Variance Reduction I

Lecture #8: Thursday, 23 October 1997
Lecturer: Eric Veach
Scribe: James Davis

1 Transformation of variables in several dimensions

Last lecture we talked about methods to sample random variables according to a given density function. One such method is transformation of variables. We discussed this method in a single dimension previously. Now let us generalize to the multidimensional case.

First, we examine how density functions are affected by transformations. Suppose we have a random variable $\mathbf{X} = (X_1, \ldots, X_n)$ on some domain $\Omega_{\mathbf{X}} \subset \mathbb{R}^n$, and let $p_{\mathbf{X}}(\mathbf{x})$ be the density function of $\mathbf{X}$ (where $\mathbf{x} = (x_1, \ldots, x_n)$ denotes an n-dimensional point). Now let $\mathbf{Y} = T(\mathbf{X})$ be another random variable, where T is a bijective, differentiable function from $\Omega_{\mathbf{X}}$ to some other domain $\Omega_{\mathbf{Y}}$. Then we can ask, what is the density function of $\mathbf{Y}$?

Similar to the single variable case, we know that

$$Pr\{\mathbf{Y} \in T(D)\} = Pr\{\mathbf{X} \in D\}$$

(since T is a bijection). We can write this in terms of the density function $p_{\mathbf{X}}$ and $p_{\mathbf{Y}}$ as

$$\int_{T(D)} p_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} = \int_D p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\ . \tag{1}$$

Now we apply the change of variables $\mathbf{y} = T(\mathbf{x})$ to the left hand side, which yields

$$\int_D p_{\mathbf{Y}}(T(\mathbf{x}))\, |J_T(\mathbf{x})|\, d\mathbf{x} = \int_D p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\ , \tag{2}$$

where $J_T(\mathbf{x})$ is the Jacobian of the transformation $\mathbf{y} = T(\mathbf{x})$,

$$J_T = \begin{bmatrix} \partial y_1/\partial x_1 & \cdots & \partial y_1/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial y_n/\partial x_1 & \cdots & \partial y_n/\partial x_n \end{bmatrix}$$

and $|\cdot|$ denotes the determinant of a matrix.

Finally, since equation (2) holds for any region D, the quantities being integrated must be the same:

$$p_{\mathbf{Y}}(T(\mathbf{x})) \cdot |J_T(\mathbf{x})| = p_{\mathbf{X}}(\mathbf{x}) \; .$$

This tells us how the density functions of $\mathbf{X}$ and $\mathbf{Y}$ are related. We can rewrite this as

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|J_T(\mathbf{x})|}\, p_{\mathbf{X}}(\mathbf{x}) \tag{3}$$

where $\mathbf{x} = T^{-1}(\mathbf{y})$. This identity is useful for figuring out what happens to densities when points are mapped from one coordinate representation to another.

1.1 Example: polar coordinates

Often it is convenient to choose points on the plane using polar coordinates (for example, when the density function we want to sample from is radially symmetric). Polar coordinates are defined by the mapping

$$\begin{array}{rrcl} & (x, y) & = & T(r, \theta) \\ \text{where} & x & = & r\cos\theta \\ & y & = & r\sin\theta \; . \end{array}$$

So, suppose that we choose points according to some density $p(r, \theta)$. What is the corresponding density $p(x, y)$?

The Jacobian of the transformation $(x, y) = (r\cos\theta, r\sin\theta)$ is

$$J_T(r, \theta) = \begin{bmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

and the corresponding determinant is

$$|J_T(r, \theta)| = r\cos^2\theta + r\sin^2\theta = r \; .$$

So according to (3), the densities are related by

$$p(x, y) = \frac{1}{r}\, p(r, \theta) \; .$$

Most often we apply this the other way around: we are given a density function $p(x, y)$ that is measured with respect to surface area ($dA = dx\, dy$), and we want to convert it into polar coordinates to make the sampling easier. In this case we get

$$p(r, \theta) = r\, p(x, y) \; , \tag{4}$$

in other words we need to add a factor of r to the given density $p(x, y)$. A special case of this is sampling uniformly with respect to area: to do this in polar coordinates, we sample according to a density that is proportional to r.

