ECO 586/PHY 560C

Modelling Financial Markets: an Introduction to Econophysics

Michael Benzaquen[1]

Ecole polytechnique – 3rd year

November 17, 2023

[1] www.econophysiX.com
michael.benzaquen@polytechnique.edu

Contents

Foreword

Ever since Bachelier's PhD thesis in 1900 [1], a theory of Brownian motion 5 years before Einstein, our understanding of financial markets has reasonably progressed. Over the past decades, financial engineering has grown tremendously and has regrettably outgrown our understanding. The inadequacy of the models used to describe financial markets is often responsible for the worst financial crises, with significant impact on everyday economy. From a physicist's perspective, understanding the price formation mechanisms – namely how markets absorb and process information of thousands of individual agents to come up to a "fair" price – is a truly fascinating and challenging problem. Fortunately, modern financial markets provide enormous amounts of data that can now be used to test scientific theories at levels of precision comparable to those achieved in physical sciences.

This course presents the approach adopted by physicists to analyse and model financial markets (see e.g. [2–5]). Our analysis shall, insofar as this is possible, always be grounded on the real financial data.[1] Rather than sticking to a rigorous mathematical formalism, we will seek to foster critical thinking and develop intuition on the "mechanics" of financial markets, the orders of magnitude, and certain open problems. By the end of this course, students should be able to:

- Answer simple questions on the subject of this course (without course notes).
- Rigorously analyse real financial data and empirical results.
- Conduct logical reasoning to answer modelling problems (ABM).
- Carry out simple calculations similar to those presented in class and during the tutorials (PDEs, Langevin equations, dimensional analysis, etc.).

The content of this course is largely inspired by references [2, 3], together with my own experience in quantitative finance as a physicist. Students interested in going the extra mile are encouraged to consult references [6–14].

[1] The tutorials (on Python) will focus on the analysis of real financial data and numerical simulations of some of the models presented in the course.

1

Empirical time series

In this Chapter we introduce some important ideas and methods for the description of one-dimensional time series and the analysis of empirical data.

1.1 A few examples

Time series can be really anything, a few examples are:

- The position of a random walker as function of time,
- The price of financial asset as function of time,
- The velocity of a turbulent flow at a certain point in space as function of time,
- The temperature in Paris as function of time,
- The seismic intensity in Lyon as function of time.

Throughout this course we shall restrict to the analysis of one-dimensional time series, denoted $x(t)$, or equivalently $x_t \in \mathbb{R}$.

1.2 Variogram

Most of the time, we shall also restrict to the analysis of *stationary processes*.

Stationary process $\Leftrightarrow$ The statistics of $\Delta_\tau x := x(t+\tau)-x(t)$ is independent of t $\Leftrightarrow$ Time translation symmetry.

Most of the time we shall consider *detrended* time series, that is $\langle \Delta_\tau x\rangle = 0$. The most natural quantity that comes to mind to characterise by how much $x(t)$ varies over a given timescale τ is the standard deviation:

$$\sigma(\tau) := \sqrt{\mathcal{V}(\tau)}\,, \quad \text{with} \quad \mathcal{V}(\tau) := \left\langle (\Delta_\tau x)^2 \right\rangle . \tag{1.1}$$

The plot of $\mathcal{V}(\tau)$ against τ is known as the *variogram*. Frequently, one rather refers to the *signature plot*, which is the plot of $\mathcal{V}(\tau)/\tau$ against τ.

Let us now consider a few illustrative examples.

- Brownian motion (normal diffusion): $\mathcal{V}(\tau) = D\tau$ with D the *diffusivity* parameter.
- Fractional Brownian motion or "fBm" (anomalous diffusion): $\mathcal{V}(\tau) \sim \tau^{2H}$ with H the *Hurst exponent*. Normal diffusion is recovered for $H = 1/2$. For $H > 1/2$, one speaks of *persistent processes* or *superdiffusion* (e.g. atmospheric mixing). For $H < 1/2$, one has *subdiffusion*, or *antipersistent processes*, or *mean-reverting processes* (e.g. intracellular transport, $H = 1/3$ velocity fluctuation in turbulent flows).

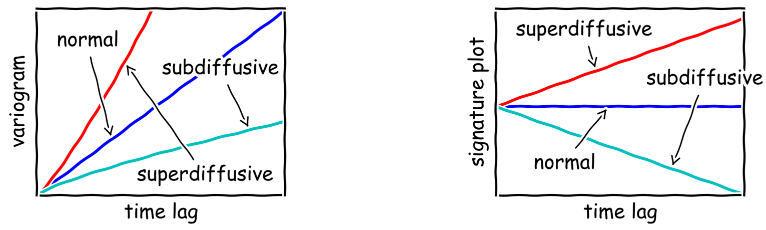


Figure 1.1: Variogram and signature plot.

- Levy flights: $\mathscr{V}(\tau) = \infty$ (e.g. ocean predators' hunting patterns).
- Ornstein-Uhlenbeck process: $\dot{x}_t = -\omega x_t + \eta_t$ with $x_0 = 0$, $\omega > 0$ and η_t a Langevin noise $\langle \eta_t \rangle = 0$, $\langle \eta_t \eta_{t'} \rangle = 2\sigma_0^2 \delta(t - t')$. The process is stationary for $t \gg \omega^{-1}$ and:[1]

$$\mathscr{V}_{\mathrm{OU}}(\tau) = \frac{\sigma_0^2}{\omega}\left(1 - e^{-\omega|\tau|}\right). \tag{1.2}$$

  Typical examples of OU processes are the velocity of a large Brownian particle in a viscous fluid or interest rates, see *Vasicek interest rate model* [15].

We also define the *correlogram* $\mathscr{C}(\tau) := \langle x_{t+\tau} x_t \rangle - \langle x_t \rangle^2$, which can be expressed as a function of the variogram.

In most cases, the variogram does not suffice to fully characterise the time series. One should compute the whole probability distribution $P_\tau(\Delta_\tau x)$, or equivalently all of its moments $\mathscr{M}_q(\tau) = \left\langle (\Delta_\tau x)^q \right\rangle$. There is however a remarkable exception: scale invariant processes.

1.3 Scale invariance

Process x_t is said to be *scale invariant* if and only if there exists a function f such that:

$$P_\tau(\Delta_\tau x) = \frac{1}{\sigma(\tau)} f\left(\frac{\Delta_\tau x}{\sigma(\tau)}\right). \tag{1.3}$$

Fractional Brownian motions are a good example of scale invariant processes, in particular one has $f_{\mathrm{fBm}}(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ for all H. Levy flights are also scale invariant.

If x_t is a scale invariant process, then all the relevant information is contained in $\sigma(\tau)$. Indeed, all the moments $\mathscr{M}_q(\tau)$ are proportional to the q^{th} power of $\sigma(\tau)$:[2]

$$\mathscr{M}_q(\tau) = \left[\sigma(\tau)\right]^q \int du\, u^q f(u)\,. \tag{1.4}$$

In addition to bringing substantial mathematical simplifications, scale invariance often bears witness of a physical property of the system (see Chapter 5).

[1]To compute this result one may proceed in discrete time: $x_{t+1} = (1-\omega)x_t + \eta_t \Rightarrow x_t - x_0 = \sum_{t'=0}^{t-1}(1-\omega)^{t-t'-1}\eta_{t'}$, and then using that $\langle \eta_{t'}\eta_{t''} \rangle = 2\sigma_0^2\delta(t'-t'')$, one can compute $\langle (x_\tau - x_0)^2 \rangle = 2\sigma_0^2 \sum_{t'=0}^{\tau-1}[(1-\omega)^2]^{\tau-t'-1} = 2\sigma_0^2 \frac{1-(1-\omega)^\tau}{1-(1-\omega)^2}$, which for $\omega \ll 1$ (continuous limit) amounts to Eq. (1.2).

[2]Using Eq. (1.3) one has $\mathscr{M}_q(\tau) = \int du\, u^q P_\tau(u) = \int du\, u^q \frac{1}{\sigma} f\left(\frac{u}{\sigma}\right) = \sigma^q \int dv\, v^q f(v)$.

1.4 Intermittency

In many situations, higher order moments carry essential information. Indeed, while the variogram of prices returns x_t in financial markets is remarkably linear, financial time series are quite far from Bachelier's random walks, or even Levy flights. Indeed, strong long range correlations appear at the level of x_t^2 which measures the activity or *volatility* of the market. Very much like in fluid turbulence, one observes *intermittency* or *volatility clustering*, that is calm periods interspersed with more agitated episodes of all sizes. While the correlogram of price changes doesn't reveal such effects, the correlogram of squared returns displays very long range correlations:

$$\langle x_t^2 x_{t+\tau}^2\rangle - \langle x_t^2\rangle^2 \sim \tau^{-\nu}\,, \quad \text{with } \nu \approx 0.2\,. \tag{1.5}$$

As a results, moments no longer trivially scale as in Eq. (1.4). One speaks of a *multifractal* time series when:

$$\mathcal{M}_q(\tau) \sim \big[\sigma(\tau)\big]^{\xi(q)\neq q}\,. \tag{1.6}$$

A rather good model for both fluid turbulence and finance is the so-called *log-normal cascades* [16, 17]. In a nutshell, one has $\xi(q) = q + \lambda^2 q(q-2)$ where λ is coined the *intermittency* parameter. The scale invariant or *monofractal* case corresponds to $\lambda = 0$.

When the variance $\langle x_t^2\rangle$ is itself a stochastic process, one speaks of *Heteroskedasticity*, see Chapter 2.

1.5 Skewness and kurtosis

Skewness and *kurtosis* are commonly used to further describe the shape of a probability distribution. The skewness ζ is a measure of the asymmetry of a probability distribution. The kurtosis κ is a measure of the "tailedness" of a probability distribution. They are given by the 3rd and 4th standardised moments:

$$\zeta := \left\langle \left(\frac{x-\langle x\rangle}{\sigma}\right)^3 \right\rangle = \frac{\mathcal{M}_3^c}{\sigma^3}\,, \qquad \kappa := \left\langle \left(\frac{x-\langle x\rangle}{\sigma}\right)^4 \right\rangle - 3 = \frac{\mathcal{M}_4^c}{\sigma^4} - 3\,, \tag{1.7}$$

with $\mathcal{M}_3^c$ and $\mathcal{M}_4^c$ the 3rd and 4th central moments respectively.[3] The Gaussian distribution has $\kappa = \zeta = 0$, and more generally all cumulants of higher order are identically zero. One speaks of *negative skew* ($\zeta < 0$) when the left tail is longer, and *positive skew* ($\zeta > 0$) when the right tail is longer, see Fig 1.2. A distribution with $\kappa = 0$ is said to be *mesokurtic* while $\kappa > 0$ (resp. < 0) is refered to as *leptokurtic* (resp. *platykurtic*). A leptokurtic (resp. platykurtic) distribution has fatter (resp. thinner) tails than the Gaussian.

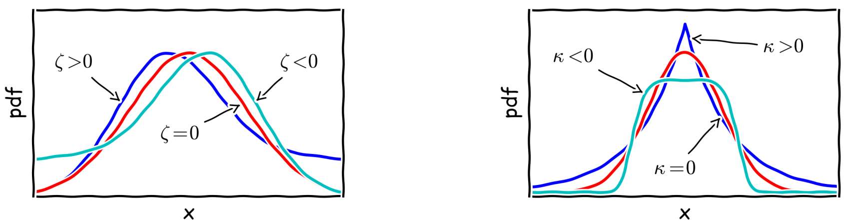


Figure 1.2: Skewness and kurtosis.

[3]Note that sometimes κ, as defined in Eq. (1.7), is called *excess kurtosis* while $\kappa+3$ is called kurtosis.

Conclusions

While the variogram is often the first quantity one will consider to analyse empirical time series, one should bear in mind that the mean and standard deviation of a time series are most often not the whole story. In certain cases, mean and variance may even not be defined (e.g. Levy flights); and yet, empirically one can always compute a mean and a variance, only the latter will be completely irrelevant and reflect boundary effects only. To avoid this, one should always compute the whole probability distribution, and more generally look at the time series directly! Intermittency or Levy flights are generally visible to the naked eye (see Fig. 1.3). Levy flights resemble a Brownian motion with occasional large jumps.

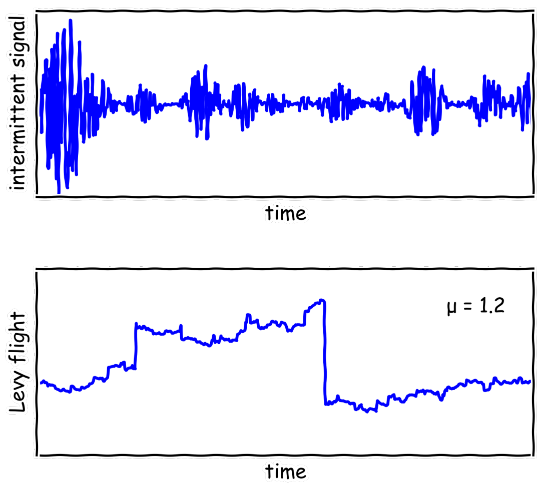


Figure 1.3: Intermittent or *heteroskedastic* signal (top), and Levy stable random walk (bottom).

2

Statistics of real prices

In this Chapter we present some important features and stylised facts on financial time series.

2.1 Bachelier's Brownian motion

Bachelier's thesis "Théorie de la spéculation" (1900) is often considered as the first serious attempt to account for price dynamics. To note, it is a theory of Brownian Motion 5 years before Einstein's [18]. The idea of Bachelier is as follows:

1. Each transaction involves a buyer and a seller, which means that there must be as many people who think the price will rise as people who think it shall decline. Therefore, price changes are unpredictable, or in the modern language, price are *Martingales*.[1]

2. Further, if one considers that price returns at a given timescale, say daily, are non other than the sum of a large number N of small price changes,

   $$p_N - p_0 = \sum_{t=0}^{N-1} r_t\,, \quad \text{with} \quad r_t = p_{t+1} - p_t\ ,$$

   then, for large N, the Central Limit Theorem (CLT) ensures that daily price changes are Gaussian random variables, and that prices thus follow Gaussian random walks.

While his first conclusion is rather accurate, the second is quite wrong as we shall see below. Note however that such reasoning is quite remarkable for that time! Be that as it may, on such grounds, Bachelier derives a series of very interesting results such as *Bachelier first law* which states that the price variogram grows linearly with time lag τ:

$$\mathscr{V}(\tau) := \langle (p_{t+\tau} - p_t)^2 \rangle = D\tau\,, \tag{2.1}$$

but also results on first passage times[2] and option pricing (precursor of Black-Scholes).

2.2 Central Limit Theorem and rare events

Let us now discuss further point 2. For the CLT to be applicable, several requirements need to be met. The returns r_t must be independent and identically distributed (iid) and have a finite variance $\sigma^2 < \infty$. While, as we shall see below, returns are not exactly iid, this is not the most problematic point in Bachelier's reasoning.[3] In most markets (yet not all) return do have a finite variance, but it

[1] A Martingale is a stochastic process x_t satisfying $\langle x_{t+s} | \{x_0, x_1, \ldots, x_t\} \rangle = \langle x_{t+s} | x_t \rangle = x_t$.

[2] Calling τ_1 the first passage time, one can show that for a Gaussian random walk the probability distribution of τ_1 is given by:

$$\mathbb{P}(\tau_1) = \frac{|\Delta x|}{\sqrt{4\pi D \tau_1^3}} \exp\left(-\frac{\Delta x^2}{4D\tau_1}\right) \sim \tau_1^{-3/2}\,,$$

such that the average first passage time diverges $\langle \tau_1 \rangle = \infty$. The *typical* first passage time $(\partial_{\tau_1} \mathbb{P}|_{\tau_{\text{typ}}} = 0)$ reads $\tau_{\text{typ}} = \Delta x / 6D$.

[3] Actually, in several cases the CLT holds much beyond the iid case.

should be noted that the CLT also applies beyond this constraint, only the aggregate distribution no longer converges to a Gaussian but to a *Levy stable law*.

Most importantly, for the error to be negligible everywhere one needs $N \to \infty$, or equivalently here, continuous time.[4] This is never the case in real life, and thus empirically the CLT only applies to a central region of width w_N, and nothing can be said for the tails of the distribution beyond this region (see Fig. 2.1). If the return distribution is power law, say $\rho(r) \sim 1/|r|^{1+\mu}$ with $\mu > 2$ such that the variance is still finite, the width of the central region scales as $w_N \sim \sqrt{N \log N}$ which is only $\sqrt{\log N}$ times wider than the natural width $\sigma\sqrt{N}$. The probability to fall in the tail region decays slowly as $1/N^{\mu/2-1}$. In fact the tail behaviour of the aggregate distribution is the very same power-law as as that of $\rho(r)$. In other words, far away from the central Gaussian region, the power-law tail survives even when N is very large.

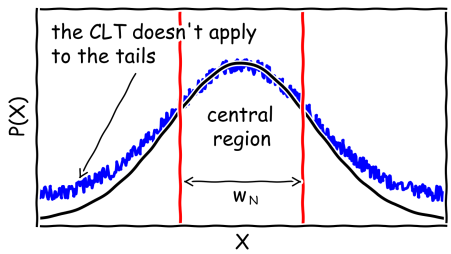


Figure 2.1: Distribution of aggregate returns ($N < \infty$).

One should thus carefully refrain from invoking the central limit theorem to describe the probability of extreme events – in most cases, the error made in estimating their probabilities is orders of magnitudes large.

2.3 Absolute or relative price changes

Are prices multiplicative or additive? Or in other words, are price changes proportional to prices? In most cases, the order of magnitude of relative fluctuations (1-2% per day for stocks) are much more stable in time and across assets than absolute price changes.[5] This is in favour of a *multiplicative* price process with relative returns, or equivalently an *additive* log-price process:

$$x_t := \frac{p_{t+1} - p_t}{p_t} \approx \log p_{t+1} - \log p_t \,. \tag{2.2}$$

In addition, the price of stocks is rather arbitrary, as it depends on the number of stocks in circulation and one may very well decide to split each share into n, thereby dividing their price by n, without *a priori* changing any fundamental properties (*splitting invariance* or *dilation symmetry*). Another symmetry exists in foreign exchange (FX) markets. The idea is that there should be no distinction between using the price π of currency A in units of B, or $1/\pi$ for currency B in units of A. Relative returns satisfy such a property: $x = \delta p/p = -\delta(1/p)/(1/p)$.

Notwithstanding, there are arguments in favour of an additive price process:

$$r_t := p_{t+1} - p_t \,. \tag{2.3}$$

Indeed, the fact that price are discrete and quoted in *ticks* which are fixed fractions of dollars (e.g. 0.01\$ for US Stocks) introduces a well-defined \$-scale for price changes and breaks the dilation symmetry.

[4] Considering Gaussian iid returns in continuous time, one is left with the standard *geometric Brownian motion* model, well established in mathematical finance since the 1960's.

[5] On a practical note, relative price changes are also more convenient since asset prices can span a few \$ to a few M\$.

Other examples in favour of an additive price process are some contract for which the dilation argument does not work, such as volatility, which is traded on option markets, and for which there is no reason why volatility changes should be proportional to volatility itself.

2.4 Typical drawdown

Average effects are in general small compared to fluctuations. Taking typical orders of magnitude, on average, individual stocks grow by $m \approx 5\%$ per year, and fluctuate by $\approx 1.5\%$ per day, that is $\sigma = 1.5 \times \sqrt{250} \approx 25\%$ per year.[6] Further, whereas average return increases linearly with time, fluctuations increase as the square root of time, see Fig. 2.2.

