15-780: Graduate AI
Lecture 17. Inference, Learning

Geoff Gordon (this lecture)
Tuomas Sandholm
TAs Sam Ganzfried, Byron Boots

Admin

- *HW3 back*
- *HW4 out*

Alexander Pope on the converse

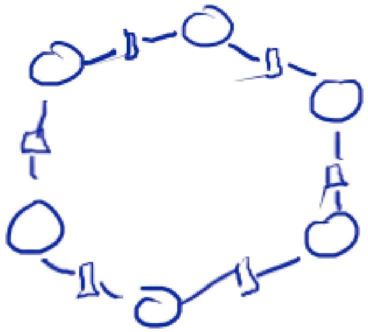
Sir, I admit your general rule
That every poet is a fool:
But you yourself may serve to show it,
That every fool is not a poet.

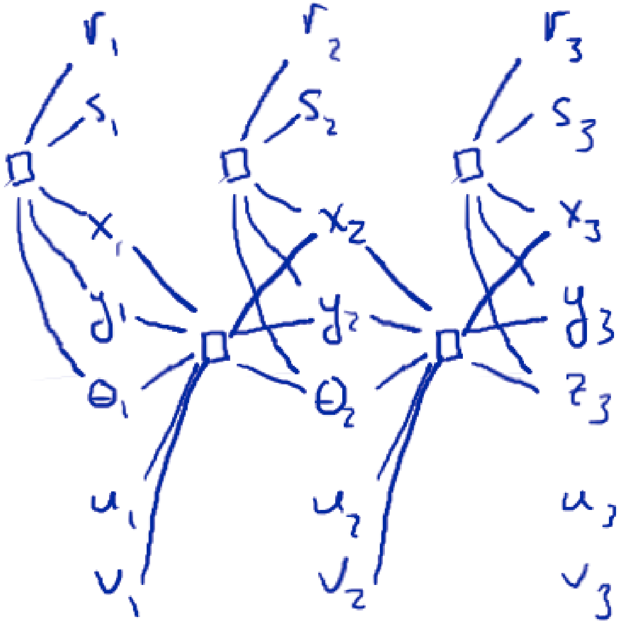
—Alexander Pope

Review

Factor graph (exact) inference

- *Variable elimination*
- *Treewidth*
- *HMMs and DBNs = factor graphs shaped like chains or parallel chains*
  - *forward-backward algorithm*





Conditional independence

- *Conditioning on a variable can break a factor graph into disconnected parts*
- *R.v.s in separate parts are* ***conditionally independent****—simplifies conditional inference*

Approximate inference

- *Uniform sampling*
- *Importance sampling*
  - *importance weights*
- *Parallel importance sampling*
  - *allows unnormalized importance $ZP(x)$*
  - *but biased (bias → 0 as samples → ∞)*

More probability

Expectation

- *Expectation* $E_P(f(X))$ *= the average value of* $f(X)$ *when* $x \sim P(X)$
- *Average: if each* $x_1, x_2, \ldots, x_N \sim P(X)$*,*

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \rightarrow E(f(X)) \quad \text{as} \quad N \rightarrow \infty$$

- *Will omit P when clear from context*

Expectation

- *Formula:*

$$
\begin{aligned}
E_P(f(\mathbf{X})) &= \sum_{\mathbf{x}} P(\mathbf{x})f(\mathbf{x}) \\
E_P(f(\mathbf{X})) &= \int P(\mathbf{x})f(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

Expectation example

15-780

HW4

	A	*A-*	*B+*
A	*.21*	*.17*	*.07*
A-	*.17*	*.15*	*.06*
B+	*.07*	*.06*	*.04*

probability

15-780

HW4

	A	*A-*	*B+*
A	*4*	*3.7*	*3.3*
A-	*4*	*3.7*	*3.3*
B+	*4*	*3.7*	*3.3*

f(HW4, 15-780)