An easy way to handle this kind of manipulation is with differentials. We start with the fundamental identity

$$p_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} = p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

(which corresponds to equation (1)), and write the relationship between the differential volumes $d\mathbf{y}$ and $d\mathbf{x}$ as

$$d\mathbf{y} = |J_T(\mathbf{x})|\, d\mathbf{x}\ .$$

So for polar coordinates we have

$$p(x, y)\, dx\, dy = p(r, \theta)\, dr\, d\theta$$

and

$$dA = dx\, dy = r\, dr\, d\theta$$

(where dA is an element of area). These equations can be manipulated to give the same result (4) between the densities.

1.2 Example: spherical coordinates

Similarly, for sampling in three dimensions we often use spherical coordinates:

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta\ . \end{aligned}$$

The transformation $(x, y, z) = T(r, \theta, \phi)$ has a Jacobian determinant of

$$|J_T| = r^2 \sin\theta\ ,$$

which is often written as the relationship

$$dV = dx\, dy\, dz = r^2 \sin\theta\, dr\, d\theta\, d\phi$$

where dV is an element of volume.

We can also use spherical coordinates to represent points on the surface of a unit sphere, by setting $r = 1$:

$$\begin{aligned} x &= \sin\theta \cos\phi \\ y &= \sin\theta \sin\phi \\ z &= \cos\theta\ . \end{aligned}$$

The reason that this is important in graphics is that a direction ω can be represented as a unit vector (x, y, z). The solid angle of a set of directions is given by the area of the corresponding set of points on the unit sphere. In spherical coordinates, solid angle is measured by[1]

$$d\omega = \sin\theta \, d\theta \, d\phi \, .$$

So, if we are given a density function $p(\omega)$ that is measured with respect to solid angle, which simply means that

$$\int_D p(\omega)\, d\omega$$

is the probability of selecting a direction $\omega \in D$, then the corresponding density with respect to (θ, ϕ) satisfies

$$\begin{array}{rrcl} & p(\theta,\phi)\, d\theta\, d\phi & = & p(\omega)\, d\omega \\ \Rightarrow & p(\theta,\phi) & = & \sin(\theta)\, p(\omega)\, . \end{array}$$

In other words, we need to add an extra factor of $\sin\theta$ in order to sample a density $p(\omega)$ that is given with respect to solid angle.

2 Sampling by transformation in two dimensions

We now give several examples of sampling two-dimensional densities: uniform points on a hemisphere, uniform points in a disc, and cosine-weighted points on a hemisphere. The first two examples use the inversion method (recall that this is a special case of transformation, where we start with uniform random variables U_i), while the third example uses a more general transformation (it starts with uniformly distributed points in a unit disc).

2.1 Marginal and conditional densities

Suppose that we have some joint density function $p(x, y)$. To draw a sample (X, Y) from this density using the inversion method, we need to introduce the idea some new concepts. The *marginal density function* $p(x)$ is defined by

$$p(x) = \int p(x, y)\, dy \, ,$$

which is simply the density function for X alone, while the *conditional density function* $p(y \mid x)$ is defined by

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \, ,$$

which gives the density function for Y once we have fixed a particular value of x.

[1] Note that this identity cannot be derived using a Jacobian determinant, since we have a mapping from $\mathbb{R}^2$ to a surface in $\mathbb{R}^3$. It can be derived from a similar formula for integration on a surface with respect to a particular parameterization.

These are the tools we need to sample from a joint probability distribution, since we can use the marginal density to isolate a single variable's distribution. Then we can determine the value of the other variable according to the conditional density. Examples will be given below.

Another useful concept which we will need later is *conditional expectation* of a random variable $F = f(X, Y)$, which is defined by

$$E_Y[F] = \int f(x, y)\, p(y \mid x)\, dy \ .$$

This is just the usual expression for the expected value, except that x is treated as a parameter (i.e. the result is a function of x).

It is easy to see that the expected value obeys

$$E[F] = E_X[E_Y[F]]$$

where $E[F]$ is the ordinary expected value

$$E[F] = \int f(x, y)\, p(x, y)\, dx\, dy \ ,$$

and E_X denotes the expected value with respect to X (using the marginal density $p(x)$).

2.2 Example: uniform sampling of a hemisphere

Recall the example from last time of sampling a hemisphere uniformly with respect to solid angle. We show how to derive this using marginal and conditional densities (which is really what we did last time, except that we didn't say so explicitly).

