Intro to Sampling Methods

CSE586 Computer Vision II
Penn State Univ
A Brief Overview of Sampling

Monte Carlo Integration
Sampling and Expected Values
Inverse Transform Sampling (CDF)
Rejection Sampling
Importance Sampling
Markov Chain Monte Carlo
Integration and Area

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?

http://videolectures.net/mlss08au_freitas_asm/
Integration and Area

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?

http://videolectures.net/mlss08au_freitas_asm/
Integration and Area

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?

http://videolectures.net/mlss08au_freitas_asm/
Integration and Area

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?

http://videolectures.net/mlss08au_freitas_asm/
Integration and Area

- As we use more samples, our answer should get more and more accurate
- Doesn’t matter what the shape looks like

arbitrary region (even disconnected)

area under curve aka integration!
Monte Carlo Integration

Goal: compute definite integral of function $f(x)$ from $a$ to $b$

count the $K$ samples that fall below the $f(x)$ curve

Generate $N$ uniform random samples in upper bound volume

Answer = $\frac{K}{N} \times \text{Area of upper bound volume}$
Monte Carlo Integration

Sampling-based integration is useful for computing the normalizing constant that turns an arbitrary non-negative function $f(x)$ into a probability density function $p(x)$.

$$ Z = \int f(x) \, dx $$

Compute this via sampling (Monte Carlo Integration). Then:

$$ P(x) = \frac{1}{Z} f(x) $$

Note: for complicated, multidimensional functions, this is the ONLY way we can compute this normalizing constant.
Sampling: A Motivation

If we can generate random samples $x_i$ from a given distribution $P(x)$, then we can estimate expected values of functions under this distribution by summation, rather than integration.

That is, we can approximate:

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

by first generating $N$ i.i.d. samples from $P(x)$ and then forming the empirical estimate:

$$\hat{E}(f(x)) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$
Expected Values and Sampling

Example:

a discrete pdf

\[ P(x) \]

\[
P(1) = \frac{1}{4} \\
P(2) = \frac{1}{4} \\
P(3) = \frac{2}{4}
\]

\[
E_P(g(x)) = \sum_{i=1}^{3} g(i)P(i)
\]

\[
= g(1)\frac{1}{4} + g(2)\frac{1}{4} + g(3)\frac{2}{4}
\]
Expected Values and Sampling (cont)

generate 10 samples from $P(x)$

$x_1$ $x_2$ $x_3$ $x_4$ $x_5$ $x_6$ $x_7$ $x_8$ $x_9$ $x_{10}$
1 3 3 2 2 3 1 3 3 1

$\hat{E}_P(g(x)) = \frac{1}{10} \sum_{i=1}^{10} g(x_i)$

$= \frac{1}{10} [g(1) + g(3) + g(3) + g(2) + g(2) + g(3) + g(1) + g(3) + g(3) + g(1)]$

$= [3g(1) + 2g(2) + 5g(3)]$

$= \frac{3}{10} g(1) + \frac{2}{10} g(2) + \frac{5}{10} g(3) \sim g(1) \frac{1}{4} + g(2) \frac{1}{4} + g(3) \frac{2}{4}$
Inverse Transform Sampling

It is easy to sample from a discrete 1D distribution, using the cumulative distribution function.

\[ c(k) = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{N} w_i} \]

cumulative distribution function

\[ F(x) = P(X \leq x) \]
Inverse Transform Sampling

It is easy to sample from a discrete 1D distribution, using the cumulative distribution function.

1) Generate uniform $u$ in the range $[0,1]$
2) Visualize a horizontal line intersecting bars
3) If index of intersected bar is $j$, output new sample $x_k = j$
Inverse Transform Sampling

Why it works:

cumulative distribution function

\[ F(x) = P(X \leq x) \]

inverse cumulative distribution function

\[ F^{-1}(t) = \min\{x : F(x) = t, 0 < t < 1\} \]

Claim: if \( U \) is a uniform random variable on (0,1) then \( X = F^{-1}(U) \) has distribution function \( F \).

