1 Importance Sampling and Extending the Random Walk Model

1.1 Sums of Independent Random Variables

If we consider a random walk with $x_0 = 0$, we can define it by

$$X_n = \sum_{i=1}^{n} Z_i,$$

where the $\{Z_i\}_{i \geq 1}$ are i.i.d. random variables with

$$\mathbb{P}(Z_1 = 1) = 1 - \mathbb{P}(Z_1 = -1) = p.$$

If we replace the $\{Z_i\}_{i \geq 1}$ with an arbitrary sequence of independent random variables, $\{Y_i\}_{i \geq 1}$ we are in the setting of sums of independent random variables. That is, we consider

$$S_n = \sum_{i=1}^{n} Y_i.$$

Now, in the case of a random walk, which is a sum of the $\{Z_i\}_{i \geq 1}$ random variables, we know the distribution of S_n (we calculated this earlier). However, this is not always the case. Normally, in order to find the distribution of a sum of n random variables, we have to calculate an n-fold convolution or use either moment generating functions or characteristic functions and hope that things work out nicely.

Recall, given two independent random variables, X and Y, the convolution of the distributions of X and Y is given by

$$\begin{aligned}\mathbb{P}(X + Y = z) &= \sum_{x=-\infty}^{\infty} \mathbb{P}(X = x)\mathbb{P}(Y = z - x) \\ &= \sum_{y=-\infty}^{\infty} \mathbb{P}(X = z - y)\mathbb{P}(Y = y)\end{aligned}$$

in the discrete case and the convolution, h, of the density of X, f, and the density of Y, g, is given by

$$h(z) = (f * g)(z) = \int_{-\infty}^{\infty} f(x)g(z - x)\mathrm{d}x = \int_{-\infty}^{\infty} f(z - y)g(y)\mathrm{d}y$$

in the continuous case. It is easy to see that the calculations can be pretty messy if lots of variables with different distributions are involved.

Sometimes, things are nice. For example, the sum of independent normal random variables is normally distributed and the sum of i.i.d. exponential random variables is distributed according to a special case of the gamma distribution (called the Erlang distribution). But most things are not so nice. For example, try to work out the distribution of n exponential random variables with parameters $\lambda_1, \dots, \lambda_n$.

There are various tools in mathematics that help us deal with sums of independent random variables. For example, we have the Lindeberg central limit theorem and a version of the strong law of large numbers. However, these do not answer all the questions we might reasonably ask and there are lots of random variables that do not satisfy the technical conditions of these theorems. Simulation is a useful tool for solving problems in these settings.

We will consider a couple of examples that use normal random variables. Recall that if $Z \sim \mathsf{N}(0,1)$ then $X = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$. This means we can simulate a normal random variable with mean μ and variance σ^2 in Matlab using the command

```
X = mu + sqrt(sigma_sqr) * randn;
```

Another new concept in the examples is *relative error*.

Definition 1.1 (Relative Error)**.** The relative error of an estimator $\widehat{\ell}$ is defined by

$$\mathsf{RE} = \frac{\sqrt{\mathrm{Var}(\widehat{\ell})}}{\ell}.$$

Basically, the relative error tell us the size of our estimator's error as a percentage of the thing we are tying to estimate (i.e., if we have a relative error of 0.01, that means that the standard deviation of our estimator is about 1 percent of ℓ). The relative error is often a more meaningful measure of error in settings where the thing we are trying to estimate, ℓ, is small. In practice, the relative error needs to be estimated.

Example 1.2 (Sums of log-normal random variables)**.** Consider a portfolio consisting of n stocks (which, for some reason, are independent of one another). At the start of the year, the stocks all have value 1. The changes in value of these stocks over a year are given by the random variables $V_1, \dots, V_n$, which are log-normal random variables, i.e., $V_1 = e^{Z_1}, ., V_n = e^{Z_n}$ with $Z_1 \sim \mathsf{N}(\mu_1, \sigma_1^2), \cdots, Z_1 \sim \mathsf{N}(\mu_n, \sigma_n^2)$. It is not so straightforward to calculate the distribution of $S_n = V_1 + \cdots + V_n$.