To sample uniformly with respect to solid angle, the density function is

$$p(\omega) = 1/(2\pi)$$

for ω in the unit hemisphere centered around around the positive z-axis. (Recall that this hemisphere has a solid angle (area) of 2π.) The corresponding joint probability density function with respect to (θ, ϕ) is

$$p(\theta, \phi) = \frac{1}{2\pi} \sin\theta \qquad 0 \le \theta \le \frac{\pi}{2}, \quad 0 \le \phi \le 2\pi \ .$$

To sample according to this density we first find the marginal density $p(\theta)$:

$$p(\theta) = \int_0^{2\pi} p(\theta, \phi)\, d\phi = \sin\theta$$

Supposing that we know how to sample a random variable θ from this density, we then need to sample ϕ from the conditional density

$$p(\phi \mid \theta) = \frac{p(\theta, \phi)}{p(\theta)} = \frac{1}{2\pi}$$

Since we now have density functions for each variable individually, we can use the single variable methods introduced last lecture to find samples. To use the *inversion method,* we first must find the cumulative distributions $P(\theta_0)$ and $P(\phi \mid \theta)$:

$$\begin{aligned} P(\theta_0) &= \int_0^{\theta_0} p(\theta)\, d\theta = \int_0^{\theta_0} \sin\theta\, d\theta = 1 - \cos\theta_0 \\ P(\phi_0 \mid \theta) &= \int_0^{\phi_0} \frac{1}{2\pi}\, d\phi = \frac{\phi_0}{2\pi} \end{aligned}$$

Now, applying inversion to θ, we set $P(\theta) = 1 - U_1$ (recalling that U_1 and $1 - U_1$ have the same distribution). This gives

$$\begin{aligned} 1 - \cos\theta &= 1 - U_1 \\ \cos\theta &= U_1 \\ \theta &= \cos^{-1} U_1 \end{aligned}$$

Now applying inversion to ϕ, we have

$$\begin{aligned} P(\phi \mid \theta) &= U_2 \\ \frac{\phi}{2\pi} &= U_2 \\ \phi &= 2\pi U_2 \end{aligned}$$

So we can find a point uniformly on a hemisphere by picking two uniform random variables $U_1$and U_2 and solving for (ϕ, θ).

2.3 Example: uniform samples in a disc

As a second example, suppose that we want to sample a point uniformly with respect to area in a unit disc. This corresponds to the joint density function

$$p(x, y) = \begin{cases} 1/\pi & \text{if } x^2 + y^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

which in terms of polar coordinates (r, θ) is

$$p(r, \theta) = \frac{1}{\pi} r$$

where $r \in [0, 1]$ and $\theta \in [0, 2\pi]$. In order to sample r and θ, we first find the marginal density for r:

$$p(r) = \int_0^{2\pi} \frac{1}{\pi} r\, d\theta = 2r$$

which is then integrated to find the cumulative distribution

$$P(r_0) = \int_0^{r_0} 2r\, dr = r_0^2\ .$$

Assuming now that r has already been chosen, the conditional density of θ is

$$p(\theta \mid r) = \frac{p(r,\theta)}{p(r)} = \frac{1}{2\pi}$$

and its cumulative distribution is

$$P(\theta_0 \mid r) = \int_0^{\theta_0} \frac{1}{2\pi}\, d\theta = \frac{\theta_0}{2\pi}$$

Applying the inversion method, we first find r_0 by setting $P(r_0) = U_1$:

$$\begin{aligned} r_0^2 &= U_1 \\ r_0 &= \sqrt{U_1} \end{aligned}$$

We then apply the inversion method again, this time to θ, by setting $P(\theta_0 \mid r_0) = U_2$:

$$\begin{aligned} \frac{\theta_0}{2\pi} &= U_2 \\ \theta_0 &= 2\pi U_2 \end{aligned}$$

A couple of comments are in order. Since both of these examples are trivial, we could easily have switched the order and found the marginal density of θ instead of r. In general, however, the calculations may be easier one way than the other (with respect to being able to do the integrations analytically). Also note that sometimes special tricks can be helpful for sampling from the density functions. For instance, the distribution $P(r_0) = r_0^2$ can be sampled by letting $R = max(U_3, U_4)$, as we showed in the last lecture.

