Probabilistic Graphical Models

Lecture 16 – Sampling

CS/CNS/EE 155
Andreas Krause

Announcements

- Homework 3 due today

- Project poster session on Friday December 4 (tentative)
- Final writeup (8 pages NIPS format) due Dec 9

Approximate inference

- Three major classes of general-purpose approaches

- **Message passing**
  - E.g.: Loopy Belief Propagation (today!)

- **Inference as optimization**
  - Approximate posterior distribution by simple distribution
  - <u>Mean field / structured mean field</u>
  - Assumed density filtering / expectation propagation

- **Sampling based inference**
  - Importance sampling, particle filtering
  - Gibbs sampling, MCMC

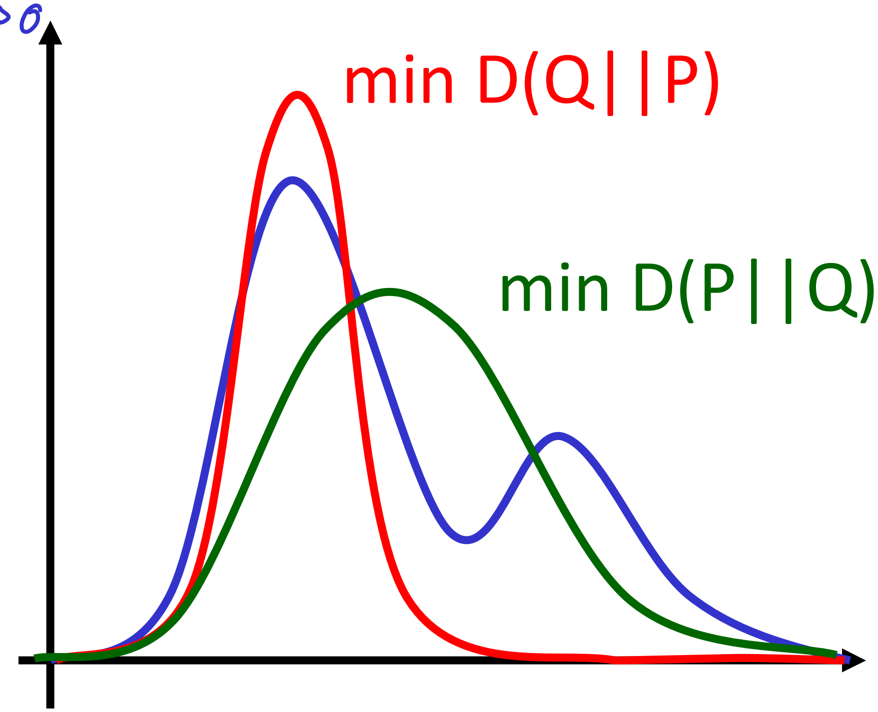
- Many other alternatives (often for special cases)

Variational approximation

- **Key idea**: Approximate posterior with simpler distribution that's as close as possible to P
  - What is a "simple" distribution?
  - What does "as close as possible" mean?
- **Simple** = efficient inference
  - Typically: factorized (fully independent, chain, tree, …)
  - Gaussian approximation
- **As close as possible** = KL divergence

Finding simple approximate distributions

- KL divergence not symmetric; need to choose directions
- P: true distribution; Q: our approximation
- D(P || Q) $P(x) > 0 \Rightarrow Q(x) > 0$
  - The "right" way
  - Often intractable to compute
  - Assumed Density Filtering
- D(Q || P) $P(x) = 0 \Rightarrow Q(x) = 0$
  - The "reverse" way
  - Underestimates support (overconfident)
  - Mean field approximation
- Both special cases of α-divergence



Approximate inference

- Three major classes of general-purpose approaches

- **Message passing**
  - E.g.: Loopy Belief Propagation (today!)

