Lecture 4

Error in Direct MC and Importance Sampling

Contents

1 Goals and Outline

We shall take a closer look at **error analysis** in computing integrals and expected values using direct Monte Carlo

$$\text{error} \sim \frac{\text{constant}}{\sqrt{n}}, \tag{1}$$

and its link to the **central limit theorem** (CLT).

Recall that integration using direct MC involves computing averages of independent samples. So we shall review the topic of **uncertainty propagation** and use it to determine the variance of the average from the variance of individual samples. This will allow us to develop an explicit expression for the "constant" in eqn 1.

Subsequently, we shall introduce the idea of **importance sampling**, which seeks to reduce the "constant" term by decreasing the rejection rate.

2 Propagation of Error

Consider independent variables x and y with errors σ_x and σ_y.[1]

- if $z = x + y$, or $z = x - y$, then the error in z is given by:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2. \tag{2}$$

- if $z = xy$, or $z = x/y$ then

$$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2. \tag{3}$$

This has the appearance of Pythagorean triples.[2]

If $z = f(x)$ is some arbitrary function of x then

$$\sigma_z = |f'(x)|\, \sigma_x. \tag{4}$$

Examples

Q1: Volume $V = s^3$, with $s = 2.00 \pm 0.02$ cm. What is the uncertainty in volume?
A1: $\sigma_V = |3s^2|\, \sigma_s$. Therefore, $V = 8 \pm 0.2$ cm^3.

Q2: $x = 100 \pm 6$, and $z = \sqrt{x}$. What is σ_z?
A2: $\sigma_z = \sigma_x/(2\sqrt{x})$ = 0.3

Q3: If $x = 1 + 0.1$, $y = 3 + 0.2$. What is the error in $x - y$ and x/y?

Q4: Write $x = z \times z$ to find the error in x using the "product rule" (eqn 3).[3]

[1] For example, σ may be the standard deviation of a random variable. Here, the label "error" $\equiv$ "uncertainty"

[2] add/subtract operates on absolute error, while multiply/divide operates on relative error

[3] Why is this conceptually incorrect? Are the two zs being multiplied together independent?

2.1 General Formula

If x_1, x_2, ..., x_n are independent random variables, and

$$q = q(x_1, x_2, ..., x_n),$$

then,

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x_1}\sigma_{x_1}\right)^2 + \left(\frac{\partial q}{\partial x_2}\sigma_{x_2}\right)^2 + ... + \left(\frac{\partial q}{\partial x_n}\sigma_{x_n}\right)^2}. \tag{5}$$

The derivation is presented in section A.1. It uses a linear approximation, and is guaranteed to work well when the errors are "small".

Exercise: Verify that the formulae 2-4 above are special cases of eqn 5.

2.2 Average of Independent Observations

Consider the average of a number of **independent and identically distributed** (iid) samples/observations,

$$z = \frac{x_1 + x_2 + ... + x_n}{n}.$$

That is, $\sigma_{x_i} = \sigma_x$ for all x_i. Using the general formula (eqn 5), we get:

$$\sigma_z = \sigma_x \sqrt{\frac{1}{n^2} + \frac{1}{n^2} + ... \frac{1}{n^2}} = \frac{\sigma_x}{\sqrt{n}}. \tag{6}$$

The error of the average is smaller than the error in an individual observation, due to cancellation of errors. We understand this intuitively, which is why we take multiple measurements when we do experiments.

3 Central Limit Theorem

While eqn. 6 is sufficient for understanding the error in direct MC,[4] we are simply too close to one of the most beautiful ideas in statistical mathematics to look away. You can thank me later.

Consider n iid random variables drawn from the distribution $X_i \sim \pi(x)$, $1 \leq i \leq n$. Let $\langle x \rangle$ and σ_x^2 be the mean and variance of $\pi(x)$.[5]

Further, let the average or sample mean

$$Z = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We know from the general formula for error propagation of iid random variables that the variance of the sample mean, $\sigma_z^2 = \sigma_x^2/n$.[6]

- $\sigma_z^2 \to 0$, as $n \to \infty$, and hence,

[4] The $1/\sqrt{n}$ term should be a giveaway.

[5] Here, we use angular brackets $\langle \cdot \rangle$ to denote the true mean. Later, we shall use an "overbar" ($\bar{x}$) to denote a sample mean; for now we represent it using z.

[6] Show that the expected value of z is $E[z] = E[x] = \langle x \rangle$. The expected value is often written in angular brackets for brevity.

- the sample mean approaches the true mean $\langle x \rangle = E[x]$.

We knew all this. The CLT tells us something more about Z, in addition to the first and second moments.

It tells us something about the entire **probability distribution** of Z, provided n is sufficiently large, for its effects to kick in. Informally:

As n increases, the distribution of the sample mean, say $g_Z(z)$ approaches a normal distribution

In the limit, $n \to \infty$, the distribution $g_Z(z) \to \mathcal{N}\left(\langle x \rangle, \sigma_z^2\right)$.

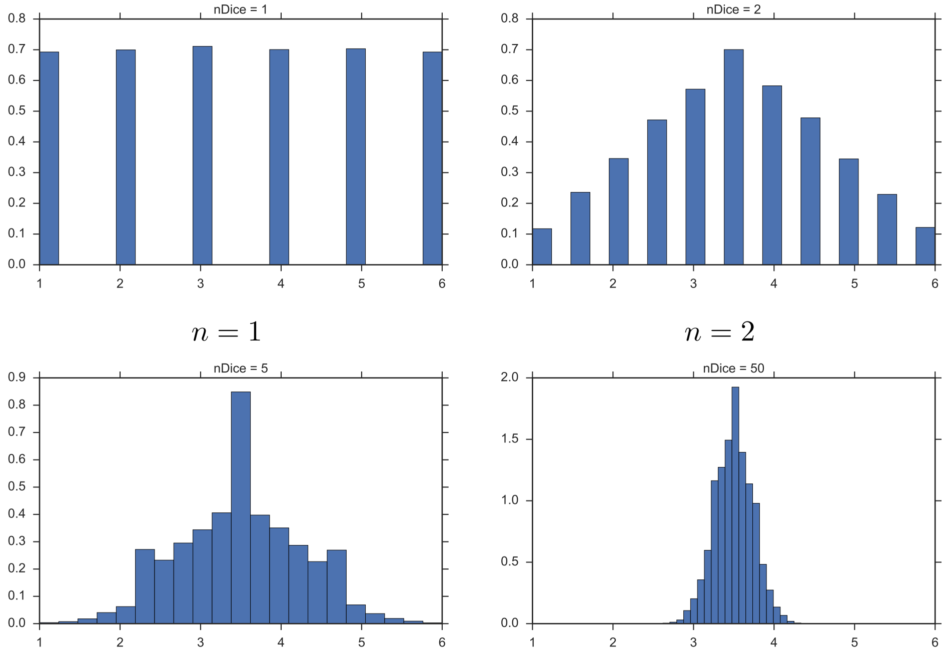
Note that the distribution $\pi(x)$ can be anything: discrete, continuous, whatever!

