Statistical Machine Learning

Christian Walder

Machine Learning Research Group
CSIRO Data61

and

College of Engineering and Computer Science
The Australian National University

Canberra
Semester One, 2020.

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

Outlines

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Part XVI

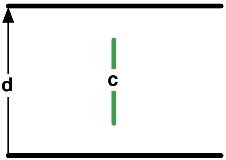
Sampling

Approximating π - Buffon's Needle (1777)

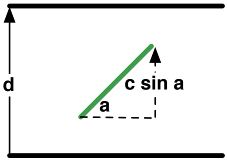
- Drop length c needle on parallel lines distance d apart
- Needle falls perpendicular:
  Probability of crossing the line is c/d.
- Needle falls at an arbitrary angle a:
  Probability of crossing the line $c\sin(a)/d$.
- Every angle is equally probable. Calculate the mean:

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

(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.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Motivation

- For most probabilistic models of practical interest, exact inference is intractable. Need approximation.
- Numerical sampling (Monte Carlo methods).
- Fundamental problem : Find the expectation of some function $f(\mathbf{z})$ w.r.t. a probability distribution $p(\mathbf{z})$

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

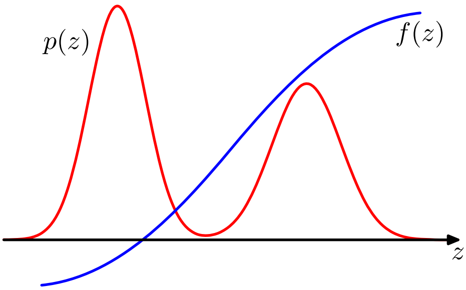
- Key idea : Draw $\mathbf{z}^{(l)}, l = 1 \dots, L$ independent samples from $p(\mathbf{z})$ and approximate the expectation by

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

- 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 apparent sample size.
- For dependent sequences of samples: 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.
- 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 $\mathcal{U}(z \mid 0,1)$ by

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

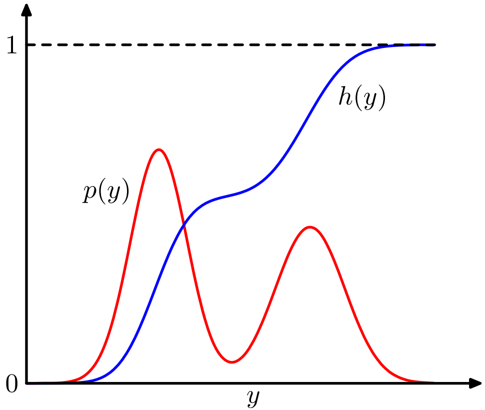
  to obtain samples y distributed according to $p(y)$.

Easy to check that h is then the CDF of y:

$$\begin{aligned} p(y < x) &= p(h^{-1}(z) < x) \\ &= p(z < h(x)) \\ &= h(x). \end{aligned}$$

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

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

- 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.

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 $\mathcal{U}(z \,|\, 0, 1)$)
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

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

Sampling the Multivariate Gaussian

$$
\begin{aligned}
\text{Generate} \quad \mathbf{x} &\sim \mathcal{N}(0, I) \\
\text{let} \quad \mathbf{y} &= A\mathbf{x} + \mathbf{b} \\
\mathbb{E}\,\mathbf{y} &= \mathbf{b} \\
\mathbb{E}\left[(\mathbf{y} - \mathbb{E}\,\mathbf{y})^\top (\mathbf{y} - \mathbb{E}\,\mathbf{y})\right] &= A^\top \mathbb{E}\left[\mathbf{x}^\top \mathbf{x}\right] A \\
&= A^\top A
\end{aligned}
$$

We know $\mathbf{y}$ is Gaussian (why?). So to sample $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ put

- $\mathbf{b} = \mathbf{m}$
- $A^\top A = \Sigma$
- A is the Cholesky factor of Σ

