Economics

A redevelopment of economic theory without parallel universes

Ole Peters and Alexander Adamou

LONDON MATHEMATICAL LABORATORY

2019/10/30 at 14:40:18

Contents

Chapter 1

Why this book?

{chapter:Why}

In this introductory chapter we lay the conceptual foundation for the rest of what we have to say about redeveloping economic theory. In Sec. 1.1 we play a simple coin-toss game and analyze it numerically, by Monte Carlo simulation, and analytically, with pen and paper. The game motivates the introduction of the expectation value and the time average, which in turn lead to a discussion of ergodic properties. The ergodicity question – whether time averages are identical to expectation values – will turn out to be the key to our redevelopment of formal economics. This is because ergodicity hadn't been established as a concept when the original formalism was developed. We note the importance of rates as ergodic observables. This section also introduces the concepts of a random variable, a stochastic process, scalars as representations of transitive preferences, logarithms and exponentials, and dimensional analysis.

In Sec. 2.10 we notice that wealth on logarithmic scales follows a random walk in our game, and we relate this to Brownian motion, as the continuous-time limit of the random walk. This allows us to introduce Brownian motion and its scaling properties that are robust enough to yield insights into more complicated models.

Finally, we ask in Sec. 2.11 what wealth in our game is doing in the continuum limit but on linear scales. This takes us to geometric Brownian motion, which will be our starting point for much of the rest of these lectures. We derive ensemble-average and time-average growth rates for geometric Brownian motion, by explicitly taking the continuous-time limit, and then state the key result of Itô calculus, (Eq. 2.64) and (Eq. 2.65), which allows an easier derivation of these growth rates and will be relied on in later chapters.

Some historical perspective is provided to understand the prevalence or absence of key concepts in modern economic theory and other fields. The emphasis is on introducing key concepts and useful machinery, with more formal treatments and applications in later chapters.

1.1 The game

{section:The_game}

Imagine we offer you the following game: we toss a coin, and if it comes up heads we increase your monetary wealth by 50%; if it comes up tails we reduce your wealth by 40%. We're not only doing this once, we will do it many times, for example once per week for the rest of your life. Would you accept the rules of our game? Would you submit your wealth to the dynamic our game will impose on it?

Your answer to this question is up to you and will be influenced by many factors, such as the importance you attach to wealth that can be measured in monetary terms, whether you like the thrill of gambling, your religion and moral convictions and so on.

In these notes we will mostly ignore these factors. We will build an extremely simple model of your wealth, which will lead to an extremely simple and powerful model of the way you make decisions that affect your wealth. We are interested in analyzing the game mathematically, which requires a translation of the game into mathematics. We choose the following translation: we introduce the key variable, $x(t)$, which we refer to as "wealth". We refer to t as "time". It should be kept in mind that "wealth" and "time" are just names that we've given to mathematical objects. We have chosen these names because we want to compare the behaviour of the mathematical objects to the behaviour of wealth over time, but we emphasize that we're building a model – whether we write $x(t)$, or wealth(time), or ☺(ϟ) makes no difference to the mathematics.

The usefulness of our model will be different in different circumstances, ranging from completely meaningless to very helpful. There is no substitute for careful consideration of any given situation, and labeling mathematical objects in one way or another is certainly none.

Having got these words of warning out of the way, we define our model[1] of the dynamics of your wealth under the rules we just specified. At regular intervals of duration δt we randomly generate a factor $r(t)$ with each possible value $r_i \in \{0.6, 1.5\}$ occurring with probability 1/2,

$$r(t) = \begin{cases} 0.6 & \text{with probability } 1/2 \\ 1.5 & \text{with probability } 1/2 \end{cases} \tag{1.1}$$

{eq:law}

and multiply current wealth by that factor, so that

$$x(t) = r(t)x(t - \delta t). \tag{1.2}$$

{eq:gamble}

We have good reasons to suspect that this model will do something interesting. It's not just a silly game because it has one important property: it's multiplicative. Every time the coin is tossed, wealth is *multiplied* by one of the two possible factors. This introduces a reference point, or state-dependence, of the absolute size of the change. Processes with this property are very common in nature – the amount of anything that reproduces changes in proportion to what's already there. In the absence of other constraints, such as limited resources, this applies to the dynamics of the biomass of a cell culture. In fact, life itself has been defined as that which produces more of itself [**?**]: life and evolution begin with the minimal chemical structure that can copy itself. The rest,

[1] For those in the know: "our" coin toss is a discrete version of geometric Brownian motion, the workhorse model of financial mathematics and much more.

in a sense, is details. We will see, especially in Chap. **??**, how multiplicativity generates interesting structure in – including but not limited to – economics.

Without discussing in depth how realistic a representation of your wealth this model is (for instance your non-gambling related income and spending are not represented in the model), and without discussing whether randomness truly exists and what the meaning of a probability is, we simply switch on a computer and simulate what might happen. You may have many good ideas of how to analyze our game with pen and paper, but we will just generate possible trajectories of your wealth and pretend we know nothing about mathematics or economic theory. Figure 1.1 is a trajectory of your wealth, according to our computer model as it might evolve over the course of 52 time steps (corresponding to one year given our original setup).

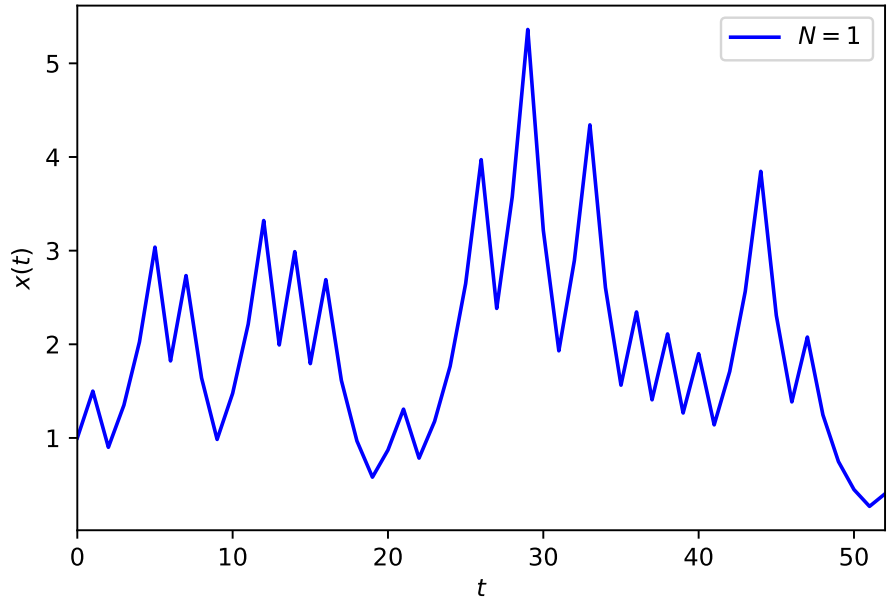


Figure 1.1: Wealth $x(t)$ resulting from a computer simulation of our game, as specified by (Eq. 1.1) and (Eq. 1.2), for 52 time steps (corresponding to one year in the given setup). {fig:1_1}

A cursory glance at the trajectory does not reveal much structure. Of course there are regularities, for instance at each time step $x(t)$ changes, but no trend is discernible – does this trajectory have a tendency to go up, does it have a tendency to go down? Neither? What are we to learn from this simulation? Perhaps we conclude that playing the game for a year is quite risky, but is the risk worth taking?

1.1.1 Averaging over many trials

The trajectory in Fig. 1.1 doesn't tell us much about overall tendencies. There is too much noise to discern a clear signal. A common strategy for getting rid of noise is to try again. And then try again and again, and look at what happens on average. An example of the technique is Shannon's error-correcting code:

instead of sending the message 0 (or 1), send the relevant digit 3 times. The recipient averages over the received digits and takes the closest possibility. If one out of three digits was miscommunicated because of noise, the code nonetheless recovers the original message: averaging gets rid of noise.

So let's try this in our case and see if we can make sense of the game. In Fig. 1.2 we average over a finite number, N, of trajectories. We call a collection of trajectories an ensemble. We shall see that in the limit $N \to \infty$ the ensemble average converges to the expectation value, and indeed the terms "ensemble average" and "expectation value" are synonyms. To avoid confusion we will be explicit when N is finite: in Fig. 1.2 we plot "finite-ensemble averages."

DEFINITION: **Finite-ensemble average**

The finite-ensemble average of the quantity z at a given time t is

$$\langle z(t) \rangle_N = \frac{1}{N} \sum_{i=1}^{N} z_i(t), \tag{1.3}$$ {eq:f_ens}

where i indexes a particular realization of $z(t)$ and N is the number of realizations included in the average.

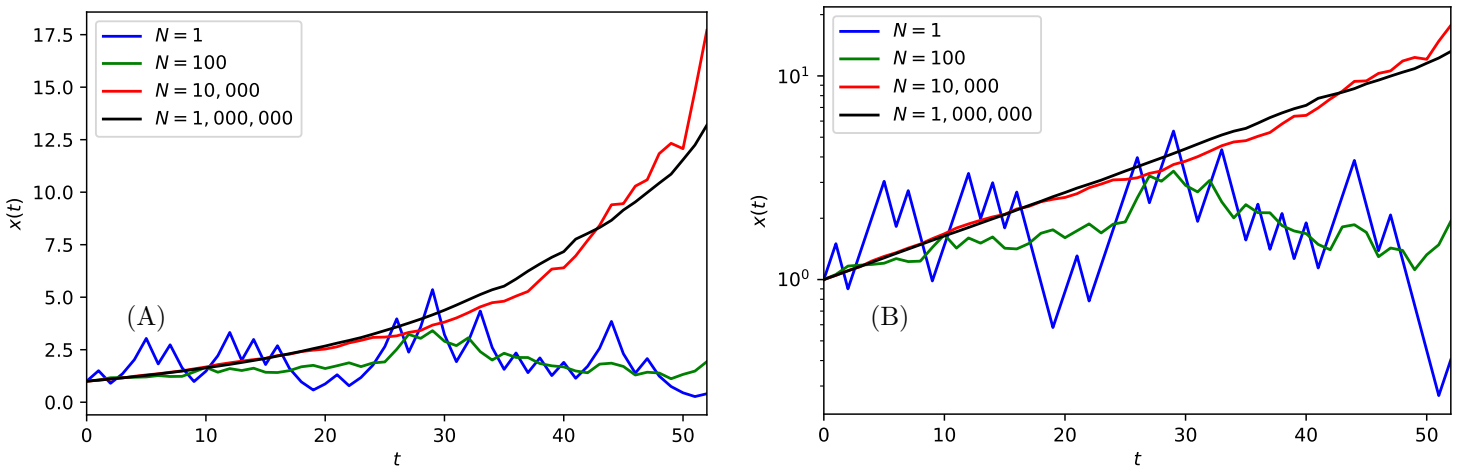


Figure 1.2: Finite-ensemble averages $\langle x(t) \rangle_N$ for ensemble sizes $N = 1, 10^2, 10^4, 10^6$. The noise in the finite-ensemble average diminishes systematically as N increases. (A) on linear scales the multiplicative (non-linear) nature of the process is apparent, (B) on logarithmic scales the multiplicative process is additive in time, and the finite-ensemble average for $N = 10^6$ is a straight line except for small fluctuations. {fig:1_2}

As expected, the more trajectories are included in the average, the smaller the fluctuations of that average. For $N = 10^6$ hardly any fluctuations are visible. Since the noise-free trajectory points up it is tempting to conclude that

my own wealth will similarly go up and conclude that the risk of the game is worth taking. This reasoning has dominated economic theory for about 350 years now. But it is flawed. The correction of this flaw and its far-reaching consequences constitute our research program.

1.1.2 Averaging over time

Does our analysis necessitate the conclusion that the gamble is worth taking? Of course it doesn't, otherwise we wouldn't be belabouring this point. Our critique will focus on the type of averaging we have applied – we didn't play the game many times in a row as would correspond to the real-world situation of repeating the game once a week for the rest of your life. Instead we played the game many times in parallel, which corresponds to a different setup[2].

We therefore try a different analysis. Figure 1.3 shows another simulation of your wealth. This time we don't show an average over many trajectories but a simulation of a single trajectory over a long time. Noise is removed also in this case but in a different way: to capture visually what happens over a long time we have to zoom out – more time has to be represented by the same amount of space on the page. In the process of this zooming-out, small short-time fluctuations will be diminished. Eventually the noise will be removed from the system just by the passage of time.

[2]This different setup could be realised by splitting your wealth into N equal parts and betting each part in a different sequence of independent coin tosses. But the rules of our game as we defined it don't allow that. Another interesting setup allows the gambler to choose what proportion of his wealth he wants to wager. These conditions were studied by Kelly [**?**] and are known to every professional poker player. In the present setup, betting 1/4 of your wealth in each coin toss will lead to the fastest possible growth if you're allowed to choose your wager. You can think of the proportion not wagered as frozen in time: it ensures that, to some extent, you can restore the conditions before the coin toss, which is a bit like allowing you to step back in time.

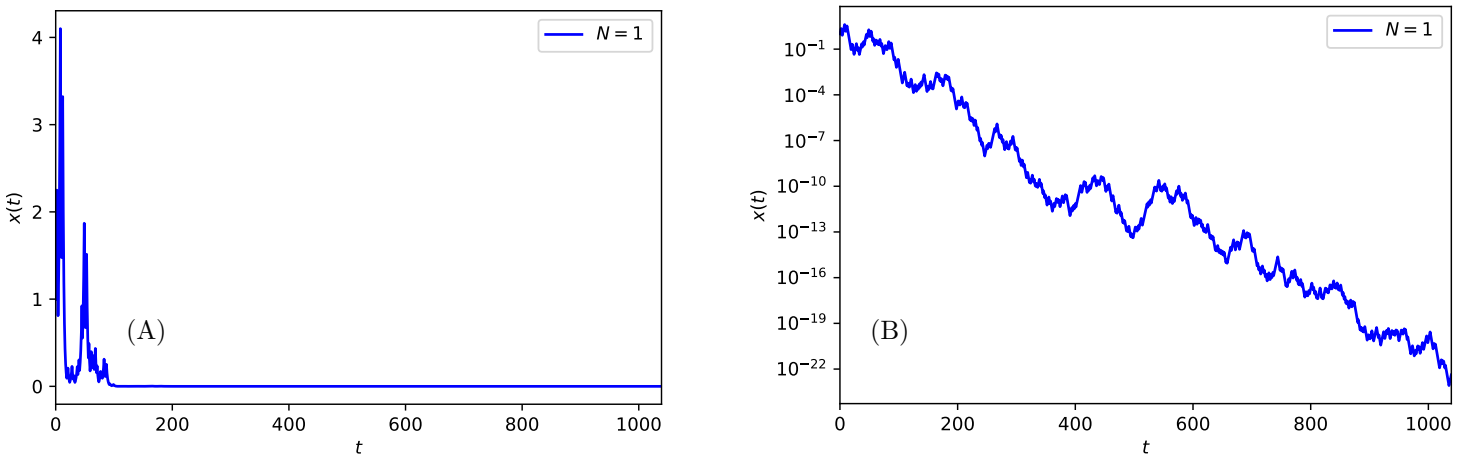


Figure 1.3: Single trajectory over 1,040 time steps, corresponding to 20 years in our setup. (A) on linear scales all we see is wealth quickly dropping to zero, (B) logarithmic scales are more appropriate for the multiplicative process and reveal systematic exponential decay. {fig:1_3}

The trajectory in Fig. 1.3 is, of course, random, but the apparent trend emerging from the randomness strongly suggests that our initial analysis in Fig. 1.2 does not reflect what happens over time in a single system. Several important messages can be derived from the observation that an individual trajectory grows more slowly (or decays faster) over time than an average of a large ensemble.

1. An individual whose wealth follows (Eq. 1.2) will make poor decisions if he uses the finite-ensemble average of wealth as an indication of what is likely to happen to his own wealth.

2. The performance of the average (or aggregate) wealth of a large group of individuals differs systematically from the performance of an individual's wealth. In our case large-group wealth grows (think gross domestic product (GDP)), whereas individual wealth decays.

3. For point 2 to be possible, *i.e.* for the average to outperform the typical individual, wealth must become increasingly concentrated in a few extremely rich individuals. The wealth of the richest individuals must be so large that the average becomes dominated by it, so that the average can grow although almost everyone's wealth decays. Inequality increases in our system.

The two methods we've used to eliminate the noise from our system are well known. The first method is closely related to the mathematical object called the "expectation value," and the second is closely related to the object called the "time average."

Summary of Chap. 1

Suppressed.

Part I

Tools

Chapter 2

Tools {chapter:Tools}

In this chapter we motivate and introduce the basic mathematical tools we will use. In Sec. 1.1 we play a simple coin-toss game and analyze it numerically, by Monte Carlo simulation, and analytically, with pen and paper. The game motivates the introduction of the expectation value and the time average, which in turn lead to a discussion of ergodic properties. As we have seen, the ergodicity question – whether time averages are identical to expectation values – is the key to our redevelopment of formal economics. This is because ergodicity hadn't been established as a concept when the original formalism was developed. The scientific search for stable structures leads to constants in deterministic settings. When randomness is introduced, the role previously played by constants is taken on by ergodic observables. We also introduce the concepts of a random variable, a stochastic process, scalars as representations of transitive preferences, logarithms and exponentials, and dimensional analysis.

In Sec. 2.10 we notice that wealth on logarithmic scales follows a random walk in our game, and we relate this to Brownian motion, as the continuous-time limit of the random walk. This allows us to introduce Brownian motion and its scaling properties that are robust enough to yield insights into more complicated models.

Finally, we ask in Sec. 2.11 what wealth in our game is doing in the continuum limit but on linear scales. This takes us to geometric Brownian motion, which will be our starting point for much of the rest of these lectures. We derive ensemble-average and time-average growth rates for geometric Brownian motion, by explicitly taking the continuous-time limit, and then state the key result of Itô calculus, (Eq. 2.64) and (Eq. 2.65), which allows an easier derivation of these growth rates and will be relied on in later chapters.

Some historical perspective is provided to understand the prevalence or absence of key concepts in modern economic theory and other fields. The emphasis is on introducing concepts and useful machinery, with applications in later chapters.

2.1 Random variable

{section:random_variable}

In economics, as elsewhere, we are often interested in 'experiments' whose outcomes we don't yet know. Examples are each coin toss in the game in Chap. 1 and the result of a football match. We might know something about the experiment, such as the possible outcomes and that some are more plausible than others, but we are ignorant of the actual outcome. Luckily, we can use probability theory, a branch of mathematics, to build models of our ignorance.

Often the experimental outcomes have, or can be mapped to, numerical values. For example, each coin toss in the game has possible outcomes heads and tails, which correspond to wealth multipliers 1.5 and 0.6. In such cases, we model the uncertain numerical value as a *random variable*. We assume you have seen random variables before and we will not give a lengthy technical account. Instead, we recommend the two-page discussion in [**?**, p. 2], whose key points we reproduce here.

A random variable, Z, is defined by:

- the set of its possible values; and
- a probability distribution over this set.

The set of values of Z may be continuous, like the interval $(3, 12)$ or the real numbers, $\mathbb{R}$; discrete, like $\{4, 7.8, 29\}$ or the integers, $\mathbb{Z}$; or a combination of the two. The probability distribution is a function which maps values to probabilities. So

$$\mathrm{P}\left[Z = z\right] = p \tag{2.1}$$

means that the outcome $Z = z$ is associated with the probability p. A specific value, z, is sometimes called an 'instance' or 'realisation' of the random variable, Z. While not obligatory, it is a common convention to denote random variables in upper case and realisations in lower case.

Forget, for the moment, what probabilities might mean in the context of an experiment. In purely mathematical terms, they are just real numbers associated with outcomes. The probability distribution has two constraints:

- the probability of any outcome must be non-negative; and
- the probability of an outcome that is certain to happen, *i.e.* that includes all possible outcomes, must be one.

The latter is a normalisation condition which fixes the scale of the probabilities.

Discrete random variable

When the set of outcomes is discrete, say $\{z_1, \ldots, z_M\}$, we assign a probability, p_j, to each outcome, z_j, such that

$$\mathrm{P}\left[Z = z\right] = \begin{cases} p_j & \text{if } z = z_j \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and

$$\sum_{j=1}^{M} p_j = 1. \tag{2.3}$$

Continuous random variable
Most of the models we will study contain random variables whose possible outcomes form a continuous set. In this case, Z can take uncountably many values, to which we can't assign non-zero probabilities while maintaining the normalisation condition. Instead, we assign probabilities to intervals. We define a probability density function (PDF), $\mathcal{P}_Z(z)$, such that the probability of Z being in the interval (a, b) is given by

$$\mathrm{P}\left[a \leq Z \leq b\right] = \int_a^b \mathcal{P}_Z(z)dz. \tag{2.4}$$

The PDF is a non-negative function, $\mathcal{P}_Z(z) \geq 0$, normalised so that the probability of the certain outcome, *i.e.* the integral over all possible outcomes, is one:

$$\int_{-\infty}^{+\infty} \mathcal{P}_Z(z)dz = 1. \tag{2.5}$$

Note the difference between subscript and argument: $\mathcal{P}_Z(z)$ is the probability density of the random variable Z at value z. You might find it helpful to think of $\mathcal{P}_Z(z)\delta z$ as the approximate probability of Z being in the small interval $(z, z + \delta z)$ close to z.

Interpretation
So far we have said nothing of the meaning of probabilities: they are simply numbers assigned to outcomes of random variables. One way to interpret probabilities is to imagine many separate experiments, of whose outcomes are identically ignorant. For example, we can imagine tossing the same coin many different times, assuming that each coin toss is equally likely to result in heads. Suppose we perform N experiments and record the number of times, n, that a particular outcome occurs. Under the so-called 'frequentist' interpretation, the appropriate probability to assign to this outcome is its relative frequency, n/N, in the limit $N \to \infty$. In the coin toss example, the probability assigned to heads would be 0.5 if the coin were unbiased; if biased, it would be some other number between 0 and 1.

Note also that time does not appear in the random variable setup. Of course, there is nothing stopping us from using a probability distribution which depends on time or, indeed, any other variable, like the day of the week or the country we are in. Such parametrisations of the random variable do not change fundamentally its mathematical structure: a set of outcomes and associated probabilities. When we consider probability distributions of random variables that do depend on time, such as wealth, $x(t)$, in the coin tossing game, we will make the time dependence explicit. By default we assume random variables are time-independent

2.2 Expectation value

The *expectation value*[1] is a property of a random variable. Denoted by $\mathrm{E}\left[Z\right]$, it is the probability-weighted average of the realisations of Z.

[1] Also known as expected value, mathematical expectation, first moment, and mean.

DEFINITION: **Expectation value**

If Z is a discrete random variable, its expectation value is the sum of the possible realisations, z_j, weighted by their probabilities, p_j:

$$\mathrm{E}[Z] = \sum_j p_j z_j. \tag{2.6}$$ {eq:exp_sum}

If Z is a continuous random variable, its expectation value is the integral over the possible realisations weighted by the probability density:

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} \mathcal{P}_Z(z) z dz. \tag{2.7}$$

The sum or the integral may not exist, in which case the random variable does not have an expectation value. Examples of this are the payout of the St Petersburg lottery in Sec. **??** and the power-law distributed random variable with PDF:

$$\mathcal{P}_Z(z) = \begin{cases} z^{-2} & z \geq 1 \\ 0 & z < 1. \end{cases} \tag{2.8}$$

2.3 Ensemble average

A conceptually different quantity from the expectation value is the *ensemble average*. The idea is that, instead of weighting the average of the possible realisations of Z by their probabilities, we take an unweighted average of a collection of realisations. For a finite number of realisations, we call this the *finite-ensemble average* and denote it by $\langle Z \rangle_N$.

DEFINITION: **Finite-ensemble average**

The finite-ensemble average of a random variable, Z, is the average over a finite number, N, of realisations,

$$\langle Z \rangle_N = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{2.9}$$ {eq:f_ens}

where z_i denotes the i^{th} realisation of Z.

You may know this as the sample average or sample mean.

The finite-ensemble average is itself a random variable: each finite ensemble of size N can contain different realisations of Z which sum to a different total. As $N \to \infty$, this random average may converge with probability one to a unique constant. If it does, we call this limiting value the ensemble average and denote it by $\langle Z \rangle$.