Proof:

\[
P(F^{-1}(U) \leq x) \\
= P(\min\{x : F(x) = U\} \leq x) \quad \text{(def of } F^{-1}) \\
= P(U \leq F(x)) \quad \text{(applied } F \text{ to both sides)} \\
= F(x) \quad \text{(def of distribution function of } U) 
\]
Efficient Generating Many Samples
(naive approach is NlogN, but we can do better)

Algorithm 2: Resampling Algorithm
\[ \{\{x^*_k, w_k, i\}\}_{j=1}^{N_s} = \text{RESAMPLE } \{\{x_k, w_k\}\}_{j=1}^{N_s} \]
- Initialize the CDF: \( c_1 = 0 \)
- FOR \( i = 2: N_s \)
  - Construct CDF: \( c_i = c_{i-1} + w_k^i \)
  - END FOR
- Start at the bottom of the CDF: \( i = 1 \)
- Draw a starting point: \( u_1 \sim \mathcal{U}[0, N_s^{-1}] \)
- FOR \( j = 1: N_s \)
  - Move along the CDF: \( u_j = u_1 + N_s^{-1}(j - 1) \)
  - WHILE \( u_j \geq c_i \)
  - \( i = i + 1 \)
  - END WHILE
  - Assign sample: \( x^*_k = x_k \)
  - Assign weight: \( w_k^i = N_s^{-1} \)
  - Assign parent: \( i^j = i \)
- END FOR

Basic idea: choose one initial small random number; deterministically sample the rest by “crawling” up the cdf function. This is O(N).

Odd property: you generate the “random” numbers in sorted order...
Example: Sampling From an Image

“Likelihood” image

500 samples 1000 samples 5000 samples 10000 samples
A Brief Overview of Sampling

Inverse Transform Sampling (CDF)

Rejection Sampling

Importance Sampling

For these two, we can sample from an unnormalized distribution function.

That is, to sample from distribution $P$, we only need to know a function $P^*$, where $P = P^* / c$, for some normalization constant $c$. 
Rejection Sampling

Need a proposal density $Q(x)$ [e.g. uniform or Gaussian], and a constant $c$ such that $c(Qx)$ is an upper bound for $P^*(x)$

Example with $Q(x)$ uniform

- generate uniform random samples in upper bound volume
- the marginal density of the $x$ coordinates of the points is then proportional to $P^*(x)$
- accept samples that fall below the $P^*(x)$ curve

Note the relationship to Monte Carlo integration.
Rejection Sampling

More generally:

1) generate sample $x_i$ from a proposal density $Q(x)$
2) generate sample $u$ from uniform $[0, cQ(x_i)]$
3) if $u \leq P^*(x_i)$ accept $x_i$; else reject
Importance “Sampling”

Not for generating samples. It is a method to estimate the expected value of a function $f(x_i)$ directly

1) Generate $x_i$ from $Q(x)$

2) an empirical estimate of $E_Q(f(x))$, the expected value of $f(x)$ under distribution $Q(x)$, is then

$$\hat{E}_Q(f(x)) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

3) However, we want $E_P(f(x))$, which is the expected value of $f(x)$ under distribution $P(x) = P^*(x)/Z$
Importance Sampling

When we generate from $Q(x)$, values of $x$ where $Q(x)$ is greater than $P^*(x)$ are overrepresented, and values where $Q(x)$ is less than $P^*(x)$ are underrepresented.

To mitigate this effect, introduce a weighting term

$$w_i = \frac{P^*(x_i)}{Q(x_i)}$$
Importance Sampling

New procedure to estimate $E_P(f(x))$:

1) Generate $N$ samples $x_i$ from $Q(x)$

2) form importance weights

$$w_i = \frac{P^*(x_i)}{Q(x_i)}$$

3) compute empirical estimate of $E_P(f(x))$, the expected value of $f(x)$ under distribution $P(x)$, as

$$\hat{E}_P(f(x)) = \frac{\sum w_i f(x_i)}{\sum w_i}$$
Resampling

Note: We thus have a set of weighted samples \((x_i, w_i | i=1,\ldots,N)\)

If we really need random samples from \(P\), we can generate them by resampling such that the likelihood of choosing value \(x_i\) is proportional to its weight \(w_i\).

This would now involve now sampling from a discrete distribution of \(N\) possible values (the \(N\) values of \(x_i\)).

Therefore, regardless of the dimensionality of vector \(x\), we are resampling from a 1D distribution (we are essentially sampling from the indices 1...\(N\), in proportion to the importance weights \(w_i\)). So we can using the inverse transform sampling method we discussed earlier.
Note on Proposal Functions

Computational efficiency is best if the proposal distribution looks a lot like the desired distribution (area between curves is small).

These methods can fail badly when the proposal distribution has 0 density in a region where the desired distribution has non-negligible density.

