7 Randomness and determinism

Recall the definition of the random walk: it is a discrete-time stochastic signal $X = (X_t)_{t\in\mathbb{Z}_+}$ with the deterministic initial condition $X_0 = 0$ and the update rule

$$X_{t+1} = X_t + U_t, \qquad t = 0, 1, 2, \ldots$$

where $U = (U_t)_{t\in\mathbb{Z}}$ is an i.i.d. stochastic signal. Let $\mu = \mathbf{E}[U_0]$ and $\sigma^2 = \mathrm{Var}[U_0]$. Then

$$\mathbf{E}[X_t] = \mathbf{E}[U_0 + \ldots + U_{t-1}] = t\mu$$

and

$$\mathrm{Var}[X_t] = \mathrm{Var}[U_0 + \ldots + U_{t-1}] = \sum_{s=0}^{t-1} \mathrm{Var}[U_s] = t\sigma^2,$$

where we have used the fact that $U_0, \ldots, U_{t-1}$ are independent. Thus, both the mean and the variance of X_t grow linearly with t, which means that, if $\mu \neq 0$, then the random walk will drift farther and farther away from the origin as time goes by: at time t, with high probability it will be somewhere in the interval $[t\mu - \sqrt{t}\sigma, t\mu + \sqrt{t}\sigma]$. If the increments of the walk are zero-mean, then the walk will stay near the origin on average, but will take longer and longer excursions as t increases. Therefore, we are justified in saying that the amount of randomness in X_t increases with t — as $t \to \infty$, the value of X_t will become less and less predictable.

On the other hand, consider the *average* displacement of the random walk at time t:

$$\overline{X}_t \triangleq \frac{X_t}{t} = \frac{U_0 + \ldots + U_{t-1}}{t}.$$

Then $\mathbf{E}[\overline{X}_t] = \frac{1}{t}\mathbf{E}[X_t] = \mu$ and $\mathrm{Var}[\overline{X}_t] = \frac{1}{t^2}\mathrm{Var}[X_t] = \frac{\sigma^2}{t}$. We notice two things:

1. The expected average displacement is constant and equal to μ, which makes sense: μ is the average displacement per time step.
2. The variance of the expected average displacement *decays* as $\frac{1}{t}$.

Therefore, if we observe the random walk for a long enough time, then we will see that it will tend to spend a great deal of time around the point $x = \mu$. In fact, as we will make precise later, as $t \to \infty$, the average displacement $\overline{X}_t$ will be in the interval $[\mu - \frac{\sigma}{\sqrt{t}}, \mu + \frac{\sigma}{\sqrt{t}}]$ with overwhelming probability. This suggests that the operation of averaging has the effect of reducing the fluctuations around the mean, and in fact drives them to zero as $t \to \infty$. It is convenient to subtract off the mean μ and to focus on the random variables

$$S_t \triangleq \overline{X}_t - \mu = \frac{U_0 + \ldots + U_{t-1}}{t} - \mu.$$

Then $\mathbf{E}[S_t] = 0$ and $\mathrm{Var}[S_t] = \frac{\sigma^2}{t}$. Since any random variable with zero variance is deterministic and equal to its mean, we see that, in the limit as $t \to \infty$, the random variables S_t will become deterministic: $S_t \to 0$. This is the Law of Large Numbers (LLN).

Now, if we want to take a better look at the fluctuations of S_t around zero, we should scale S_t in such a way that the variance of the scaled quantity remains *constant* as t increases. Since $S_t = \frac{X_t - t\mu}{t}$, we see that multiplying S_t by $\sqrt{t}$ will have the desired effect:

$$\mathrm{Var}[\sqrt{t}S_t] = t\mathrm{Var}[S_t] = t \cdot \frac{\sigma^2}{t} = \sigma^2.$$

With this in mind, let us define

$$Z_t \triangleq \frac{\sqrt{t}S_t}{\sigma} = \frac{U_0 + \ldots + U_{t-1} - t\mu}{\sqrt{t\sigma^2}}.$$

Then $\mathbf{E}[Z_t] = 0$ and $\mathrm{Var}[Z_t] = \frac{t}{\sigma^2}\mathrm{Var}[S_t] = 1$. This scaling helps us "zoom in" on the flucutations of S_t around 0. As we will see shortly, at this scale the flucutations are Gaussian, with zero mean and unit variance. That is, as $t \to \infty$, the distribution of Z_t will approach $N(0,1)$. This is the Central Limit Theorem (CLT).

7.1 The LLN, the CLT, and stability of linear systems

Before proving the LLN and the CLT, it is instructive to look at these results through the lens of linear systems. We have two stochastic signals, $S = (S_t)_{t\in\mathbb{N}}$ and $Z = (Z_t)_{t\in\mathbb{N}}$, that are given by linear transformations of the i.i.d. stochastic signal U. In particular,

$$\begin{aligned} S_{t+1} &= \frac{U_0 + \ldots + U_t}{t+1} - \mu \\ &= \frac{U_0 + \ldots + U_{t-1}}{t+1} + \frac{U_t}{t+1} - \mu \\ &= \frac{t(S_t + \mu)}{t+1} + \frac{U_t}{t+1} - \mu \\ &= \frac{t}{t+1}S_t + \frac{U_t - \mu}{t+1}. \end{aligned}$$

Introducing the centered version of U_t, $V_t \triangleq U_t - \mu$, we can write so we can write

$$S_{t+1} = f_t(S_t, V_t), \qquad \text{where } f_t(s,v) \triangleq \frac{t}{t+1}s + \frac{v}{t+1}. \tag{7.1}$$

