Lecture 7: Basic Sampling Methods

Notes on random numbers

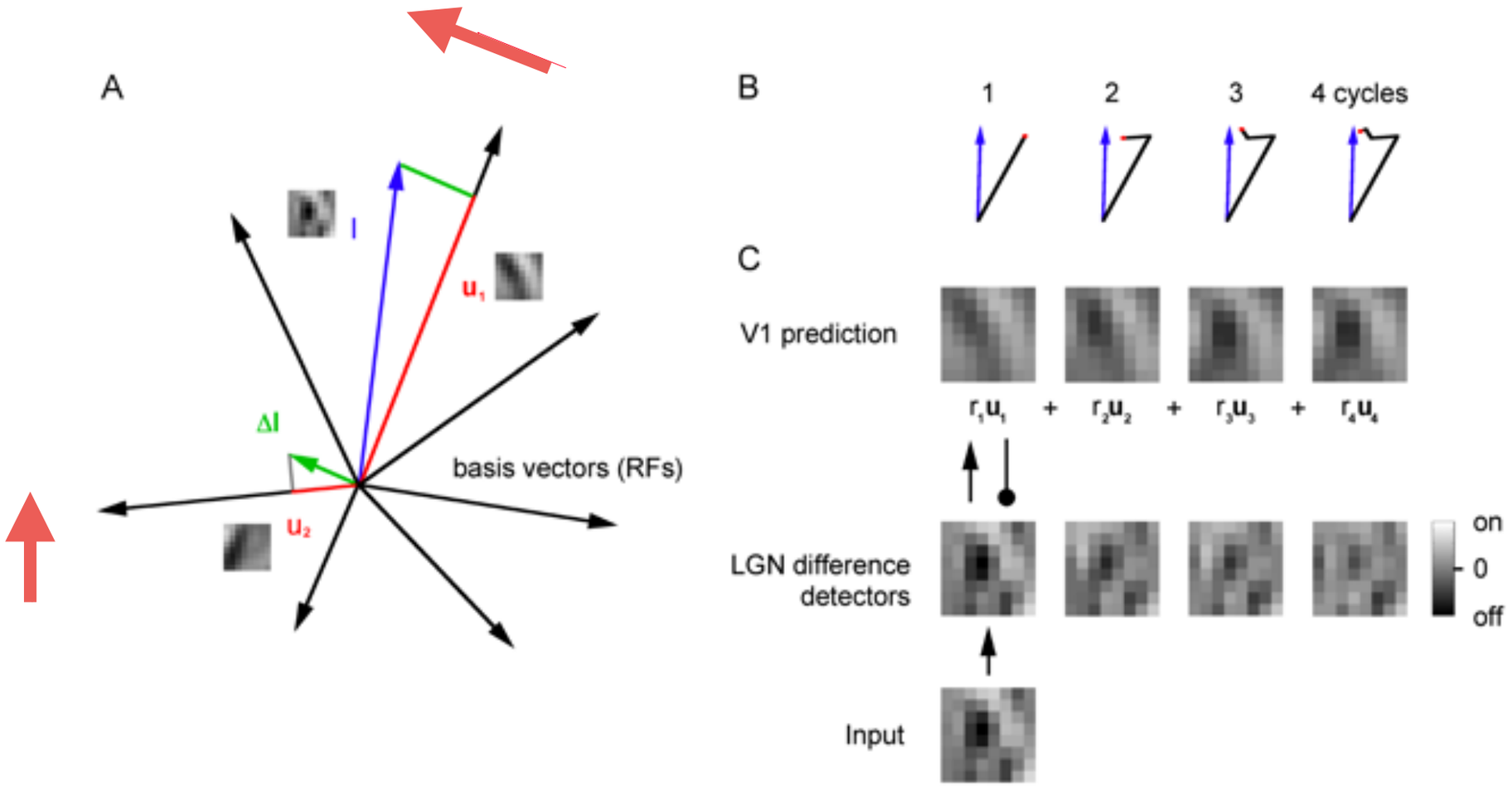
Generating random numbers is not easy.
In fact a sequence is random if a Turing Machine cannot decide whether or not it is random. Reduces to the Halting problem