DEFINITION: **Ensemble average**

The ensemble average of a random variable, Z, is the $N \to \infty$ limit of the finite-ensemble average,

$$\langle Z \rangle = \lim_{N\to\infty} \langle Z \rangle_N = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{2.10}$$ {eq:ens}

where z_i denotes the i^{th} realisation of Z.

The limit is not guaranteed to exist, in which case Z has no ensemble average, but finite-ensemble averages can always be computed. We will sometimes refer to the imagined infinite collection of realisations as the "ensemble" of Z.

Note that, unlike the expectation value, the ensemble average of a random variable is not defined in terms of probabilities. It is the quantity to which averages over realisations converge as the number of realisations grows. Recall that we followed this procedure when analysing the coin game. At each round of the gamble, we found finite-ensemble averages of simulated wealth for increasingly large samples, from one to one million, plotted against time in Fig. 1.2.

It is laborious to compute averages of ever-larger finite ensembles, in the hope of discerning their convergence to a limit. Fortunately, we can show, under the frequentist interpretation of probability introduced in Sec. 2.1, that the ensemble average of a random variable is equal to its expectation value.

Proof. Denote by n_j the number of times the value z_j is observed in N realisations of Z. The finite-ensemble average can be written as

$$\langle Z \rangle_N = \frac{1}{N} \sum_i z_i = \sum_j \frac{n_j}{N} z_j, \tag{2.11}$$

where the subscript i indexes a particular realisation of z and the subscript j indexes a possible value of z. Under the frequentist interpretation, the relative frequency of each possible value, n_j/N, converges almost surely in the limit $N \to \infty$ to its probability, p_j. Thus, we find

$$\lim_{N\to\infty} \langle Z \rangle_N = \sum_j p_j z_j = \mathrm{E}[Z]\,. \tag{2.12}$$

□

This result, commonly known as the *law of large numbers*, is exceedingly useful. It means that we no longer need large numbers of realisations of a random variable to compute its ensemble average. Instead, if we know the probability distribution, we can compute its expectation value straightforwardly as a weighted average or integral. It also means that, from now on, we can use ensemble average and expectation value interchangeably.

History: The invention of the expectation value

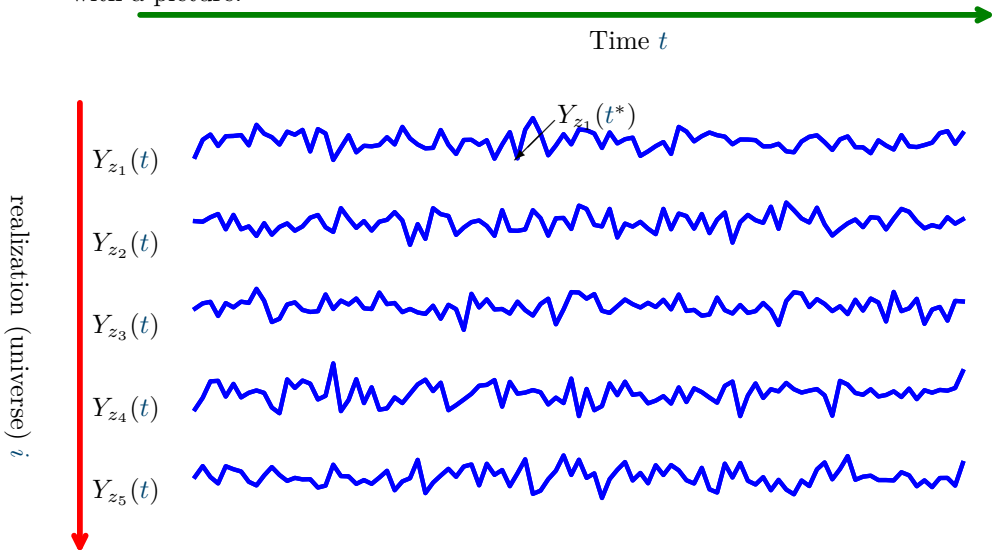
Suppressed.

2.4 Stochastic processes

Once again, we recommend van Kampen [**?**, p. 52] for a simple definition – this time of stochastic processes. Imagine we've defined a random variable, Z. Any function $Y(Z)$ is then also a random variable. A stochastic process is a special case of such a function, namely one that depends on an additional variable t, a simple scalar parameter, a number, which is interpreted as time, so we write

$$Y_Z(t) = f(Z, t). \tag{2.13}$$ {eq:st_pr}

This may not be how you think of a stochastic process, so let's illustrate this with a picture.



When we simulate stochastic processes, we often start with some value and modify it iteratively, for example in each step of a for-loop. In each step we generate a new instance of a random number and thereby construct the trajectory of the stochastic process. In (Eq. 2.13) it's not generated that way. Instead, in this picture, we generate an instance z of the random variable Z only once and insert that into (Eq. 2.13). The value z specifies a simple function of time

$$Y_z(t) = f(z, t) \tag{2.14}$$ {eq:st_pr_r}

meaning that all the randomness is contained in z. Once z is specified, $Y_z(t)$ is specified for all time, and we call it a "realization" or "trajectory" of the stochastic process. Note the use of capital Z for the random variable in (Eq. 2.13) and small z for a realization of it in (Eq. 2.14). As an example you can think of drawing at random a single uniformly distributed real number from the interval $(0, 1)$. With probability 1, this number will be irrational and correspond to an infinite sequence of random decimal digits, which can be interpreted as a stochastic process, where t is given by the decimal place of the digit.

We can also do this: fix a specific time, t^*, and consider the stochastic process at that time, $Y_Z(t^*)$. That's again a random variable, an instance of which may be $Y_{z_1}(t^*)$.

Just as a function of a random variable is another random variable, a function of a stochastic process is another stochastic process. We will often use the noun "observable" to refer to a quantity that is derived from a stochastic process. For example, the growth rate of wealth is an observable of the wealth process.

We will suppress the random variable Z in our notation, and just write $x(t)$ for the stochastic wealth process (instead of writing $x(Z,t)$, *c.f.* (Eq. 2.13)). We will also write $x(t)$ for a specific realization of this process, or $x_i(t)$ when it's important to distinguish different realizations.

2.5 Time averages

An observable that neatly captures the two different aspects of multiplicative growth we have illustrated is the exponential growth rate, $g_{\mathrm{m}}(\langle x(t)\rangle_N, \Delta t)$ observed over finite time Δt, in a finite ensemble of N realisations. Exponential growth rates are ubiquitous and may be familiar, but because they are the origin of the logarithmic function, which will be important for us later on, we will intro them properly in a little excursion that will also clarify what a logarithm is.

Excursion: Compounding growth, exponentials, and the logarithm

Suppressed.

The exponential growth rate of average wealth in an ensemble of N systems, observed over time Δt is

$$g_{\mathrm{m}}(\langle x(t)\rangle_N, \Delta t) = \frac{\Delta \ln \langle x\rangle_N}{\Delta t}, \tag{2.15}$$ {eq:gest}

where the Δ in the numerator corresponds to the change over the Δt in the denominator. For N and Δt finite this is a random variable. The relevant scalars arise as two different limits of the same stochastic object. The exponential growth rate of the expectation value (that's also $\frac{1}{\delta t}\ln\langle r\rangle$) is

$$g_{\mathrm{m}}(\langle x\rangle) = \lim_{N\to\infty} g_{\mathrm{m}}, \tag{2.16}$$

and the exponential growth rate followed by every trajectory when observed for a long time (that's also $\frac{1}{\delta t}\ln\overline{r}$) is

$$\overline{g} = \lim_{\Delta t\to\infty} g_{\mathrm{m}}. \tag{2.17}$$ {eq:gt}

We can also write (Eq. 2.15) as a sum of the logarithmic differences in the T individual rounds of the gamble that make up the time interval $\Delta t = T\delta t$

$$g_{\mathrm{m}}(\langle x(t)\rangle_N, \Delta t) = \frac{1}{T\delta t}\sum_{\tau=1}^{T} \Delta \ln \langle x(t+\tau\delta t)\rangle_N . \tag{2.18}$$

This leads us to a technical definition of the time average.

DEFINITION: **Finite-time average**

The "finite-time average" of the quantity $x(t)$ is

$$\overline{x}_{\Delta t} = \frac{1}{\Delta t}\int_t^{t+\Delta t} x(s)ds. \tag{2.19}$$ {eq:t_ave_f}

If x only changes at $T = \Delta t/\delta t$ discrete times $t+\delta t$, $t+2\delta t$, *etc.*, then this can be written as

$$\overline{x}_{\Delta t} = \frac{1}{T\delta t}\sum_{\tau=1}^{T} x(t+\tau\delta t). \tag{2.20}$$ {eq:t_ave_f_disc}

DEFINITION: **Time average**

The "time average" is the long-time limit of the finite-time average

$$\overline{x} = \lim_{\Delta t\to\infty} \overline{x}_{\Delta t}. \tag{2.21}$$ {eq:t_ave}

According to this definition, $\overline{g}$ is the time average of the observable $\frac{\delta \ln x}{\delta t}$. It can be shown that the time-average growth rate of a single trajectory is the same as that of a finite-ensemble average of trajectories, $\lim_{\Delta t\to\infty}\frac{\Delta \ln x}{\Delta t} = \lim_{\Delta t\to\infty}\frac{\Delta \ln\langle x\rangle_N}{\Delta t}$, [**?**]. In Sec. **??** we will derive this result as well as growth rates in finite ensembles and finite time.

Excursion: Dimensional analysis

Suppressed.

2.6 The game – revisited

We pretended to be mathematically clueless when we ran the simulations, with the purpose to gain a deeper conceptual understanding of the expectation value. We now compute exactly the expectation value of the stochastic process $x(t)$, instead of approximating it numerically. Consider the expectation value of (Eq. 1.2)

$$\langle x(t+\delta t)\rangle = \langle x(t)r(t+\delta t)\rangle\,. \tag{2.22}$$ {eq:step_1}

We've just learned what to call objects like $r(t)$: it's another stochastic process, or an observable. This one is especially simple: in a given realization $x(t)$ it's one instance of the same random variable for each time t. one, namely one that is ergodic. We note here that its ensemble average is time-independent (and in Sec. 2.7 we will see that it's an example of an ergodic observable). Since $r(t+\delta t)$ is independent of $x(t)$, (Eq. 2.22) can be re-written as

$$\langle x(t+\delta t)\rangle = \langle x(t)\rangle\,\langle r\rangle\,. \tag{2.23}$$

Therefore, we can solve recursively for the wealth after T rounds, corresponding to a playing time of $\Delta t = T\delta t$:

$$\langle x(t+\Delta t)\rangle = \langle x(t+T\delta t)\rangle = x(t)\,\langle r\rangle^T\,. \tag{2.24}$$

δt is the duration of a single round of a gamble, while Δt is the amount of time spent gambling.

The expectation value $\langle r \rangle$ is easily found from (Eq. 1.1) as $\langle r \rangle = \frac{1}{2} \times 0.6 + \frac{1}{2} \times 1.5 = 1.05$. Since this number is greater than one, $\langle x(t) \rangle$ grows exponentially in time by a factor 1.05 each time unit, or expressed as a continuous growth rate, at $\frac{1}{\delta t} \ln \langle r \rangle \approx 4.9\%$ per time unit. This is what might have led us to conclude that the gamble is worth taking. Figure 2.1 compares the analytical result for the infinite ensemble to the numerical results of Fig. 1.2 for finite ensembles.

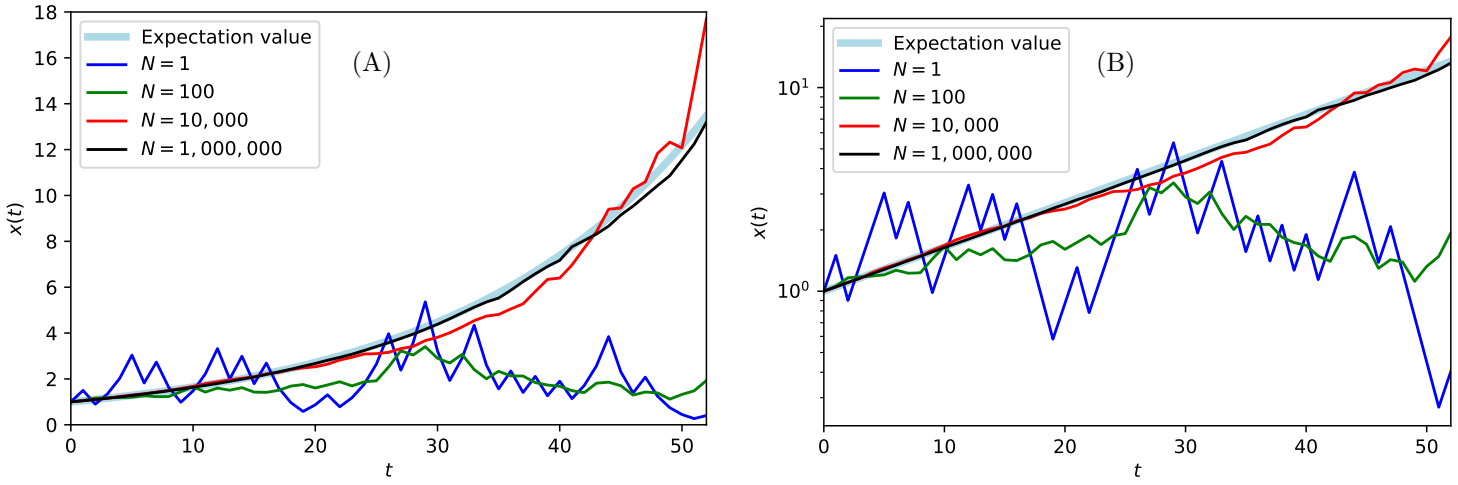


Figure 2.1: Expectation value (thick light blue line) and finite-ensemble averages. (A) linear scales, (B) logarithmic scales. {fig:cf_exp}

In this section we validate Fig. 1.3 and compute analytically what happens in the long-time limit. The blue line in Fig. 1.3 is not completely smooth, there's still some noise (see panel B). It has some average slope, but that slope will vary from realisation to realisation. The longer we observe the system, *i.e.* the more time is represented in a figure like Fig. 1.3, the smoother the line will be. In the long-time limit, $\Delta t \to \infty$, the line will be completely smooth, and the average slope will be a deterministic number – in any realization of the process it will come out identical.

The dynamic is set up such that wealth at time $t + \Delta t$, where $\Delta t = T\delta t$ as before, is

$$x(t + \Delta t) = x(t) \prod_{\tau=1}^{T} r(t + \tau \delta t), \tag{2.25}$$

with the dummy variable τ indicating the round of the game. We can split this into two products, one for each possible value of $r(t)$, which we call r_1 and r_2, *i.e.*

$$r(t) = \begin{cases} r_1 & \text{with probability } p_1 \\ r_2 & \text{with probability } p_2 = 1 - p_1. \end{cases} \tag{2.26}$$

Let's denote the number of occurrences of r_1 by n_1 and of r_2 by n_2, so that

$$x(t+\Delta t) = x(t) r_1^{n_1} r_2^{n_2}. \tag{2.27}$$

We denote by $\overline{r}$ the effective factor by which $x(t)$ is multiplied per round when the change is computed over a long time, *i.e.* $x(t+\Delta t) \sim x(t)(\overline{r})^T$ as $\Delta t \to \infty$. This quantity is found by taking the T^{th} root of $\frac{x(t+\Delta t)}{x(t)}$ and considering its long-time limit:

$$\overline{r} = \lim_{\Delta t\to\infty} \left(\frac{x(t+\Delta t)}{x(t)}\right)^{1/T} \tag{2.28}$$

$$= \lim_{T\to\infty} r_1^{n_1/T} r_2^{n_2/T}. \tag{2.29}$$

Identifying $\lim_{T\to\infty} n_1/T$ as the probability p_1 for r_1 to occur (and similarly $\lim_{T\to\infty} n_2/T = p_2$) this is

$$\lim_{T\to\infty} \left(\frac{x(t+T\delta t)}{x(t)}\right)^{1/T} = (r_1 r_2)^{1/2}, \tag{2.30}$$ {eq:long_t}

or $\sqrt{0.9} \approx 0.95$, *i.e.* a number smaller than one, reflecting decay in the long-time limit for the individual trajectory. The trajectory in Fig. 1.3 was not a fluke: *every* trajectory will decay in the long run at a rate of $(r_1 r_2)^{1/2}$ per round.

Figure 2.2 (B) compares the trajectory generated in Fig. 1.3 to a trajectory decaying exactly at rate $\overline{r}$ and places it next to the average over a million systems.

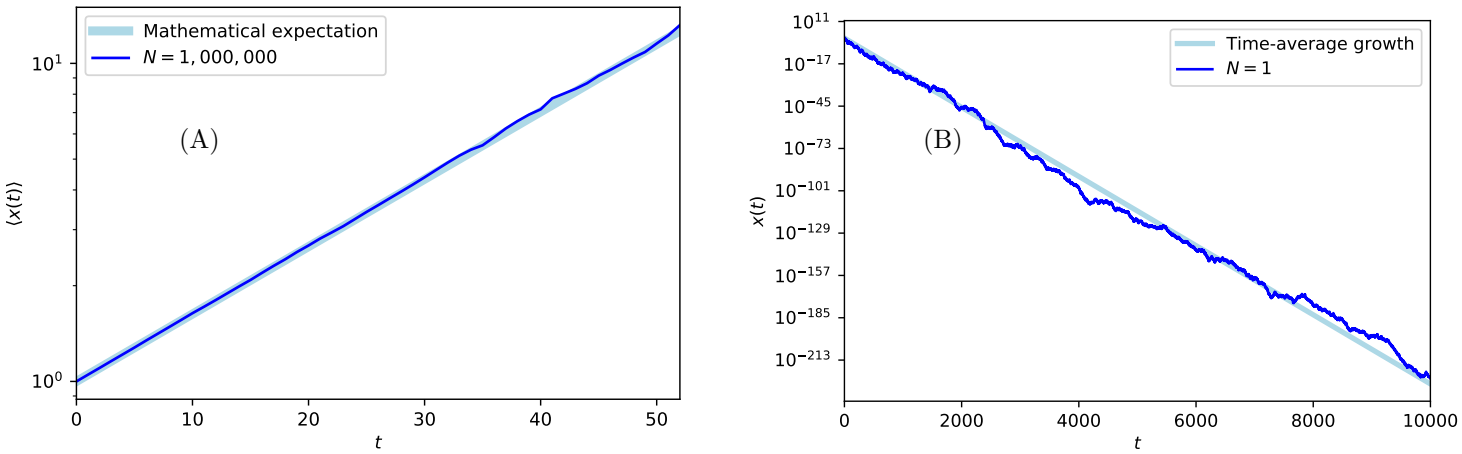


Figure 2.2: (A) Finite-ensemble average for $N = 10^6$ and 52 time steps, the light blue line is the expectation value. (B) A single system simulated for 10,000 time steps, the light blue line decays exponentially with the time-average decay factor $\overline{r}$ in each time step. {fig:1_4}

Excursion: Scalars

Suppressed.

There are two averages, $r_{\langle\rangle}$ and $\overline{r}$ that we have determined numerically and analytically. Neither average is "wrong" in itself; instead each average corresponds to a different property of the system. Each average is the answer to a different question. Saying that "wealth goes up, on average" is clearly meaningless and should be countered with the question "on what type of average?"

History: William Allen Whitworth

Suppressed.

2.7 Ergodicity

{section:Ergodic_observab

We have encountered two types of averaging – the ensemble average and the time average. In our case – assessing whether it will be good for you to play our game, the time average is the interesting quantity because it tells you what happens to your wealth as time passes. The ensemble average is irrelevant because you do not live your life as an ensemble of many yous who can average over their wealths. Whether you like it or not, you will experience yourself owning your own wealth at future times; whether you like it not, you will never experience yourself owning the wealth of a different realization of yourself. The different realizations, and therefore the expectation value, are fiction, fantasy, imagined.

We are fully aware that it can be counter-intuitive that with probability one, a different rate is observed for the expectation value than for any trajectory over time. It sounds strange that the expectation value is completely irrelevant to the problem. A reason for the intuitive discomfort is history: since the 1650s we have been trained to compute expectation values, with the implicit belief that they will reflect what happens over time. It may be helpful to point out that all of this trouble has a name that's well-known to certain people, and that an entire field of mathematics is devoted to dealing with precisely this problem. The field of mathematics is called "ergodic theory." It emerged from the question under what circumstances the expectation value is informative of what happens over time, first raised in the development of statistical mechanics by Maxwell and Boltzmann starting in the 1850s. These lecture notes are our attempt to use precisely the insights of these physicists to re-develop economic theory from the foundations up.

History: Randomness and ergodicity in physics

Suppressed.

To convey concisely that we cannot use the expectation value and the time average interchangeably in our game, we would say "the observable x is not ergodic."

DEFINITION: **Ergodic property**

In these notes, an observable A is called ergodic if its expectation value is constant in time and its time average converges to this value with probability one[a]

$$\lim_{\Delta t\to\infty}\frac{1}{\Delta t}\int_t^{t+\Delta t}A(s)ds=\lim_{N\to\infty}\frac{1}{N}\sum_i^N A_i(t). \tag{2.31}$$ {eq:def_ergodic}

[a]Some researchers would call A "mean ergodic" and require further observables derived from it to be (mean) ergodic in order to call A "wide-sense ergodic." This extra nomenclature is not necessary for our work, but we leave a footnote here to avoid confusion.

The right-hand side (RHS) of (Eq. 2.31) is evaluated at time t, and unlike the left-hand side (LHS) could be a function of time. For now, we restrict our definition of ergodicity to a setup where that is not the case, *i.e.* where the ergodic property holds at all times. In Sec. **??** we will discuss transient behavior, where the distribution of A is time dependent. We then also consider an observable "ergodic" if its expectation value only converges to the time-average in the $t\to\infty$ limit.

In terms of random variables, Z, and stochastic processes, $Y_Z(t)$, the ergodic property can be visualized as in Fig. 2.3. Averaging a stochastic process over time or over the ensemble are completely different operations, and only under very rare circumstances (namely under ergodicity) can the two operations be interchanged. In our coin-tossing game the operations are clearly not interchangeable. An implicit assumption of interchangeability in the early days is the Original Sin of economic theory.

We stress that in a given setup, some observables may have the ergodic property even if others do not. Language therefore must be used carefully. Saying our game is non-ergodic really means that some key observables of interest, most notably wealth x, are not ergodic. Wealth $x(t)$, defined by (Eq. 1.1), is clearly not ergodic – with $A=x$ the LHS of (Eq. 2.31) is zero, and the RHS is not constant in time but grows. The expectation value $\langle x\rangle(t)$ simply doesn't give us the relevant information about the temporal behavior of $x(t)$.

This does not mean that no ergodic observables exist that are related to x. Such observables do exist, and we have already encountered two of them. In fact, we will encounter a particular type of them frequently – in our quest for an observable that tells us what happens over time in a stochastic system we will find them automatically. However, again, the issue is subtle: an ergodic observable may or may not tell us what we're interested in. It may be ergodic but not indicate what happens to x. For example, the multiplicative factor $r(t)$ is an ergodic observable that reflects what happens to the expectation value of x, whereas per-round changes in the logarithm of wealth, $\delta\ln x=\ln r$, are also ergodic and reflect what happens to x over time.

Proposition: $r(t)$ and $\delta\ln x$ are ergodic for the wealth dynamic defined by (Eq. 1.1) and (Eq. 1.2).