2.4 Example: cosine-weighted hemisphere

We would often like to evaluate an integral such as the *reflectance equation*:

$$L_o(\omega_o) = \int f_r(\omega \to \omega_o) L_i(\omega) \cos\theta \, d\omega \ ,$$

where $L_o(\omega_o)$ is the radiance reflected toward the viewer, L_i is the incoming radiance, f_r is the bidirectional reflectance distribution function, and θ is the angle between the incident vector ω and the surface normal $\mathbf{N}$.

In the case of a diffuse surface $f_r = K_d$, so the integral reduces to

$$L_o(\omega_o) = \int K_d \, L_i(\omega) \, \cos\theta \, d\omega \ .$$

Since the function $L_i(\omega)$ is unknown, a reasonable importance sampling strategy is to choose ω according to a density proportional to $\cos\theta$:

$$p(\omega) \propto \cos\theta \ .$$

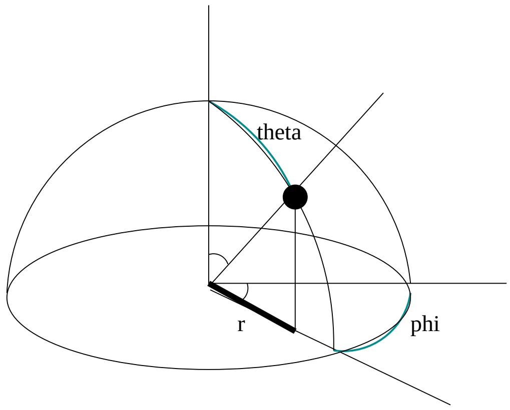
Converting this into a density with respect to (θ, ϕ) we get

$$p(\theta, \phi) \propto \cos\theta \, \sin\theta \; ,$$

where as usual $\theta \in [0, \pi/2]$ and $\phi \in [0, 2\pi]$. By integrating this density we find that the normalization constant is $1/\pi$, so that the desired joint density function is

$$p(\theta, \phi) = \frac{1}{\pi} \cos\theta \, \sin\theta \; .$$

It turns out that we can sample from this density by picking a point uniformly at random on the unit disc, and then projecting it vertically up onto the hemisphere.



Let (r, ϕ) be the polar coordinates of a point in the unit disc (note that we have switched from θ to ϕ). Recall that uniform sampling of the unit disc is represented by the joint density function

$$p(r, \phi) = \frac{1}{\pi} r$$

where $r \in [0, 1]$ and $\phi \in [0, 2\pi]$. We map this into a point ω on the hemisphere with spherical coordinates (θ, ϕ), where the vertical projection implies that r and θ are related by

$$\sin\theta = r \; .$$

(This can be verified by inspecting the diagram above, or by noting that $\sin\theta$ is the distance of ω from the z-axis.)

The Jacobian determinant of the transformation $(r, \phi) = (\sin\theta, \phi)$ is:

$$|J_T| = \begin{vmatrix} \partial r/\partial\theta & \partial r/\partial\phi \\ \partial\phi/\partial\theta & \partial\phi/\partial\phi \end{vmatrix} = \begin{vmatrix} \cos\theta & 0 \\ 0 & 1 \end{vmatrix} = \cos\theta$$

so the densities $p(r, \phi)$ and $p(\theta, \phi)$ are related by

$$p(r, \phi) = \frac{1}{|J_T|} p(\theta, \phi) \; .$$

This can be rearranged to give

$$p(\theta, \phi) \;=\; |J_T|\, p(r, \phi) \;=\; \cos\theta\, \frac{r}{\pi} \;=\; \frac{1}{\pi}\, \sin\theta\, \cos\theta\ .$$

So amazingly enough, this generates the desired distribution over the hemisphere. To summarize, we can generate cosine-weighted samples on a hemisphere by first choosing (X, Y) uniformly within the unit disc (such that $X^2 + Y^2 \leq 1$), and then letting

$$Z = \sqrt{1 - (X^2 + Y^2)}\ .$$

Note that this works for any method of generating samples in the disc, e.g. rejection sampling, or Shirley's mapping from a square to the unit disc (as discussed below).