It is possible that your mind has not been completely blown away, and you are wondering: "What is this "distribution of the sample mean" business?"

Recognize that the average of n random variables (Z) is itself a random variable. That is, if we repeat the "experiment" of calculating the average two different times, we don't expect the answers to be identical. Let's illustrate this with an example.

Illustration: Suppose we roll n dice ($x_i \sim U[1,6]$) and define Z as the average. What is the distribution of Z?

We could try to do an analytical calculation to figure out $g(z)$, but let's just do a MC simulation, since it is much easier (see code `avgDiceRoll` in sec A.3).



$n = 1$

$n = 2$

$n = 5$

$n = 50$

As n increases,[7] observe how an initially flat distribution, increasingly starts to peak at its mean. Also note that the distribution gets increasingly narrow.

For "large" n the distribution looks unmistakably Gaussian. The distribution starts looking like a continuous distribution, despite the underlying distribution being discrete in this example.

[7] `nDice` in the code

Exercise: Set Z as the average of n exponentially distributed variables. Plot the distribution of z for different choices of n, to affirm the CLT.

Next, we link CLT and "Integration by Darts".

4 Error in Direct Monte Carlo

We are ready to circle back to MC integration, where we compute integrals by estimating the mean or average value of the integrand over the domain.

We draw a large number of samples. Hence, we can trust the law of large numbers to apply. Consider a 1D integral,

$$\begin{aligned} I &= \int_a^b f(x)dx \\ &\approx \sum_{i=1}^{n} f(x_i)\Delta x \\ &\approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) \\ &\approx (b-a)\bar{f}. \end{aligned}$$

The question we ask ourselves is: "if we use n darts to estimate $\bar{f}$, treating f_i as iid random variables with variance σ_f^2, what is the variance σ_I^2?"

Using eqn 6,[8]

$$\begin{aligned} \sigma_I^2 &= (b-a)^2\frac{\sigma_f^2}{n} \\ &= (b-a)^2\frac{\langle f^2\rangle - \langle f\rangle^2}{n}. \end{aligned}$$

In general, for integrals of any dimension,

$$\sigma_I^2 = \frac{V^2}{n}\sigma_f^2, \tag{7}$$

where $V = \int d\mathbf{x}$ is the volume of the domain. Notice that the variance in the integral is proportional to:

(i) the variance of the integrand $f(x)$

(ii) the square of the domain size

(iii) $1/n$

Thus, the "constant" term we talked about at the start (eqn 1) is constant $= V\sigma_f$. That is, the error in MC integration $\sigma_I = V\sigma_f/\sqrt{n}$.

Does this have any practical relevance? It sure does!

[8] Note that the var$(ax) = a^2$var(x)

In MC, we evaluate the average of independently sampled $f_i = f(x_i)$ to get $\bar{f}$.[9] We can estimate the variance of f simply by keeping track of $\bar{f^2}$, in addition to $\bar{f}$.

This will let us put **error-bounds** on any **single direct MC estimate**.

4.1 Example

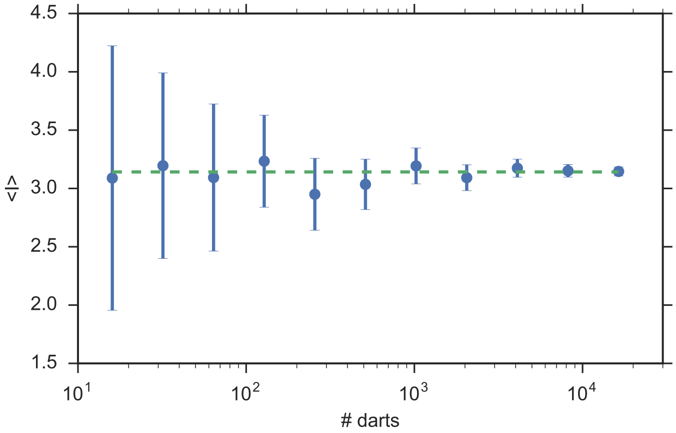
Consider the previous problem,

$$I = \int_0^\pi x \sin x \, dx = \pi.$$

We can write a program (see `simpleIntegral` in sec. A.3) to evaluate the variance in f_i.

```
I, errI = simpleIntegral(1000)
3.16046202212 0.0198906574291
```

We can vary the number of darts.



Each estimate and error bar in the figure above is obtained using a single MC run. The dashed green line is the true value of the integral, π. As you see, the error estimates are quite sharp!

Normally, we would mark this problem as solved and move on. However, given the analytical tractability of this example, it is worthwhile to scrape one layer deeper.

4.1.1 Post-Mortem

Given an integration problem,

$$I = \int_a^b f(x) \, dx$$

the key results using direct MC,

$$I = (b-a)\langle f \rangle$$
$$\sigma_I = \frac{(b-a)\sigma_f}{\sqrt{n}}.$$

In a MC simulation, we estimate $\langle f \rangle$ and σ_f.

[9] We picked x_i from a (uniform) distribution $\pi(x)$. The corresponding distribution of f depends on $f(x)$, and can be obtained explicitly via $\pi(f) = \pi(x)|dx/df|$.

