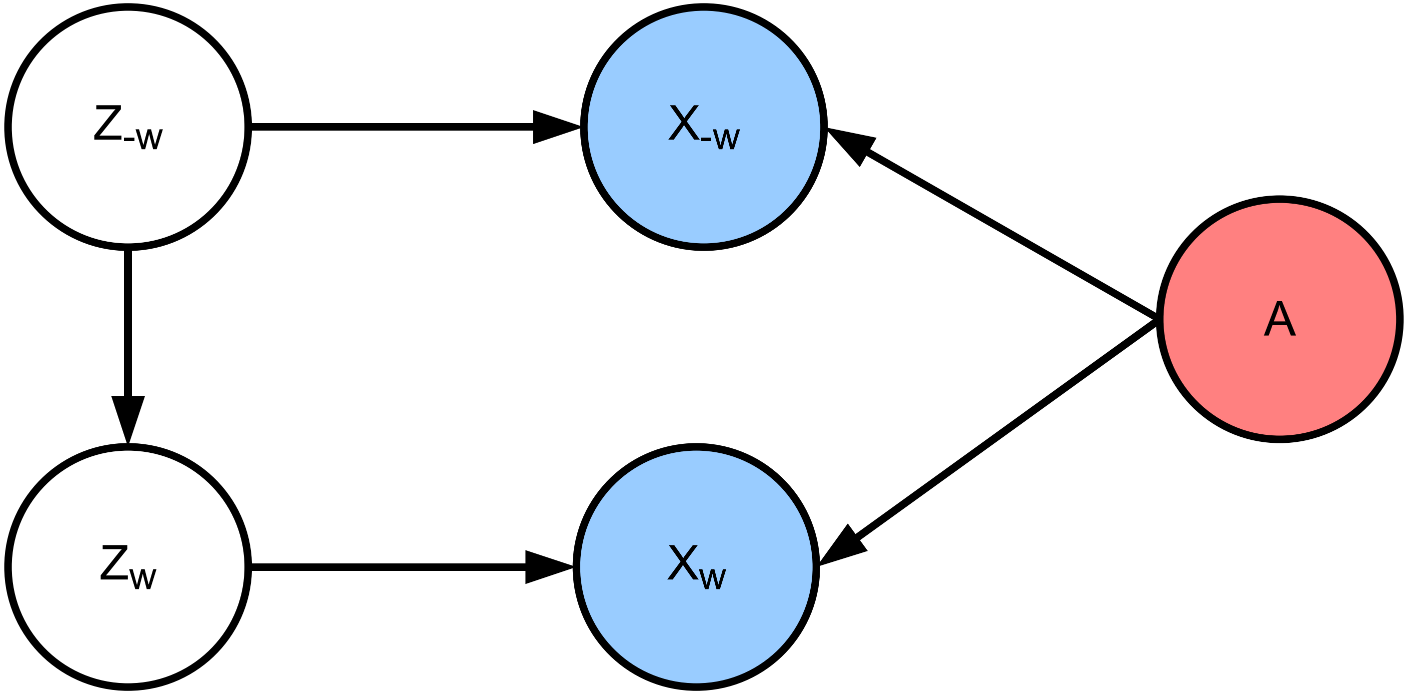
Advantage: Faster to mix.

Disadvantage: Inference no longer scales!

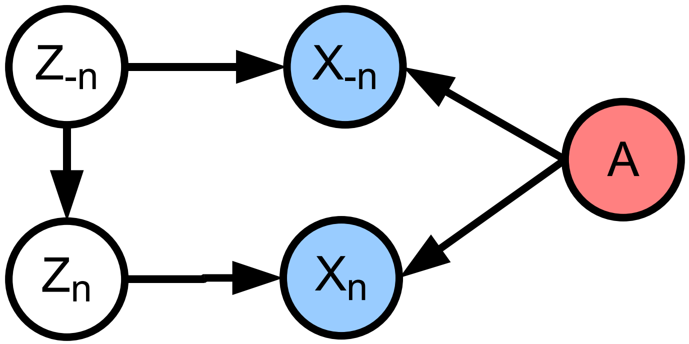


Our solution: Accelerated Sampling

Keep a posterior on A. Observations stay independent!