3 Variance Reduction

The goal of variance reduction is to improve the efficiency of our sampling. The *efficiency* of a Monte Carlo estimator F depends on both its variance $V[F]$, and the time $T[F]$ that is required to evaluate it:

$$\epsilon[F] = \frac{1}{V[F]\, T[F]}$$

According to this definition, F_1 is more efficient than F_2 if it yields a smaller variance in the same running time. (Similarly, an efficient estimator requires less time to achieve a given fixed variance.)

This definition is motivated by the fact that when we take N independent samples, variance is reduced by a factor of N, but computation time is increased by a factor of N. Letting

$$F_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

where the Y_i are independent and identically distributed samples, we have

$$\begin{aligned} V[F_N] &= \frac{1}{N}\, V[F_1] \\ T[F_n] &= N\, T[F_1]\ . \end{aligned}$$

The definition of efficiency takes into account this tradeoff, by saying that F_N and F_1 are equally efficient:

$$\epsilon[F_N] = \epsilon[F_1]\ .$$

The main goal of unbiased Monte Carlo integration is to maximize efficiency: to find estimators whose variance is small and which are fast to evaluate. Methods for doing this are often simply called *variance reduction methods*.

Some of the most powerful variance reductions techniques are based on the idea of analytically integrating a function that is similar to the integrand. There are several ways to take advantage of this idea, of which the most important are *the use of expected values*, *importance sampling*, and *control variates*.

3.1 The use of expected values

Perhaps the most obvious way to reduce variance is to use analytic integration where possible. This idea is commonly referred to as *the use of expected values.* Specifically, it consists of replacing an estimator of the form

$$F = f(X,Y) \,/\, p(X,Y) \tag{5}$$

with one of the form

$$F' = f'(X) \,/\, p(X) \;, \tag{6}$$

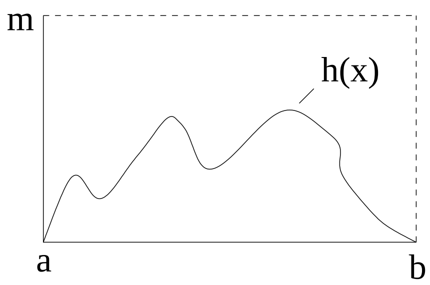
where $f'(x)$ and $p(x)$ are defined by

$$\begin{aligned} f'(x) &= \int f(x,y)\, dy \\ p(x) &= \int p(x,y)\, dy \;. \end{aligned}$$

Thus, to apply this technique we must be able to integrate both f and p with respect to y. We also must be able to generate samples according to the marginal density $p(x)$.

3.1.1 Example

Suppose that we want to integrate a function $h(x)$ on the interval $[a,b]$, such that $0 \le h(x) \le m$ for all $x \in [a,b]$.



Perhaps the simplest technique for doing this is the *hit-or-miss method.* The idea is to pick uniform random points inside the rectangle $[a,b] \times [0,m]$, and count how many fall under the curve $h(x)$. Formally, we can represent this strategy by the two-dimensional function

$$f(x,y) = \begin{cases} 1 & \text{if } 0 \le y \le h(x), \\ 0 & \text{otherwise,} \end{cases}$$

and letting $p(x,y)$ be the uniform density

$$p(x,y) = \frac{1}{m(b-a)}$$

on the rectangle $[a, b] \times [0, m]$. We can then estimate the integral

$$I = \int f(x, y)\, dx\, dy$$

by picking N sample points (X_i, Y_i) according to the density $p(x, y)$, and using the standard estimator

$$\begin{aligned}\hat{I} &= \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i,Y_i)}{p(X_i, Y_i)}\\ &= m(b-a)\cdot\frac{1}{N}\sum_{i=1}^{N} f(X_i, Y_i)\ ,\end{aligned}$$

which is simply the area of the rectangle multiplied by the fraction of the sample points that lie under the curve $h(x)$.