If $n = 5$, with $\boldsymbol{\mu} = (-0.1, 0.2, -0.3, 0.1, 0)$ and $\boldsymbol{\sigma^2} = (0.3, 0.3, 0.3, 0.2, 0.2)$, what is the probability that the portfolio is worth more than 20 at the end of the year? It is straightforward to use Monte Carlo to get an estimate of $\ell = \mathbb{P}(S_n > 20)$. Here, we check that the mean and variance of our simulation output correspond to the theoretical mean and variance (it never hurts to check things seem to be working properly). The mean of a log-normal random variable is given by $\exp\{\mu + \sigma^2/2\}$ and the variance is given by $(\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\}$. In addition, we estimate $\mathbb{E}\left[\max(V_1, \dots, V_5) \,|\, S_5 > 20\right]$ (that is, the average value of the largest portfolio component when the portfolio has a value bigger than 20) and $\mathbb{E}\left[S_5 \,|\, S_5 > 20\right]$.

Listing 1: Matlab code

```
N = 5*10^7; S = zeros(N,1); threshold = 20;
V_max = zeros(N,1); V_mean = zeros(N,1);

mu = [-0.1 0.2 -0.3 0.1 0]; sigma_sqr = [.3 .3 .3 .2 .2];

for i = 1:N
    Z = mu + sqrt(sigma_sqr) .* randn(1,5);
    V = exp(Z);
    V_max(i) = max(V);
    S(i) = sum(V);
end

est_mean = mean(S)
actual_mean = sum(exp(mu + sigma_sqr/2))

est_var = var(S)
actual_var = sum((exp(sigma_sqr) - 1) .* exp(2 * mu + sigma_sqr))

ell_est = mean(S>threshold)
ell_RE = std(S>threshold) / (ell_est * sqrt(N))

[event_occurs_index dummy_var] = find(S > threshold);
avg_max_v = mean(V_max(event_occurs_index))
avg_S = mean(S(event_occurs_index))
```

Running this one time produced the following output

```
est_mean = 5.6576
actual_mean = 5.6576
est_var = 1.9500
actual_var = 1.9511
ell_est = 2.3800e-06
ell_RE = 0.0917
avg_max_v = 15.9229
avg_S = 21.5756
```

Notice that, on average, the rare event seems to be caused by a single portfolio component taking a very large value (rather than all the portfolio components taking larger than usual values). This is typical of a class of random variables called heavy tailed random variables, of which the log-normal distribution is an example.

Example 1.3 (A gambler's ruin problem)**.** Consider an incompetent businessman. His company starts off with €10000 but makes a loss, on average, each day. More precisely, the profit or loss on the ith day is given by $Y_i \sim \mathsf{N}(-20, 10000)$. If his company can get €11000 in the bank, he is able to sell his company to a competitor. If his company's bank account drops below €0 he goes bankrupt. What is the probability that he is able to sell the company?

We can formulate this as a problem about hitting times. Define $S_n = \sum_{i=1}^n Y_i$, as the company bank account (minus the initial €10000) on the nth day. Define the time at which he can sell by

$$\tau_S = \inf\{n \geq 1 : S_n \geq 1000\}$$

and the time at which he can go bankrupt by

$$\tau_B = \inf\{n \geq 1 : S_n \leq -10000\}.$$

We want to know $\ell = \mathbb{P}(\tau_S < \tau_B)$. This is easy to simulate, we just increase n by one until either $S_n \leq -10000$ or $S_n \geq 1000$. We might as well find out $\mathbb{E}[\tau_S \,|\, \tau_S < \tau_B]$ and $\mathbb{E}[\tau_B \,|\, \tau_B < \tau_S]$ while we are doing that.

Listing 2: Matlab code

```
N = 10^7; sold = zeros(N,1); days = zeros(N,1);
mu = -20; sigma_sqr = 10000; sigma = sqrt(sigma_sqr);
up = 1000; low = -10000;

for i = 1:N
    S = 0; n = 0;
    while S > low && S< up
        S = S + (mu + sigma * randn);
        n = n + 1;
    end
    sold(i) = S > up;
    days(i) = n;
end

ell_est = mean(sold)
est_RE = sqrt(sold) / (sqrt(N)*ell_est)

[event_occurs_index dummy_var] = find(sold == 1);
[event_does_not_occur_index dummy_var] = find(sold == 0);

avg_days_if_sold = mean(days(event_occurs_index))
avg_days_if_bankrupt = mean(days(event_does_not_occur_index))
```

Running this one time produced the following output

```
ell_est = 0.0145
ell_RE = 0.0026
avg_days_if_sold = 52.9701
avg_days_if_bankrupt = 501.5879
```

1.2 Importance Sampling

In examples 1.2 and 1.3, the probabilities we were interested in were quite small. Estimating such quantities is usually difficult. If you think about it, if something only happens on average once every 10^6 times, then you will need a pretty big sample size to get many occurrences of that event. We can be a bit more precise about this. Consider the relative error of the estimator $\widehat{\ell}$ for $\ell = \mathbb{P}(X > \gamma) = \mathbb{E}\mathbb{I}(X > \gamma)$. This is of the form

$$\mathsf{RE} = \frac{\sqrt{\mathbb{P}(X > \gamma)\,(1 - \mathbb{P}(X > \gamma))}}{\mathbb{P}(X > \gamma)\sqrt{N}}.$$

