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

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

\[ \mathcal{D} = \{ \mathbf{x}^{(n)} \}_{n=1}^{N} \]

2. Model

\[ p(\mathbf{x} | \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \]

3. Objective

\[ \ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)} | \theta) \]

4. Learning

\[ \theta^* = \arg\max_{\theta} \ell(\theta; \mathcal{D}) \]

5. Inference

1. Marginal Inference

\[ p(\mathbf{x}_C) = \sum_{\mathbf{x}' : \mathbf{x}'_C = \mathbf{x}_C} p(\mathbf{x}' | \theta) \]

2. Partition Function

\[ Z(\theta) = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \]

3. MAP Inference

\[ \hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x} | \theta) \]
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?
   \( P(H \mid C = c) = \ldots \)
Suppose we took many samples from the distribution over taggings:

\[ p(x) = \frac{1}{Z} \prod_{\alpha} \psi_{\alpha}(x_{\alpha}) \]

Sample 1:

- \( n \)
- \( v \)
- \( p \)
- \( d \)
- \( n \)

Sample 2:

- \( n \)
- \( n \)
- \( v \)
- \( d \)
- \( n \)

Sample 3:

- \( n \)
- \( v \)
- \( p \)
- \( d \)
- \( n \)

Sample 4:

- \( v \)
- \( n \)
- \( p \)
- \( d \)
- \( n \)

Sample 5:

- \( v \)
- \( n \)
- \( v \)
- \( d \)
- \( n \)

Sample 6:

- \( n \)
- \( v \)
- \( p \)
- \( d \)
- \( n \)
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.

Sample 1: $n$, $v$, $p$, $d$, $n$
Sample 2: $n$, $n$, $v$, $d$, $n$
Sample 3: $n$, $v$, $p$, $d$, $n$
Sample 4: $v$, $n$, $p$, $d$, $n$
Sample 5: $v$, $n$, $v$, $d$, $n$
Sample 6: $n$, $v$, $p$, $d$, $n$

<START> time flies like an arrow
Marginals by Sampling on Factor Graph

Estimate the marginals as:

\[
\begin{array}{c|cc}
   X_0 & n/4 & v/6 \\
   \psi_0 & n/4 & v/6 \\
   X_1 & n/3 & v/3 \\
   \psi_2 & n/3 & v/3 \\
   X_2 & n/4 & p/6 \\
   \psi_4 & n/4 & p/6 \\
   X_3 & n/4 & d/6 \\
   \psi_6 & n/4 & d/6 \\
   X_4 & n/6 \\
   \psi_8 & n/6 \\
   X_5 & n/6 \\
   \psi_9 & n/6 \\
\end{array}
\]

Sample 1:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

Sample 2:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

Sample 3:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

Sample 4:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

Sample 5:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

Sample 6:

\[
\begin{array}{c}
   n \\
   \psi_0 \\
   \psi_2 \\
   \psi_4 \\
   \psi_6 \\
   \psi_8 \\
   \psi_9 \\
\end{array}
\]

<START>

\[
\begin{array}{c}
   \text{time} \\
   \psi_1 \\
   \psi_3 \\
   \psi_5 \\
   \psi_7 \\
   \psi_9 \\
\end{array}
\]

flies

like

an

arrow

Estimate the marginals as:
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: \[ \int f(x)P(x) \, dx \approx \hat{f} \equiv \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x) \]

Estimator is unbiased:

\[ \mathbb{E}_{P(\{x^{(s)}\})} \left[ \hat{f} \right] = \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_{P(x)} [f(x)] = \mathbb{E}_{P(x)} [f(x)] \]

Variance shrinks \( \propto 1/S \):

\[ \text{var}_{P(\{x^{(s)}\})} \left[ \hat{f} \right] = \frac{1}{S^2} \sum_{s=1}^{S} \text{var}_{P(x)} [f(x)] = \frac{\text{var}_{P(x)} [f(x)]}{S} \]

“Error bars” shrink like \( \sqrt{S} \)
A dumb approximation of $\pi$

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

$$\pi = 4 \int \int \mathbb{1} \left((x^2 + y^2) < 1\right) P(x, y) \, dx \, dy$$

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

Gives $\pi$ 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 $\sim \text{Uniform}[0,1]$
Sampling from distributions

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

\[
h(y) = \int_{-\infty}^{y} p(y') \, dy'
\]

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

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

Although we can’t always compute and invert \(h(y)\)
Rejection sampling

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

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

- 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 \( Q \):

\[
\int f(x) P(x) \, dx = \int f(x) \frac{P(x)}{Q(x)} Q(x) \, dx,
\]

\[\approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}) \frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x)\]

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 \( P(x) = \frac{\tilde{P}(x)}{Z_P} \)

\[
\int f(x)P(x) \, dx \approx \left( \frac{Z_Q}{Z_P} \right) \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}) \frac{\tilde{P}(x^{(s)})}{\tilde{Q}(x^{(s)})}, \quad x^{(s)} \sim Q(x)
\]

\[
\approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}) \tilde{r}(s) \equiv \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}) w^{(s)}
\]

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

**Exercise:** Prove that \( Z_P/Z_Q \approx \frac{1}{S} \sum_{s} \tilde{r}^{(s)} \)
• 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
MCMC (BASIC METHODS)

Metropolis, Metropolis-Hastings, Gibbs Sampling
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?
   \[ P(H \mid C = c) = \ldots \]