Sampling Random Variables

Lecture #7: Tuesday, 21 October 1997
Lecturer: Eric Veach
Scribe: Menelaos Karavelas
Reviewer: Lucas Pereira

1 Properties of Monte Carlo estimators

Recall from the previous lecture that the purpose of a Monte Carlo estimator is to approximate the value of some quantity of interest Q. Most commonly Q will be an integral of the form

$$Q = \int_\Omega f(x)\, dx \tag{1}$$

where Ω is some arbitrary domain. An *estimator* is then defined to be a function of the form

$$F_N = F_N(X_1, \ldots, X_N)$$

(where the X_i are random variables), such that the mean of F_N is a usable approximation of the quantity we want to estimate.

The estimator F_N is said to be *unbiased* if

$$E[F_N] = Q \quad \text{for all } N.$$

For example, suppose that Q is given by (1), and let X be a random variable on Ω with density $p(x)$. Furthermore, suppose that $p(x) > 0$ whenever $f(x) \neq 0$. Then

$$F = \frac{f(X)}{p(X)}$$

is an unbiased estimator of Q, since

$$E[F] = \int_\Omega \frac{f(x)}{p(x)}\, p(x)\, dx = I.$$

Otherwise, the quantity

$$\beta[F_N] = E[F_N] - Q$$

is called the *bias*. If the bias goes to zero as the number of samples N is increased, then the estimator is called *consistent*:

$$Pr\left\{ \lim_{N \to \infty} F_N = Q \right\} = 1,$$

i.e., F_N converges in probability to Q.

1.1 Example: a biased, consistent estimator

Recall the example of antialiased sampling of a 1D pixel from the previous lecture. The quantity to be estimated is the pixel value I, which is defined by an integral

$$I = \int_0^1 w(x) f(x)\, dx. \tag{2}$$

Here f is the image function on the domain $[0,1]$, and w is the filter function, which satisfies

$$\int_0^1 w(x)\, dx = 1.$$

A common way to evaluate the integral (2) is to use the estimator:

$$F_N = \frac{\sum_{i=1}^N w(X_i) f(X_i)}{\sum_{i=1}^N w(X_i)}, \tag{3}$$

where the X_i are independent uniformly distributed samples on $[0,1]$.

We first show that this estimator is in general biased. Suppose that just one sample is taken ($N = 1$). Then

$$F_1 = \frac{w(X_1) f(X_1)}{w(X_1)} = f(X_1),$$

and

$$E[F_1] = \int_0^1 f(x) p(x)\, dx = \int_0^1 f(x)\, dx,$$

noting that $p(x) = 1$ is the density function for X_1 on the domain $[0,1]$. Since this expected value is in general different from I (see (2)), the estimator F_1 is biased.

However, the estimator F_N is consistent since by the law of large numbers

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_i w(X_i) f(X_i) &= \int_0^1 w(x) f(x)\, dx, \\ \lim_{N\to\infty} \frac{1}{N} \sum_i w(X_i) &= \int_\Omega w(x)\, dx \;=\; 1. \end{aligned}$$

Thus we have

$$\lim_{N\to\infty} E[F_N] = \int_0^1 w(x) f(x)\, dx,$$

i.e. we obtain the right answer as N becomes large.

A numerical example. To further understand how this estimator behaves, suppose that we have an image composed of identical pixels, where the left half of each pixel is white and the right half of is black. This corresponds to the image function

$$f(x) = \left\{ \begin{array}{ll} 1, & x \in [0, \frac{1}{2}) \\ 0, & x \in [\frac{1}{2}, 1] \end{array} \right. ,$$

where 1 represents white and 0 represents black. Suppose now that we use a filter $w(x)$ that puts a lot of weight on the right half of the each pixel, specifically:

$$w(x) = \begin{cases} 2 - 2\epsilon, & x \in [0, \frac{1}{2}) \\ 2\epsilon, & x \in [\frac{1}{2}, 1] \end{cases}, \tag{4}$$

where ϵ is a small number in the range $0 < \epsilon < 1$. Notice that this weighting function integrates to one, as required.