More formally: Consider one element



$$P(Z_{nk}=1|Z_{-nk},X)\propto$$

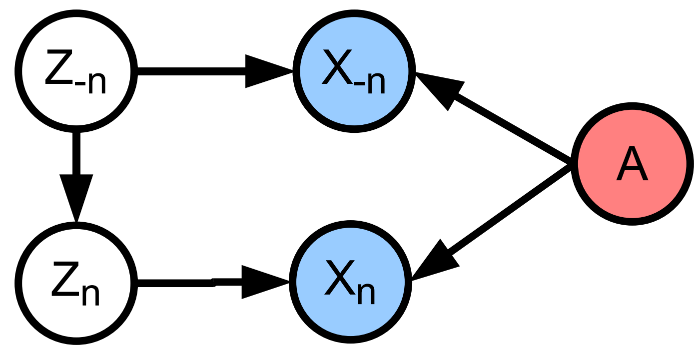
$$P(Z_{nk}=1|Z_{-nk})P(X|Z)$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X|Z,A)P(A)\,dA$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)P(X_{-n}|Z_{-n},A)P(A)\,dA$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)\,{\color{darkred}P(A|Z_{-n},X_{-n})}\,dA$$

More formally: Consider one element



$$P(Z_{nk}=1|Z_{-nk},X)\propto$$

Bayes Rule

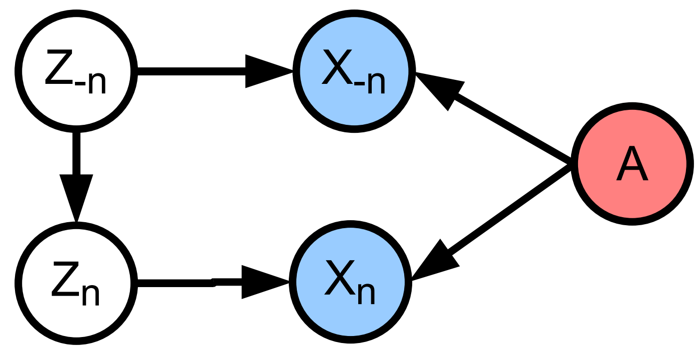
$$P(Z_{nk}=1|Z_{-nk})\,P(X|Z)$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X|Z,A)\,P(A)\,dA$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)\,P(X_{-n}|Z_{-n},A)\,P(A)\,dA$$

$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)\,P(A|Z_{-n},X_{-n})\,dA$$

More formally: Consider one element



$$P(Z_{nk}=1|Z_{-nk}, X) \propto$$

$$P(Z_{nk}=1|Z_{-nk})\, P(X|Z)$$

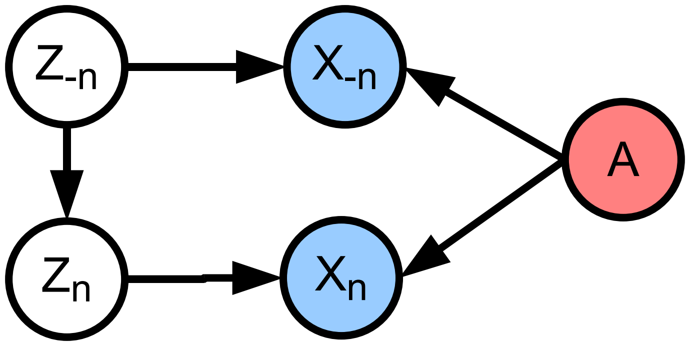
$$P(Z_{nk}=1|Z_{-nk}) \int_A P(X|Z, A)\, P(A)\, dA$$

Joints and conditionals

$$P(Z_{nk}=1|Z_{-nk}) \int_A P(X_n|Z_n, A)\, P(X_{-n}|Z_{-n}, A)\, P(A)\, dA$$

$$P(Z_{nk}=1|Z_{-nk}) \int_A P(X_n|Z_n, A)\, P(A|Z_{-n}, X_{-n})\, dA$$

More formally: Consider one element

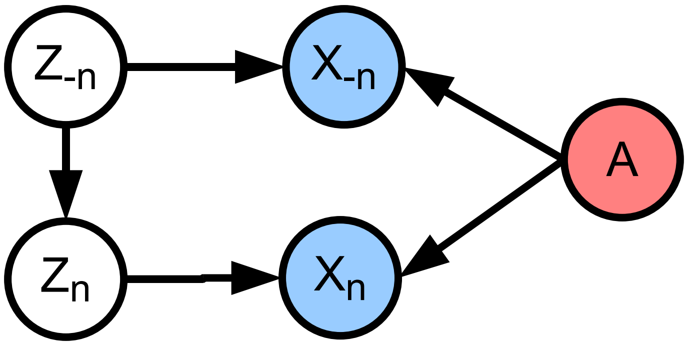


$$P(Z_{nk}=1|Z_{-nk},X)\propto$$

$$P(Z_{nk}=1|Z_{-nk})P(X|Z)$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X|Z,A)P(A)\,dA$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)P(X_{-n}|Z_{-n},A)P(A)\,dA$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)P(A|Z_{-n},X_{-n})\,dA$$

