10-418 / 10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Monte Carlo Methods

Matt Gormley
Lecture 17
Oct. 23, 2019

Q&A

Q: Is this ILP for MAP inference from Lecture 13 correct?

ILP: Goal: $\hat{y} = \underset{y}{\operatorname{argmax}} \log p(\hat{y})$

variables

$$\max_{\vec{y}} \left[\sum_{t=1}^{T} y_t \phi_{t,y_t} + (1-y_t)\phi_{t,(1-y_t)} \right] + \sum_{(s,t)\in E} y_s y_t \phi_{s,t,y_s,y_t} + (1-y_s) y_t \phi_{s,t,(1-y_s),y_t} + (1-y_t) y_s \phi_{s,t,y_s,(1-y_t)} + (1-y_s)(1-y_t)\phi_{s,t,(1-y_s),(1-y_t)}$$

s.t. $y_t \in \{0,1\} \;\forall t$

A: No! The indexing here is incorrect. It should be…

Reminders

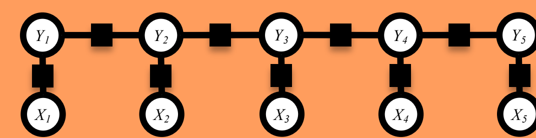
- Homework 3: Structured SVM
  - Out: Tue, Oct. 18
  - Due: Mon, Nov. 4 at 11:59pm
- Midterm Exam Viewing
- Project Milestones

1. Data

$$\mathcal{D} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$$

2. Model

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$



3. Objective

$$\ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})$$

5. Inference

1. Marginal Inference

$$p(\boldsymbol{x}_C) = \sum_{\boldsymbol{x}' : \boldsymbol{x}'_C = \boldsymbol{x}_C} p(\boldsymbol{x}' \mid \boldsymbol{\theta})$$

2. Partition Function

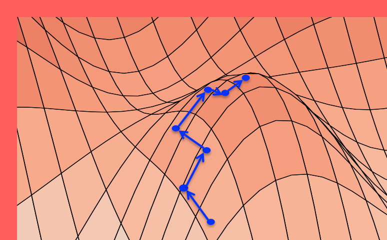
$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$

3. MAP Inference

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\ p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