For numbers between 0 and *m* the following sequence can do a good job for large numbers a and b.

$$X_{n+1} = (aX_n + b) \mod m$$

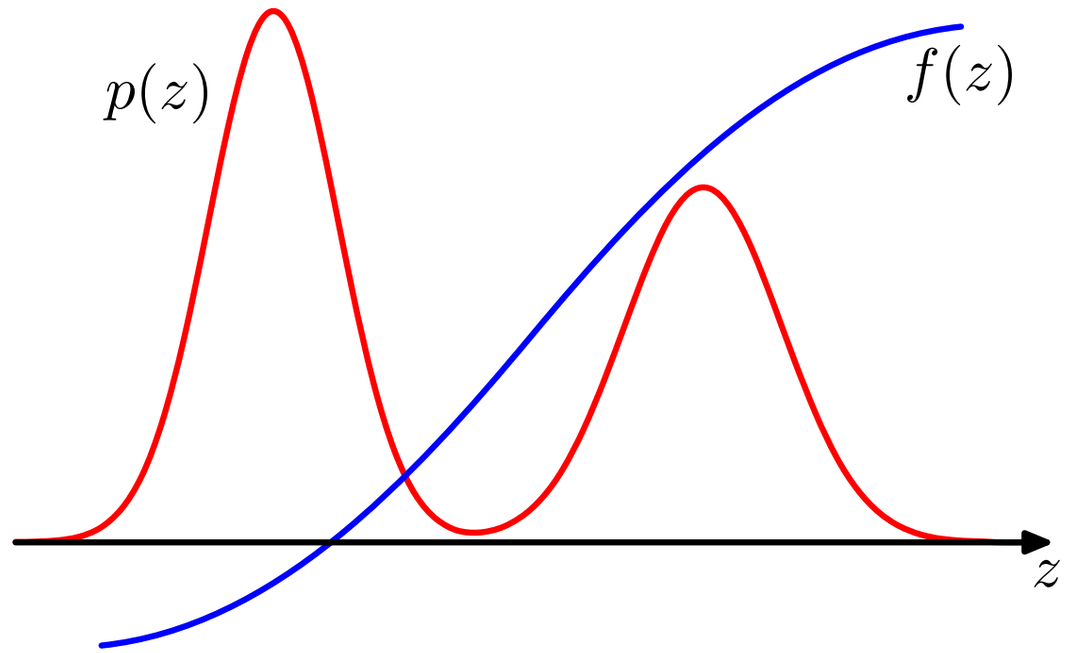
In our subsequent development we'll assume we can sample from a constant probability density function (**pdf**).

A standard Algorithm: Matching Pursuit



Learning algorithm: move RFs
of winning neurons towards inputs

The primary motivation is that computing the expectation of a function takes just a few samples



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

But we want to evaluate *f* from samples

$$\widehat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)}) \qquad\qquad \mathbb{E}\left[\widehat{f}\right] = \ ?$$

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

And variance is given by $\mathbf{var}\left[\widehat{f}\right] = \frac{1}{L}E[f - E[f])^2]$

$$\mathrm{var}\left[\widehat{f}\right] = \mathbb{E}\left[\frac{1}{L}\sum_{m=1}^{L} f(z^{(m)})\frac{1}{L}\sum_{k=1}^{L} f(z^{(k)})\right] - \mathbb{E}[f]^2$$