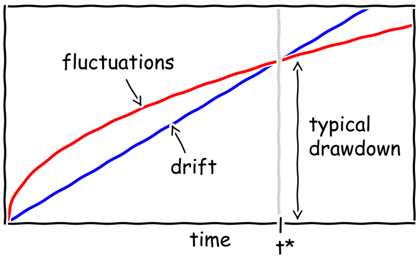


Figure 2.2: Typical drawdown in a Gaussian world.

The typical timescale below which fluctuations dominate drift is given by $t^\star$ such that $mt^\star = \sigma\sqrt{t^\star}$, that is $t^\star = \sigma^2/m^2 \approx 25$ years. Needless to say, very few investors are ready to hold their position for 25 years, but even it were the case, the *typical drawdown*, conventionally defined as $\Delta^\star := \sigma\sqrt{t^\star} = \sigma^2/m \approx 125\%$ illustrating that, even if markets were Gaussian, it is highly probable to loose a lot with such an investment.[7] Further, it illustrates that fluctuations are more important to understand than long term drift (a reduction of the risk by factor 2 leads to a reduction of the typical drawdown by a factor 4, whereas an increase of the expected gain by a factor 2 decreases the typical drawdown by a factor 2 only).

2.5 Fat tails

In contradiction with textbook mathematical finance, the real price statistics of any kind of financial asset (stocks, futures, currencies, interest rates, commodities etc.) are very far from Gaussian. Instead:

- The unconditional distributions of returns have fat power law tails. Recall that power law functions are scale invariant, which here corresponds to micro-crashes of all sizes.

- The empirical probability distribution function of returns on short to medium timescales (from a few minutes to a few days) is best fitted by a symmetric *Student t-distribution*, or simply the *t-distribution*:

$$\mathbb{P}(x) := \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{1+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)} \frac{a^\mu}{(x^2+a^2)^{\frac{1+\mu}{2}}} \sim \frac{1}{|x|^{1+\mu}}\,, \tag{2.4}$$

  with typically $3 < \mu < 5$. Its variance, given by $\sigma^2 = \frac{a^2}{\mu-2}$, diverges as $\mu \downarrow 2$ from above.

- On longer timescales (months, years) the returns distribution becomes quite asymmetric. While the CLT starts to kick (very slowly, see Section 2.2) for the positive tail, the negative remains very fat. In other words, downward price jumps are on average larger than their upward counterparts.

[6] Excluding week-ends and public holidays, there are ≈ 250 trading days per year.
[7] For real (non-Gaussian) markets, it is even worse. An obvious improvement is diversification.

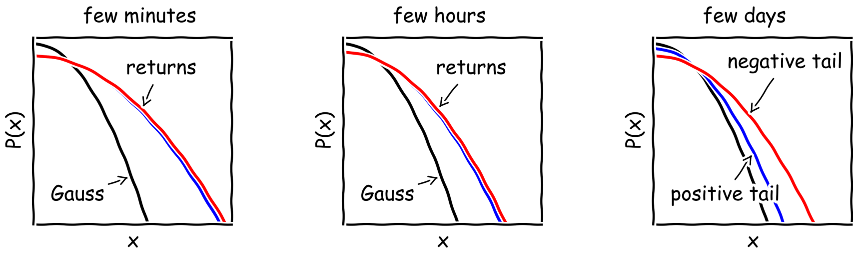


Figure 2.3: Distribution of returns on different timescales (log-scale).

- In extreme markets, one can have $\mu < 3$ or even $\mu < 2$ (e.g. MXN/\$ rate) such that $\sigma = \infty$!

The daily returns of the MXN/\$ rate are actually very well fitted by a pure Levy distribution with no obvious truncation ($\mu \approx 1.4$). The case of short term interest rates is also of interest. The latter (say 3-month rates) are strongly correlated to the decision of central banks to increase or decrease the day-to-day rate; kurtosis is rather high as a result of the short rate often not changing at all but sometimes changing a lot.

2.6 Heteroskedasticity and volatility dynamics

As alluded to above and shortly discussed in Chapter 1, relative price returns x_t are not exactly iid. While they are indeed quasi-uncorrelated:

$$\langle x_t x_{t'} \rangle \approx \sigma^2 \delta(t - t'), \tag{2.5}$$

for timescales above a few minutes and below a few days,[8] or else statistical arbitrages would be possible, one observes activity *intermittency* (see Chapter 1).

Actually, the volatility is itself a dynamic variable evolving at short and long timescales (multiscale). One says that price returns are *heteroskedastic* random variables, from ancient Greek *hetero*: different, and *skedasis*: dispersion. A common model is given by:

$$x_t = \sigma_t \xi_t, \tag{2.6}$$

where the ξ_t are centred iid random variables with unit variance encoding sign of returns and unpredictability $\langle \xi \rangle$=0, while σ_t is a positive random variable with fast and slow components, see Eq. (1.5). The squared volatility variogram is given by:

$$\mathcal{V}_{\sigma^2}(\tau) := \left\langle (\sigma^2_{t+\tau} - \sigma^2_t)^2 \right\rangle \approx A - B\tau^{-\nu}, \quad \text{with } \nu \approx 0.2. \tag{2.7}$$

To validate such a scaling, one would need $1/\nu \approx 5$ decades of data which is inaccessible. Actually, it is difficult to be sure that $\mathcal{V}_{\sigma^2}(\tau)$ converges to a finite value A at all. Multifractal models suggest instead:

$$\mathcal{V}_{\log \sigma}(\tau) := \left\langle (\log \sigma_{t+\tau} - \log \sigma_t)^2 \right\rangle \approx \lambda^2 \log \tau. \tag{2.8}$$

Volatility appears to be marginally stationary as its long term average can hardly be defined. The very nature of the volatility process is still a matter of debate (see [19] for recent insights) highlighting the complexity of price change statistics.

[8] At very high frequency price are mechanically mean reverting. At long timescales systematic inefficiencies exist (trend, value).

2.7 Volatility fluctuations and kurtosis

Heteroskedasticity actually has direct consequences on the return distribution. Denoting $g(\tau)$ the autocorrelation function of squared volatility,

$$g(\tau) := \frac{\left\langle \sigma_{t+\tau}^2 \sigma_t^2 \right\rangle - \langle \sigma_t^2 \rangle^2}{\langle \sigma_t^2 \rangle^2} \sim \tau^{-\nu},$$

one can show that the kurtosis κ_N of the aggregate return $X_N = \sum_{t=0}^{N-1} x_t = \sum_{t=0}^{N-1} \sigma_t \xi_t$ writes:

$$\kappa_N = \frac{1}{N}\left[\kappa_\xi + (3+\kappa_\xi) g(0) + 6 \sum_{\tau=1}^{N} \left(1 - \tfrac{\tau}{N}\right) g(\tau)\right], \tag{2.9}$$

with κ_ξ the kurtosis of ξ_t. Interestingly enough $\kappa_1 = \kappa_\xi + (3+\kappa_\xi)g(0) > \kappa_\xi$ which mean that – even within a Gaussian model ($\kappa_\xi = 0$) – a fluctuating volatility suffices to create a little kurtosis. For large N, the CLT kicks in and κ_N decays to zero, but it does so extremely slowly as $\kappa_N \sim \kappa_1 / N^\nu$.

2.8 Leverage effect and skewness

Financial time series break *Time Reversal Symmetry* (TRS). One can show that ξ_t and σ_t are not independent and in particular:

- Negative past returns tend to increase future volatility,
- Positive past returns tend to lower future volatility,
- Past volatility is not informative of the sign of future returns (or else there would be trivial profitable statistical arbitrage strategies).

This is the so-called *leverage effect*. Consistently, the response function $\langle \xi_t \sigma_{t+\tau} \rangle$ is negative for $\tau > 0$, and $= 0$ for $\tau < 0$, see Fig.2.4.

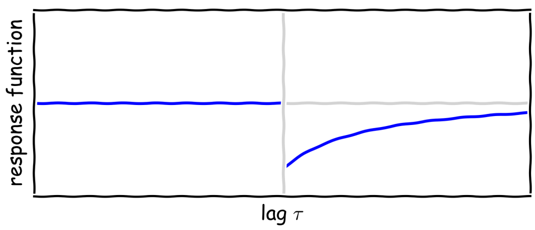


Figure 2.4: Leverage effet. Plot of $\langle \xi_t \sigma_{t+\tau} \rangle$ as function of the lag τ.

The leverage effect has direct implications on the skewness of the return distribution. Analogous to Eq. (2.9), one can show that the skewness of the aggregate return writes:

$$\zeta_N = \frac{1}{\sqrt{N}}\left[\zeta_\xi h(0) + 3 \sum_{\tau=1}^{N} \left(1 - \tfrac{\tau}{N}\right) h(\tau)\right], \tag{2.10}$$

where $h(\tau) = \left\langle x_t \sigma_{t+\tau}^2 \right\rangle \leq 0$ is the return-volatility response function. Here again, even within a Gaussian model ($\zeta_\xi = 0$), the leverage effect suffices to create a little negative skewness, consistent with the empirical return distributions, see Section 2.5. Actually one can show that ζ_N actually increases with N and reached a maximum value at the typical timescale of the leverage effect before the CLT kicks in.

Conclusion

Real financial time series display a number of properties not accounted for within the geometric (continuous time) Brownian motion standard framework. Accounting for of all these effects is of outmost importance for risk control and derivatives pricing. Different assets differ in the value of their higher cumulants (skewness, kurtosis); for this reason a description where the volatility is the only parameter is bound to miss a great deal of reality.

3

Why do prices change?

In this Chapter we confront two rather opposing views of prices in financial markets: *price discovery* and *price formation*.

3.1 The Efficient Market Hypothesis

According to Eugene Fama's *Efficient Market Hypothesis* (EMH) or *Efficient Market Theory* (EMT), asset prices reflect all available information. This is the classical and dominant view (since the mid-1960s) which can be coined *price discovery*, and which, as we shall see below, is poorly supported empirically. It further relies on the idea that markets are at equilibrium (demand and supply are balanced), agents have rational expectations etc.

> *"I can't figure out why anyone invests in active management [...]. Since I think everything is appropriately priced, my advice would be to avoid high fees. So you can forget about hedge funds."*
>
> – Eugene Fama

The market is seen as an objective measuring instrument which provides a reliable assessment p_t of the *fundamental value* v_t of the exchanged assets.[1] Consistently, in most Economics 101 textbooks one shall find the following equation:

$$p_t = \mathbb{E}[v_t|\mathscr{F}_t], \tag{3.1}$$

with $\mathscr{F}_t$ the *common knowledge*.[2] Immediate consequences of the EMH are as follows.

- Prices can only change with the arrival of new exogenous information (e.g. new iPhone release, discovery of a new gold mine, diplomatic crisis). As a results, price moves are unpredictable because news are, by definition, unpredictable. While consistent with Bachelier's findings discussed in Chapter 2, nothing says that the EMH it is the only possible explanation.
- Large price moves should be due to important news that change significantly the fundamental value of the asset. Crashes must be exogenous.
- Markets are fundamentally efficient. Small mispricings are rapidly corrected by "those who know the fundamental price" (whoever that is).

> *"Professor Fama is the father of modern efficient-markets theory, which says financial prices efficiently incorporate all available information and are in that sense perfect. In contrast, I have argued that the theory makes little sense, except in fairly trivial ways. Of course, prices reflect available information. But they are far from perfect. [...] I emphasise the enormous role played in markets by human error."*
>
> – Robert Shiller [3]

[1] The fundamental value or *intrinsic value* is, according to Wikipedia, the "true, inherent, and essential value" of a financial asset. Other definitions vary with the asset class, but clearly, it is a very ill-defined concept.

[2] Clearly, also a very ill-defined concept.

[3] Fama and Shiller shared the 2013 Nobel prize in Economics...

3.2 Empirical data and "puzzles"

A liquid stock counts 10^5 to 10^6 trades a day (over 1000 trades a minute). Clearly, news feeds frequency is much lower. So, if prices really reflect value and are unpredictable, why do people trade so much? This has been coined the *excess trading puzzle*. In the same vein, the price moves frequency is too high to be explained by fluctuations of fundamentals (news arrival rate is much lower). This is the *excess volatility puzzle*.

The latter puzzle suggests that a significant fraction of the volatility is of endogenous nature, in contradiction with Fama's theory. To a physicist, nontrivial endogenous dynamics is a natural feature of a complex system made of a large number of interacting agents, very much like a bird flock or a fish school. Imitation and feedback loops induce instabilities and intricate behavior consistent with the statistical anomalies described in Chapter 2.

Empirical data actually suggests that over 90% of the volatility is of endogenous nature. Indeed, restricting to large price jumps ($> 4\sigma$) and using a large news database, Joulin *et al.* [20] showed that only 5 to 10% of 4σ-jumps can be explained by news. One may rightfully argue that however large, there is no database which contains "all the news". Interestingly enough, one can show that exogenous jump are less persistent than endogenous jumps, and thus cross-validate the jumps identified as endogenous/exogenous. In particular, the volatility decay after a jump follows (see Omori law):

$$\sigma_{t>t_{\text{jump}}} - \sigma_0 \sim (t - t_{\text{jump}})^{-a}, \tag{3.2}$$

with $a = 1$ for an exogenous jump and $a = 1/2$ for an endogenous jump. To note, slow relaxation is a characteristic feature of complex systems.

3.3 Continuous double auction

The vast majority of modern markets use a *continuous-time double auction* (CDA) mechanism implemented through an *electronic limit order book* (LOB) updated in real time and observable by all traders.[4] In this setup each market participant may (i) provide firm trading opportunities to the rest of the market by posting a *limit order* at a given price (liquidity provision),[5] or (ii) accept such trading opportunities by placing a *market order* (liquidity taking). The LOB stores, for a given asset on a given platform, the limit orders until they are executed against incoming market orders or cancelled. The price b_t (resp. a_t) at time t of the highest buy (resp. lowest sell) limit order is coined the *best bid* (resp. *best ask*). Buy (resp. sell) *market orders* are executed upon arrival against limit orders at the best ask (resp. best bid). If the volume of an incoming market order is larger than that available at the best, some of it will get executed against the best quote, and the rest of it against the next best quote in line, see Fig. 3.1.

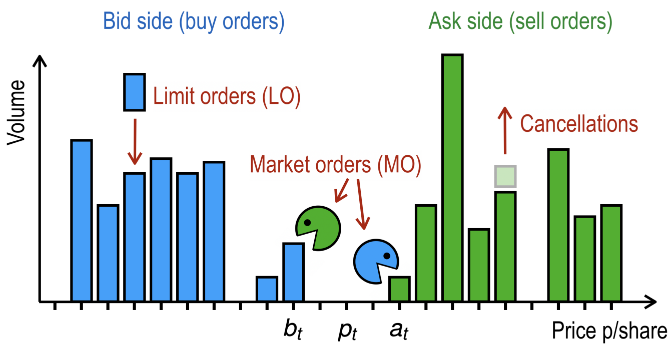


Figure 3.1: Limit order book (LOB).

[4] Fama's arguments disregard the way in which markets operate and how the trading is organised.

[5] High frequency liquidity providers acting near the trade price are called *market makers*.

We define the *midprice* $p_t := (a_t + b_t)/2$ and the *bid-ask spread* $s_t := a_t - b_t$.[6] The price axis is discrete and the step-size is coined the *tick size*, typically 0.01$ for US stocks. When the average spread is of the order of (resp. larger than) the tick size, one speaks of a *large tick* (resp. *small tick*) asset.

3.4 Liquidity and market impact

From the very nature of the trading mechanism, one can easily deduce that *trades mechanically impact prices*. The market *liquidity* can be defined as the capacity of the market to accommodate a large market order. Large trades consume liquidity and may "eat up" several queues in the order book. If there is substantial volume in the best queues, the mid-price won't be too affected. If on the other hand the LOB stores very little volume (sparse LOB), the mid-price will be very sensitive to trades. Liquidity is difficult to define because it is a dynamical concept, limit orders are continuously deposited, cancelled and executed against incoming market orders.

Be that as it may, trades consume liquidity and impact prices, this is called *market impact*, or *price impact*, or simply *impact*, commonly denoted I. It corresponds to the average price move induced by a trade of sign ϵ ($\epsilon = +1$ for buy trades, and -1 for sell trades):

$$I := \langle \epsilon_t \cdot (p_{t+1} - p_t) \rangle \,. \tag{3.3}$$

Note that $I > 0$ since, on average, buy trades push prices up while sell trades drag prices down.

At this point, it should be stressed that the available volume in the order book at a given instant in time (the *instantaneous liquidity*) is a very small fraction of the total daily traded volume, typically $< 1\%$.[7] As a result, large trades must necessarily be cut in small pieces (order splitting), and can take hours, days or even weeks to get executed. Such large orders executed sequentially or long periods of time are coined *metaorders*.

From the perspective of the EMH, market impact is a substantial paradigm shift, prices appear to move mostly because of trades themselves, very little because of new public information. One speaks of *price formation*, rather than *price discovery*. Actually, because of the small outstanding liquidity, private information (if any) can only be very slowly incorporated in prices.

Of interest for academics and practitioners, market impact is indeed both of fundamental and practical relevance. Indeed, in addition to being at the very heart of the price formation process, it is also the source of substantial trading costs due to price *slippage*[8] – also called *execution shortfall*.[9] Significant progress has been made in understanding market impact during the past decades [22–25].

3.5 The diffusivity puzzle

As a result of order splitting and metaorders, strong autocorrelations in trade signs arise. These autocorrelations decay very slowly with time as:

$$\langle \epsilon_t \epsilon_{t'} \rangle \sim |t - t'|^{-\gamma} \,, \quad \text{with } \gamma \approx 0.5 \,, \tag{3.4}$$

encoding that the order flow has long range predictability. But if trades are indeed the reason of price changes, how is this compatible with the fact that returns are nearly unpredictable? This *a priori* paradox coined the *diffusivity puzzle* [23, 26] refers to the *a priori* incompatibility of diffusive prices and super-diffusive order flow. Indeed, if we are to believe that prices are driven by trades, one would naïvely expect that the impact of correlated orders should result in persistent price dynamics.

[6] The spread represents the cost of an immediate round trip: buy then sell a small quantity results in a cost per share of s_t; it also turns out to set the order of magnitude of the volatility per trade, that is the scale of price changes per transaction [21].

[7] The daily traded volume is itself also very small compared to the total market capitalization.

[8] The slippage is the difference between the expected price of a trade and the price at which the trade is actually executed.

[9] Slippage is usually of the order a few basis points ($1\,\text{bp} = 10^{-4}$).

3.6 Short-term mean reversion, trend and value

"Nowadays people know the price of everything and the value of nothing."

– Oscar Wilde

Bachelier's first law discussed in Chapter 2 holds for timescales typically spanning from a few minutes to a few days. Below and above several market "anomalies" arise. At very short timescales, prices tend to mean revert.[10] At longer timescales (few weeks to few months) prices returns tend to be positively autocorrelated (trend), and at even longer timescales (few years) mean-revert. Actually, on such timescales the log-price π is well described by an Ornstein-Uhlenbeck process driven by a positively correlated (trending) noise:

$$\frac{\mathrm{d}\pi}{\mathrm{d}t} = -\kappa \pi_t + \eta_t\,, \quad \text{with } \langle \eta_t \eta_{t'} \rangle \sim e^{-\gamma|t-t'|}\,, \tag{3.5}$$

where $\gamma^{-1} \approx$ few months and $\kappa^{-1} \approx$ few years. The intuitive explanation of this phenomenon is that when trend signals become very strong it is very likely that the price is far away from the fundamental value. Fundamentalists (investors believing in value) then become more active, causing price mean-reversion, overriding the influence of chartists or trend-followers.