4. Learning

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; \mathcal{D})$$



A Few Problems for a Factor Graph

Suppose we already have the parameters of a Factor Graph…

1. How do we compute the probability of a specific assignment to the variables?
   $P(T=t, H=h, A=a, C=c)$

2. How do we draw a sample from the joint distribution?
   $t,h,a,c \sim P(T, H, A, C)$

3. How do we compute marginal probabilities?
   $P(A) = \ldots$

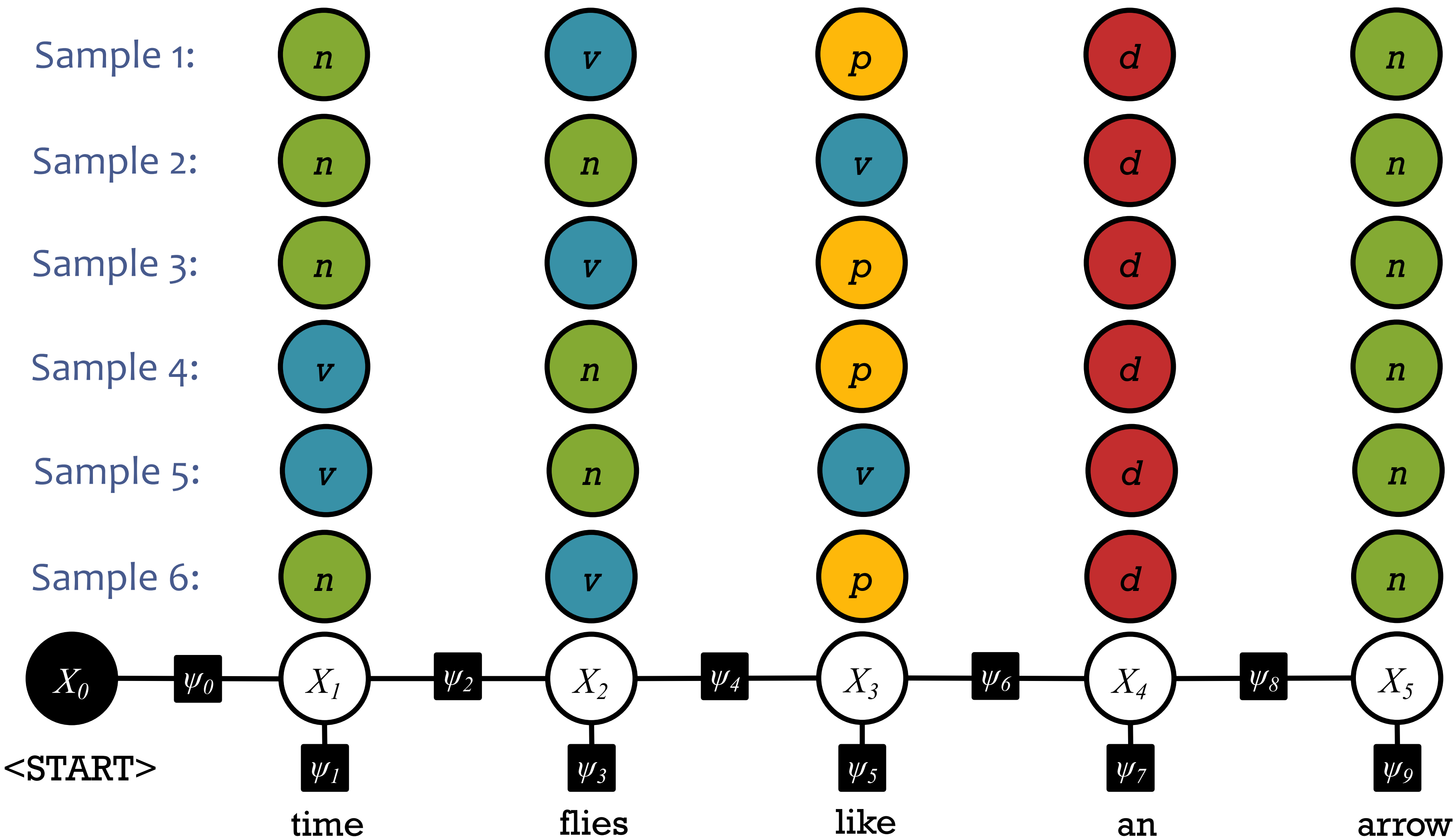
4. How do we draw samples from a conditional distribution?
   $t,h,a \sim P(T, H, A \mid C = c)$

5. How do we compute conditional marginal probabilities?
   $P(H \mid C = c) = \ldots$

Can we use samples ?