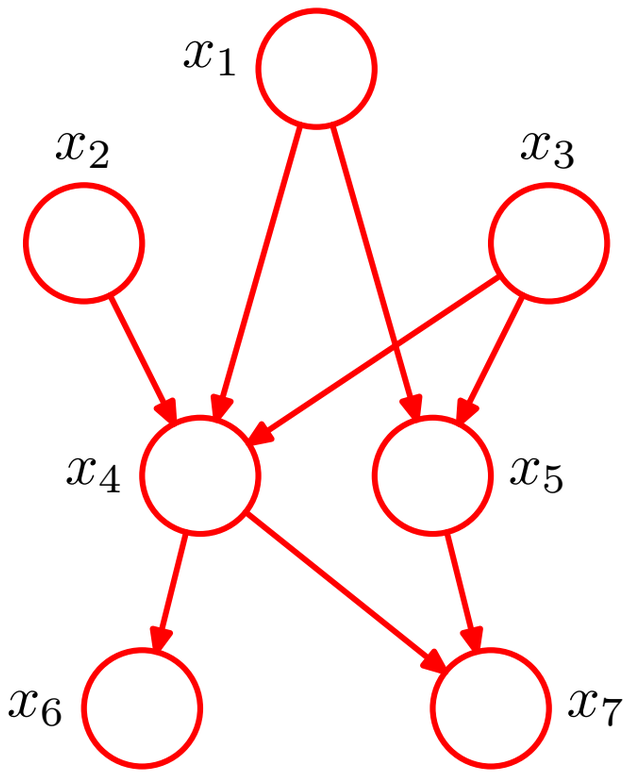
$$\mathbb{E}\left[f(z^{(k)}), f(z^{(m)})\right] = \begin{cases} \mathrm{var}[f] + \mathbb{E}[f^2] & \text{if } n = k, \\ \mathbb{E}[f^2] & \text{otherwise,} \end{cases}$$
$$= \mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f],$$

$$= \frac{1}{L^2}\sum_{m=1}^{L}\sum_{k=1}^{L}\left\{\mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f]\right\} - \mathbb{E}[f]^2$$
$$= \frac{1}{L}\mathrm{var}[f]$$
$$= \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}\left[f\right])^2\right].$$

Ancestral sampling

Faced with sampling from $p(\mathbf{x})$

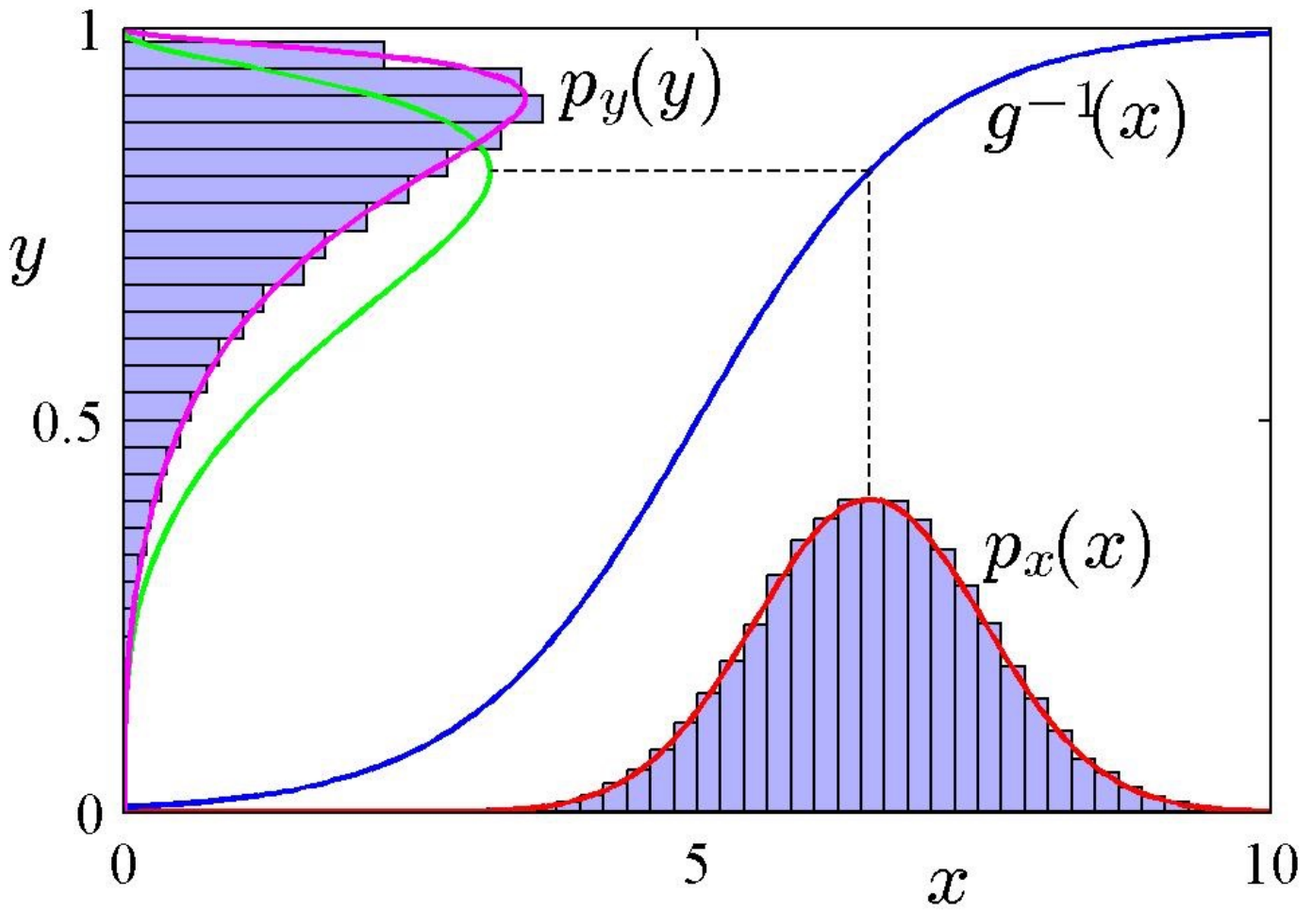
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$