Bayes Rule

More formally: Consider one element



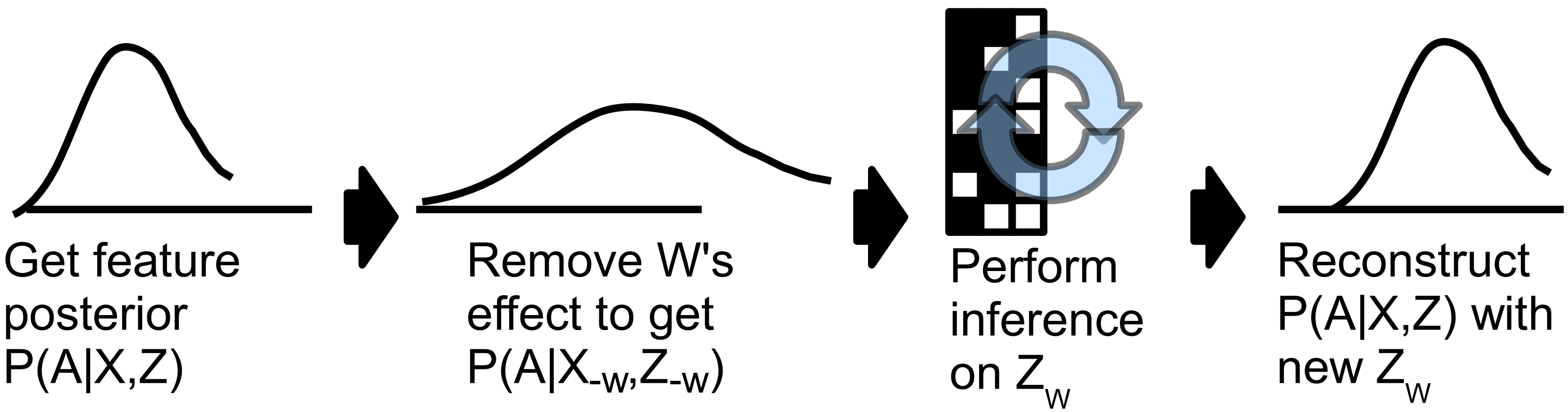
$$P(Z_{nk}=1|Z_{-nk},X)\propto$$

$$P(Z_{nk}=1|Z_{-nk})P(X|Z)$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X|Z,A)P(A)\,dA$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)P(X_{-n}|Z_{-n},A)P(A)\,dA$$
$$P(Z_{nk}=1|Z_{-nk})\int_A P(X_n|Z_n,A)P(A|Z_{-n},X_{-n})\,dA$$

EXACT!

Accelerated Gibbs Sampling

1. Initialise some Z, feature posterior
2. For each window of observations W



Considerations: how many observations should we consider at once? Depends on the cost of computing $P(A|X,Z)$ and $P(X|Z,A)$, numerical errors.

Details for the IBP Model

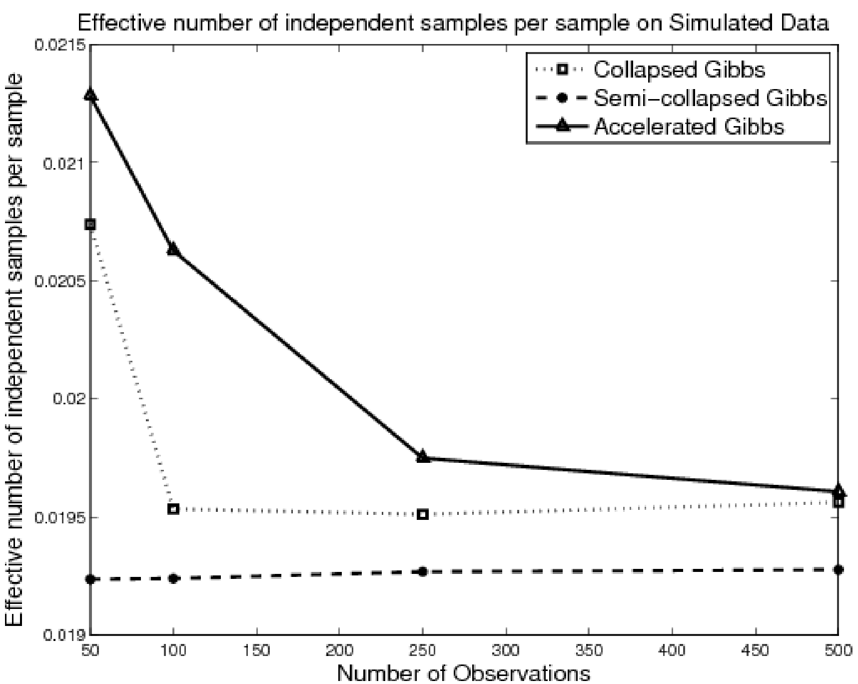
If the prior on A, noise is Gaussian, then