Expectation example

		15-780		
		A	A-	B+
HW4	A	.21*4	.17*3.7	.07*3.3
	A-	.17*4	.15*3.7	.06*3.3
	B+	.07*4	.06*3.7	.04*3.3

$$\sum_{\mathbf{x}} P(\mathbf{x}) f(\mathbf{x})$$

$$= 3.767$$

Conditional expectation

- $E(f(X) \mid event) = E_{P(x \mid event)}(f(X))$
- *Expectation under conditional distribution*

Conditional expectation example

15-780

HW4

	A	A-	B+
A	.21	.17	.07
A-	.17	.15	.06
B+	.07	.06	.04

15-780

A	.41
A-	.35
B+	.24

f(X)

A	4.0
A-	3.7
B+	3.3

$E(f \mid HW4=B+) = 3.73$

Expecting expectations

- *Interpret:* $E(E(f(X) \mid Y))$
- ***Law of iterated expectations:***
  - $E(E(f(X) \mid Y)) = E(f(X))$
- *Proof is algebra on definition of* $E(\,)$*:*

Proof of iterated expectations

$$
\begin{aligned}
E(E(f(X) \mid Y)) &= E\left(\sum_x P(x \mid Y) f(x)\right) \\
&= \sum_y P(y) \sum_x P(x \mid y) f(x) \\
&= \sum_y \sum_x P(x, y) f(x) \\
&= E(f(X))
\end{aligned}
$$

Sample vs. population

- *P(X) describes "true" probability dist'n for a* ***population*** *(all realizations of X)*
- *If* $x_1 \sim P(X)$, $x_2 \sim P(X)$, …, $x_N \sim P(X)$, *all independent, the* x_i *are a* ***sample*** *from P(X)*
- *Finite N means that actual proportion of any outcome* $X=x$ *may differ from* $P(X=x)$
  - *E.g., flip 100 coins: may get 53 heads or 46 heads instead of 50*

Estimator

- *Want a **population** quantity, e.g., $E_P(f(X))$*
- *If we don't know P, or if P is hard to compute, population value is inaccessible*
- *But if we have a sample $x_1, \ldots, x_N \sim P(X)$, we can use (e.g.) the **estimator***

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

as a proxy for population value $E_P(f(X))$

Estimator

- *Estimator = r.v. M that tells us about a population value μ*
  - *often, M_N depends on sample $x_1, \ldots, x_N$*
- *Desirable property: **consistency***

$$M_N \to \mu \quad \text{as} \quad N \to \infty$$

- *E.g., sample mean is a consistent estimate of population expectation**

MCMC

Integration problem

- *Recall: wanted*

$$E(f(X)) = \int f(x)P(x)dx$$

- *And therefore, wanted good importance distribution Q(x)*

Back to high dimensions

- *Picking a good importance distribution is hard in high-D*
- *Major contributions to integral can be hidden in small areas*
    - *recall, want (P big ==> Q big)*
- *Would like to search for areas of high P(x)*
- *But searching could bias our estimates*

Markov-Chain Monte Carlo

- *Design a randomized search procedure M which tends to increase P(x) if it is small*
- *Run M for a while, take resulting x as a sample*
- *Importance distribution Q(x)?*

Markov-Chain Monte Carlo

- *Design a randomized search procedure M which tends to increase P(x) if it is small*
- *Run M for a while, take resulting x as a sample*
- *Importance distribution Q(x)?*
  Q = stationary distribution of M…

Stationary distribution

- *If* x_t *is a sample from Q(x), then* x_{t+1} *is also a sample from Q(x)*

$$
\begin{aligned}
Q(x_{t+1}) &= \mathbb{P}(x_{t+1}) \\
&= \int \mathbb{P}(x_{t+1}, x_t) dx_t \\
&= \int \mathbb{P}(x_{t+1} \mid x_t) \mathbb{P}(x_t) dx_t \\
&= \int \mathbb{P}(x_{t+1} \mid x_t) Q(x_t) dx_t
\end{aligned}
$$

Stationary distribution

- *If we run M a long time, eventually we won't* be able to tell where we started*
- *Limit is* ***stationary distribution*** *of M*
- *…which is why we use Q = stationary distribution in importance weight*

Designing a search chain

$$\int f(x)dx = \int P(x)g(x)dx = E_P(g(x))$$

- *Would like Q(x) = P(x)*
  - *makes importance weight = 1*
- *Turns out we can get this exactly, using* ***Metropolis-Hastings***