Now, suppose that we apply the expected values technique by integrating analytically with respect to y. We have

$$f'(x) = \int f(x, y)\, dy = \int_0^{h(x)} 1\, dy = h(x)$$

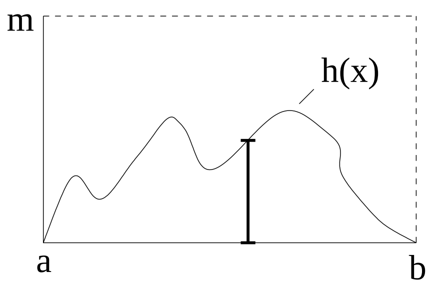
and the marginal density of X is

$$p(x) = \int_0^m p(x, y)\, dy = \frac{1}{b-a}\ .$$

This leads to the new estimator

$$\hat{I}' = \frac{1}{N}\sum_{i=1}^{N}\frac{f'(X_i)}{p(X_i)} = \frac{b-a}{N}\sum_{i=1}^{N} h(X_i)$$

which is simply the standard Monte Carlo estimator for one-dimensional integration that we have been using all along. (It is sometimes called the *sample-mean method.*)



Below, we will show that the expected values method always reduces variance. So in particular, this means that the sample-mean method always gives a lower variance than the hit-or-miss method.

3.1.2 Variance analysis

The name *expected values* comes from the fact that the estimator F' is simply the expected value of F with respect to Y:

$$\begin{aligned} E_Y[F] &= E_Y\left[\frac{f(X,Y)}{p(X,Y)}\right] \\ &= \int \frac{f(X,y)}{p(X,y)}\, p(y \mid X)\, dy \\ &= \int \frac{f(X,y)}{p(X,y)}\, \frac{p(X,y)}{p(X)}\, dy \\ &= \frac{\int f(X,y)\, dy}{p(X)} \\ &= f'(X) \,/\, p(X) \\ &= F' \,. \end{aligned}$$

To determine the variance of F' relative to F, we use the following identity (homework problem #1), where V_Y is the conditional variance with respect to Y, and E_Y is the conditional mean with respect to Y:

$$V[F] = E_X[V_Y[F]] + V_X[E_Y[F]] \,.$$

Using the fact that $F' = E_Y F$, we immediately obtain

$$\begin{aligned} V[F] - V_X[E_Y[F]] &= E_X V_Y[F] \\ V[F] - V[F'] &\geq 0 \\ V[F] &\geq V[F'] \end{aligned}$$

To get from the first to the second line, we used the fact that $E_X V_Y[F]$ is always non-negative (since the variance of a random variable is always non-negative). Furthermore $E_Y[F]$ is a function only of X, so $V[F'] = V_X[F']$.

So, we find that the variance of F' is always less than the variance of F. This suggests that we should always integrate those parts that we can, as we have done in the example above.

3.2 Importance Sampling

Suppose we want to evaluate $I = \int f(x)\, dx$. As we have seen, one method is to choose N random variables X_i uniformly over the domain, and estimate $\hat{I} = (1/N) \sum f(X_i)$. Suppose that instead, we choose the X_i according to some other density function $p(x)$, and define a new random variable

$$F = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad \text{where} \qquad Y_i = \frac{f(X_i)}{p(X_i)} \,.$$

This is unbiased, since

$$E[Y_i] = \int \frac{f(x)}{p(x)}\, p(x)\, dx = \int f(x)\, dx$$

provided that $p(x) > 0$ whenever $f(x) \neq 0$.

Now why would we want to do this? Suppose that $p(x)$ were exactly proportional to $f(x)$. Then we would have:

$$\begin{aligned} p(x) &= cf(x) \\ Y_i &= \frac{f(X_i)}{p(X_i)} = \frac{1}{c} \end{aligned}$$

so that Y_i always has the same value. This means that the variance of Y_i (and F) is zero, which is the best possible.

Unfortunately we can't do this exactly, since it turns out that in order for p to be normalized, we need

$$c = \frac{1}{\int f(x)\, dx}\,.$$

So if we could evaluate c, then we would already know the answer to our original problem, and there would be no need for sampling.

However, if we can find a density p that is approximately proportional to f (and that we can sample from), we can often reduce the variance of our estimator. This principle is called *importance sampling*. It should be noted that the argument above assumes that $f(x) \geq 0$, since $p(x)$ can't be negative. If $f(x)$ is sometimes negative, it turns out that the best thing is to make $p(x)$ proportional to $|f(x)|$, but this does not give zero variance even if we can sample this way exactly.

Importance sampling is a very useful technique in graphics. It is used to sample functions that are non-uniform, e.g. a peaked specular lobe of a BRDF, or a solid angle containing small light source. The cosine-weighted hemisphere was a simple example of this, where we sampled according to the $\cos\theta$ factor that appears in the integral for reflected radiance. Other examples are given in the homework problems.