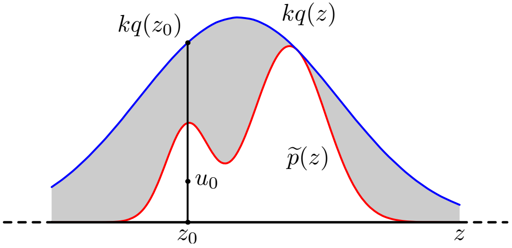
Rejection Sampling

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

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

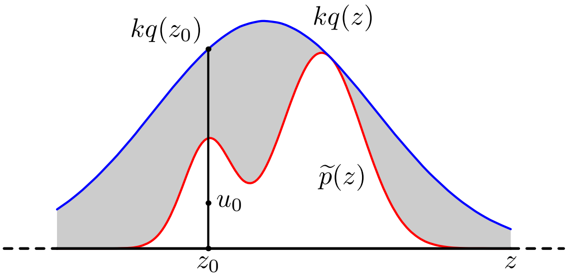
- Assumption 2 : We can draw samples from a simpler distribution $q(z)$ and for some constant k and all z holds

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



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

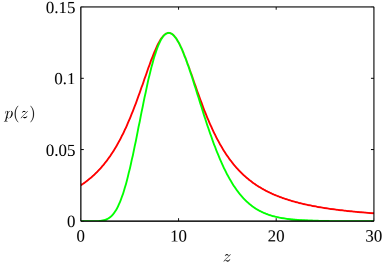
- Consider the Gamma Distribution for $a > 1$

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

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

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

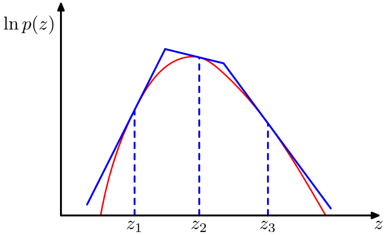
- Samples z from $q(z)$ by using uniformly distributed y and transformation $z = b \tan y + c$ for $c = a - 1$, $b^2 = 2a - 1$ and k as small as possible for $kq(z) \geq \widetilde{p}(z)$.



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.

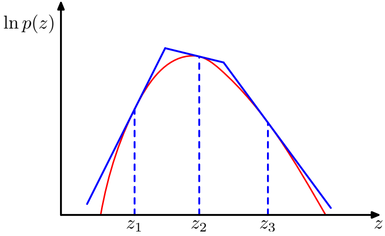


Adaptive Rejection Sampling - Example

- The piecewise exponential distribution is defined as

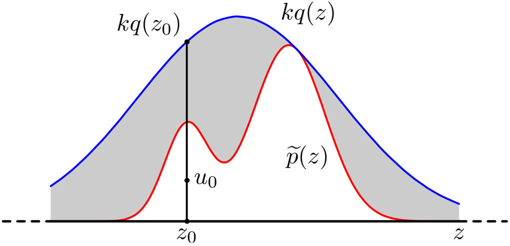
$$p(z) = k_m \lambda_m e^{-\lambda_m (z - z_{m-1})} \qquad \widehat{z}_{m-1,m} < z \leq \widehat{z}_{m,m+1}$$

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

- Directly calculate $\mathbb{E}_p[f(z)]$ w.r.t. some $p(z)$.
- Does not sample $p(z)$ as an intermediate step.
- Again use samples $z^{(l)}$ from some proposal $q(z)$.

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

Importance Sampling - Unnormalised

- Consider unnormalised $p(z) = \frac{\widetilde{p}(z)}{Z_p}$ and $q(z) = \frac{\widetilde{q}(z)}{Z_q}$.
- It follows then that, with $\widetilde{r}_l = \frac{\widetilde{p}(z^{(l)})}{\widetilde{q}(z^{(l)})}$,

$$\begin{aligned}
\mathbb{E}\left[f\right] &= \int f(z)\, p(z)\, \mathrm{d}z \\
&= \frac{Z_q}{Z_p} \int f(z)\, \frac{\widetilde{p}(z)}{\widetilde{q}(z)} q(z)\, \mathrm{d}z \\
&\approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} \widetilde{r}_l f(z^{(l)}).
\end{aligned}$$

- Using the same samples (and $f(z) = 1$) gives

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

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

Importance Sampling - Key Points

- 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

- Goal: sample from $p(z)$.
- Generate a sequence using a Markov chain.
  1. Generate a new sample $z^\star \sim q(z \mid z^{(l)})$, conditional on the previous sample $z^{(l)}$.
  2. Accept or reject the new sample according to some appropriate criterion.