E.g.
Sampling x_1, x_2 and x_3
allows the sampling of x_4

So if we have a way of sampling the conditional distributions,
then we can use the graph structure to drastically reduce the work

Transformed Densities



$$
\begin{aligned}
p_y(y) &= p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= p_x(g(y)) \, |g'(y)|
\end{aligned}
$$

Suppose that z is uniformly distributed over the interval $(0, 1)$, and that we transform the values of z using some function $f(\cdot)$ so that $y = f(z)$. The distribution of y will be governed by

$$p(y) = p(z)\left|\frac{dz}{dy}\right| \tag{11.5}$$

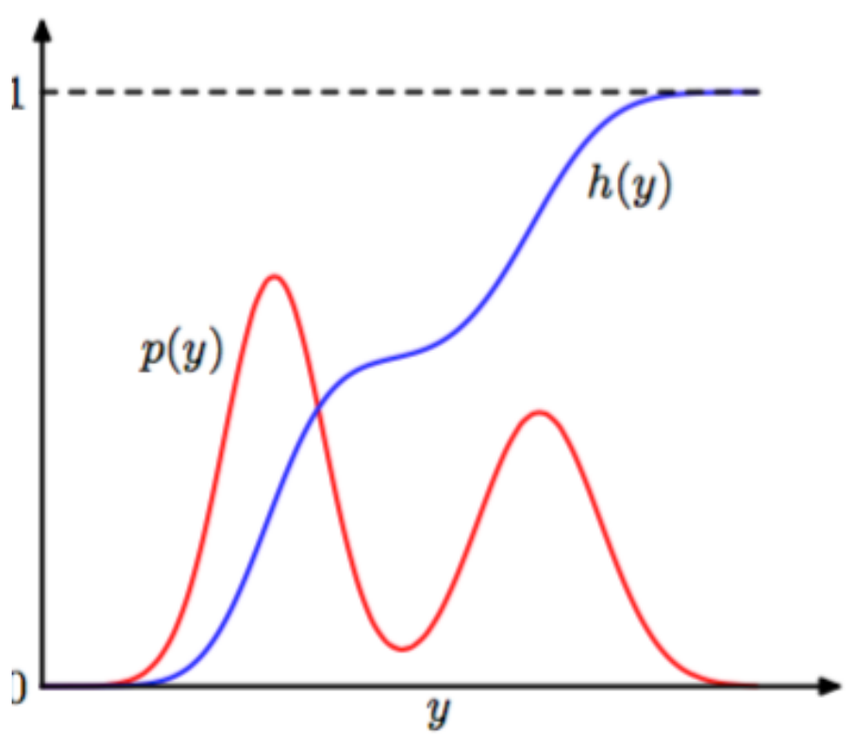
where, in this case, $p(z) = 1$. Our goal is to choose the function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$. Integrating (11.5) we obtain

$$z = h(y) \equiv \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y} \tag{11.6}$$

Now suppose we want to sample from the exponential distribution

$$p(y) = \begin{cases} \lambda e^{-\lambda y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

??