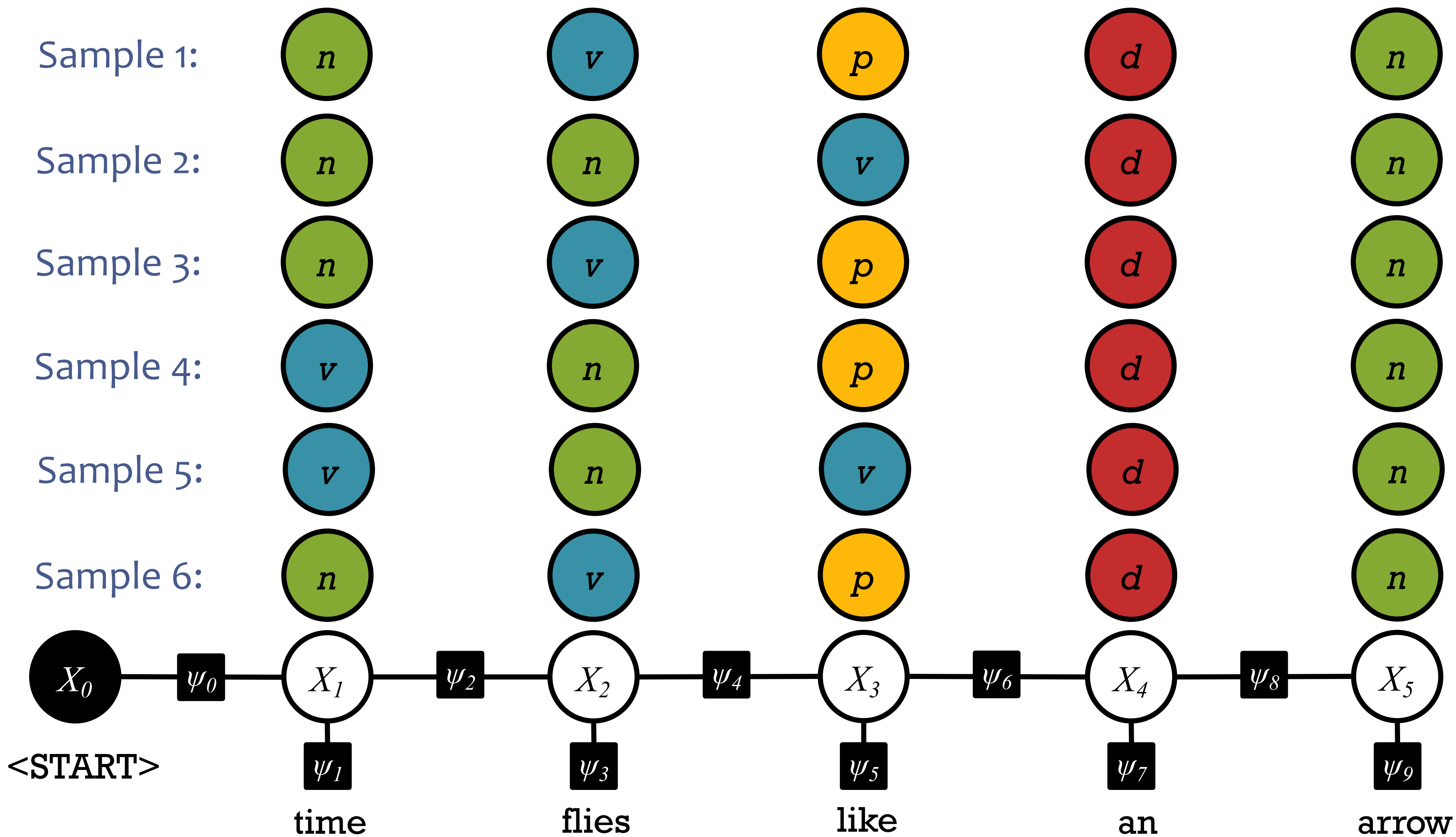
Marginals by Sampling on Factor Graph

Suppose we took many samples from the distribution over taggings: $p(\boldsymbol{x}) = \frac{1}{Z} \prod_{\alpha} \psi_\alpha(\boldsymbol{x}_\alpha)$