For this last reason, it is said that the proposal distribution should have *heavy tails*. 
Sequential Monte Carlo Methods

Sequential Importance Sampling (SIS) and the closely related algorithm Sampling Importance Sampling (SIR) are known by various names in the literature:

- bootstrap filtering
- particle filtering
- Condensation algorithm
- survival of the fittest

General idea: Importance sampling on time series data, with samples and weights updated as each new data term is observed. Well-suited for simulating Markov chains and HMMs!
Markov-Chain Monte Carlo
References

Markov Chain Monte Carlo for Computer Vision
A tutorial at the 10th Int'l Conf. on Computer Vision
October, 2005, Beijing
by
Song-Chun Zhu,  UCLA
Frank Dellaert,  Gatech
Zhuowen Tu,  UCLA

INTRODUCTION TO MONTE CARLO METHODS
D.J.C. MACKAY
Department of Physics, Cambridge University.
Cavendish Laboratory, Madingley Road,
Cambridge, CB3 0HE. United Kingdom.

An Introduction to MCMC for Machine Learning
CHRISTOPHE ANDRIEU  C.Andrieu@bristol.ac.uk
Department of Mathematics, Statistics Group, University of Bristol, University Walk, Bristol BS8 1TW, UK

NANDO DE FREITAS  nando@cs.ubc.ca
Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver,
BC V6T 1Z4, Canada

ARNAUD DOUCET  doucet@ee.uma.oz.au
Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia

MICHAEL I. JORDAN  jordan@cs.berkeley.edu
Departments of Computer Science and Statistics, University of California at Berkeley, 587 Soda Hall, Berkeley,
CA 94720-1776, USA
Recall: Sampling Motivation

If we can generate random samples $x_i$ from a given distribution $P(x)$, then we can estimate expected values of functions under this distribution by summation, rather than integration.

That is, we can approximate:

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

by first generating $N$ i.i.d. samples from $P(x)$ and then forming the empirical estimate:

$$\hat{E}(f(x)) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$
Standard methods fail:

- **Rejection Sampling**
  - Rejection rate increase with $N \to 100\%$
- **Importance Sampling**
  - Same problem: vast majority weights $\to 0$

**Intuition:** In high dimension problems, the “Typical Set” (volume of nonnegligible prob in state space) is a small fraction of the total space.
High-Dimensional Spaces

consider ratio of volumes of hypersphere inscribed inside hypercube

\[
\frac{V(S_d(r))}{V(H_d(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \approx 75\%
\]

\[
\frac{V(S_3(r))}{V(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} \approx 50\%
\]

Asymptotic behavior:

\[
\lim_{d \to \infty} \frac{V(S_d(r))}{V(H_d(2r))} = \lim_{d \to \infty} \frac{\pi^{d/2}}{2^d \Gamma\left(\frac{d}{2} + 1\right)} \to 0
\]

most of volume of the hypercube lies outside of hypersphere as dimension \( d \) increases

http://www.cs.rpi.edu/~zaki/Courses/dmcourse/Fall09/notes/highdim.pdf
High Dimensional Spaces

Segmentation Example

- Binary Segmentation of image

each pixel has two states: on and off
Probability of a Segmentation

- Very high-dimensional
- $256 \times 256$ pixels = 65536 pixels
- Dimension of state space $N = 65536$ !!!!

- # binary segmentations = finite, but...
- $2^{65536} = 2 \times 10^{19728} >> 10^{79} = \text{atoms in universe}$
Robert Collins
CSE

Representation $P(\text{Segmentation})$

- Histogram? No!
- Assume pixels independent?
  $P(x_1 x_2 x_3 ...) = P(x_1) P(x_2) P(x_3) ...$
  
  ignores neighborhood structure of pixel lattice and empirical evidence that images are “smooth”

- Approximate solution: samples !!!
Brilliant Idea!

- Published June 1953
- Top 10 algorithm!
- Set up a Markov chain
- Run the chain until stationary
- All subsequent samples are from stationary distribution
Recall: Markov Chains

Markov Chain:

- A sequence of random variables $X_1, X_2, X_3, ...$
- Each variable is a distribution over a set of states (a, b, c...)
- Transition probability of going to next state only depends on the current state. e.g. $P(X_{n+1} = a | X_n = b)$

transition probs can be arranged in an NxN table of elements

$k_{ij} = P(X_{n+1} = j | X_n = i)$

where the rows sum to one
General Idea: MCMC Sampling

Start in some state, and then run the simulation for some number of time steps. After you have run it “long enough” start keeping track of the states you visit.