So, for a fixed RE, we need

$$\sqrt{N} = \frac{\sqrt{\mathbb{P}(X > \gamma)\,(1 - \mathbb{P}(X > \gamma))}}{\mathbb{P}(X > \gamma)\mathsf{RE}} \Rightarrow N = \frac{1 - \mathbb{P}(X > \gamma)}{\mathbb{P}(X > \gamma)\mathsf{RE}^2}.$$

So $N = O(1/\mathbb{P}(X > \gamma))$ which means N gets big very quickly as $\ell = \mathbb{P}(X > \gamma) \to 0$. This is a big problem in areas where events with small probabilities are important. There are lots of fields where such events are important: for example, physics, finance, telecommunication, nuclear engineering, chemistry and biology. One of the most effective methods of estimating these probabilities is called *importance sampling*.

In the case of sums of independent random variables, the basic idea is to change the distributions of the random variables so that the event we are interested in is more likely to occur. Of course, if we do this, we will have a biased estimator. So, we need a way to correct for this bias. It is easiest to describe these things using continuous random variables and densities but, as we will see in the examples, everything works for discrete random variables as well.

Consider a random variable X taking values in $\mathbb{R}$ with density f. Suppose we wish to estimate $\ell = \mathbb{E}S(X)$. Note that we can write

$$\mathbb{E}S(X) = \int_{-\infty}^{\infty} S(x)\, f(x)\, \mathrm{d}x.$$

Now, this suggests the natural estimator

$$\widehat{\ell} = \frac{1}{N}\sum_{i=1}^{N} S(X^{(i)}),$$

where $X^{(1)}, \ldots, X^{(N)}$ are i.i.d. draws from the density f. Now, suppose the expectation of $S(X)$ is most influenced by a subset of values with low probability. For example, if $S(X) = \mathbb{I}(X > \gamma)$ and $\mathbb{P}(X > \gamma)$ is small, then this set of values would be $\{x \in \mathcal{X} : S(x) > \gamma\}$. We want to find a way to make this 'important' set of values happen more often. This is the idea of *importance sampling*. The idea is to sample $\{X^{(i)}\}_{i=1}^{N}$ according to another density, g, that ascribes much higher probability to the important set. Observe that, given a density g such that $g(x) = 0 \Rightarrow f(x)S(x) = 0$, and being explicit about the density used to calculate the expectation,

$$\begin{aligned}\mathbb{E}_f S(X) &= \int_{-\infty}^{\infty} S(x)\, f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} S(x)\frac{g(x)}{g(x)} f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} S(x)\frac{f(x)}{g(x)} g(x)\,\mathrm{d}x = \mathbb{E}_g \frac{f(X)}{g(X)} S(X).\end{aligned}$$

We call $f(x)/g(x)$ the *likelihood ratio*. This suggests, immediately, the importance sampling estimator

Definition 1.4 (The Importance Sampling Estimator)**.** The importance sampling estimator, $\widehat{\ell}_{\mathsf{IS}}$, of $\ell = \mathbb{E}S(X)$ is given by

$$\widehat{\ell}_{\mathsf{IS}} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X^{(i)})}{g(X^{(i)})} S(X^{(i)}),$$

where the $\{X^{(i)}\}_{i=1}^{N}$ are i.i.d. draws from the importance sampling density g.

Because we wish to use this estimator for variance reduction, it makes sense for us to calculate its variance.

Lemma 1.5. The variance of the importance sampling estimator, $\widehat{\ell}_{\mathsf{IS}}$, is given by

$$\mathrm{Var}(\widehat{\ell}_{\mathsf{IS}}) = \frac{1}{N}\left(\mathbb{E}_f\left[\frac{f(X)}{g(X)} S(X)^2\right] - \ell^2\right).$$