Similarly,

$$\begin{aligned} Z_{t+1} &= \frac{\sqrt{t+1}}{\sigma}S_{t+1} \\ &= \frac{\sqrt{t+1}}{\sigma}\left(\frac{t}{t+1}S_t + \frac{U_t - \mu}{t+1}\right) \\ &= \frac{\sqrt{t+1}}{\sigma}\left(\frac{t}{t+1}\frac{\sigma Z_t}{\sqrt{t}} + \frac{U_t - \mu}{t+1}\right) \\ &= \sqrt{\frac{t}{t+1}}Z_t + \frac{U_t - \mu}{\sigma\sqrt{t+1}}, \end{aligned}$$

which allows us to write

$$Z_{t+1} = g_t(Z_t, V_t), \qquad \text{where } g_t(z,u) \triangleq \sqrt{\frac{t}{t+1}}z + \frac{v}{\sigma\sqrt{t+1}}. \tag{7.2}$$

Noting that V_t is independent of S_t and Z_t, we see from (7.1) and (7.2) that both S and Z are discrete-time, continuous-state, time-inhomogeneous Markov chains, that S_{t+1} is a linear function of S_t and V_t, and that Z_{t+1} is a linear function of Z_t and V_t. Moreover, note that $f_t(0,0) = g_t(0,0) = 0$, so we can think about $s = 0$ and $z = 0$ as "equilibrium points" of these two time-varying linear systems. We can now restate the LLN and the CLT as follows:

1. As $t \to \infty$, S_t will converge to the equilibrium point $s = 0$.
2. As $t \to \infty$, the distribution of the fluctuations of Z_t around the equilibrium point $z = 0$ will approach that of a zero-mean, unit-variance Gaussian random variable.

7.2 Proving the LLN and the CLT

We now sketch the proofs of the LLN and the CLT by characterizing the limiting behavior of the probability distributions of S_t and Z_t. To that end, we will look at their characteristic functions. Our goal is to show that

$$\lim_{t\to\infty} \Phi_{S_t}(u) = 1 \tag{7.3}$$

and

$$\lim_{t\to\infty} \Phi_{Z_t}(u) = e^{-u^2/2}. \tag{7.4}$$

That is, as $t \to \infty$, the characteristic functions of S_t will converge to the characteristic function of the deterministic random variable taking the value 0, while those of Z_t will converge to the characteristic function of a $N(0,1)$ random variable. Since the distribution of a random variable is uniquely determined by its characteristic function, Eq. (7.3) gives the Law of Large Numbers, while Eq. (7.4) gives the Central Limit Theorem.

We will first express everything in terms of the characterisitc function $\Phi(u) = \mathbf{E}[e^{\mathrm{i}uU_0}]$ of U_0. Using the fact that the U_t's are i.i.d., we have

$$\begin{aligned}
\Phi_{S_t}(u) &= \mathbf{E}[e^{\mathrm{i}uS_t}] \\
&= \mathbf{E}\left[\exp\left\{\mathrm{i}u\left(\frac{U_0 + \ldots + U_{t-1}}{t} - \mu\right)\right\}\right] \\
&= e^{-\mathrm{i}u\mu}\mathbf{E}\left[\exp\left(\frac{\mathrm{i}u}{t}U_0 + \ldots + \frac{\mathrm{i}u}{t}U_{t-1}\right)\right] \\
&= e^{-\mathrm{i}u\mu}\mathbf{E}[e^{\mathrm{i}(u/t)U_0}\ldots e^{\mathrm{i}(u/t)U_{t-1}}] \\
&= e^{-\mathrm{i}u\mu}\mathbf{E}[e^{\mathrm{i}(u/t)U_0}]\ldots\mathbf{E}[e^{\mathrm{i}(u/t)U_{t-1}}] \\
&= e^{-\mathrm{i}u\mu}\left(\Phi\left(\frac{u}{t}\right)\right)^t
\end{aligned} \tag{7.5}$$

and

$$\begin{aligned}\Phi_{Z_t}(u) &= \mathbf{E}[e^{\mathrm{i}u\sqrt{t}S_t/\sigma}] \\ &= \Phi_{S_t}\left(\frac{u\sqrt{t}}{\sigma}\right) \\ &= e^{-\mathrm{i}u\sqrt{t}/\sigma}\left(\Phi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t. \end{aligned} \tag{7.6}$$

Now we will investigate the limits as $t \to \infty$. To that end, we will use first- and second-order Taylor approximations of the characteristic function $\Phi(u)$ around $u = 0$. Recall that the first-order Taylor approximation of a differentiable function f around the point $u = 0$ is given by

$$f(u) = f(0) + f'(0)u + R_1(u),$$

where the remainder term $R_1(u)$ has the property that $\lim_{u\to 0} \frac{R_1(u)}{u} = 0$. Similarly, the second-order Taylor approximation of a twice-differentiable function f around $u = 0$ is given by

$$f(u) = f(0) + f'(0)u + \frac{1}{2}f''(0)u^2 + R_2(u),$$

where the remainder $R_2(u)$ is such that $\lim_{u\to 0} \frac{R_1(u)}{u^2} = 0$. In the special case of f being the characteristic function of some random variable Z, i.e., $f(u) = \mathbf{E}[e^{\mathrm{i}uZ}]$, we have

$$f(0) = \mathbf{E}[e^{\mathrm{i}uZ}]\Big|_{u=0} = 1,$$

and the first two derivatives at $u = 0$ are

$$f'(0) = \frac{\mathrm{d}}{\mathrm{d}u}\mathbf{E}[e^{\mathrm{i}uZ}]\Big|_{u=0} = \mathrm{i}\mathbf{E}[Ze^{\mathrm{i}uZ}]\Big|_{u=0} = \mathrm{i}\mathbf{E}[Z]$$

and

$$f''(0) = \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}u}\mathbf{E}[Ze^{\mathrm{i}uZ}]\Big|_{u=0} = -\mathbf{E}[Z^2] = \mathbf{E}[Z]^2 - \mathrm{Var}[Z].$$