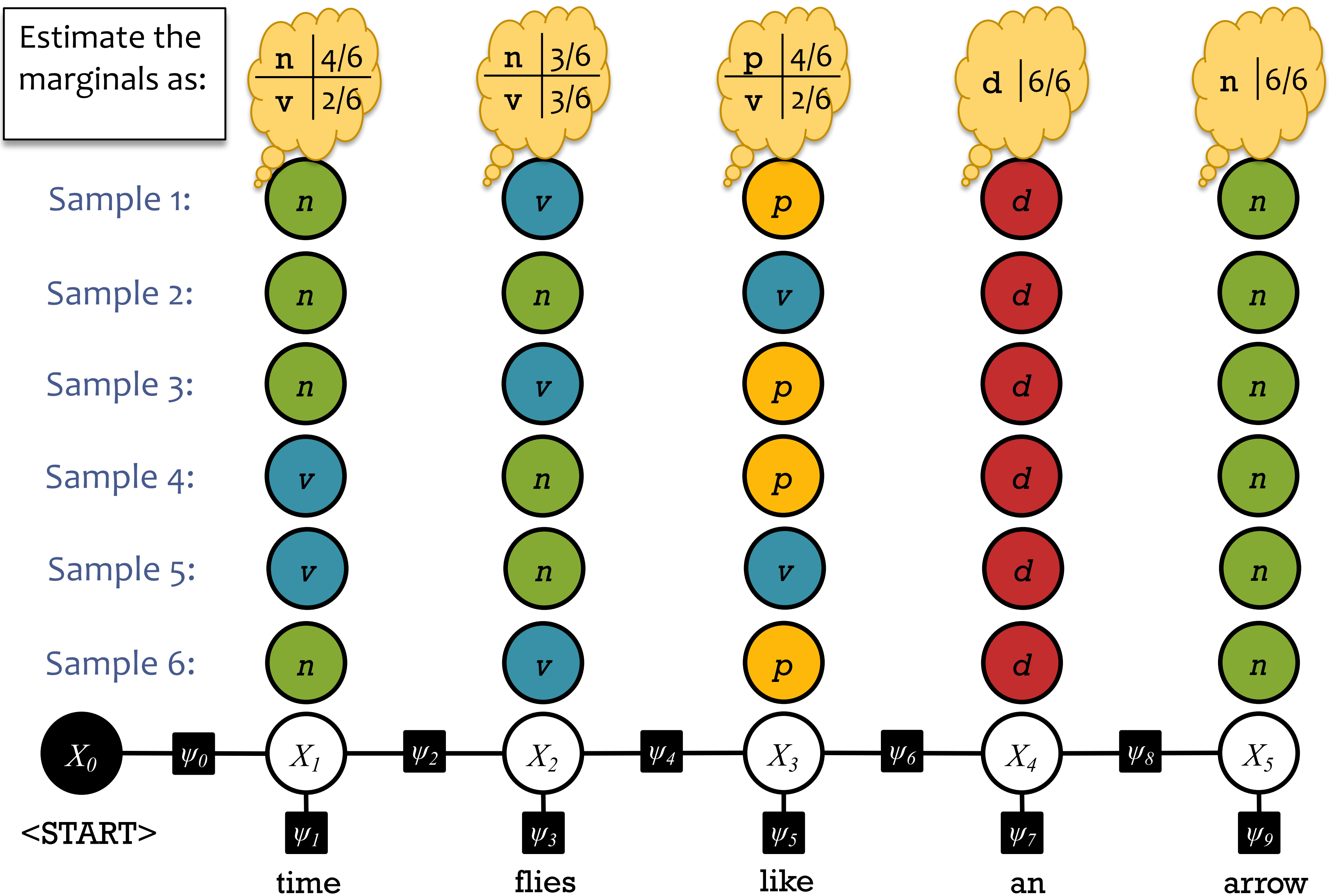


Marginals by Sampling on Factor Graph

The marginal $p(X_i = x_i)$ gives the probability that variable X_i takes value x_i in a random sample



Marginals by Sampling on Factor Graph



MONTE CARLO METHODS

Monte Carlo Methods

Whiteboard

- Problem 1: Generating samples from a distribution
- Problem 2: Estimating expectations
- Why is sampling from p(x) hard?
- Example: estimating plankton concentration in a lake
- Algorithm: Uniform Sampling
- Example: estimating partition function of high dimensional function

Properties of Monte Carlo

Estimator: $$\int f(x)P(x)\,\mathrm{d}x \;\approx\; \hat{f} \;\equiv\; \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

Estimator is unbiased:

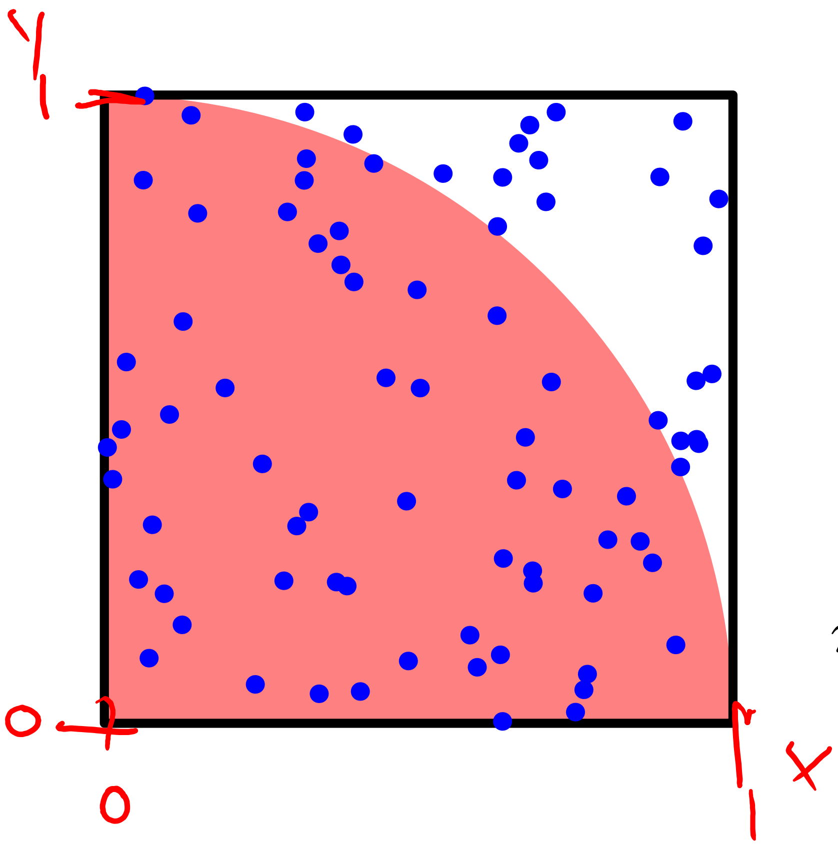
$$\mathbb{E}_{\underbrace{P(\{x^{(s)}\})}_{\text{sample dist.}}}\left[\hat{f}\right] \;=\; \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{P(x)}[f(x)] \;=\; \mathbb{E}_{P(x)}[f(x)]$$

Variance shrinks $\propto 1/S$:

$$\mathrm{var}_{P(\{x^{(s)}\})}\left[\hat{f}\right] \;=\; \frac{1}{S^2}\sum_{s=1}^{S}\mathrm{var}_{P(x)}[f(x)] \;=\; \mathrm{var}_{P(x)}[f(x)]\,/S$$