Proof. We have that

$$\begin{aligned}\mathrm{Var}(\widehat{\ell}_{\mathsf{IS}}) &= \frac{1}{N}\mathrm{Var}\left(\frac{f(X)}{g(X)} S(X)\right) \\ &= \frac{1}{N}\left(\mathbb{E}_g\left[\frac{f(X)^2}{g(X)^2} S(X)^2\right] - \left(\mathbb{E}_g \frac{f(X)}{g(X)} S(X)\right)^2\right) \\ &= \frac{1}{N}\left(\left[\int_{-\infty}^{\infty} \frac{f(x)^2}{g(x)^2} S(x)^2\, g(x)\,\mathrm{d}x\right] - \ell^2\right) \\ &= \frac{1}{N}\left(\left[\int_{-\infty}^{\infty} \frac{f(x)}{g(x)} S(x)^2\, f(x)\,\mathrm{d}x\right] - \ell^2\right) \\ &= \frac{1}{N}\left(\mathbb{E}_f\left[\frac{f(X)}{g(X)} S(X)^2\right] - \ell^2\right)\end{aligned}$$

□

Comparing $\mathrm{Var}(\widehat{\ell})$, the variance of the normal Monte Carlo estimator, to $\mathrm{Var}(\widehat{\ell}_{\mathsf{IS}})$, the variance of the importance sampling estimator, we see that

$$\mathrm{Var}(\widehat{\ell}_{\mathsf{IS}}) < \mathrm{Var}(\widehat{\ell}) \Leftrightarrow \mathbb{E}_f\left[\frac{f(X)}{g(X)} S(X)^2\right] < \mathbb{E}_f S(X)^2.$$

When we are estimating probabilities, $S(x)$ is an indicator function. For example, it could be $S(x) = \mathbb{I}(x > \gamma)$. Then,

$$\mathbb{E}S(X) = \mathbb{P}(X > \gamma) = \mathbb{E}\mathbb{I}(X > \gamma) = \mathbb{E}\mathbb{I}(X > \gamma)^2 = \mathbb{E}S(X)^2,$$

so the condition above reduces to requiring that $\mathbb{E}_f\left[\frac{f(X)}{g(X)} S(X)^2\right] < \ell$.

The above technology is easily combined to problems involving discrete random variables. Just replace integrals with sums and densities with probability mass functions.

Example 1.6 (Importance sampling with a normal random variable)**.** Consider the problem of estimating $\ell = \mathbb{P}(X > \gamma)$, where $X \sim \mathsf{N}(0,1)$. If γ is big, for example $\gamma = 5$, then ℓ is very small. The standard estimator of $\ell = \mathbb{P}(X > \gamma)$ is

$$\widehat{\ell} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(X^{(i)} > \gamma\right),$$

where the $\{X^{(i)}\}_{i=1}^N$ are i.i.d. $\mathsf{N}(0,1)$ random variables. This is not a good estimator for large γ. We can code this as follows.

Listing 3: Matlab code

```
gamma = 5; N = 10^7;
X = randn(N,1);
ell_est = mean(X > gamma)
RE_est = std(X > gamma) / (sqrt(N) * ell_est)
```

For $\gamma = 5$ with a sample size of 10^7, an estimate of the probability is 2×10^{-7} and an estimate of the relative error is 0.7071. So, this problem is a good candidate for importance sampling. An obvious choice of an importance sampling density is a normal density with variance 1 but with mean γ. The likelihood ratio $f(x)/g(x)$ is given by

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

Thus, the estimator will be of the form

$$\widehat{\ell}_{\mathsf{IS}} = \frac{1}{N}\sum_{i=1}^{N} \exp\left\{\frac{\gamma^2}{2} - X^{(i)}\gamma\right\} \mathbb{I}\left(X^{(i)} > \gamma\right),$$

where the $\{X^{(i)}\}_{i=1}^N$ are i.i.d. $\mathsf{N}(\gamma,1)$ random variables. The code for this follows.

Listing 4: Matlab code

```
gamma = 5; N = 10^7;
X = gamma + randn(N,1);
values = exp(gamma^2 / 2 - X*gamma) .* (X > gamma);
ell_est = mean(values)
RE_est = std(values) / (sqrt(N) * ell_est)
```

For $\gamma = 5$ with a sample size of 10^7 an estimate of the probability is 2.87×10^{-7} and an estimate of the relative error is 7.53×10^{-4}. We can check the true value in this case using the Matlab command

```
1 - normcdf(5)
```

This gives a value of 2.87×10^{-7} which is more or less identical to the value returned by our estimator.

If we have a sum of n independent variables, $X_1, \ldots, X_n$, with densities $f_1, \ldots, f_n$, we can apply importance sampling using densities $g_1, \ldots, g_n$. We would then have a likelihood ratio of the form

$$\prod_{i=1}^{n} \frac{f_i(x)}{g_i(x)}.$$

Everything then proceeds as before.

Example 1.7 (A rare event for a random walk)**.** Given a random walk, $\{X_n\}_{n \geq 0}$, with $x_0 = 0$ and $p = 0.4$, what is $\ell = \mathbb{P}(X_{50} > 15)$? We can estimate this in Matlab using standard Monte Carlo.