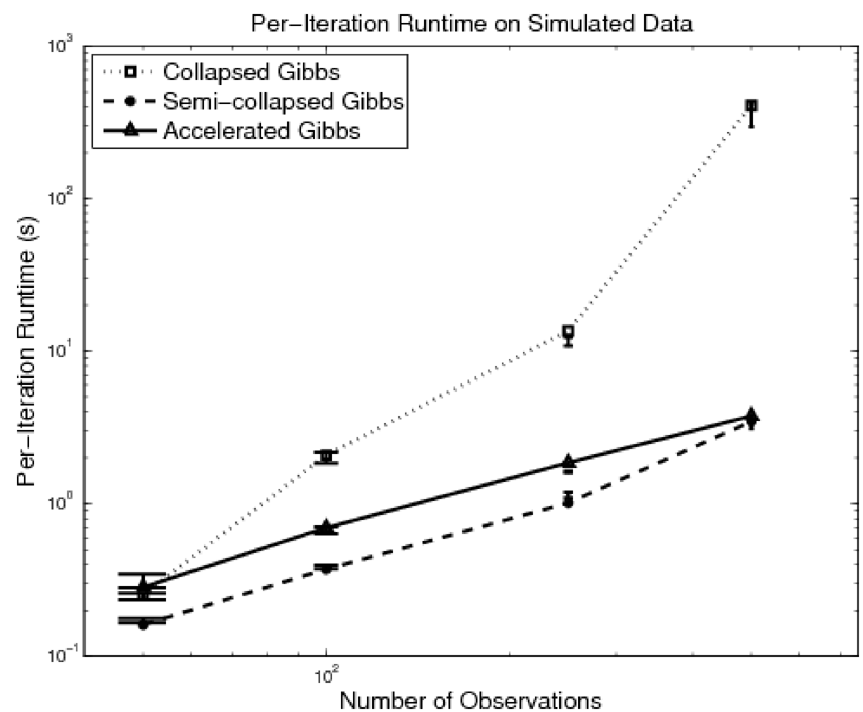
- Posterior on A is Gaussian.
- Posterior can be updated with rank-one updates.
- Optimal window is 1.

Also, intelligently choosing to represent Gaussians in information form (h, Σ^{-1}) or covariance form (μ, Σ) helps maintain numerical precision. Details in the paper.

Experiments on Synthetic Data

Data generated from the prior; D=10, N = {50,100,250, 500}.