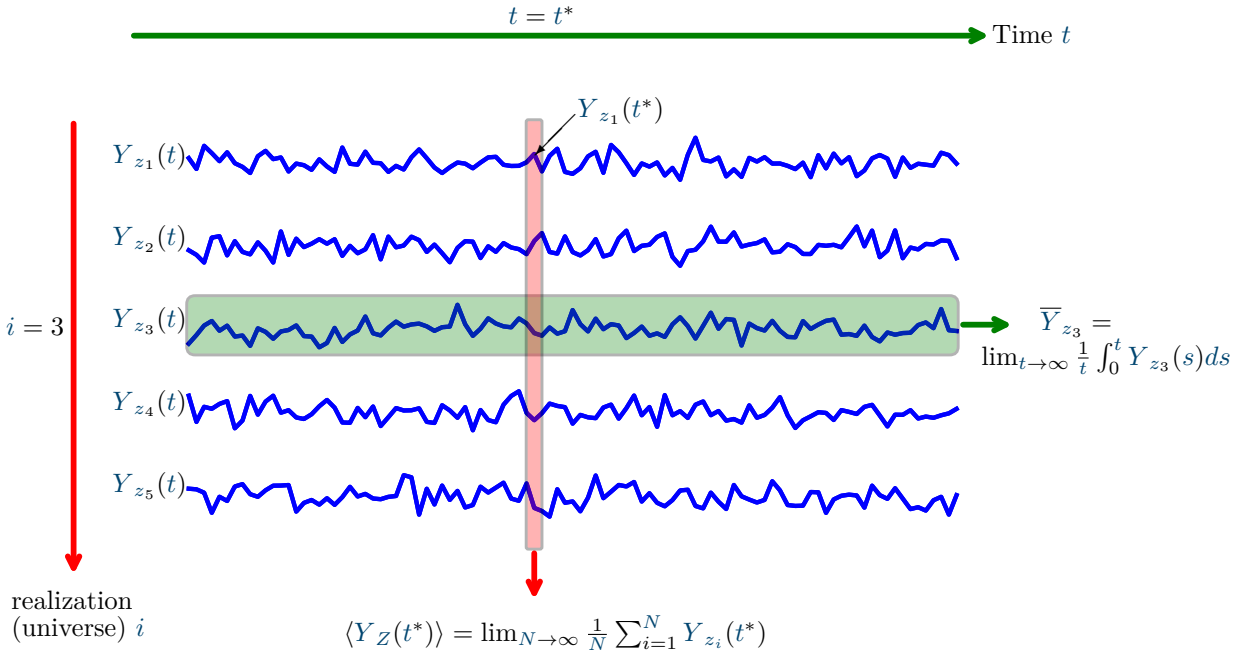


Figure 2.3: Extending the figure on p. 16, averaging over time means averaging along one trajectory from left to right; averaging over the ensemble means averaging at a fixed time across different trajectories from top to bottom. {fig:ergodic_grid}

Proof. According to (Eq. 2.10) and (Eq. 2.9), the expectation value of $r(t)$ is

$$\langle r\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_i^N r_i, \tag{2.32}$$ {eq:e_r}

and, according to (Eq. 2.20), the time average of $r(t)$ is

$$\overline{r} = \lim_{T\to\infty} \frac{1}{T}\sum_\tau^T r_\tau, \tag{2.33}$$ {eq:t_r}

where we have written $r_\tau = r(t+\tau\delta t)$ to make clear the equivalence between the two expressions. The only difference is between the labels we have chosen for the dummy variable (i in (Eq. 2.32) and τ in (Eq. 2.33)). Clearly, the expressions yield the same value.

The same argument holds for $\delta \ln x$. □

Whether we consider (Eq. 2.33) an average over time or over an ensemble is only a matter of our choice of words.

The expectation value $\langle \delta \ln x\rangle$ is important, historically. Daniel Bernoulli noticed in 1738 [**?**] that people tend to optimize $\langle \delta \ln x\rangle$, whereas it had been assumed that they should optimize $\langle \delta x\rangle$. Unaware of the issue of ergodicity (200 years before the concept was discovered and the word was coined), Bernoulli had no good explanation for this empirical fact and simply stated that people tend

to behave as though they valued money non-linearly. We now know what is actually going on: multiplicative dynamics are a fairly realistic model for real wealth, and under those dynamics δx is not ergodic, and $\langle \delta x \rangle$ is of no interest – it doesn't tell us what happens over time. However, $\delta \ln x$ *is* ergodic, and $\langle \delta \ln x \rangle$ does tell us what happens to x over time, wherefore seeing people optimise $\langle \delta \ln x \rangle$ just means seeing them optimise wealth over the one trajectory that describes a financial life, rather than across the ensemble of possibilities.

Ergodicity is not the same concept as stationarity. As an illustration of the difference, consider the following process: $f(t) = z_i$, where z_i is an instance of a random variable Z. Explicitly, this means a realisation of the stochastic process $f(t)$ is generated as follows: we generate the random instance z_i once, and then fix $f(t)$ at that value for all time. The distribution of $f(t)$ is independent of t and in that sense $f(t)$ is stationary. But it is not ergodic: averaging over the ensemble, we obtain $\langle f(t) \rangle = \langle z \rangle$, whereas averaging over time in the i^{th} trajectory gives $\overline{f} = z_i$. Thus the process is stationary but not ergodic.

2.8 Changes and stability

{section:Rates}

Deleted – the material now appears in the next chapter. However, we must say enough about rates to consider the growth rates of Brownian motion (BM) and geometric Brownian motion (GBM).

2.9 Normal distribution

{section:Normal_distribut

2.10 Brownian motion

{section:Brownian_motion}

We motivate the model called BM as a limiting process, the continuous-time limit, that arises from random walks. In the previous section we established that the discrete increments of the logarithm of x, which we called v, are instances of a time-independent random variable in our game. A quantity making such random steps over time is said to perform a "random walk." Indeed, the blue line for a single system in Fig. 1.2 (B) shows 52 steps of a random walk trajectory. Random walks come in many forms – in all of them v changes discontinuously by an amount δv drawn from a time-independent distribution, over time intervals which may be regular or which may be drawn from a time-independent distribution themselves.

We are interested only in the simple case where v changes at regular intervals, $\delta t, 2\delta t, \ldots$. For the distribution of increments we only insist on the existence of the variance, meaning we insist that $\mathrm{var}(\delta v) = \langle \delta v^2 \rangle - \langle \delta v \rangle^2$ be finite. Increments whose distributions are heavier-tailed do not lead to BM (BM has continuous paths, and that continuity is broken by such increments).

The change in v after a long time is the sum of many independent increments,

$$v(t + T\delta t) - v(t) = \sum_i^T \delta v_i. \tag{2.34}$$

The Gaussian central limit theorem tells us that such a sum will become Gaussian-distributed as we add more terms to the sum and re-scale it appropriately,

namely so as to keep the width of the distribution finite and remove any systematic drift,

$$\lim_{T\to\infty} \overbrace{\frac{1}{\sqrt{T}}}^{\text{keep width finite}} \sum_{i}^{T} (\delta v_i \underbrace{- \langle \delta v \rangle}_{\text{remove systematic drift}}) \sim \mathcal{N}(0, \text{var}(\delta v)), \tag{2.35}$$ {eq:CLT}

where we call $\frac{\langle \delta v \rangle}{\delta t}$ the "drift term." The notation $\sim \mathcal{N}(0, \text{var}(\delta v))$ is short-hand for "is Gaussian distributed, with mean 0 and variance $\text{var}(\delta v)$." The logarithmic change in the long-time limit that was of interest to us in the analysis of the coin toss game is thus Gaussian distributed.

Let's also ask about the re-scaling that was applied in (Eq. 2.35). Scaling properties are very robust, and especially the scaling of random walks for long times will be useful to us.

We work with the simplest setup: at time zero we start at zero, $v(0) = 0$, and in each time step, we either increase or decrease v by 1, with probability 1/2. To avoid notation clutter, we'll set the duration of a time step to $\delta t = 1$, so that T is both the number of steps and the time.

We are interested in the variance of the distribution of v as T increases, which we obtain by computing the first and second moments of the distribution.

The first moment (the expectation value) of v is $\langle v \rangle (T) = 0$, by symmetry for all times.

We obtain the second moment by induction[2]: Whatever the second moment, $\langle v(T)^2 \rangle$, is at time T, we can write down its value at time $T + 1$ as

$$\begin{aligned} \langle v(T+1)^2 \rangle &= \frac{1}{2} \left[\langle (v(T)+1)^2 \rangle + \langle (v(T)-1)^2 \rangle \right] && (2.36) \\ &= \frac{1}{2} \left[\langle v(T)^2 + 1 + 2v(T) \rangle + \langle (v(T)^2 + 1 - 2v(T)) \rangle \right] && (2.37) \\ &= \langle v(T)^2 \rangle + 1. && (2.38) \end{aligned}$$

In addition, we know the initial value of $v(0) = 0$. By induction it follows that the second moment is

$$\langle v(T)^2 \rangle = T \tag{2.39}$$

and, since the first moment is zero, the variance is

$$\text{var}(v(T)) = T. \tag{2.40}$$ {eq:BM_var}

The standard deviation – the width of the distribution – of changes in a quantity following a random walk thus scales as the square-root of the number of steps that have been taken, $\sqrt{T}$.

This square-root behaviour leads to many interesting results. It can make averages stable (because $\sqrt{T}/T$ converges to zero for large T), and sums unstable (because $\sqrt{T}$ diverges for large T). Consequently, we may expect that as the size of some system increases, some properties become stable and others unstable.

Imagine simulating a single long trajectory of v and plotting it on paper[3].

[2] The argument is nicely illustrated in [**?**, Volume 1, Chapter 6-4], where we first came across it.

[3] This argument is inspired by a colloquium presented by Wendelin Werner in the mathematics department of Imperial College London in January 2012. Werner started the colloquium with a slide that showed a straight horizontal line and asked: what is this? Then answered that it was the trajectory of a random walk, with the vertical and horizontal axes scaled equally.

The amount of time that has to be represented by a fixed length of paper increases linearly with the simulated time because the paper has a finite width to accommodate the horizontal axis. If $\langle \delta v \rangle \neq 0$ then the amount of variation in v that has to be represented by a fixed amount of paper also increases linearly with the simulated time. However, the departures of Δv from its expectation value $T \langle \delta v \rangle$ only increase as the square-root of T. Thus, the amount of paper-space given to these departures scales as $T^{-1/2}$, and for very long simulated times the trajectory will look like a straight line on paper.

In an intermediate regime, fluctuations will still be visible but they will also be approximately Gaussian distributed. In this regime it is often easier to replace the random walk model with the corresponding continuous process. That process – finally – is BM.

We think of BM as the limit of a random walk where we shorten the duration of a step $\delta t \to 0$, and scale the width of an individual step so as to maintain the random-walk scaling of the variance, meaning $|\delta v| = \sqrt{\delta t}$. In the limit $\delta t \to 0$, this implies that the local slope of a BM trajectory diverges, $\frac{\delta v}{\delta t} \to \infty$. This means that BM trajectories are infinitely jagged, or – in mathematical terms – they are not differentiable. However, the way in which they become non-differentiable, through the $\sqrt{\delta t}$ factor, just leaves the trajectories continuous (this isn't the case for $|\delta v| = \delta t^{\alpha}$, where α is less than 0.5).

Continuity of v means that it is possible to make the difference $|v(t) - v(t + \epsilon)|$ arbitrarily small by choosing ϵ sufficiently small. Trajectories (of non-BM processes) that don't have this property contain what are appropriately called "jumps." Continuity therefore means that there are no jumps. These subtleties make BM a topic of great mathematical interest, and many books have been written about it. We will pick from these books only what is immediately useful to us. To convey the universality of BM we define it formally as follows:

DEFINITION: **Brownian motion i**

If a stochastic process has continuous paths, stationary independent increments, and is distributed according to $\mathcal{N}(\mu t, \sigma^2 t)$ then it is a Brownian motion.

The process can be defined in different ways. Another illuminating definition is this:

DEFINITION: **Brownian motion ii**

If a stochastic process is continuous, with stationary independent increments, then the process is a Brownian motion.

We quote from [**?**]: *"This beautiful theorem shows that Brownian motion can actually be defined by stationary independent increments and path continuity alone, with normality following as a consequence of these assumptions. This may do more than any other characterization to explain the significance of Brownian motion for probabilistic modeling."*

Indeed, BM is not just a mathematically rich model but also – due to its emergence through the Gaussian central limit theorem – a model that represents a large universality class, *i.e.* it is a good description of what happens over long times in many other models that produce random trajectories.

The power of BM lies in its simplicity and analytic tractability, involving only two parameters, μ and σ. We will often work with its representation as a stochastic differential equation (SDE)

$$dv = \mu dt + \sigma dW \tag{2.41}$$ {eq:BM_dx}

where dW is the so-called "Wiener increment," the beating heart of many SDEs. The Wiener increment can be defined by two properties: its distribution and its auto-correlation,

$$dW \sim \mathcal{N}(0, dt) \tag{2.42}$$
$$\langle dW(t) dW(t') \rangle = dt\ \delta(t, t'), \tag{2.43}$$

where $\delta(t, t')$ is the Kronecker delta – zero if its two arguments differ ($t \neq t'$), and one if they are identical ($t = t'$).[4] In simulations BM paths can be constructed from a discretized version of (Eq. 4.4)

$$v(t + \delta t) = v(t) + \mu \delta t + \sigma \sqrt{\delta t} \xi_t, \tag{2.44}$$ {eq:BM_d}

where ξ_t are instances of a standard normal distribution ($\mathcal{N}(0, 1)$).

BM itself is not a time-independent random variable – it is a non-ergodic stochastic process. This is easily seen by comparing expectation value and time average. We start with the expressions (stated without proof here) for the finite-ensemble average and the finite-time average of BM. The finite-ensemble average (easy to derive) is distributed as

$$\langle v \rangle_N \sim \mu t + \mathcal{N}(0, t/N), \tag{2.45}$$ {eq:fin_ens_BM}

and the finite-time average (a little harder to derive) of a single BM trajectory is distributed as

$$\bar{v}_t \sim \mu t/2 + \sigma \mathcal{N}(0, t/3). \tag{2.46}$$ {eq:fin_tim_BM}

The expectation value, *i.e.* the limit $N \to \infty$ of (Eq. 2.45), converges to μt with probability one, so it depends on time, and it's unclear how to compare that to a time average (which cannot depend on time). Its limit $t \to \infty$ does not exist.

The time average, the limit $t \to \infty$ of (Eq. 2.46) diverges unless $\mu = 0$, but even with $\mu = 0$ the limit is a random variable with diverging variance – something whose density is zero everywhere. In no meaningful sense do the two expressions converge to the same scalar in the relevant limits.

Clearly, BM, whose increments are ergodic, is itself not ergodic. However, that doesn't make it unmanageable or unpredictable – we know the distribution of BM at any moment in time. But the non-ergodicity has surprising consequences of which we mention one now. We already mentioned that if we plot a Brownian trajectory with non-zero drift on a piece of paper it will turn into a straight line for long enough simulation times. This suggests that the randomness of a Brownian trajectory becomes irrelevant under a very natural rescaling.

[4] Physicists often write $dW = \eta dt$, where $\langle \eta \rangle = 0$ and $\langle \eta(t) \eta(t') \rangle = \delta(t - t')$, in which case $\delta(t - t')$ is the Dirac delta function, defined by the integral $\int_{-\infty}^{\infty} f(t) \delta(t - t') dt = f(t')$. Because of its singular nature ($\eta(t)$ does not exist ("is infinite"), only its integral exists) it can be difficult to develop an intuition for this object, and we prefer the dW notation.

Inspired by this insight let's hazard a guess as to what the time-average of zero-drift BM might be.

The simplest form of zero-drift BM starts at zero, $v(0) = 0$ and has variance $\mathrm{var}(v(t)) = t$ (this process is also known as the "Wiener process"). The process is known to be recurrent – it returns to zero, arbitrarily many times, with probability one in the long-time limit. We would not be mad to guess that the time average of zero-drift BM,

$$\bar{v} = \lim_{t\to\infty} \frac{1}{t} \int_0^t dt' v(t'), \tag{2.47}$$

will converge to zero with probability one. But we would be wrong. Yes, the process has no drift, and yes it returns to zero infinitely many times, but its time average is not a delta function at zero. It is, instead normally distributed with infinite variance according to the following limit

$$\bar{v} \sim \lim_{t\to\infty} \mathcal{N}(0, t/3). \tag{2.48}$$

Averaging over time, in this case, does not remove the randomness. A sample trajectory of the finite-time average (not of BM but of the average over a BM) is shown in Fig. 2.4. In the literature this process, $\frac{1}{t}\int_0^t dt' v(t')$, is known as the

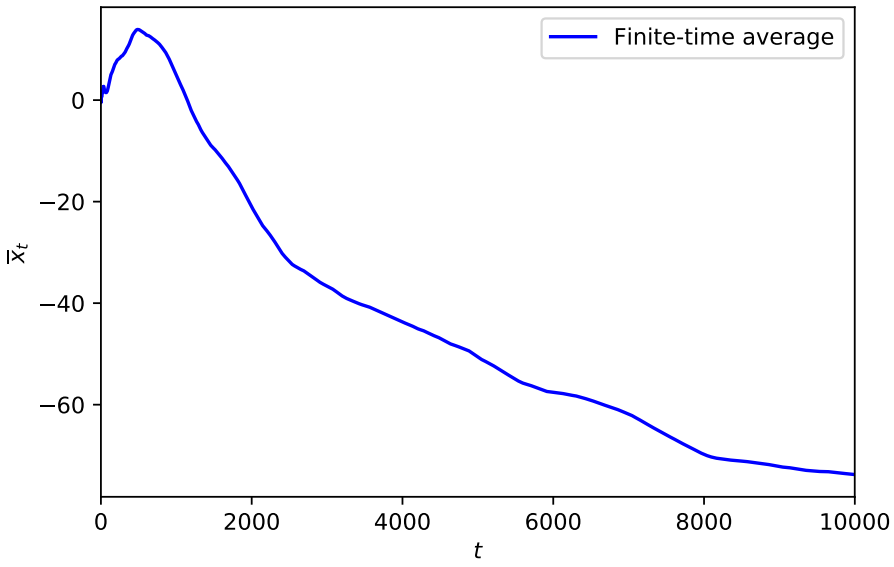


Figure 2.4: Trajectory of the finite-time average of a zero-drift BM. The process is not ergodic: the time average does not converge to a number, but is instead distributed according to $\mathcal{N}(0, t/3)$ for all times, while the expectation value is zero. It is the result of integrating a BM; integration is a smoothing operation, and as a consequence the trajectories are smoother than BM (unlike a BM trajectory, they are differentiable). {fig:1_6}

2.11 Geometric Brownian motion {section:Geometric_Browni

DEFINITION: **Geometric Brownian motion**

If the logarithm of a quantity performs Brownian motion, the quantity itself performs "geometric Brownian motion."

While in Sec. 2.10 $v(x) = \ln(x)$ performed BM, x itself performed GBM. The change of variable from x to $v(x) = \ln(x)$ is trivial in a sense but it has interesting consequences. It implies, for instance, that

- $x(t)$ is log-normally distributed
- increments in x are neither stationary nor independent
- $x(t)$ cannot become negative
- the most likely value of x (the mode) does not coincide with the expectation value of x.

These and other properties of the log-normal distribution will be discussed in detail in Sec. **??**.

Again, it is informative to write GBM as a stochastic differential equation.

$$dx = x(\mu dt + \sigma dW). \tag{2.49}$$ {eq:GBM_c}

Similarly to BM, trajectories for GBM can be simulated using the discretized form (*cf.* (Eq. 2.44))

$$\delta x = x(\mu \delta t + \sigma\sqrt{\delta t}\xi_t), \tag{2.50}$$ {eq:GBM_d}

where $\xi_t \sim \mathcal{N}(0,1)$ are instances of a standard normal variable. In such simulations we must pay attention that the discretization does not lead to negative values of x. This happens if the expression in brackets in (Eq. 2.50) is smaller than -1 (in which case x changes negatively by more than itself). To avoid negative values we must have $\mu\delta t + \sigma\sqrt{\delta t}\xi_t > -1$, or $\xi_t < \frac{1+\mu\delta t}{\sigma\sqrt{\delta t}}$. As δt becomes large it becomes more likely for ξ_t to exceed this value, in which case the simulation fails. But ξ_t is Gaussian distributed, meaning it has thin tails, and choosing a sufficiently small value of δt makes these failures essentially impossible.

GBM on logarithmic vertical scales looks like BM on linear vertical scales. Figure 1.2 is, in fact, and example of a very coarse discretisation of GBM. But it's useful to look at a more finely discretised trajectory of GBM on linear scales to develop an intuition for this important process.

The basic message of the game from Sec. 1.1 is that we may obtain different values for growth rates, depending on how we average – an expectation value is one average, a time average is quite another. The game itself is sometimes called the multiplicative binomial process [**?**], we thank S. Redner for pointing this out to us. GBM is the continuous version of the multiplicative binomial process, and it shares the basic feature of a difference between the growth rate of the expectation value and time-average growth.

The expectation value is easily computed – the process is not ergodic, but that does not mean we cannot compute its expectation value. We simply take the expectations values of both sides of (Eq. 2.49) to get

$$\langle dx \rangle = \langle x(\mu dt + \sigma dW) \rangle \tag{2.51}$$

$$= d\langle x \rangle = \langle x \rangle \mu dt. \tag{2.52}$$

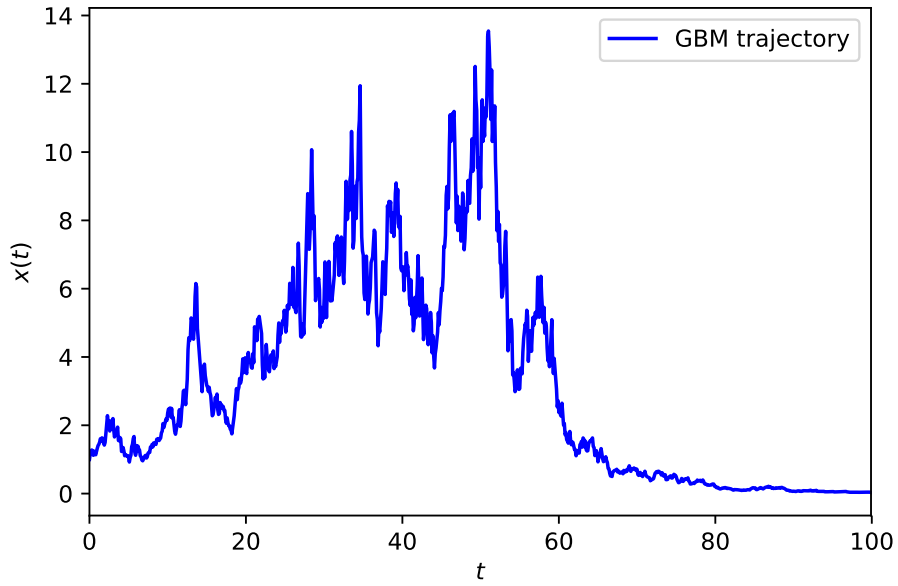


Figure 2.5: Trajectory of a GBM. What happens to the trajectory tomorrow depends strongly on where it is today – for instance, unlike for BM, it is difficult to recover from a low value of x, and trajectories are likely to get stuck near zero. Occasional excursions are characterised by large fluctuations. Parameters are $\mu = 0.05$ per time unit and $\sigma = \sqrt{2\mu}$, corresponding to zero growth rate in the long run. It would be easy to invent a story to go with this (completely random) trajectory – perhaps something like "things were going well in the beginning but then a massive crash occurred that destroyed morale." {fig:1_7}

This differential equation has the solution

$$\langle x(t)\rangle = x(t_0)\exp(\mu t), \tag{2.53}$$

which determines the growth rate of the expectation value as

$$g_{\mathrm{m}}(\langle x\rangle) = \mu. \tag{2.54}$$ {eq:expectation_g}

As we know, this growth rate is different from the growth rate that materializes with probability 1 in the long run. Computing the time-average growth rate is only slightly more complicated, and it will get even simpler once we've introduced Itô calculus in Sec. 2.12. But for now we will follow this plan: consider the discrete process (Eq. 2.50) and compute the changes in the logarithm of x, then we will let δt become infinitesimal and arrive at the result for the continuous process. We know $\delta \ln(x(t))$ to be ergodic and reflective of performance over time, wherefore we will proceed to take its expectation value to compute the time average of the exponential growth rate of the process.