Then it is easy to see that the correct value for the filtered pixel is:

$$I = \int_0^1 f(x)w(x)\,dx = \frac{1}{2}(1)(2\epsilon) + \frac{1}{2}(0)(2 - 2\epsilon) = \epsilon,$$

and this is exactly the value that we want our Monte Carlo method to estimate (i.e. $Q = I$). However, if we use the biased estimator described above to evaluate the integral I at each pixel, then using just one sample X_1 per pixel leads to

$$F_1 = \begin{cases} 1, & X_1 \in [0, \frac{1}{2}) \\ 0, & X_1 \in [\frac{1}{2}, 1] \end{cases}.$$

This produces an image with a checkerboard appearance, where approximately half the pixels are black and half are white (see Figure 1). The resulting image has a mean value of 1/2, as opposed to the correct filtered image which has a mean value of ϵ (a uniform image of low gray-level).

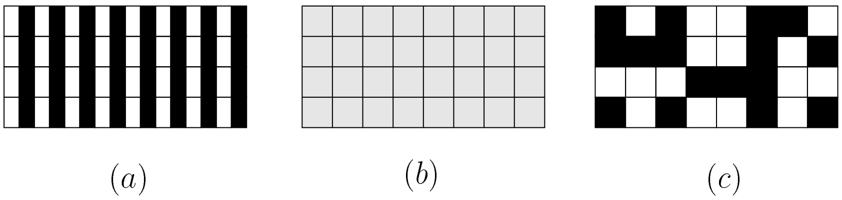

Figure 1: (a) A 4×8-pixel image, each pixel of which is half white and half black. (b) The filtered image using the filter w in (4), which has a mean value of ϵ. (c) The filtered image using the Monte Carlo estimator (3) with $N = 1$, which has a mean value of 1/2.

1.2 Mean squared error

The main reason for preferring unbiased estimators is that it is easier to estimate the error. Typically our goal is to minimize the *mean squared error* (MSE), defined by

$$MSE[F] \;=\; E[(F - Q)^2] \tag{5}$$

(where we have dropped the subscript N). We would like to find an estimator such that the mean squared error is small, relative to Q^2:

$$MSE[F_N] \ll Q^2.$$

In general, the mean squared error can be rewritten as

$$\begin{aligned} MSE[F] &= E[(F-Q)^2] \\ &= E[(F-E[F]+E[F]-Q)^2] \\ &= E[(F-E[F])^2] + 2E[F-E[F]](E[F]-Q) + (E[F]-Q)^2 \\ &= V[F] + \beta[F]^2 \end{aligned}$$

(since $E[F-E[F]]=0$). Thus, to estimate the error we must have an upper bound on the possible bias. In general, this requires additional knowledge about the estimand Q, and it is often difficult to find a suitable bound.

On the other hand, for unbiased estimators we have $E[F]=Q$, so that the mean squared error is identical to the variance:

$$MSE[F] \;=\; V[F] \;=\; E[(F-E[F])^2] \;.$$

This makes it far easier to obtain error estimates, by simply taking several independent samples. Letting $Y_1,\ldots,Y_N$ be independent samples of an unbiased estimator Y, and letting

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

as before (which is also an unbiased estimator), then the quantity

$$\hat{V}[F_N] \;=\; \frac{1}{N-1}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^2\right) - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2\right\}$$

is an unbiased estimator of the variance $V[F_N]$ (see Kalos and Whitlock). Thus, error estimates are easy to obtain for unbiased estimators.

2 Sampling random variables

There are a variety of techniques for sampling random variables, including:

- the method of *transformation of variables* (also known as the *inversion method*),
- the *rejection method*, and
- the *Metropolis method* (which will be discussed in another lecture).