However, for this particular example $f(x) = x \sin x$, we can analytically evaluate I and σ_I.

$$\langle f \rangle = \frac{1}{\pi} \int_0^{\pi} x \sin x \, dx = 1$$
$$\langle f^2 \rangle = \frac{1}{\pi} \int_0^{\pi} x^2 \sin^2 x \, dx = \frac{\pi^2}{6} - \frac{1}{4}$$

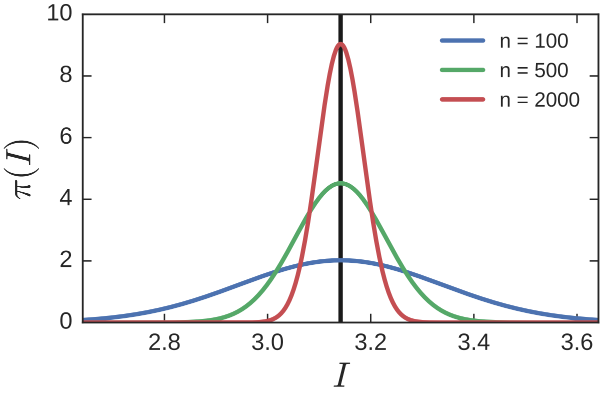
Theoretically,

$$\sigma_f^2 = \langle f^2 \rangle - \langle f \rangle^2 = \frac{\pi^2}{6} - \frac{5}{4} \approx 0.3949$$

The CLT tells us that any particular MC simulation or experiment involves sampling I[10] from its own distribution $\pi(I)$. For sufficiently large n, this distribution is Gaussian with mean $\langle f \rangle V = \pi$ and variance $\sigma_I^2 = \sigma_f^2 V^2/n$,

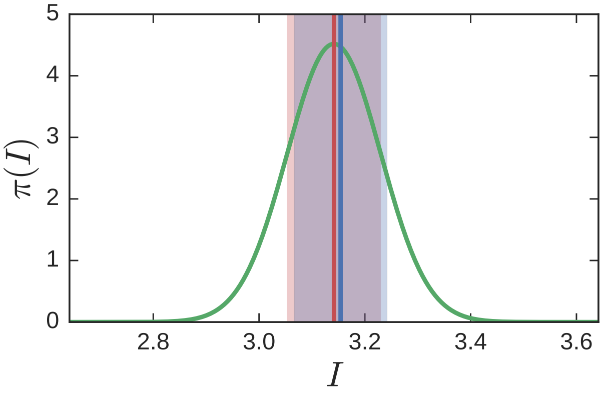
$$\pi(I) = \mathcal{N}(\pi, \pi^2 \sigma_f^2/n). \tag{8}$$

The figure below plots $\pi(I)$ for different n.



When we run an MC simulation with a particular n, the resulting estimate of the integral is sampled from $\pi(I, n)$. As n increases, $\pi(I, n)$ becomes narrower, and our estimates are closer to the expected value $I = (b - a)\langle f \rangle$ more often.

Our on-the-fly estimate of σ_I depends on our on-the-fly estimate of σ_f. For $n = 500$, the true $\sigma_f \approx \sqrt{0.3949} = 0.6284$.



[10] which may be thought of as a random variable

Using a single MC run, I estimated $\sigma_f = 0.6168$.[11] The red and blue areas in the figure above show range of the true and estimated σ_I.

What can you do if you have multiple replicas? See appendix A.2.

5 Importance Sampling

On its face, direct MC seems very fatalistic: the only knob one can turn to reduce error (σ_I) is to throw more darts (n). The numerator is determined by the integrand and the domain of integration, both of which are beyond our control.

Are they really?

Importance sampling is a technique which tries to work on this numerator. It can be thought of as a variance reduction technique for a certain subset of commonly encountered problems. It is best to start with a particular example.

Consider the expected value of a function $f(x)$,[12]

$$E_\pi[f] = \int_a^b f(x)\pi(x)\,dx, \tag{9}$$

where $\pi(x)$ is normalized probability distribution with

$$\int_a^b \pi(x)\,dx = 1.$$

Setting $g(x) = f(x)\pi(x)$

$$E_\pi[f] = \int_a^b g(x)\,dx = (b-a)\bar{g}. \tag{10}$$

There are two ways to calculate this expected value.

5.1 Two Methods to Compute Expected Values

The two methods are:

(i) **Method 1**: Sampling from uniform distribution

(ii) **Method 2**: Sampling from target distribution

5.1.1 Method 1: Sampling from Uniform distribution

This is our standard way of computing integrals with MC.

- draw n samples $X_1, X_2, ..., X_n$ from $u(x) = U[a,b]$
- estimate $\bar{g}$ in eqn 10 by averaging,

$$\bar{g} \approx \frac{1}{n}\sum_{j=1}^{n} g(X_j) = \frac{1}{n}\sum_{j=1}^{n} \pi(X_j)f(X_j) \tag{11}$$

[11] This estimate will change from simulation to simulation.

[12] integrals of this kind occur very frequently in practical applications

- hence, estimate $E_\pi[f] = \bar{g}(b-a)$:

$$E_\pi[f] = \frac{b-a}{n} \sum_{j=1}^{n} \pi(X_j) f(X_j) \tag{12}$$

We sample X_j from a uniform distribution on the domain.

5.1.2 Method 2: Sampling from target distribution

If it is convenient to sample from the target distribution $\pi(x)$ directly, we can propose another method:

- draw n samples $X_1, X_2, ..., X_n$ from $\pi(x)$
- estimate $E_\pi[f]$ directly by a simple average:

$$E_\pi[f] \approx \frac{1}{n} \sum_{j=1}^{n} f(X_j) \tag{13}$$

Note the absence of $(b-a)$ and $\pi(x_j)$ in eqn 13 compared to 12.

Exercise: Using eqns 12 and 13 develop methods to estimate the error/variance.

Let us consider a simple example to demonstrate these two methods.

Example: Consider,

$$E[f] = \int_{-10}^{10} f(x)\,\pi(x)\,dx = \int_{-10}^{10} g(x)\,dx \tag{14}$$

with $f(x) = x^2$, and $\pi(x) = \mathcal{N}(0,1)$, the unit normal distribution,

$$\pi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

Note that $\pi(x)$ is *almost* normalized,

$$\int_{-10}^{10} \pi(x)\,dx \approx 1,$$

The true solution of the problem is $E[f] = 1.0$.[13]

Solution:

We can define the appropriate python functions, and test them by verifying that the built-in integrator yields an answer close to the true solution.

```
f = lambda x: x**2        # f(x)
p = lambda x: 1./np.sqrt(2.*np.pi) * np.exp(-x**2/2.) # target distribution
g = lambda x: f(x) * p(x)  # combination g(x)

# numerical integral using built-in integrator
scipy.integrate.quad(g, -10, 10)
(1.0, 7.348033275594856e-10)
```

[13] This is not “news”; the variance of $\mathcal{N}(0,1)$ is $\sigma^2 = E[x^2] - E[x]^2 = 1 - 0 = 1$.

We can now implement the two methods.

```
# Method 1
x     = np.random.uniform(-10, 10, size=10000)        # uniform distribution
gbar1 = np.mean(g(x))        # average
fbar1 = gbar1 * (10. - (-10.))  # E[f] for method 1
1.00636345198

# Method 2
x = np.random.normal(0, 1, size=10000) # target distribution
fbar = np.mean(f(x)) # simple average E[f] for method 2
1.00961719771
```

Both the estimates are decent. However note that unlike the first method, the second method:

(i) does not involve multiplication with $(b - a)$

(ii) works well even when the distribution is not properly normalized.