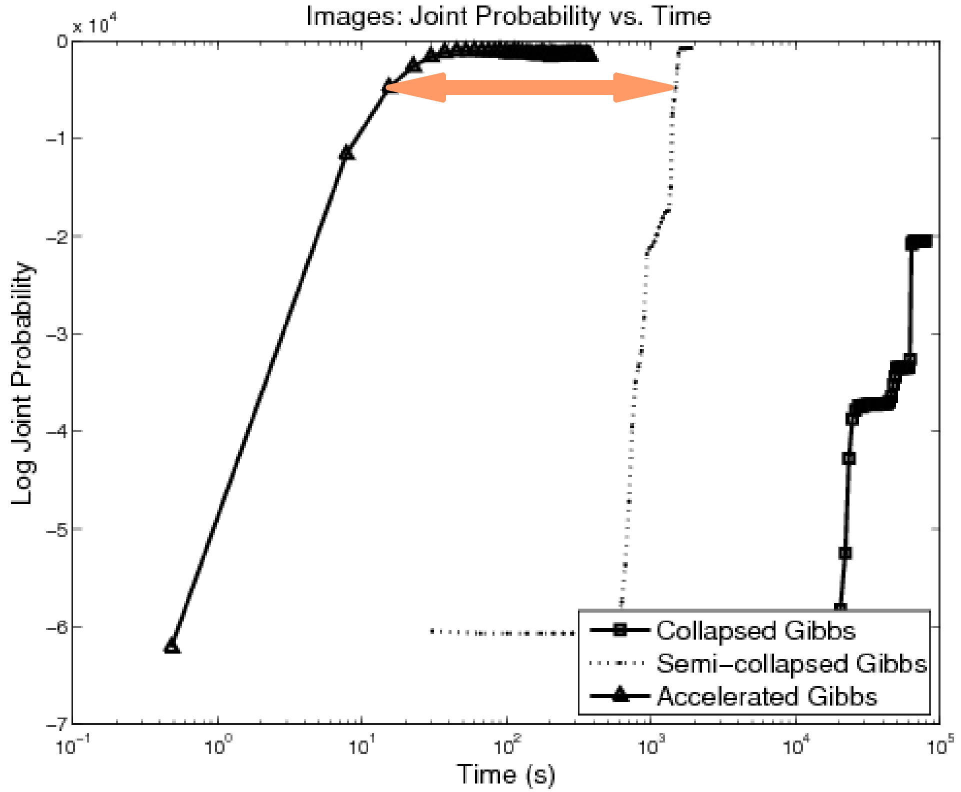


Mixing similar to collapsed sampler

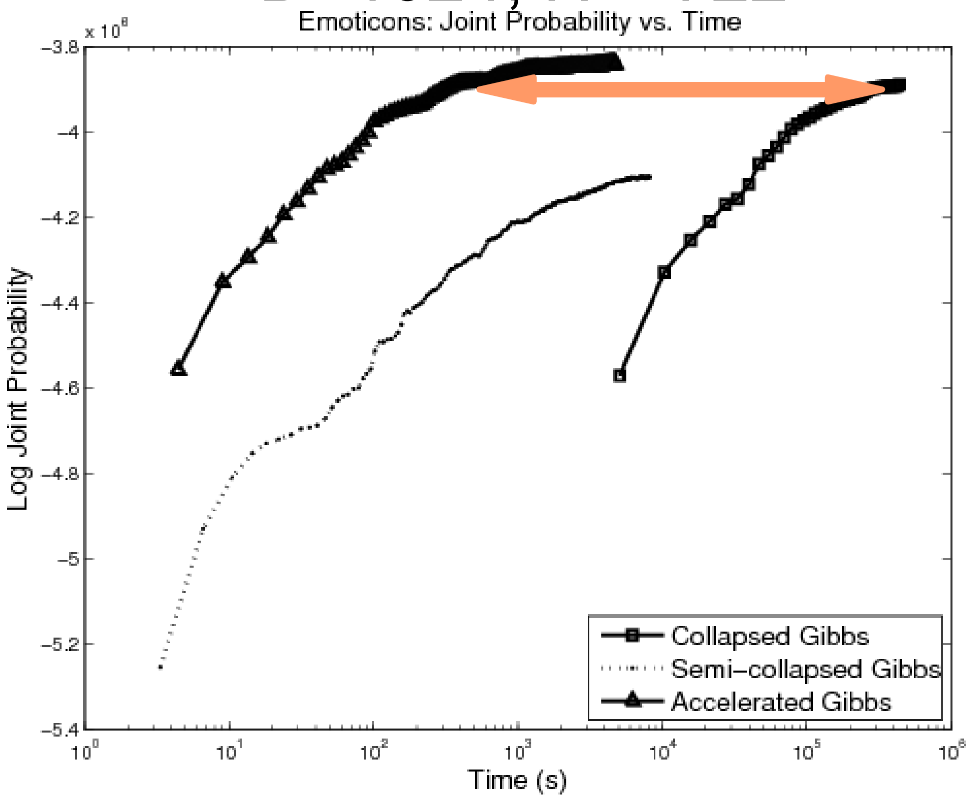
Runtime similar to semi-collapsed sampler