Trends are one of the most statistically significant and universal anomalies in financial markets. One can actually show that the overall performance of say, the 5-month trend, is in fact positive over every decade since 200 years [27]. Trends are clearly hard to reconcile with EMH, as it would mean that some (obvious) public information is not included in the current price! Inline with Shiller's ideas, trend and value seem inherent to human nature.[11]

"The generally accepted view is that markets are always right. [...]. I tend to believe markets are always wrong."

– George Soros

Assuming the existence of a fundamental value, one can even show that, not only is the market price quite dispersed about the fundamental value, but that the histogram of price-value distortions is actually bi-modal, indicating that assets are most often either over-valued or under-valued long periods of time [28, 29] (see Fig. 3.2).

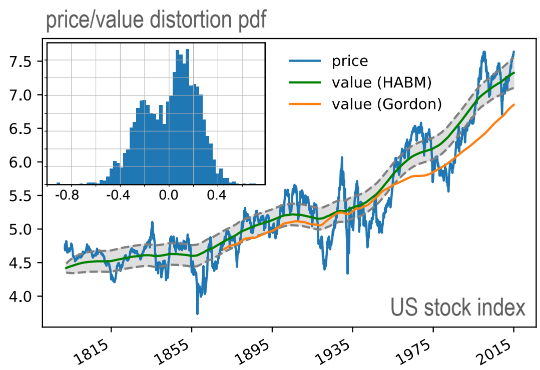


Figure 3.2: Price/value distortions on a US stock index for over two centuries, from [28].

[10]For market makers mean-reversion is favourable while trending is detrimental (adverse selection).

[11]Artificial market experiments show that even when the fundamental value is known to all, one is tempted to forecast the behaviour of their fellow traders which ends up creating trends, bubbles and crashes. The temptation to outsmart one's peers is too strong to resist.

3.7 Paradigm shifting

Two different scenarios for price changes have been exposed: fundamental value driven prices, and order flow driven prices.

1. Within the EMH or *price discovery* framework, prices are *exogenous*. Prices reflect fundamental values, up to small and short-lived mispricings (quick and efficient digestion). Market impact is non other than information revelation, the order flow adjusts to changes in fundamental value, regardless of how the trading is organised.

   While consistent with martingale prices and *fundamentally efficient* markets (by definition new information cannot be anticipated), this view of markets comes with some major puzzles. In addition to the whole idea of markets almost immediately digesting the information content of news being rather hard to believe, one counts in particular the *excess trading*, the *excess volatility* and the *trend-following* puzzles. The concept of high frequency non-rational agents, *noise traders*, was artificially introduced in the 80's to cope with excess trading and excess volatility. But however noisy, noise traders cannot account for excess long-term volatility and trend-following.

2. Within the *order-driven prices* or *price formation* framework, prices are *endogneous*, mostly affected by the process of trading itself. The main driver of price changes is the order flow, regardless of its information content. Impact is a mechanical statistical effect, very much like the response of a physical system.

   Here, prices are thus perfectly allowed to err away from the fundamentals (if any). Further, excess volatility is a direct consequence of excess trading! This scenario is also consistent with self-exciting feedback effects, expected to produce clustered volatility (see Chapter 2): the activity of the market itself leads to more trading which, in turn, impacts prices and generates more volatility and so on and so forth. As we shall see in Chapter 5, such mechanisms are also expected to produce power law tailed returns.

   While probably more convincing – and more inline with real data – than the EMH view, two caveats remain at this stage: the diffusivity puzzle[12] and the market efficiency. Indeed, let's recall that despite the anomalies discussed above, for reasonable timescales prices are approximately martingales. How can the order-driven prices perspective explain why signature plots are so universally flat? Fundamental efficiency is replaced with *statistical efficiency*, the idea being that efficiency results from competition: traders seek to exploit statistical arbitrage opportunities, which, as a result, mechanically disappear,[13] by that flattening the signature plot.

Finally, note that the very meaning of what a good trading strategy is varies with one's view. Within the EMH, a good strategy is one which predicts well moves in the fundamental value. With mechanical impact, a good strategy aims at anticipating the actions of fellow traders, the order flow, rather than fundamentals. Further empirical support of the "order-driven prices" view is given in Chapters 4 and 7.

[12] In Chapter 4, we will provide an econometric resolution for the diffusivity puzzle.
[13] See e.g. the *Minority Game* presented in Chapter 5.

4

Econometric models for price changes

In this Chapter, we present some *econometric models*[1] aimed at reproducing (sometimes predicting) some of the stylised facts unveiled in Chapters 2 and 3. Such models need to be calibrated on real data, but one should always be extra-careful and look at their conclusions with a critical eye. Indeed, anything can be calibrated on data,[2] and, despite the natural temptation to do so, the fact that one can put a number on it should not give extra-credit to the model whatsoever.

4.1 The propagator model

As anticipated by Bachelier over a hundred years ago, prices are very close to unpredictable diffusive processes, or in modern language *martingales*. One has $\mathscr{C}_r(t-t') := \langle r_t r_{t'} \rangle \sim \delta(t-t')$. The order flow, on the other hand, is a highly persistent process presenting long-range autocorrelation resulting from directional metaorder flow, see Chapter 3. Denoting ϵ_t the trade sign, empirical data shows that $\mathscr{C}_\epsilon(t-t') := \langle \epsilon_t \epsilon_{t'} \rangle \sim |t-t'|^{-\gamma}$ with $\gamma \approx 0.5$. The *diffusivity puzzle* [23, 26] refers to the *a priori* incompatibility of diffusive prices and super-diffusive order flow. Indeed, if we are to believe that prices are driven by trades, one would naively expect that the impact of correlated orders should result in persistent price dynamics. The *propagator model* was initially introduced to solve the diffusivity puzzle [30].

4.1.1 The impact of trades

The most simple and naive way to translate mathematically that trades impact prices is:

$$r_t = G_1 \epsilon_t + \eta_t \,, \tag{4.1}$$

with G_1 a constant measuring the impact amplitude and η_t a noise term accounting for non-trade related price changes (exogenous quote changes). We assume $\langle \eta_t \rangle = 0$, $\langle \eta_t \eta_{t'} \rangle = \sigma_0^2 \delta(t-t')$ and $\langle r_t \eta_{t'} \rangle = 0$. One can easily check that such a model violates market efficiency and does not solve the diffusivity puzzle: $\langle r_t r_{t'} \rangle = G_1^2 \mathscr{C}_\epsilon(t-t') \not\sim \delta(t-t')$. The resulting price dynamics is super-diffusive, with a Hurst exponent $H = 1-\gamma/2 > 1/2$. Using Eq. (4.2) and writing:

$$p_t = p_0 + \sum_{t'=0}^{t-1} r_{t'} = p_0 + G_1 \sum_{t'=0}^{t-1} \epsilon_{t'} + \sum_{t'=0}^{t-1} \eta_{t'} \,, \tag{4.2}$$

allows to identify the problem. Here, each trade suffices to shift the supply and demand curves permanently, which seems a bit too extreme. There is no reason why the impact of each trade should be imprinted in the price forever.

[1]As opposed to *microscopic* or *microfounded models*, see Chapter 5.

[2]For example, one can always calibrate a Gaussian random walk on a given time series and output a standard deviation σ, but this does not prove that the data are Gaussian!

4.1.2 Transient impact

Market efficiency seems to impose that a trade's impact must relax over time. Precisely, in order to compensate for order flow correlation and restore price diffusivity, the impact of a trade must be transient and can be described by a time-decaying kernel or *propagator* $G(t)$:

$$p_t = p_0 + \sum_{t'=0}^{t-1} G(t-t')\epsilon_{t'} + \sum_{t'=0}^{t-1} \eta_{t'} \,. \tag{4.3}$$

Non-parametric calibration of this model onto real data indicates that G decays as a power law. Letting $G(t) \sim t^{-\beta}$ into Eq. (4.3) and enforcing $\langle (p_t - p_0)^2 \rangle \sim t$ yields:[3]

$$\beta = \frac{1-\gamma}{2} \,, \tag{4.4}$$

which formalises that persistent order flow and diffusive prices can make peace provided the impact of single trades decays as a slow power law of time with a particular exponent $\beta \approx 0.25$. Impact decay must be fine-tuned to compensate the long memory of order flow, and allow the price to be close to a martingale. The very slow decay of impact (so slow that its sum is divergent) is sometimes referred to as *long-range resilience*. Note also that this model predicts zero permanent impact $\lim_\infty G = 0$.

A few points deserve further discussion:

- Calibrating Eq. (4.3) on real data reveals indeed that $\beta \approx (1-\gamma)/2$, and that up to $\approx 80\%$ of price moves can be explained by trades!

- Using trade volume q_t instead of trade signs ϵ_t is also common, in particular for pricing purposes, see below. Actually, the most statistically significant proxy for order flow is $\epsilon_t \log|q_t|$ indicating that the most important feature is the sign but with a residual dependence on volume.

- In practice, more complex multi-event propagator models are used, often coined *generalized propagator models*, with several kernels and feedback on limit orders, cancellations etc.

- In some situations, it is more convenient to write the propagator model for single returns. Using Eq. (4.3), one has:
$$r_t = p_{t+1} - p_t = G(1)\epsilon_t + \sum_{t'=0}^{t-1} \big[G(t+1-t') - G(t-t')\big]\epsilon_{t'} + \eta_t \,.$$
Defining the discrete derivative $\dot{G}(t) = G(t+1) - G(t)$ and using the convention $G(0) = 0$ such that $\dot{G}(0) = G(1)$ yields:
$$r_t = \sum_{t' \le t} \dot{G}(t-t')\epsilon_{t'} + \eta_t \,. \tag{4.5}$$

Most importantly bear in mind that the propagator model is *a* solution to the diffusivity puzzle, it is not *the* solution. Further, at this stage the model is purely econometric, and finding a deeper reason for Eq. (4.4), possibly from a microfounded game-theoretic scenario appears essential to improve our understanding of price formation.

[3] First, note that $\langle (p_t - p_0)^2 \rangle \sim t$ is tantamount to $\mathscr{C}_r(t-t') \sim \delta(t-t')$, indeed $\langle (p_t - p_0)^2 \rangle = \sum_{t',t''=0}^{t-1} \langle r_{t'} r_{t''} \rangle \sim \sum_{t',t''=0}^{t-1} \delta(t'-t'') \sim t$. Then, using Eq. (4.3), $\langle (p_t - p_0)^2 \rangle = \sum_{\ell,s=1}^{t} G(\ell)G(s)\mathscr{C}_\epsilon(s-\ell) \sim \iint^t \mathrm{d}\ell\,\mathrm{d}s (\ell s)^{-\beta} |\ell - s|^{-\gamma} \sim t^{2-2\beta-\gamma} \underbrace{\iint^1 \mathrm{d}u\,\mathrm{d}v (uv)^{-\beta} |u-v|^{-\gamma}}_{\text{constant}}$.

4.1.3 A note on cross-impact

In this course we focus mainly on the price impact of single assets with no regard of inter-asset interactions, a strong approximation that may lead to an underestimation of trading costs and possible contagion effects. As shall be briefly discussed in Chapter 9 most market participants trade large portfolios that combine hundreds or thousands of correlated assets. By calibrating multivariate propagator models [31] of the form:[4]

$$\boldsymbol{p}_t = \boldsymbol{p}_0 + \sum_{t'=0}^{t-1} \boldsymbol{G}(t-t')\boldsymbol{\epsilon}_{t'} + \sum_{t'=0}^{t-1} \boldsymbol{\eta}_t\,, \tag{4.6}$$

one can show that inter-asset price impact, coined *cross-impact*, is significant, and that transactions mediate a significant part of the cross-correlation between different instruments. The intuition is that in two related products the order flow of one may reveal information, or communicate excess supply/demand regarding the other. The analysis of cross-impact effects falls beyond the scope of this course, for more details see e.g. [32–38].

4.1.4 Slippage costs and optimal execution

From the industrial perspective, market impact means trading costs. The extra cost due to impact paid for executing a given order sequence $\{v_t\}_{t\in[1,T]}$ is called *slippage cost*, or *implementation shortfall*. It is computed as:

$$C_{\text{slip}} = \sum_{t=1}^{T} v_t(p_t - p_0) \tag{4.7}$$

Using a linear propagator model, as that of Eq. (4.3), but replacing ϵ_t with v_t yields:

$$C_{\text{slip}} = \sum_{t=1}^{T}\sum_{t'=1}^{t} v_t G(t-t')v_{t'} = \frac{1}{2}\sum_{t,t'=1}^{T} v_t G(|t-t'|)v_{t'}\,, \tag{4.8}$$

which is a convenient quadratic form in v_t. Note that here we have omitted spread costs arising from the difference between midprice and transaction price. In full generality one should write:

$$C_{\text{slip}} = \frac{1}{2}\sum_{t,t'=1}^{T} v_t G(|t-t'|)v_{t'} + \sum_{t=1}^{T}\frac{v_t s_t}{2}\,. \tag{4.9}$$

Solving the *optimal execution* problem amounts to finding the optimal schedule $\{v_t^\star\}$ which minimises C_{slip} under the constraint $\sum_t v_t = Q$ with Q the total order volume. Assuming constant spread $s_t = s$ allows to disregard the spread term (linear in v_t) in the optimisation problem.[5] Switching to continuous time, $v_t \to v_t \mathrm{d}t$ and $\sum\sum \to \iint$, one is left with a standard problem of variational calculus, where one shall introducing a Lagrange multiplier λ to enforce the constraint $\int v_t \mathrm{d}t = Q$. Setting $\delta C_{\text{slip}}/\delta[v_t] = 0$, it follows that for all t:

$$\int_0^T \mathrm{d}t' G(|t-t'|)v_{t'} = \lambda\,. \tag{4.10}$$

In the general case this equation is difficult to solve.[6] For an exponentially decaying kernel $G(t) = G_0 e^{-\omega t}$, one can show that $v_t^\star$ is given by the bucket shaped function:

$$v_t^\star = \frac{Q}{1+\omega T/2}\Big[\delta(t) + \delta(T-t) + \tfrac{\omega}{2}\Big], \tag{4.11}$$

[4] Bold lower (resp. upper) cases are vectors (resp. matrices).

[5] If the spread is non-constant, traders try to take advantage of moments where the spread is small, and as a result the two terms on the RHS of Eq. (4.9) interact in a non-trivial way, making the optimisation problem much more complicated.

[6] One can prove that v_t must be symmetric about $T/2$: $v_t = v_{T-t}$. Indeed, one can easily check that v_{T-t} also solves Eq. (4.10) (change of variables $t' \to T - t'$).

with $\delta(t)$ such that $\int_0^T \mathrm{d}t\,\delta(t) = 1/2$, see Fig. 4.1. One finds that a fraction $1/(2+\omega T)$ should be executed at the open and at the close, while the rest should be executed at a constant speed throughout trading the interval.[7] The corresponding slippage writes:

$$C^\star_{\text{slip}} = G_0 \left(\frac{Q}{2+\omega T} \right)^2 (1+\omega T) + \frac{Qs}{2}, \tag{4.12}$$

which a decreasing function of T, in favour of the slowest possible execution,[8] as could be expected from the quadratic nature of the cost model.

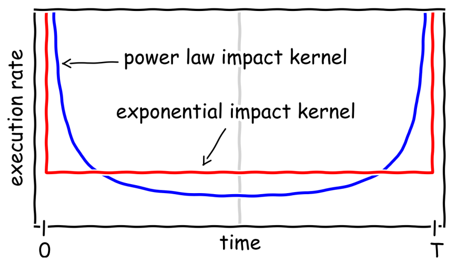


Figure 4.1: Optimal execution rate as function of time.

In the more realistic case of a power-law kernel $G(t) \sim (1+\omega t)^{-\beta}$, the problem can be solved numerically. The optimal profile is well approximated by the following U-shaped function (see Fig. 4.1):

$$v^\star_t \approx \frac{\Gamma(2\beta)}{\Gamma^2(\beta)} Q T^{1-2\beta} t^{\beta-1} (T-t)^{\beta-1}. \tag{4.13}$$

Note that, minimising the variance of the expected slippage can also be achieved with an extra penalty term (see the famous *Almgren-Chriss problem* [39]).

An important warning is that, throughout this section, we have implicitely assumed that our order flow $\{v^\star_t\}$ is uncorrelated with the order flow from the rest of the market, and that in the absence of our investor, the price is a martingale. These assumptions might very well fail, since the information used to decide on such a trade could well be shared with other investors, who may trade in the same direction.

4.2 The GARCH framework

Generalised autoregressive conditional heteroscedasticity (GARCH) models are meant to capture time-varying volatility effects. They are an attempt to reproduce some stylized facts of price statistics (fat-tails, volatility clustering, leverage effect) but without even considering trades, and thus completely abstracting from the price formation process. They most often lack sound micro-foundations and bring little intuition about the mechanisms leading to the proposed mathematical representation of reality.

Fat tails The idea that large realised returns in the past create extra volatility in the present can be encoded as:

$$\sigma^2_{t+1} = \sigma^2_0 + \kappa\left[\alpha x^2_t + (1-\alpha)\sigma^2_t\right], \tag{4.14}$$

[7] In the limit $\omega T \to \infty$ (rapid impact decay), the optimal profile is $v^\star_t = Q/T$, called a *time-weighted average price* (TWAP).

[8] In practice execution cannot be infinitely slow as the expected gain that motivates trading generally has a finite prediction horizon.

with σ_0 the *bare* volatility level, κ the feedback intensity, and α the relative weight of the realized volatility to the true underlying volatility in the feedback process. With the specification $x_t = \sigma_t \xi_t$ (see Chapter 2) one obtains $\sigma_{t+1}^2 = \sigma_0^2 + \kappa\sigma_t^2[1+\alpha(\xi_t^2-1)]$.

In the case $\alpha = 1$, provided the volatility reaches a stationary state with mean $\langle \sigma_t^2 \rangle = \bar{\sigma}^2$, one has for $\kappa < 1$:

$$\bar{\sigma}^2 = \frac{\sigma_0^2}{1-\kappa} \underset{\kappa\uparrow 1}{\longrightarrow} +\infty\,. \tag{4.15}$$

For $\kappa > 1$ the volatility diverges. For $\kappa < 1$, one can show that such a model leads to power-law distributed returns:[9]

$$\mathbb{P}(x) \sim \frac{1}{|x|^{1+2\zeta}}\,, \tag{4.16}$$

with $\zeta(\kappa) \mid (2\kappa)^{\zeta}\Gamma\left(\zeta+\frac{1}{2}\right) = \Gamma\left(\frac{1}{2}\right)$.

Volatility clustering Multiplying both sides of Eq. (4.14) by $\sigma_{t-\tau}$ (still with $\alpha = 1$) and taking the average yields $g(\tau+1) = \sigma_0^2 - (1-\kappa)\bar{\sigma}^2 + \kappa g(\tau)$ where we have defined $\bar{\sigma}^2 g(\tau) := \langle \sigma_t^2 \sigma_{t+\tau}^2 \rangle - \bar{\sigma}^4$. Using Eq. (4.15) one obtains: $g(\tau+1) = \kappa g(\tau)$ and thus $g(\tau) \sim \exp(-\tau/\tau_c)$ with $\tau_c = |\log\kappa|^{-1}$. While the decay time of the exponential volatility correlation diverges as $\kappa \uparrow 1$, there is a single characteristic timescale, whereas empirical data clearly shows the existence of multiple time scales encoded in a (scale invariant) power-law decay $g(\tau) \sim \tau^{-\nu}$ with $\nu \approx 0.2$ (see Chapter 2).