See sec. A.4 for a modification of this problem, where the domain is $[-1, 1]$, instead of $[-10, 10]$.[14]

Side Note

We can rephrase our old friend in light of the two methods,

$$I = \int_a^b f(x)\,dx = (b-a)\langle f \rangle.$$

Here,

$$\begin{aligned} \langle f \rangle &= \frac{1}{b-a}\int_a^b f(x)\,dx \\ &= \int_a^b f(x)\,u(x)\,dx \\ &= E_u[f], \end{aligned} \tag{15}$$

since, $u(x) = U[a, b] = 1/(b - a)$. To compute $\bar{f}$,[a] we "effectively" used method 2. The algorithm we used was:

- draw samples $X_1, X_2, ..., X_n$ from $\pi(x) = u(x) = U[a, b]$
- estimate $E_u[f]$ by a simple average:

$$E_u[f] \approx \frac{1}{n}\sum_{j=1}^{n} f(X_j)$$

To find I we multiplied $E_u[f]$ by the domain size $(b - a)$.

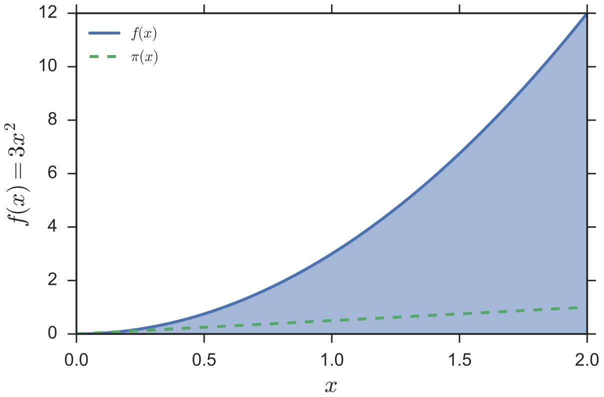
[a] which is an estimate of $\langle f \rangle$

[14] It may be best to defer that section until after the full discussion on importance sampling.

5.2 Illustrating Importance Sampling

Now suppose, instead of choosing points uniformly, we "focused" on important regions of the domain. This idea is best illustrated with an example. Consider the simple integral:

$$I = \int_0^2 3x^2 dx.$$



I is the shaded area under the curve. Note that points close to $b = 2$ "contribute" more to the integral than points close to $a = 0$. Therefore instead of choosing points uniformly between a and b, suppose we draw points from the distribution $\pi(x) = x/2$, which is shown by the dashed line.[15]

Of course, we will need to unbias our estimate.[16]

Informally, instead of taking a simple average of uniformly weighted points, we will take a weighted average. This is the key intuition behind importance sampling.

Let's see how the math behind the idea works out. Since, $I = E_u[f](b - a)$, where

$$\begin{aligned} E_u[f] &= \int_a^b f(x)\, u(x)\, dx \\ &= \int_a^b f(x)u(x)\frac{\pi(x)}{\pi(x)}dx \\ &= \frac{1}{b-a}\int_a^b \frac{f(x)}{\pi(x)}\pi(x)dx. \end{aligned}$$

Therefore,

$$I = E_u[f](b-a) = \int_a^b \frac{f(x)}{\pi(x)}\pi(x)dx == E_\pi\left[\frac{f}{\pi}\right], \tag{16}$$

where $\pi(x)$ is normalized over the domain. The form of eqn 16 indicates that it is an ideal candidate to use method 2 on.

[15] We "select" this $\pi(x)$ because it loosely follows the shape of the integrand. This choice is clearly not unique.

[16] You can consider an extreme case for the integrand,

$$f(x) = \begin{cases} 0 & 0 \leq x < 1 \\ 1 & 1 \leq x < 2, \end{cases}$$

where sampling points $x < 1$ is completely wasted effort.

Note that we could directly have written eqn 16 from $I = \int f(x)dx$, by multiplying and dividing the integrand by $\pi(x)$!

Exercise: We showed above that $E_u[g](b-a) = E_\pi[g/\pi]$ for any $g(x)$. Can you justify the equivalence between method 1 and method 2 by setting, $g(x) = f(x)\pi(x)$?

5.2.1 Algorithm

This observation yields the algorithm for importance sampling MC integration:

1. Draw $X_1, X_2, ..., X_n$ from $\pi(x)$
2. Estimate $I = E_\pi(f/\pi)$ via
$$I \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{\pi(X_i)} \tag{17}$$

Note that this formula works gracefully for $\pi(x) = u(x)$, where we essentially revert back to method 1.[17]

5.2.2 Error Estimate for Importance Sampling

Can we get an on-the-fly error estimate for importance sampling, like we did for method 1?

Sure. From eqn. 16 note that,

$$\begin{aligned} I &= \int_a^b \frac{f(x)}{\pi(x)}\pi(x)dx, \\ &= E_\pi[f/\pi], \end{aligned}$$

the integral is the expected value of f/π.[18] One can estimate the error in I by keeping track of the variance $\sigma^2_{f/\pi}$ of f/π.[19]

For uniform sampling, we saw earlier that

$$\sigma_I^2 = \frac{V^2}{n}\sigma_f^2.$$

That is, the error $\sigma_I \approx 0$, as $f(x) \approx$ constant.

For importance sampling,

$$\sigma_I^2 = \frac{\sigma^2_{f/\pi}}{n},$$

where

$$\sigma^2_{f/\pi} = \left\langle \left(\frac{f}{\pi}\right)^2 \right\rangle - \left\langle \left(\frac{f}{\pi}\right) \right\rangle^2$$

[17] Method 2 does not require $\pi(x)$ to be normalized. However, eqn 17 **does require** $\pi(x)$ to be normalized, otherwise the estimate of I will be off by a multiplicative factor of the normalization constant. However, there is a *self-normalizing importance sampling* algorithm listed at the end of A.4. which can deal with cases where the normalization constant is not available, or easy to estimate.

[18] The subscript reminds us that the points are sampled from the distribution $\pi(x)$ and not $u(x)$.

[19] This follows directly from the CLT, where the underlying distribution doesn't matter; the average (or expected value) is normally distributed!

That is, the error $\sigma_I \approx 0$, when $\pi(x) \approx f(x)$. This gives us a guideline on how to pick $\pi(x)$:[20] It suggests that we should pick a distribution $\pi(x)$ which resembles the shape of $f(x)$, and is easy to sample from.

