Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

ISML 2013

Outlines

Part XIX

Sampling

Motivation

- For most probabilistic models of practical interest, exact inference is intractable. Need approximation.
- Last lecture: deterministic approximations (which can not be exact in principle).
- Now : Numerical sampling (Monte Carlo methods).
- Fundamental problem : Find the expectation of some function $f(\mathbf{z})$ w.r.t. a probability distribution $p(\mathbf{z})$

$$\mathbb{E}\left[f\right] = \int f(\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

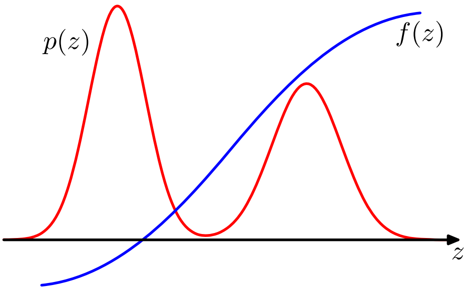
- Key idea : Draw $\mathbf{z}^{(l)}$, $l = 1 \dots, L$ independent samples from $p(\mathbf{z})$ and approximate the expectation by

$$\widehat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)})$$

- Problem: How to obtain independent samples from $p(\mathbf{z})$?

Approximating the Expectation of $f(\mathbf{z})$

- Samples must be independent, otherwise the effective sample size is much smaller than the appearent sample size.
- If $f(\mathbf{z})$ is small in regions where $p(\mathbf{z})$ is large (or vice versa) : need large sample sizes to catch contributions from all regions.



Sampling from the Uniform Distribution

- In a computer usually via pseudorandom number generator : an algorithm generating a sequence of numbers that approximates the properties of random numbers.
- Example : linear congruential generators

$$z^{(n+1)} = (a\, z^{(n)} + c) \mod m$$

for modulus $m > 0$, multiplier $0 < a < m$, increment $0 \leq c < m$, and seed z_0.

- Other classes of pseudorandom number generators:
  - Lagged Fibonacci generators
  - Linear feedback shift registers
  - Generalised feedback shift registers

Pseudorandom Number Generators - Problems

Careful mathematical analysis required to avoid problems like

- Shorter than expected periods for some seed states
- Lack of uniformity of distribution
- Correlation of successive values
- Poor dimensional distribution of the output sequence
- The distances between where certain values occur are distributed differently from those in a random sequence distribution.