"Error bars" shrink like $\sqrt{S}$

A dumb approximation of π



$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \underbrace{\mathbb{I}\left((x^2+y^2)<1\right)}_{f(x,y)} P(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

```
octave:1> S=12; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> S=1e7; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

Aside: don't always sample!

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

Example: numerical solutions to (nice) 1D integrals are fast

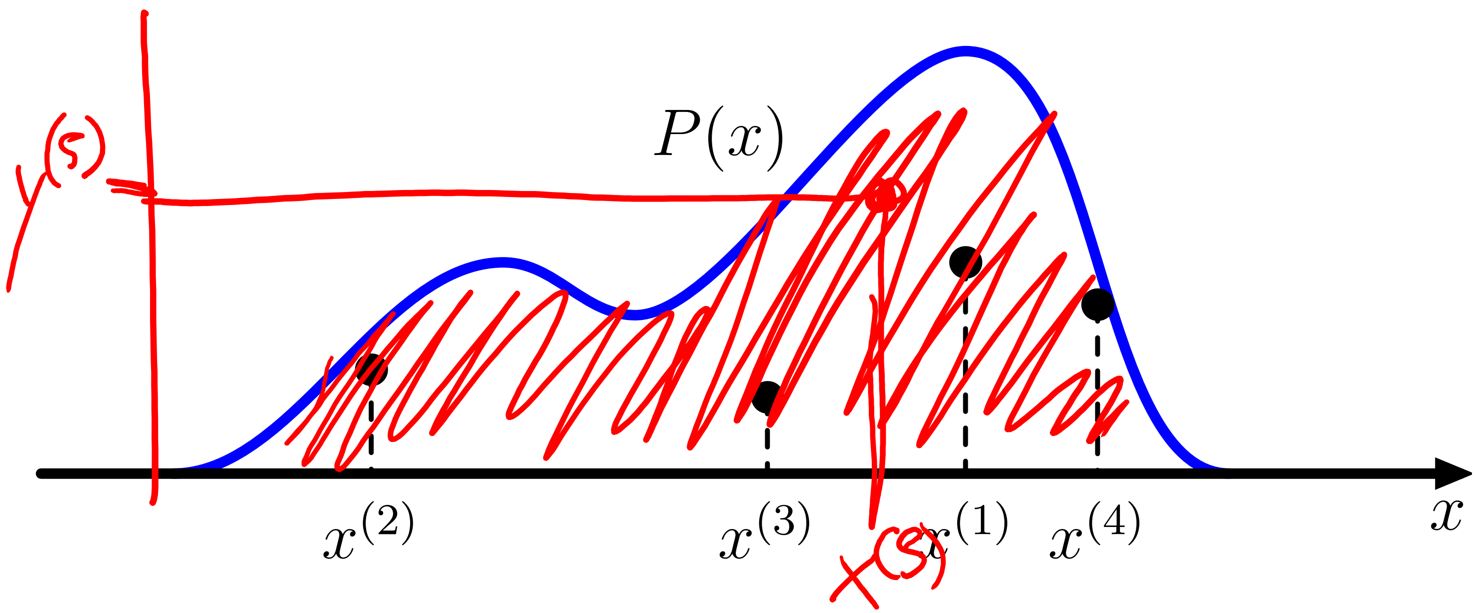
```
octave:1> 4 * quadl(@(x) sqrt(1-x.^2), 0, 1, tolerance)
```

Gives π to 6 dp's in 108 evaluations, machine precision in 2598.

(NB Matlab's `quadl` fails at zero tolerance)

Sampling from distributions

Draw points uniformly under the curve:



Probability mass to left of point $\sim$ Uniform[0,1]

Slide from Ian Murray

Sampling from distributions

How to convert samples from a Uniform[0,1] generator:

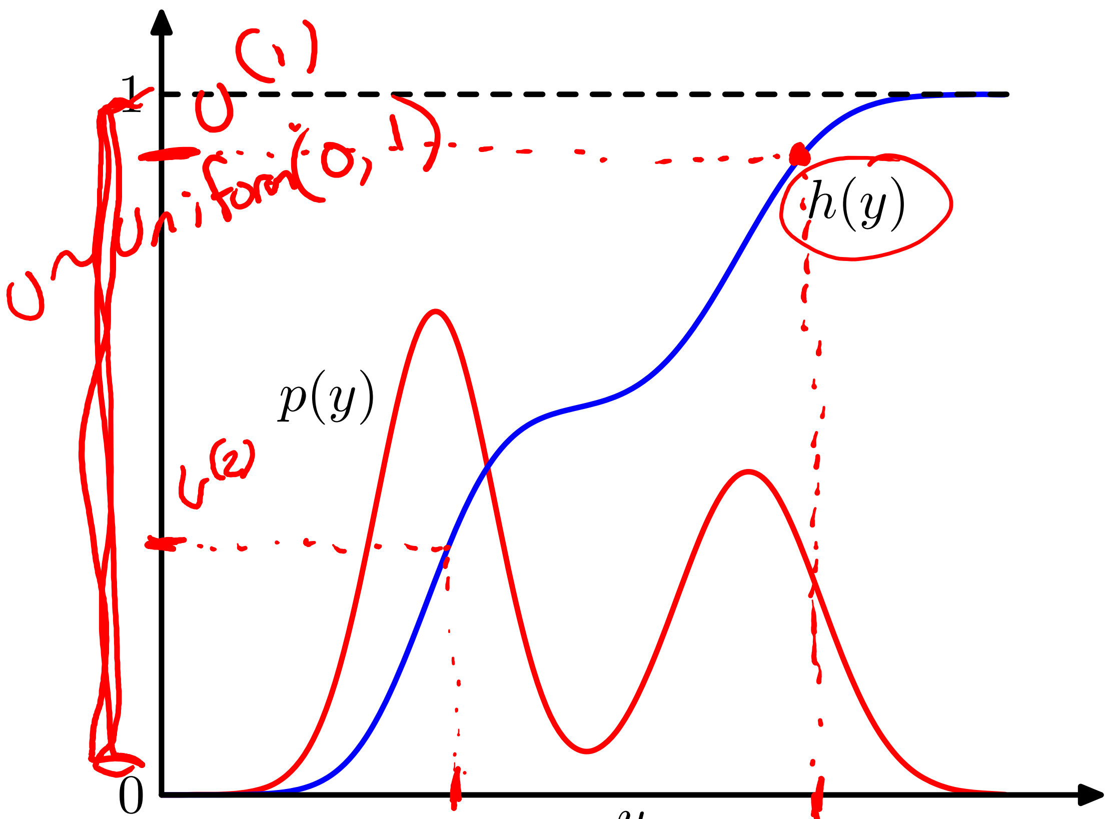


Figure from PRML, Bishop (2006)

CDF

$h(y) = \int_{-\infty}^{y} p(y')\,\mathrm{d}y'$

Draw mass to left of point:
$u \sim$ Uniform[0,1]

Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

Slide from Ian Murray

Rejection sampling

★ don't know Z

can compute

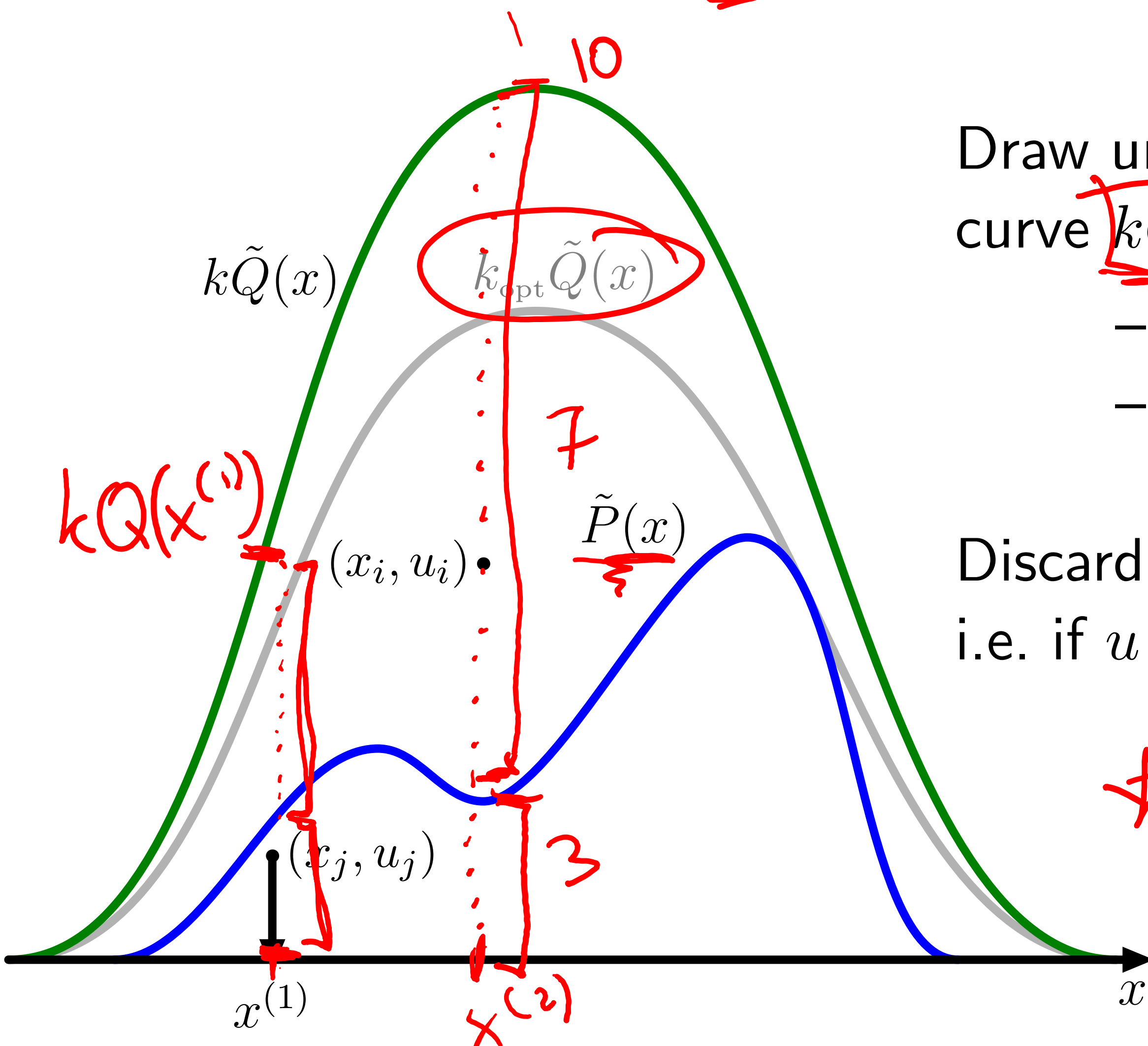
cannot compute

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid

Proposal Distribution

e.g. Gaussian

easy to sample from



Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$: $\forall x$

– Draw $x \sim Q(x)$

– height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