Example: Let's get back to the original integral,

$$I = \int_0^2 3x^2 dx.$$

Find I using importance sampling from the distribution $\pi(x) = x/2$. Use uniform sampling to solve the same problem, and compare error estimates from the two methods.

Solution: First, we can use the transformation rule to figure out how to sample from $\pi(x) = x/2$. This involves finding the CDF of $\pi(x)$,

$$F(x) = \int_0^x \pi(x')\, dx' = x^2/4. \tag{18}$$

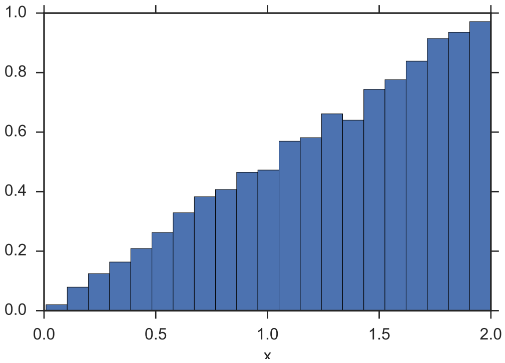
We set $u = F(x)$, and try to solve for x in terms of u, where $u \sim U[0,1]$.
Here,

$$u = F(x) = x^2/4 \implies x = +\sqrt{4u}.$$

We can write a python program to implement this.

```python
def drawLinearDist(npts):
    u = np.random.rand(npts)
    return np.sqrt(4.0*u)
```

It is always a good idea to test the sampler by plotting the histogram.



Now let's compare uniform and importance sampling for this problem. We begin by writing python code for each method.

[20] remember, this is completely the user's choice

Uniform Sampling

```
def uniformSampling(npts):
    """x uniform dist"""
    b   = 2.
    a   = 0.
    x   = np.random.uniform(a, b, npts)
    f   = 3.*x**2

    intg = (b-a) * np.mean(f)
    stdI = (b-a) * np.std(f)/np.sqrt(npts)

    return intg, stdI
```

Importance Sampling

```
def importanceSampling(npts):
    """x linear dist"""

    x   = drawLinearDist(npts)
    f   = 3.*x**2
    p   = x/2.0    #pi(x)

    intg = np.mean(f/p)
    stdI = np.std((f/p))/np.sqrt(npts)

    return intg, stdI
```

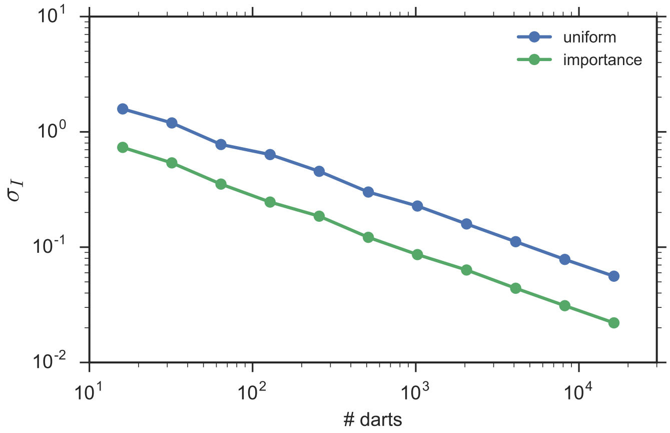
We can compare the two methods by using, say, $n = 50$ darts. The true solution is $I = 8$.

```
print(uniformSampling(50))
(8.8636221283119916, 1.0940151684845272)

print(importanceSampling(50))
(7.817575629020471, 0.40950491361094382)
```

The first argument is the estimated value of the integral,[21] and the second argument is the estimated error. Notice that both error estimates are reasonable[22]; however, the error is smaller for importance sampling.

This is true for different n. In both cases, $\sigma_I \sim 1/\sqrt{n}$.



However, the error is offset by a constant term ("variance reduction"). Importance sampling tries to reduce the error by modifying the numerator of σ_I.

5.3 Intuition

Practically, for "flat" integrands (the area of a circle example), the principal goal is to cover all parts of the domain; uniform sampling works perfectly fine.

Importance sampling shines when the integrand has only a few sharp peaks, which contribute most of the "area under the curve".

[21] which is close to the true value, as expected

[22] in that the true value is within ±$2.5\sigma_I$ of the estimated integral

Analogy: If you want to find the average height of a palm tree between the US and Japan, you don't want to randomly throw darts, since most of them will land in the Pacific Ocean. You want to focus on islands and other land masses.

There are some truly needle-in-the-haystack type problems, where the important region is well-hidden. These are truly pathological problems. Fortunately, true pathology is not an everyday occurence.

Exercise: Use importance sampling with a poor choice[23] of $\pi(x) = 1 - x/2$ to evaluate the integral $I = \int_0^2 3x^2 \, dx$. How does σ_I vary with the number of "darts"?

5.4 Summary

- **Error Propagation**: If $x_1, x_2, \ldots, x_n$ are independent random variables, and $q = q(x_1, x_2, ..., x_n)$, then,

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x_1}\sigma_{x_1}\right)^2 + \left(\frac{\partial q}{\partial x_2}\sigma_{x_2}\right)^2 + ... + \left(\frac{\partial q}{\partial x_n}\sigma_{x_n}\right)^2.$$

- This implies that the error in the average of n iids ($\sigma_{x_i}^2 = \sigma_x^2$), $z = (x_1 + x_2 + ... + x_n)/n$ is,

$$\sigma_z^2 = \frac{\sigma_x^2}{n}.$$

- The **CLT** makes an even stronger claim. It says that the *distribution* of z approaches a normal distribution,

$$g(z) \to \mathcal{N}\left(\langle x \rangle, \sigma_z^2\right).$$

  for large n (the law of large numbers).

- These results help explain why the error in the MC integration estimate

$$\epsilon = \left| \int f(x) \, dx - V\bar{f} \right|,$$

  decreases as $V\sigma_f/\sqrt{n}$. It also gives us a method for estimating the error in the integral on the fly.

- A commonly encountered integration problem in MC is computation of **expected values**,

$$E_\pi[f] = \int_a^b f(x)\pi(x) \, dx.$$

  We can use **two methods** to solve the integral: sample (1) $X_i \sim U[a, b]$ and estimate $E_\pi[f] \approx (b - a)\overline{f(x)\pi(x)}$, and (2) $X_i \sim \pi(x)$, and estimate $E_\pi[f] \approx \bar{f}$.

- It seems that the only way to decrease the error in MC integration is to increase the number of samples n. However, importance sampling is a variance reduction technique that seeks to decrease the "numerator".