Generalising the model to take into account the past realized returns over many time scales as:

$$\sigma_{t+1}^2 = \sigma_0^2 + \kappa \sum_{\tau \le 1} \mathcal{K}(\tau) X_{t-\tau,\tau}^2\,, \tag{4.17}$$

with $X_{t,\tau} := \log p_{t+\tau} - \log p_t = \sum_{t=t'}^{t+\tau-1} x_{t'}$ and $\mathcal{K}(\tau) \sim t^{-\delta}$, yields instead $g(\tau) \sim \tau^{-\nu}$ with $\nu = 3-2\delta$.

Leverage effect To capture the leverage effect one needs to include a sign sensitive term to dissymmetrise the dynamics. The simplest model is encoding that volatility should increase with negative past returns:

$$\sigma_{t+1}^2 = \sigma_0^2 + \kappa x_t^2 - \kappa_{\text{lev}} x_t\,, \tag{4.18}$$

with $\kappa > 0$ and $\kappa_{\text{lev}} < 2\kappa\sigma_0$ to ensure positivity of the RHS.

4.3 Multifractal models

Alternative options to capture long-range volatility correlations are the *multifractal stochastic volatility models*. The *Bacry-Muzy-Delour* (BMD) model [40] is defined by taking the volatility to be a log-normal random variable as:

$$\begin{gathered} \sigma_t := \sigma_0 \exp \omega_t\,, \quad \text{with } \omega_t \text{ Gaussian}\,, \\ \langle \omega_t \rangle = -\lambda^2 \log \alpha\,, \quad \text{and } \langle \omega_t \omega_{t+\tau} \rangle - \langle \omega_t \rangle^2 = -\lambda^2 \log(\alpha(|\tau|+1))\,. \end{gathered} \tag{4.19}$$

The interesting regime is $\tau \ll \alpha^{-1}$ where $\alpha^{-1} \gg 1$ is a large cut-off time scale, beyond which the correlations of ω_t vanish; λ is the *intermittency parameter*. One can show that $\bar{\sigma}^2 = \sigma_0^2$ and that the rescaled correlation $g(\tau)$ of the squared volatilities, as defined above, writes:

$$g(\tau) = [\alpha(1+\tau)]^{-\nu}\,, \quad \text{with } \nu = 4\lambda^2\,. \tag{4.20}$$

[9] Ajouter la preuve.

Furthermore, all even moments of price changes can be computed and one finds a *multifractal* behaviour:

$$\langle |\log p_{t+\tau} - \log p_t|^q \rangle \sim \tau^{\xi(q)}, \quad \text{with } 2\xi(q) = q\big[1 - \lambda^2(q-2)\big]. \tag{4.21}$$

Some remarks can be made.

- The empirical distribution of volatility is compatible with log-normality. But this does not mean that it is *the* right model.[10]
- For $q\lambda^2 > 1$, the moments of price changes diverge, suggesting power-law tailed returns with tail exponent $\mu = 1/\lambda^2$.
- Fitting Eq. (4.19) to real data yields $\lambda \approx 0.01$ to 0.1 and $\alpha^{-1} \approx$ a few years. This yields $\mu \approx 10$ to 100, much larger than the empirical tails ($3 < \mu < 5$) shown in Chapter 2.

Conclusions

The danger of overconfidence in econometric models, as those presented in this Chapter, is important. Indeed, it is quite human when rather simple equations give non-trivial results consistent with empirical data to adopt them as if they were the one and only, forgetting that there is no microfounded justification whatsoever for postulating them in the first place.

[10] An inverse gamma distribution also fits the data very well, and is actually motivated by simple models of volatility feedback such as GARCH.

5

Microscopic (agent-based) models for price changes

If, as suggested in Chapter 3, volatility is indeed mostly endogenous then clearly seeking for a microscopic interpretation is probably the best way to go. Imitation, risk aversion and feedback loops (trading impacts prices, which in turn influence trading and so on and so forth) are good candidates to account for some of the non-Gaussian statistics revealed by empirical data. Let us stress that, to this day, there exists no complete agent-based model (ABM), simple and universal, able to account for all of the empirical features of price statistics. In this Chapter, we provide a few instructive toy examples or *toy models* to analyse the possible effects of interactions, feedback or heterogeneities, to name a few, and help develop an intuition about the *complex systems* underlying market dynamics.

5.1 Collective behaviour

Statistical physics' very *raison d'être* is to bridge the gap between the microscopic world and the macroscopic laws of nature. It provides an intellectual scheme and tools to describe the collective behaviour of large systems of interacting *particles*, which can be very diverse objects from molecules and spins to information bits or individuals in a crowd, or market participants. As put by Anderson in his exquisite paper *More is different* (Science, 1972) [41]:

> *"The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other."*
>
> – Philip W. Anderson

Let us also mention that, while such ideas were mainly developed by them, they do not belong to statistical physicists only. For example, as Poincaré commented on Bachelier's thesis:

> *"When men are in close touch with each other, they no longer decide independently of each other, they each react to the others. Multiple causes come into play which trouble them and pull them from side to side, but there is one thing that these influences cannot destroy and that is their tendency to behave like Panurges sheep."*
>
> – Henri Poincaré

Nature counts many examples of collective behaviour, from the dynamics of bird flocks or fish schools to the synchronisation of fireflies (see [42] for a great holiday reading). Interactions are key in understanding collective phenomena. For example, by no means could anyone pretend to account for the complex dynamics of a bird flock by simply extrapolating the behaviour of one bird. Only interactions can explain how over a thousand birds can change direction in a fraction of a second, without

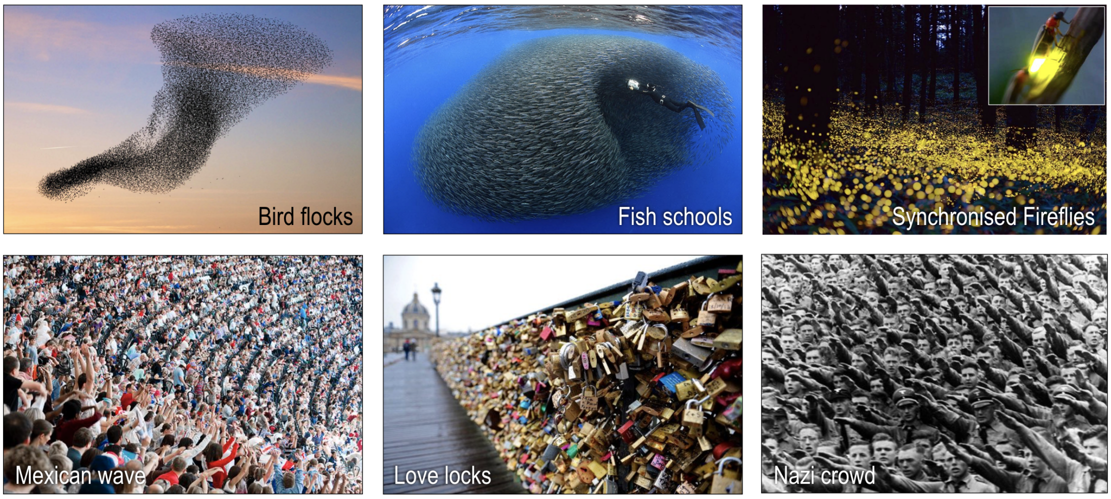


Figure 5.1: Examples of collective phenomena in the animal world and human systems.

having a leader giving directions (nonlinear amplification of local fluctuations or avalanches). Further, it appears that the features of the individual may be inessential to understand aggregate behaviour. Indeed, while a bird and a fish are two quite different specimens of the animal world, bird flocks and fish schools display numerous similarities. Human systems naturally display collective behaviour as well, for the better or for the worse, see Fig. 5.1. Other examples are clapping, fads and fashion, mass panics, vaccination campaigns, protests, or stock market crashes.

5.2 Power laws, scale invariance and universality

Physicists are quite keen on power laws as they are most often the signature of collective, complex and fascinating phenomena. Further, power laws are most often universal, in the sense that the power law exponents do not depend on the microscopic details of the system at hand.

Power laws are interesting because they are scale invariant functions.[1] This means that, contrary to e.g. exponential functions, systems described by power laws have no characteristic length scales or timescales. Many examples can be found in the physics of *phase transitions*. At the critical temperature of the paramagnetic-ferromagnetic phase transition, Weiss magnetic domains become scale invariant. So do the fluctuations of the interface at the critical point of the liquid-gas phase transition. Of particular interest to our purpose is the percolation phase transition, see below. Fluid turbulence, already mentioned in Chapters 1 and 2 for its similarities with financial time series statistics, displays a number of power laws. In particular the statistics of the velocity field in a turbulent flow are often made of scale invariant power laws, independent of the fluid's nature, the geometry of the flow and even the injected energy. As we have seen, and shall see in Chapters 7 and 6, power laws are also very present in finance: probability distribution of returns ($3 < \mu < 5$), correlations of volatility ($\nu \approx 0.2$), correlations of the order flow ($\gamma \approx 0.5$), volatility decay after endogenous shocks ($a \approx 0.5$), etc. Their universality means that they are independent of the asset, the asset class, the time period, the market venue, etc.

All this indicates that financial markets can probably benefit from the sort of analysis and modelling conducted in physics to understand complex collective systems. In a nutshell, this amount to modelling the system at the microscopic level with its interactions and heterogeneities, most often through agent-based modelling (ABM), and carefully scaling up to the aggregate level where the generic power laws arise. One remark is in order here: this approach goes against the whole *representative agent* idea which nips in the bud all heterogeneities, often essential to account for in the description of the phenomena at hand, see [43]. Actually, simplifying the world to representative agents poses a real dimensionality issue: while there is only one way to be the same, there are an infinity of ways to be different.

5.3 Mimicry and opinion changes

Below, we present four insightful toy models to illustrate the effects of interactions in collective socioeconomic systems.

5.3.1 Herding and percolation

Here we present a simple model to understand how herding can affect price fluctuations [2]. The ingredients of this model are:

- Consider a large number N of agents.
- Assume that returns r are proportional to demand-supply imbalance as:

$$r = \frac{1}{\lambda} \sum_{i=1}^{N} \varphi_i := \frac{\phi}{\lambda}, \tag{5.1}$$

  where $\varphi_i \in \{-1, 0, 1\}$ signifies agent i selling, being inactive, or buying, and λ is a measure of market depth.
- Agents i and j are connected (or interact) with probability p/N (and ignore each other with probability $1 - p/N$), such that the average number of connections per agent is $\sum_{j \neq i} p/N \approx p$.
- If two agents i and j are connected, they agree on their strategy: $\varphi_i = \varphi_j$.

[1] A function f is said to be scale invariant if and only if there exists a function g such that for all x, y, $\frac{f(x)}{f(y)} = g\left(\frac{x}{y}\right)$. One can easily show that the only scale invariant functions are power laws.

Percolation theory teaches us that the population clusters into connected groups sharing the same opinion, see e.g. [44]. Denoting by n_α the size of cluster α, one has:

$$r = \frac{1}{\lambda}\sum_\alpha n_\alpha \varphi_\alpha\,, \tag{5.2}$$

where φ_α is the common strategy within cluster α. Price statistics thus conveniently reduce to the statistics of cluster sizes for which many results are known. In particular, on can distinguish three regimes.

1. As long as $p < 1$ all cluster are small compared to the total number of agents N, and the probability distribution of cluster sizes scales as:

$$\mathbb{P}(n) \sim \frac{1}{n^{5/2}}\exp(-\epsilon^2 n)\,, \tag{5.3}$$

   with $\epsilon = 1 - p \ll 1$. The market is unbiased $\langle r\rangle = 0$ (as long as $\varphi = \pm 1$ play identical roles).

2. When $p = 1$, equivalently $\epsilon = 0$ (percolation threshold), $\mathbb{P}(n)$ becomes a pure power-law with exponent $\mu = 3/2$, and the generalised CLT implies that the distribution of returns converges to a pure symmetric Levy distribution of index $\mu = 3/2$.

3. If $p > 1$, there exists a percolation cluster of size $O(N)$, or in other words there exists a finite fraction of agents with the same strategy, $\langle \phi\rangle \neq 0 \Rightarrow \langle r\rangle \neq 0$, and the market crashes. A *spontaneous symmetry-breaking* occurs.

This quite instructive model gives the "good" distribution of returns ($\mu = 3/2$ was observed in Chapter 2 for the MXN/$ rate) and the possibility for crashes when the connectivity of the interaction network increases. However, this story holds only if the system sits below, but very close to the instability threshold. But what ensures such *self-organised criticality* where the value of p would stabilise near $p \lesssim 1$? In section 5.5, we give a stylised example illustrating why such systems could be naturally attracted to the critical point. Further, note that this model is static and thus not relevant to account for volatility dynamics. How do these clusters evolve with time? How to model opinion dynamics? An interesting extension of this model was proposed in [45]: by allowing each agent to probe the opinion of a subset of other agents and either conform to the dominant opinion or not if the majority is too strong, one obtains a richer variety of market behaviors, from periodic to chaotic.

5.3.2 The random field Ising model

The *random field Ising model* (RFIM) was introduced in the 90's to model hysteresis loops in disordered magnets [46]. Such magnets called *spins* flip collectively in response to a quasi-steady external solicitation, and, depending on the parameters, the flips may take place smoothly, or be organized in intermittent bursts, or *avalanches*, leading to a specific acoustic pattern called Barkhausen noise. Because the model is so generic, it has been used to account for many other physical situations, such as earthquakes, fracture in disordered materials, failures in power grids, etc.

Here we present its application to socio-economical systems [47], and in particular to binary decision situations under both social pressure and the influence of some global information. The ingredients of the model are as follows:

- Consider a large number N of agents who must make a binary choice (e.g. to buy or to sell, to trust or not to trust, to vote yes or no, to cheat or not to cheat, to evade tax or not, to clap or to stop clapping, to attend or not to attend a seminar, to join or not to join a riot, etc.).
- The outcome of agent i's choice is denoted by $s_i \in \{-1, +1\}$.
- The incentive u_i of agent i to choose $+1$ over -1 is given by:[2]

$$u_i(t) = f_i + h(t) + \sum_{j=1}^{N} J_{ij} s_j(t-1)\,, \tag{5.4}$$

[2]The incentive u_i is the difference between the utilities choices $+1$ and -1.

encoding that the decision of agent i depends on three distinct factors:

1. The personal inclination, which we take to be time independent and which is measured by $f_i \in \mathbb{R}$ with distribution $\rho(f)$. Large positive (resp. negative) f indicates a strong *a priori* tendency to decide $s_i = +1$ (resp. $s_i = -1$).
2. Public information, affecting all agents equally, such as objective information on the scope of the vote, the price of the product agents want to buy, or the advance of technology, etc. The influence of this exogenous signal is measured by the *incentive field*, $h(t) \in \mathbb{R}$.
3. Social pressure or imitation effects. Each agent i is influenced by the previous decision made by a certain number of other agents j. The influence of j on i is measured by the *connectivity matrix* J_{ij}. If $J_{ij} > 0$, the decision of agent j to e.g. buy reinforces the attractiveness of the product for agent i, who is now more likely to buy. This reinforcing effect, if strong enough, can lead to an unstable feedback loop.

- Agents decide according to the so-called *logit rule* or *quantal response* in the choice theory literature [48] (see Appendix A) which makes the decision a random variable, with probability:

$$\mathbb{P}(s_i = +1|u_i) = \frac{1}{1+e^{-\beta u_i}}, \qquad \mathbb{P}(s_i = -1|u_i) = 1 - \mathbb{P}(s_i = +1|u_i), \tag{5.5}$$

  where β quantifies the level of irrationality in the decision process, analogous to the inverse temperature in physics. When $\beta \to 0$ incentives play no role and the choice is totally random/unbiased, whereas $\beta \to \infty$ corresponds to deterministic behaviour.

We focus on the *mean-field* case where $J_{ij} := J/N$ for all i, j,[3] and $f_i = f = 0$ for all i. Note that this does not mean that each agent consults all the others, but rather that the average opinion $m := N^{-1} \sum_i s_i$, or e.g. the total demand, is public information and influences the behaviour of each individual agent equally.[4] Defining the average incentive $u := N^{-1} \sum_i u_i$ and the fraction $\phi = N_+/N$ of agent choosing $+1$, one can easily show that:

$$m(t) = 2\phi(t) - 1, \quad \text{and} \quad u(t) = h(t) + J\big[2\phi(t-1) - 1\big]. \tag{5.6}$$

Using Eqs. (5.5) and denoting $\zeta = e^{\beta u}$ yields the following updating rules:

$$\mathbb{P}(N_+ \to N_+ + 1) = (1-\phi)\frac{\zeta}{1+\zeta}, \qquad \mathbb{P}(N_+ \to N_+ - 1) = \phi\frac{1}{1+\zeta}, \tag{5.7}$$

which naturally lead to the following evolution equation: $\langle N_+ \rangle_{t+1} = \langle N_+ \rangle_t + 1 \times \mathbb{P}(N_+ \to N_+ + 1) - 1 \times \mathbb{P}(N_+ \to N_+ - 1) + 0 \times \mathbb{P}(N_+ \to N_+)$ that is:

$$\mathrm{d}\langle N_+ \rangle = \frac{\zeta}{1+\zeta} - \phi\,. \tag{5.8}$$

The equilibrium state(s) of the system are such that $\mathrm{d}\langle N_+ \rangle = 0$ which yields:

$$\phi^\star = \frac{\zeta^\star}{1+\zeta^\star}, \quad \text{with} \quad \zeta^\star = e^{\beta[h+J(2\phi^\star-1)]}. \tag{5.9}$$

Noting that $m^\star = 2\phi^\star - 1$ yields the well known *Curie-Weiss* (self-consistent) *equation*:

$$m^\star = \tanh\left[\tfrac{\beta}{2}(h + Jm^\star)\right]. \tag{5.10}$$

The solutions of Eq. (5.10) are well known (see Fig. 5.2). When $h = 0$, there is a critical value $\beta_c = 2/J$ separating a high temperature (equivalently weak interactions) regime $\beta < \beta_c$ where agents shift randomly between the two choices, with $\phi^\star = 1/2$; this is the *paramagnetic* phase. A spontaneous *polarization* (symmetry-breaking) of the population occurs in the low temperature (equivalently strong

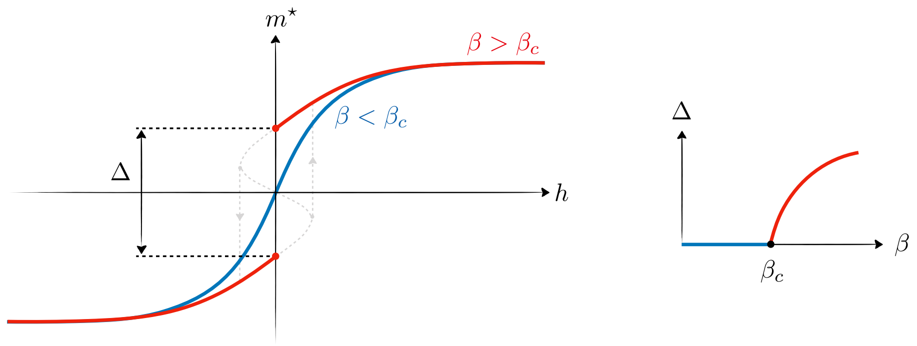


Figure 5.2: Average opinion (or aggregate demand, or overall trust etc.) as a function of the external incentive field in the high and low temperature limits.

interactions) regime $\beta > \beta_c$, that is $\phi^\star \neq 1/2$; this is the *ferromagnetic* phase.[5] When $h \neq 0$, one of the two equilibria becomes exponentially more probable than the other.