The change in the logarithm of x in a time interval δt is

$$\begin{aligned}
\ln x(t+\delta t) - \ln x(t) &= \ln[x(1+\mu\delta t + \sigma\sqrt{\delta t}\xi_t)] - \ln x(t) && (2.55)\\
&= \ln x + \ln(1+\mu\delta t + \sigma\sqrt{\delta t}\xi_t) - \ln x(t) && (2.56)\\
&= \ln(1+\mu\delta t + \sigma\sqrt{\delta t}\xi_t), && (2.57)
\end{aligned}$$

which we Taylor-expand as ln(1+something small) because we will let δt become small. Expanding to second order,

$$\ln x(t+\delta t) - \ln x(t) = \mu\delta t + \sigma\sqrt{\delta t}\xi_t - \frac{1}{2}\left(\mu\sigma\delta t^{3/2}\xi_t + \sigma^2\delta t\xi_t^2\right) + o(\delta t^2), \quad (2.58)$$

using "little-o notation" to denote terms that are of order δt^2 or smaller. Finally, because $\delta \ln x(t)$ is ergodic, by taking the expectation value of this equation we find the time average of $\delta \ln x(t)$

$$\langle \ln x(t+\delta t) - \ln x(t)\rangle = \mu\delta t - \frac{1}{2}\left(\mu^2\delta t^2 + \sigma^2\delta t\right) + o(\delta t^2). \quad (2.59)$$

Letting δt become infinitesimal the higher-order terms in δt vanish, and we find

$$\langle \ln x(t+dt) - \ln x(t)\rangle = \mu dt - \frac{1}{2}\sigma^2 dt, \quad (2.60)$$

so that the time-average growth rate is

$$\overline{g} = \frac{d\langle \ln x\rangle}{dt} = \mu - \frac{1}{2}\sigma^2. \quad (2.61) \quad \texttt{\{eq:time\_g\}}$$

The non-ergodicity of GBM leads to a difference between the behaviour of the expectation value (which grows at $g_{\rm m}(\langle x\rangle)$) and the long-time behaviour of any given trajectory (which grows at $\overline{g}$). Because people experience their wealth over time (which may be described by GBM) and have not access to the ensemble of other possible trajectories, they quite reasonably behave closer to optimising $g_{\rm m}(\langle x\rangle)$ than to $\overline{g}$.

We could have guessed the result by combining Whitworth's argument on the disadvantage of gambling with the scaling of BM. Let's re-write the factor $1-\epsilon$ in (Eq. **??**) as $1-\sigma\sqrt{\delta t}$. According to the scaling of the variance in a random walk, (Eq. 2.40), this would be a good coarse-graining of some faster process (with shorter time step) underlying Whitworth's game. To find out what happens over one single time step we take the square root of (Eq. **??**),

$$[(1+\sigma\sqrt{\delta t})(1-\sigma\sqrt{\delta t})]^{1/2} = [1-\sigma^2\delta t]^{1/2}. \quad (2.62)$$

Letting δt become infinitesimally small, we replace δt by dt, and the first-order term of a Taylor-expansion becomes exact,

$$[(1+\sigma\sqrt{\delta t})(1-\sigma\sqrt{\delta t})]^{1/2} \to 1 - \frac{\sigma^2}{2}dt, \quad (2.63)$$

in agreement with (Eq. 2.61) if the drift term $\mu = 0$, as assumed by Whitworth.

2.12 Itô calculus

{section:Ito}

We have chosen to work with the discrete process here and have arrived at a result that is more commonly shown using Itô's formula. We will not discuss Itô calculus in depth but we will use some of its results. The key insight of Itô was that the non-differentiability of so-called Itô processes leads to a new form

of calculus, where in particular the chain rule of ordinary calculus is replaced. An Itô process is a SDE of the following form

$$dx = a(x,t)dt + b(x,t)dW. \tag{2.64}$$ {eq:Ito_process}

If we are interested in the behaviour of some other quantity that is a function of x, let's say $v(x)$, then Itô's formula tells us how to derive the relevant SDE as follows:

$$dv = \left(\frac{\partial v}{\partial t} + a(x,t)\frac{\partial v}{\partial x} + \frac{b(x,t)^2}{2}\frac{\partial^2 v}{\partial x^2} \right) dt + b(x,t)\frac{\partial v}{\partial x} dW. \tag{2.65}$$ {eq:Ito}

Derivations of this formula can be found on Wikipedia. Intuitive derivations, such as [**?**], use the scaling of the variance, (Eq. 2.40), and more formal derivations, along the lines of [**?**], rely on integrals. We simply accept (Eq. 2.65) as given. It makes it very easy to re-derive (Eq. 2.61), which we leave as an exercise: use (Eq. 2.65) to find the SDE for $\ln(x)$, take its expectation value and differentiate with respect to t. We will use (Eq. 2.65) in Sec. **??**. The above computations are intended to give the reader intuitive confidence that Itô calculus can be trusted[5]. We find that, though phrased in different words, our key insight – that *the growth rate of the expectation value is not the time-average growth rate* – has appeared in the literature not only in 1870 but also in 1944. And in 1956 [**?**], and in 1966 [**?**], and in 1991 [**?**], and at many other times. Yet the depth of this insight remained unprobed.

Equation (2.61), which agrees with Itô calculus, may be surprising. Consider the case of no noise $dx = x\mu dt$. Here we can identify $\mu = \frac{1}{x}\frac{dx}{dt}$ as the infinitesimal increment in the logarithm, $\frac{d\ln(x)}{dt}$, using the chain rule of ordinary calculus. A naïve application of the chain rule to (Eq. 2.49) would therefore also yield $\frac{d\langle\ln(x)\rangle}{dx} = \mu$, but the fluctuations in GBM have a non-linear effect, and it turns out that the usual chain rule does not apply. Itô calculus is a modified chain rule, (Eq. 2.65), which leads to the difference $-\frac{\sigma^2}{2}$ between the expectation-value growth rate and the time-average growth rate.

This difference is sometimes called the "spurious drift", but at the London Mathematical Laboratory (LML) we call it the "Weltschmerz" because it is the difference between the many worlds of our dreams and fantasies, and the one cruel reality that the passage of time imposes on us.

Summary of Chap. ??

Suppressed.

[5] Itô calculus is one way of interpreting the non-differentiability of dW. Another interpretation is due to Stratonovich, which is not strictly equivalent. However, the key property of GBM that we make extensive use of is the difference between the growth rate of the expectation value, $g_{\rm m}(\langle x\rangle)$, and the time-average growth rate, $\overline{g}$. This difference is the same in the Stratonovich and the Itô interpretation, and all our results hold in both cases.

Part II

Microeconomics

Chapter 3

Decisions in a riskless world {chapter:Riskless}

Decision theory is a cornerstone of formal economics. As the name suggests, it models how people make decisions. In this chapter we will generalise and formalise the treatment of the coin tossing game to introduce our approach to decision theory. Our central axiom will be that people attempt to maximize the rate at which wealth grows when averaged over time. This is a surprisingly powerful idea. In many cases it eliminates the need for well established but epistemologically troublesome techniques, such as utility functions.

3.1 Models and science fiction

{section:Models_and}

We will do decision theory by using mathematical models, and since this can be done in many ways we will be explicit about how we choose to do it. We will define a wealth process – a model of how wealth changes with time – and a decision criterion. The wealth process and the decision criterion may or may not remind you of the real world. We will not worry too much about the accuracy of these reminiscences. Instead we will "shut up and calculate" – we will let the mathematical model create its world. Writing down a mathematical model is like laying out the premise for a science-fiction novel. We may decide that people can download their consciousness onto a computer, that medicine has advanced to eliminate ageing and death – these are premises we are at liberty to invent. Once we have written them down we begin to explore the world that results from those premises. A decision criterion is really a model of human behaviour – what makes us who we are if not our decisions? It therefore implies a long list of specific behaviours that will be observed in a given model world. For example, some criteria will lead to cooperation, others will not, some will lead to the existence of insurance contracts, others will not *etc.* We will explore the worlds created by the different models. Once we have done so we invite you to judge which model you find most useful for your understanding of the world. Of course, having spent many years thinking about these issues we have come to our own conclusions, and we will put them forward because we believe them to be helpful.

To keep the discussion to a manageable volume we will only consider a setup that corresponds to making purely financial decisions. We may bet on a horse or take out personal liability insurance. This chapter will not tell you whom you should marry or even whose economics lectures you should attend.

3.2 The decision axiom

A "decision theory" is a model of human behaviour. We will write down such a model phrased as the following simple axiom:

Decision axiom

People optimize the growth rate of their wealth.

Without discussing why people might do this, let's step into the world created by this axiom. To do that, we need to be crystal clear about what a growth rate is, so we'll discuss that first, in Sec. 3.3. Traditionally, decision theory deals with an uncertain future: we have to decide on a course of action now although we don't know with certainty what will happen to us in the future under any of our choices. We will systematically work our way towards this setup, beginning with trivial decisions where neither time nor uncertainty matters Sec. 3.4.1, next introducing time Sec. 3.4.2 (where we will shed light on what's called "discounting"). In the next chapter we will introduce uncertainty, see Sec. **??** (where we will shed light on what's called "expected utility theory").

3.3 Growth rates {section:Growth_rates}

You may have wondered why both

$$g_{\mathrm{a}} = \frac{\mathsf{x}(t+\Delta t) - \mathsf{x}(t)}{\Delta t} \tag{3.1}$$ {eq:add_rate}

and

$$g_{\mathrm{e}} = \frac{\ln \mathsf{x}(t+\Delta t) - \ln \mathsf{x}(t)}{\Delta t} \tag{3.2}$$ {eq:exp_rate}

are sometimes called a growth rate – they're different objects, why the same name? By the end of this section, the answer to this question should be clear.

When we say that $\mathsf{x}(t)$ is a growth process, we mean that it is a monotonic function of t. If you're thinking about randomness – don't, we'll come to that later. For now, we will just work with a deterministic function $\mathsf{x}(t)$ – we even use an unusual font to indicate that this is a deterministic function.

For a given process, the appropriate growth rate, g, solves the following problem for us: how do we characterise how fast x grows? A growth rate is a mathematical object of the form

$$g = \frac{\Delta v(\mathsf{x})}{\Delta t}, \tag{3.3}$$ {eq:gen_rate}

were $v(\mathsf{x})$ is a monotonically increasing function of wealth x. Comparing to (Eq. 3.1) and (Eq. 3.2), we find that $v(\mathsf{x}) = \mathsf{x}$ for additive dynamics and $v(\mathsf{x}) = \ln \mathsf{x}$ for multiplicative dynamics. The transformation $\mathsf{x} \to v(\mathsf{x})$ ensures temporal stability of g.

How does this work, and what does it mean? Specifically,

1. why the transformation?
2. how do we know which transformation to use?

We will start with the mathematically simple case of the additive growth rate, (Eq. 3.1), and discuss its properties by applying it to additive growth. Next we will ask under what conditions the exponential growth rate, (Eq. 3.2), is appropriate, and that will lead us to the general growth rate, (Eq. 3.3).

3.3.1 Additive growth rate {section:add_rate}

If I want to know how fast $\mathsf{x}(t)$ grows, the most obvious thing to compute is its rate of change – that's the additive growth rate, (Eq. 3.1). This tells me by how much x grows in the interval $[t, t+\Delta t]$. If $\mathsf{x}(t)$ is linear in t, so that

$$\mathsf{x}(t) = \mathsf{x}(0) + \gamma t \tag{3.4}$$ {eq:linx}

then this is a very informative quantity. We'll now state carefully why it is informative in this case. That may seem pedantic at this point, but it will become useful when we generalise in Sec. 3.3.2 and Sec. 3.3.3.

The additive growth rate (Eq. 3.1) is informative of how fast x grows under additive dynamics (Eq. 3.4) because in this case the t-dependence drops out: we can measure g_{a} whenever we want, and we'll always get the same value, $g_{\mathrm{a}} = \gamma$. Not to get too philosophical about it, but this kind of time-translation invariance

(fancy word) is a key concept in science: the search for laws is the search for universal structure – especially for time-translation invariant structure, for something "timeless."

Let's re-write the linear dynamic (Eq. 3.4) in differential form

$$d\mathsf{x} = \gamma dt \tag{3.5}$$

Because γ depends neither on t nor on x, we can re-write this as

$$d\mathsf{x} = d(\gamma t) \tag{3.6}$$

This way of writing it tells us that the growth rate γ is really a sort of clock speed. There's no difference between rescaling t and rescaling γ (by the same factor).

We make a mental note: *the growth rate is a clock speed.* But what kind of clock speed are we talking about? What's a clock speed anyway?

Or: what's a clock? A clock is a process that we believe does something repeatedly at regular intervals. It lets us measure time by counting the repetitions. By convention, after 9,192,631,770 cycles of the radiation produced by the transition between two levels of the caesium 133 atom we say "one second has elapsed." That's just something we've agreed on. But any other thing that does something regularly would work as a clock – like the Earth completing one full rotation around its axis *etc.*

When we say "the growth rate of the process is γ," we mean that x advances by γ units on the process-scale (meaning in x) in one standard time unit (in finance we often choose one year as the unit, Earth going round the Sun). So it's a conversion factor between the time scales of a standard clock and the process clock.

Of course, a clock is no good if it speeds up or slows down. For processes other than additive growth we have to be quite careful before we can use them as clocks, i.e. before we can state their growth rates.

3.3.2 Exponential growth rate {section:exp_rate}

Now what about the exponential growth rate, (Eq. 3.2)? This first thing to notice is that it's not time-translation invariant for additive growth, (Eq. 3.4). Substituting (Eq. 3.4) in (Eq. 3.2) gives

$$\begin{aligned} g_{\mathrm{e}} &= \frac{\ln \mathsf{x}(t+\Delta t) - \ln \mathsf{x}(t)}{\Delta t} && (3.7)\\ &= \frac{\ln\left[\mathsf{x}(t) + \gamma \Delta t\right] - \ln \mathsf{x}(t)}{\Delta t} && (3.8)\\ &= \frac{\ln\left(1 + \frac{\gamma \Delta t}{\mathsf{x}(t)}\right)}{\Delta t}. && (3.9) \quad \{\texttt{eq:exp\_lin}\} \end{aligned}$$

That means the exponential growth rate does not extract the clock speed γ from linear growth. There's a mismatch between the process and the form of the rate with which we're measuring its speed. The exponential growth rate of additive growth is not a constant but (see RHS of (Eq. 3.9)) depends on $\mathsf{x}(t)$, *i.e.* on the time when we started measuring. It also depends on how long we measured, Δt. If we used it to characterise the growth in described by (Eq. 3.4), we would

find lots of contradictions – some people would say the growth is faster, others slower because they measured at different times or for different periods.

But the exponential growth rate is commonly used, and for good reasons. Let's see what it's good for, by imposing that it's useful and then working backwards to find the process we should use it for (we expect to find exponential growth).

We require that (Eq. 3.2) yield a constant, let's call that γ again, irrespective of when we measure it.

$$g_e = \frac{\Delta \ln \mathsf{x}}{\Delta t} = \gamma, \tag{3.10}$$

or

$$\Delta \ln \mathsf{x} = \gamma \Delta t, \tag{3.11}$$

or indeed, in differential form, and revealing that again the growth rate is a clock speed: γ plays the same role as t,

$$d \ln \mathsf{x} = d(\gamma t). \tag{3.12}$$

This differential equation can be directly integrated and has the solution

$$\ln \mathsf{x}(t) - \ln \mathsf{x}(0) = \gamma t. \tag{3.13}$$

We solve for the dynamic $\mathsf{x}(t)$ by writing the log difference as a fraction

$$\ln \left[\frac{\mathsf{x}(t)}{\mathsf{x}(0)} \right] = \gamma t, \tag{3.14}$$

and exponentiating

$$\mathsf{x}(t) = \mathsf{x}(0) \exp(\gamma t) \tag{3.15}$$ {eq:expx}

As expected, we find that the *exponential* growth rate, (Eq. 3.2), is the appropriate growth rate (meaning time-independent) for *exponential* growth.

In terms of clocks, what just happened is this: we insisted that (Eq. 3.2) be a good definition of a clock speed. That requires it to be constant, meaning that the process has to advance on the logarithmic scale, specified in (Eq. 3.2), by the same amount in every time interval (measured on the standard clock, of course – Earth or caesium).

3.3.3 General growth rate {section:gen_rate}

Finally let's be more ambitious and posit a general process, $\mathsf{x}(t)$, of which we only assume that it grows according to a dynamic that can be written down as a separable differential equation. We could be even more general, but this is bad enough.

How do we define a growth rate now?

Well, as in Sec. 3.3.2, we insist that the thing we're measuring will be a clock speed, *i.e.* a time-independent rescaling of time. We enforce this by writing down the dynamic in differential form, containing the growth rate as a time rescaling factor. Then we'll work backwards and solve for g:

$$d\mathsf{x} = f(\mathsf{x}) d(gt) \tag{3.16}$$ {eq:gen_diff_x}

(for linear growth, like in (Eq. 3.4), $f(\mathsf{x})$ would just be $f(\mathsf{x}) = 1$, and for exponential growth, (Eq. 3.15), it would be $f(\mathsf{x}) = \mathsf{x}$, but we're leaving it general). We separate variables in (Eq. 3.16) and integrate the differential equation

$$\int_{\mathsf{x}(t)}^{\mathsf{x}(t+\Delta t)} \frac{1}{f(\mathsf{x})} \delta\mathsf{x} = g\Delta t, \tag{3.17}$$

and we've got what we want, namely the functional form of g:

$$g = \frac{\int_{\mathsf{x}(t)}^{\mathsf{x}(t+\Delta t)} \frac{1}{f(\mathsf{x})} d\mathsf{x}}{\Delta t}. \tag{3.18}$$ {eq:g_int}

This doesn't quite look like our stated aim: the general expression for a growth rate, (Eq. 3.3). But we get there, by simplifying (Eq. 3.18) and denoting the definite integral with the letter v, so that

$$\Delta v = \int_{\mathsf{x}(t)}^{\mathsf{x}(t+\Delta t)} \frac{1}{f(\mathsf{x})} d\mathsf{x}. \tag{3.19}$$ {eq:Dv}

Equation (3.3) now follows exactly by substituting (Eq. 3.19) in (Eq. 3.18). This answers the second question of Sec. 3.3: "how do we know which transformation to use?"

But there's a simpler way of finding the transformation $v(\mathsf{x})$ that doesn't involve integrals and differential equations. Let $\mathsf{x}(t)$ be whatever function it wants – we know one transformation of $\mathsf{x}(t)$ that's linear in time, namely the inverse function of $\mathsf{x}(t)$, which we denote

$$\mathsf{x}^{(-1)}(\mathsf{x}) = t \tag{3.20}$$

$\mathsf{x}^{(-1)}(\mathsf{x})$ is the transformation that pulls t out of x: I give you the value of x, you take $\mathsf{x}^{(-1)}(\mathsf{x})$, and you know what t is.

So far, so good – now we know how to get t, which is of course linear in t. But that no longer tells us how fast something grows: we can't use $\mathsf{x}^{(-1)}(\mathsf{x})$ as $v(\mathsf{x})$ because $\frac{\mathsf{x}^{(-1)}[\mathsf{x}(t+\Delta t)] - \mathsf{x}^{(-1)}[\mathsf{x}(t)]}{\Delta t} = 1$, always. So something is missing.

If we use $\mathsf{x}^{(-1)}$ as the transformation in the general growth rate, we're in effect measuring the speed of the process on the scale of the process, which is why the answer is trivial: we will always find a growth rate of 1. The growth rate is a *conversion factor* between time measured on the standard clock (one that ticks once a second, say), and time measured on the process clock (one that advances γ units on the $v(x)$-scale in each second).

So $\mathsf{x}^{(-1)}$ has the right form but not the right scale. Instead, let's try the following: take the process $\mathsf{x}(t)$ at unit rate on the standard clock. We'll denote this as $\mathsf{x}_1(t)$. If we now take its inverse as the transformation, $v(\mathsf{x}) = \mathsf{x}_1^{(-1)}(\mathsf{x})$, it will of course produce a rate 1 if $\gamma = 1$. But if γ is something else, it will extract that something else for us!

Here's the algorithm for measuring the growth rate for a general process $\mathsf{x}(t)$.

- Write down the process at rate 1 on the standard clock, $\mathsf{x}_1(t)$.
- Invert it, to find the transformation $v = \mathsf{x}_1^{(-1)}(\mathsf{x})$.

- Finally, evaluate the rate of change of the transformation of the process at the unknown growth rate,

$$g = \frac{\mathrm{x}_1^{(-1)}[\mathrm{x}(t + \Delta t)] - \mathrm{x}_1^{(-1)}[\mathrm{x}(t)]}{\Delta t} \tag{3.21}$$ {eq:g_inv}

The key conceptual message from this section is this: any growth process defines an appropriate functional form of a growth rate. If we measure a process with the wrong form of growth rate, we obtain something that's not stable in time. Measurements will be irreproducible or inconsistent, depending on arbitrary circumstances: the time of measurement or how long we measured for.

The key formal result is this: (Eq. 3.3) tells us there is a specific transformation of wealth that is required to state meaningfully how fast wealth grows. Equation (3.21) tells us what that transformation is.

The transformation is a linearisation. At the moment we could call it the stationarity transformation because it appropriately removes time dependence. Later – when we generalise to random growth processes – we will call it the ergodicity mapping. In the economics literature, the closest thing to this transformation is called the utility function – a term we will mostly avoid because it comes with unhelpful connotations.

3.4 Decisions in a deterministic world {section:Decisions_in_a_d

Having clarified what a growth rate is, we can now apply our decision axiom to different situations: act so as to maximize the growth rate of your wealth. In other words, we can explore the world generated by this axiom. We will build from the ground up. First, in Sec. 3.4.1, we will look at comparing two simultaneous payments of different magnitude – which one will the model human choose who obeys our axiom? This is a sanity check: does the model human choose the bigger payment? Next, in Sec. 3.4.2 we add time – what if the model human chooses between two payments of different magnitudes that will occur at different times? This is already a far more complex situation, where the decision criterion requires knowledge of the dynamic. It will shed light on what's called "discounting."

In the next chapter, Chap. 4, we will add fluctuations, noise, uncertainty: what if the model human doesn't know the magnitude (or time) of the payments with certainty? But for now, everything will be perfectly known.

3.4.1 Different magnitudes {section:Different_magni

I'm off to the bank to withdraw some money for you. I offer to give you either

(1) $10 when I get back or

(2) $25 when I get back. You tell me what you prefer.

Let's see what our decision axiom says you'll do. Remember there's no uncertainty, I'm not lying to you, no one will rob me on my way from the bank *etc.*

I haven't told you how long it will take me to get to the bank, so we have to keep that general. We'll call that time interval Δt. Because we know that Δt is the same under options (1) and (2) we don't actually need to know its value to compare the growth rates for the two options. Nor do we have to know the functional form of the growth rate. In this simple case, we can work with a general $v(\mathsf{x})$ in (Eq. 3.3), and any growth rate will give the same answer. Let's see. Under option (1) we have

$$g^{(1)} = \frac{v(\mathsf{x} + \$10) - v(\mathsf{x})}{\Delta t}, \tag{3.22}$$

and under option (2) we have

$$g^{(2)} = \frac{v(\mathsf{x} + \$25) - v(\mathsf{x})}{\Delta t}. \tag{3.23}$$

To find out which growth rate is larger, we subtract $g^{(1)}$ from $g^{(2)}$

$$g^{(2)} - g^{(1)} = \frac{v(\mathsf{x} + \$25) - v(\mathsf{x})(-v(\mathsf{x} + \$10) - v(\mathsf{x}))}{\Delta t} \tag{3.24}$$

$$= \frac{v(\mathsf{x} + \$25) - v(\mathsf{x} + \$10)}{\Delta t}. \tag{3.25}$$

Because $v(\mathsf{x})$ is monotonically increasing, any proper growth rate will be greater under option (2), and our model humans will always go for option (2). That's good – because we would have chosen option (2) if we were you, and our model reproduces this intuitive result.

More generally, our model says: of two certain payments of different sizes at the same time, choose the bigger one.

3.4.2 Different magnitudes and times: discounting

{section:Different_magnit

Let's make the decision a little harder: what if we offer you the same amounts as before, but now at different times:

(1) \$10 in a month or

(2) \$25 in two months?

Again, we will compute the two growth rates corresponding to options (1) and (2), and then choose the bigger one – that's how we have been programmed to behave in the world that our axiom is creating. But unlike in the previous case, the functional form of the growth rate will now be important.