Experiments on Smaller Datasets

D=36, N = 1000



D=1024, N = 722

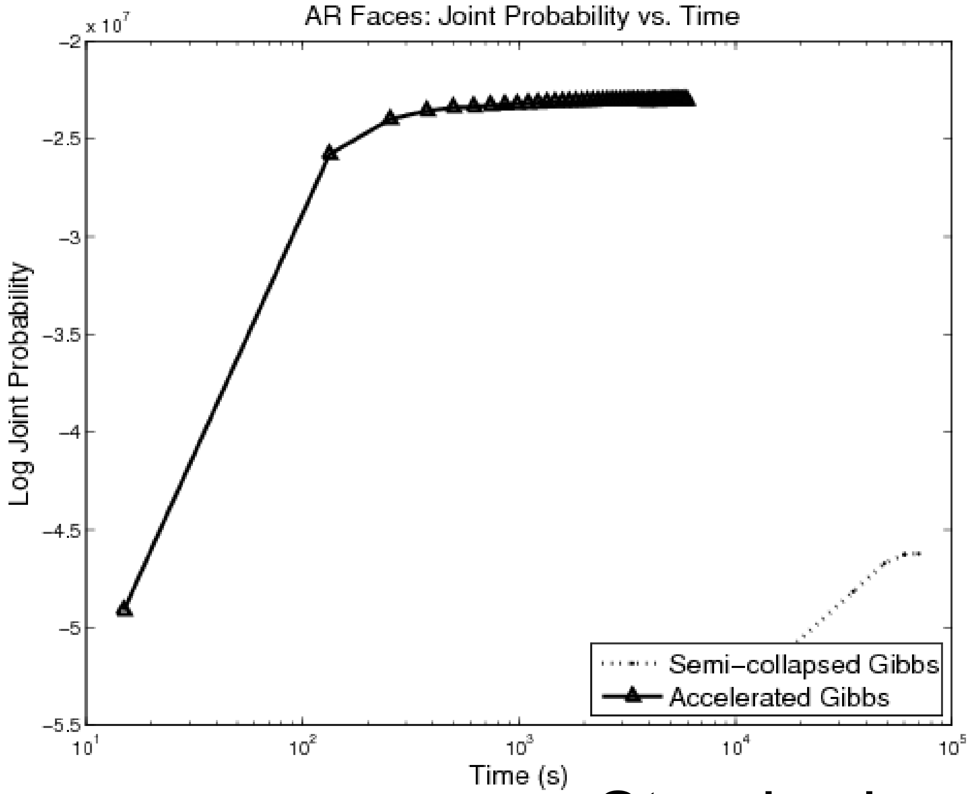


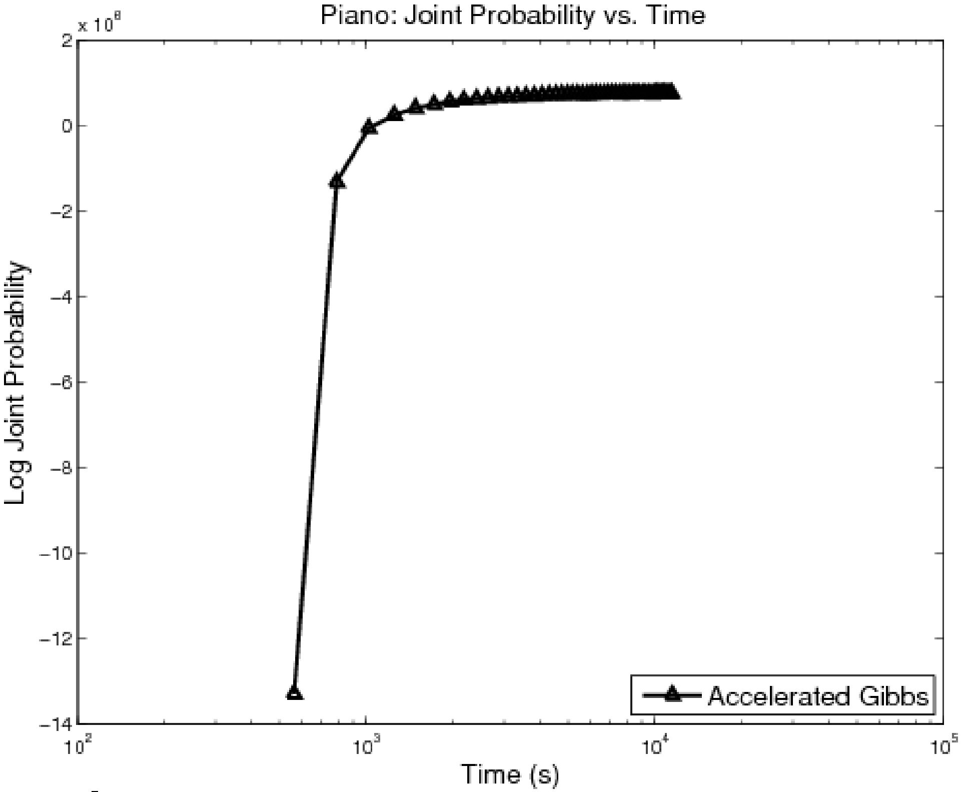
Reach mode orders of magnitude faster!

Experiments on Larger Datasets

D=1598, N = 2600

D=161, N = 10000





Standard samplers
become impractical...

Returning to an age-old question...

To marginalize or not marginalize, that is the question:

Whether 'tis more tractable for the sampler to suffer the hills and valleys of local optima,

Or to take expectations against a set of variables, and by integrating collapse them?

Returning to an age-old question...