- **Inference as optimization**
  - Approximate posterior distribution by simple distribution
  - Mean field / structured mean field
  - Assumed density filtering / expectation propagation

- **Sampling based inference**
  - Importance sampling, particle filtering
  - Gibbs sampling, MCMC

- Many other alternatives (often for special cases)

Sampling based inference

- So far: deterministic inference techniques
  - Loopy belief propagation
  - (Structured) mean field approximation
  - Assumed density filtering

- Will now introduce stochastic approximations
  - Algorithms that "randomize" to compute expectations
  - In contrast to the deterministic methods, can sometimes get approximation guarantees
  - More exact, but slower than deterministic variants

Computing expectations

- Often, we're not necessarily interested in computing marginal distributions, but certain expectations:

- Moments (mean, variance, …)

$$\mathbb{E}_P[X^k] = \int x^k P(x)dx$$

- Event probabilities

$$P(X > c) = \mathbb{E}_P[I_{X>c}] = \int [x > c] P(x)dx$$

Sample approximations of expectations

- $x_1,\ldots,x_N$ samples from RV X
- Law of large numbers:

$$\mathbb{E}_P[f(X)] = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- Hereby, the convergence is with probability 1 (almost sure convergence)
- Finite samples: $\mathbb{E}_P[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$

How many samples do we need?

- Hoeffding inequality
  Suppose f is bounded in [0,C]. Then

$$P\left(\left|\mathbb{E}_P[f(X)] - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right| > \varepsilon\right) \leq \underbrace{2\exp(-2N\varepsilon^2/C^2)}_{\text{error probability}}$$

- Thus, probability of error decreases exponentially in N!

want error ε with probability $1-\delta$

$$2\exp(-2N\varepsilon^2/C^2) < \delta$$

$$2N\varepsilon^2/C^2 > \log\frac{2}{\delta}$$

$$N > \frac{1}{2}\,\frac{1}{\varepsilon^2}\cdot C^2\cdot\log\frac{2}{\delta}$$

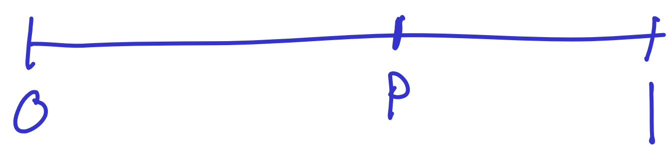
- Need to be able to draw samples from P

Sampling from a Bernoulli distribution

- X ~ Bernoulli(p)

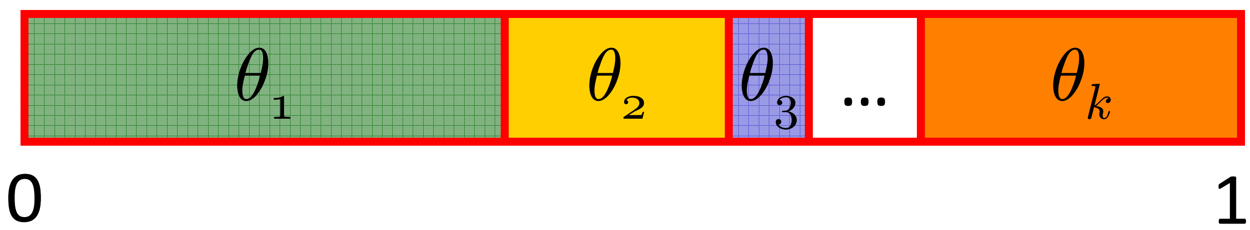
- How can we draw samples from X?

Assume we can draw uniform distribution [0,1]



Sampling from a Multinomial

- X ~ Mult($[\theta_1,\ldots,\theta_k]$)
  where θ_i = P(X=i); $\sum_i \theta_i = 1$