Now let us examine the term involving Φ in Eq. (7.5). Using the first-order Taylor approximation of $\Phi(u/t)$ around 0, we have

$$\begin{aligned}\left(\Phi\left(\frac{u}{t}\right)\right)^t &= \left(\Phi(0) + \Phi'(0)\frac{u}{t} + R_1\left(\frac{u}{t}\right)\right)^t \\ &= \left(1 + \frac{\mathrm{i}u\mu}{t} + R_1\left(\frac{u}{t}\right)\right)^t \\ &= \left(1 + \frac{1}{t}\left(\mathrm{i}u\mu + \xi_t\right)\right)^t, \end{aligned}$$

where we have defined $\xi_t \triangleq tR_1(u/t)$. Since R_1 is the remainder term in the first-order Taylor approximation, we have $\lim_{t\to\infty} \xi_t = 0$ (recall that u is fixed). Now we will use the following result: If $(a_t)_{t\in\mathbb{N}}$ is any sequence of complex numbers, such that the limit $a = \lim_{t\to\infty} a_t$ exists, then

$$\lim_{t\to\infty}\left(1+\frac{a_t}{t}\right)^t = e^a. \tag{7.7}$$

Applying (7.7) to the sequence $a_t = \mathrm{i}u\mu + \xi_t$, we get

$$\lim_{t\to\infty}\left(1+\frac{1}{t}(\mathrm{i}u\mu+\xi_t)\right)^t = e^{\mathrm{i}u\mu},$$

and therefore

$$\lim_{t\to\infty}\Phi_{S_t}(u) = e^{-\mathrm{i}u\mu}\lim_{t\to\infty}\left(\Phi\left(\frac{u}{t}\right)\right)^t = e^{-\mathrm{i}u\mu}e^{\mathrm{i}u\mu} = 1.$$

Next, we turn to (7.6). Writing

$$\Phi(u) = \mathbf{E}[e^{\mathrm{i}uU_0}] = \mathbf{E}[e^{\mathrm{i}u(V_0+\mu)}] = e^{\mathrm{i}u\mu}\Psi(u),$$

where $\Psi(u) \triangleq \mathbf{E}[e^{\mathrm{i}uV_0}]$ is the characteristic function of $V_0 = U_0 - \mu$, we can express the right-hand side of (7.6) as

$$\begin{aligned} e^{-\mathrm{i}u\sqrt{t}/\sigma}\left(\Phi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t &= e^{-\mathrm{i}u\sqrt{t}/\sigma}\left(e^{\mathrm{i}u/\sqrt{t}\sigma}\Psi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t \\ &= \left(\Psi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t. \end{aligned}$$

Using the second-order Taylor approximation of $\Psi(u/\sqrt{t}\sigma)$ at 0 in the above expression, we have

$$\begin{aligned} \left(\Psi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t &= \left(\Psi(0) + \frac{\Psi'(0)u}{\sqrt{t}\sigma} + \frac{\Psi''(0)}{2}\left(\frac{u}{\sqrt{t}\sigma}\right)^2 + R_2\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t \\ &= \left(1 - \frac{\sigma^2}{2}\left(\frac{u}{\sqrt{t}\sigma}\right)^2 + R_2\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t \\ &= \left(1 - \frac{1}{t}\left(\frac{u^2}{2} + \eta_t\right)\right)^t, \end{aligned}$$

where we have defined $\eta_t \triangleq tR_2(u/\sqrt{t}\sigma)$. Since R_2 is the remainder term in the second-order Taylor approximation, we have $\lim_{t\to\infty}\eta_t = 0$. Therefore, using (7.7) with $a_t = -\left(\frac{u^2}{2} + \eta_t\right)$, we get

$$\lim_{\to\infty}\left(1 - \frac{1}{t}\left(\frac{u^2}{2} + \eta_t\right)\right)^t = e^{-u^2/2}.$$

Consequently,

$$\lim_{t\to\infty}\Phi_{Z_t}(u) = \lim_{t\to\infty} e^{-\mathrm{i}u\sqrt{t}/\sigma}\left(\Phi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t = \lim_{t\to\infty}\left(\Psi\left(\frac{u}{\sqrt{t}\sigma}\right)\right)^t = e^{-u^2/2}.$$

7.3 Variance reduction by averaging

Informally, the Law of Large Numbers says that, if we average a large number of independent random variables $U_0, U_1, \ldots, U_{t-1}$ with common mean μ, then the resulting quantity $\overline{X}_t \triangleq \frac{1}{t}(U_0 + \ldots + U_{t-1})$ will be nearly constant (and equal to μ). Moreorver, the Central Limit Theorem says that, provided all the U_t's have the same finite variance σ^2, then, for all sufficiently large t, the rescaled average $\sqrt{t} \cdot \overline{X}_t$ will resemble a $N(\mu, \sigma^2)$ random variable. These two fundamental results of probability theory have many important consequences, and we will discuss them in what follows.

7.3.1 The Monte Carlo method

As we have discussed earlier, there are many cases where randomness can be beneficial. One such instance is the problem of numerical integration. Suppose that we wish to compute the definite integral

$$I = \int_a^b g(w) \mathrm{d}w,$$

where g is some function of interest, and where $-\infty \leq a < b \leq +\infty$. We assume that $g(w)$ is easy to evaluate for any given $w \in [a, b]$, but computing the integral in closed form is not possible. An ingenious idea, which has its origins in the Manhattan Project during World War II, is as follows: Pick a well-behaved pdf f supported on the interval $[a, b]$, i.e., $f(w) > 0$ for $w \in [a, b]$ and $f(w) = 0$ for $w \notin [a, b]$ and write

$$I = \int_a^b f(w) \frac{g(w)}{f(w)} \mathrm{d}w. \tag{7.8}$$

For example, if $[a, b]$ is a finite interval, then we can take f to be the pdf of a $U(a, b)$ random variable, in which case $f(w) = \frac{1}{b-a}$ for $w \in [a, b]$ and 0 otherwise; if $[a, b] = \mathbb{R}$, we can take the Gaussian pdf with mean 0 and variance 1. Now, if we define the function $h(w) \triangleq \frac{g(w)}{f(w)}$ for all $w \in [a, b]$, then Eq. (7.8) shows that the value I of the integral is equal to the *expectation* $\mathbf{E}[h(W)]$ with $W \sim f$:

$$I = \mathbf{E}[h(W)] = \mathbf{E}\left[\frac{g(W)}{f(W)}\right], \qquad W \sim f. \tag{7.9}$$

Now let us make two additional assumptions:

- We can easily generate i.i.d. samples $W_0, W_1, \ldots$ with pdf f.
- Given any point $w \in [a, b]$, it is easy to compute the value $h(w) = g(w)/f(w)$.