Metropolis-Hastings

- *Way of designing chain w/ $Q(x) = P(x)$*
- *Basic strategy: start from arbitrary x*
- *Repeatedly tweak x to get x'*
- *If $P(x') \geq P(x)$, move to x'*
- *If $P(x') << P(x)$, stay at x*
- *In intermediate cases, randomize*

Proposal distribution

- *Left open: what does “tweak” mean?*
- *Parameter of MH: $Q(x' \mid x)$*
  - *one-step proposal distribution*
- *Good proposals explore quickly, but remain in regions of high $P(x)$*
- *Optimal proposal?*

MH algorithm

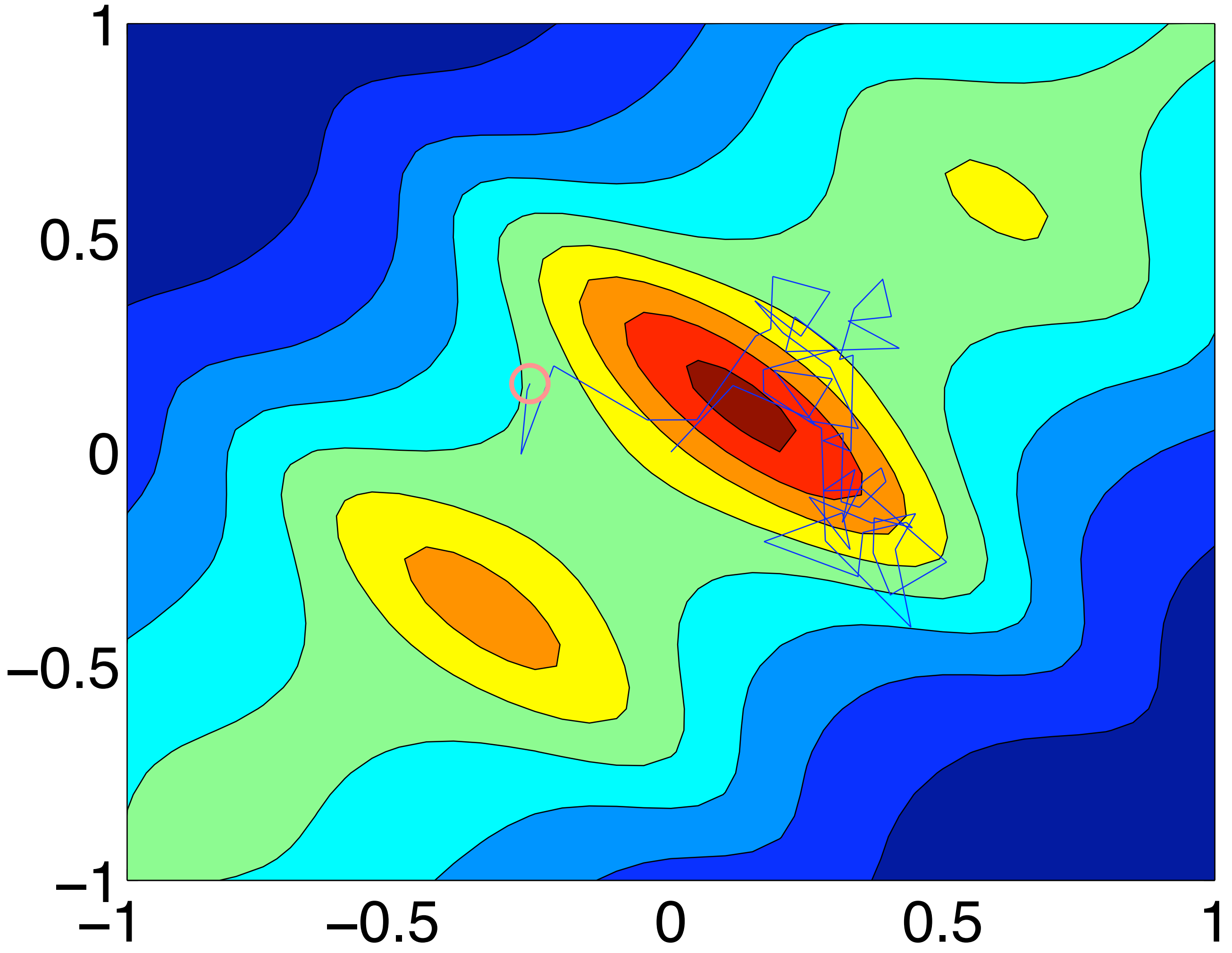
- *Sample* $x' \sim Q(x' \mid x)$
- *Compute* $p = \dfrac{P(x')}{P(x)} \dfrac{Q(x \mid x')}{Q(x' \mid x)}$
- *With probability* $\min(1, p)$*, set* $x := x'$
- *Repeat for T steps; sample is* $x_1, \ldots, x_T$ *(will usually contain duplicates)*