To marginalize or not marginalize, that is the question:

Whether 'tis more tractable for the sampler to suffer the hills and valleys of local optima,

Or to take expectations against a set of variables, and by integrating collapse them?

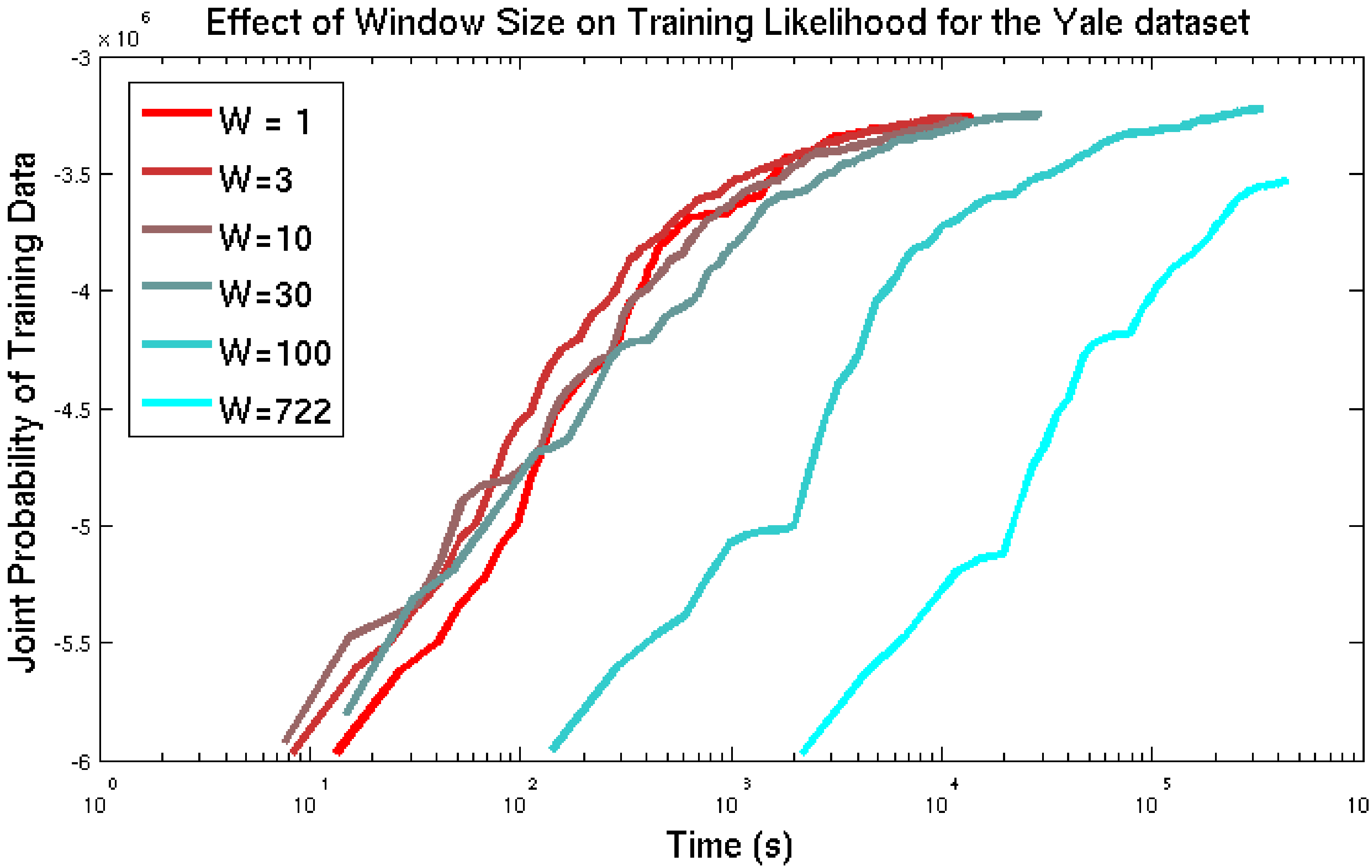
In answer: of a third example...