There are also a variety of tricks that are useful for sampling particular distributions, such as the normal distribution. We give some simple examples of such tricks below.

With all of these methods, it is assumed that a supply of random variables $U_1, U_2, \ldots$ is available, where U_i is independent and uniformly distributed on $[0,1]$ ($U_i \sim U[0,1]$).

2.1 Transformation of variables

Our goal is to generate a random variable Y that is distributed according to a given density function $p(y)$. Let $P(y)$ be the cumulative distribution of Y, that is,

$$P(y) = Pr\{Y \leq y\}.$$

The *inversion method* consists of first choosing a random variable U that is uniformly distributed on $[0, 1]$, and then finding a value of Y such that $P(Y) = U$. Equivalently, we can write

$$Y = P^{-1}(U),$$

where P^{-1} denotes the inverse function of P. This procedure is shown graphically in Figure 2.

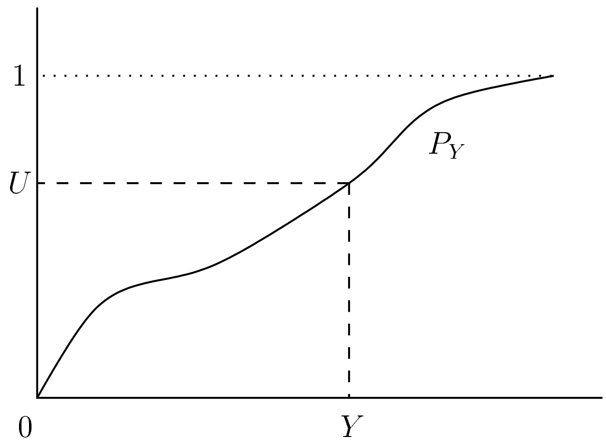

Figure 2: The inversion method.

It is easy to check that Y has the desired distribution, since

$$\begin{aligned} Pr\{Y \leq y\} &= Pr\{P(Y) \leq P(y)\} \\ &= Pr\{U \leq P(y)\} \\ &= P(y), \end{aligned}$$

where the first equality holds because P is a non-decreasing function, and the last equality holds because U is uniform on $[0, 1]$.

An example. To see how the method works, suppose we want to evaluate the integral

$$I = \int_0^\infty e^{-cx} g(x)\, dx.$$

Integrals of this type arise in volume rendering: x represents the distance along a ray, $g(x)$ represents the amount of light sent toward the viewer at the distance x, and e^{-cx} represents the attentuation of this light as it travels through a medium of constant opacity.

To evaluate this integral, we would like to sample a random distance X according to the density

$$p(x) = ce^{-cx}$$

(the factor of c is required so that p integrates to one). The cumulative distribution of X is

$$P(x_0) = \int_0^{x_0} p(x)\,dx = \int_0^{x_0} ce^{-cx}\,dx = -\frac{c}{c}\,e^{-cx}\Big|_0^{x_0} = 1 - e^{-cx_0}.$$

Thus using the inversion method, we have

$$\begin{aligned} 1 - e^{-cX} = U \quad &\Rightarrow \quad -cX = \ln(1-U) \\ &\Rightarrow \quad X = -(1/c)\,\ln(1-U). \end{aligned}$$

Equivalently, we could have started by setting $P(X) = 1 - U$, since if U is uniformly distributed on $[0,1]$ then so is $1-U$. This yields the slightly simpler result

$$X = -(1/c)\,\ln U.$$

Note that using this strategy, there is a high probability of sampling small values of X (due to the logarithm calculation). This corresponds to the fact that attenuation factor e^{-cx} goes to zero rapidly with increasing values of x.