Example 2 Cauchy distribution

$$p(y) = \frac{1}{\pi}\frac{1}{1+y^2}$$

Given $$\int \frac{1}{a^2+u^2}\,\mathrm{d}u = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right) + C$$

where C is a constant, we can integrate the r.h.s. of (11.8) to obtain

$$z = h(y) = \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y} = \frac{1}{\pi}\tan^{-1}(y) + \frac{1}{2}$$

where we have chosen the constant $C = 1/2$ to ensure that the range of the cumulative distribution function is $[0, 1]$.

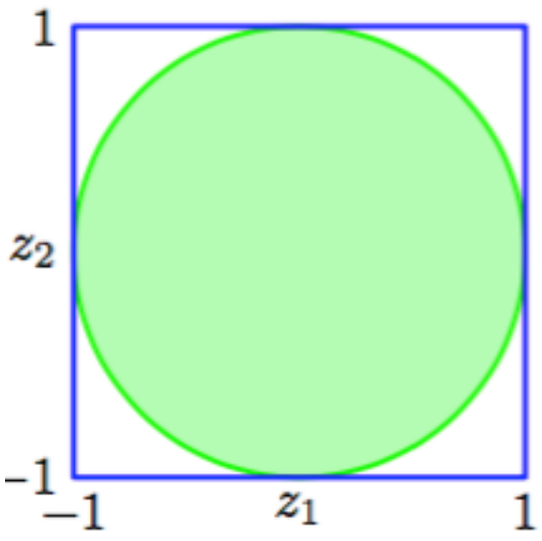
Thus the required transformation function becomes

$$y = h^{-1}(z) = \tan\left(\pi\left(z - \frac{1}{2}\right)\right).$$

Sampling from a Gaussian

First sample uniformly from the square using $z \in [0, 1] \qquad z \longrightarrow 2z - 1$

and only keep the samples if they satisfy $z_1^2 + z_2^2 \leq 1$



Next form

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

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

where $r^2 = z_1^2 + z_2^2$. Then the joint distribution of y_1 and y_2 is given by

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]$$

In doing so, we will find it helpful to make use of intermediary variables in polar coordinates

$$\theta = \tan^{-1}\frac{z_2}{z_1} \tag{299}$$

$$r^2 = z_1^2 + z_2^2 \tag{300}$$

from which it follows that

$$z_1 = r\cos\theta \tag{301}$$

$$z_2 = r\sin\theta. \tag{302}$$

From (301) and (302) we have

$$\frac{\partial(z_1, z_2)}{\partial(r, \theta)} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}$$

and thus

$$\left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\right| = r(\cos^2\theta + \sin^2\theta) = r. \tag{303}$$

From (11.10), (11.11) and (300)–(302) we have

$$y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\cos\theta \tag{304}$$

$$y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\sin\theta \tag{305}$$

which give

$$\frac{\partial(y_1, y_2)}{\partial(r, \theta)} = \begin{pmatrix} -2\cos\theta\,(-2\ln r^2)^{-1/2}\,r^{-1} & -2\sin\theta\,(-2\ln r^2)^{-1/2}\,r^{-1} \\ -\sin\theta\,(-2\ln r^2)^{1/2} & \cos\theta\,(-2\ln r^2)^{1/2} \end{pmatrix}$$

and thus

$$\left|\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| = \left|\frac{\partial(y_1, y_2)}{\partial(r, \theta)}\right|^{-1} = \left(-2r^{-1}(\cos^2\theta + \sin^2\theta)\right)^{-1} = -\frac{r}{2}.$$

Combining this with (303), we get

$$\begin{aligned} \left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| &= \left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| \\ &= \left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\right|\left|\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| = -\frac{r^2}{2} \end{aligned} \tag{306}$$