Conclusions

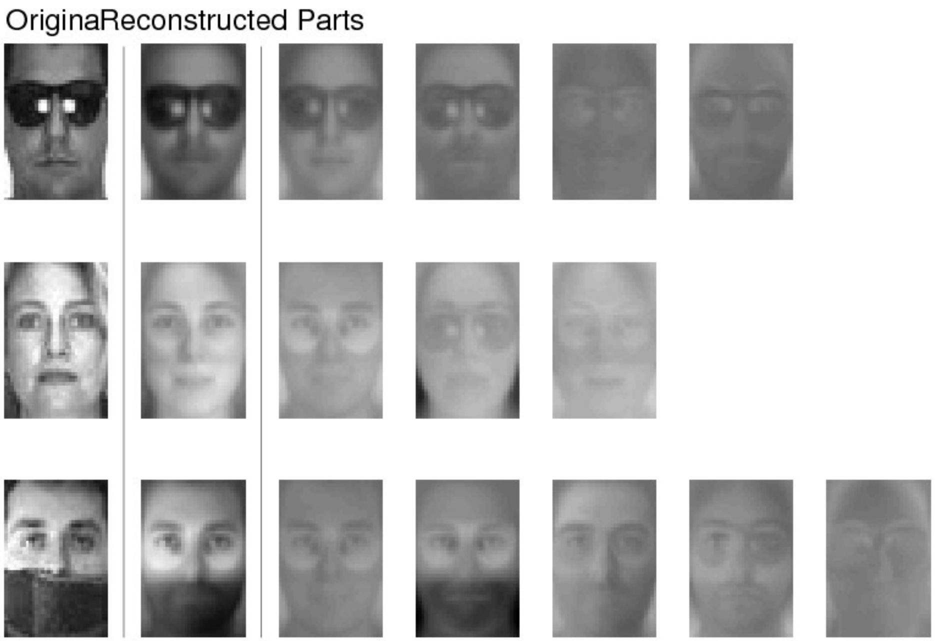
- Maintaining a posterior within a sampler allows us to perform fast inference in an important class of models
- In particular, our approach allows us to scale inference to large Indian Buffet Process models.

... code available on my website:
http://mlg.eng.cam.ac.uk/finale/wiki

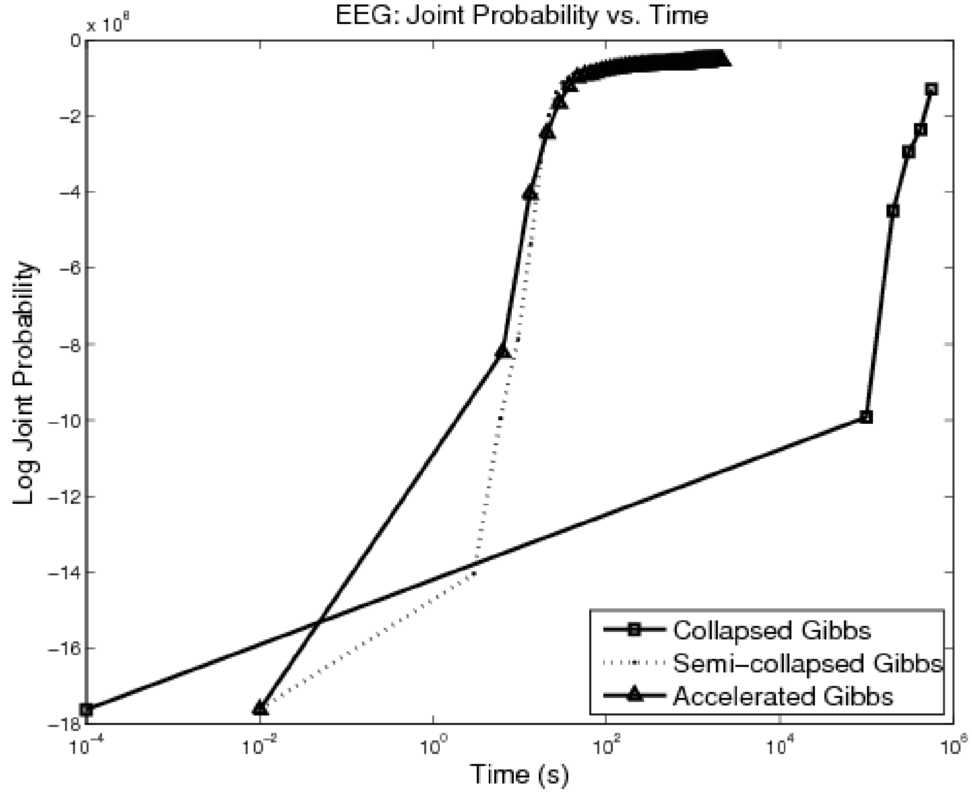
Effect of Window Size



Experiments on Real Data



EEG Dataset



EEG: Joint Probability vs. Time
x 10^6
0
−2
−4
−6
−8
−10
−12
−14
−16
−18
Log Joint Probability
10^-4
10^-2
10^0
10^2
10^4
10^6
Time (s)
Collapsed Gibbs
Semi−collapsed Gibbs
Accelerated Gibbs