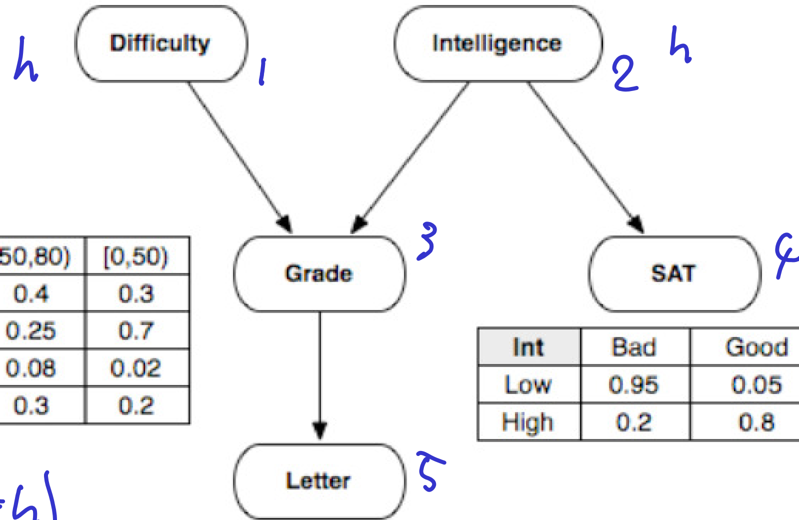


- Function g: [0,1]→{1,…,k} assigns state g(x) to each x
- Draw sample from uniform distribution on [0,1]
- Return $g^{-1}(x)$

Forward sampling from a BN

Easy	Hard
0.6	0.4

Low	High
0.7	0.3



← $P(S \mid I=h)$

Int	Diff	[80,100]	[50,80)	[0,50)
Low	Easy	0.3	0.4	0.3
Low	Hard	0.05	0.25	0.7
High	Easy	0.9	0.08	0.02
High	Hard	0.5	0.3	0.2

Int	Bad	Good
Low	0.95	0.05
High	0.2	0.8

g

$P(G \mid D=h, I=h)$

[80,100]

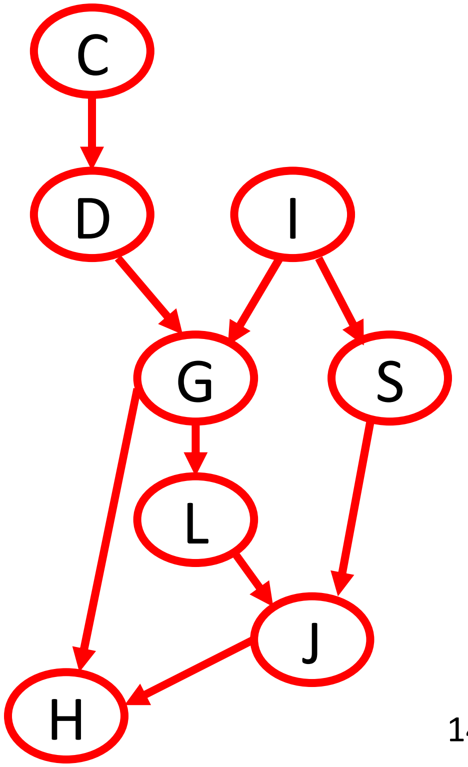
Grade	Fail	Pass
[80,100]	0.1	0.9
[50,80)	0.4	0.6
[0,50)	0.99	0.01

$P(L \mid \ldots)$

Monte Carlo sampling from a BN

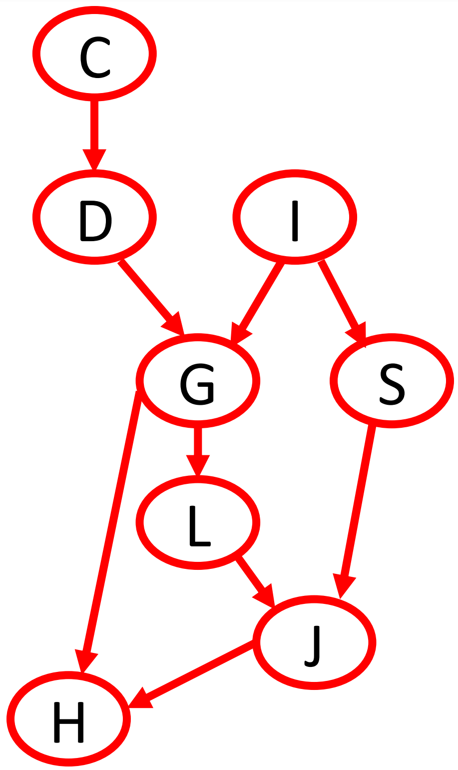
- Sort variables in topological ordering $X_1,\ldots,X_n$
- For i = 1 to n do
  - Sample $x_i \sim P(X_i \mid X_1=x_1, \ldots, X_{i-1}=x_{i-1})$ $= P(X_i \mid Pa_{X_i})$

- Works even with high-treewidth models!