If we sample N different distances X_i in this way, i.e.

$$X_i = -(1/c)\,\ln U_i$$

for independent values of $U_i \sim U[0,1]$, then corresponding Monte Carlo estimator for the integral I is

$$\begin{aligned} F_N &= \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-cX_i}g(X_i)}{ce^{-cX_i}} \\ &= \frac{1}{cN}\sum_{i=1}^{N} g(X_i). \end{aligned}$$

So, the whole technique is very easy to apply.

Note that the inversion technique can easily be extended to several dimensions, by computing marginal and conditional distributions and inverting each dimension separately. (Examples will be given in the next lecture.)

Transformation of variables. There is a generalization of the inversion method called *transformation of variables*. Suppose that we start with a random variable X, with a known density function $p_X(x)$. (The inversion method was a special case of this with $X = U$.) Now let $Y = y(X)$, where y is some known function. Then it is natural to ask, what is the density function of Y, in terms of the density function of X? (Note that Y is also a random variable.)

Let P_Y be the cumulative distribution of Y, that is,

$$P_Y(y) = Pr\{Y \leq y\}.$$

We will assume that $y(x)$ is a continuous, non-decreasing function of x (which implies that $y'(x) \geq 0$). Then clearly we have

$$Pr\{Y \leq y(x)\} = Pr\{X \leq x\},$$

which implies that

$$P_Y(y) = P_X(x)$$

where y = y(x). Differentiating with respect to x, we obtain the following expression that relates the density function of X with that of Y:

$$p_Y(y)\,(dy/dx) = p_X(x)$$

which implies that

$$p_Y(y) = |dy/dx|^{-1}\ p_X(x), \tag{6}$$

where the absolute value signs allow for possibility that y is non-increasing rather than non-decreasing. This relationship is often written as

$$|p_Y(y)\,dy| = |p_X(x)\,dx|,$$

where as always, x and y are implicitly related by $y = y(x)$.

To illustrate how this identity is used, suppose that X is distributed according to the density $p_X(x) = 2x$ over the domain $[0, 1]$, and let $Y = \sin(X)$. What is the density function for Y? We have

$$\begin{aligned} p_Y(y) &= |d\sin(x)/dx|^{-1} p_X(x) \\ &= |\cos(x)|^{-1} 2x \\ &= 2\arcsin(y) \Big/ \cos(\arcsin(y)) \\ &= 2\arcsin(y) \Big/ (1 - y^2)^{1/2}, \end{aligned}$$

where we have used the facts that (a) $\cos(x) > 0$ on $[0, 1]$, (b) $y = \sin(x)$ implies $x = \arcsin(y)$, and (c) $\cos(x) = (1 - \sin^2(x))^{1/2}$ on $[0, 1]$.

Deriving a transformation. Equation (6) suggests that given a random variable X, we can sample from an arbitrary density p_Y by finding an appropriate transformation $Y = y(X)$. In one dimension, there is a straightforward procedure for deriving such a transformation: we simply find a value of Y such that

$$P_Y(Y) = P_X(X),$$

or in other words we let y be the function

$$y(x) = P_Y^{-1}(P_X(x)).$$

(This assumes that we can find analytic expressions for the cumulative distributions P_X and P_Y.) The reason this works is that if X is any random variable, then the random variable $Z = P_X(X)$ has the uniform distribution:

$$Pr\{Z \leq x\} = Pr\{P_X(X) \leq x\} = Pr\{X \leq P_X^{-1}(x)\} = P_X(P_X^{-1}(x)) = x,$$

for any $x \in [0, 1]$. Thus, setting $P_Y(Y) = P_X(X)$ is equivalent to setting $P_Y(Y) = U$.

In one dimension, the main benefit of the transformation method over the simpler inversion method is that we can start with non-uniform samples. So for example, if we already know to generate a sample X with a normal distribution, then we can apply transformations to obtain various other distributions p_Y. In higher dimensions, the transformation method can be used to map random points from one domain to another (e.g. from a disc to a hemisphere). In this case, the factor $|dy/dx|$ in equation (6) becomes the *Jacobian determinant*, as we will see in the next lecture.