However, we only retain the absolute value of this, since both sides of (11.12) must be non-negative. Combining this with

$$p(z_1, z_2) = \frac{1}{\pi}$$

which follows from the fact that z_1 and z_2 are uniformly distributed on the unit circle, we can rewrite (11.12) as

$$p(y_1, y_2) = \frac{1}{2\pi}r^2. \tag{307}$$

By squaring the left- and rightmost sides of (304) and (305), adding up the results and rearranging, we see that

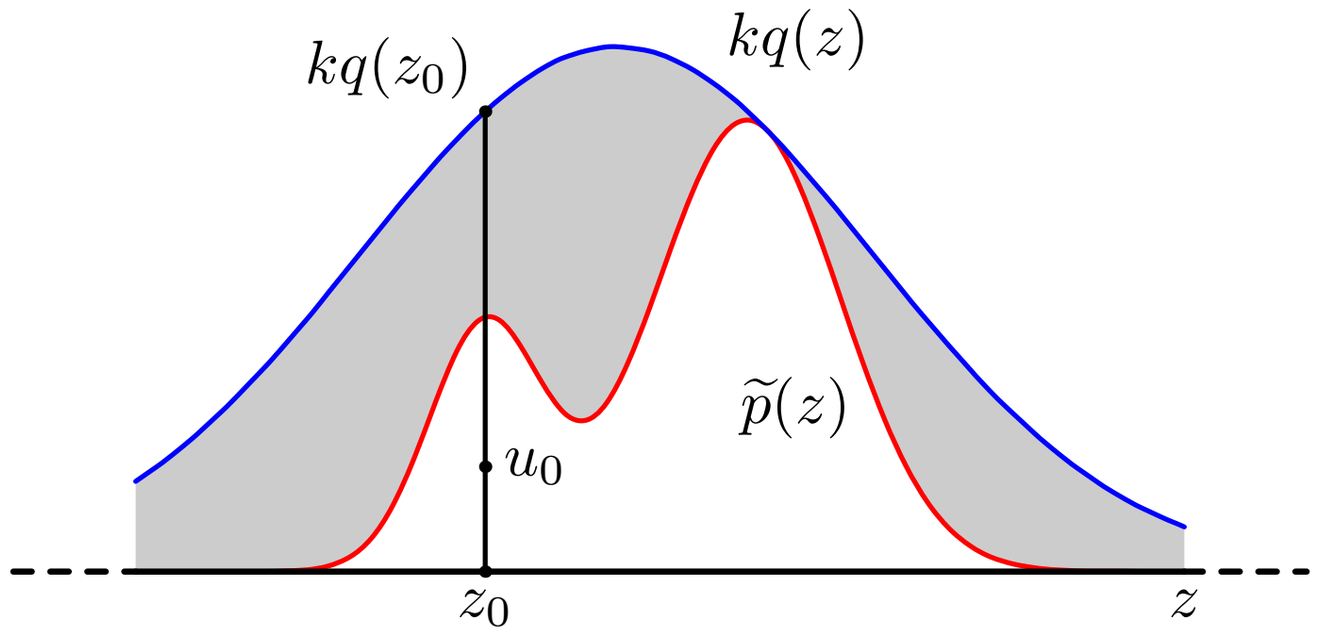
$$r^2 = \exp\left(-\frac{1}{2}\left(y_1^2 + y_2^2\right)\right)$$

$$
\begin{aligned}
p(y_1, y_2) &= p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| \\
&= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]
\end{aligned}
$$

What is the mean and variance?

Rejection Sampling

Know $\hat{p}(z)$ but don't know the normalization factor



$$
\begin{aligned}
p(\text{accept}) &= \int \{\widetilde{p}(z)/kq(z)\}\, q(z)\, \mathrm{d}z \\
&= \frac{1}{k} \int \widetilde{p}(z)\, \mathrm{d}z.
\end{aligned}
$$

Does this strategy do the right thing?

$$p(\text{acceptance}|\mathbf{z}) = \int_0^{\widetilde{p}(\mathbf{z})} \frac{1}{kq(\mathbf{z})}\,\mathrm{d}u = \frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})}.$$

Therefore, the probability of drawing a sample, $\mathbf{z}$, is

$$q(\mathbf{z})p(\text{acceptance}|\mathbf{z}) = q(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})} = \frac{\widetilde{p}(\mathbf{z})}{k}.$$