Computing probabilities through sampling

- Want to estimate probabilities
- Draw N samples from BN



- Marginals

$$P(H=y) = \mathbb{E}_P\left[I_{H=y}\right] = \sum_x P(x)\cdot \underbrace{I_{H=y}(x)}_{=1 \text{ iff } x_H = y}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} I_{H=y}(x^{(i)}) = \frac{\text{Count}(H=y)}{N}$$

- Conditionals

$$P(D=h \mid H=m) = \frac{P(D=h, H=m)}{P(H=m)} = \frac{\text{Count}(D=h, H=m)}{\text{Count}(H=m)}$$

Rejection sampling

Rejection sampling

- Collect samples over all variables

$$\widehat{P}(\mathbf{X}_A = \mathbf{x}_A \mid \mathbf{X}_B = \mathbf{x}_B) \approx \frac{Count(\mathbf{x}_A, \mathbf{x}_B)}{Count(\mathbf{x}_B)}$$

- Throw away samples that disagree with $\mathbf{x}_\mathbf{B}$
- Can be problematic if P($\mathbf{X}_\mathbf{B}$ = $\mathbf{x}_\mathbf{B}$) is rare event

Sample complexity for probability estimates

- Absolute error:

$$Prob\Big(\big|\widehat{P}(\mathbf{x}) - P(\mathbf{x})\big| > \varepsilon\Big) \leq 2\exp(-2N\varepsilon^2)$$

- Relative error:

$$Prob\Big(\widehat{P}(\mathbf{x}) < (1+\varepsilon)P(\mathbf{x})\Big) \leq 2\exp(-N\underline{P(\mathbf{x})}\varepsilon^2/3)$$

Estimating low probability events is hard

Sampling from rare events

- Estimating conditional probabilities $P(X_A \mid \mathbf{X_B}=\mathbf{x_B})$ using rejection sampling is hard!
  - The more observations, the unlikelier $P(\mathbf{X_B} = \mathbf{x_B})$ becomes

- Want to directly sample from posterior distribution!

Sampling from intractable distributions

$P(X_A \mid X_B = x_B) \propto P(X_A, X_B = x_B)$

- Given unnormalized distribution

  $P(X) \propto Q(X) = P(X, X_{obs} = x_{obs})$

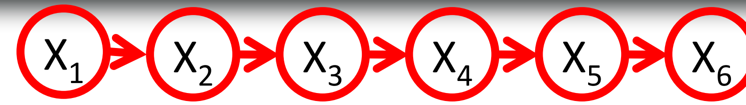
- Q(X) efficient to evaluate, but normalizer intractable
- For example, $Q(X) = \prod_j \Psi(C_j)$

- Want to sample from P(X) $= \frac{1}{Z} Q(x)$

- **Ingenious idea**:

  Can create Markov chain that is efficient to simulate and that has stationary distribution P(X)

Markov Chains



- A Markov chain is a sequence of RVs, $X_1,\ldots,X_N,\ldots$ with
  - Prior $P(X_1)$
  - Transition probabilities $P(X_{t+1} \mid X_t)$

- A Markov Chain with $P(X_{t+1} \mid X_t)>0$ has a unique **stationary distribution** $\mu(X)$, such that for all x

$$\lim_{N\to\infty} P(X_N=x) = \mu(x)$$

The stationary distribution is independent of $P(X_1)$

Simulating a Markov Chain

- Can sample from a Markov chain as from a BN:

- Sample $x_1 \sim P(X_1)$
- Sample $x_2 \sim P(X_2 \mid X_1 = x_1)$
- …
- Sample $x_N \sim P(X_N \mid X_{N-1} = x_{N-1})$
- …

- If simulated "sufficiently long", sample X_N is drawn from a distribution "very close" to stationary distribution μ

Markov Chain Monte Carlo

- Given an unnormalized distribution Q(x)
- Want to design a Markov chain with stationary distribution

  $\pi(x) = 1/Z\ Q(x)$

- Need to specify transition probabilities P(x | x')!