To summarise the results let us consider the following *gedankenexperiment*. Suppose that one starts at $t = 0$ from a euphoric state (e.g. confidence in economic growth), where $h \gg J$, such that $\phi^\star = 1$ (everybody wants to buy). As confidence is smoothly decreased, the question is: will sell orders appear progressively, or will there be a sudden panic where a large fraction of agents want to sell? One finds that for small enough influence, the average opinion varies continuously (until $\phi^\star = 0$ for $h \ll -J$), whereas for strong imitation discontinuity appears around a *crash time*, when a finite fraction of the population simultaneously change opinion. Empirical evidence of such nonlinear opinion shifts can be found in the adoption of cell phones in different countries in the 90s [49], the drop of birth rates by the end of the glorious thirty [50], crime statistics in different US states [51], or the way clapping dies out at the end of music concerts [49].

For β close to β_c, one finds that opinion swings (e.g. sell orders) organise as avalanches of various sizes, distributed as a power-law with an exponential cut-off which disappears as $\beta \to \beta_c$. The power law distribution indicates that most avalanches are small, but some may involve an extremely large number of individuals, without any particularly large change of external conditions. In this framework, it is easy to understand that, provided the external confidence field $h(t)$ fluctuates around zero, bursts of activity and power-laws (e.g. in the distribution of returns) are natural outcomes. In other words, a slowly oscillating $h(t)$ leads to a succession of bull and bear markets, with a strongly non-Gaussian, intermittent behaviour.

5.3.3 The limits of copy-cat strategies

The Marsili & Curty forecasting game [52] provides a very simple framework to understand why there is some value in following the trend (*herding* or *crowding*), and why beyond a certain number of copy-cats such a game becomes quite dangerous.The ingredients of this model are as follows.

- Consider a large number N of agents who must make a binary choice.
- A fraction z of the population is made of *fundamentalists* who process information to make a choice. They see right with probability $p > 1/2$ (encoding value in objective information).
- A fraction $1 - z$ of the population is lazy, made of *followers* who merely observe the action of others to make up their mind.
- At each time step t, one of the lazy agent adopts the majority opinion at time $t - 1$ among a group of m agents randomly chosen (including *a priori* both followers and fundamentalists). The

[3] Local networks effects relevant in the socio-economic context can be addressed, e.g. taking J_{ij} to be a regular tree or a random graph, but falls beyond the scope of this course.

[4] In the case of financial markets, the price change itself can be seen as an indicator of the aggregate demand.

[5] Using the Taylor expansion of the tanh function in Eq. (5.10) in the vicinity of the critical point ($\beta \to \beta_c$) allows to determine the critical exponents of this phase transition.

choices of the followers are initialised with equal probability and we assume m odd to avoid draws.

- We denote by q_t (resp. π_t) the probability that a follower (resp. an agent chosen at random) makes the right choice at time t. We repeat this process until it converges: $q_t, \pi_t \to q, \pi$.

An equal-time equation can be easily obtained by noting that agents are either fundamentalists or followers, such that:

$$\pi_t = zp + (1-z)q_t \,. \tag{5.11}$$

A dynamical equation can be obtained by noting that the probability that a follower sees right at time $t+1$ is equal to the probability that the majority among m saw right at time t, which in turn equal to the probability that at least $(m+1)/2$ agents saw right at time t, such that:

$$q_{t+1} = \sum_{\ell=(m+1)/2}^{m} C_m^\ell \pi_t^\ell (1-\pi_t)^{m-\ell} \,. \tag{5.12}$$

Combining Eqs. (5.11) and (5.12) yields a dynamical equation of the form:

$$q_{t+1} = F_z(q_t)\,, \tag{5.13}$$

from which the fixed points $q^\star(z), \pi^\star(z)$ can be computed, see Fig. 5.3.

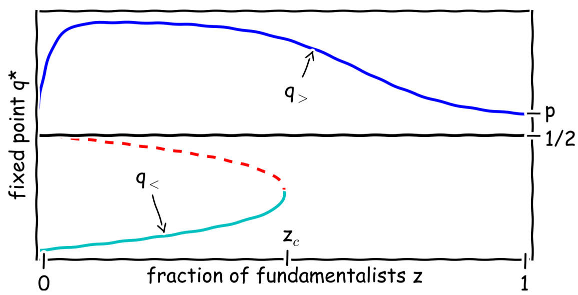


Figure 5.3: Fixed points of the Marsili & Curty game.

- For z large, there is only one attractive fixed point $q^\star = q_> \geq p$. Followers actually increase their probability of being right, herding is efficient as it yields more accurate predictions than information seeking. Further, the performance of followers increases with their number! This *a priori* counterintuitive result comes from the fact that while fundamentalists do not interact, followers benefit from the observation of aggregate behavior. Herders use the information of other herders who have themselves a higher performance than information forecasters.
- However, below a certain critical point z_c, two additional solutions appear, one stable $q_< < 1/2$ and one unstable. The upper solution $q_>$ keeps increasing as z decreases, until it decreases abruptly towards $1/2$ at $z = 0$. The lower solution $q_<$ is always very bad: there is a substantial probability that the initial condition will drive the system towards $q_<$, i.e. the probability to be right is actually lower than a fair coin toss. If herders are trapped in the bad outcome, adding more herders will only self-reinforce the effect, by that making things even worse.

Quite naturally, the next step is to allow agents to choose whether to be followers or fundamentalists, that is allowing for z to depend on time: $z \to z_t$. We consider selfish agents following game theory and aiming at reaching a correct forecast. Further, given that information processing has a cost, as long as $\langle q \rangle > p$, agents will prefer switching from the fundamentalist strategy to the follower strategy ($z_t \downarrow$). Conversely, $z_t \uparrow$ when $\langle q \rangle < p$, and hence we expect that the population will self-organize to a state $z^\dagger$ in which that no agent has the incentive to change his strategy, that is $\langle q(z^\dagger) \rangle = p$. The state $z^\dagger$ is called a *Nash equilibrium*. One can show that $z^\dagger \sim N^{-1/2}$, see [52]. Most importantly, here, such an equilibrium is the result of the simple dynamics of adaptive agents with limited rationality, in contrast with the standard interpretation with forward looking rational agents, who correctly anticipate the behavior of others, and respond optimally given the rules of the game.

This model captures very well the balance between private information seeking and exploiting information gathered by others (herding). When few agents herd, information aggregation is highly efficient. Herding is actually the choice taken by nearly the whole population, setting the system in a *phase coexistence* region ($z^\dagger < z_c$) where the population as a whole adopts either the right or the wrong forecast. See [52] for the details and in particular the effects of including heterogeneity in the agents' characteristics.

5.3.4 Collective decision making with heterogeneities

Aside from his famous segregation model, Thomas Schelling introduced another model which attracted a lot of attention. It is a model for collective decision making with heterogeneous agents having to choose between two options (e.g. joining a riot or not, attending a seminar or not, trading or not, etc.) under the influence of the number of people who made a certain decision at the previous time step. It is an interesting setting to understand some possible nontrivial implications of heterogeneity in collective decision making, such as thresholds. The ingredients of the model are as follows.

- Consider a large number N of agents who must make a binary choice, say join a riot or not.
- Call N_t^+ the number of agents deciding to join at time t and $\phi_t = N_t^+/N$.
- Each agent i makes his mind according to his *conformity threshold* $c_i \in [0,1]$, heterogeneous across agents and distributed according to $\rho(c)$.
- If the number of agents N_t^+ exceeds $N(1-c_i)$, then agent i joins at $t+1$.

In mathematical terms the last point translates into $N_{t+1}^+ = \sum_i \mathbb{1}_{N_t^+ > N(1-c_i)}$ or equivalently

$$\phi_{t+1} = \frac{1}{N}\sum_i \mathbb{1}_{\phi_t > (1-c_i)} = \frac{1}{N}\sum_i \mathbb{1}_{c_i > 1-\phi_t}\,. \tag{5.14}$$

In the continuous limit (large N) this is

$$\phi_{t+1} = \int_{1-\phi_t}^{\infty} \rho(c)\mathrm{d}c = 1 - P_<(1-\phi_t)\,. \tag{5.15}$$

where we have introduced $P_>(u) = 1 - P_<(u) := \int_u^\infty \rho(v)\mathrm{d}v$. The equilibrium fraction of adopters $\phi^\star$ is such that $\phi^\star = 1 - P_<(1-\phi^\star)$. If the only solutions of this equation are the trivial $\phi^\star = 0$ and $\phi^\star = 1$ (see left panel of Fig. 5.4), then any small initial adoption will induce the whole population to adopt the same decision. This corresponds to a situation in which $\rho(c)$ is a monotonous function. If on the other hand some non trivial solution $0 < \phi^\star < 1$ exists, then the fraction of adopters must exceed a certain tipping point to induce full adoption (see right panel of Fig. 5.4). This is the case if $\rho(c)$ is a non-monotonous function, say a Gaussian distribution centred at $x = 1/2$.

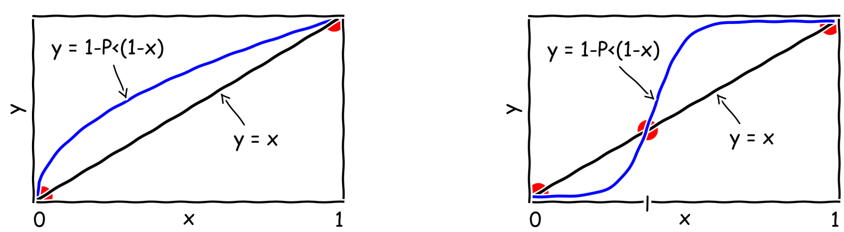
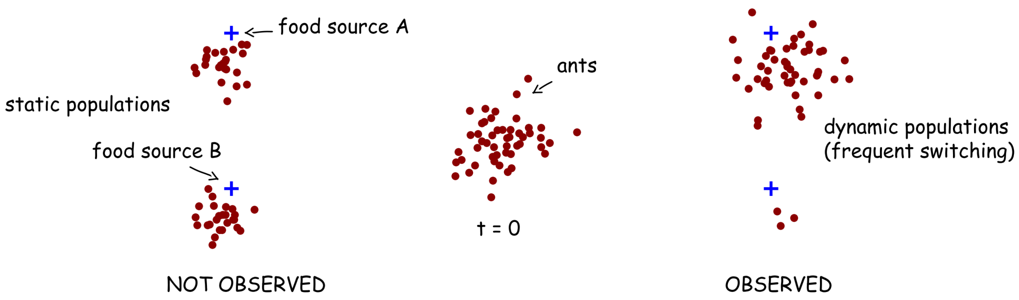


Figure 5.4: Fixed points of the 1st Shelling model.

5.3.5 Kirman's ants, herding and switching

Several decades ago entomologists were puzzled by the following observation. Ants, faced with two identical and inexhaustible food sources, tend to concentrate more on one of them, but periodically switch from one to the other. Such intermittent herding behavior is also observed in humans choosing between equivalent restaurants [53], and in financial markets [54, 55].



Such asymmetric exploitation does not seem to correspond to the equilibrium of a representative ant with rational expectations. The explanation is rather to be looked after in the interactions, or as put by biologists: *recruitment dynamics*. Kirman proposed a simple and insightful model [56] based on *tandem recruitment* to account for such behavior.

- Consider N ants and denote by $n(t) \in [0, N]$ the number of ants feeding on source A at time t.
- When two ants meet, the first one converts the other with probability $1-\delta$.[6]
- Each ant can also change its mind spontaneously with probability ϵ.[7]

The probability p_{n^+} for $n \to n+1$ can be computed as follows. Provided there are $1-n$ ants feeding on B, either one of them changes its mind with probability ϵ or she meets one of the n ants from A and gets recruited with probability $1-\delta$. The exact same reasoning can be held to compute the probability p_{n^-} for $n \to n-1$. Mathematically this translates into:

$$p_{n^+} = \left(1-\frac{n}{N}\right)\left[\epsilon + (1-\delta)\frac{n}{N-1}\right] \tag{5.16a}$$

$$p_{n^-} = \frac{n}{N}\left[\epsilon + (1-\delta)\frac{N-n}{N-1}\right], \tag{5.16b}$$

the probability of $n \to n$ being given by $1 - p_{n^+} - p_{n^-}$. Two interesting limit cases can be addressed.

- In the $\epsilon = 1/2$, $\delta = 1$ (no interaction) case, the problem at hand is tantamount to the *Ehrenfest urn model* or *dog-flea model* [57], proposed in the 1900's to illustrate certain results of the emerging statistical mechanics theory. In this limit, n follows a binomial distribution at equilibrium $\mathbb{P}(n) = C_N^n \epsilon^n (1-\epsilon)^{N-n} = C_N^n / 2^N$.
- When $\delta = \epsilon = 0$, the first ant always adopts the position of the second, and since first/second are drawn with equal probability, the n process is a martingale with absorption at $n = 0$ or $n = N$. Indeed, once all the ants are at the same food source, nothing can convert them ($\epsilon = 0$).[8]

In the general case and large N limit, one can show (see [56]) that there exists an order parameter $O = \epsilon N/(1-\delta)$ such that for $O < 1$, the distribution is bimodal (corresponding to the situation observed in the experiments), for $O = 1$ the distribution is uniform, and for $O > 1$ the distribution is unimodal, see Fig. 5.5.[9] Note that the interesting $O < 1$ regime can be obtained even for weakly persuasive agents ($\delta \lesssim 1$) provided self-conversion ϵ is low enough.

---

[6] Of course who "the first one" is, is unimportant since they could have been drawn in the other order with the same probability.

[7] In the framework of trading, ϵ can represent either exogenous news, or the replacement of the trader by another one.

[8] The probability for absorption at $n = N$ is simply given by n_0/N with n_0 the number of ants feeding on A at $t = 0$.

[9] In the $O < 1$ regime with $\delta = 2\epsilon$ and $N \to \infty$ limit one can prove that the distribution of the fraction $x = n/N$ is given by a *symmetric Beta distribution* $\mathbb{P}(x) \sim x^{\alpha-1}(1-x)^{\alpha-1}$ with $\alpha = \epsilon N$.

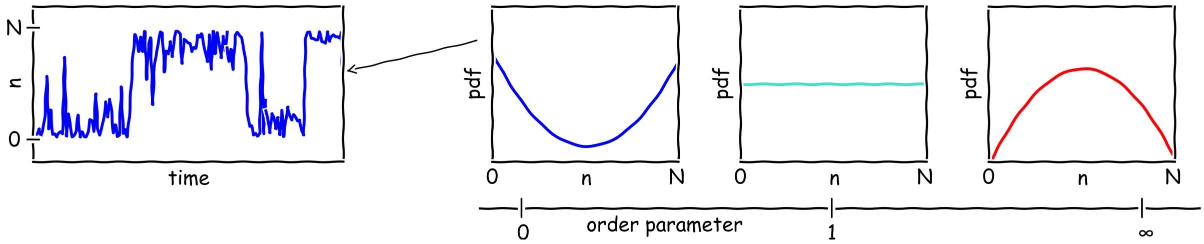


Figure 5.5: (Left) Switching behavior in the $O < 1$ regime. (Right) Ant population distributions for three different value of the order parameter $O = \epsilon N/(1-\delta)$.

When $O > 1$ the system fluctuates around $n = N/2$. When $O < 1$, while the average value of n is also $N/2$, this value has little relevance since the system spends most of the time close to the extremes $n = 0, N$, regularly switching from one to the other.[10]

The most important point is that in the $O < 1$ regime, none of the n states is, itself, an equilibrium. While the system can spend a long time at $n = 0, N$ (locally stationary) these states are by no means equilibria. It is not a situation with multiple equilibria, all the states are always revisited, and there is no convergence to any particular state. In other words, there is perpetual change, the system's natural endogenous dynamics is *out of equilibrium*.[11] Most economic models focus on finding the equilibrium to which the system will finally converge, and the system can only be knocked of its path by large exogenous shocks. Yet, financial markets or even larger economies display a number of regular large switches (correlations, mood of investors etc.) which do not seem to be always driven by exogenous shocks. In the stylised setting presented here such switches are understood endogenously.

Several extensions of the model have been proposed. In particular, the version presented here does not take into account the influence of the proximity of agents; but one can easily limit the scope of possible encounters according to a given communication network J_{ij} (see RFIM model above).

5.4 Feedback effects

Agents are sensitive to past price moves. For many investors, past trends are incentives to buy or sell, as such moves might be reflection of information that other investors possess. This is a rational strategy provided the other agents indeed have extra information, but if not, it is clearly a mechanism for bubbles and excess volatility.

Feedback, or the action of the outputs of a system on its inputs (cause-and-effect loop), is well understood in physics and engineering. Here, we analyse in a stylised setting the possible effects of feedback in financial markets. See also Appendix B for other intricate effects related to memory, such as habit formation and self-fulfilling prophecies.

5.4.1 Langevin dynamics

We take, as above, that returns are proportional to imbalance ϕ_t as:

$$r_t = \frac{\phi_t}{\lambda}, \tag{5.17}$$

where λ is a measure of market depth, and we suppose the following model for the evolution of imbalance:

$$\phi_{t+1} - \phi_t \;=\; a r_t - b r_t^2 - a' r_t - k(p_t - p_{\mathrm{F}}) + \chi \xi_t\,. \tag{5.18}$$

The interpretation of each term in the right hand side of this equation is a follows.

[10]Check `https://rf.mokslasplius.lt/kirman-ants` to play with the model.
[11]The only thing that can be said is that there exists an *equilibrium distribution*.

- $a > 0$ accounts for trend following, past returns amplify the average propensity to buy or sell.
- $b > 0$ accounts for risk aversion. Negative returns have a larger effect than positive returns. Indeed the two first terms can be re-written as $(a - br_t)r_t$ such that it becomes clear that the effective trend following effect increases when $r_t < 0$. The term $-br_t^2$ also accounts for the effect of short term volatility, the increase of which (apparent risk) is expected to decrease demand.
- $a' > 0$ accounts for the market clearing mechanism. It is a stabilising term: price moves clear orders and reduce imbalance.
- $k > 0$ accounts for mean-reversion towards a hypothetic fundamental value p_F; if the price wanders too far above (resp. below) sell (resp. buy) orders will be generated.
- $\chi > 0$ accounts for the sensitivity to random exogenous news ξ_t, with $\langle \xi_t \rangle = 0$, $\langle \xi_t \xi_{t'} \rangle = 2\varsigma_0^2 \delta(t - t')$.

Combining Eqs. (5.17) and (5.18) and taking the continuous time limit $r_t \approx u = \partial_t p$, one obtains a *Langevin equation* for the price *velocity* u of the form:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{\partial V}{\partial u} + \tilde{\xi}_t , \quad \text{avec} \quad V(u) = \kappa(p_t - p_\mathrm{F})u + \alpha \frac{u^2}{2} + \beta \frac{u^3}{3} , \tag{5.19}$$

where we have introduced $\alpha = (a' - a)/\lambda$, $\beta = b/\lambda$, $\kappa = k/\lambda$, and $\tilde{\xi}_t = \chi \xi_t / \lambda$. The variable u thus follows the dynamics of a damped fictitious particle evolving in a potential $V(u)$ with a random forcing $\tilde{\xi}_t$.

5.4.2 Excess volatility, bubbles and crashes

Here, we address different particular cases of increasing complexity (and interest), see Fig. 5.6.