★ Samples from P(x)

Importance sampling

Computing $\tilde{P}(x)$ and $\tilde{Q}(x)$, then *throwing* x *away* seems wasteful
Instead rewrite the integral as an expectation under Q:

$\cdot 1 = \frac{Q(x)}{Q(x)}$

$$E_p\{f(x)\} = \int f(x) P(x)\, \mathrm{d}x = \int f(x) \frac{P(x)}{Q(x)} Q(x)\, \mathrm{d}x, \qquad (Q(x) > 0 \text{ if } P(x) > 0)$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}) \boxed{\frac{P(x^{(s)})}{Q(x^{(s)})}}, \qquad x^{(s)} \sim Q(x)$$

e.g. 3/10

This is just simple Monte Carlo again, so it is unbiased.

Importance sampling applies when the integral is not an expectation.
Divide and multiply any integrand by a convenient distribution.

Importance sampling (2)

Previous slide assumed we could evaluate $P(x) = \tilde{P}(x)/\mathcal{Z}_P$

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{\mathcal{Z}_Q}{\mathcal{Z}_P}\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\underbrace{\frac{\tilde{P}(x^{(s)})}{\tilde{Q}(x^{(s)})}}_{\tilde{r}^{(s)}}, \quad x^{(s)} \sim Q(x)$$

$$\approx \frac{\cancel{1}}{\cancel{S}}\sum_{s=1}^{S} f(x^{(s)})\frac{\tilde{r}^{(s)}}{\frac{\cancel{1}}{\cancel{S}}\sum_{s'}\tilde{r}^{(s')}} \equiv \sum_{s=1}^{S} f(x^{(s)})\,w^{(s)}$$

importance weight.

This estimator is **consistent** but **biased**

Exercise: Prove that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s \tilde{r}^{(s)}$

Summary so far

- Sums and integrals, often expectations, occur frequently in statistics
- **Monte Carlo** approximates expectations with a sample average