The main advantage of the inversion/transformation technique compared to the methods below is that it allows samples to be stratified easily (see the next lecture). This is done by stratifying the canonical parameter space $[0, 1]^s$, and then mapping these samples into the given domain Ω with the desired density p_Y. Another advantage is that the technique has a fixed cost per sample, which can easily be estimated. The main disadvantage is that the density $p(x)$ must be integrated analytically, which is not always possible. It is also preferable for the cumulative distribution to have an analytic inverse, since numerical inversion is typically slow.

2.2 The rejection method

The idea behind the rejection method is to propose a trial value for the random variable that we want to sample. The value is then subjected to a test, and it may be either accepted or rejected. If rejection takes place, another trial value is chosen and tested, and this procedure goes on until an acceptable value is chosen.

Specifically, let $p(x)$ be the desired density function. Suppose that we can find another density $q(x)$ that we know how to sample, and such that

$$p(x) \leq M q(x)$$

for some constant M. Then the rejection method consists of the following algorithm:

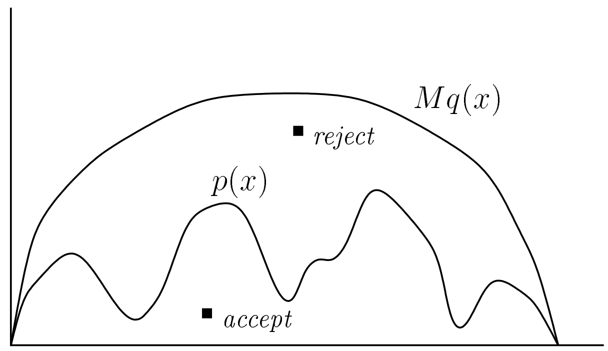

Figure 3: The rejection method.

Algorithm Reject;
    for $i = 1$ **to** ∞ **do begin**
        draw $X_i \sim q$;
        draw $U_i \sim U[0,1]$;
        if $U_i \leq \dfrac{p(X_i)}{M\,q(X_i)}$ **then return** X_i;
    end;
end.

This is illustrated graphically in Figure 3. A pair of random variables (X_i, U_i) is sampled, and if the point $(X_i, U_i \cdot Mq(X_i))$ lies under the curve $p(x)$, then it is accepted (by returning X_i). Otherwise, we choose a new pair (X_i, U_i) and try again.

Here is why the method works: on each iteration, the density of generating a sample at X_i is $q(X_i)$. This sample is returned if

$$U_i \leq \frac{p(X_i)}{Mq(X_i)},$$

which happens with probability $p(X_i)/(Mq(X_i))$. So, the probability density of returning a sample at x on a particular iteration is

$$q(x)\,\frac{p(x)}{Mq(x)} = \frac{p(x)}{M},$$

which integrates to $1/M$. Thus, when a sample is returned (with probability $1/M$), then X_i is distributed according to $p(x)$.

On the other hand, there is a $1 - 1/M$ probability that the sample will be rejected. Thus the average number of samples required until acceptance is the sum of a geometric series,

$$\sum_{i=0}^{\infty} \left(1 - \frac{1}{M}\right)^i = M,$$

where the term $(1 - 1/M)^i$ is the probability that at least $i + 1$ samples will be required.

The main advantage of rejection sampling is that it can be used with any density function, even those that cannot be integrated analytically. However, if it is applied naively then it can be very inefficient, since we may iterate a lot of times until we find an acceptable value. Clearly, the efficiency depends on the size of the constant M, i.e. how tightly $Mq(x)$ bounds $p(x)$. The goal is to find a density $q(x)$ that we know how to sample from, and that is also a reasonably good match for $p(x)$. Another disadvantage of the rejection method is that it is difficult to apply with stratification.