Then we can consider the following *randomized* procedure for *approximating* I: pick a sufficiently large t, generate random samples $W_0, W_1, \ldots, W_{t-1} \overset{\text{i.i.d.}}{\sim} f$, and compute

$$\widehat{I}_t \triangleq \frac{1}{t} \sum_{s=0}^{t-1} h(W_s) = \frac{1}{t} \sum_{s=0}^{t-1} \frac{g(W_s)}{f(W_s)}. \tag{7.10}$$

This is known as the *Monte Carlo method*[1], and $\widehat{I}_t$ is referred to as the *Monte Carlo estimate* of I.

To get an idea of how good of an estimate $\widehat{I}_t$ is, we will use the LLN and the CLT. First, note that, if we define the random variable $U \triangleq h(W) = \frac{g(W)}{f(W)}$, then (7.10) can be rewritten as

$$\widehat{I}_t \triangleq \frac{1}{t}\sum_{s=0}^{t-1} U_s.$$

Since $W_0, W_1, \ldots$ are i.i.d., so are $U_0, U_1, \ldots$, and moreover

$$\mathbf{E}[U_0] = \mathbf{E}[h(W_0)] = \int_a^b f(w)\frac{g(w)}{f(w)}\mathrm{d}w = I.$$

Therefore, by the LLN, $\widehat{I}_t$ will converge to I as $t \to \infty$. In other words, the more samples from the pdf f we generate, the better our Monte Carlo approximation will be. On the other hand, $\widehat{I}_t - I \neq 0$ for any finite t, but we know that

$$\mathbf{E}[(\widehat{I}_t - I)^2] = \mathrm{Var}[\widehat{I}_t] = \frac{1}{t}\mathrm{Var}[U],$$

so, provided $\mathrm{Var}[U] = \mathrm{Var}[g(W)/h(W)]$ is small, the absolute error $|\widehat{I}_t - I|$ will be small on average. With the help of the CLT, we can say even more — when t is sufficiently large, the probability distribution of the quantity

$$Z_t = \frac{\widehat{I}_t - I}{\sqrt{t\mathrm{Var}[U]}}$$

will be approximately normal with zero mean and unit variance. In particular, if we introduce the so-called *Q-function*

$$Q(z) \triangleq \frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-x^2/2}\mathrm{d}x, \tag{7.11}$$

then the CLT says, roughly, that

$$\mathbf{P}\left[\widehat{I}_t - I \geq a\sqrt{t\mathrm{Var}[U]}\right] \approx Q(a) \qquad \text{and} \qquad \mathbf{P}\left[\widehat{I}_t - I \leq -a\sqrt{t\mathrm{Var}[U]}\right] \approx Q(a)$$

for any $a > 0$ and all sufficiently large t. As we will see shortly, $Q(a) \leq e^{-a^2/2}$, and therefore, for all sufficiently large t and for all $a > 0$,

$$\begin{aligned}\mathbf{P}\left[\left|\widehat{I}_t - I\right| \geq a\sqrt{t\mathrm{Var}[U]}\right] &= \mathbf{P}\left[\left\{\widehat{I}_t - I \geq a\sqrt{t\mathrm{Var}[U]}\right\} \cup \left\{\widehat{I}_t - I \leq -a\sqrt{t\mathrm{Var}[U]}\right\}\right] \\ &\leq \mathbf{P}\left[\widehat{I}_t - I \geq a\sqrt{t\mathrm{Var}[U]}\right] + \mathbf{P}\left[\widehat{I}_t - I \leq -a\sqrt{t\mathrm{Var}[U]}\right] \\ &\approx 2e^{-a^2/2}.\end{aligned}$$

Of course, the variance of U is determined by the function g and on the pdf f, and it may not be easy to compute it exactly.

[1] The name "Monte Carlo," which is a reference to the famous Monte Carlo Casino in Monaco, was used as a code name at the Los Alamos Laboratory during World War II.

7.3.2 The benefit of diversification in financial portfolios

As we have seen above, taking the average of many independent random variables reduces the overall variance. If the random variables are not independent, simple averaging may not achieve this effect because of correlations, but it may be possible to reduce the variance using a *weighted* average. We will now see an illustration of this in the context of financial risk management.

Consider an investor who wishes to invest in several assets. Suppose that the assets are purchased at $t = 0$ and then sold at $t = 1$ (the units can be days, weeks, months, years, ...). The performance of each asset is measured by its *return*, which is typically defined as the difference between the sale and the purchase prices, normalized by the purchase price. Thus, if the return is positive, the investor nets a profit; if the return is negative, the investor suffers a loss; if the return is zero, the investor breaks even. The return of each asset is determined by the market, so it is reasonable to think of it as a random variable. The mean of that random variable is the *expected return*, and the variance tells us how much the actual return will tend to fluctuate around its mean value. So, the question is: how should one invest? One obvious idea is to purchase the asset with the best expected performance, i.e., the largest expected return. However, that asset may be highly volatile, i.e., the variance of the return may be rather large, which could result in big gains or big losses. Betting on the least volatile asset is not the most sensible thing either — as an old Russian proverb goes, "if you don't take risks, you don't get to drink champagne." The question is, can we quantify risk and use this to guide our investment strategy? The first quantitative approach, based on weighted averaging of assets to balance the variance of the return against the mean was proposed in 1952 by Harry Markowitz[2] who was awarded the Nobel Prize in Economics in 1990.