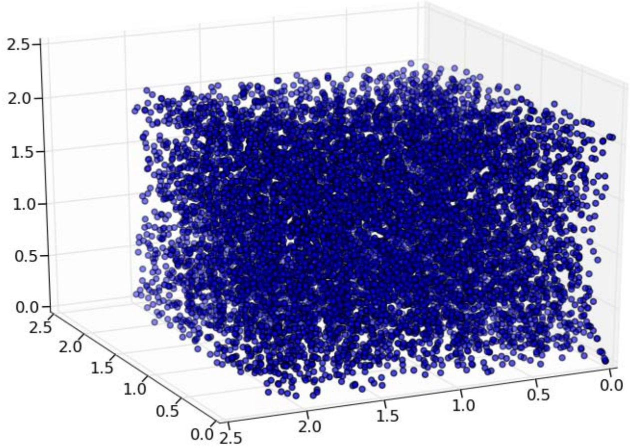
A Bad Generator - RANDU

- Used since the 1960s on many machines
- Defined by the recurrence

$$z^{(n+1)} = (2^{16} + 3)\, z^{(n)} \mod 2^{31}$$

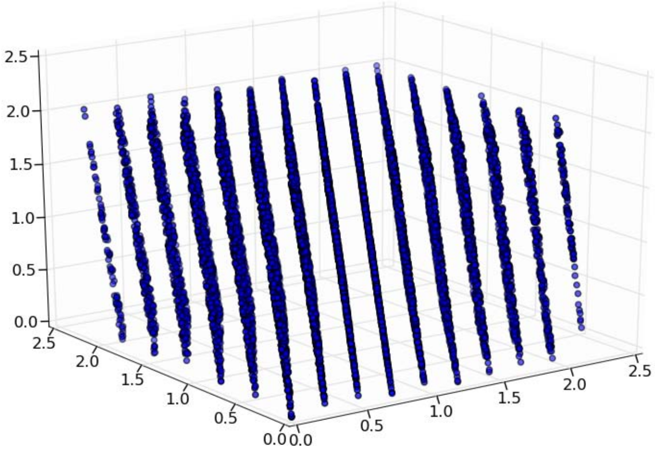
A Bad Generator - RANDU

- Plotting $\left(z^{(n+2)}, z^{(n+1)}, z^{(n)}\right)^T$ in 3D ...



A Bad Generator - RANDU

- Plotting $\left(z^{(n+2)}, z^{(n+1)}, z^{(n)}\right)^T$ in 3D ... and changing the viewpoint results in 15 planes.



A Bad Generator - RANDU

- Analyse the recurrence

$$z^{(n+1)} = (2^{16} + 3)\, z^{(n)} \quad \bmod\ 2^{31}$$

- Assuming every equation to be modulo 2^{31}, we can correlate three samples

$$\begin{aligned} z^{(n+2)} &= (2^{16} + 3)^2\, z^{(n)} \\ &= (2^{32} + 6 \cdot 2^{16} + 9) z^{(n)} \\ &= (6(2^{16} + 3) - 9) z^{(n)} \\ &= 6 z^{(n+1)} - 9 z^{(n)} \end{aligned}$$

- Marsaglia, George "Random Numbers Fall Mainly In The Planes", Proc National Academy of Sciences 61, 25-28, 1968.

Sampling from the Uniform Distribution

- Use a mathematically well crafted pseudorandom number generator.
- From now on we will assume that we have a good pseudorandom number generator for uniformly distributed data available.
- If you don't trust any algorithm :
  Three carefully adjusted radio receivers picking up atmospheric noise to provide real random numbers at
  `http://www.random.org/`

Sampling from Standard Distributions

- Goal: Sample from $p(y)$ which is given in analytical form.
- Suppose uniformly distributed samples of z in the interval $(0, 1)$ are available.
- Calculate the cumulative distribution function

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

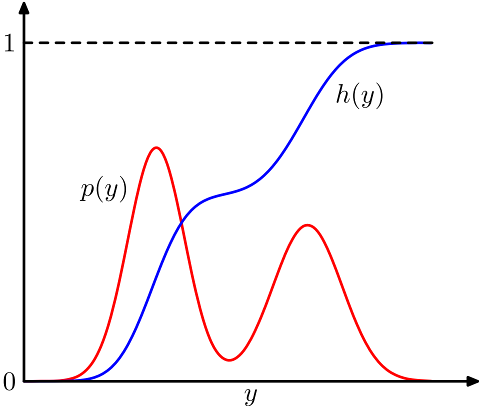
- Transform the samples from $\mathcal{U}(z \,|\, 0, 1)$ by

$$y = h^{-1}(z)$$

to obtain samples y distributed according to $p(y)$.

Sampling from Standard Distributions

- Goal: Sample from $p(y)$ which is given in analytical form.
- If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$ then y will be distributed according to $p(y)$.



Sampling from the Exponential Distribution

- Goal: Sample from the exponential distribution

$$p(y) = \begin{cases} \lambda e^{-\lambda y} & 0 \leq y \\ 0 & y < 0 \end{cases}$$

  with rate parameter $\lambda > 0$.
- Suppose uniformly distributed samples of z in the interval $(0, 1)$ are available.
- Calculate the cumulative distribution function

$$h(y) = \int_{-\infty}^{y} p(x) \,\mathrm{d}x = \int_{0}^{y} \lambda e^{-\lambda y} \,\mathrm{d}x = 1 - e^{-\lambda y}$$

- Transform the samples from $\mathcal{U}(z \,|\, 0, 1)$ by

$$y = h^{-1}(z) = -\frac{1}{\lambda} \ln(1 - z)$$

  to obtain samples y distributed according to the exponential distribution.