Discounting under additive dynamics

Let's start with the additive growth rate, (Eq. 3.1), which is nothing but using the identity function in the general growth rate, $v(\mathsf{x}) = x$ in (Eq. 3.3). Which payment do the model humans choose according to this rate? We've got all the parameters, so this is just a matter of substitution

$${g_{\mathrm{a}}}^{(1)} = \frac{\mathsf{x} + \$10 - \mathsf{x}}{1 \text{ month}} = \$120 \text{ p.a.}, \tag{3.26}$$

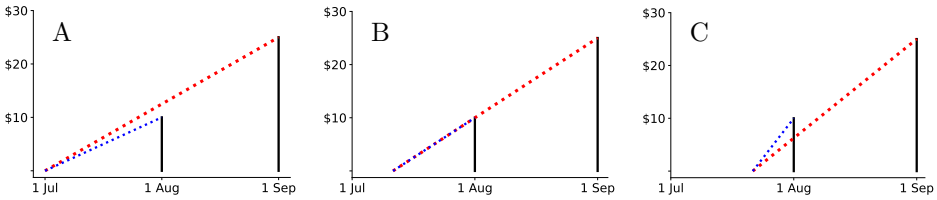


Figure 3.1: Slopes with linear vertical scales are *additive* growth rates. (A) At the beginning option (2) yields the highest additive growth rate; (B) after 1/3 of a month, the options are equally good; (C) after 2/3 of a month, preference reversal has taken place, and option (1) now yields the highest growth rate. As the first payment approaches, the associated growth rate diverges. {fig:hyp_disc}

and

$$g_{\mathrm{a}}{}^{(2)} = \frac{\mathsf{x} + \$25 - \mathsf{x}}{2 \text{ months}} = \$150 \text{ p.a.}. \tag{3.27}$$

The result is clear: the decision axiom, using this growth rate, produces model humans that prefer payment (2). The additive growth rate has a unique feature: initial wealth, x cancels out. I didn't need to know your initial wealth to compute the rate! Only under additive dynamics does initial wealth not enter into the computation of the growth rate, and growth rates can be computed with knowledge of only the payouts and waiting times.

But perhaps the setup is more interesting than it seems at first glance. In the economics literature, decision-making based on additive growth rates is called "hyperbolic discounting" because this case is mathematically equivalent to discounting payments in the future with the hyperbolic function $\frac{1}{\Delta t}$.

An interesting feature of optimizing additive growth rates is what's called "preference reversal:" let's keep our example as it is, except we now let time march forward, holding fixed the moments in time when the payments are to be made. Under these conditions, there comes a time, precisely after a third of a month, when option (2) is no longer preferred, see Fig. 3.2.

You may wonder why someone might model wealth as an additive process. Here is one possibility: if wealth is mostly affected by income and expenses then it will be described by an additive dynamic. Imagine you have a monthly salary of $\$1,000$, and you spend \$900 every month on all your expenses. So long as any investment income, like interest payments *etc.* , is negligible, your wealth will follow (Eq. 3.4) with $\gamma = \$100$ per month.

Discounting under multiplicative dynamics

What about the exponential growth rate, with $v(\mathsf{x}) = \ln \mathsf{x}$ in (Eq. 3.3)? We now have growth rates

$$g_{\mathrm{m}}{}^{(1)} = \frac{\ln(\mathsf{x} + \$10) - \ln(\mathsf{x})}{1 \text{ month}}, \tag{3.28}$$

and

$$g_{\mathrm{m}}{}^{(2)} = \frac{\ln(\mathsf{x} + \$25) - \ln(\mathsf{x})}{2 \text{ months}}. \tag{3.29}$$

Curiously, which is greater depends on your initial wealth in our model world. If your wealth is \$100, then $g_{\mathrm{m}}{}^{(1)} \approx 114\%$ p.a., and $g_{\mathrm{m}}{}^{(2)} \approx 134\%$ p.a., wherefore you will choose option (2).

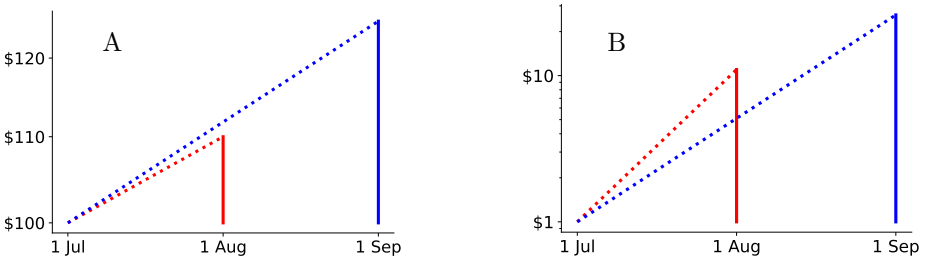


Figure 3.2: Slopes with logarithmic vertical scales. (A) If you have a lot of money (here \$100), exponential growth-rate optimization tells you to be patient and choose the later, larger, payment of \$25. (B) If you have little money (here \$1), the same criterion – exponential growth-rate optimization – tells you to get the cash as fast as possible, and choose the earlier, smaller, payment of \$10. {fig:hyp_disc}

But if your initial wealth is \$1, then $g_{\mathrm{m}}{}^{(1)} \approx 2,877\%$ p.a. and $g_{\mathrm{m}}{}^{(2)} \approx 1,955\%$ p.a., and you'll choose option (1).

Notice how the poorer decision maker seems to be more impatient, despite his use of the exact same decision axiom. Using $v(x) = \ln(x)$ in (Eq. 3.3) is related to what's called "exponential discounting" in the economics literature [**?**]. The word "exponential" is there because we're discounting under multiplicative dynamics (which means exponential growth), and the logarithm is the inverse function of the exponential, which we need to define the appropriate growth rate, see (Eq. 3.21).

Again, let's ask why one might want to model wealth as a multiplicative process. Multiplicative processes have the property that how much you gain tomorrow is proportional to how much you currently have. Increases in wealth are proportional to wealth – this multiplicative property is virtually ubiquitous in Nature. You can imagine many reasons why it applies to wealth. For instance, wealth can be invested in interest-paying bonds. Or it can be invested in oneself: I may pay for a roof over my head, which transforms my life and earning potential from that of someone living on the streets to that of someone with a home address. Similarly, I can invest in my health and education. In Nature, multiplicativity and exponential growth occurs whenever something lives. That's because the most successful definition we have of life is "that which self-reproduces," and self-reproduction implies multiplicative growth.

We learn: in this slightly more complex though still fully deterministic case, which option is preferable does not only depend on the options available but also on the personal circumstances of the decision maker. Both the way the decision maker thinks about wealth as a dynamical process, and his wealth influence his preferences.

Perhaps the most significant message is the richness of this problem. We're applying nothing but our simple axiom, but it forces us to choose how we think about the dynamics of our wealth, and in reality that may depend strongly on many difficult to specify circumstances. In real life payments are not just offered at some point in time, but usually in return for something – an asset or work. Depending on the specific exchange, an additive, multiplicative, or more general model will be appropriate. Such models are explored in even greater detail in [**?**].

Importantly, we need not resort to psychology to generate a host of behaviours, such as impatience of poorer individuals or preference reversal as time

ticks on.

Chapter 4

Decisions in a risky world {chapter:Risky}

In the previous section we saw some interesting types of behaviour that occur in a model world generated by our decision axiom. We were able to relate these to phenomena that already have names in the more complex model worlds of classical economics, for instance "discounting" and "preference reversal."

In the present section we will introduce randomness to our model, which will resemble situations where a decision maker is not completely sure about what the consequences of his decisions will be. We will do this in a way that's natural from the perspective of growth rate optimization, and it will lead us to discover precisely what the meanings are in our new framework, of concepts in classical economics, such as gambles and utility theory.

We clearly have to do something about our axiom: with randomness the growth rates that are the basis of decisions in our model will also be random. We get around that problem by interpreting "growth rate" in the axiom as "time-average growth rate" when there's randomness involved. Finding these time averages will be simple: the growth rate that's time-invariant for a give deterministic growth process is ergodic when we introduce randomness. Its time average can therefore be computed as its expectation value, which makes this a local (in time) operation.

Decision axiom *with randomness*

People optimize the *time-average* growth rate of their wealth.

4.1 Perturbing the process {section:Perturbing}

XXX new structure

perturb by adding noise to growth rate (no change).

point out $v(x)$ does BM (no change)

point out expectation of x is $\left\langle v^{(-1)}(v)\right\rangle$

because v is being perturbed symmetrically, if $v^{(-1)}$ is convex, then $\langle x\rangle > \mathsf{x}$, meaning perturbation increases expectation value, and is not neutral.

In these cases, what happens over time will underperform what happens to expected wealth (non-ergodicity of x).

convexity of $v^{(-1)}$ is guaranteed by concavity of $v(x)$ [also need $v(x)$ to be monotonically increasing, which it is by assumption].

concavity of $v(x)$ is called "risk aversion" in economics: that's because it corresponds to dynamics where the expectation value of x is misleading w.r.t what happens over time.

In other words: show that $\langle x \rangle$ is misleading. That's all. No need to compute correction, just make the structural statement.

Alternative: compute the correction. the magnitude of Jensen's inequality must grow as $x \to$ does that mean we can always define a simple correction term?

Compute expectation value of square-root normal distribution, *i.e.* check special case of Cramer explicitly. square-root u gives chi-squared distribution for x.

XXX

We begin by introducing noise into the wealth dynamic. This has to be done carefully because we want the significance of the noise to stay the same as time passes. That doesn't necessarily mean that the amplitude of the noise – the absolute size of the typical perturbation – will stay the same. But we don't want to have to adjust the perturbation by hand, either – we're looking for a systematic way of perturbing the process that automatically takes into account the way in which $x(t)$ changes with time.

Without further ado, here's the solution: introduce a constant-amplitude perturbation to something about the process that is otherwise unchanging. Of course – you've guessed it – the growth rate fits the bill. In general – namely unless $x(t)$ is additive – such a perturbation will change the expectation value $\langle x(t) \rangle$, that is, in general we will have

$$\langle x \rangle (t) \neq \mathsf{x}(t) \tag{4.1}$$ {eq:exp_changed}

and we will explore the consequences of this inequality. Incidentally, when we say $x(t)$ "is additive," we mean that time is an additive operation. Adding some δt to time t is then equivalent to adding some δx to x.

We will follow the familiar structure, and first try out this recipe for additive dynamics, $v(x) = x$, then for multiplicative dynamics, $v(x) = \ln x$, and finally for general dynamics.

4.1.1 Perturbed additive dynamics

{section:Perturbed additi

We expect additive dynamics to be a trivial case because the additive growth rate is just the rate of change, and the stationarity transformation is the identity, $v(x) = x$. We start from deterministic additive growth, written in differential form

$$g_{\mathrm{a}} = \frac{d\mathsf{x}}{dt} = \gamma \tag{4.2}$$

then rearrange and add the perturbation. This makes sure that the dynamic significance of the perturbation doesn't change over time: the constant growth rate becomes the ergodic growth rate. Specifically, we choose a standard Wiener

perturbation.

$$\begin{aligned} dx &= g_{\mathrm{a}}dt + \sigma dW(t) && (4.3)\\ &= \gamma dt + \sigma dW(t). && (4.4) \end{aligned}$$ {eq:BM_dx}

We integrate this (setting $x(0) = 0$ for simplicity)

$$\begin{aligned} x(t) &= \int_0^t g_{\mathrm{a}}ds + \sigma dW(s) && (4.5)\\ &= \gamma t + \sigma W(t). && (4.6) \end{aligned}$$ {eq:BM_x}

Because of additivity, in this special case we expect $\mathsf{x}(t) = \langle x(t)\rangle$ – meaning (Eq. 4.1) is not true here. So let's check by taking expectations

$$\begin{aligned} \langle x(t)\rangle &= \langle \gamma t + \sigma W(t)\rangle && (4.7)\\ &= \gamma t && (4.8) \end{aligned}$$

Comparing to x, we confirm that the perturbation in this case does not change the expectation value of the process.

To recap: first, the noise in this dynamic has constant dynamic significance because it is a constant-amplitude perturbation applied to something that is unchanging in time in the deterministic case. Second, the expectation value of this dynamic is identical to the unperturbed, deterministic, case ($\sigma = 0$).

4.1.2 Perturbed multiplicative dynamics

{section:Perturbed multip

Multiplicative dynamics will be less trivial but shouldn't be too hard. The multiplicative growth rate is just rate of change of the logarithm, meaning the stationarity transformation is the logarithm, $v(\mathsf{x}) = \ln \mathsf{x}$. We start from deterministic multiplicative growth, written in differential form

$$g_{\mathrm{e}} = \frac{d\ln \mathsf{x}}{dt} = \gamma \qquad (4.9)$$ {eq:gexp}

then, as before, rearrange and add the perturbation to the ergodic growth rate. Again this ensures that the dynamic significance of the perturbation doesn't change over time. Like in the additive case, we choose a standard Wiener perturbation.

$$d\ln x = \gamma dt + \sigma dW(t). \qquad (4.10)$$ {eq:GBM_dlnx}

Just as we did to arrive at (Eq. 4.6), we have to integrate a Brownian motion. Because we're working in the ergodically transformed variable, this is a recurring theme: whatever the process $x(t)$, once it's been put through the appropriate transformation, we will end up with Brownian motion. Integrating (Eq. **??**) (setting $\ln x(0) = 0$ for simplicity),

$$\begin{aligned} \ln x(t) &= \int_0^t \gamma ds + \sigma dW(s) && (4.11)\\ &= \gamma t + \sigma W(t). && (4.12) \end{aligned}$$ {eq:GBM_lnx}

Because the stationarity transformation is no longer the identity, we now have to invert the logarithm (apply $v^{(-1)}(\cdot) = \exp(\cdot)$) to find the actual process

$$x(t) = \exp\left[\gamma t + \sigma W(t)\right]. \tag{4.13}$$

Multiplicative dynamics are not additive, meaning there is a mismatch between the additive expectation value and the multiplicative effects of time. The expectation value of an exponentiated Wiener noise is boosted by the fluctuations: the expectation value is linear, but the exponential generates disproportionately large contributions for positive values of $W(t)$. We leave it as an exercise to compute the expectation value, $\langle x(t)\rangle$, (hint: $x(t)$ is log-normally distributed) and only state the well-known result here

$$\begin{aligned} \langle x(t)\rangle &= \langle \exp\left[\gamma t + \sigma W(t)\right]\rangle & (4.14)\\ &= \exp\left[\left(\gamma + \frac{\sigma^2}{2}\right) t\right] & (4.15) \end{aligned}$$

Comparing to x, we see that in this case, as in most cases, (Eq. 4.1) applies. If we want a multiplicative process whose expectation value is identical to its zero-noise limit, we have to include a correction. Indeed, when we consider multiplicative dynamics, we will work with this parameterization

$$d\ln x = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma dW(t). \tag{4.16}$$ {eq:log_inc_noise}

Again: the noise in this dynamic has constant dynamic significance because it is a constant-amplitude perturbation applied to something that is unchanging in time in the deterministic case. Thanks to the correction term $-\frac{\sigma^2}{2}$, the expectation value of this dynamic, (Eq. 4.16), is identical to the unperturbed, deterministic, case ($\sigma = 0$).

Equation (4.16) is solved for x by integrating and then exponentiating,

$$x(t) = \exp\left[\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W(t)\right]. \tag{4.17}$$ {eq:mult_x}

4.1.3 Perturbed general dynamics

{section:Perturbed_genera

The simplest, and probably most important, models to describe real-life wealth are additive, $v(x) = x$, and multiplicative, $v(x) = \ln x$. But it helps to think of those two cases as examples of something more general. In this section, we will re-do what we did for additive and multiplicative dynamics but keep things more general – within reason $x(t)$ may now follow any dynamic you like. To confirm we're really just generalising, you can go through the following steps and verify that Sec. 4.1.1 and Sec. 4.1.2 are special cases.

We find it intuitive to start with a deterministic process and add an appropriate perturbation. However, because this approach doesn't always produce the smoothest mathematics, we will return to the problem in Sec. 4.5 from a slightly different angle. That will allow us to use Itô calculus and lead to a deeper analysis, but first things first.

The starting point is now a general deterministic growth rate

$$\begin{aligned} g &= \frac{dv(\mathsf{x})}{dt} & (4.18)\\ &= \gamma. & (4.19) \end{aligned}$$ {eq:growth_gen}

We re-arrange this and add the perturbation

$$dv = \gamma dt + \sigma dW. \tag{4.20}$$ {eq:dv_gen}

As before, the perturbation is guaranteed to have constant dynamic significance because it is applied to the otherwise unchanging growth rate – that's how v is defined: it's the transformation of x that grows linearly in time (so that $g = \frac{dv}{dt}$ is constant, taking the value γ in the absence of perturbations). We integrate (assuming $v(0) = 0$ for simplicity),

$$\begin{aligned} v[x(t)] &= \int_0^t \gamma ds + \sigma dW && (4.21)\\ &= \gamma t + \sigma W(t) && (4.22)\\ & && (4.23) \end{aligned}$$ {eq:v_int}

...and apply the inverse function $v^{(-1)}$...

$$x(t) = v^{(-1)}[\gamma t + \sigma W(t)]. \tag{4.24}$$ {eq:inversion}

Equation (4.24) can be read as x being a non-linear function, $v^{(-1)}$, of a symmetrically perturbed time, $\gamma t + \sigma W(t)$. If $v^{(-1)}$ is convex, then Jensen's inequality tells us that the expectation value of this perturbed function is greater than the unperturbed case, meaning at any time

$$\langle x(t) \rangle > \mathsf{x}(t). \tag{4.25}$$ {eq:exp_x_gen}

It can be shown that $v^{(-1)}$ is convex whenever v itself is concave (to prove this v has to be monotonically increasing, which is true by construction in our case).

The expectation value is misleading

This means, whenever the ergodicity transformation is concave, the expectation value of x is misleadingly large: it is larger than x (the unperturbed case), and grows faster (at a higher rate) than any individual trajectory of x will grow in the long run. Put simply: as regards an individual trajectory, the expectation value is pure fiction in this case. Its performance over time is not indicative of the performance of real trajectories.

We will see below that a concave ergodicity transformation (called a utility function in economics) is associated with what economists call "risk aversion." I am deemed risk averse if I prefer the risk-free process $\mathsf{x}(t)$ to a risky process $x(t)$ whose expectation value is $\langle x(t) \rangle = \mathsf{x}(t)$. Because the expectation value is misleadingly large, it makes perfect sense to be risk averse: in the long run, the risk-free process – by Jensen's inequality above – is guaranteed to outperform the risky one.

4.2 The appropriate growth rate is ergodic

Equation (3.3) defines deterministic dynamics by specifying a growth rate. Knowledge of the growth rate is thus knowledge of the dynamic. In the deterministic case, we defined growth rates by insisting that they not change over

time. In the risky, or noisy, case, of course growth rates do change over time but – crucially – the appropriately defined growth rate changes only because of the noise and nothing else – it fluctuates from one time interval to another, but it does not systematically increase or decrease. Specifically, the way we constructed noisy dynamics, starting from the growth rate, guarantees that the noisy growth rates are ergodic: their expectation values are also their long-time averages. We will now prove this result. We will only prove it for the general case because that's easily done, and it implies the veracity of the statement for any special case.

Proof. To avoid problems with differentiating the non-differentiable Wiener process, we begin with the expectation value of gdt,

$$\begin{aligned} \langle gdt \rangle &= \langle dv \rangle && (4.26)\\ &= \langle \gamma dt + \sigma dW \rangle && (4.27)\\ &= \gamma dt && (4.28) \end{aligned}$$

Dividing by dt, we find the expectation value of g to be

$$\langle g \rangle = \gamma. \qquad (4.29) \quad \texttt{\{eq:exp\_value\_growth\_gen\}}$$

Next, we compute the time average of g. By the definition of time averages, that's

$$\begin{aligned} \overline{g} &= \lim_{T\to\infty} \frac{1}{T} \int_0^T g(t)dt && (4.30)\\ &= \lim_{T\to\infty} \frac{1}{T} \int_0^T dv && (4.31)\\ &= \lim_{T\to\infty} \frac{1}{T} \int_0^T [\gamma dt + \sigma dW] && (4.32)\\ &= \gamma + \lim_{T\to\infty} \frac{1}{T} \sigma W(T) && (4.33)\\ &= \gamma, && (4.34) \quad \texttt{\{eq:time\_average\_growth\_g} \end{aligned}$$

where the last line follows with probability one from the scaling properties of the Wiener process.

Comparing (Eq. 4.29) and (Eq. 4.34), we conclude that the expectation value and time average of g are identical for general dynamics, and hence that the growth rate, appropriately defined for a given process, is an ergodic observable. □

The result shows that the two averages are identical, irrespective of whether (Eq. 4.1) applies or not. So, while the expectation value of x usually does not reflect what actually happens over time in a single system, the expectation value of the growth rate g is *guaranteed* to reflect what happens over time in a single system.

We cannot overstate the importance of this fact: *the appropriate growth rate for a stochastic wealth process is an ergodic observable.* Economic theory, historically, is based on expectation values. This is simply because when the foundations of economic theory were laid, expectation values had been invented

as tools of analysis, whereas time averages had not. The ergodicity of growth rates allows us to use their expectation values, even in cases where we're actually interested in time averages. This, and its science-historical significance, will be worked out in detail in Sec. 4.4.

4.3 Ergodicity economics decision algorithm

We're now in a position to specify how humans in our model will make decisions under uncertainty. In Chap. 3 we worked out precisely how a growth rate must be defined for a given deterministic dynamic. In Sec. 4.1 we used our knowlegde of such growth rates to generate consistently perturbed dynamics.

It is possible to write down dynamics that don't allow the definition of a growth rate. It is also possible to perturb a dynamic in an inconsistent way. But in these lecture notes, all deterministic dynamics will have proper growth rates, and all stochastic dynamics are obtained as consistently perturbed deterministic dynamics.

Given these constraints on the processes we consider, we can specify the algorithm our model humans will follow when making decisions under risk.

Ergodicity economics decision algorithm

When deciding which option, m^*, to choose

1. Specify the wealth dynamic, $dx(t)$, or equivalently its ergodicity transformation, $v(x)$, or equivalently the form of the relevant growth rate g;
2. Determine the time-average growth rates, $\overline{g}^{(m)}$, either directly by averaging over time; or by invoking the ergodic property and averaging over the ensemble;
3. Choose the option, m^*, with the largest time-average growth rate.

4.4 Relation to earlier economic theories {section:Relation_to_ear

In the previous section we presented the ergodicity economics model of decision-making: maximise the time-average growth rate of wealth. That's it; everything in ergodicity economics can be related back to this simple idea.

In the present section, we will detail precisely how our model is related to models that were developed earlier, specifically at a time before the ergodicity question had been asked. This section is thus about history: how did we model human decision making under risk before we knew we had to find appropriate growth rates?

The story is fascinating: despite the absence of appropriate tools and concepts, early researchers invented ingenious mathematical representations of human behaviour. With not very much work, it is possible to tweak these models, at least in special cases, to coincide with ergodicity economics.

The roadmap for this section is as follows. We will first introduce the concept of a *gamble* – in essence that's a little piece of a stochastic process, a random

wealth change realised over some short time. Next, we will mention the *gamble problem*, namely the problem of choosing between different gambles when we have to. To connect to ergodicity economics, we point out how different ways of repeating a gamble can generate different dynamics, very similar to the types considered in Sec. 4.1. Finally, we will present the classic solution to the gamble problem, which assigns a utility $u(x)$ – non-linear in wealth – to monetary wealth. This will be summarised in the expected-utility-theory decision algorithm.

4.4.1 Gamble: random number and duration

{section:Gamble:}

One fundamental building block of mathematical decision theory is the gamble. This is a mathematical object that resembles a number of situations in real life, namely situations where we face a decision whose consequences will be purely financial and are somewhat uncertain when we make the decision. An example is buying a lottery ticket. We define the gamble mathematically as follows.

DEFINITION: **Gamble**

A gamble is a pair of a random variable, Q, and a duration, δt.

Q is called the payout and takes one of K (mutually exclusive) possible monetary values, $\{q_1, \ldots, q_K\}$, associated with probabilities, $\{p_1, \ldots, p_K\}$, where $\sum_{j=1}^{K} p_j = 1$. Payouts can be positive, associated with a monetary gain, or negative, associated with a loss. We order them such that $q_1 < \ldots < q_K$.