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

  3. For an appropriate proposal and corresponding acceptance criterion, as $l \to \infty$, $z^{(l)}$ approaches an independent sample of $p(z)$.

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

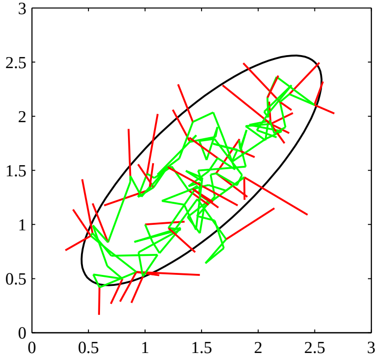
$$A(z^\star, z^{(r)}) = \min\left(1, \frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})}\right)$$

   e.g., let $u \sim$ Uniform$(0, 1)$ and accept if $\frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})} > u$.
3. Unlike rejection sampling we include the previous sample on rejection of the proposal:

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

Metropolis Algorithm - Illustration

- Sampling from a Gaussian Distribution (black contour shows one standard deviation).
- Proposal is $q(\mathbf{z}|\mathbf{z}^{(l)}) = \mathcal{N}(\mathbf{z}|\mathbf{z}^{(l)}, \sigma^2 I)$ with $\sigma = 0.2$.
- 150 candidates generated; 43 rejected.



accepted steps, rejected steps.

Markov Chain Monte Carlo - Why it works

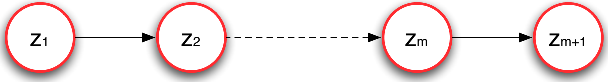
- A First Order Markov Chain is a series of random variables $z^{(1)}, \dots, z^{(M)}$ such that the following property holds

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

- 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

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- 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

$$p^\star(z) = \sum_{z'} T(z', z)\, p^\star(z'). \tag{1}$$

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

Markov Chain Monte Carlo - Why it works

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Detailed balance

$$p^\star(z)\,T(z,z') = p^\star(z')\,T(z',z). \tag{2}$$

  is sufficient (but not necessary) for $p^\star(z)$ to be invariant (to check, put (2) into (1)). (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-Hastings

- 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-Hastings

- Transition probability of this Markov chain is

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

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

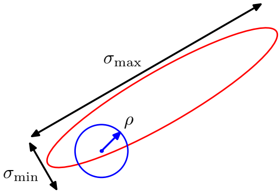
$$p(z)\, T(z, z') = T(z', z)\, p(z').$$

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

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

Markov Chain Monte Carlo - Metropolis-Hastings

- 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 $(\sigma_{max}/\sigma_{min})^2$.



Gibbs Sampling

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Goal: sample from a joint distribution $p(\mathbf{z}) = p(z_1, \dots, z_M)$
- How? sample one variable from the distribution conditioned on all the other variable
- Example : given $p(z_1, z_2, z_3)$
- At step τ we have samples $z_1^{(\tau)}$, $z_2^{(\tau)}$ and $z_3^{(\tau)}$.
- Get samples for the next step $\tau + 1$

$$
\begin{aligned}
z_1^{(\tau+1)} &\sim p(z_1^{(\tau)} \mid z_2^{(\tau)}, z_3^{(\tau)}) \\
z_2^{(\tau+1)} &\sim p(z_2^{(\tau)} \mid z_1^{(\tau+1)}, z_3^{(\tau)}) \\
z_3^{(\tau+1)} &\sim p(z_3^{(\tau)} \mid z_1^{(\tau+1)}, z_2^{(\tau+1)})
\end{aligned}
$$

Gibbs Sampling - Why does it work?

1. $p(\mathbf{z})$ is invariant of each of the Gibbs sampling steps and hence of the whole Markov chain : a) $p(z_i \mid \{\mathbf{z}_{\setminus i}\})$ is invariant because the marginal distribution $p(\mathbf{z}_{\setminus i})$ does not change, and b) by definition each step samples from $p(z_i \mid \{\mathbf{z}_{\setminus i}\})$.
2. Ergodicity: sufficient that none of the conditional distribution is zero anywhere.
3. Gibbs sampling is a Metropolis-Hastings sampling in which each step is accepted.