\{... \text{X1 X2 X1 X3 X3 X2 X1 X2 X1 X1 X3 X3 X2 ...} \}

These are samples from the distribution you want, so you can now compute any expected values with respect to that distribution empirically.
If the Markov chain is positive recurrent, there exists a stationary distribution. If it is positive recurrent and irreducible, there exists a unique stationary distribution. Then, the average of a function $f$ over samples of the Markov chain is equal to the average with respect to the stationary distribution.

We can compute this empirically as we generate samples. This is what we want to compute, and is infeasible to compute in any other way.
K = transpose of transition prob table \( \{ k_{ij} \} \) (cols sum to one. We do this for computational convenience (next slide).
Question:

Assume you start in some state, and then run the simulation for a large number of time steps. What percentage of time do you spend at X1, X2 and X3?
Robert Collins
CS

four possible initial distributions

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\frac{1}{3} \\
\frac{1}{3} \\
\frac{1}{3}
\end{bmatrix}
\]

initial distribution

distribution after one time step
\[
\begin{align*}
q_0 &= \mathbf{K} q_0 \\
q_1 &= \mathbf{K} q_1 = \mathbf{K}^2 q_0 \\
q_2 &= \mathbf{K} q_2 = \mathbf{K}^2 q_1 = \mathbf{K}^3 q_0 \\
q_3 &= \mathbf{K} q_3 = \mathbf{K}^2 q_2 = \mathbf{K}^3 q_1 = \mathbf{K}^4 q_0 \\
q_{10} &= \mathbf{K} q_9 = \ldots = \mathbf{K}^{10} q_0
\end{align*}
\]

all eventually end up with same distribution -- this is the stationary distribution!
# Eigen-analysis

\[
K = \begin{bmatrix}
0.1000 & 0.5000 & 0.6000 \\
0.6000 & 0.2000 & 0.3000 \\
0.3000 & 0.3000 & 0.1000
\end{bmatrix}
\]

\[
K \mathbf{e}_1 = \mathbf{e}_1 \mathbf{D}
\]

(\text{Perron-Frobenius theorem})

**Stationary distribution**

\[
\pi = \frac{\mathbf{e}_1}{\text{sum}(\mathbf{e}_1)}
\]

i.e. \( K \pi = \pi \)

\[
E = \begin{bmatrix}
0.6396 & 0.7071 & -0.2673 \\
0.6396 & -0.7071 & 0.8018 \\
0.4264 & 0.0000 & -0.5345
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
1.0000 & 0 & 0 \\
0 & -0.4000 & 0 \\
0 & 0 & -0.2000
\end{bmatrix}
\]
The PageRank of a webpage as used by Google is defined by a Markov chain. It is the probability to be at page $i$ in the stationary distribution on the following Markov chain on all (known) webpages. If $N$ is the number of known webpages, and a page $i$ has $ki$ links then it has transition probability $(1-q)/ki + q/N$ for all pages that are linked to and $q/N$ for all pages that are not linked to. The parameter $q$ is taken to be about 0.15.
Google Pagerank

Pagerank == First Eigenvector of the Web Graph!

Computation assumes a 15% “random restart” probability

Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual {Web} search engine, Computer Networks and ISDN Systems, 1998
Another Question:

Assume you want to spend a particular percentage of time at X1, X2 and X3. What should the transition probabilities be?

\[ P(x1) = 0.2 \]
\[ P(x2) = 0.3 \]
\[ P(x3) = 0.5 \]

\[
K = \begin{bmatrix}
? & ? & ? \\
? & ? & ? \\
? & ? & ?
\end{bmatrix}
\]
Thought Experiment

Consider only two states. What transition probabilities should we use so that we spend roughly equal time in each of the two states? (i.e. 50% of the time we are in state 1 and 50% of the time we are in state 2)
Detailed Balance

• Consider a pair of configuration nodes r, s
• Want to generate them with frequency relative to their likelihoods \( L(r) \) and \( L(s) \)
• Let \( q(r, s) \) be relative frequency of proposing configuration \( s \) when the current state is \( r \) (and vice versa)

A sufficient condition to generate \( r, s \) with the desired frequency is

\[
L(r) \cdot q(r, s) = L(s) \cdot q(s, r)
\]

“detailed balance”
Detailed Balance

• In practice, you just propose some transition probabilities.
• They typically will NOT satisfy detailed balance (unless you are extremely lucky).
• Instead, you “fix them” by introducing a computational fudge factor

Detailed balance:

\[ a \cdot L(r) \cdot q(r,s) = L(s) \cdot q(s,r) \]

Solve for \(a\):

\[ a = \frac{L(s) \cdot q(s,r)}{L(r) \cdot q(r,s)} \]
diff with rejection sampling: instead of throwing away rejections, you replicate them into next time step.