In economics, the duration of a gamble is rarely discussed but it's clearly crucial information, as a trivial example shows. Let's say you get to choose between two gambles. One pays \$1 with probability 1 and takes one second to complete; the other also pays \$1 also with probability 1 but takes one year to complete. Clearly the former is more attractive. Knowing the duration will also be necessary to relate the gamble to ergodicity economics – the latter being based on time, this should come as no surprise.

The following situations may be modelled as gambles:

Example: Betting on a fair coin

Imagine betting \$10 on the toss of a fair coin. We would model this with the following payouts and probabilities:

$$q_1 = -\$10, \qquad p_1 = 1/2; \tag{4.35}$$
$$q_2 = +\$10, \qquad p_2 = 1/2. \tag{4.36}$$

The duration may be the time until you receive the payout, or if you participate in one coin toss every week, say, we may want to make it $\delta t = 1$ week.

Example: Playing the lottery

We can also imagine a gamble akin to a lottery, where we pay an amount,

F, for a ticket which will win the jackpot, J, with probability, p. The corresponding payouts and probabilities are:

$$q_1 = -F, \qquad p_1 = 1 - p; \tag{4.37}$$
$$q_2 = J - F, \qquad p_2 = p. \tag{4.38}$$

Note that we deduct the ticket price, F, in the payout q_2. The duration may be $\delta t = 1$ week.

Example: The null gamble

It is useful to introduce the null gamble, in which a payout of zero is received with certainty: $q_1 = \$0$, $p_1 = 1$. This represents the 'no bet' or 'do nothing' option.

As in the examples above, the duration, δt, has to be chosen appropriately. The meaning of the duration will become clearer later on – often it is the time between two successive rounds of a gamble.

The gamble we have presented is discrete, in that the payout, Q, is a random variable with a countable (and, we usually assume, small) number of possible values. The extension to continuous random variables is natural and used frequently to model real-world scenarios where the number of possible outcomes, *e.g.* the change in a stock price over one day, is large.

This presents a natural connection to ergodicity economics: given a stochastic wealth process dx, the wealth change generated by that process over a certain time interval $[t, t + \delta t]$ is a gamble. The random variable is

$$Q = \int_t^{t+\delta t} dx, \tag{4.39}$$ {eq:gamble_from_process}

and the duration is δt.

Suppose now that you have to choose between two options that you've modelled as two gambles (possibly including the null gamble). Which should you choose, and why? This is the gamble problem, the central question of decision theory, and the basis for much of mainstream economics.

DEFINITION: **The gamble problem**

The gamble problem is the problem to choose between two gambles.

We stress here that the gamble alone is not enough information to answer this question. The value of a gamble – clearly – depends on more facts than probabilities, payouts, and duration. Crucially, it depends on

1. how the gamble affects our future. For instance: if we go bankrupt as a result, can we recover from that?

2. our circumstances. When you're very rich you can afford to take risks that you can't afford when you're poor.

3. our personality. Some like the thrill of gambling, others find it unpleasant.

Mainstream economics focuses on point 3. Ergodicity economics, on the other hand, focuses on points 1 and 2, where we would compute the average growth rates of the processes corresponding to the gambles. Even in the simple case of multiplicative dynamics, this requires knowledge of the individual's wealth before the gamble, $x(t)$. It also requires knowledge of the process itself. Without that we wouldn't know the form of the ergodic growth rate; we wouldn't know what to compute.

Although the gamble problem is underspecified by gambles alone, the gamble is a useful conceptual unit because it specifies that part of the model of evolving wealth that is independent of individual circumstances. It thus splits an easily observable part of the problem from information that's much harder to obtain. We can all buy the same lottery ticket with publicly specified prizes, probabilities, duration – for some of us that will be attractive, for others it won't, for a variety of reasons.

4.4.2 Repeated gamble: towards a wealth process

{section:Repeated_gamble}

We already know one touch point between ergodicity economics and the classic gamble setup: we know how to construct a gamble as the random variable corresponding to a time interval of a stochastic process – that's (Eq. 4.39). Ergodicity economics specifies how to evaluate a wealth process, and if it's justifiable to identify the gamble with the process, then we know how to evaluate the gamble under ergodicity economics.

We will now strengthen this connection by showing what it might mean to identify the gamble and the process. To do this we will take a gamble and construct from it a wealth process. That means we have to extend the gamble over an arbitrarily long time. A principled way of doing that (one that keeps extra assumptions to a minimum and clearly visible) is to imagine the gamble is being repeated over and over again.

Crucially, *the mode of repetition is not specified in the gamble itself.* It is the only additional assumption we have to make to arrive at a wealth process and unlock the power of ergodicity economics. We shall focus on two modes: *additive* and *multiplicative* repetition, which correspond to additive and multiplicative dynamics. Thus *the same gamble can correspond to different dynamics.*

Additive repetition

DEFINITION: **Additive repetition**

If a gamble is repeated additively, then a newly generated realization of the random payout, q, is added to $x(t)$ in each round. We define the change in wealth occurring over a single round as

$$\delta x(t) \equiv x(t + \delta t) - x(t). \tag{4.40}$$ {eq:DW_def}

In the additive case, we have

$$\delta x(t) = q. \tag{4.41}$$ {eq:DW_add}

In other words, under additive repetition, δx is a stationary random variable, meaning the ergodicity transformation is the identity, $v(x) = x$, and we're in the case of additive dynamics. Starting at time, t_0, wealth after T rounds is

$$x(t_0 + T\delta t) = x(t_0) + \sum_{\tau=1}^{T} q(\tau), \tag{4.42}$$ {eq:Wt_add}

where $q(\tau)$ is the realisation of the random variable in round τ. This is an evolution equation for wealth following a noisy additive dynamic. Note that $x(t_0 + T\delta t)$ is itself a random variable.

Example: Additive repetition of a \$10 bet

We return to our first example of a gamble: a \$10 bet on a coin toss. Under additive repetition, successive bets will always be \$10, regardless of how rich or poor you become. Suppose your starting wealth is $x(t_0) = \$100$. Then, following (Eq. 4.42), your wealth after T rounds will be

$$\begin{aligned} x(t_0 + T\delta t) &= \$100 + \$10k - \$10(T - k) && (4.43)\\ &= \$[100 + 10(2k - T)], && (4.44) \end{aligned}$$

where $0 \leq k \leq T$ is the number of tosses you've won. Note that we have assumed your wealth is allowed to go negative. If not, then the process would stop when $x < \$10$, since you would be unable to place the next \$10 bet.

Multiplicative repetition

An alternative is multiplicative repetition. In the example above, let us imagine that the first \$10 bet were viewed not as a bet of fixed monetary size, but as a fixed fraction of the starting wealth (\$100). Under multiplicative repetition, each successive bet is for the same fraction of wealth which, in general, will be a different monetary amount.

The formalism is as follows.

DEFINITION: **Mulitplicative repetition**

The payout, q_j, in the first round is expressed as a random wealth multiplier,

$$r_j \equiv \frac{x(t_0) + q_j}{x(t_0)}. \tag{4.45}$$ {eq:R_def}

This multiplier is another random variable, and multiplicative repetition means drawing a new instance of it every δt and multiplying wealth $x(t)$ accordingly. Wealth after T rounds of the multiplicatively repeated gamble is

$$x(t_0 + T\delta t) = x(t_0) \prod_{\tau=1}^{T} r(\tau), \tag{4.46}$$

where $r(\tau)$ is the realisation of the random multiplier in round τ. The ergodicity transformation is now $v(x) = \ln x$, by which we mean that logarithmic wealth changes, $\delta \ln x$, are stationary. The exponential growth rate is ergodic, and its expectation value

$$\frac{\langle \delta \ln x \rangle}{\delta t} \tag{4.47}$$

is the time-average growth rate under this mode of repetition.

Example: Multiplicative repetition

The \$10 bet on a coin toss is now re-expressed as a bet of a fixed fraction of wealth at the start of each round. Following (Eq. 4.45), the random multiplier, r, has two possible outcomes:

$$r_1 = \frac{\$100 - \$10}{\$100} = 0.9, \qquad p_1 = 1/2; \tag{4.48}$$

$$r_2 = \frac{\$100 + \$10}{\$100} = 1.1, \qquad p_2 = 1/2. \tag{4.49}$$

The wealth after T rounds is, therefore,

$$x(t_0 + T\delta t) = \$100\ (1.1)^k\ (0.9)^{T-k}, \tag{4.50}$$

where $0 \leq k \leq T$ is the number of winning tosses. In this example there is no need to invoke a 'no bankruptcy' condition, since we can lose no more than 10% of our wealth in each round.

We have defined a gamble and clarified how it is related to ergodicity economics. It can be derived from a wealth dynamic; conversely a wealth dynamic can be constructed from a gamble provided we know how to repeat the gamble. But early decision theory didn't use the concepts of repetition, time averages, or ergodicity transformations. In the next section we will present how the gamble problem was addressed before the advent of ergodicity economics, and precisely how this earlier treatment is related to ours.

4.4.3 Expected wealth and expected utility

In ergodicity economics, we solve the gamble problem by maximizing the time (or ensemble) average of the ergodic growth rate for the process we consider the gamble to be part of.

But until recently, this was not how economists treated the gamble problem. It's instructive to mention two earlier criteria for gamble evaluation: the expected wealth change, invented around 1654; and the expected utility change, invented around 1738. Despite the conceptual error embedded in these criteria – confusing expectation values with time averages – they are easy to express in terms of time averages and ergodicity economics. In later developments, such as extensions of expected-utility theory beginning in the 1930s, and of prospect theory in the 1970s, the original error is compounded, and it is unclear how to ascribe physical meaning to the resulting models.

Expected wealth

The first gamble evaluation criterion, which emerged in the early days of probability theory in the 17$^{\text{th}}$ century, was this:

Expected wealth decision algorithm

1. Specify $Q^{(m)}$ for the gambles offered;
2. Determine the change of expected wealth induced by each gamble,

$$\langle \delta x \rangle^{(m)} = \langle Q(^m) \rangle \,; \tag{4.51}$$ {eq:EUT_criterion}

3. Choose the gamble, m^*, with the largest $\langle \delta x \rangle^{(m)}$.

The history of this criterion is linked to finding what was perceived as a fair value of an uncertain prospect. Say we're playing a game of chance, for some amount of money in a pot, maybe we're rolling dice, best out of three. If you're currently ahead, say after two throws, and someone wants to take over your position, it may be fair to sell your position for the expectation value of your winnings. One reason this criterion makes some sense is conservation: the sum of the expected winnings of all players is precisely the pot, so buying everyone out of his position is equivalent to buying the pot, as perhaps it should be.

Connecting this back to ergodicity economics: under what conditions would we evaluate a gamble by computing $\langle \delta x \rangle$? The answer is: under additive dynamics. The ergodicity economics maximand is then $\langle g_{\text{a}} \rangle = \frac{\langle \delta x \rangle}{\delta t}$. Apart from the δt in the denominator (which cancels if we assume it's the same for all considered gambles), this is expected-wealth maximisation. We conclude that expected-wealth maximisation is equivalent to ergodicity economics under the assumption of additive wealth dynamics: one of the simplest wealth models we can write down.

There is even a good reason why additive dynamics may have been assumed in the early developments. Let's Taylor-expand the time average of the general ergodic growth rate (which we may write as an ensemble average because of

ergodicity)

$$\langle g \rangle = \frac{\langle \delta v(x) \rangle}{\delta t} \tag{4.52}$$

$$= \frac{1}{\delta t} \left\langle \left[\frac{dv}{dx} \delta x + \frac{1}{2} \frac{d^2 v}{dx^2} \delta x^2 + \cdots \right] \right\rangle \tag{4.53}$$

$$= \frac{1}{\delta t} \left[\frac{dv}{dx} \langle \delta x \rangle + \frac{1}{2} \frac{d^2 v}{dx^2} \langle \delta x^2 \rangle + \cdots \right] \tag{4.54}$$

The last line follows from the fact that the derivatives of v are known functions, evaluated at the known wealth x before the gamble takes place. In other words, they are just constants and can be taken out of the expectation operator, $\langle \cdot \rangle$. If δx is small, keeping only the first term in the expansion is often a valid approximation, in which case we will call the dynamic "linearisable," and the expression becomes

$$\approx \frac{\langle \delta x \rangle}{\delta t} \frac{dv}{dx}, \tag{4.55}$$

which is proportional to the expected wealth change. In behavioural terms, being proportional means yielding the same ranking of any set of gambles as just the expected wealth $\langle \delta x \rangle$.

Summary: the first formal decision theory is expected wealth maximization. This theory is equivalent to ergodicity economics under additive dynamics, and additive dynamics are equivalent to any linearisable dynamic in the small-stakes limit.

Expected utility

In the language of economics, the expected-wealth paradigm treats humans as 'risk neutral', *i.e.* they have no preference between gambles whose expected changes in wealth are identical (over the same time interval). For example, people would be indifferent to either keeping what they have (the null gamble), or tossing a coin to win or lose $1,000.

At least since 1713 [**?**, p. 402], this has been known to be a flawed model. Nicolas Bernoulli pointed out that gambles can be constructed – at least in theory – whose expected wealth change does not exist. What then, would this criterion mean? Moreover, expected wealth maximisation does not always accord well with observed behaviour. For instance, anyone who buys an insurance contract prefers the certain loss of the insurance fee to a random loss of smaller expectation value. Such a person does not maximise expected wealth, but insurance contracts have been signed since Phoenecian times.

In 1738 Daniel Bernoulli put forward a new model of human behaviour that addressed some of the empirical failures of expected wealth maximisation[1].

Bernoulli noted that the value to an individual of a possible change in wealth depends on how much wealth the individual already has and on his psychological attitude to taking risks. In other words, people do not treat equal amounts of extra money equally. This makes intuitive sense: an extra $10 is much less significant to a rich man than to a pauper for whom it represents a full belly;

[1] Bernoulli's original paper contains an error. The theory as we present it here is a corrected version that can be found in [**?**], see also [**?**] for a discussion of the error

an inveterate gambler has a different attitude to risking \$100 on the spin of a roulette wheel than a prudent saver, their wealths being equal.

In 1738 Bernoulli [**?**], after correspondence with Cramer, devised the 'expected-utility paradigm' to model these considerations. He observed that money may not translate linearly into usefulness and assigned to an individual an idiosyncratic utility function, $u(x)$, that maps his wealth, x, into usefulness, u. He claimed that this was the true quantity whose expected change, $\langle \delta u(x) \rangle$, is maximised in a choice between gambles.

This is the axiom of utility theory. It leads to yet another decision algorithm.

Expected-utility decision algorithm

1. Specify $Q^{(m)}$ for the gambles offered;
2. Specify the individual's idiosyncratic utility function, $u(x)$, which maps his wealth to his utility;
3. Determine the change of expected utility induced by the gamble,
$$\langle \delta u \rangle = \left\langle u\left(x + q^{(m)}\right)\right\rangle - u(x); \tag{4.56}$$ {eq:EUT_criterion}
4. Choose the gamble, m^*, with the largest $\langle \delta u \rangle^{(m)}$.

Let's ask the same question as for the expected-wealth paradigm: under what conditions would ergodicity economics maximise the quantity in (Eq. 4.56)? The utility function $u(x)$ is defined – somewhat circularly, as *e.g.* von Neumann and Morgenstern pointed out [**?**, p. 28] – as the object whose expected changes are maximised by a person. Under ergodicity economics, what's being maximised is the expected (rate of) change of the ergodicity transformation, $v(x)$.

Mapping: expected-utility theory ⇔ ergodicity economics

We conclude that expected-utility theory is equivalent to ergodicity economics if gambles of equal duration, δt are considered and *if the utility function, u, coincides with the ergodicity transformation v.*

But it gets even better. Bernoulli did not just write down a general function $u(x)$ but argued that the logarithm is a plausible candidate for this function. People, he claimed, tend to behave as if they were optimising expected changes in the logarithm of wealth. Quite why that was the case, he didn't know. We do: using the logarithm as the ergodicity transformation, *i.e.* equating $u = v$, we find that Bernoulli's observation can be rephrased: people commonly maximise the time average of the ergodic growth rate under multiplicative dynamics. That's the second important wealth model we identified!

This is quite an astonishing correspondence, and we believe it is not coincidental. In the 18th century, researchers discovered elements of the mathematical structure of ergodicity economics. Just by careful observation – the appropriate mathematical tools had yet to be invented. Because mathematical concepts were immature at the time, a nomenclature emerged ("utility," "risk preferences" *etc.*) that seems quaint from today's conceptual context.

Summary: historically, the second formal decision theory is expected utility maximisation. This theory is equivalent to ergodicity economics if the utility function is the ergodicity transformation. Each dynamic thus has a corresponding growth-optimal utility function. The special case treated by Bernoulli in detail – logarithmic utility – is equivalent to ergodicity economics under multiplicative dynamics. Expected wealth maximisation is a special case of expected-utility maximisation, namely using a linear utility function (corresponding to ergodicity economics under additive dynamics).

4.5 From growth rate to dynamic and back – Itô

{section:From_growth}

In this section we will use Itô calculus to consolidate and extend our results a little. In addition to what we already know (converting ergodicity mappings into wealth processes), we will arrive at a recipe for going the other way: converting wealth processes, dx, into ergodicity mappings v. This is an important part of the connection to expected-utility theory: give us the growth process, and we will tell you which utility function will outperform all others in the long run.

Furthermore, we will arrive at a set of conditions for these mapping to be possible. We illustrate the procedure with one example for each direction: we derive the growth process that corresponds to a square-root ergodicity mapping; and we derive the utility function (which will be exponential) that corresponds to a curious-looking dynamic. In other words, we go explicitly beyond linear and logarithmic ergodicity mappings.

4.5.1 Itô setup

{section:Ito_setup}

We have seen that specifying a process $x(t)$ is equivalent to specifying an ergodicity transformation $v(x)$. When we mapped one onto the other, we had to invert v. This tells us one restriction: v has to be invertible. In this section, we will do the mapping again, but using Itô calculus, which will allows us to say more about the restrictions on v and x.

We will also revisit the inequality (Eq. 4.25) and discover ageing – a dependence on t in its magnitude.

We start with the self-imposed restriction that wealth dynamics are an Itô process, which you may remember from (Eq. 2.64) in Sec. 2.12. We will further restrict ourselves to coefficient functions $a_x(x)$ and $b_x(x)$ without explicit t dependence, meaning wealth will follow

$$dx = a_x(x)dt + b_x(x)dW. \tag{4.57}$$ {eq:dx_Ito}

In this phrasing, we can implement additive dynamics by setting $a_x = \mu$ and $b_x = \sigma$ as constants (Brownian motion, (Eq. 4.4)). We can also choose multiplicative dynamics, with $a_x = \mu x$ and $b_x = \sigma x$ (geometric Brownian motion, (Eq. 4.10)). So Itô processes include the most important models we've already seen, and many others.

Our aim is to find pairs of processes dx and ergodicity mappings $v(x)$, meaning we're after a connection between an Itô process and a function of an Itô

process. That's exactly what Itô's formula is for. By construction, v follows a Brownian motion with drift, which is another Itô process, namely

$$dv = a_v(v)dt + b_v(v)dW \tag{4.58}$$
$$= \gamma dt + \sigma dW \tag{4.59}$$ {eq:dv}

4.5.2 From ergodicity transformation to wealth process

Previously, we wrote x in terms of the inverse function of v as $x = v^{(-1)}(\gamma t + \sigma W(t))$. Itô's formula allows us to write dx in terms of the coefficient functions of dv in (Eq. 4.59) as follows

$$dx = \underbrace{\left(\frac{\partial x}{\partial t} + a_v\frac{\partial x}{\partial v} + \frac{1}{2}{b_v}^2\frac{\partial^2 x}{\partial v^2}\right)}_{a_x(x)}dt + \underbrace{b_v\frac{\partial x}{\partial v}}_{b_x(x)}dW \tag{4.60}$$ {eq:dx}

This expression involves partial derivatives of $x(v)$, *i.e.* of the inverse function of v (of course, $x(v) = v^{-1}(v)$). This confirms the constraint we already knew: v has to be invertible. Another constraint – clearly – is that $x(v)$ has to be twice-differentiable.

We have thus shown that

Invertible ergodicity mappings (utility functions) have dynamic interpretations

For any invertible ergodicity mapping $v(x)$ a class of corresponding wealth processes dx can be obtained such that the rate of change (*i.e.* the additive growth rate) in the expectation value of net changes in utility is the time-average growth rate of wealth.

Optimising the expected changes of the ergodicity mapping, $\langle \Delta v \rangle$, is equivalent to optimising time-average wealth growth for the corresponding wealth process, $\overline{g}(x)$.

We caution that it will be impossible for some processes $x(t)$ to find a $v(x)$ that satisfies (Eq. 4.59). In this case we cannot interpret expected utility theory dynamically, and such processes are likely to be pathological.

In the language of utility theory, every invertible utility function is an encoding of a wealth dynamic. Under that dynamic, behaving according to the corresponding utility function is optimal over time. The dynamic arises as utility – or rather: v – performs Brownian motion, and wealth is the transformation $v^{(-1)}(v(t))$.

Equation (4.60), creates the now familiar pairs of ergodicity mappings (utility functions) $v(x)$ and dynamics dx. Below we state the two familiar examples and work out a third one to illustrate the generality and ease of using Itô calculus.

Examples

- The linear ergodicity mapping (utility function) corresponds to additive wealth dynamics (Brownian motion),

$$v(x) = x \quad \leftrightarrow \quad dx = a_v dt + b_v dW, \tag{4.61}$$

as is easily verified by substituting $x(v) = v$ in (Eq. 4.60).

- The logarithmic ergodicity mapping (utility function) corresponds to multiplicative wealth dynamics (geometric Brownian motion),

$$v(x) = \ln(x) \quad \leftrightarrow \quad dx = x\left[\left(a_v + \frac{1}{2}{b_v}^2\right)dt + b_v dW\right]. \tag{4.62}$$

- To demonstrate the generality of our procedure, we carry it out for another special case that is historically important. The first utility function ever to be suggested was the square-root function $u(x) = x^{1/2}$, by Cramer in a 1728 letter to Daniel Bernoulli, partially reproduced in [**?**]. What would be the dynamic under which it is optimal to behave according to this utility function? In other words, what dx corresponds to the ergodicity mapping $v(x) = x^{1/2}$? This case will display ageing – a new phenomenon that we haven't encountered yet. So we'll got through it slowly.

  The square-root function is invertible, namely $x(v) = v^2$, and this inverse is twice-differentiable. So we can use (Eq. 4.60). In the next three lines we

  – substitute v^2 for $x(v)$ in (Eq. 4.60)
  – carry out the differentiations
  – substitute $x^{1/2}$ for v.

$$\begin{aligned} dx &= \left(a_v \frac{\partial v^2}{\partial v} + \frac{1}{2}{b_v}^2 \frac{\partial^2 v^2}{\partial v^2}\right)dt + b_v \frac{\partial v^2}{\partial v} dW && (4.63) \\ &= (2\gamma v + \sigma^2)dt + 2\sigma v dW && (4.64) \quad \texttt{\{eq:dx\_in\_v\}} \\ &= (2\gamma x^{1/2} + \sigma^2)dt + 2\sigma x^{1/2} dW. && (4.65) \end{aligned}$$

  We've thus established the third mapping in our collection:

$$v(x) = x^{1/2} \quad \leftrightarrow \quad dx = \left(2a_v x^{1/2} + {b_v}^2\right)dt + 2b_v x^{1/2} dW. \tag{4.66} \quad \texttt{\{eq:dx\_2\}}$$

  The new phenomenon we've mentioned – ageing – is observed when we compare the time average growth rate and that of the expectation value. By construction, the time-average growth rate is $\overline{g}(x) = \frac{\langle \Delta v \rangle}{\Delta t} = \gamma$.

  To compare this to the growth rate of the expectation value, we compute $\langle x \rangle$ as follows. Take the expectation value of (Eq. 4.64)

$$\begin{aligned} \langle dx \rangle &= (2\gamma \langle v \rangle + \sigma^2)dt && (4.67) \\ &= (2\gamma t + \sigma^2)dt && (4.68) \end{aligned}$$

  This can be integrated to yield

$$\langle x \rangle = \gamma t^2 + \sigma^2 t + \langle x(0) \rangle \tag{4.69}$$

  and we find its growth rate by differentiating,

$$\begin{aligned} \frac{dv\langle x \rangle}{dt} &= \frac{\partial v(\langle x \rangle)}{\partial \langle x \rangle} \times \frac{d\langle x \rangle}{dt} && (4.70) \\ &= \frac{1}{2}\langle x \rangle^{-1/2}(2\gamma t + \sigma^2) && (4.71) \\ &= \frac{(\gamma^2 t + \frac{1}{2}\sigma^2)}{(\gamma^2 t^2 + \sigma^2 t)^{1/2}} && (4.72) \end{aligned}$$

The first thing to notice here is that the expectation-value growth rate, appropriately defined, is not constant but depends on when we measure it! This is what is meant by "ageing:" the system, the dynamic itself, changes with time.