MH algorithm

note: we don't need to know Z

- *Sample x' ~ Q(x' | x)*
- *Compute p =* $\dfrac{P(x')}{P(x)}\dfrac{Q(x \mid x')}{Q(x' \mid x)}$
- *With probability min(1, p), set x := x'*
- *Repeat for T steps; sample is* $x_1, \ldots, x_T$
  (will usually contain duplicates)

MH example



Acceptance rate

- *Moving to new x' is* ***accepting***
- *Want* ***acceptance rate*** *(avg p) to be large, so we don't get big runs of the same x*
- *Want Q(x' | x) to move long distances (to explore quickly)*
- *Tension between Q and P(accept):*

$$p = \frac{P(x')}{P(x)} \frac{Q(x \mid x')}{Q(x' \mid x)}$$

Mixing rate, mixing time

- *If we pick a good proposal, we will move rapidly around domain of P(x)*
- *After a short time, won't be able to tell where we started*
- *This is short **mixing time** = # steps until we can't tell which starting point we used*
- ***Mixing rate** = 1 / (mixing time)*

MH estimate

- *Once we have our samples* $x_1, x_2, \ldots$
- *Optional: discard initial "burn-in" range*
  - *allows time to reach stationary dist'n*
- *Estimated integral:* $\frac{1}{N}\sum_{i=1}^{N} g(x_i)$

In example

- $g(x) = x^2$
- True $E(g(X)) = 0.28\ldots$
- Proposal: $Q(x' \mid x) = N(x' \mid x, 0.25^2 I)$
- Acceptance rate 55–60%
- After 1000 samples, minus burn-in of 100:

```
final estimate 0.282361
final estimate 0.271167
final estimate 0.322270
final estimate 0.306541
final estimate 0.308716
```

Annealing

- *MH acceptance probability:*

$$\min\left(1, \frac{P(x')}{P(x)} \frac{Q(x \mid x')}{Q(x' \mid x)}\right)$$

- *What if we use instead:*

$$\min\left(1, \frac{P(x')^{\beta}}{P(x)^{\beta}} \frac{Q(x \mid x')}{Q(x' \mid x)}\right)$$

*$T = 1/\beta$ is **temperature***

Effect of temperature

- *What if we had done MH with P' instead?*

$$P'(x) = P(x)^\beta / Z$$

- *Acceptance probability would be*

$$\min\left(1, \frac{P'(x')}{P'(x)} \frac{Q(x \mid x')}{Q(x' \mid x)}\right) = \min\left(1, \frac{P(x')^\beta}{P(x)^\beta} \frac{Q(x \mid x')}{Q(x' \mid x)}\right)$$