Sampling from Multivariate Distributions

- Generalisation to multiple variables is straightforward
- Consider change of variables via the Jacobian

$$p(y_1, \dots, y_M) = p(z_1, \dots, z_M) \left| \frac{\partial(z_1, \dots, z_M)}{\partial(y_1, \dots, y_M)} \right|$$

- Technical challenge: Multiple integrals; inverting nonlinear functions of multiple variables.

NICTA
The Australian National University

Sampling the Gaussian Distribution - Box-Muller

1. Generate pairs of uniformly distributed random numbers $z_1, z_2 \in (-1, 1)$ (e.g. $z_i = 2z - 1$ for z from $\mathcal{U}(z \mid 0, 1)$)
2. Discard any pair (z_1, z_2) unless $z_1^2 + z_2^2 \leq 1$. Results in a uniform distribution inside of the unit circle $p(z_1, z_2) = 1/\pi$.
3. Evaluate $r^2 = z_1^2 + z_2^2$ and

$$y_1 = z_1 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

$$y_2 = z_2 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

4. y_1 and y_2 are independent with joint distribution

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| = \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2}$$

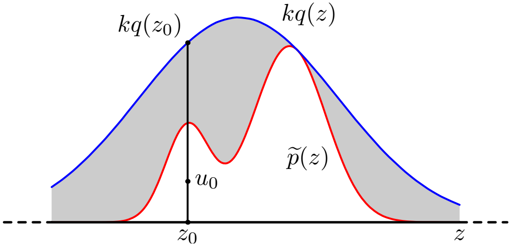
Rejection Sampling

- Assumption 1 : Sampling directly from $p(z)$ is difficult, but we can evaluate $p(z)$ up to some unknown normalisation constant Z_p

$$p(z) = \frac{1}{Z_p} \widetilde{p}(z)$$

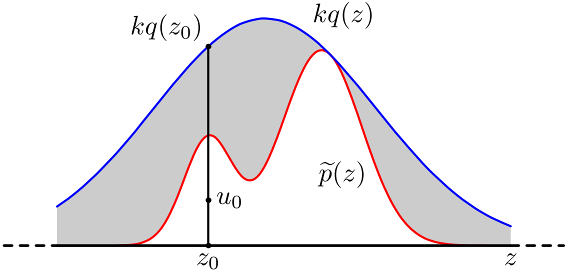
- Assumption 2 : We can draw samples from a simpler distribution $q(z)$ and for some constant k and all z holds

$$kq(z) \geq \widetilde{p}(z)$$



Rejection Sampling

1. Generate a random number z_0 from the distribution $q(z)$.
2. Generate a number from the u_0 from the uniform distribution over $[0, k\,q(z_0)]$.
3. If $u_0 > \widetilde{p}(z_0)$ then reject the pair (z_0, u_0).
4. The remaining pairs have uniform distribution under the curve $\widetilde{p}(z)$.
5. The z values are distributed according to $p(z)$.



Rejection Sampling - Example

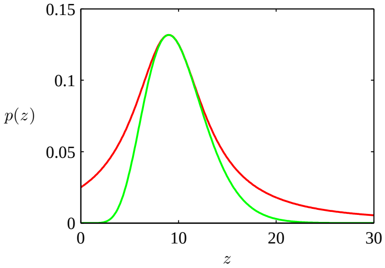
- Consider the Gamma Distribution for $a > 1$

$$\mathrm{Gam}(z \mid a, b) = \frac{b^a z^{a-1} \exp(-bz)}{\Gamma(a)}$$

- Suitable $q(z)$ could be like the Cauchy distribution

$$q(z) = \frac{k}{1 + (z-c)^2/b^2}$$

- Samples z from $q(z)$ by using uniformly distributed y and transformation $z = b \tan y + c$ for $c = a - 1$, $b^2 = 2a - 1$ and k as small as possible for $kq(z) \geq \widetilde{p}(z)$.



Approximating π *- Buffon's Needle (1777)*

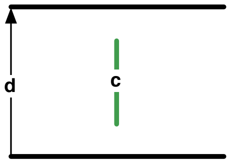
(i) Needle falls perpendicular:
Probability of crossing the line is c/d.