The second thing to notice is the asymptotic value. In the long run, the growth rate of the expectation value, for the Cramer dynamic converges to the ergodic growth rate,

$$\lim_{t\to\infty} \frac{dv(\langle x\rangle)}{dt} = \gamma. \tag{4.73}$$

It's quite subtle. While Jensen's inequality still holds,

$$\frac{dv(\langle x\rangle)}{dt} > \frac{d\,\langle v(x)\rangle}{dt}, \tag{4.74}$$

the magnitude of this inequality vanishes in the long-time limit. We leave the discussion of the Cramer dynamic here. It shows that ergodicity economics is full of open interesting avenues. Try out your own dynamics, and discover new phenomena!

History: Bounded utility functions in mainstream economics

Curiously, a celebrated but erroneous paper by Karl Menger [**?**] "proved" that all utility functions must be bounded (the proof is simply wrong). Boundedness makes utility functions non-invertible and precludes the developments we present here. Influential economists lauded Menger's paper, including Paul Samuelson [**?**, p. 49] who called it "a modern classic that [...] stands above all criticism." This is one reason why mainstream economics has failed to use the optimisation of wealth growth over time to understand human behavior – a criterion we consider extremely simple and natural. A discussion of Menger's precise errors can be found in [**?**, p. 7]. Although mainstream economics still considers boundedness of utility to be formally required, it is such an awkward restriction that John Campbell noted recently [**?**] that "this requirement is routinely ignored."

4.5.3 From wealth process to ergodicity transformation

{section:From_wealth}

We now ask under what circumstances the procedure in (Eq. 4.60) can be inverted: when can we find an ergodicity mapping for a given dynamic? Where this is possible, optimisation over time can be represented by optimisation of expected-utility changes.

We ask whether a given dynamic can be mapped into a $v(x)$ that follows Brownian motion, (Eq. 4.59).

In Sec. 4.5.1 we restricted ourselves to wealth following an Itô process, so that (Eq. 4.57) applies, with $a_x(x)$ and $b_x(x)$ as arbitrary functions of x. For this dynamic to translate into a Brownian motion for $v(x)$, (Eq. 4.59) must be satisfied. The tricky part here – the part that gives us constraints – is that the coefficients a_v and b_v are constants, namely γ and σ. Let's write this in an

equation

$$dv = \underbrace{\left(a_x(x)\frac{\partial v}{\partial x} + \frac{1}{2}{b_x}^2(x)\frac{\partial^2 v}{\partial x^2} \right)}_{a_v} dt + \underbrace{b_x(x)\frac{\partial v}{\partial x}}_{b_v} dW. \tag{4.75}$$ {eq:du_2}

To avoid clutter, let's use Lagrange notation, namely a dash – ′ – to denote a derivative. Explicitly, we arrive at two equations for the coefficients

$$a_v = a_x(x)v' + \frac{1}{2}{b_x}^2(x)v'' \tag{4.76}$$ {eq:A}

and

$$b_v = b_x(x)v'. \tag{4.77}$$ {eq:b_u}

Differentiating (Eq. 4.77), it follows that

$$v''(x) = -\frac{b_v {b_x}'(x)}{{b_x}^2(x)}. \tag{4.78}$$

Substituting in (Eq. 4.76) for v' and v'' and solving for $a_x(x)$ we find the drift term as a function of the noise term,

$$a_x(x) = \frac{a_v}{b_v}b_x(x) + \frac{1}{2}b_x(x){b_x}'(x). \tag{4.79}$$ {eq:consistency}

This equation is a consistency check of the wealth dynamic dx. An ergodicity mapping exists if and only if the coefficient functions $a_x(x)$ and $b_x(x)$ satisfy (Eq. 4.79). We do not need to construct the mapping explicitly to know whether a pair of drift term and noise term is consistent or not.

But once we know that it exists we can construct it by substituting for $b_x(x)$ in (Eq. 4.76). This yields a differential equation for v

$$a_v = a_x(x)v' + \frac{{b_v}^2}{2v'^2}v'' \tag{4.80}$$

or

$$0 = -a_v v'^2 + a_x(x)v'^3 + \frac{{b_v}^2}{2}v''. \tag{4.81}$$

Overall, then the triplet noise term, drift term, ergodicity mapping is interdependent. Given a noise term we can find consistent drift terms, and given a drift term we find a consistency condition (differential equation) for the ergodicity mapping. These arguments may seem a little esoteric when first encountered, using bits and pieces from different fields of mathematics. But they constitute the actual physical story behind the fascinating history of decision theory. To prevent the discussion from getting too dry, let's illustrate the procedure with an example.

Example: a curious-looking dynamic

We will work with the following wealth dynamic

$$dx = \left(\frac{a_v}{b_v}e^{-x} - \frac{1}{2}e^{-2x} \right) dt + e^{-x}dW. \tag{4.82}$$ {eq:test_dyn}

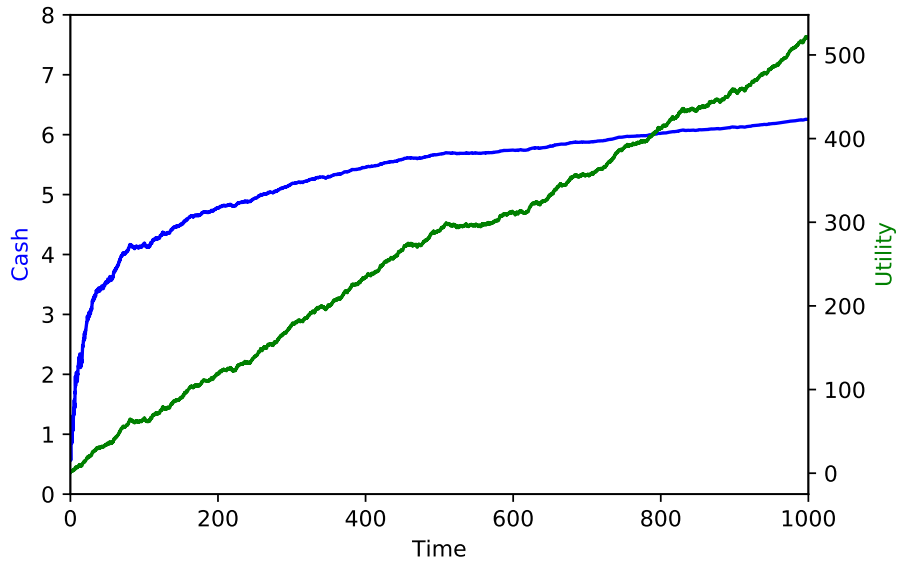


Figure 4.1: Typical trajectory $x(t)$ of the wealth dynamic (Eq. 4.82), with parameter values $a_v = 1/2$ and $b_v = 1$, and the corresponding Brownian motion $v(t)$. Note that the fluctuations in $x(t)$ become smaller for larger wealth. {fig:test_dyn}

A typical trajectory of (Eq. 4.82) is shown in Fig. 4.1.

We will check whether an ergodicity mapping exists for it, find that to be the case, then construct the ergodicity mapping explicitly, and discuss some of its properties.

Check consistency: In (Eq. 4.82), we have $a_x(x) = \frac{a_v}{b_v} e^{-x} - \frac{1}{2} e^{-2x}$ and $b_x(x) = e^{-x}$. Equation (4.79) imposes conditions on the drift term $a_x(x)$ in terms of the noise term $b_x(x)$. Substituting in (Eq. 4.79) reveals that the consistency condition is satisfied by the dynamic in (Eq. 4.82).

Construct the ergodicity mapping: Because (Eq. 4.82) is internally consistent, it is possible to derive the corresponding ergodicity mapping. Equation (4.77) is a first-order ordinary differential equation for $v(x)$

$$v'(x) = \frac{b_v}{b_x(x)}, \tag{4.83}$$ {eq:diff_eq_u}

which can be integrated to

$$v(x) = \int_0^x d\tilde{x} \frac{b_v}{b_x(\tilde{x})} + C, \tag{4.84}$$

with C an arbitrary constant of integration.

Substituting for $b_x(x)$ from (Eq. 4.82), (Eq. 4.83) becomes

$$v'(x) = b_v e^x, \tag{4.85}$$

which is easily integrated to

$$v(x) = b_v e^x + C, \tag{4.86}$$ {eq:test_dyn_u}

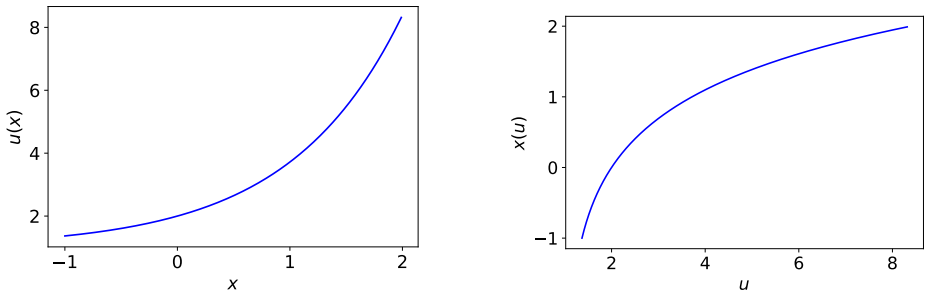


Figure 4.2: The exponential ergodicity mapping (or utility) $v(x)$, (Eq. 4.86) with $b_v = 1$ and $C = 1$, is monotonic and unbounded and therefore invertible. Left panel: $v(x)$. Right panel: inverse $x(v)$. {fig:u_of_x}

plotted in Fig. 4.2.

Discussion: This exponential ergodicity mapping is monotonic and therefore invertible – we knew that because the consistency condition is satisfied.

The ergodicity mapping is convex. From the perspective of expected-utility theory (where the ergodicity mapping is called a utility function), an individual behaving optimally according to this function would be labelled "risk-seeking." This behavior would commonly be observed under the dynamic (Eq. 4.82) because Jensen's inequality in (Eq. 4.25) now points the other way from what we had in all previous examples: the expectation value of x *understates* what typically happens to a single trajectory over time.

Another new phenomenon: under the dynamic (Eq. 4.82), optimal behaviour is "risk-seeking," in the sense that it will lead to faster wealth growth than risk-averse behaviour. This dynamic has the feature that fluctuations in wealth become smaller as wealth grows. High wealth is therefore sticky – an individual will quickly fluctuate out of low wealth and into higher wealth. It will then tend to stay there.

Once more, we see the conceptual difference to mainstream economics and utility theory: from the perspective of ergodicity economics what's optimal is determined by the dynamic, not by the individual. Of course the individual may choose not to behave optimally.

We end here the abstract discussion of dynamics and ergodicity mappings. Much remains to be discovered in this space, but we'd like to get on to real-world applications. Before we do that, we can't help ourselves but to hint at another instructive example, which we invite the reader to explore. Auto-elastic ergodicity mappings of the form

$$v(x) = \frac{x^{1-\eta} - 1}{1 - \eta}. \tag{4.87}$$

These can be concave or convex, depending on the parameter η. For $\eta > 1$ they are bounded, and that leads to curious problems. Because $v(x)$ is bounded, no matter how large x becomes, some values of $v(x)$ cannot occur. At the same time, our framework assumes that $v(t)$ performs a Brownian motion with drift and that it will eventually reach any value. All of this creates an internal conflict that corresponds to finite-time singularities in wealth. In other words: assuming bounded utility functions, when translated into dynamic terms, amounts to assuming finite-time singularities in wealth. Beyond these singularities, wealth becomes a complex number – mathematically, this is fun, but of course physical

realism is then lost. It is interesting, of course, that Menger, with his bounded utility functions, and those who endorsed him inadvertently argued for all of economics to take place in this non-physical realm outside the real numbers. A formal disaster.

4.6 The St Petersburg paradox

Sec. **??** The problem known today as the St Petersburg paradox was suggested by Nicolaus Bernoulli[2] in 1713 in his correspondence with Montmort [**?**]. It involves a hypothetical lottery for which the rate of change of expected wealth diverges for any finite ticket price. The expected-wealth paradigm would predict, therefore, that people are prepared to pay any price to enter the lottery. However, when the question is put to them, they rarely want to wager more than a few dollars. This is the paradox. It is the first well-documented example of the inadequacy of the expected-wealth paradigm as a model of human rationality. It was the primary motivating example for Daniel Bernoulli's and Cramer's development of the expected-utility paradigm [**?**].

In some sense it is a pity that this deliberately provocative and unrealistic lottery has played such an important role in the development of classical decision theory. It is quite unnecessary to invent a gamble with a diverging change in expected wealth to expose the flaws in the expected-wealth paradigm. The presence of infinities in the problem and its variously proposed solutions has caused much confusion, and permits objections on the grounds of physical impossibility. These objections don't much advance decision theory: they address only the gamble and not the decision paradigm. Nevertheless, the paradox is an indelible part not only of history but also of the current debate [**?**], and so we recount it here. We'll start by defining the lottery.

Example: St Petersburg lottery

Imagine a starting prize of \$1 (originally the prize was in ducats). A fair coin is tossed: if it lands heads, the player wins the prize and the lottery ends; if it lands tails, the prize is doubled and the process is repeated. Therefore, the player wins \$2, \$4, \$8 if the first head lands on the second, third, fourth toss, and so on. The player must buy a ticket, at price F, to enter the lottery. The question is: what is the largest F the player is willing to pay?

The lottery can be translated neatly into our gamble formalism:

$$q_j = \$2^{j-1} - F, \quad p_j = 2^{-j}, \tag{4.88}$$ {eq:lottery_def}

for $j \in \{1, 2, 3, \ldots\}$, *i.e.* the set of positive integers. The vast majority of observed payouts are small, but occasionally an extremely large payout (corresponding to a very long unbroken sequence of tails in the classical description) occurs. This is shown in the example trajectories in Fig. 4.3, where the lottery has been repeated additively.

[2] Daniel's cousin. The Bernoulli family produced a remarkable number of famous mathematicians in the 17th and 18th centuries, who helped lay the foundations of applied mathematics and physics.

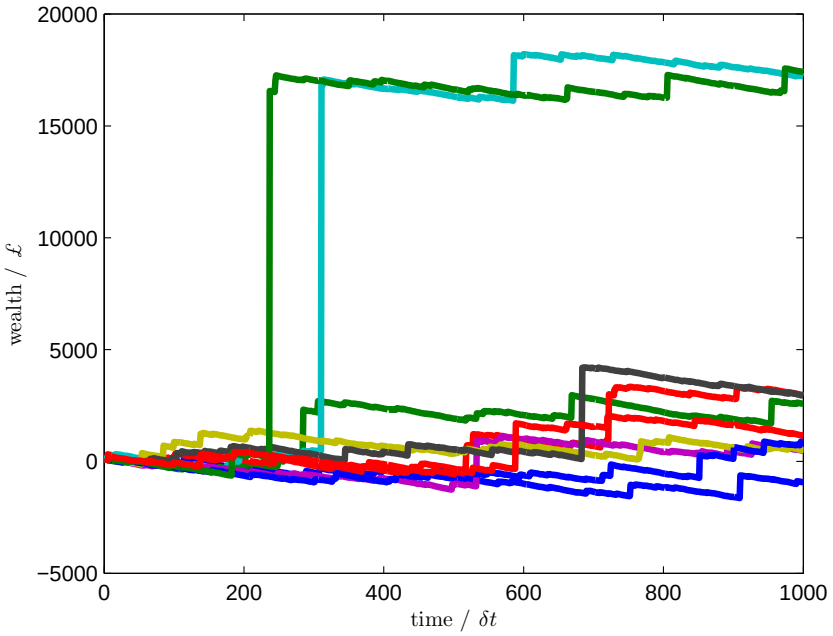


Figure 4.3: Wealth trajectories for the additively repeated St Petersburg lottery, with starting wealth, $x(0) = \$100$, and ticket price, $F = \$10$. Ten trajectories are plotted over 1,000 rounds. fig:lottery_add_traj

From now on we will forget about the coin tosses, which are simply a mechanism for selecting one of the possible payouts. They are nothing but an 18th-century random number generator. Instead we shall work with the compact definition of the lottery in (Eq. 4.87) and assume it takes a fixed amount of time, δt, to play.

The rate of change of expected wealth is

$$\frac{\langle \delta x \rangle}{\delta t} = \frac{1}{\delta t} \sum_{j=1}^{\infty} p_j q_j \tag{4.89}$$

$$= \frac{1}{\delta t} \left(\$ \sum_{j=1}^{\infty} 2^{-j}\, 2^{j-1} - \sum_{j=1}^{\infty} 2^{-j} F \right) \tag{4.90}$$

$$= \frac{1}{\delta t} \left(\$ \sum_{j=1}^{\infty} \frac{1}{2} - F \right). \tag{4.91}$$ {eq:lottery_ex_wealth}

This diverges for any finite ticket price. Under the expected-wealth paradigm, this means that the lottery is favourable at any price.

This implausible conclusion, which does not accord with human behaviour, exposes the weakness of judging a gamble by its effect on expected wealth. Daniel Bernoulli suggested to resolve the paradox by adopting the expected-utility paradigm. His choice of utility function was the logarithm, $u(x) = \ln x$, which, as we now know, produces a decision rule equivalent to growth-rate

optimisation under multiplicative repetition. This correspondence was not appreciated by Bernoulli: indeed 18$^{\text{th}}$-century mathematics did not possess the concepts and language required to distinguish between averages over time and across systems, even though it had the basic arithmetic tools.

Unfortunately, Bernoulli made a mathematical error in the implementation of his own paradigm – accidentally he proposed two mutually inconsistent versions of utility theory in the paper that established the paradigm. Initially, the error had little impact, and it was corrected by Laplace in 1814 [**?**]. But Laplace didn't openly say he'd corrected an error, he just worked with what he thought Bernoulli had meant. This politeness had awful consequences. In 1934 Menger [**?**], keen to get the story right, went back to the original text by Bernoulli. He didn't notice the error but rather got confused by it which led him to introduce a further error. Based on this car crash of scientific communication, Menger derived the infamous (wrong) claim we encountered in the history segment in Sec. **??**: utility functions must be bounded, with disastrous consequences for the budding neoclassical formalism. We will leave this most chequered part of the paradox's history alone – details can be found in [**?**, **?**]. Instead we will focus on what's usually presumed Bernoulli meant to write.

Example: Resolution by logarithmic utility

Instead of (Eq. 4.90), we calculate the rate of change of expected logarithmic utility,

$$\frac{\langle \delta \ln x \rangle}{\delta t} = \frac{1}{\delta t} \sum_{j=1}^{\infty} p_j \left[\ln(x + q_j) - \ln x \right] \tag{4.92}$$

$$= \frac{1}{\delta t} \sum_{j=1}^{\infty} 2^{-j} \ln \left(\frac{x + \$2^{j-1} - F}{x} \right), \tag{4.93}$$ {eq:lottery_ex_util}

where x is the ticket buyer's wealth.

This is finite for all finite ticket prices less than the buyer's wealth plus the smallest prize: $F < x + \$1$. This can be shown by applying the ratio test.[3] It may be positive or negative, depending on the values of F and x. Fig. 4.4 shows the locus of points in the (x, F)-plane for which the sum is zero.

The utility paradigm is a model that resolves the paradox, in the sense that creates a world where players may decline to buy a ticket. Bernoulli argued for this resolution framework in plausible terms: the usefulness of a monetary gain depends on how much money you already have. He also argued specifically for the logarithm in plausible terms: the gain in usefulness should be proportional to the fractional gain it represents, $\delta u = \delta x / x$. Yet, the framework has left many unsatisfied: why does usefulness have this functional form? We provide this deeper reason by connecting the problem to dynamics and time, unlike Bernoulli. Had Bernoulli made the connection, he might have been less willing to accept Cramer's square-root utility function as an alternative, which, as we've seen, corresponds to a rather less intuitive dynamic.

[3] The ratio of the $(j+1)^{\text{th}}$ term to the j^{th} term in the sum tends to $1/2$ as $j \to \infty$.

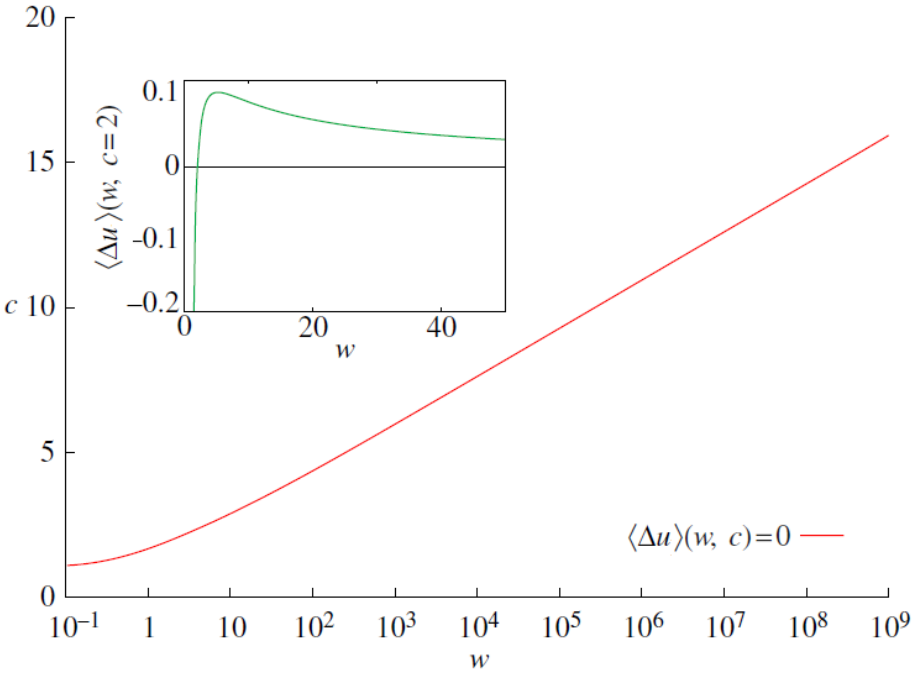


Figure 4.4: Locus of points in the (x, F)-plane for which the expected change in logarithmic utility is zero. The inset shows the expected change in utility as a function of x for $F = \$2$. Adapted from [**?**].fig:gbar_zero

Turning to our decision algorithm, we will assume that the lottery is assessed by the growth rate it would impart on the player were it repeated multiplicatively. This means, in effect, that the prizes and ticket price are treated as fractions of the player's wealth, such that the effect of each lottery is to multiply current wealth by a random factor,

$$r_j = \frac{x + \$2^{j-1} - F}{x}, \quad p_j = 2^{-j}. \tag{4.94}$$

This follows precisely our earlier treatment of a gamble with multiplicative dynamics, and we can apply our results directly. The time-average (exponential) growth rate is

$$\overline{g}_{\mathrm{m}} = \frac{1}{\delta t} \lim_{T\to\infty} \left\{ \frac{1}{T} \sum_{\tau=1}^{T} \ln r(\tau) \right\} = \frac{1}{\delta t} \sum_{j=1}^{\infty} 2^{-j} \ln r_j, \tag{4.95}$$ {eq:lottery_gbar}

which is identical to the expression for the rate of change of expected log-utility, (Eq. 4.92). This is, as we've discussed, because $v(x) = \ln(x)$ is the appropriate ergodicity mapping for multiplicative dynamics. The result is the same, but the interpretation is different: we have assumed less, only that our player is interested in the growth rate of his wealth and that he gauges this by imagining the outcome of an indefinite sequence of repeated lotteries.

Thus the locus in Fig. 4.4 also marks the decision threshold *versus* the null gamble under our decision axiom. The player can sensibly decline the gamble, even though it results in a divergent change in expected wealth. This is illustrated by comparing Fig. 4.5, which shows trajectories of multiplicatively repeated lotteries, with the additively repeated lotteries already seen in Fig. 4.3. The trajectories are based on the same sequences of lottery outcomes, only the mode of repetition is different. The simulation shows us visually what we have already gleaned by analysis: what appears favourable in the expected-wealth paradigm (corresponding to additive repetition) results in a disastrous decay of the player's wealth over time under a realistic dynamic.