Effect of temperature

- *What happens to acceptance probability as T grows (β shrinks)?*

$$\frac{P(x')^\beta}{P(x)^\beta}$$

Annealing in MH

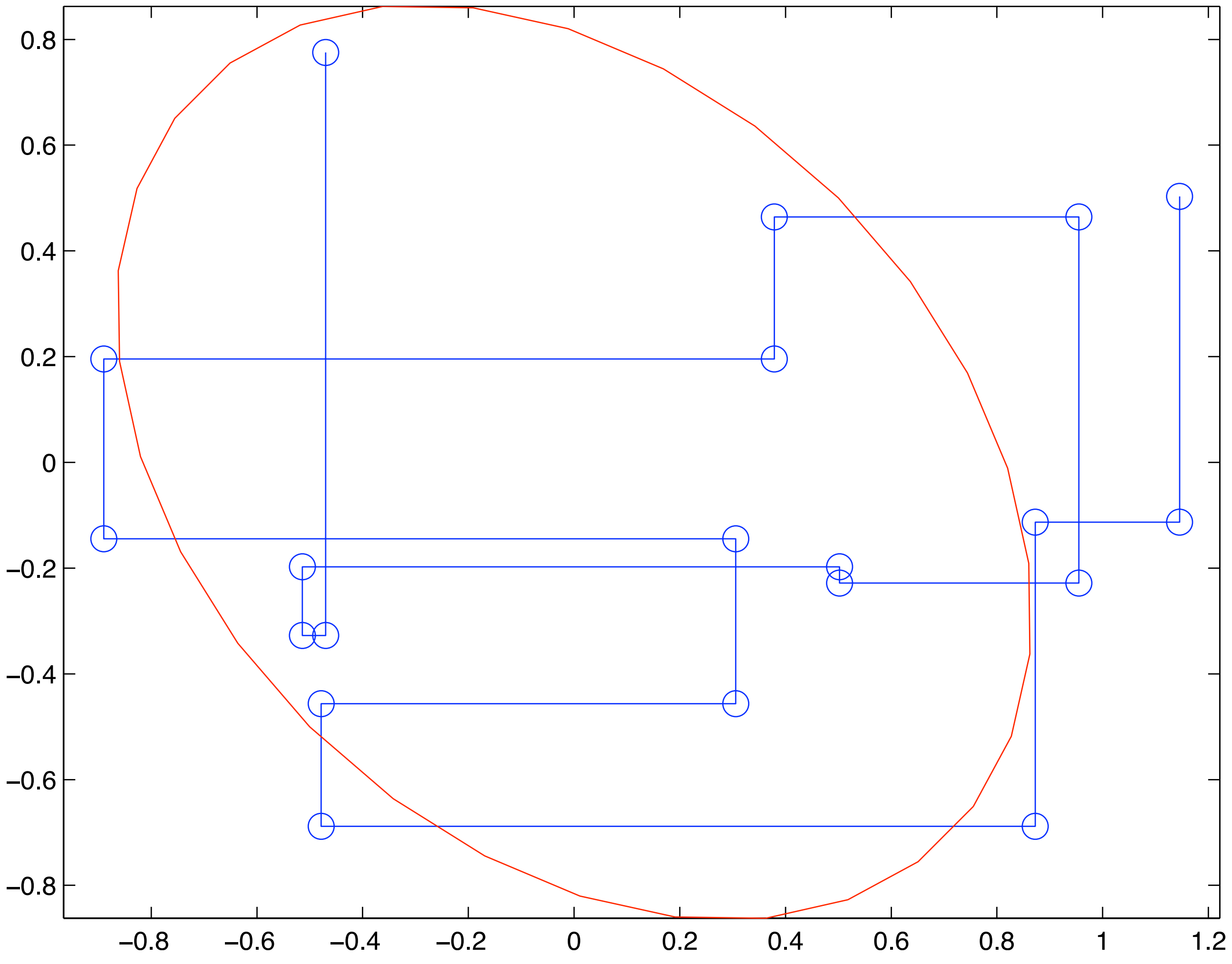
- *Start with T big (easy to sample)*
- *Let T shrink slowly until T = 1*
- *When T reaches 1, MH has already mixed!*