[23] This choice places a higher density of darts in regions that do not contribute significantly to the integral.

- **Importance Sampling**: Rewrite the integral,

$$I = \int_a^b f(x)\,dx = \int_a^b \frac{f(x)}{\pi(x)} \pi(x)\,dx$$

and use method 2 to estimate $I \approx E_\pi(f/\pi)$. The error with importance sampling is:

$$\sigma_I^2 = \frac{\sigma^2_{f/\pi}}{n}.$$

6 Problems

6.1 Thought Questions

(i) Consider independent random variables x_1, x_2, x_3, with variance $\sigma^2_{x_1}, \sigma^2_{x_2}$, and $\sigma^2_{x_3}$.

The geometric average of these variables is given by,

$$z = (x_1 x_2 x_3)^{1/3}.$$

Use the error-propagation formula, to find the variance of z, σ_z^2.

(ii) Suppose you conduct two independent experiments to measure a quantity z, and obtain the estimates, $z_1 = 6 \pm 2$ and $z_2 = 4 \pm 1$. The two measurements do not arise from one experiment repeated twice, but rather from two completely different experimental setups.

Your goal is to combine these two estimates to come with an improved consensus estimate.

(a) Suppose your first attempt is the simple mean $\bar{z} = (z_1 + z_2)/2$. Find $\bar{z}$ and the corresponding error $\sigma_{\bar{z}}$ using the standard propagation of error formula.

(b) You think about the problem, and notice that z_2 has a smaller error bar than z_1. You wonder if you should consider a weighted mean,

$$\bar{z}_w = w z_1 + (1 - w) z_2,$$

where $0 \leq w \leq 1$ is a weight. Find the corresponding error $\sigma_{\bar{z}_w}$ using the standard propagation of error formula, as a function of w.

(c) Find the value of w that minimizes the error $\sigma_{\bar{z}_w}$. For this example, show that the weights $w_1 = w$ and $w_2 = (1 - w)$ associated with z_1 and z_2 are inversely proportional to $\sigma^2_{z_1}$ and $\sigma^2_{z_2}$, respectively.

(d) What is the corresponding $\bar{z}_w$ and $\sigma_{\bar{z}_w}$?

(iii) Suppose X_i are iid random variables drawn from the power law distribution $\pi(x) = 1.5x^{-2.5}$, which has infinite variance. Let Z be the average of n such random variables. Plot the histogram of Z for different values of n. Comment on the applicability of the CLT.

(iv) How would you provide an error estimate for $E_\pi[f]$ evaluated using methods 1 and 2? Implement this for the example in the text, illustrating the two methods.

(v) True or False: In importance sampling, $\pi(x)$ has to be nonzero over the domain of integration. Discuss.

6.2 Numerical Questions

(i) Importance Sampling

Consider the integral,

$$I = \int_0^\infty x^{1/2} e^{-x} \, dx.$$

(a) Develop an expression and program to sample from the exponential distribution, $\pi(x) = e^{-x}, x \geq 0$ (10 points).

(b) Use importance sampling with 10^3 points to sample from $\pi(x)$, and estimate the integral.

- Describe the formulae you use, and your strategy in sufficient detail. (10 points)
- Report the estimated error, based on a single simulation run. (15 points)

(ii) Transformation Method and Importance Sampling

Consider evaluating the integral,

$$I = \int_0^\pi f(x) \, dx = \int_0^\pi \frac{dx}{x^2 + \cos^2 x},$$

by drawing points from the distribution $\pi(x) \sim e^{-ax}$, with $a > 0$.

(a) Normalize $\pi(x)$ so that $\int_0^\pi \pi(x) \, dx = 1$.

(b) Use the transformation method to sample from $\pi(x)$, given uniformly distributed random numbers $u \sim U[0,1]$. Test your formula for $a = 1$ by plotting a normalized histogram and $\pi(x, a = 1)$.

(c) We can rewrite the integral (for importance sampling) as,

$$I = \int_0^\pi \frac{f(x)}{\pi(x)} \pi(x) \, dx.$$

The resulting variance of the integral $\sigma_I^2 = \sigma_{f/\pi}^2/n$, where n is the number of Monte Carlo points. With $n = 10^5$, vary a between 0.05 and 2.0 and make a plot of $\sigma_I(a)$. From the plot, estimate the value of a (to within ± 0.1) which minimizes $\sigma_I(a)$.

(iii) Importance Sampling

Let $h(x)$ be the step function,

$$h(x) = \begin{cases} 1, & \text{for } x \geq 4, \\ 0, & \text{elsewhere,} \end{cases} \tag{19}$$

and let $p(x)$ be the unit normal distribution,

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Suppose we want to use Monte Carlo to find the integral,

$$E(h) = \int_{-\infty}^\infty h(x) p(x) \, dx. \tag{20}$$

The true value of the integral is $E(h) = 3.167 \times 10^{-5}$.

(a) Use “method 2” discussed in class, with $N = 10^5$ or 10^6 samples from $p(x)$ to estimate $E(h)$, and the associated error.

(b) Let us now consider using importance sampling. Instead of sampling from $p(x)$, let us sample from a shifted exponentially distributed $q(x)$ given by,

$$q(x) = \begin{cases} e^{-(x-4)}, & \text{for } x \geq 4, \\ 0, & \text{elsewhere,} \end{cases} \tag{21}$$

Note that I can write,

$$E(h) = \int_{-\infty}^{\infty} \frac{h(x)p(x)}{q(x)} q(x)\, dx. \tag{22}$$

(c) Sample from $q(x)$ to estimate $E(h)$ from the equation above, and the associated error.

(d) Comment on your observations.

(iv) Signal to Noise Ratio

Consider two functions,

$$f(x_1, x_2) = x_1 + x_2$$
$$g(x_1, x_2) = x_1/x_2,$$

of normally distributed variables $x_i \sim \mathcal{N}(\mu_i, \sigma_i)$. In this example, we want to explore the boundary of the error propagation formula (eqn 5) when the signal to noise ratio is large.

For specificity, assume $\mu_1 = 5.0, \sigma_1 = 0.1$, and $\mu_2 = 0.0, \sigma_2 = 0.2$. Draw 10,000 normally distributed samples for x_1 and x_2 and compute the mean and standard deviation of the corresponding functions f and g.

Compare these empirical estimates with the estimates from the error propagation formula. Report your observations. An important lesson is that the error propagation formula is guaranteed to work when σ/μ is small. But sometimes it will work even when this condition is not strictly met (for f, but not for g).