Detailed balance equation

- A Markov Chain satisfies the **detailed balance equation** for unnormalized distribution Q if for all x, x':

  Q(x) P(x'|x) = Q(x') P(x | x')

- In this case, the Markov chain has stationary distribution 1/Z Q(x)

$$\underbrace{\frac{1}{Z} Q(x)}_{\mu(x)} = \frac{1}{Z}\sum_{x'} P(x'|x)\, Q(x) = \frac{1}{Z}\sum_{x'} Q(x')\, P(x|x')$$
$$= \sum_{x'} \mu(x')\, P(x|x')$$

Designing Markov Chains

1) Proposal distribution R(X' | X)

- Given X_t = x, sample "proposal" x'~R(X' | X=x)
- Performance of algorithm will strongly depend on R

2) Acceptance distribution:

- Suppose X_t = x
- With probability set X_{t+1} = x'

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

- With probability 1-α, set X_{t+1} = x

Theorem [Metropolis, Hastings]: The stationary distribution is Z^{-1} Q(x)

- Proof: Markov chain satisfies detailed balance condition!

MCMC for Graphical Models

- Random vector $X=(X_1,\ldots,X_n)$ is high-dimensional
- Need to specify proposal distributions $R(x' \mid x)$ over such random vectors
  - x~~’~~: old state
  - x’: proposed state, $x' \sim R(X' \mid X=x)$
- Examples

– $R(X' \mid \cancel{X=x}) = R(X')$

– $R(x' \mid x)$ $\quad x'_i \sim R_i(x'_i \mid x)$

$x'_{\neg i} = x_{\neg i}$

Gibbs sampling

- Start with initial assignment $\mathbf{x}^{(0)}$ to all variables
- For $t = 1$ to ∞ do
  - Set $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)}$
  - For each variable X_i
    - Set $\mathbf{v}_i$ = values of all $\mathbf{x}^{(t)}$ except x_i
    - Sample $x^{(t)}_i$ from $P(X_i \mid \mathbf{v}_i)$

- Gibbs sampling satisfies detailed balance equation for P
- **Key challenge**: Computing conditional distributions $P(X_i \mid \mathbf{v}_i)$

Computing $P(X_i \mid \mathbf{v}_i)$

$$Q(x) = \prod_i \psi_i(C_i)$$

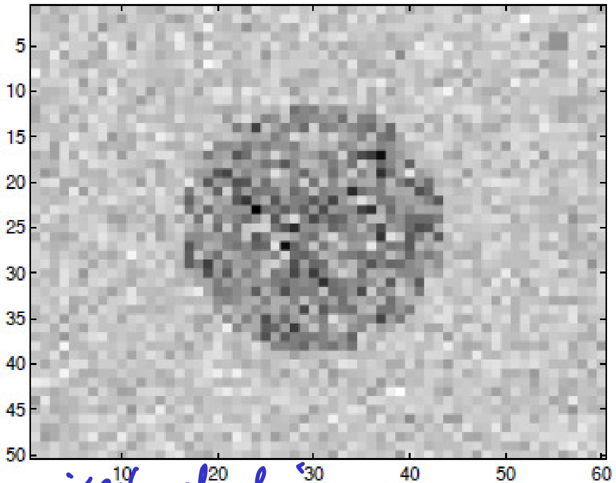
$$P(X_i \mid X_1 \ldots X_{i-1}, X_{i+1}, \ldots X_n) = \frac{P(X_1 \ldots X_n)}{P(X_1 \ldots X_{i-1}, X_{i+1} \ldots X_n)}$$

$$= \frac{\cancel{\tfrac{1}{Z}} Q(x_1 \ldots x_n)}{\sum_{x_i} \cancel{\tfrac{1}{Z}} Q(x_1 \ldots x_n)} = \frac{\prod_j \psi_j(C_j)}{\sum_{x_i} \prod_j \psi_j(C_j)}$$