As $F \to x + \$1$ from above in (Eq. 4.94), $\overline{g}_{\mathrm{m}}$ diverges negatively, since the first term in the sum is the logarithm of a quantity approaching zero. This corresponds to a lottery which can make the player bankrupt. The effect is also shown in the inset of Fig. 4.4.

Treatments based on multiplicative repetition have appeared sporadically in the literature, starting with Whitworth in 1870 [**?**, App. IV].[4] It is related to the famous Kelly Criterion [**?**][5], although Kelly did not explicitly treat the St Petersburg game, and tangentially to Itô's lemma [**?**]. It appears as an exercise

[4] Whitworth was dismissive of early utility theory: "The result at which we have arrived is not to be classed with the arbitrary methods which have been again and again propounded to evade the difficulty of the Petersburg problem. . . . Formulae have often been proposed, which have possessed the one virtue of presenting a finite result. . . but they have often had no intelligible basis to rest upon, or. . . sufficient care has not been taken to draw a distinguishing line between the significance of the result obtained, and the different result arrived at when the mathematical expectation is calculated." Sadly he chose to place these revolutionary remarks in an appendix of a college probability textbook.

[5] Kelly was similarly unimpressed with the mainstream and noted in his treatment of decision theory, which he developed from the perspective of information theory and which is identical to ergodicity economics with multiplicative dynamics, that the utility function is "too general to shed any light on the specific problems of communication theory."

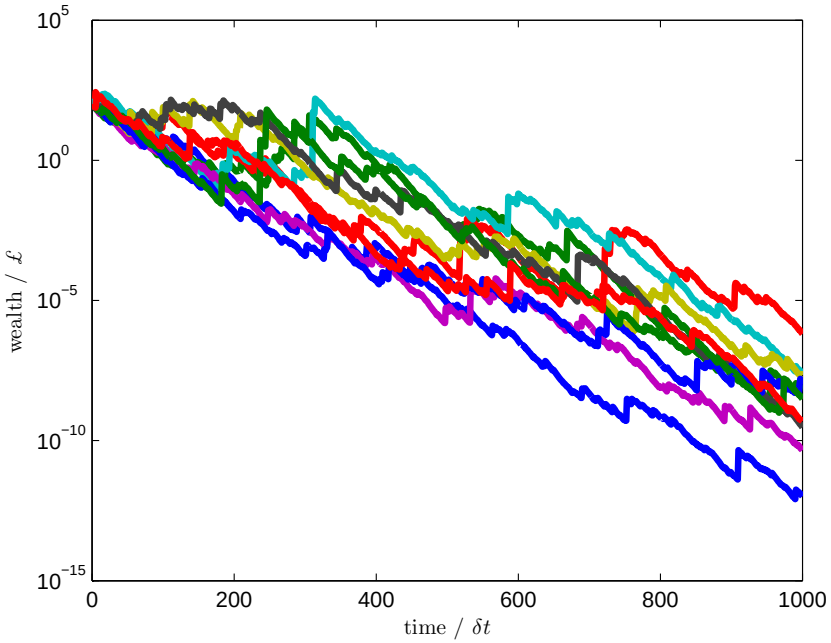


Figure 4.5: Wealth trajectories for the multiplicatively repeated St Petersburg lottery, with starting wealth, $x(0) = \$100$, and ticket price, $F = \$10$. Ten trajectories are plotted over 1,000 rounds. The realisations of the individual lotteries are the same as in Fig. 4.3 but the mode of repetition is different. fig:lottery_mult_traj

in a well-known text on information theory [**?**, Ex. 6.17]. Mainstream economics has ignored all this. A full and rigorous resolution of the paradox, including the epistemological significance of the shift from ensemble to time averages, was published recently by one of the present authors [**?**].

Chapter 5

Decisions in the real world {chapter:Real}

5.1 The Copenhagen experiment

5.2 Insurance {section:Insurance}

The insurance contract is an important and ubiquitous type of economic transaction, which can be modelled as a gamble. However, it poses a puzzle [**?**]. In the expected-wealth paradigm, insurance contracts shouldn't exist, because buying insurance would only be rational at a price at which it would be irrational to sell. More specifically:

1. To be viable, an insurer must charge an insurance premium of at least the expectation value of any claims that may be made against it, called the "net premium" [**?**, p. 1].

2. The insurance buyer therefore has to be willing to pay more than the net premium so that an insurance contract may be successfully signed.

3. Under the expected-wealth paradigm it is irrational to pay more than the net premium, and therefore insurance contracts should not exist.

In this picture, an insurance contract can only ever be beneficial to one party. It has the anti-symmetric property that the expectation value of one party's gain is the expectation value of the other party's loss.

The puzzle is that insurance contracts are observed to exist.[1] Why? Classical resolutions appeal to utility theory (*i.e.* psychology) and asymmetric information (*i.e.* deception). However, our decision theory naturally predicts contracts with a range of prices that increase the time-average growth rate for both buyer and seller. We illustrate this with an example drawn from maritime trade, in which the use of insurance has a very long history.[2] A similar example was used by Bernoulli [**?**].

[1] Something of an understatement. The Bank for International Settlements estimated the market value of all the world's derivatives contracts, which are essentially insurance contracts, as $15 trillion in the first half of 2015 (see http://www.bis.org/statistics/d5_1.pdf). That's six times the gross domestic product of the United Kingdom.

[2] Contracts between Babylonian traders and lenders were recorded around 1750 BC in the Code of Hammurabi. Chinese traders practised diversification by spreading cargoes across multiple vessels even earlier than this, in the third millennium BC.

Example: A shipping contract

We imagine a shipowner sending a cargo from St Petersburg to Amsterdam, with the following parameters:

- owner's wealth, $x_{\text{own}} = \$100,000$;
- gain on safe arrival of cargo, $G = \$4,000$;
- probability ship will be lost, $p = 0.05$;
- replacement cost of the ship, $C = \$30,000$; and
- voyage time, $\delta t = 1$ month.

An insurer with wealth $x_{\text{ins}} = \$1,000,000$ proposes to insure the voyage for a fee, $F = \$1,800$. If the ship is lost, the insurer pays the owner $L = G + C$ to make him good on the loss of his ship and the profit he would have made.

We phrase the decision the owner is facing as a choice between two gambles.

Definition **The owner's gambles**

Sending the ship uninsured corresponds to gamble o1

$$q_1^{(\text{o1})} = G, \qquad p_1^{(\text{o1})} = 1 - p; \tag{5.1}$$
$$q_2^{(\text{o1})} = -C, \qquad p_2^{(\text{o1})} = p. \tag{5.2}$$

Sending the ship fully insured corresponds to gamble o2

$$q_1^{(\text{o2})} = G - F \qquad p_1^{(\text{o2})} = 1. \tag{5.3}$$

This is a trivial "gamble" because all risk has been transferred to the insurer.

We also model the insurer's decision whether to offer the contract as a choice between two gambles

Definition **The insurer's gambles**

Not insuring the ship corresponds to gamble i1, which is the null gamble

$$q_1^{(\text{i1})} = 0 \qquad p_1^{(\text{i1})} = 1. \tag{5.4}$$

Insuring the ship corresponds to gamble i2

$$q_1^{(\text{i2})} = +F, \qquad p_1^{(\text{i2})} = 1 - p; \tag{5.5}$$
$$q_2^{(\text{i2})} = -L + F, \qquad p_2^{(\text{i2})} = p. \tag{5.6}$$

We ask whether the owner should sign the contract, and whether the insurer should have proposed it.

Example: Expected-wealth paradigm

In the expected-wealth paradigm (corresponding to additive repetition under the time paradigm) decision makers maximise the rate of change of

the expectation values of their wealths, according to (Eq. **??**): Under this paradigm the owner collapses gamble o1 into the scalar

$$\begin{aligned} \overline{g}_a^{(\mathrm{o1})} &= \frac{\langle \delta x \rangle}{\delta t} && (5.7)\\ &= \frac{\langle q^{(\mathrm{o1})} \rangle}{\delta t} && (5.8)\\ &= \frac{(1-p)G + p(-C)}{\delta t} && (5.9)\\ &= \$2,300 \text{ per month}, && (5.10) \end{aligned}$$

and gamble o2 into the scalar

$$\begin{aligned} \overline{g}_a^{\mathrm{o2}} &= \frac{\langle q^{(\mathrm{o2})} \rangle}{\delta t} && (5.11)\\ &= \frac{(G-F)}{\delta t} && (5.12)\\ &= \$2,200 \text{ per month}. && (5.13) \end{aligned}$$

The difference, $\delta\overline{g}_a^{\mathrm{o}}$, between the expected rates of change in wealth with and without a signed contract is the expected loss minus the fee per round trip,

$$\delta\overline{g}_a^{\mathrm{o}} = \overline{g}_a^{\mathrm{o2}} - \overline{g}_a^{\mathrm{o1}} = \frac{pL - F}{\delta t}. \qquad (5.14)$$ {eq:dro}

The sign of this difference indicates whether the insurance contract is beneficial to the owner. In the example this is not the case, $\delta\overline{g}_a^{\mathrm{o}} = -\$100$ per month.

The insurer evaluates the gambles i1 and i2 similarly, with the result

$$\overline{g}_a^{(\mathrm{i1})} = \$0 \text{ per month}, \qquad (5.15)$$

and

$$\begin{aligned} \overline{g}_a^{(\mathrm{i2})} &= \frac{F - pL}{\delta t} && (5.16)\\ &= \$100 \text{ per month}. && (5.17) \end{aligned}$$ {eq:r}

Again we compute the difference – the net benefit to the insurer that arises from signing the contract

$$\delta\overline{g}_a^{\mathrm{i}} = \overline{g}_a^{\mathrm{i2}} - \overline{g}_a^{\mathrm{i1}} = \frac{F - pL}{\delta t}. \qquad (5.18)$$ {eq:dri}

In the example this is $\delta\overline{g}_a^{\mathrm{i}} = \$100$ per month, meaning that in the world of the expected-wealth paradigm the insurer will offer the contract.

Because only one party (the insurer) is willing to sign, no contract will come into existence. We could think that we got the price wrong, and the contract would be sigend if offered at a different fee. But this is not the case, and that's the fundamental insurance puzzle: in the world created by expected-wealth

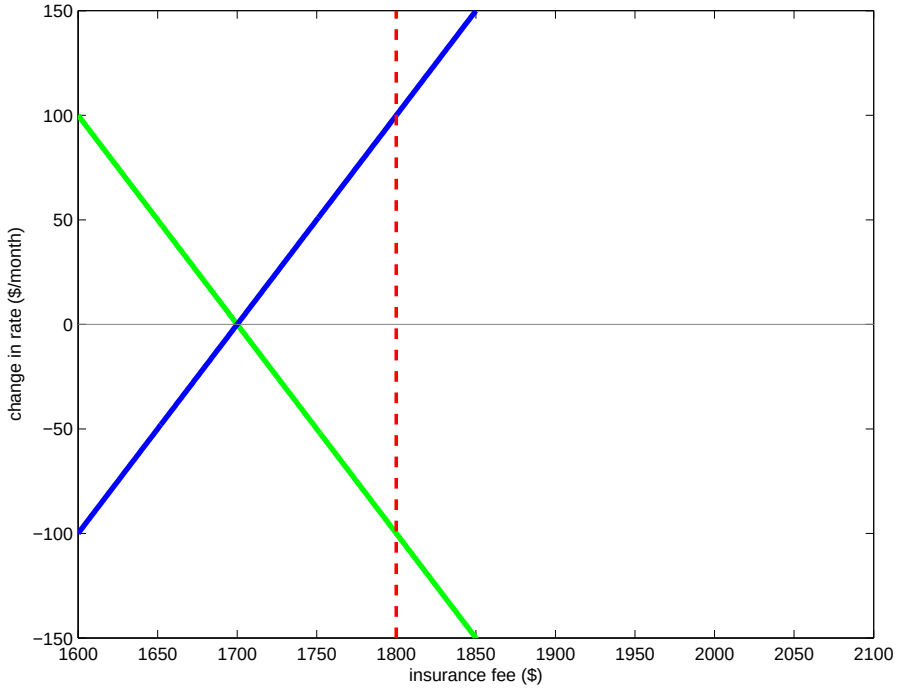


Figure 5.1: Change in the rate of change of expected wealth for the shipowner (green) and the insurer (blue) as a function of the insurance fee, F. fig:ins_lin

maximisation no price exists at which both parties will sign the contract.

Looking at (Eq. 5.14) and (Eq. 5.18) we notice the anti-symmetric relationship between the two expressions, $\delta\overline{g}_a^{\mathrm{o}} = -\delta\overline{g}_a^{\mathrm{i}}$. By symmetry, there can be no fee at which both expressions are positive. Hence there are no circumstances in the world created by the expected-wealth paradigm under which both parties will sign. Insurance contracts cannot exist in this world.

One party winning at the expense of the other makes insurance an unsavoury business in the expected-wealth paradigm. This is further illustrated in Fig. 5.1, which shows the change in the rate of change of expected wealth (the decision variable) for both parties as a function of the fee, F. There is no price at which the decision variable is positive for the both parties. The best they can do is to pick the price at which neither of them cares whether they sign or not.

In this picture, the existence of insurance contracts requires some asymmetry between the contracting parties, such as:

- different attitudes to bearing risk;
- different access to information about the voyage;
- different assessments of the riskiness of the voyage;
- one party to deceive, coerce, or gull the other into a bad decision.

It is difficult to believe that this is truly the basis for a market of the size and global reach of the insurance market.

5.2.1 Solution in the time paradigm

Example: Time paradigm

The insurance puzzle is resolved in the 'time paradigm', *i.e.* using the growth-optimal decision theory we have developed in this lecture and multiplicative repetition. Again, we pause to reflect what multiplicative repetition means compared to additive repetition. This is important because additive repetition is equivalent to the expected-wealth paradigm that created the insurance puzzle. Multiplicative repetition means that the ship owner sends out a ship and a cargo whose values are proportional to his wealth at the start of each voyage. A rich owner who has had many successful voyages will send out more cargo, a larger ship, or perhaps a *flotilla*, while an owner to whom the sea has been a cruel mistress will send out a small vessel until his luck changes. Under additive repetition, the ship owner would send out the same amount of cargo on each journey, irrespective of his wealth. Shipping companies of the size of Evergreen or Maersk would be inconceivable under additive repetition, where returns on successful investments are not reinvested.

The two parties seek to maximise

$$\overline{g}_m = \lim_{\Delta t\to\infty} \frac{\Delta v(x)}{\Delta t} = \frac{\langle \delta \ln x\rangle}{\delta t}, \tag{5.19}$$

where we have used the ergodic property of $\Delta v(x) = \Delta \ln x$ under multiplicative repetition.

The owner's time-average growth rate without insurance is

$$\overline{g}_m^{\mathrm{o1}} = \frac{(1-p)\ln(x_{\mathrm{own}}+G) + p\ln(x_{\mathrm{own}}-C) - \ln(x_{\mathrm{own}})}{\delta t} \tag{5.20}$$

or 1.9% per month. His time-average growth rate with insurance is

$$\overline{g}_m^{\mathrm{o2}} = \frac{\ln(x_{\mathrm{own}}+G-F) - \ln(x_{\mathrm{own}})}{\delta t} \tag{5.21}$$

or 2.2% per month. This gives a net benefit for the owner of

$$\delta\overline{g}_m^{o} = \overline{g}_m^{\mathrm{o1}} - \overline{g}_m^{\mathrm{o2}} \approx +0.24\% \text{ per month.} \tag{5.22}$$

The time paradigm thus creates a world where the owner will sign the contract.

What about the insurer? Without insurance, the insurer plays the null gamble, and

$$\overline{g}_m^{\mathrm{i1}} = \frac{0}{\delta t} \tag{5.23}$$

or 0% per month. His time-average growth rate with insurance is

$$\overline{g}_m^{\mathrm{i2}} = \frac{(1-p)\ln(x_{\mathrm{ins}}+F) + p\ln(x_{\mathrm{ins}}+F-L) - \ln(x_{\mathrm{ins}})}{\delta t} \tag{5.24}$$

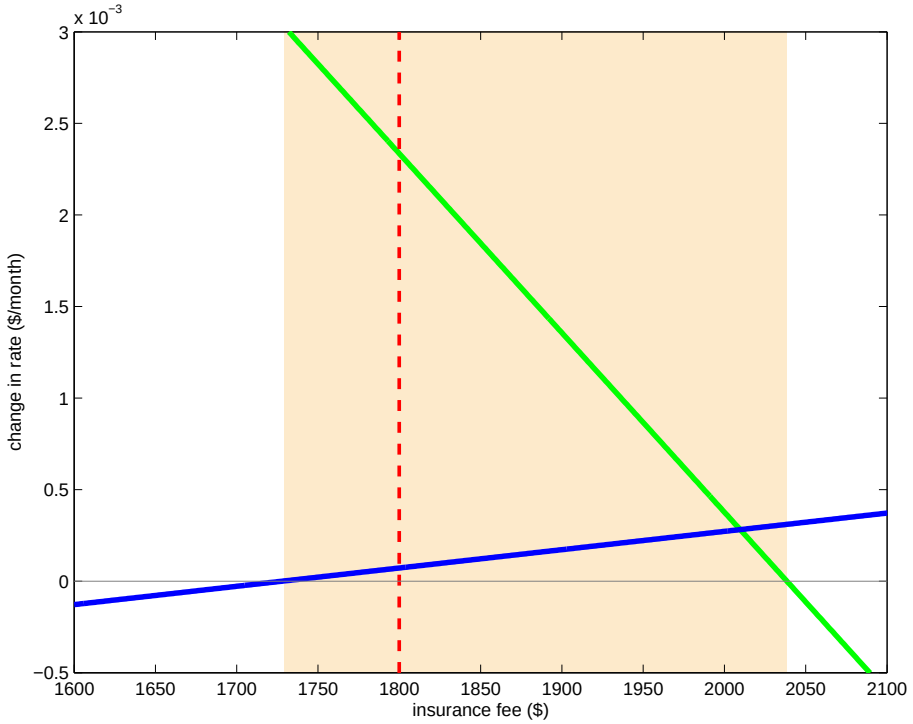


Figure 5.2: Change in the time-average growth rate of wealth for the shipowner (green) and the insurer (blue) as a function of the insurance fee, F. The mutually beneficial fee range is marked by the beige background. `fig:ins_log`

or 0.0071% per month. The net benefit to the insurer is therefore also

$$\delta\bar{g}^{\mathrm{i}}_m = \bar{g}^{\mathrm{i2}}_m - \bar{g}^{\mathrm{i1}}_m \tag{5.25}$$

i.e. 0.0071% per month. Unlike the expected wealth paradigm, the time paradigm with multiplicative repetition creates a world where an insurance contract can exist – there exists a range of fees F at which both parties gain from signing the contract!

We view this as the

Fundamental resolution of the insurance puzzle:

The buyer and seller of an insurance contract both sign when it increases the time-average growth rates of their wealths.

It requires no appeal to arbitrary utility functions or asymmetric circumstances, rather it arises naturally from the model of human decision-making that we have set out. Fig. 5.2 shows the mutually beneficial range of insurance fees predicted by our model. Generalizing, the message of the time paradigm is that business happens when both parties gain. In the world created by this model any agreement, any contract, any commercial interaction comes into existence because it is mutually beneficial.

5.2.2 The classical solution of the insurance puzzle

{section:The classical so

The classical solution of the insurance puzzle is identical to the classical solution of the St Petersburg paradox. Wealth is replaced by a non-linear utility function of wealth, which breaks the symmetry of the expected-wealth paradigm. While it is always true that $\delta \langle r \rangle_{\text{own}} = -\delta \langle r \rangle_{\text{ins}}$, the expected growth rates of non-linear utility functions don't share this anti-symmetry. A difference in the decision makers' wealths is sufficient, though often different utility functions are assumed for owner and insurer, which is a model that can create pretty much any behavior. The downside of a model with this ability is, of course, that it makes no predictions – nothing is ruled out, so the model cannot be falsified.

Acronyms

BM Brownian motion.

GBM geometric Brownian motion.

GDP gross domestic product.

LHS left-hand side.

LML London Mathematical Laboratory.

PDF probability density function.

RHS right-hand side.

SDE stochastic differential equation.

List of Symbols

A An observable.

a_v Itô coefficient function for ergodicity mapping $v(x)$.

a_x Itô coefficient function for wealth process x.

b_v Itô coefficient function for ergodicity mapping $v(x)$.

b_x Itô coefficient function for wealth process x.

C Replacement cost of a ship.

d Differential operator in Leibniz notation, infinitesimal.

δt A time interval corresponding to the duration of one round of a gamble or, mathematically, the period over which a single realisation of the constituent random variable of a discrete-time stochastic process is generated..

δ Most frequently used to express a difference, for instance δx is a difference between two wealths x. It can be the Kronecker delta function, a function of two arguments with properties $\delta(i,j) = 1$ if $i = j$ and $\delta(i,j) = 0$ otherwise. It can also be the Dirac delta function of one argument, $\int f(x)\delta(x - x_0)dx = f(x_0)$.

Δ Difference operator, for instance Δv is a difference of two values of v, for instance observed at two different times.

Δt A general time interval..

η Langevin noise with the properties $\langle \eta \rangle = 0$ and $\langle \eta(t_1)\eta(t_2) \rangle = \delta(t_1 - t_2)$.

f Generic function.

F Fee to be paid.

g Growth rate.

G Gain from one round trip of the ship.

$g_{\mathbf{e}}$ Ergodic growth rate for exponential growth.

$g_{\mathbf{a}}$ Ergodic growth rate under additive dynamics, *i.e.* rate of change, $g_{\mathrm{a}}(x; t, \Delta t) = \frac{\Delta x(t)}{\Delta t}$.

γ Parameter specifying the value of a time-average growth rate. This enables statements like $g_t ime = \gamma$, *i.e.* the time average growth rate take the value γ.

$g_{\mathbf{m}}$ Ergodic growth rate for multiplicative dynamics, *i.e.* exponential growth rate, $g_{\mathrm{m}}(x; t, \Delta t) = \frac{\Delta \ln x}{\Delta t}$.

$\overline{g}$ Time-average ergodic growth rate.

i Label for a particular realization of a random variable.

j Label of a particular outcome.

J Size of the jackpot.

k dummy.

K Number of possible values of a random variable.

L Insured loss.

m Index specifying a particular gamble.

μ Drift term in BM.

n n_j is the number of times outcome j is observed in an ensemble.

N Ensemble size, number of realizations.

$\mathcal{N}$ Normal distribution, $x \sim \mathcal{N}(\langle p \rangle, \mathrm{var}(p))$ means that the variable p is normally distributed with mean $\langle p \rangle$ and variance $\mathrm{var}(p)$.

o Little-o notation.

p Probability, p_j is the probability of observing event j in a realization of a random variable.

$\mathcal{P}$ Probability density function.

q Possible values of Q. We denote by q_i the value Q takes in the i^{th} realization, and by q_j the j^{th}-smallest possible value.

Q Random variable defining a gamble through additive wealth changes.

r Random factor whereby wealth changes in one round of a gamble.

$r_{\langle\rangle}$ Expectation value of growth factor r.

$\overline{r}$ Average growth factor over a long time.

s Dummy variable in an integration.

σ Magnitude of noise in a Brownian motion.

t Time.

T Number of sequential iterations of a gamble, so that $T\delta t$is the total duration of a repeated gamble.

t_0 Specific value of time t, usually the starting time of a gamble..

τ Dummy variable indicating a specific round in a gamble.

u Utility function.

v Stationarity mapping function, so that $v(x)$ has stationary increments.

var Variance.

W Wiener process, $W(t) = \int_0^t dW$ is continuous and $W(t) \sim \mathcal{N}(0, t)$.

x Wealth.

$\overline{x}$ Time-average of x. With subscript Δt, this is a finite-time average, without the subscript it refers to the infinite-time average.

x Deterministic wealth.

ξ A standard normal variable, $\xi \sim \mathcal{N}(0, 1)$.

Y Random variable that is a function of another random variable, Z.

z Generic value of a random variable.

Z Generic random variable.