- **Rejection sampling** draws samples from complex distributions
- **Importance sampling** applies Monte Carlo to 'any' sum/integral

Pitfalls of Monte Carlo

Rejection & importance sampling scale badly with dimensionality

Example:

$$P(x) = \mathcal{N}(0, \mathbb{I}), \quad Q(x) = \mathcal{N}(0, \sigma^2\mathbb{I})$$

Rejection sampling:

Requires $\sigma \geq 1$. Fraction of proposals accepted $= \sigma^{-D}$

Importance sampling:

Variance of importance weights $= \left(\frac{\sigma^2}{2-1/\sigma^2}\right)^{D/2} - 1$

Infinite / undefined variance if $\sigma \leq 1/\sqrt{2}$

Outline

- **Monte Carlo Methods**
- **MCMC (Basic Methods)**
  - Metropolis algorithm
  - Metropolis-Hastings (M-H) algorithm
  - Gibbs Sampling
- **Markov Chains**
  - Transition probabilities
  - Invariant distribution
  - Equilibrium distribution
  - Markov chain as a WFSM
  - Constructing Markov chains
  - Why does M-H work?
- **MCMC (Auxiliary Variable Methods)**
  - Slice Sampling
  - Hamiltonian Monte Carlo

Metropolis, Metropolis-Hastings, Gibbs Sampling

MCMC (BASIC METHODS)

A Few Problems for a Factor Graph

Suppose we already have the parameters of a Factor Graph…

1. How do we compute the probability of a specific assignment to the variables?
   P(T=t, H=h, A=a, C=c)

2. How do we draw a sample from the joint distribution?
   t,h,a,c ~ P(T, H, A, C)

3. How do we compute marginal probabilities?
   P(A) = …

4. How do we draw samples from a conditional distribution?
   t,h,a ~ P(T, H, A | C = c)

5. How do we compute conditional marginal probabilities?
   P(H | C = c) = …

Can we use samples?