3.3 Control Variates

Again suppose we want to evaluate $I = \int f(x)\, dx$. The idea of *control variates* is find some function $\tilde{f}$ similar to f that we can integrate analytically, and then subtract it. This is done by rewriting the integral as

$$I = \int \tilde{f}(x)\, dx + \int f(x) - \tilde{f}(x)\, dx$$

where the integral on the left can be evaluated exactly. This leads to a new estimator of the form

$$F' = \int \tilde{f}(x)\, dx + \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i) - \tilde{f}(X_i)}{p(X_i)}$$

where $p(x)$ is the density that the X_i are sampled from.

The new estimator F' will have a lower variance than the basic estimator F whenever

$$V\left[\frac{f(x_i) - \tilde{f}(X_i)}{p(X_i)}\right] \leq V\left[\frac{f(X_i)}{p(X_i)}\right] \tag{7}$$

Notice that given some analytic approximation $\tilde{f}$, we also have the option of using it for importance sampling. This is done by letting $p(x)$ be the density function

$$p(x) = \frac{\tilde{f}(x)}{\int \tilde{f}(x)\,dx}\ ,$$

assuming that we can find a way to generate samples with this density.

However, it doesn't make sense to use $\tilde{f}$ for control variates and importance sampling at the same time, since if $\tilde{f}$ is proportional to p, then the two estimators in equation (7) differ by a constant and their variance is the same. Thus if $\tilde{f}$ is already being used for importance sampling, it is not helpful to use it as a control variate as well.

So if we have found a function $\tilde{f}$ that is similar to f, which should we use: importance sampling, or control variates? The answer depends on the exact functions involved, but in general, if $f - \tilde{f}$ is a nearly constant function, then it is better to use control variates. If instead the ratio $f/\tilde{f}$ is nearly constant, then it is better to use importance sampling.

Another consideration is that with control variates, $\tilde{f}$ needs to be an approximation to the *entire integrand*. For example, if we wanted to apply control variates to the integral

$$\int_\Omega f_r(\omega \to \omega_o)\, L_i(\omega)\, \cos\theta\, d\omega\ ,$$

then we would need to approximate both the BRDF and incident light. It would not make sense to use the BRDF alone as a control variate, since it is scaled by an arbitrary factor L_i in the integral. (Subtracting the BRDF effectively replaces $L_i(\omega)$ by $L_i(\omega) - 1$, and with no knowledge of the magnitude of L_i there is no reason to think that this is an improvement.) With importance sampling, on the other hand, using the BRDF as a sampling density would be perfectly reasonable.

3.3.1 Example

In ray tracing, biased results can be avoided by choosing the recursion depth randomly. Typically this is done by defining a probability $1 - q$ that the recursion will stop at each depth, and increasing the sample weight by a factor of $1/q$ in the event that the recursion continues. (This technique is called *Russian roulette.*)

Typically, when the recursion stops we just return a value of 0. But suppose that we had some estimate L^* of the ambient light in the scene. In this case, we could apply the control variates technique, by returning the value

$$\int f_r(\omega \to \omega_o)\, L^*\, \cos\theta\, d\omega$$

instead. We compensate for this by rewriting the integral as

$$\begin{aligned} L_o(\omega_o) &= \int f_r(\omega \to \omega_o)\, L_i(\omega)\, \cos\theta\, d\omega \\ &= \int f_r(\omega \to \omega_o)\, L^*\, \cos\theta\, d\omega + \int f_r(\omega \to \omega_o)\, (L_i(\omega) - L^*)\, \cos\theta\, d\omega\ , \end{aligned}$$

i.e. if the recursion continues, we subtract L^* from the actual radiance samples $L_i(\omega)$.

One needs to be careful of course, since $L_i(\omega) - L^*$ could produce a negative value, and some systems may be unhappy with the concept of negative radiance. Another consequence is that control variates can lead to large relative errors when estimating values close to zero, for instance pixels in a dark region of the image.

4 Stratified Sampling

Stratified sampling is another way to reduce variance. We only introduce the basic idea here, and discuss its variance reduction properties in the next lecture.