Suppose we have n assets whose returns $R_1, \ldots, R_n$ are random variables with known means $\mu_R(i) \triangleq \mathbf{E}[R_i]$ and covariances $C_R(i,j) \triangleq \mathrm{Cov}(R_i, R_j) = \mathbf{E}[(R_i - \mu_R(i))(R_j - \mu_R(j))]$. Here, $\mu_R(i)$ is the expected return of asset i, $C_R(i,i) = \mathrm{Var}[R_i]$ is the volatility of asset i (which is fancy finance-speak for the variance of the asset's return), and $C_R(i,j)$ for $j \neq i$ measures the correlation between assets i and j. A *portfolio* is a vector $p = (p_1, \ldots, p_n)^T$ of nonnegative weights that sum to 1, and it represents the allocation of the investor's wealth to the n assets. For example, if we have a budget of $\$x$ and each asset costs \$1, then, for each i, we invest the p_i fraction of $\$x$ in the ith asset. For future convenience, we introduce the random vector $R = (R_1, \ldots, R_n)^T$, the vector of expected returns $\mu = (\mu_R(1), \ldots, \mu_R(n))^T$, and the covariance matrix $C = (C_R(i,j))_{1 \leq i,j \leq n}$. We will assume that the expected returns are nonnegative: $\mu_R(i) \geq 0$ for all i.

With this notation, we can write down the actual return, the expected return, and the volatility of the portfolio p. The actual return is given by

$$R_p \triangleq \sum_{i=1}^{n} p_i R_i = p^T R,$$

the expected return is

$$\mu_p \triangleq \mathbf{E}[R_p] = \mathbf{E}\left[\sum_{i=1}^{n} p_i R_i\right] = \sum_{i=1}^{n} p_i \mu_R(i) = p^T \mu,$$

[2] Harry Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, March 1952.

and the volatility is

$$v_p \triangleq \text{Var}[R_p] = \text{Var}\left[\sum_{i=1}^{n} p_i R_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j C_R(i,j) = p^T C p.$$

According to Markowitz, for a given target value of the expected return, the best portfolio is the one whose volatility is the smallest. We will now *derive* the optimal portfolio $p^* = (p_1^*, \ldots, p_n^*)^T$ for a given target expected return r — i.e., we will find p^* to achieve

$$v^*(r) \triangleq \min_p \left\{ p^T C p : p^T \mu = r \right\},$$

where the minimum is over all valid portfolios.

This is a *constrained optimization problem*, which is solved using the method of Lagrange multipliers. Introduce the *Lagrangian*

$$L(p, \lambda, \nu) \triangleq p^T C p - \lambda(p^T \mu - r) - \nu(p^T e - 1), \tag{7.12}$$

where λ is the Lagrange multiplier corresponding to the expected return constraint $p^T \mu = r$ and ν is the Lagrange multiplier corresponding to the portfolio constraint $p^T e = p_1 + \ldots + p_n = 1$. Strictly speaking, we should also enforce the nonnegativity constraints $p_1, \ldots, p_n \geq 0$, but, as we will see, this constraint will be automatically satisfied by our solution. The optimal solution will be given by the triple (p^*, λ^*, ν^*), satisfying the conditions

$$\left.\frac{\partial L}{\partial p_i}\right|_{(p,\lambda,\nu)=(p^*,\lambda^*,\nu^*)} = 0, \qquad i = 1, \ldots, n \tag{7.13a}$$

$$\left.\frac{\partial L}{\partial \lambda}\right|_{(p,\lambda,\nu)=(p^*,\lambda^*,\nu^*)} = 0 \tag{7.13b}$$

$$\left.\frac{\partial L}{\partial \nu}\right|_{(p,\lambda,\nu)=(p^*,\lambda^*,\nu^*)} = 0. \tag{7.13c}$$

The first condition (7.13a) is given by

$$\sum_{j=1}^{n} C_R(i,j) p_j^* = \lambda^* \mu_R(i) + \nu^*, \qquad i = 1, \ldots, n$$

or, in a more succinct matrix form,

$$C p^* = \lambda^* \mu + \nu^* e. \tag{7.14}$$

The two remaining conditions (7.13b) and (7.13c) are

$$\mu^T p^* = r \qquad \text{and} \qquad e^T p^* = 1. \tag{7.15}$$

We will assume that the covariance matrix C is nonsingular, i.e., $\det C \neq 0$. Then we can invert (7.14) to get

$$p^* = C^{-1}(\lambda^* \mu + \nu^* e). \tag{7.16}$$

Notice that we have two unknowns, λ^* and ν^*, and two equations in (7.15). Substituting the expression for p^* from (7.16) into (7.15), we get

$$a\lambda^* + b\nu^* = r \tag{7.17a}$$
$$b\lambda^* + c\nu^* = 1 \tag{7.17b}$$

where we have defined $a \triangleq \mu^T C^{-1} \mu$, $b \triangleq \mu^T C^{-1} e$, and $c \triangleq e^T C^{-1} e$. This is a system of two linear equations in two unknowns, and it can be solved to get

$$\lambda^* = \frac{rc - b}{\Delta} \qquad \text{and} \qquad \nu^* = \frac{a - br}{\Delta}, \tag{7.18}$$

where we have defined $\Delta \triangleq ac - b^2$. This completes the computation of the optimal portfolio p^*, and we can also compute the volatility: since C is a symmetric matrix [i.e., $C^T = C$], so is its inverse C^{-1}, and therefore

$$\begin{aligned} v^*(r) &= (p^*)^T C p^* \\ &= \left(C^{-1}(\lambda^* \mu + \nu^* e)\right)^T C \left(C^{-1}(\lambda^* \mu + \nu^* e)\right) \\ &= (\lambda^* \mu + \nu^* e)^T (CC^{-1})^T C^{-1} (\lambda^* \mu + \nu^* e) \\ &= (\lambda^* \mu + \nu^* e)^T C^{-1} (\lambda^* \mu + \nu^* e) \\ &= (\lambda^*)^2 \mu^T C^{-1} \mu + 2\lambda^* \nu^* \mu^T C^{-1} e + (\nu^*)^2 e^T C^{-1} C. \end{aligned}$$