(ii) Needle falls at an arbitrary angle a :
Probability of crossing the line $c \sin(a)/d$.

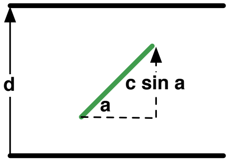
(iii) Every angle is equally probable. Calculate the mean.

$$p(\text{crossing}) = \frac{c}{d} \int_0^{\pi} \sin(a)\, \mathrm{d}p(a) = \frac{1}{\pi} \frac{c}{d} \int_0^{\pi} \sin(a)\, \mathrm{d}a = \frac{2}{\pi} \frac{c}{d}$$

(iv) n crossings in N experiments results in $\frac{n}{N} \approx \frac{2}{\pi} \frac{c}{d}$



(i) Needle falls perpendicular ($a = \pi/2$).

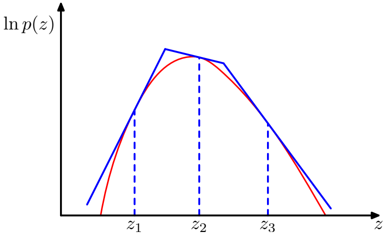


(ii) Needle falls at arbitrary angle a.

Adaptive Rejection Sampling

- Suitable form for the proposal distribution $q(z)$ might be difficult to find.
- If $p(z)$ is log-concave ($\ln p(z)$ has nonincreasing derivatives), use the derivatives to construct an envelope.

1. Start with an initial grid of points $z_1, \dots, z_M$ and construct the envelope using the tangents at the $p(z_i)$, $i = 1, \dots, M$.
2. Draw a sample from the envelop function and if accepted use it to calculate $p(z)$. Otherwise, use it to refine the grid.

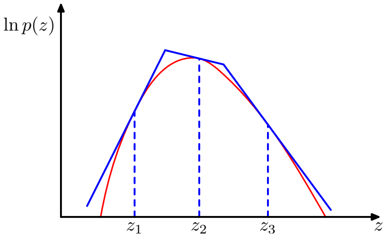


Adaptive Rejection Sampling - Example

- The piecewise exponential distribution is defined as

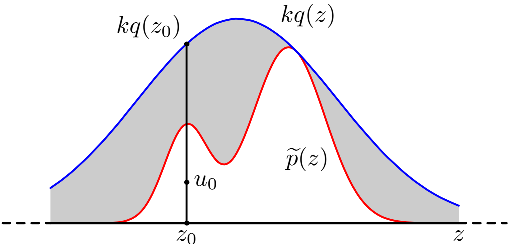
$$p(z) = k_m \lambda_m e^{-\lambda_m (z - z_{m-1})} \qquad \widehat{z}_{m-1,m} < z \leq \widehat{z}_{m,m+1}$$

where $\widehat{z}_{m-1,m}$ is the point of intersection of the tangent lines at z_{m-1} and z_m, λ_m is the slope of the tangent at z_m and k_m accounts for the corresponding offset.



Rejection Sampling - Problems

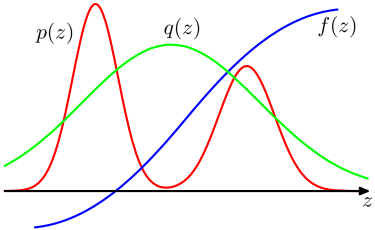
- Need to find a proposal distribution $q(z)$ which is a close upper bound to $p(z)$; otherwise many samples are rejected.
- Curse of dimensionality for multivariate distributions.