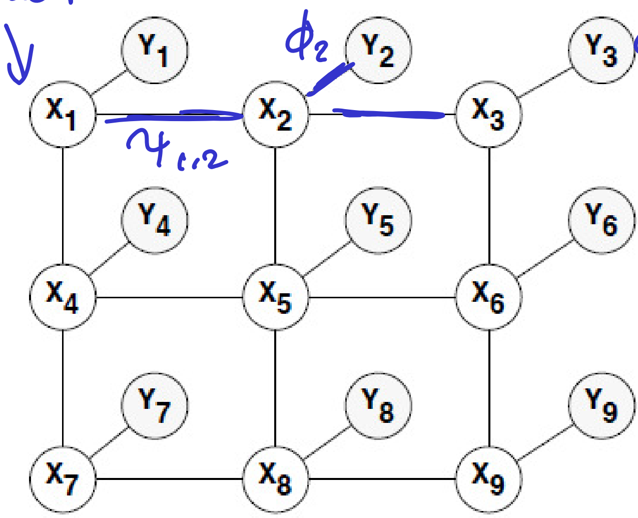
$$= \frac{\prod_{j \in N(i)} \psi_j(C_j)}{\sum_{x_i} \prod_{j \in N(i)} \psi_j(C_j)}$$

all factors that contain X_i

Example: (Simple) image segmentation



True pixel val of i



noisy obs.

$$P(x) = \frac{1}{Z} \prod_i \Phi_i(x_i) \prod_{(j,k)\in E} \Psi_{jk}(x_j, x_k)$$

$$\Phi(x_i) = exp\left\{-\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2}\right\}$$

μ_{x_i} = mean for true pix. val x_i

$$\Psi(x_i, x_j) = \exp\left\{-\beta(x_i - x_j)^2\right\}$$

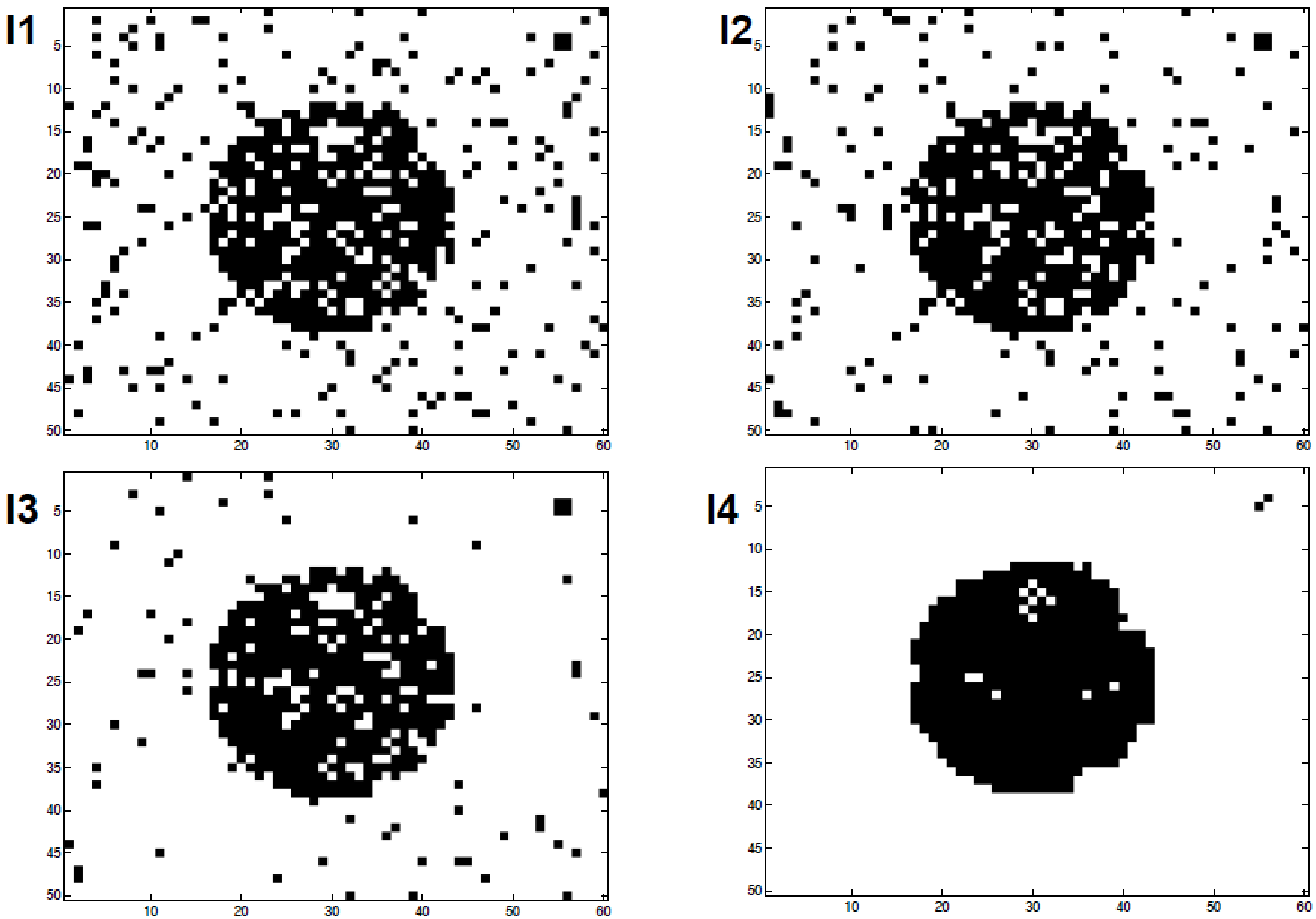
Gibbs sampling:

$$P(x_i \mid x_1 \ldots x_{i-1}, x_{i+1}, \ldots x_n) =$$

$$= \frac{\phi_i(x_i) \prod_{j\in N(i)} \psi_{ij}(x_i, x_j)}{\sum_{x_i} \phi_i(x_i) \prod_{j\in N(i)} \psi_{ij}(x_i, x_j)}$$

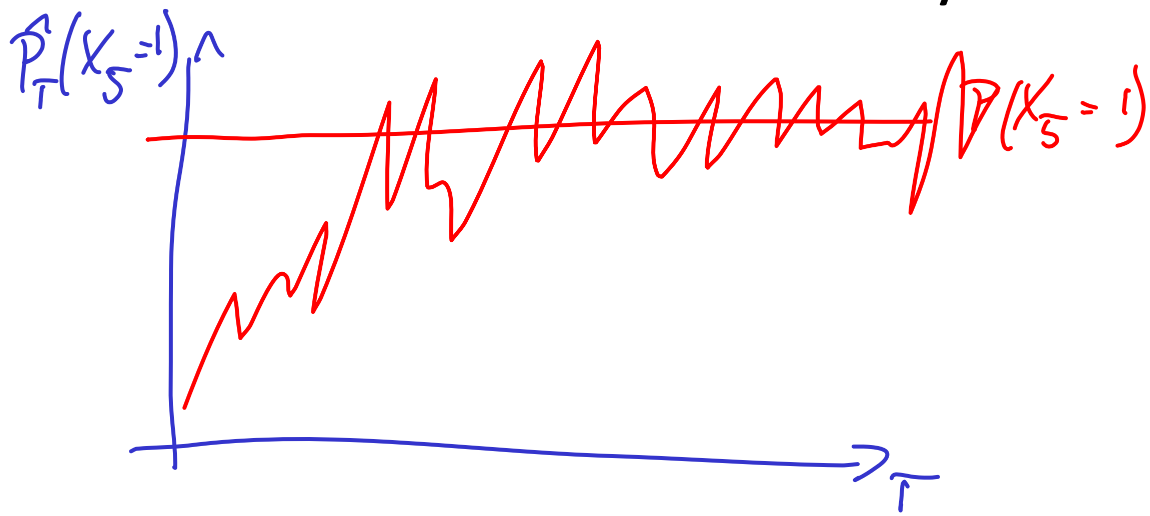
[see Singh '08]

Gibbs Sampling iterations



Convergence of Gibbs Sampling

- When are we close to stationary distribution?



Summary of Sampling

- Randomized approximate inference for computing expections, (conditional) probabilities, etc.
- Exact in the limit
  - But may need ridiculously many samples
- Can even directly sample from intractable distributions
  - Disguise distribution as stationary distribution of Markov Chain
  - Famous example: Gibbs sampling