Recalling the definitions of a, b, c and using Eqs. (7.17) and (7.18), we can rewrite the last expression as

$$\begin{aligned} (\lambda^*)^2 a + \lambda^* \nu^* b + \lambda^* \nu^* b + (\nu^*)^2 e^T C^{-1} C &= \lambda^*(a\lambda^* + b\nu^*) + \nu^*(b\lambda^* + c\nu^*) \\ &= r\lambda^* + \nu^* \\ &= \frac{(rc - b)r + a - br}{\Delta} \\ &= \frac{c^2 r - 2br + a}{\Delta}. \end{aligned}$$

Altogether, this gives us the explicit expression for $v^*(r)$:

$$v^*(r) = \frac{cr^2 - 2br + a}{ac - b^2}. \tag{7.19}$$

This is a quadratic function of r, and it gives us the smallest volatility that can be achieved by any portfolio p for a given expected return $\mu_p = r$. Since the coefficient of r^2 is nonnegative, the

minimum value of $v^*(r)$ can be computed by setting the derivative $\frac{\mathrm{d}}{\mathrm{d}r}v^*(r)$ to 0, giving $r_{\min} = b/c$ and

$$v^*_{\min} = \min_r v^*(r) = \frac{c(b/c)^2 - 2b(b/c) + a}{ac - b^2} = \frac{a - b^2/c}{ac - b^2} = \frac{1}{c},$$

which characterizes the minimum-variance portfolio. In general, though, the expression (7.19), called the Markowitz frontier, allows one to precisely characterize the trade-off between risk and return.

The mean-variance portfolio selection of Markowitz was the first investment strategy that attempted to quantify risk. This approach can be criticized on several points, the main one being that the mean and the variance may not tell the whole story, unless the returns $R_1, \ldots, R_n$ are jointly Gaussian. In many situations, this is a good approximation to reality, but not in highly interconnected markets like the one we have today. Moreover, the construction of an optimal portfolio requires knowledge of the expected returns, the volatilities, and the correlations among different assets. Typically, these are estimated from a combination of historical data and forecasts, but, obviously, historical data can never be completely reliable. In addition, many finance managers may advise in favor of volatile assets, provided the probability that the asset's return falls below a given benchmark value is suitably small. To appreciate this, we need to look at the probability of *large deviations*, the subject we will briefly touch upon next.

7.4 Large deviations and the Chernoff bound

The LLN and the CLT tell us what happens when we average a sufficiently large number t of independent random variables. However, when is t large enough? For example, if $U_0, U_1, \ldots, U_{t-1}$ are i.i.d. random variables with mean μ, what is the *probability* that their average $\overline{X}_t = \frac{U_0 + \ldots + U_{t-1}}{t}$ rises far above or dips far below the mean μ? In other words, what can we say about the probability

$$\mathbf{P}\left[\left|\frac{U_0 + \ldots + U_{t-1}}{t} - \mu\right| \geq a\right] \tag{7.20}$$

for a given tolerance $a > 0$?

If the only piece of information we have is the mean and the variance of U_0, we can already show that the probability in (7.20) decreases at least as fast as $1/a^2$. To see this, let us forget for the moment that we are interested in the average $\overline{X}_t$ and ask for the probability that an arbitrary random variable Z takes values outside the interval $[\mathbf{E}Z - a, \mathbf{E}Z + a]$. Assume that $\mathrm{Var}[Z] < \infty$. Then the answer is given by *Chebyshev's inequality*, which says

$$\mathbf{P}\left[|Z - \mathbf{E}Z| \geq a\right] \leq \frac{\mathrm{Var}[Z]}{a^2}. \tag{7.21}$$

To prove (7.21), we will first establish another result, the so-called *Markov's inequality* — if Y is a random variable taking nonnegative values, then

$$\mathbf{E}[Y \geq a] \leq \frac{\mathbb{E}\,Y}{a}. \tag{7.22}$$

Before proving (7.22), let us first derive (7.21) as a consequence. We simply apply Markov's inequality to $Y = |Z - \mathbf{E}Z|^2$, for which we have $\mathbf{E}Y = \mathrm{Var}[Z]$:

$$\begin{aligned}\mathbf{P}[|Z - \mathbf{E}Z| \geq a] &= \mathbf{P}[|Z - \mathbf{E}Z|^2 \geq a^2] \\ &= \mathbf{P}[Y \geq a^2] \\ &\leq \frac{\mathbf{E}Y}{a^2} \\ &= \frac{\mathrm{Var}[Z]}{a^2}.\end{aligned}$$

Now we prove (7.22). To that end, we first express the probability $\mathbf{P}[U \geq a]$ as an expectation:

$$\mathbf{P}[Y \geq a] = \mathbf{E}[u(Y - a)],$$

where $u(\cdot)$ is the unit step function. We now use the following simple but important fact: if f and g are two functions such that $f(y) \leq g(y)$ for all y in their common domain, then, for any random variable Y on that domain, $\mathbf{E}[f(Y)] \leq \mathbf{E}[g(Y)]$. Thus, consider the functions $f(y) = u(y - a)$ and $g(y) = y/a$ on the positive half-line $\mathbb{R}_+$. Then $f(y) \leq g(y)$. Indeed, since $y/a \geq 0$, $u(y - a) \leq y/a$ for $0 \leq y \leq a$ (with equality at $y = a$), and $y/a \geq 1 = u(y - a)$ for $y \geq a$. Thus, for any random variable Y taking nonnegative real values,

$$\mathbf{P}[Y \geq a] = \mathbf{E}[u(Y - a)] \leq \mathbf{E}\left[\frac{Y}{a}\right] = \frac{\mathbf{E}Y}{a},$$

which is exactly what we wanted to prove.