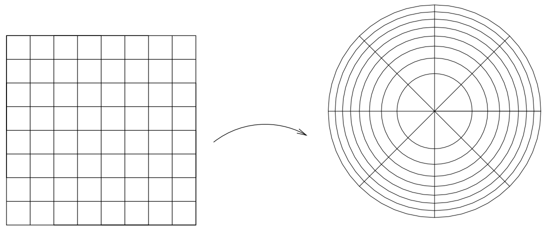
Given a domain Ω, the idea of stratified sampling is to split the domain into non-overlapping subregions $S_1, ..., S_k$ called *strata* such that

$$\bigcup_{i=1}^{n} S_i = \Omega\ .$$

We then take a fixed number of samples n_i to estimate the integral in each region, and add up the results. For example in graphics we often supersample a pixel, by breaking it into a $k \times k$ grid and choosing a random point in each of these strata. This is better than choosing k^2 points completely at random within the pixel, as we will see next time.

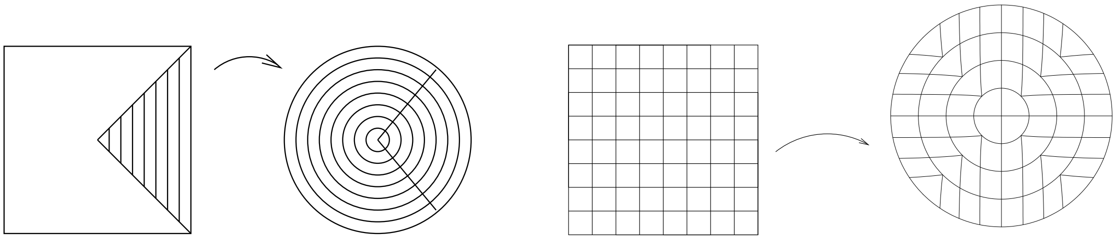
For now, let's consider how to pick a good set of strata. For instance, we found earlier that we could sample a disc uniformly with the mapping

$$\begin{aligned} r &= \sqrt{U_i} \\ \theta &= 2\pi U_2 \end{aligned}$$



We can see that this results in some long thin regions. As we will show next time, it is better for the strata to have compact shapes, since this tends to result in the integrand having a more constant value over each stratum.

Peter Shirley proposed the following alternative mapping from a square to the unit disc, which results in more uniform strata. This mapping takes concentric squares into concentric circles, while preserving their relative area.



Mathematically, the mapping takes a point (u, v) in the square $[-1, 1] \times [-1, 1]$ to a point (r, θ) in the unit disc. It is defined as

$$\left.\begin{array}{rcl} r &=& u \\ \theta &=& \frac{\pi}{4}\frac{v}{u} \end{array}\right\} \quad \text{when } |u| \geq |v|,$$
$$\left.\begin{array}{rcl} r &=& v \\ \theta &=& \frac{\pi}{4}\left(2 - \frac{u}{v}\right) \end{array}\right\} \quad \text{otherwise.}$$

This certainly looks like a better stratification of a disc, but how do we know that these strata have uniform area? We start with a sample (U, V) whose density function is

$$p(u, v) = 1/4 \ ,$$

since (U, V) is chosen uniformly on $[-1, 1] \times [-1, 1]$. We then transform this to a sample (R, θ) in polar coordinates using the mapping defined above. If we consider just the case $|u| \geq |v|$, the Jacobian determinant for the transformation $(r, \theta) = T(u, v)$ is

$$|J_T| = \begin{vmatrix} \partial r/\partial u & \partial r/\partial v \\ \partial \theta/\partial u & \partial \theta/\partial v \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \text{irrelevant} & \pi/(4u) \end{vmatrix} = \frac{\pi}{4r} \ .$$

So, the density $p(r, \theta)$ in terms of $p(u, v)$ is

$$p(r, \theta) = \frac{1}{|J_T|}\, p(u, v) = \frac{4r}{\pi}\frac{1}{4} = \frac{1}{\pi} r \ ,$$

which is the density for sampling uniformly on a unit disc, as we saw in Section 2.3. If we take this one step further and map (r, θ) to a point (x, y), the corresponding density function is

$$p(x, y) = \begin{cases} 1/\pi & \text{if } x^2 + y^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

So, Shirley's mapping sends strata of equal area in (u, v)-space into strata of equal area in (x, y)-space. This is called a *constant Jacobian mapping*, since the Jacobian determinant of the combined transformation $(x, y) = T(u, v)$ is $|J_T| = \pi/4$.