Integrating both sides w.r.t. $\mathbf{z}$, we see that $kp(\text{acceptance}) = Z_p$, where

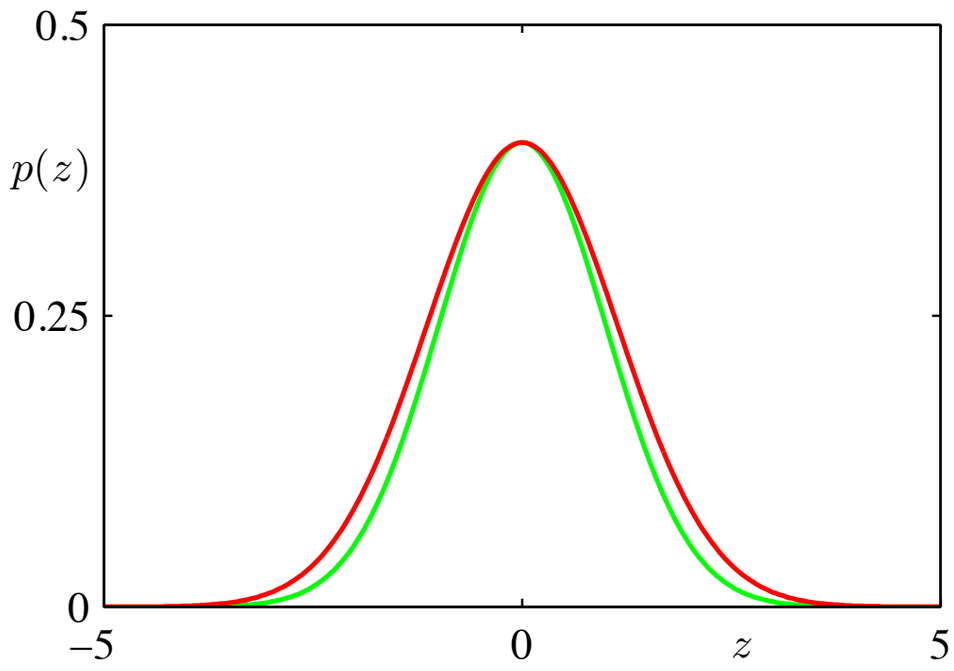
$$Z_p = \int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Using * again,

$$\frac{q(\mathbf{z})p(\text{acceptance}|\mathbf{z})}{p(\text{acceptance})} = \frac{1}{Z_p}\widetilde{p}(\mathbf{z}) = p(\mathbf{z})$$