Now, Chebyshev's inequality is often very loose. One way to improve on it is to consider other functions $g(y)$ that dominate the unit step $f(y) = u(y - a)$. A good choice is the exponential function $g_\lambda(y) \triangleq \exp(\lambda(y - a))$, where $\lambda > 0$ is a free paramter. The inequality $f(y) \leq g_\lambda(y)$ holds for all $y \in \mathbb{R}$, so we can consider the tail probability $\mathbf{P}[Y \geq a]$ without restricting Y or a to take nonnegative values. Then

$$\begin{aligned}\mathbf{P}[Y \geq a] &= \mathbf{E}[f(Y)] \\ &\leq \mathbf{E}[g_\lambda(Y)] \\ &= \mathbf{E}[e^{\lambda(Y-a)}] \\ &= e^{-\lambda a}\mathbf{E}[e^{\lambda Y}].\end{aligned}$$

Taking a closer look at the quantity $\mathbf{E}[e^{\lambda Y}]$, we see that it is reminiscent of the characteristic function $\Phi_Y(u) = \mathbf{E}[e^{\mathrm{i}uY}]$, and indeed equals $\Phi_Y(-\mathrm{i}a)$. In fact, if Y has a pdf f_Y, then $\mathbf{E}[e^{\lambda Y}]$ is given by the *Laplace transform* of f_Y at λ. At any rate, defining $\Lambda_Y(\lambda) \triangleq \log \mathbf{E}[e^{\lambda Y}]$ (this quantity is called the *cumulant generating function* of Y), we can write

$$\mathbf{P}[Y \geq a] \leq \exp\Big\{-(\lambda a + \Lambda_Y(\lambda))\Big\}. \tag{7.23}$$

We now observe that the left-hand side of (7.23) is a function of the threshold $a > 0$, while the right-hand side is a function of a and the free paramter λ. Since (7.23) holds for every $\lambda \geq 0$, we can take the minimum of both sides over all such λ to get the tightest inequality:

$$\begin{aligned}\mathbf{P}[Y \geq a] &\leq \min_{\lambda \geq 0} \exp\Big\{ -(\lambda a + \Lambda_Y(\lambda)) \Big\} \\ &= \exp\Big\{ -\max_{\lambda \geq 0} (\lambda a + \Lambda_Y(\lambda)) \Big\}, \end{aligned} \tag{7.24}$$

where in the second line we have used the fact that the function e^{-x} is decreasing. This inequality is known as the *Chernoff bound*, after the great statistician Hermann Chernoff, who was one (but not the only one) of hte inventors of this technique.

The benefit of (7.24) is that we can often carry out the maximization in (7.24) in closed form. As an example, consider the case of $Y \sim N(0,1)$. Then

$$\begin{aligned}\mathbf{E}[e^{\lambda Y}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda y} e^{-y^2/2} \mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2 - 2\lambda y)} \mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}} e^{\lambda^2/2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y-\lambda)^2} \mathrm{d}y \\ &= e^{\lambda^2/2}, \end{aligned}$$

so, for $Y \sim N(0,1)$, we have $\Lambda_Y(\lambda) = \lambda^2/2$ (a similar calculation gives $\Lambda_Y(\lambda) = e^{\lambda^2 \sigma^2/2}$ for $Y \sim N(0, \sigma^2)$). Then it is a simple exercise in calculus to prove that

$$\max_{\lambda \geq 0} \left\{ \lambda a + \frac{\lambda^2}{2} \right\} = \frac{a^2}{2}$$

(prove this!), which yields the famous Gaussian tail bound

$$Q(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \mathrm{d}y \leq e^{-a^2/2}.$$

One can get tighter bounds on the Q-function using more refined techniques, but this already gives us an idea of the power of the method.

Now, closer to home, let us consider the case when Y is the sum of t i.i.d. Bern(p) random variables: $Y = U_0 + \ldots + U_{t-1}$, where $U_0, \ldots, U_{t-1} \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$ — that is, $Y \sim \text{Bin}(t,p)$. Consider any $a \in [p,1]$. Then

$$\begin{aligned}\mathbf{P}[Y \geq at] &\leq \min_{\lambda \geq 0} e^{-\lambda a t} \mathbf{E}[e^{\lambda Y}] \\ &= \min_{\lambda \geq 0} e^{-\lambda a t} \mathbf{E}[e^{\lambda (U_0 + \ldots + U_{t-1})}] \\ &= \min_{\lambda \geq 0} e^{-\lambda a t} \mathbf{E}[e^{\lambda U_0} e^{\lambda U_1} \ldots e^{\lambda U_{t-1}}]. \end{aligned} \tag{7.25}$$

Since the U's are i.i.d., we have

$$\mathbf{E}[e^{\lambda U_0} e^{\lambda U_1} \dots e^{\lambda U_{t-1}}] = \left(\mathbf{E}[e^{\lambda U_0}]\right)^t,$$

and we can compute the expectation inside the parentheses directly:

$$\mathbf{E}[e^{\lambda U_0}] = 1 - p + pe^{\lambda}.$$

Hence, the Chernoff bound for $Y \sim \text{Bin}(t, p)$ takes the form

$$\begin{aligned}\mathbf{P}[Y \geq a] &\leq \left(\min_{\lambda \geq 0}\left(e^{-\lambda a}(1 - p + pe^{\lambda})\right)\right)^t \\ &= \left(\min_{\lambda \geq 0}\left((1-p)e^{-\lambda a} + pe^{\lambda(1-a)}\right)\right)^t.\end{aligned}$$

Thus, we need to compute the mininum value of the function $F(\lambda) \triangleq (1-p)e^{-\lambda a} + pe^{\lambda(1-a)}$ over all $\lambda \geq 0$. To that end, we first find the critical points by setting the derivative to zero:

$$F'(\lambda) = -a(1-p)e^{-\lambda a} + (1-a)pe^{\lambda(1-a)} = 0,$$

which gives

$$\lambda^* = \log\frac{a}{1-a} - \log\frac{p}{1-p},$$

which is nonnegative when $a \geq p$. Now, since $F''(\lambda) = a^2(1-p)e^{-\lambda a} + (1-a)^2 pe^{\lambda(1-a)} \geq 0$, $F(\lambda^*) = \min_{\lambda \geq 0} F(\lambda)$, and a straightforward calculation shows that

$$F(\lambda^*) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}.$$

Substituting this into (7.25), we get the Chernoff bound for $Y \sim \text{Bin}(t, p)$:

$$\mathbf{P}[Y \geq ta] \leq \exp\left(-t\left(a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}\right)\right), \qquad a \in [p, 1]. \tag{7.26}$$