(v) Integrable Singularities: Consider the following innocuous-looking integral from Krauth.

$$I(\gamma) = \int_0^1 x^\gamma dx = \frac{1}{\gamma + 1}, \text{ for } \gamma > -1.$$

- Plot the integrand $f(x) = x^\gamma$ for $\gamma = -0.8, -0.2$ and 0.5.
- Write a program to evaluate the integral using uniform sampling, given γ and the number of samples n. Use $n = 1000$ darts to evaluate I and the corresponding error-bars for $\gamma = -0.8, -0.2$ and 0.5.
- Compare these estimates with analytical results to show that for $\gamma = -0.8$, I is often inaccurate, and the error-bars often underestimate the true error.

- The error estimate $\sigma_I^2 = (\langle f^2 \rangle - \langle f \rangle^2)/n$ is erratic for a very subtle reason. Show by integrating analytically that,

  $$\langle f^2 \rangle = \frac{1}{1-0} \int_0^1 x^{2\gamma} dx = \begin{cases} \frac{1}{2\gamma+1} & \gamma > -1/2 \\ \infty & \text{otherwise} \end{cases}.$$

  Discuss your previous observations in light of this fact.

- **Importance Sampling**: Instead of picking points uniformly, suppose we draw points from a distribution $\pi(x) 0.3x^{-0.7}$ which is chosen to mimic the shape of the integrand. Write a program to sample from this distribution, and use importance sampling from this distribution to determine the integral and its corresponding error for $\gamma = -0.8$.

- Show that the error estimate is reliable because the variance of $\langle f/\pi \rangle$ is no longer unbounded. Do this calculation analytically.

A Appendices

A.1 Derivation of Error Propagation Formula

Without loss of generality, assume $q(x_1, x_2)$. Let X_1 and X_2 be particular random variables and $Q = q(X_1, X_2)$ be the resultant.

Further suppose that (μ_1, μ_2), and $(\sigma_{x_1}^2, \sigma_{x_2}^2)$ are the expected values, and variances, respectively, of X_1 and X_2.

Step 1: Taylor series expansion

$$\begin{aligned} Q &= q(X_1, X_2) \\ &= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(X_1 - \mu_1) + \frac{\partial q}{\partial x_2}(X_2 - \mu_2) + \text{h.o.t} \end{aligned} \tag{23}$$

We can ignore higher order terms if $(X_i - \mu_i)$ are small. In experiments, this often requires measurement error to be “reasonably” small.

Step 2: Take Expected Value

Note for independent variables:

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2).$$

Using this property, we can write eqn 23,

$$\begin{aligned} E(Q) &= E(q(\mu_1, \mu_2)) + E\left(\frac{\partial q}{\partial x_1}(X_1 - \mu_1)\right) + E\left(\frac{\partial q}{\partial x_2}(X_2 - \mu_2)\right) \\ &= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(E(X_1) - \mu_1) + \frac{\partial q}{\partial x_2}(E(X_2) - \mu_2)) \\ &= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(\mu_1 - \mu_1) + \frac{\partial q}{\partial x_2}(\mu_2 - \mu_2)) \\ &= q(\mu_1, \mu_2) \end{aligned}$$

Step 3: Take Variance

Note for independent variables:

$$V(aX_1 + bX_2) = a^2 V(X_1) + b^2 V(X_2),$$

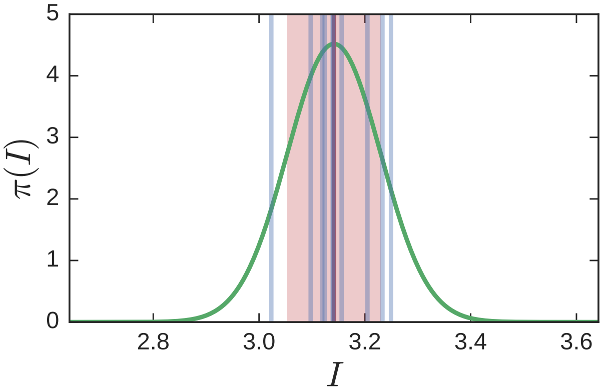
where V is used as a short-hand for variance. Using this property, we can write eqn 23,

$$\begin{aligned} V(Q) &= V(q(\mu_1, \mu_2)) + V\left(\frac{\partial q}{\partial x_1}(X_1 - \mu_1)\right) + V\left(\frac{\partial q}{\partial x_2}(X_2 - \mu_2)\right) \\ &= 0 + \left(\frac{\partial q}{\partial x_1}\right)^2 V((X_1) - \mu_1)) + \left(\frac{\partial q}{\partial x_2}\right)^2 V((X_2) - \mu_2)) \\ &= \left(\frac{\partial q}{\partial x_1}\right)^2 (V(X_1) - 0) + \left(\frac{\partial q}{\partial x_2}\right)^2 (V(X_2) - 0) \\ \sigma_q^2 &= \left(\frac{\partial q}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial q}{\partial x_2}\right)^2 \sigma_{x_2}^2 \qquad \text{QED} \end{aligned}$$

A.2 Multiple Replicas in Direct MC

Consider the example in sec. 4.1. As shown there, we can obtain an on-the-fly error estimate without the need to run multiple replicas. But if we have N independent replicas we can get an even better estimator of the value of the integral I.

Suppose you have $N = 8$ independent estimates using $n = 500$ each. These are shown in the figure below by blue vertical lines.



We can combine these estimates by averaging:

$$\bar{I} = \frac{1}{N}\sum_{i=1}^{N} I_i,$$

where I_i are independent replicates. From propagation of error we know that,

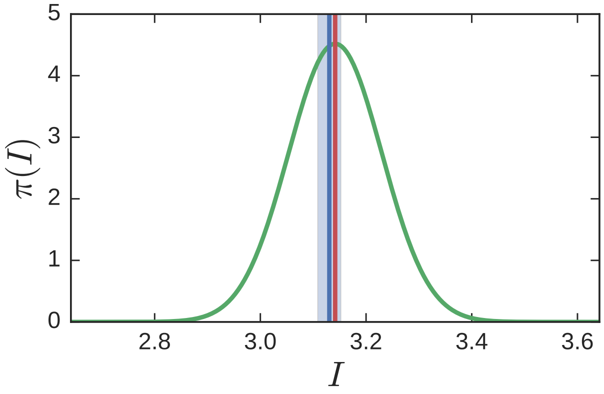
$$\sigma_{\bar{I}} = \frac{\sigma_I}{\sqrt{N}},$$

will provide a more accurate estimate of the integral.

Is there a free lunch here somewhere?