An important special case. Let $p(x)$ be defined on the domain $[0, 1]$, and suppose that $p(x) \leq M$ for all x. In this case we can let $q(x) = 1$ (that is, X_i is drawn from the uniform distribution $U[0, 1]$). The acceptance probability for X_i then becomes

$$\frac{p(X_i)}{Mq(X_i)} = \frac{p(X_i)}{M}.$$

Graphically, the acceptance probability is simply the ratio of the heights of the curves $p(x)$ and $Mq(x) = M$ at the given point X_i (see Figure 4). It is clear that the accepted samples have a density proportional to $p(x)$.

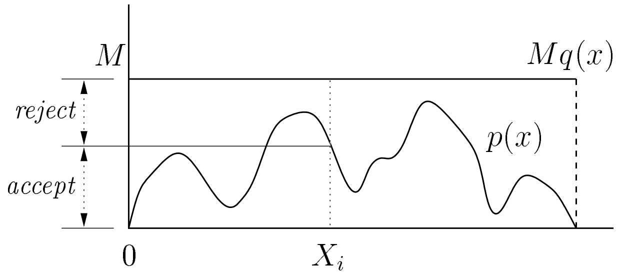

Figure 4: Each sample is accepted with probability $p(X_i)/M$, and rejected otherwise.

A geometric example. The rejection method can also be applied in geometric settings that do not correspond exactly to the general framework above. For example, suppose that we want to sample uniformly from a unit disc centered at the origin (see Figure 5).

Letting U_1 and U_2 be independent samples from $U[0, 1]$, the rejection method consists of the following:

Algorithm Reject-Circle;
    do forever
        $X = 2U_1 - 1$;
        $Y = 2U_2 - 1$;
        if $X^2 + Y^2 \leq 1$ **then return** (X, Y);
    end;
end.

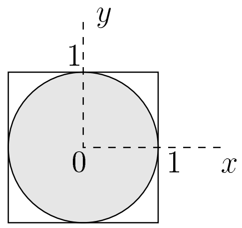

Figure 5: Sampling from a unit disc.

This can be considered a method for sampling the two-dimensional density function

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

by drawing samples from the uniform density $q(x)$ on $[-1, 1] \times [-1, 1]$. We do not need an extra random variable U_3 to decide whether to accept the samples, since the acceptance probability is always 0 or 1.

2.3 Tricks

There are a variety of tricks for sampling from specific distributions, usually by mapping several uniform random variables into a single output variable Y.

For example, let $U_1, \ldots, U_k$ be independent samples from $U[0, 1]$, and let

$$Y = \max\{U_1, U_2, \ldots, U_k\}.$$

Then the cumulative distribution of Y is:

$$P_Y(x) = Pr\{Y \leq x\} = \prod_{i=1}^{k} Pr\{U_i \leq x\} = x^k,$$

and thus the density function of Y is

$$p_Y(x) = P'_Y(x) = kx^{k-1}.$$

Therefore if we want to sample from a density proportional to x^k, we can do this by choosing $k+1$ random variables uniformly distributed on $[0, 1]$ and taking their maximum.

As another example, consider the random variable $Y = U_1 + U_2$. We can find the cumulative distribution of Y graphically, by thinking of U_1 and U_2 as the x and y coordinates of a unit square (see Figure 6). Specifically, we have

$$P_Y(z) = Pr\{Y \leq z\} = Pr\{U_1 + U_2 \leq z\},$$

which is simply the area below the line $U_1 + U_2 = z$. Using this fact we can easily derive the density function of Y, which is

$$p(x) = \begin{cases} x, & x \in [0, 1) \\ 2 - x, & x \in [1, 2] \end{cases}.$$

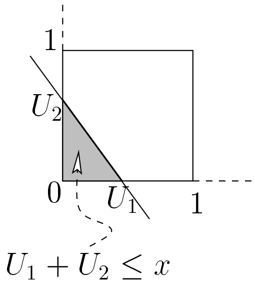

Figure 6: The density of a sum of two uniform random variables.