Listing 5: Matlab code

```
N = 10^5; threshold = 15;
n = 50; p = 0.4; X_0 = 0;
X_50 = zeros(N,1);

for i = 1:N
    X = X_0;
    for j = 1:n
        Y = rand <= p;
        X = X + 2*Y - 1;
    end
    X_50(i) = X;
end
ell_est = mean(X_50 > threshold)
RE_est = std(X_50 > threshold) / (sqrt(N) * ell_est)
```

Running this program once, we get an estimated probability of 1.2×10^{-4} and an estimated relative error of 0.29. This is not so great, so we can try using importance sampling. A good first try might be to simulate a random walk, as before, but with another parameter, $q = 0.65$. If we write

$$X_n = \sum_{i=1}^{n} Z_i,$$

then, the original random walk is simulated by generating the $\{Z_i\}_{i \geq 1}$ according to the probability mass function $p\,\mathbb{I}(Z = 1) + (1-p)\,\mathbb{I}(Z = -1)$. Generating the new random walk means generating the $\{Z_i\}_{i \geq 1}$ according to the probability mass function $q\,\mathbb{I}(Z=1)+(1-q)\,\mathbb{I}(Z=-1)$. This then gives a likelihood ratio of the form

$$\prod_{i=1}^{n} \frac{p\,\mathbb{I}(Z_i=1)+(1-p)\,\mathbb{I}(Z_i=-1)}{q\,\mathbb{I}(Z_i=1)+(1-q)\,\mathbb{I}(Z_i=-1)} = \prod_{i=1}^{n}\left[\frac{p}{q}\mathbb{I}(Z_i=1)+\frac{1-p}{1-q}\mathbb{I}(Z_i=-1)\right].$$

We can implement the estimator in Matlab as follows.

Listing 6: Matlab code

```
N = 10^5; threshold = 15;
n = 50; p = 0.4; X_0 = 0; q = 0.65;
X_50 = zeros(N,1); LRs = zeros(N,1);

for i = 1:N
    X = X_0; LR = 1;
    for j = 1:n
        Y = rand <= q;
        LR = LR * (p/q * (Y == 1) + (1-p) / (1 - q) * (Y == 0));
        X = X + 2*Y - 1;
    end
    LRs(i) = LR;
    X_50(i) = X;
end
ell_est = mean(LRs .* (X_50 > threshold))
RE_est = std(LRs .* (X_50 > threshold)) / (sqrt(N) * ell_est)
```

Running this program, we get an estimated probability of 1.81×10^{-4} and a relative error of 0.0059. We can check this makes sense by using the standard Monte Carlo estimator with a much bigger sample size. Using a sample size of $N = 10^8$, we get an estimate of 1.81×10^{-4}, confirming the importance sampling gives the right result.

1.2.1 Rules of Thumb for Effective Importance Sampling

Recall the definition of a *moment generating function*.

Definition 1.8 (Moment Generating Function)**.** We define the moment generating function of a random variable X by

$$M(\theta) = \mathbb{E}e^{\theta X}.$$

For $\theta = 0$, $M(\theta) = 1$. However, for other values of θ, it may not be the case that $M(\theta) < \infty$. In order for $M(\theta)$ to be finite for some $\theta \neq 0$, the probability of X taking very large (or small) values has to go to zero exponentially fast. This leads to the definition of *light-tailed* random variables. Usually, people assume that light-tailed means right light-tailed.

Definition 1.9 (Light-tailed random variable)**.** We say a random variable X is (right) light-tailed if $M(\theta) < \infty$ for some $\theta > 0$. We say X is left light-tailed if $M(\theta) < \infty$ for some $\theta < 0$.

The rules of thumb, which only apply when dealing with light-tailed random variables, are as follows.

- For sums of i.i.d. random variables, e.g., $\mathbb{P}(X_1 + \cdots + X_n > \gamma)$ choose the importance sampling density, g, so that $\mathbb{E}_g X_1 = \gamma/n$.
- For stopping time problems, e.g., $\mathbb{P}(\tau_A < \tau_B)$ or $\mathbb{P}(\tau_A < \infty)$, where the process is drifting away from A, the set of interest, choose g so that the drift of the stochastic process is reversed. For example, if $A = \{10, 11, \ldots\}$ and $S_n = \sum_{i=1}^n Y_i$, with $Y_i \sim \mathsf{N}(-1, 1)$, then choose g so that $\mathbb{E}_g Z_1 = 1$.