Unfortunately, no. For the same effort ($N \times n$ trials) a single MC provides an equally accurate estimate since,

$$\sigma_{\bar{I}} = \frac{\sigma_I}{\sqrt{N}} = \pi \frac{\sigma_f}{\sqrt{Nn}}.$$



A.3 Python Programs

A.3.1 Average of n Dice Rolls with Histogram

```
def avgDiceRoll(nDice):
    """nDice = n in the description above"""

    nSamples = 50000 # num of indep experiments
    z = np.zeros((nSamples))

    for i in range(nSamples):
        z[i] = np.mean(np.random.randint(1,7,nDice))

    # histogram
    n, bins, patches = plt.hist(z, 21, density=True)
    plt.xlim(1,6)
    plt.title('nDice = {0:d}'.format(nDice))
```

A.3.2 Error Estimation in direct MC Example

```
def simpleIntegral(npts):

    xi   = np.random.uniform(0, np.pi, npts)
    fi   = xi * np.sin(xi)

    intg   = np.pi * np.mean(fi)
    stdInt = np.pi/np.sqrt(npts) * np.std(fi)

    return intg, stdInt
```

A.4 Unnormalized Distributions and Importance Sampling

What happens when the probability distribution $\pi(x)$ is not normalized on the interval $[a, b]$?

Here, I shall denote such an unnormalized distribution by $\pi_u(x)$, and the underlying normalized distribution by $\pi(x)$

$$\pi_u(x) = c\pi(x)$$

Their (non)normalizations imply,

$$\int_a^b \pi_u(x)dx = c$$
$$\int_a^b \pi(x)dx = 1$$

We will consider how this affects method 2 on two problems we studied in this lecture,

- The standard integral:
$$E_\pi[f] = \int_a^b f(x)\pi(x)\,dx$$
- The importance formula integral
$$I = \int_a^b \frac{f(x)}{\pi(x)}\pi(x)dx$$

when we can sample from $\pi_i(x)$ and do not know the normalization constant c.

A.4.1 First Problem: $E_\pi[f]$

For the first problem, it turns out that we don't have to do anything different. The actual samples you draw from $\pi(x)$ or $\pi_u(x)$ are indistinguishable. Using method 2, we estimate

$$E_\pi[f] = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad X_i \sim \pi(x) \text{ or } \pi_u(x)$$

We never have to compute the value of $\pi(x)$ or $\pi_u(x)$ explicitly.

We can illustrate this by reconsidering a modified form of a previous example,

$$E_\pi[f] = \int_{-1}^{1} f(x)\,\pi(x)dx = \int_{-1}^{1} g(x)\,dx \tag{24}$$

with $f(x) = x^2$, and $\pi(x)$ is a truncated unit normal distribution, $\mathcal{N}(\mu = 0, \sigma = 1| - 1 < x < 1)$, and

$$\pi_u(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2}).$$

Thus, $\pi(x) = \pi_u(x)/c$, where
$$\int_{-1}^{1} \pi_u(x)dx = c = 0.6827.$$

Thus $\pi_u(x)$ is not normalized over the domain $[-1, 1]$. We can encode this problem in python.

```
f  = lambda x: x**2
pu = lambda x: 1./np.sqrt(2.*np.pi) * np.exp(-x**2/2.) # unnormalized
c  = scipy.integrate.quad(pu, -1, 1) # value of c
(0.682689492137086, 7.579375928402476e-15)

p  = lambda x: pu(x)/c        # normalized over [-1, 1]
g  = lambda x: f(x) * p(x)
```

Compute $E_\pi[f]$ over $[-1, 1]$, using method 1 and 2.

$$E_\pi[f] = \int_{-1}^{1} f(x)(\pi_u(x)/c)\, dx = \int_{-1}^{1} g(x)\, dx. \tag{25}$$

```
# method 1: uniform samples over the domain
x = np.random.uniform(-1, 1, size=10000)
gbar1 = np.mean(g(x))
fbar1 = gbar1 * (1 - (-1))
0.290944073298
```

Method 2 doesn't need to know the normalization constant explicitly, as long as it samples points from the correct distribution.

```
# method 2
x    = np.random.normal(0, 1, size=10000)
cond = np.logical_and(x > -1., x < 1.)
x    = x[cond]  # fewer points
fbar = np.mean(f(x))
0.286776441847
```

Note that this method did not use the function $\pi(x)$ explicitly.

A.4.2 Second Problem: $E_\pi[f/\pi]$

The second problem required a little more attention, when $\pi(x)$ is not normalized

$$I = \int_a^b \frac{f(x)}{\pi(x)} \pi(x) dx,$$

because we estimate the integral in method 2 by,

$$I = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{\pi(X_i)}, \qquad X_i \sim \pi(x)$$

Since we actually compute the value of $\pi(x)$ during the computation, it matters what $c = \pi_u(x)/\pi(x)$ is. This restriction can be overcome by using **self-normalizing importance sampling**.

Mathematically, the idea is straight-forward. We write $I = (b - a)E_\pi[f]$ where,

$$\begin{aligned}
E_\pi[f] &= \int_a^b f(x)u(x)dx \\
&= \int_a^b f(x)u(x)\frac{\pi_u(x)}{\pi_u(x)}dx \\
&= \frac{\dfrac{1}{b-a}\int\limits_a^b \dfrac{f(x)}{c\pi(x)}\pi_u(x)dx}{\dfrac{1}{b-a}\int\limits_a^b \dfrac{1}{c\pi(x)}\pi_u(x)dx}, && \text{denominator = 1} \\
&= \frac{\int_a^b w_u(x)f(x)\pi_u(x)dx}{\int_a^b w_u(x)\pi_u(x)dx}, && c \text{ cancels N/D,}
\end{aligned}$$

where $w_u(x) = 1/\pi_u(x)$ is a weighting term. This gives us a method to compute $E_\pi[f]$ using self-normalizing importance sampling.

- draw X_1, X_2, ..., X_n from $\pi_u(x)$ [same as $\pi(x)$]
- estimate $E_\pi[f]$ by,

  $$E_\pi[f] \approx \frac{\sum_{i=1}^n w_u(X_i) f(X_i)}{\sum_{i=1}^n w_u(X_i)} = \mu$$

  where $w_u(X_i) = 1/\pi_u(X_i)$.
- estimate $I = (b - a)E_\pi[f]$

An estimator for the variance σ_I for self-normalized importance sampling can be shown to be,

$$\sigma_I^2 = (b-a)^2 \frac{\sum_{i=1}^n w_u(X_i)^2 (f(X_i) - \mu)^2}{\left(\sum_{i=1}^n w_u(X_i)\right)^2}.$$