Importance Sampling

- Provides a framework to directly calculate the expectation $\mathbb{E}_p\left[f(z)\right]$ with respect to some distribution $p(z)$.
- Does NOT provide $p(z)$.
- Again use a proposal distribution $q(z)$ and draw samples z from it.
- Then

$$\mathbb{E}\left[f\right] = \int f(z)\, p(z)\, \mathrm{d}z = \int f(z)\, \frac{p(z)}{q(z)} q(z)\, \mathrm{d}z \approx \frac{1}{L} \sum_{l=1}^{L} \frac{p(z^{(l)})}{q(z^{(l)})} f(z^{(l)})$$



Importance Sampling - Unnormalised

- Consider both $\widetilde{p}(z)$ and $\widetilde{q}(z)$ to be not normalised.

$$p(z) = \frac{\widetilde{p}(z)}{Z_p} \qquad\qquad q(z) = \frac{\widetilde{q}(z)}{Z_q}.$$

- It follows then that

$$\mathbb{E}\left[f\right] \approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} \widetilde{r}_l f(z^{(l)}) \qquad\qquad \widetilde{r}_l = \frac{\widetilde{p}(z^{(l)})}{\widetilde{q}(z^{(l)})}.$$

- Use the same set of samples to calculate

$$\frac{Z_p}{Z_q} \approx \frac{1}{L} \sum_{l=1}^{L} \widetilde{r}_l,$$

- resulting in the formula for unnormalised distributions

$$\mathbb{E}\left[f\right] \approx \sum_{l=1}^{L} w_l f(z^{(l)}) \qquad\qquad w_l = \frac{\widetilde{r}_l}{\sum_{m=1}^{L} \widetilde{r}_m}$$

ISML 2013

Importance Sampling - Key Points

- Try to choose sample points in the input space where the product $f(z)\,p(z)$ is large.
- Or at least where $p(z)$ is large.
- Importance weights r_l correct the bias introduced by sampling from the proposal distribution $q(z)$ instead of the wanted distribution $p(z)$.
- Success depends on how well $q(z)$ approximates $p(z)$.
- If $p(z) > 0$ in same region, then $q(z) > 0$ necessary.

Markov Chain Monte Carlo

ISML 2013

- Goal : Generate samples from the distribution $p(z)$.
- Idea : Build a machine which uses the current sample to decide which next sample to produce in such a way that the overall distribution of the samples will be $p(z)$.
  1. Current sample $z^{(r)}$ is known. Generate a new sample $z^{\star}$ from a proposal distribution $q(z \mid z^{(r)})$ we know how to sample from.
  2. Accept or reject the new sample according to some appropriate criterion.

$$z^{(l+1)} = \begin{cases} z^{\star} & \text{if accepted} \\ z^{(r)} & \text{if rejected} \end{cases}$$

  3. Proposal distribution depends on the current state.

Metropolis Algorithm

1. Choose a symmetric proposal distribution $q(z_A \mid z_B) = q(z_B \mid z_A)$.
2. Accept the new sample $z^\star$ with probability

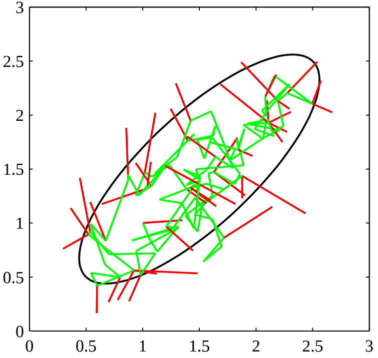
$$A(z^\star, z^{(r)}) = \min\left(1, \frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})}\right)$$

3. How? Choose a random number u with uniform distribution in $(0,1)$. Accept new sample if $A(z^\star, z^{(r)}) > u$.
4. $$z^{(l+1)} = \begin{cases} z^\star & \text{if accepted} \\ z^{(r)} & \text{if rejected} \end{cases}$$

Rejection of a point leads to inclusion of the previous sample. (Different from rejection sampling.)

Metropolis Algorithm - Illustration

- Sampling from a Gaussian Distribution (black contour shows one standard deviation).
- Proposal distribution is isotropic Gaussian with standard deviation 0.2.
- 150 candidates generated; 43 rejected.



accepted steps, rejected steps.