The curse of dimensionality revisited

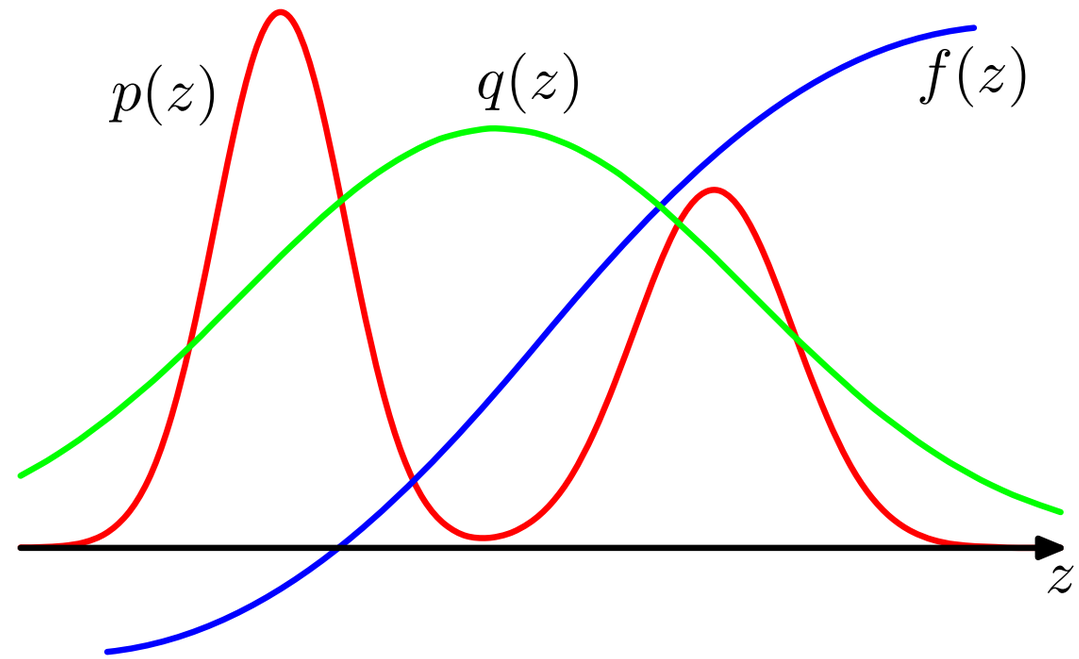
Example: Rejection sampling



Clearly for rejection sampling to be of practical value, we require that the comparison function be close to the required distribution so that the rate of rejection is kept to a minimum. Now let us examine what happens when we try to use rejection sampling in spaces of high dimensionality. Consider, for the sake of illustration, a somewhat artificial problem in which we wish to sample from a zero-mean multivariate Gaussian distribution with covariance $\sigma_p^2\mathbf{I}$, where $\mathbf{I}$ is the unit matrix, by rejection sampling from a proposal distribution that is itself a zero-mean Gaussian distribution having covariance $\sigma_q^2\mathbf{I}$. Obviously, we must have $\sigma_q^2 \geqslant \sigma_p^2$ in order that there exists a k such that $kq(z) \geqslant p(z)$. In D-dimensions the optimum value of k is given by $k = (\sigma_q/\sigma_p)^D$, as illustrated for $D = 1$ in Figure 11.7. The acceptance rate will be the ratio of volumes under $p(z)$ and $kq(z)$, which, because both distributions are normalized, is just $1/k$. Thus the acceptance rate diminishes exponentially with dimensionality. Even if σ_q exceeds σ_p by just one percent, for $D = 1,000$ the acceptance ratio will be approximately $1/20,000$. In this illustrative example the comparison function is close to the required distribution. For more practical examples, where the desired distribution may be multimodal and sharply peaked, it will be extremely difficult to find a good proposal distribution and comparison function.

Importance Sampling

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



When we use the sum for the estimate, q(**z**) d**z** disappears

Importance Sampling Cont.

It will often be the case that the distribution $p(\mathbf{z})$ can only be evaluated up to a normalization constant, so that $p(\mathbf{z}) = \widetilde{p}(\mathbf{z})/Z_p$ where $\widetilde{p}(\mathbf{z})$ can be evaluated easily, whereas Z_p is unknown. Similarly, we may wish to use an importance sampling distribution $q(\mathbf{z}) = \widetilde{q}(\mathbf{z})/Z_q$, which has the same property. We then have

$$\begin{aligned}
\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l f(\mathbf{z}^{(l)}). \qquad (11.20)
\end{aligned}$$

where $\widetilde{r}_l = \widetilde{p}(\mathbf{z}^{(l)})/\widetilde{q}(\mathbf{z}^{(l)})$.

Importance Sampling Cont.

The quantities $r_l = p(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})$ are known as *importance weights*, and they correct the bias introduced by sampling from the wrong distribution. Note that, unlike rejection sampling, all of the samples generated are retained.

where $\widetilde{r}_l = \widetilde{p}(\mathbf{z}^{(l)})/\widetilde{q}(\mathbf{z}^{(l)})$. We can use the same sample set to evaluate the ratio Z_p/Z_q with the result

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int \frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})} q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq \frac{1}{L}\sum_{l=1}^{L} \widetilde{r}_l \end{aligned} \tag{11.21}$$

and hence

$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} w_l f(\mathbf{z}^{(l)}) \tag{11.22}$$

$l=1$

where we have defined

$$w_l = \frac{\widetilde{r}_l}{\sum_m \widetilde{r}_m} = \frac{\widetilde{p}(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})}{\sum_m \widetilde{p}(\mathbf{z}^{(m)})/q(\mathbf{z}^{(m)})}.$$