- $\beta = 0$ – In this linear limit, the model reduces to a simple damped harmonic oscillator for the price:
$$\partial_{tt} p + \alpha \partial_t p + \kappa(p - p_\mathrm{F}) = \tilde{\xi}_t .$$
For $\alpha < 0$ the system is naturally unstable (trend following is too strong), but for $\alpha > 0$ one recovers well-known results. In all the following we assume that the system starts at $p_0 = p_\mathrm{F}$. On very short time scales (for which mean reversion to the fundamental value can be neglected) returns simply follow an Ornstein-Uhlenbeck (OU) process with:
$$\langle u_t u_{t+\tau} \rangle = \frac{D}{2\alpha} e^{-\alpha|\tau|} ,$$
where $D \propto (\chi/\lambda)^2$, such that volatility increases with sensitivity to the news, decreases with market depth, and diverges as trend following outweighs liquidity. For $t < \alpha^{-1}$ prices superdiffuse, but as far as $t \gg \alpha^{-1}$ (but small enough that mean reversion is still negligible) returns are uncorrelated and price diffusion is recovered. On larger time scales, mean reversion kicks in and the process becomes an OU process on the price itself. For $\tau \gg \alpha^{-1}$ one has:
$$\left\langle (p_{t+\tau} - p_t)^2 \right\rangle = \frac{D}{\kappa\alpha} \left(1 - e^{-\kappa|\tau|/\alpha}\right) ,$$
that is subdiffusive prices. Note however that on large time scales we expect p_F to be time-dependent and follow a random walk.

- $\beta > 0$, $\alpha > 0$ – Risk aversion is responsible for a local maximum at $u = u^\star = -\alpha/\beta < 0$ and a local minimum at $u = 0$ in the potential $V(u)$ (see left panel of Fig. 5.6). A potential barrier $V^\star := V(u^\star) - V(0) = \alpha u^{*2}/6$ separates a *metastable* region around $u = 0$ from an unstable region $u < u^\star$. Starting from $p_0 = p_\mathrm{F}$ the particle oscillates around $u = 0$ until an *activated* event driven by $\tilde{\xi}_t$ brings the particle to $u^\star$ after which $u \to -\infty$, that is a *crash* induced by the amplification of sell orders due to the risk aversion term. The typical time before the crash scales as $\exp(V^\star/D)$. Note that, here, a crash occurs due to a succession of unfavourable events which add up to push the system below the edge, and not due to a single large event in particular. Also note that, in practice, volatility feedback effects would increase fluctuations before the crash, by that increasing D and thus lowering further the crash time.

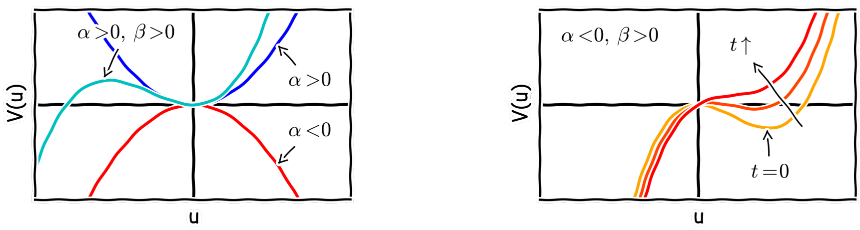


Figure 5.6: Langevin potential $V(u)$ for different values of the parameters.

- $\beta > 0$, $\alpha < 0$ – Taking now trend following to be large compared with the stabilising effects while keeping risk aversion yields a very interesting situation. Starting at $t = 0$ from $p_0 = p_F$, the potential $V(u)$ displays a local maximum at $u = 0$ and a local minimum at $u = u^\dagger = -\alpha/\beta$ with $V^\dagger = \left|V(u^\dagger)\right|$. In the beginning the particle oscillates around $u^\dagger > 0$ and the price increases linearly on average $\langle p_t - p_F \rangle \sim u^\dagger t$ with no economic justification whatsoever. This is a *speculative bubble*, growth is self-sustained by the trend following effect. However, as time passes the potential is modified (due to the increasing slope of the linear term $\kappa(p_t - p_F)u$) and $V^\dagger$ decreases accordingly (see right panel of Fig. 5.6). When the local minimum ceases to exist, the bubble bursts, that is $u^\dagger \to -\infty$). The bubble lifetime $t^\dagger$ is such that $V^\dagger = u^\dagger t^\dagger$ which is $t^\dagger = -\alpha/(4\kappa)$; as expected it increases with the amplitude of trend following and decreases with that of mean reversion.

For a deeper analysis of destabilising feedback loops in financial markets, see our recent work on endogenous liquidity crises [58].

5.5 The minority game

Here we present the *Minority Game* [59] to illustrate why complex systems made of heterogeneous and interacting agents can often be drawn to criticality. It is inspired from the theory of spin-glasses and disordered systems.

The ingredients of the model (in its simplest setting) are as follows:

- Consider a large number N of agents who must make a binary choice, say ± 1.
- At each time step, a given agent wins if and only if he chooses the option that the *minority* of his fellow players also choose. By definition, the number of winners is thus always $< N/2$.
- At the beginning of the game, each agent is given a set of strategies fixed in time (he cannot try new ones or slightly modify the ones he has in order to perform better). A strategy takes as input the string of, say M past outcomes of the game, and maps it into a decision. The total number of possible strategies is 2^{2^M} (the number of strings is 2^M and to each of them can be associated $+1$ or -1). Each agent's set of strategies is randomly drawn from the latter. While some strategies may be by chance shared, for moderately large M, the chance of repetition is exceedingly small.
- Agents make their decision based on past history. Each agent tries to rank his strategies according to their past performance, e.g. by giving them scores, say $+1$ every time a strategy gives the correct result, -1 otherwise. *A crucial point here is that he assigns these scores to all his strategies depending on the outcome of the game, as if these strategies had been effectively played, by that neglecting the fact that the outcome of the game is in fact affected by the strategy that the agent is actually playing.*[12]

The game displays very interesting dynamics which fall beyond the scope of this course, see [59–61]. Here, we focus on the most striking and generic result. We introduce the degree of predictability H

[12] Note that this is tantamount to neglecting impact and crowding effects when backtesting an investment strategy.

(inspired from the *Edwards-Anderson* order parameter in spin-glasses) as:

$$H = \frac{1}{2^M} \sum_{h=1}^{2^M} \langle w|h \rangle^2 , \tag{5.20}$$

where $\langle w|h \rangle$ denotes the average winning choice conditioned to a given history (M-string) h. One can show that the number of strategies does not really affect the results of the model, and that in the limit $N, M \gg 1$ the only relevant parameter is $\alpha = 2^M/N$. Further, one finds that there exists a critical point α_c (≈ 0.34 when the number of strategies is equal to 2) such that for $\alpha < \alpha_c$ the game is unpredictable ($H = 0$) whereas for $\alpha > \alpha_c$ the game becomes predictable in the sense that conditioned to a given history h, the wining choice w is statistically biased towards $+1$ or -1 ($H > 0$). In the vocabulary of financial markets, the unpredictable and predictable phases can be called the *efficient* and *inefficient* phases respectively.

At this point, it is easy to see that, by allowing the number of players to vary, the system self-organises such as to lie in the immediate vicinity of the critical point α_c. Indeed, for $\alpha > \alpha_c$, corresponding to a relatively small number of agents N, the game is to some extent predictable and thus exploitable. This is an incentive for new agents to join the game, that is $N \uparrow$ or equivalently $\alpha \downarrow$. On the other hand, for $\alpha < \alpha_c$, corresponding to a large number of agents N, the game is unpredictable and thus uninteresting to extract profits. Agents leave the game, that is $N \downarrow$ or equivalently $\alpha \uparrow$. This mechanism spontaneously tunes $\alpha \to \alpha_c$.

By adapting the rules the Minority Game can be brought closer to real financial markets, see [61]. The critical nature of the problem around $\alpha = \alpha_c$ leads to interesting properties, such as fat tails and clustered volatility. Most importantly, the conclusion of an attractive critical point (or self-organised criticality) is extremely insightful: it suggests that markets operate close to criticality, where they can only be marginally efficient!

6

Dimensional analysis in finance

Benoit Mandelbrot was the first to propose the idea of *scaling* in the context of financial markets [62], a concept that blossomed in statistical physics well before getting acceptance in economics, for a review see [63]. In the last thirty years, many interesting scaling laws have been reported, concerning different aspects of price and volatility dynamics. In particular, the relation between volatility and trading activity has been the focus of many studies, see e.g. [64–70] and more recently [21, 71, 72].

6.1 Vaschy-Bukingham π-theorem

Dimensional analysis states that any law relating different observables must express one particular dimensionless (or unit-less) combination of these observables as a function of one or several other such dimensionless combinations. More precisely, the *Vaschy-Bukingham π-theorem* states that *if* a physical equation involves n physical variables and m non-linearly dependent fundamental units, *then* there exists an equivalent equation involving $n - m$ dimensionless variables constructed from the original variables.

6.2 An example in physics: The ideal gas law

The simplest illustrative example might be the *ideal gas law*, that amounts to realising that pressure P times volume $\mathcal{V}$ has the dimension of an energy. Hence $P\mathcal{V}$ must be divided by the thermal energy RT of a mole of gas to yield a dimensionless combination. The right-hand side of the equation must be a function of other dimensionless variables, but in the case of non-interacting point-like particles, there is none – hence the only possibility is $P\mathcal{V}/RT = \text{cst}$.

Deviations from the ideal gas law are only possible because of the finite radius of the molecules, or the strength of their interaction energy, that allows one to create other dimensionless combinations, and correspondingly new interesting phenomena such as the liquid-gas transition.

6.3 An example in finance: The *ideal market law*

Kyle and Obizhaeva recently proposed a bold but inspiring hypothesis, coined the *trading invariance principle* [73]. In the search of an *ideal market law*, several possible observables that characterise trading come to mind:

- the share price p in \$ per share,
- the square volatility of returns σ^2 in $\%^2$ per unit time,
- the volume of individual orders Q in shares,
- the trading volume V in shares per unit time,
- and the trading cost C in \$.[1]

[1] Other more microstructural quantities might come into play, such as the spread (in \$ per share), the tick size (in \$) that sets the smallest possible price change, the lot size (in shares) that sets the smallest amount of exchanged shares, the average volume available at the best quotes, and perhaps other quantities as well.

Let us assume there exists an equation relating these $n = 5$ variables of the form:

$$f(p, \sigma^2, Q, V, C) \;=\; 0\,. \tag{6.1}$$

The number of non-linearly dependent units m can be computed as the rank of the following matrix:

$$\begin{array}{c} \\ p \\ \sigma^2 \\ Q \\ V \\ C \end{array} \begin{array}{c} \begin{array}{ccc} \$ & \textit{shares} & T \end{array} \\ \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$

Here $m = 3$ such that according to the Vaschy-Bukingham π-theorem there exists an equation equivalent to Eq. (6.1) involving $n - m = 2$ dimensionless variables. One can for example choose:

$$\frac{pQ}{C} \;=\; g\left(\frac{Q\sigma^2}{V}\right), \tag{6.2}$$

with g a dimensionless function that cannot be determined on the basis of dimensional analysis only.

Invoking the Modigliani-Miller theorem which argues that capital restructuring between debt and equity should keep $p \times \sigma$ constant, while not affecting the other variables. This suggests that $g(x) \sim x^{-1/2}$, finally leading to the so-called *3/2 law*:

$$W \sim CN^{3/2}, \tag{6.3}$$

where we introduced the *trading activity* or *exchanged risk* $W := PV\sigma$ and the trading rate $N = V/Q$.

6.4 Empirical evidence

Several empirical studies (see e.g. [73–76]) have shown that the 3/2 *law* holds remarkably well on average at the single-trade scale and metaorder scale on different asset classes, see Fig. 6.1.

Days of anomalously high volatility display deviations from the trading invariance principle, which could be exploited by regulators as yet another indicator for market anomalies. Similar to the deviations away from the ideal gas law in the example discussed above, deviations from trading invariance may suggest that additional microstructural variables must be involved in the search of a relation generalising the ideal market law to all market conditions. In some regimes, the bid-ask spread and the

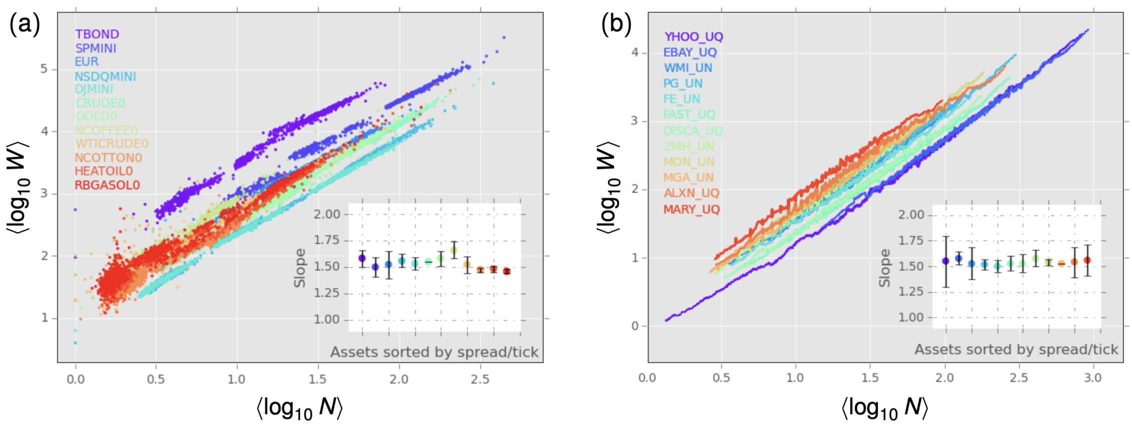


Figure 6.1: Trading activity against trading frequency for (a) 12 futures contracts and (b) 12 US stocks, from [75]. The insets show the slopes α obtained from linear regression of the data, all clustered around 3/2.

tick size, among other things, could play an important role – like the molecular size in the ideal gas analogy. Besides, note that while the Modigliani-Miller theorem is intended for stocks, the 3/2 law unexpectedly does hold also for futures (see Fig. 6.1) for which the Modigliani-Miller argument does not have any theoretical grounds.

Conclusions

Dimensional analysis is a powerful tool that has proved its worth in physics, but that is yet under-exploited in finance and economics. The prediction that *exchanged risk* W, also coined *trading activity*, namely $W =$ price × volume × volatility, scales like the 3/2 power of trading frequency N, is well supported by empirical data, both at the trade-by-trade and metaorder levels. The dimensionless quantity $W/(CN^{3/2})$, where C denotes the trading costs, is a good candidate for a trading invariant across assets and time. Finally, let us stress that unveiling the mysteries of the enigmatic 3/2 law from the microscopic 3/2 law is yet to be done: is the 3/2 an approximate happy coincidence, or is there a deep principle behind it?

7

Market impact of metaorders

As discussed in Chapter 3, the way trades affect prices (market impact) stands at the very heart of the price formation process. Here we address the impact of *metaorders*.

7.1 Measuring metaorder impact

To cope with the lack of liquidity (see Chapter 3), institutional investors must split their large orders into small pieces, called *child orders* and traded incrementally during a given time interval T (typically a few hours but can be as long as a few days or even weeks). Such a succession of small trades, all executed in the same direction (either buys or sells) and originating from the same market participant is called a *metaorder*.

In the following we shall denote by $\epsilon = \pm 1$ the direction (or sign) of the metaorder, and by Q its total volume:

$$Q = \sum_{t=1}^{T} q_t \approx \int_0^T m_t \, \mathrm{d}t \,, \tag{7.1}$$

with q_t the volumes of the child orders, m_t the *execution rate*, and T the *execution horizon*.

The ideal experiment to measure the impact of a metaorder would be to compare two different versions of history: one in which the metaorder was executed, and one in which it was not, all other things being equal. Of course this is not possible in reality and one estimates the impact as the mean price difference between the beginning ($t = 0$) and the end ($t = T$) of the metaorder:

$$I(Q, T) := \langle \epsilon \cdot (p_T - p_0) | Q \rangle \,. \tag{7.2}$$

This quantity is often called the *peak impact*. As argued in Chapter 3, one expects $I > 0$. It is also convenient to define the *impact path* as:

$$I_t := I(Q_t, t)\,, \quad \text{with } Q_t = \int_0^t m_t \mathrm{d}t \,, \tag{7.3}$$

and where one can also have $t > T$. Indeed the impact path after the execution is an important quantity showing nontrivial behavior on which we shall comment below. We define the *permanent impact* as:

$$I_\infty := \lim_{t \to \infty} I(Q, t) \,. \tag{7.4}$$

Note that the permanent impact receives two types of contribution: one depending on the motivation to execute the metaorder in the first place (often called *prediction impact*), and the other coming from the possibly permanent mechanical reaction of the market to the trades. Here we are interested in the latter. Finally, we also define the *execution cost* (also called *slippage cost* or *execution shortfall*,[1] see Chapter 4) as the volume weighted average premium paid by the trader executing the metaorder:

$$C(Q, T) := \left\langle \textstyle\sum_{t=1}^{T} q_t (p_t - p_0) | Q \right\rangle \approx \left\langle \textstyle\int_0^T m_t (p_t - p_0) \mathrm{d}t | Q \right\rangle \,. \tag{7.5}$$

[1]The execution shortfall is often defined as C/Q.

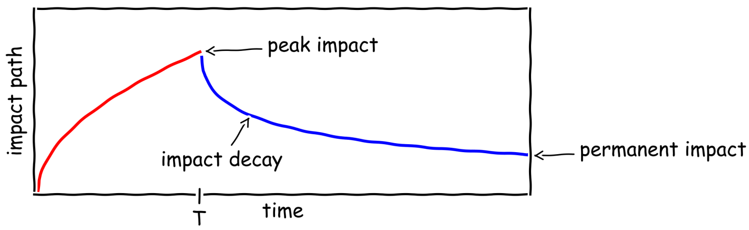


Figure 7.1: Average impact path during a metaorder execution (red) and after (blue).

Note that one would expect all these quantities to depend on the whole execution profile $\{q_t\}$ and not just on the aggregate volumes Q or Q_t. However, in practice one observes impact is only very weakly affected by the precise shape of the execution schedule.

Also note that measuring metaorder impact requires proprietary data listing which child orders belong to which metaorders, together with q_t, p_t. In addition, to be safe from possible idiosyncratic biases one would need metaorders from a large number of different market participants (broker data). Available data sets with such rich and detailed data are very scarce (and quite pricey).

7.2 The square root law

Naively, one would expect the impact of a metaorder to scale linearly with its volume. Many simple models of price impact predict precisely this behaviour, see e.g. the famous *Kyle* model [77], or the *Santa Fe* model [78]. However, empirical data reveals that this scaling is not linear, but concave and close to a square-root.

This *square-root law*, which establishes that metaorders impact the price as $\sim \sqrt{Q}$ and not proportionally to Q, quite indisputably stands amongst the most remarkable and well established stylised facts of modern finance [79–88]. In the following we present it and discuss several of its important consequences.

More precisely, for a given asset the absolute expected average price return between the beginning and the end of a metaorder follows:

$$I(Q,T) = Y\sigma_{\mathrm{T}}\left(\frac{Q}{V_{\mathrm{T}}}\right)^{\delta}, \tag{7.6}$$

where Y is a numerical factor of order one, σ_{T} and V_{T} respectively denote volatility on scale T and total traded volume during the execution, and $\delta \approx 0.5 < 1$ bears witness of the concave nature of price impact. Actually, the impact does not depend on T.[2] Indeed, taking that $\sigma_T = \sigma_0\sqrt{T}$ and $V_T = V_0 T$ yields:

$$\sigma_{\mathrm{T}}\left(\frac{Q}{V_{\mathrm{T}}}\right)^{\delta} = \sigma_0\left(\frac{Q}{V_0}\right)^{\delta} T^{1/2-\delta},$$

so that with $\delta = 1/2$ all time dependence disappears. Conventionally the square root law is written with daily volatility σ_{d} and daily traded volume V_{d}:

$$I(Q) = Y\sigma_{\mathrm{d}}\left(\frac{Q}{V_{\mathrm{d}}}\right)^{\delta}. \tag{7.7}$$

Equation (7.7) is surprisingly universal: it is found to be to a large degree independent of details such as the asset class (including equities, futures, FX, options, and even Bitcoin), market venue, execution style (limit orders or market orders or both), microstructure (small and large ticks), and time period.