Markov Chain Monte Carlo - Why it works

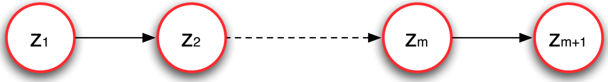
- A First Order Markov Chain is a series of random variables $z^{(1)}, \ldots, z^{(M)}$ such that the following property holds

$$p(z^{(m+1)} \mid z^{(1)}, \ldots, z^{(m)}) = p(z^{(m+1)} \mid z^{(m)})$$

- Marginal probability

$$\begin{aligned} p(z^{(m+1)}) &= \sum_{z^{(m)}} p(z^{(m+1)} \mid z^{(m)})\, p(z^{(m)}) \\ &= \sum_{z^{(m)}} T_m(z^{(m)} \mid z^{(m+1)})\, p(z^{(m)}) \end{aligned}$$

where $T_m(z^{(m)} \mid z^{(m+1)})$ are the transition probabilities.



Markov Chain Monte Carlo - Why it works

- Marginal probability

$$p(z^{(m+1)}) = \sum_{z^{(m)}} T_m(z^{(m)} \mid z^{(m+1)}) \, p(z^{(m)})$$

- A Markov chain is called homogeneous if the transition probabilities are the same for all m, denoted by $T(z', z)$.
- A distribution is invariant, or stationary, with respect to a Markov chain if each step leaves the distribution invariant.
- For a homogeneous Markov chain, the distribution $p^\star(z)$ is invariant if

$$p^\star(z) = \sum_{z'} T(z', z) \, p^\star(z').$$

(Note: There can be many. If T is the identity matrix, every distribution is invariant.)

Markov Chain Monte Carlo - Why it works

- Detailed balance

$$p^\star(z)\,T(z,z') = p^\star(z')\,T(z',z).$$

  is sufficient (but not necessary) for $p^\star(z)$ to be invariant. (A Markov chain that respects the detailed balance is called reversible.)
- A Markov chain is ergodic if it converges to the invariant distribution irrespective of the choice of the initial conditions. The invariant distribution is then called equilibrum.
- An ergodic Markov chain can have only one equilibrium distribution.
- Why is it working? Choose the transition probabilities T to satisfy the detailed balance for our goal distribution $p(z)$.

Markov Chain Monte Carlo - Metropolis-Hasting

- Generalisation of the Metropolis algorithm for nonsymmetric proposal distributions q_k.
- At step τ, draw a sample $z^\star$ from the distribution $q_k(z \mid z^{(\tau)})$ where k labels the set of possible transitions.
- Accept with probability

$$A_k^\star(z, z^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(z^\star)\, q_k(z^{(\tau)} \mid z^\star)}{\widetilde{p}(z^{(\tau)})\, q_k(z^\star \mid z^{(\tau)})}\right)$$

- Choice of proposal distribution critical.
- Common choice : Gaussian centered on the current state.
  - small variance $\rightarrow$ high acceptance rate, but slow walk through the state space; samples not independent
  - large variance $\rightarrow$ high rejection rate

Markov Chain Monte Carlo - Metropolis-Hasting

- Transition probability of this Markov chain is

$$T(z, z') = q_k(z' \mid z)\, A_k(z', z)$$

- Prove that $p(z)$ is the invariant distribution if the detailed balance holds

$$p(z)\, T(z, z') = T(z', z)\, p(z').$$

- Using the symmetry $\min(a, b) = \min(b, a)$ it can be shown that the detailed balance holds

$$\begin{aligned} p(z)\, q_k(z' \mid z)\, A_k(z', z) &= \min\left(p(z)\, q_k(z' \mid z), p(z')\, q_k(z \mid z')\right) \\ &= \min\left(p(z')\, q_k(z \mid z'), p(z)\, q_k(z' \mid z)\right) \\ &= p(z')\, q_k(z \mid z')\, A_k(z, z'). \end{aligned}$$

Markov Chain Monte Carlo - Metropolis-Hasting

- Isotropoic Gaussian proposal distribution (blue)
- In order to keep the rejection rate low, use the smallest standard deviation σ_{min} of the multivariate Gaussian (red) for the proposal distribution.
- Leads to random walk behaviour $\rightarrow$ slow exploration of the state space.
- Number of steps separating states that are approximately independent is $(\sigma_{max}/\sigma_{min})^2$.