3 Sampling uniformly over a hemisphere

In graphics we often want to sample functions $f(\omega)$ which depend on a direction ω. We will represent ω as a unit vector (x, y, z), so that the set of all possible directions is the unit sphere

$$\mathcal{S}^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}.$$

Normally we are only interested in directions on one side of a given surface, which form a unit hemisphere centered around the surface normal $\mathbf{N}$. Such a hemisphere will be denoted Ω.

As a simple example, suppose that we want to sample a direction ω uniformly with respect to solid angle. Recall that the solid angle of a set of directions is simply the area of the corresponding set of points on the unit sphere. For example, the solid angle of a hemisphere is 2π.

For convenience, suppose that the hemisphere Ω is centered around the normal $\mathbf{N} = (0, 0, 1)$. A direction $\omega = (x, y, z)$ can then be parameterized by its spherical coordinates (θ, ϕ), where

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

With these coordinates, an element of area (solid angle) on the hemisphere is

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

A *cap* of the hemisphere is defined as the set of directions where $0 \leq \theta \leq \theta_0$, for some given value of θ_0. The area $A(\theta_0)$ of this cap is:

$$\begin{aligned} A(\theta_0) &= \int_0^{2\pi} \int_0^{\theta_0} \sin(\theta) \, d\theta \, d\phi \\ &= \left(\int_0^{2\pi} d\phi \right) \left(\int_0^{\theta_0} \sin(\theta) \, d\theta \right) \\ &= 2\pi \, (-\cos\theta) \Big|_0^{\theta_0} = 2\pi(1 - \cos\theta_0). \end{aligned}$$

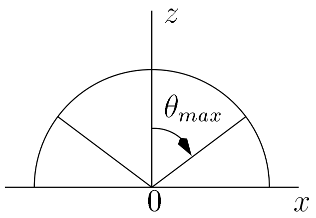

Figure 7: Sampling a cap of the hemisphere.

Recall that our goal is to sample a direction ω uniformly with respect to solid angle (i.e. area). Thus, we want the probability of sampling an angle θ to obey

$$Pr\{\theta \leq \theta_0\} \propto A(\theta_0).$$

Thus we can obtain the desired cumulative distribution for θ by normalizing $A(\theta)$:

$$P(\theta) = \frac{A(\theta)}{A(\pi/2)} = \frac{2\pi(1 - \cos\theta)}{2\pi} = 1 - z,$$

recalling that $z = \cos(\theta)$. This result implies that area on the hemisphere is uniformly distributed with respect to z, i.e. if we look at any two horizontal slices with vertical thickness dz, then both slices have the same surface area.

Using the inversion method, we can now sample uniformly on the hemisphere by setting $P(\theta) = U_1$, or equivalently $z = U_1$ (since $1 - U_1$ is also uniform). If we then choose ϕ uniformly, we get

$$\begin{aligned} Z &= U_1 \\ X &= R\cos(2\pi U_2), \\ Y &= R\sin(2\pi U_2) \end{aligned}$$

where U_1 and U_2 are uniform, and $R = \sqrt{1 - U_1^2}$. This will generate a uniform distribution of points on the hemisphere with respect to solid angle.

We can also use this technique to sample uniformly on a cap of the hemisphere (i.e. the set of directions lying with some infinite cone). Let θ_{max} be the cap angle (see Figure 7), so that $\theta \in [0, \theta_{max}]$ and $\phi \in [0, 2\pi]$. To sample this cap uniformly, we only need to change the way that z is evaluated:

$$\begin{aligned} Z &= 1 - U_1(1 - \cos\theta_{max}) \\ X &= R\cos(2\pi U_2), \\ Y &= R\sin(2\pi U_2) \end{aligned}$$

where as before $R = \sqrt{1 - U_1^2}$. This technique can also be used to generate a random point on the entire sphere, by letting $\theta_{max} = \pi$.