[2]As we shall see in the following Chapter, there is a residual dependence on T in the very small participation ratio regime.

In particular, the advent of electronic markets and High Frequency Trading (HFT) has not altered the square-root behaviour, in spite of radical changes: before 2005 liquidity was mostly provided by market makers, and after dominated by HFT. A few additional remarks are in order:

- The square root law is everything but intuitive. Indeed, it means that impact is not additive, or in other words that "$2Q \neq Q+Q$" (the first half of the trade has a much larger impact than the second).
- Concave impact also means that small orders have very large impact (relatively speaking). A metaorder taking 0.1% of the daily volume moves the price by $\sqrt{0.1\%} \approx 3\%$ of the daily volatility.
- For reasons that shall become clear in the following Chapter, the square root is expected to fail when $Q < V_{\text{best}}$ (with V_{best} the average volume at the best quotes) and $T <$ a few seconds (the typical microstructure time) and $T >$ a few days.
- As we shall see in Chapter 8, stylised agent-based models (ABM) are of great value to gain insight into the origins of the square root law.

7.3 Slippage costs, orders of magnitude

Trading costs are conventionally divided into (i) direct costs (fees) and (ii) indirect costs (spread and market impact). In general, the fees that must be paid to access a given market account for a very small fraction and shall therefore be neglected in the following. Impact or slippage costs are convex functions of traded volume $\sim QI(Q)$, and thus outweigh the linear spread costs $\sim QS$ (with S the average spread) for large enough volumes. For large enough volumes impact costs consume a substantial fraction of the expected profits.[3] Using Eq. (7.7) for constant execution rate $m_t = m_0 := Q/T$ one can right for $t = \phi T < T$:

$$I_t = I(\phi Q) = \sqrt{\phi} I(Q),$$

where we have used Eq. (7.7). Combining with Eq. (7.5), one obtains:

$$C(Q) = \int_0^T m_0 I_t \mathrm{d}t = Q \int_0^1 \sqrt{\phi} I(Q) \mathrm{d}\phi = \frac{2}{3} Q I(Q). \tag{7.8}$$

Square-root impact therefore leads to slippage costs in $Q^{3/2}$. To note, the execution shortfall is 2/3 of the peak impact (with linear impact one finds $1/2 < 2/3$, by that underestimation trading costs).

While not proven to be rigorously impossible in reality, the *no free lunch* theorem states that there is no strategy that can mechanically pump money out of markets. In other words, the cost of a round-trip trade should on average be positive: $\langle \mathscr{C} \rangle \geq 0$, very much like the entropy of an isolated thermodynamic system. This property, also referred to as *absence of price manipulation* or *arbitrage-freeness*, is a constraint for any viable model of price impact.

Institutional trading is generally divided in two stages: the decision stage during which one decides to buy or sell based on some information about the future price of the asset (say a signal μ), and the execution stage during which the metaorder is conducted. Neglecting fees, the total gain $\mathscr{G}$ can be computed as:

$$\mathscr{G}(Q) \approx \mu Q - \frac{2}{3} Q I(Q) - Q \frac{S}{2} = \tilde{\mu} Q - \frac{2}{3} Y \sigma_{\mathrm{d}} Q \left(\frac{Q}{V_{\mathrm{d}}} \right)^{\delta}, \tag{7.9}$$

with $\tilde{\mu} = \mu - S/2$. Therefore, there exists a volume $Q_{\max}(\tilde{\mu})$ above which trading will always generate losses, $\mathscr{G}(Q > Q_{\max}) < 0$. Maximum gain is obtained for $Q^\star(\tilde{\mu})$ such that $\partial_Q \mathscr{C}\big|_{Q^\star} = 0$. One obtains:

$$Q^\star = V_{\mathrm{d}} \left(\frac{\tilde{\mu}}{Y \sigma_{\mathrm{d}}} \right)^2, \quad Q_{\max} = \frac{9}{4} Q^\star . \tag{7.10}$$

[3] With $\sigma_{\mathrm{d}} \approx 1\%$ and a participation ratio of say 10^{-3} to 10^{-2}, one obtains using Eq. (7.8) that the average slippage is as large as 2 to 6 basis points. Taking an institutional investor with say $10B AUM, a turnover of say 10 days and a leverage factor 10, this means that by just trading randomly one looses statistically $500M to $1.5B per year (there are $\approx$ 250 trading days in one year).

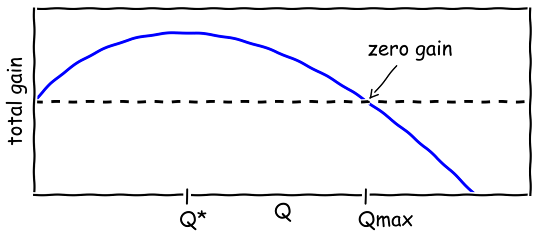


Figure 7.2: Average gain in the presence of impact.

7.4 The permanent impact conundrum

Upon completion of a metaorder execution, the buying or selling pressure stops and the price reverts first abruptly and then slowly towards a plateau value I_∞. Whether such a plateau is zero (no permanent impact $I_\infty = 0$), or some impact persists permanently ($I_\infty \neq 0$) is a question without a unique answer.

Using proprietary data, Brokmann *et al.* [86] concluded that permanent impact is vanishingly small, see also [87]. Farmer's fair pricing theory [89] states that permanent impact should be equal to 2/3 of its peak value, with some empirical support [90–92]. In addition, no-arbitrage arguments [77, 93, 94] imply that permanent impact must be *linear* in the executed volume Q. It should be noted that measuring permanent impact is a tricky business due to the very slow impact decay (often called *long-range resilience*) and the increasingly large noise level for $t \gg T$, which notably requires a very large metaorder database to circumvent possible biases. For recent empirical insights on impact decay and permanent impact, see [95]. Using the ANcerno dataset,[4] we show that while the impact at the end of the same day is on average $\approx 2/3$ of the peak impact, the decay continues the next days, initially following a power-law function, and converges to a non-zero asymptotic value at long time scales (≈ 50 days) equal to $\approx 1/2$ of the impact at the end of the first day, that is $I_\infty \approx I_T/3$.

Conclusions

The square root impact law contradicts that predicted by many models. While it took several years to push its way up (there are still a few dissenting voices), the square root law is an interesting example in which empirical data compelled the community to accept that reality was fundamentally different from theory.

[4]The ANcerno Ltd (formerly the Abel Noser Corporation) database is a very large dataset containing more than 10 million metaorders, executed on the US equity market and issued by a diversified set of institutional investors, see `www.ancerno.com` for details.

8

Latent order book models for price changes

Statistical physics aims at overcoming the gap between microscopic dynamics and aggregate behavior (e.g. deriving state equations such as the ideal gas law from kinetic theory). Analysing the dynamics of the order book is precisely looking at the system at the microscopic level, from which we may model agent's actions (order flow events), carefully upscale to the aggregate level, and hopefully understand stylised facts on price dynamics such as the square root impact law (see Chapter 7).

8.1 Coarse-graining

Let us recall that while the price impact of individual trades is non universal and strongly depends on the microstructure, the impact of metaorders is highly universal and quite insensitive to microstructural changes.

This indicates that microscopic details are likely irrelevant to account for the square root impact law, and suggests a *coarse-grained* approach.[1] The continuum Navier-Stokes equations in hydrodynamics can be obtained by coarse-graining over the microscopic molecular degrees of freedom of the liquid molecules; one obtains that one emergent scalar parameter – the viscosity – encodes all the complexity of the microscopic scale and suffices to describe the dynamics of the macroscopic systems. Here, we apply a similar approach by coarse-graining over the microscopic degrees of freedom of the order book in order to build a "hydrodynamic model" for low frequency market dynamics.

8.2 Revealed and latent liquidity

Liquidity is often defined as the resistance of a given asset price to move in response to incoming orders. In practice liquidity equivalently refers to the volume of limit orders near the best quotes. The order book of a liquid (resp. illiquid) asset has substantial (resp. little) limit order volume near the price, acting as a strong barrier (resp. a mere hump) to incoming market orders. This is the instantaneous liquidity publicly displayed in the limit order book, or *revealed liquidity*. In Chapter 3 we argued that revealed liquidity is very small, typically less than 1% of the daily traded volume in stock markets.

The concept of *latent liquidity* builds upon the idea that revealed liquidity chiefly reflects the activity of high frequency market participants that act as intermediaries between much larger *latent* volume imbalances of low frequency actors. Latent liquidity is progressively disclosed during the day as the revealed liquidity gets consumed, very much like the melting tip of an iceberg. In some sense financial markets can be seen as the arena of a collective hide-and-seek game between buyers and sellers who keep their intentions secret to avoid giving away precious private information, until one's reservation

[1] A *fine-grained* description of a system aims at describing in detail its microscopic dynamics. A *coarse-grained* description is one in which the irrelevant microscopic details are smoothed over to retain only the minimal relevant parameters to account for the aggregate behavior.

price is such that the probability to get executed is large enough to warrant posting one's order. The insensitivity of the square-root law to the high frequency dynamics of prices suggests that its origin should lie in some general properties of the low frequency, large scale dynamics of latent liquidity, rather than in its short-lived revealed counterpart.

8.3 Geometrical arguments

We define the coarse-grained latent volume densities of limit orders in the order book $\rho_{\mathrm{B}}(x,t)$ (bid side) and $\rho_{\mathrm{A}}(x,t)$ (ask side) at price x and time t.

Before we engage in the modeling of the dynamics of latent liquidity, we ask the following insightful question: in a static world ($\rho(x,t)=\rho(x)$) what should the shape of the latent order book be to recover square root impact? Simple geometry shows that while a flat latent order book $\rho(x)=v_0$ with v_0 a constant would yield linear impact $I(Q)=Q/v_0$, a linear order book $\rho_{\mathrm{B/A}}(x)=\pm\mathscr{L}x$ would be consistent with square root impact:

$$I(Q)=\sqrt{\frac{2Q}{\mathscr{L}}}\,, \tag{8.1}$$

where $\mathscr{L}$ denotes the *latent liquidity* of the market. Indeed the steeper the latent order book (large $\mathscr{L}$) the weaker the impact.

8.4 A reaction-diffusion model

As a precise mathematical incarnation of the latent order book idea [81, 85, 96], the zero-intelligence model of Donier *et al.* [97] was quite successful at providing a theoretical underpinning to the square root impact law.

8.4.1 Latent liquidity dynamics

In this model, the latent volume densities $\rho_{\mathrm{B}}(x,t)$ and $\rho_{\mathrm{A}}(x,t)$ evolve according to the following rules (see Fig. 8.1a).[2]

- Latent orders diffuse with diffusivity constant D (random re-evaluation of the reservation price).
- Latent orders are canceled with multiplicative rate ν (participants reduce their trading intentions or leave the market).
- New intentions are deposited with additive rate λ (new arrivals).
- When a buy intention meets a sell intention they are instantaneously matched: $\mathrm{A}+\mathrm{B}=\varnothing$. We implicitly assume that latent orders are revealed in the vicinity of the trade price p_{t}.
- The trade price p_{t} is conventionally defined through the equation $\rho_{\mathrm{B}}(p_{\mathrm{t}},t)=\rho_{\mathrm{A}}(p_{\mathrm{t}},t)$.

According to this set of rules, the reduced latent order book density $\phi(x,t)=\rho_{\mathrm{B}}(x,t)-\rho_{\mathrm{A}}(x,t)$ solves:

$$\begin{aligned}\partial_t\phi &= D\partial_{xx}\phi-\nu\phi+s(x,t)\,,\quad \text{with}\quad s(x,t)=\lambda\,\mathrm{sgn}(p_t-x)\,, && (8.2)\\ \phi(p_t,t) &= 0\,, && (8.3)\end{aligned}$$

where the sign function $\mathrm{sgn}(x)=\mathbb{1}_{x\geq 0}-\mathbb{1}_{x<0}$ indicates that buy (resp. sell) latent orders can only be deposited in the bid (resp. ask) side of the book. Setting $\xi=x-p_t$, the resulting stationary latent order book reads:

$$\phi^{\mathrm{st}}(\xi)=-\frac{\lambda}{\nu}\,\mathrm{sgn}(\xi)\big[1-e^{-|\xi|/\xi_{\mathrm{c}}}\big], \tag{8.4}$$

[2] Note that the variable x denotes the reservation price relative to the informational price component $\hat{p}_t$ such that the true reservation price reads $p=\hat{p}_t+x$. We here assume that $\hat{p}_t$ encodes all informational aspects of prices and itself performs an additive random walk, see [97] for a detailed discussion.

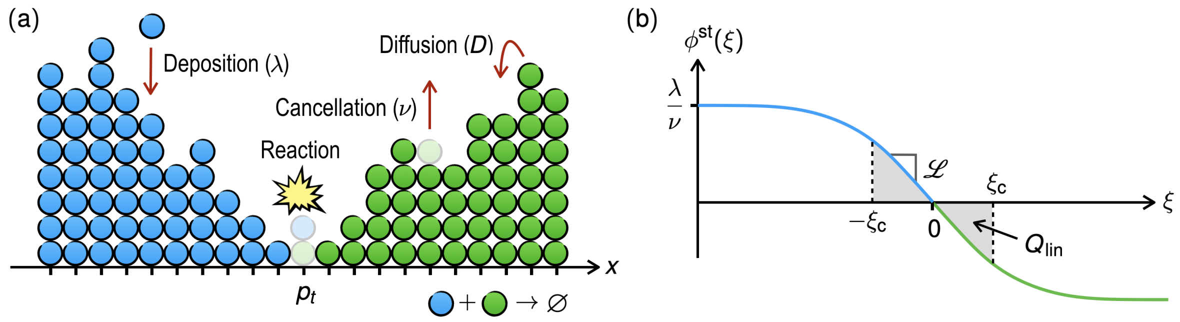


Figure 8.1: (a) Reaction-diffusion setup for the latent order book. (b) Stationary latent order book densities.

where $\xi_c = \sqrt{D\nu^{-1}}$ denotes the typical length scale below which the order book can be considered to be linear (see Fig. 8.1b):

$$\phi^{\text{st}}(\xi) \approx -\mathscr{L}\xi\,. \tag{8.5}$$

The slope $\mathscr{L} = \lambda/\sqrt{\nu D}$ is directly related to the total transaction rate J through:

$$J := D\partial_\xi \phi^{\text{st}}(\xi)|_{\xi=0} = D\mathscr{L}\,. \tag{8.6}$$

Below, we focus on the *infinite memory* limit, namely $\nu, \lambda \to 0$ while keeping $\mathscr{L} \sim \lambda\nu^{-1/2}$ constant, such that the latent order book becomes exactly linear since in that limit $\xi_c \to \infty$. This limit considerably simplifies the mathematical analysis.

8.4.2 Microscopic derivation

Here we provide a microscopic derivation of the diffusion equation for the latent order book (see Eq. 8.2). The contributions of cancellations and depositions being rather trivial, we focus on the diffusion term ($\nu, \lambda = 0$). Assume that between t and $t + \delta t$, each agent i revises his reservation price according to

$$p_i \to p_i + \beta_i f_t + \eta_{i,t}\,,$$

with:

- f_t common to all agents representing e.g. some public information,
- β_i the sensitivity of agent i to the exogenous signal with distribution $P_\beta(\beta)$ and $\langle\beta\rangle_i = 1$,
- and $\eta_{i,t}$ an independent random variable both across different agents (idiosyncratic contribution) and time with distribution $P_\eta(\eta)$ centred $\langle\eta\rangle = 0$ and variance Σ^2.

Assuming that within each price interval $x, x + \mathrm{d}x$ lie latent orders from a large number of agents, the density of latent orders $\rho(x,t)$ therefore evolves according to:

$$\begin{aligned} \rho(x, t+\delta t) &= \int \mathrm{d}\beta \int \mathrm{d}\eta \int \mathrm{d}x' P_\beta(\beta) P_\eta(\eta)\rho(x',t)\delta(x - x' - \beta f_t - \eta) \\ &= \int \mathrm{d}\beta \int \mathrm{d}\eta P_\beta(\beta) P_\eta(\eta)\rho(x - \beta f_t - \eta, t)\,. \end{aligned} \tag{8.7}$$

Performing a second order Kramers-Moyal expansion of the above equation yields:

$$\rho(x,t+\delta t) - \rho(x,t) = -f_t\partial_x\rho + \frac{1}{2}\left(\langle\beta^2\rangle f_t^2 + \Sigma^2\right)\partial_{xx}\rho + \ldots\,, \tag{8.8}$$

and assuming that formally $f_t = V_t\delta t$ and $\Sigma^2 = 2D\delta t$ leads to:

$$\partial_t\rho = -V_t\partial_x\rho + D\partial_{xx}\rho\,. \tag{8.9}$$

Note that the drift term can be absorbed through the change of variables $x \to x - \int_0^t V_{t'}\mathrm{d}t'$, which amounts to changing the frame of reference to that of the exogenous signal, by that allowing to focus on the purely mechanical component.

8.4.3 Market impact

To compute the market impact, a buy (resp. sell) meta-order of volume Q is introduced as an extra point-like source of buy (resp. sell) particles with intensity rate m_t. The source term in Eq. (8.2) then becomes: $s(x,t) = \lambda\,\mathrm{sgn}(p_t - x) + m_t\delta(x - p_t)\cdot \mathbb{1}_{[0,T]}$, where T denotes the time horizon of the execution. In all the following we shall focus on buy meta-orders – without loss of generality since within the present framework everything is perfectly symmetric. The general solution of Eq. (8.2) reads:

$$\phi(x,t) \quad = \quad \left[\mathcal{G}_\nu * (\phi_0\delta(t) + s)\right](x,t)\,, \tag{8.10}$$

where $*$ denotes the space and time convolution product, $\phi_0(x) = \phi(x,0)$ is the initial condition, and $\mathcal{G}_\nu(x,t) = e^{-\nu t}\mathcal{G}(x,t)$ with $\mathcal{G}$ the diffusion kernel:

$$\mathcal{G}(x,t) = \frac{\mathbb{1}_{t>0}}{\sqrt{4\pi D t}}\exp\left[-\frac{x^2}{4Dt}\right]. \tag{8.11}$$

The price trajectory can then be computed from the combination of Eqs. (8.3) and (8.10). With infinite memory $\nu, \lambda \to 0$, and taking $\phi_0(x) = \phi^{\mathrm{st}}(x)$, one can show that the impact path $I_t = p_t - p_0$ solves the following self consistent equation:

$$I_t = \frac{1}{\mathcal{L}}\int_0^{t\wedge T} \frac{m_{t'}}{\sqrt{4\pi D(t-t')}}\exp\left[\frac{(I_t - I_{t'})^2}{4D(t-t')}\right]\mathrm{d}t'\,. \tag{8.12}$$

Further, focusing on the case of constant participation rate ($m_t = m_0 = Q/T$) one can show that market impact reduces to:

$$I(Q) = \sqrt{\frac{Q}{\mathcal{L}}}\,\mathcal{F}(\eta)\,, \tag{8.13}$$

where $\eta := m_0/J$ is the *participation ratio* and the scaling function $\mathcal{F}(\eta) \approx \sqrt{\eta/\pi}$ for low participation ($\eta \ll 1$) and $\approx \sqrt{2}$ for high participation ($\eta \gg 1$)[3] with a smooth crossover at $\eta^\star \sim 1$. Hence, recalling $Q = m_0 T$, $I(Q)$ is linear in Q for small Q at fixed T, and crosses over to a square-root for large Q. Provided $\nu T \ll 1$, finite memory corrections ($\nu \neq 0$) are easily computed, see [98]. For $\nu T \gg 1$ the latent liquidity is very short-lived (Markovian limit) and the impact becomes linear regardless of participation ratio.