This is *much* tighter than Chebyshev's inequality. For example, for $p = 0.1$, $t = 1000$, and $a = 0.2$, the Chernoff bound (7.26) gives the value of 5.2×10^{-20}, whereas Chebyshev's inequality gives the value of 2.25×10^{-3}. Moreover, by weakening (7.26) slightly, we can get the Gaussian-like tail bound

$$\mathbf{P}\left[\frac{Y - tp}{\sqrt{tp(1-p)}} \geq r\right] \lesssim \exp\left(-\frac{r^2}{2p(1-p)}(a-p)^2\right), \qquad r \geq 0, \tag{7.27}$$

which is consistent with the CLT. To derive this, let us consider the second-order Taylor approximation of the function

$$D_p(a) \triangleq a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$$

around the point $a = p$. A straightforward calculation shows that

$$D_p(p) = \frac{\mathrm{d}}{\mathrm{d}a} D_p(a)\Big|_{a=p} = 0 \qquad \text{and} \qquad \frac{\mathrm{d}^2}{\mathrm{d}a^2} D_p(a)\Big|_{p=a} = \frac{1}{p(1-p)},$$

so, for a sufficiently close to p, we can write

$$D_p(a) \approx \frac{1}{2p(1-p)}(a-p)^2. \tag{7.28}$$

Substituting (7.28) into (7.26), we get

$$\mathbf{P}[Y \geq ta] \lesssim \exp\left(-\frac{t(a-p)^2}{2p(1-p)}\right).$$

In particular, for any $r \geq 0$,

$$\begin{aligned}
\mathbf{P}\left[\frac{Y - tp}{\sqrt{tp(1-p)}} \geq r\right] &= \mathbf{P}\left[Y \geq tp + r\sqrt{tp(1-p)}\right] \\
&= \mathbf{P}\left[Y \geq t\Big(\underbrace{p + r\sqrt{\frac{p(1-p)}{t}}}_{=a}\Big)\right] \\
&\lesssim \exp\left(-\frac{r^2}{2tp(1-p)}\right),
\end{aligned}$$

so we recover (7.27).

7.5 Chernoff bound and statistical multiplexing

We close this lecture with a nice practical application of the Chernoff bound to digital telephony. A voice call has the bandwidth of 4 kHz. To digitize it, we sample it at the Nyquist rate of 8 kHz and then represent each sample using 1 byte. This means that each call requires a line capable of transmitting at the rate of 64 kbps (kilobits per second), and a naive calculation would suggest that, in order to transmit 100 calls, we would need a line with capacity of 6.4 Mbps (meagbits per second). However, this assumes that each call would be active all the time, while in reality 64 kbps is the *peak rate*, and most of the time there are silences. This means that we can multiplex many more calls onto a single line.

To cast this in statistical terms, suppose that we wish to multiplex n calls onto a single 64 Mbps line. If we divide time into slots, we can think of each call as a Bern(p) ranodm variable that takes the value 1 when the caller is speaking and 0 when the caller is silent. Since the calls originate from different places, we can assume that these random variables are mutually independent. Thus, the number of active calls in each time slot is a Bin(n, p) random varible Y, and the probability of

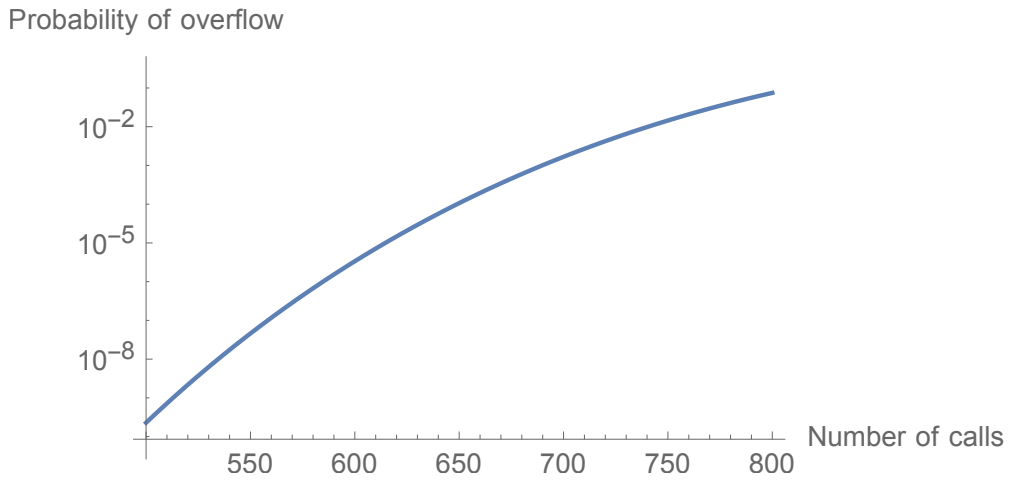

Figure 1: The probability of going over capacity versus the number of multiplexed calls.

going over capacity (i.e., when all n calls are active simultaneously) can be estimated using the Chernoff bound:

$$\mathbf{P}[Y \geq 100] \leq \exp\left(-n\left(\frac{100}{n}\log\frac{100/n}{p} + \left(1-\frac{100}{n}\right)\log\frac{1-100/n}{1-p}\right)\right). \tag{7.29}$$

Fig. 1 shows a plot of the bound in (7.29) (with the log scale on the vertical axis) for $p = 0.1$. For example, we can multiplex 600 calls if we are willing to accept exceeding capacity with probability of 3×10^{-6} and 800 calls if we are willing to tolerate going over capacity with probability of 0.07.