Metropolis-Hastings Algorithm

This leads to the following algorithm:

0. Start with $x^{(0)}$, then iterate:

1. propose $x'$ from $q(x^{(t)}, x')$

2. calculate ratio

$$a = \frac{\pi(x')q(x', x^{(t)})}{\pi(x^{(t)})q(x^{(t)}, x')}$$

3. if $a > 1$ accept $x^{(t+1)} = x'$
else accept with probability $a$
if rejected: $x^{(t+1)} = x^{(t)}$

Note: you can just make up transition probability $q$ on-the-fly, using whatever criteria you wish.
Note: the transition probabilities $q(x^{(i)}, x)$ can be arbitrary distributions. They can depend on the current state and change at every time step, if you want.
Metropolis Hastings Example

\[ P(x_1) = 0.2 \]
\[ P(x_2) = 0.3 \]
\[ P(x_3) = 0.5 \]

Proposal distribution
\[ q(x_i, (x_i-1) \text{mod} 3) = 0.4 \]
\[ q(x_i, (x_i+1) \text{mod} 3) = 0.6 \]
Variants of MCMC

- there are many variations on this general approach, some derived as special cases of the Metropolis-Hastings algorithm
The Metropolis Algorithm

When \( q \) is symmetric, i.e., \( q(x,x') = q(x',x) \): e.g. Gaussian

0. Start with \( x^{(0)} \), then iterate:
1. propose \( x' \) from \( q(x^{(t)},x') \)
2. calculate ratio
   \[
   a = \frac{\pi(x')}{\pi(x^{(t)})} \cdot \frac{q(x',x)}{q(x,x')}
   \]
   cancels
3. if \( a > 1 \) accept \( x^{(t+1)} = x' \)
   else accept with probability \( a \)
   if rejected: \( x^{(t+1)} = x^{(t)} \)
simpler version, using 1D conditional distributions

Gibbs Sampling

- Example: target \( \pi(x_1, x_2) \)
- Algorithm:
  - alternate between \( x_1 \) and \( x_2 \)
  - 1. sample from \( x_1 \sim P(x_1|x_2) \)
  - 2. sample from \( x_2 \sim P(x_2|x_1) \)
- After a while: samples from target density!

- Sampler equivalent of “Gauss-Seidel” iterations or line search, or ...
1D conditional distr

(a) $\pi(x_1, x_2)$

(b) $\pi(x_1|x_2^{(t)})$

(c) $\pi(x_2|x_1)$

(d) $\pi(x_2|x_1)$

interleave
Gibbs Sampler

Special case of MH with acceptance ratio always 1 (so you always accept the proposal).

\[ q(x, y) = \begin{cases} \pi(y_i | x_{(i)}) & y_{(i)} = x_{(i)}, \ i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \]

where \( x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k), \ i = 1, \ldots, k, \ 1 < k \leq p, \)

With this proposal, the corresponding acceptance probability is given by

\[
\alpha(x, y) = \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} = \frac{\pi(y) / \pi(y_i | x_{(i)})}{\pi(x) / \pi(x_i | y_{(i)})}
\]

\[
= \frac{\pi(y) / \pi(y_i | y_{(i)})}{\pi(x) / \pi(x_i | x_{(i)})}, \quad \text{since } y_{(i)} = x_{(i)},
\]

\[
= \frac{\pi(y_{(i)})}{\pi(x_{(i)})}, \quad \text{by definition of conditional probability for } \theta = (\theta_i, \theta_{(i)}),
\]

\[
= 1, \quad \text{since } y_{(i)} = x_{(i)}.
\]

S. Brooks, “Markov Chain Monte Carlo and its Application”
Simulated Annealing

- introduce a “temperature” term that makes it more likely to accept proposals early on. This leads to more aggressive exploration of the state space.
- Gradually reduce the temperature, causing the process to spend more time exploring high likelihood states.
- Rather than remember all states visited, keep track of the best state you’ve seen so far. This is a method that attempts to find the global max (MAP) state.
Trans-dimensional MCMC

• Exploring alternative state spaces of differing dimensions (example, when doing EM, also try to estimate number of clusters along with parameters of each cluster).

• Green’s reversible-jump approach (RJMCMC) gives a general template for exploring and comparing states of differing dimension.