- *If we care about most likely x, we can let T shrink still further*

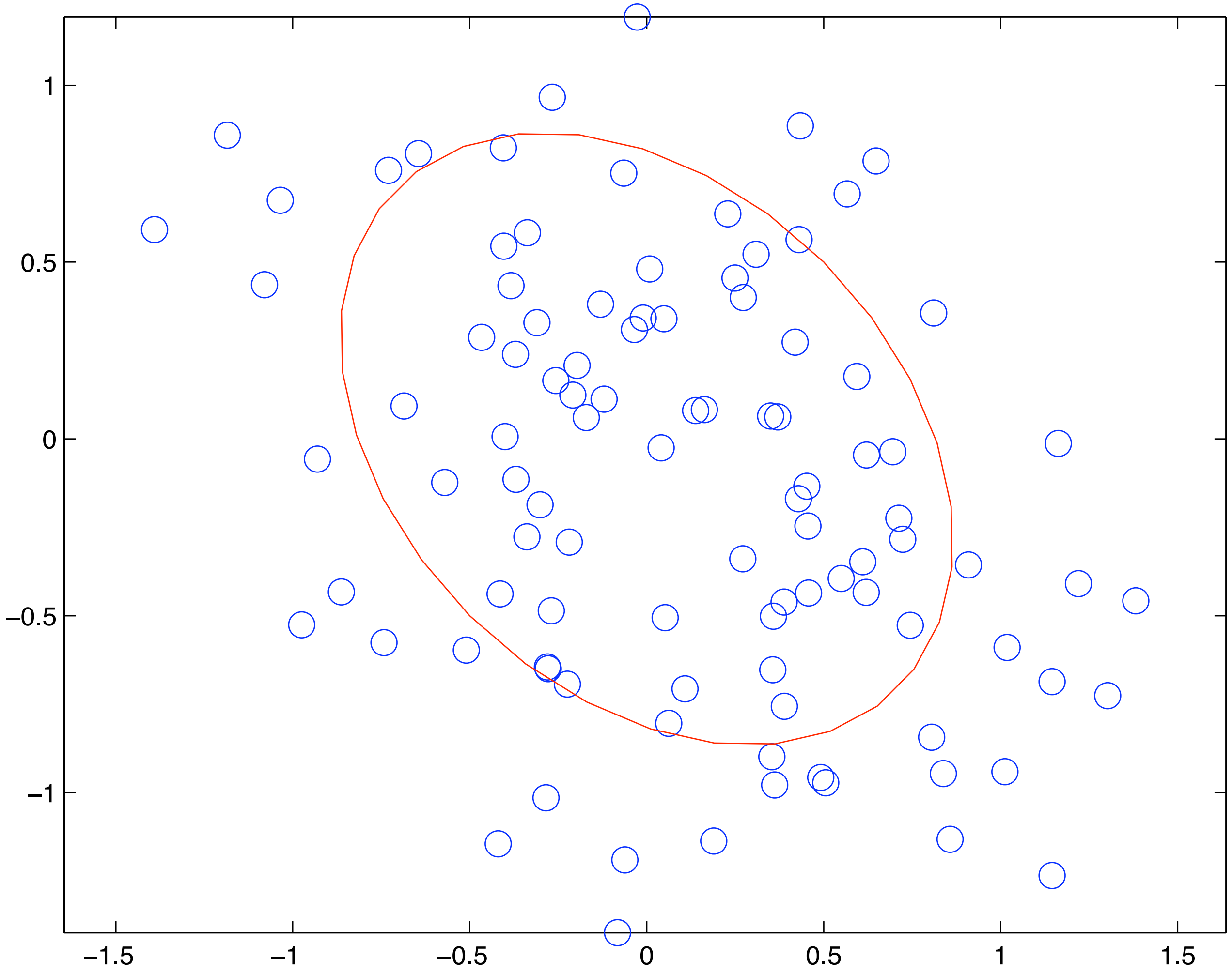
Gibbs sampler

- *Special case of MH*
- *Divide* $\boldsymbol{X}$ *into blocks of r.v.s B(1), B(2), …*
- *Proposal Q:*
  - *pick a block i uniformly*
  - *sample* $\boldsymbol{X}_{B(i)} \sim P(\boldsymbol{X}_{B(i)} \mid \boldsymbol{X}_{\neg B(i)})$

Gibbs example



Gibbs example



Why is Gibbs useful?

- $For\ Gibbs,\ p = \frac{P(x'_i, x'_{\neg i})}{P(x_i, x_{\neg i})} \frac{P(x_i \mid x'_{\neg i})}{P(x'_i \mid x_{\neg i})}$