8.4.4 Finite memory and permanent impact

Regarding impact dynamics after the metaorder execution ($t > T$, see Sec. 7.4), the infinite memory limit yields inverse square root relaxation of impact as function of time and zero permanent impact. For finite memory ($\nu T \ll 1$), we find that permanent impact is nonzero. This result can be interpreted as follows. At the end of execution (when the peak impact is reached), the impact starts decaying towards zero in a slow power law fashion until approximately $t \sim \nu^{-1}$, beyond which all memory is lost (since the latent book has been globally renewed). Impact cannot decay anymore, since the previous reference price has been forgotten. Furthermore, the permanent impact is found to be linear in the executed volume:

$$I_\infty = \xi_{\mathrm{c}}\frac{Q}{2Q_{\mathrm{lin}}}\,, \tag{8.14}$$

consistent with some theoretical predictions [77, 93, 94]. In the Markovian limit ($\nu T \gg 1$) all memory is already lost at the end of the execution and the permanent impact trivially matches the peak impact.

[3] This is precisely the regime obtained with the simple static geometrical arguments above.

8.4.5 Price manipulation

One can show that the price impact model presented here is free of price manipulation (or equivalently, verifies the *no free lunch theorem* discussed in Chapter 7). The average cost C of a closed trajectory writes:

$$C = \int_0^T m_t I_t \mathrm{d}t\,, \quad \text{with} \quad \int_0^T m_t \mathrm{d}t = 0\,.$$

Using Eq. (8.12) one obtains that C can be identically rewritten as a quadratic form:

$$C = \frac{1}{2} \iint_0^T m_t H(t,t') m'_t \mathrm{d}t \mathrm{d}t'\,, \tag{8.15}$$

where H is a non-negative operator (see [97]), thereby showing that $C \geq 0$ for any execution schedule m_t.

8.5 Timescale heterogeneity

While the reaction-diffusion model we just presented is quite insightful in many aspects, it suffers from at least two major difficulties when confronted with data.

First, a strict square-root law (with no dependence on the participation η) is only recovered in the limit where the execution rate m_0 of the meta-order is larger than the normal execution rate J of the market itself – whereas most meta-order impact data is in the opposite limit $m_0 < 0.1J$. This issue can be resolved by noting that the total market turnover is actually dominated by HFTs/market makers, while resistance to slow meta-orders can only be provided by slow participants on the other side of the book. This is implemented in the model by introducing fast and slow traders, see [98].

Second, the theoretical inverse square-root impact decay is too fast to solve the diffusivity puzzle. Indeed, in the slow execution limit and for infinite memory, one recovers the propagator model:

$$p_t = p_0 + \int_0^t \mathrm{d}t' G(t-t') m_{t'}\,, \quad \text{with} \quad G(t) = \frac{1}{\mathscr{L}} \frac{1}{\sqrt{4\pi D t}}\,. \tag{8.16}$$

The kernel $G(t)$ decays as $t^{-\beta}$ with $\beta = 1/2 \neq (1-\gamma)/2$ (see Chapter 4). As a result, the model generates mean-reverting price dynamics, inconsistent with real data. Introducing timescale heterogeneities for the renewal of liquidity – in particular *fractional diffusion* instead of normal diffusion for latent order – allows one to cure such deficiencies, see [99]. We assume waiting times for the diffusion of latent orders to be distributed according to a power-law function with tail exponent α of the form $\Psi(t) \sim 1/t^{1+\alpha}$. For $\alpha > 1$ one recovers normal diffusion, but for $\alpha < 1$, the mean waiting time diverges and Eq. (8.2) becomes:

$$\partial_t \phi \quad = \quad K \mathfrak{D}_t^{1-\alpha} (\partial_{xx}\phi - \tilde{\nu}\phi) + s(x,t)\,, \tag{8.17}$$

where K is a generalised diffusion coefficient, $\tilde{\nu}$ is a reduced cancellation rate, and where $\mathfrak{D}_t^{1-\alpha} = \partial_t \mathfrak{D}_t^{-\alpha}$ with $\mathfrak{D}_t^{-\alpha}$ the fractional Riemann-Liouville operator [100, 101].[4] Similar to Eq. (8.2), Eq. (8.17) can be solved in Fourier space in the infinite memory limit to obtain the corresponding stationary order book and market impact of metaorders, see [99]. In particular, Eq. (8.16) becomes:

$$p_t = p_0 + \frac{1}{\mathscr{L}_\alpha} \int_0^t \mathrm{d}t' \frac{m_{t'}}{\sqrt{4\pi K (t-t')^\alpha}}\,, \tag{8.18}$$

with $\mathscr{L}_\alpha$ a liquidity parameter, analogous to $\mathscr{L}$ in the normal diffusion case. Equation (8.18) allows to identify the propagator decay exponent $\beta = \min(1/2, \alpha/2)$. Thus, for $\alpha < 1$ the equality $\beta = (1-\gamma)/2$ can be achieved by the choice $\alpha = 1-\gamma$; and recalling $\gamma \approx 0.5$ implies $\alpha \approx 0.5$. In other words, a

[4] The fractional Riemann-Liouville operator is defined as $\mathfrak{D}_t^{-\alpha} f(t) = \Gamma[\alpha]^{-1} \int_0^t \mathrm{d}u\,(t-u)^{\alpha-1} f(u)$.

fractional latent order book model enables the price to be diffusive in the presence of a persistent order flow thereby solving the diffusivity puzzle.

Market participants are indeed highly heterogeneous, and display a broad spectrum of volumes and timescales, from low frequency institutional investors to High Frequency Traders (HFT). Timescale heterogeneity is often a crucial ingredient in complex systems.

8.6 Beyond mean field

We now take the deposition rate equal to $\lambda + \xi(x,t)$ with $\langle \xi \rangle = 0$, $\langle \xi(x,t)\xi(x',t')\rangle = 2\varsigma^2\delta(x - x')\delta(t - t')$, and focus on the limit $\lambda, \nu \to 0$ in the absence of a metaorder. Using Eq. (8.10) and defining $\delta J(t) := -D\partial_x \phi\big|_{x=0} - D\mathscr{L}$ the stochastic contribution to the order flow, one can compute:

$$
\begin{aligned}
\langle \delta J(t)\delta J(s)\rangle &= D^2 \int \mathrm{d}x \int \mathrm{d}y \int \mathrm{d}t' \int \mathrm{d}s' \partial_x \mathscr{G}(-x, t-t')\partial_x \mathscr{G}(-y, s-s') 2\varsigma^2 \delta(x-y)\delta(t'-s') \\
&= \frac{\varsigma^2}{2\pi D} \int \mathrm{d}x \int_0^{t\wedge s} \mathrm{d}t' \frac{4x^2}{[(t-t')(s-t')]^{3/2}} \exp\left[-\frac{x^2}{4D}\frac{t+s-2t'}{(t-t')(s-t')}\right] \\
&= \varsigma^2 \sqrt{\frac{D}{\pi}} \int_{|t-s|}^{t+s} \frac{\mathrm{d}u}{u^{3/2}} \\
&\underset{t,s\to\infty}{\longrightarrow} \frac{1}{\sqrt{|t-s|}},
\end{aligned}
\tag{8.19}
$$

which is consistent with the empirical autocorrelation of the order flow, see Chapter 3.

Conclusions

A microfounded theory based on a linear (latent) order book, inspired by diffusion-reaction models from physics and chemistry, is able to account for the square-root impact law. This is without relying on any equilibrium or fair-pricing conditions, but rather on purely statistical considerations.

9

Financial engineering

In a sense, financial engineering could be defined as the art of *chiselling* PnL (Profit & Loss) distributions. Indeed, the work of financial engineers is to devise products that suit the investors, more/less risk, more/less skew etc. In this Chapter we present different ways used to optimise investing, including optimal portfolio composition, dynamical optimisation, and option hedging.

9.1 Optimal portfolios

So far, in most of this course we have addressed asset prices as if they were uncorrelated. In practice, most market participants trade large portfolios that combine hundreds or thousands of correlated assets. Here we address the question of *optimal portfolios,* that is how the trade-off between risk and return can be dealt with optimally.

Restricting to the case of Gaussian statistics, consider N assets with arbitrary correlations described by their *correlation matrix*:

$$C_{ij} = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle \,, \tag{9.1}$$

with r_i, r_j the returns of assets $i, j \in [1, N]$. Since C is a symmetric matrix, it can be diagonalised (*principal component analysis* or PCA). In particular, returns can be written as a weighted sum of uncorrelated Gaussian variables $\{e_a\}_{a\in[1,N]}$ (often called *explicative factors* or *principal components*) with zero mean and variance given by the eigenvalue of C and denoted σ_a^2. One has:

$$r_i = \langle r_i \rangle + \sum_{a=1}^{N} v_{ia} e_a \,, \quad \text{with} \quad \langle e_a e_b \rangle = \sigma_a^2 \delta_{ab} \,. \tag{9.2}$$

This decomposition often has a simple economic interpretation. For stocks, the eigenvector associated to the highest eigenvalue, coined the *market mode*, is $\sim \frac{1}{\sqrt{N}}\{1, \dots, 1\}$. The next modes correspond to one or several economic sectors against others. The returns of a global portfolio with weights $\{w_i\}_{i\in[1,N]}$ are also Gaussian with mean $\langle r_p \rangle = \sum_i w_i \langle r_i \rangle$ and variance:

$$\sigma_p^2 = \sum_{i,j=1}^{N} w_i C_{ij} w_j \,. \tag{9.3}$$

The optimal portfolio $\{w_i^\star\}_{i\in[1,N]}$ in the sense of minimal risk given a target gain $\mathscr{G}$ is obtained by minimising $\sigma_p^2 - \lambda \langle r_p \rangle$ with respect to w_i. This gives $2\sum_j C_{ij} w_j^\star = \lambda \langle r_i \rangle$, which, provided C can be inverted yields Markowitz' famous result [102]:

$$w_i^\star = \frac{\lambda}{2} \sum_{j=1}^{N} C_{ij}^{-1} \langle r_j \rangle \,. \tag{9.4}$$

The Lagrange multiplier λ is determined by the equation $\sum_i w_i^\star \langle r_i \rangle = \mathcal{G}$. One finally obtains:

$$w_i^\star = \mathcal{G} \frac{\sum_{j=1}^N C_{ij}^{-1} \langle r_j \rangle}{\sum_{i,j=1}^N C_{ij}^{-1} \langle r_i \rangle \langle r_j \rangle}. \tag{9.5}$$

Making use of Eq. (9.4), the average return and variance of the optimal portfolio write:

$$\langle r_p \rangle = \sum_i w_i^\star \langle r_i \rangle = \frac{\lambda}{2} \sum_{i,j} C_{ij}^{-1} \langle r_i \rangle \langle r_j \rangle, \qquad \sigma_p^2 = \sum_{i,j} w_i^\star C_{ij} w_j^\star = \frac{\lambda^2}{4} \sum_{k,\ell} C_{k\ell}^{-1} \langle r_k \rangle \langle r_\ell \rangle.$$

Eliminating λ yields that the set of optimal portfolios is described by a parabola, called the *efficient frontier*, in the risk-return plane $\sigma_p^2, \langle r_p \rangle$, see Fig. 9.1. This line separates 'possible' portfolios (below) from 'impossible' ones (above). The Sharpe ratio $\mathcal{S} := \langle r_p \rangle / \sigma_p$ is constant and maximal along this line, such that all optimal portfolios have the same Sharpe. The Lagrange multiplier λ sets the risk (or equivalently the average return) along this line and can be interpreted as the typical drawdown (see Chap. 2) $\Delta^\star := \sigma_p^2 / \langle r_p \rangle = \lambda/2$.

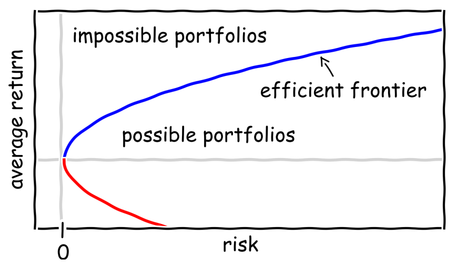


Figure 9.1: Efficient frontier in the $\sigma_p, \langle r_p \rangle$ plane.

A few important remarks are in order.

- As we have seen above, the correlation matrix C is a crucial input for managing portfolio risk. However the empirical determination of C is very noisy due to the length of the available time series not being very large compared to the number of assets in the portfolio, see e.g. [103, 104]. In Fig. 9.2 we plot a typical density of eigenvalues of C together with the corresponding *Marchenko-Pastur* distribution (pure random matrix) [105]. Anything below this curve must be considered as noise. Typically, only 5 to 10% of the eigenvalues are outside the noise band; but they account for 20 to 30% of the total volatility.[1] Here, given that the above formulas involve the inverse of the correlation matrix, one is concerned with the smallest eigenvalues of C, the high noise level of which leads to large numerical errors in C^{-1}. In practice this leads to strong underestimation of risk of optimal Markowitz-like portfolios by over-investing in artificially low risk modes. Several methods to clean correlation matrices and improve risk estimation have been proposed [106].

- Often there can be other constraints on portfolio composition. In particular, non linear constraints make the optimisation procedure highly nontrivial. Interesting similarities with the physics of spin glasses can be devised. Examples of such constraints are *long only* portfolios in which $w_i^\star \geq 0$ for all i (no short positions), or *margin calls* on futures markets where a certain deposit is required regardless of the position (long or short) that is $\sum_i |w_i| = f$ with f the fraction of wealth invested as a deposit, often called a *leverage constraint*.

- The hypothesis of stationarity underlies the use of the correlation matrix for trading optimal portfolios. But we do not live in a stationary world, described by a time-invariant covariance

[1]These values correspond to an estimation of the correlation matrix of daily returns of a few hundreds of stocks over a few years.

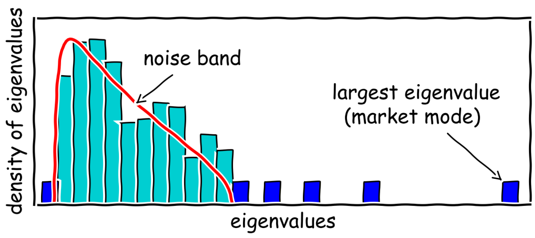


Figure 9.2: Typical density of eigenvalues of the correlation matrix.

matrix. This induces the out-of-sample risk to be larger than expected. The covariance between assets evolves not only because the volatility of each asset changes over time [107] and react to the recent market trend [108–110], but also because correlations themselves increase or decrease, depending on market conditions [111–113]. Sometimes these correlations jump quite suddenly, due to an unpredictable geopolitical event. The arch example of such a scenario is the Asian crisis in the fall of 1997, when the correlation between bonds and stocks indexes abruptly changed sign and became negative – a *flight to quality* mode that has prevailed ever since [114, 115].[2]

- Markowitz optimal portfolios are all proportional to one another (they only differ by the choice of λ) and since the problem is linear, a superposition of optimal portfolios is still optimal. If all market participants trade Markowitz optimal portfolios, the *market portfolio* made of the aggregate positions of all the agents is also optimal, and thus satisfies Eq. (9.4). Recall the equation of the optimal portfolio line (or efficient frontier) $\sigma_p^2/\langle r_p\rangle = \frac{\lambda}{2}$ and eliminate λ with the inverted Eq. (9.4) applied to a portfolio made of two 'assets': the optimal market portfolio + an infinitesimal fraction of asset i ($w_i = 1 - w_{\text{mkt}} = \epsilon \to 0$), that is $\langle r_i\rangle = \frac{2}{\lambda} C_{i,\text{mkt}}$. One obtains an expression of the average return of asset i as function of its covariance with the market portfolio $C_{i,\text{mkt}}$:

$$\langle r_i\rangle = \beta_i \langle r_{\text{mkt}}\rangle\,, \quad \text{with} \quad \beta_i = \frac{C_{i,\text{mkt}}}{\sigma_p^2}\,, \tag{9.6}$$

where β_i is called the *beta* of asset i. This is the famous *Capital Asset Pricing Model* (CAPM). Of course this does not work very well because not all agents have the same definition of an optimal portfolio, nor do they have the same estimates for average returns and risks.

- If price statistics aren't Gaussian, the variance may not be an adequate measure of risk, in the sense that minimising variance is not equivalent to an optimal control of large fluctuations. With power-law tailed returns, the *Value-at-Risk* (VaR) is more suited and the corresponding optimisation problem can be tracked analytically, see e.g. [2].[3]

- Finally, slippage costs can be very large when trading large portfolios. Accounting cross-impact effects (see Chapter 4) is essential, as not doing so leads to an incorrect estimation of liquidity which results in suboptimal execution strategies, see e.g. [32].

[2]A *Flight to quality* or *flight-to-safety* is the action of investors moving their capital from riskier assets to safer ones, such as treasuries and other bonds.

[3]The *Value-at-Risk* or VaR corresponds to the level of loss associated to a certain probability of loss, say $p = 1\%$, over a certain time interval τ. Mathematically this translates into:

$$\int_{-\infty}^{-\text{VaR}} \mathbb{P}_\tau(x)\mathrm{d}x = p\,,$$

with $P_\tau(x)$ the probability distribution of returns on timescale τ. In other words: over a time τ, the probability that I loose more than VaR is p.

9.2 Optimal trading

Here we show that by dynamically optimising one's positions one can modify the PnL distribution and minimise risk. Consider an investor holding a certain number of shares of a simple asset (say a stock) during a certain time $T = N\tau$, in order to realise a certain target gain $\mathcal{G}$. We denote by ϕ_n the number of shares, and g_n the realised gains at time $t_n = n\tau$, such that one can write:

$$g_N = \sum_{n=0}^{N-1} \phi_n r_n \,, \tag{9.7}$$

with $r_n = p_{n+1} - p_n$ are assumed to be iid with mean $\langle r_n \rangle = m\tau$ and variance $\langle r_n^2 \rangle - \langle r_n \rangle^2 = \sigma^2 \tau$. The optimal strategy in this setting is the set of successive position $\phi_n^\star(g_n)$ which ensure $\langle g_N \rangle = \mathcal{G}$ while minimising variance of final wealth:

$$\mathcal{R}^2 = \left\langle (g_N - \mathcal{G})^2 \right\rangle . \tag{9.8}$$

To determine $\phi_n^\star(g_n)$ one can work backwards in time; this is *Bellmann's method*. Optimising the last position ϕ_{N-1} is done by noting that $\mathcal{R}^2 = \left\langle (g_{N-1} + \phi_{N-1} r_{N-1} - \mathcal{G})^2 \right\rangle$, and that r_{N-1} is independent of the value of g_{N-1}, such that:

$$\mathcal{R}^2 = \phi_{N-1}^2 \langle r_{N-1}^2 \rangle + 2\phi_{N-1} \langle r_{N-1} \rangle (g_{N-1} - \mathcal{G}) + (g_{N-1} - \mathcal{G})^2 . \tag{9.9}$$

$\partial_{\phi_{N-1}} \mathcal{R}^2 = 0$ yields:

$$\phi_{N-1}^\star = \frac{m}{m^2\tau + \sigma^2} (\mathcal{G} - g_{N-1}) \,, \tag{9.10}$$

and proceeding similarly in a recursive way for all $n \in [0, N-1]$ leads to:

$$\phi_n^\star = \frac{m}{m^2\tau + \sigma^2} (\mathcal{G} - g_n) \,, \tag{9.11}$$

where we have used the already known strategy for $\ell > n$.[4] The optimal strategy thus consists in taking positions proportional to the distance to the target: invest more when one is far from the target, and reduce the investment when the gains approach the target. One can show that the resulting risk in the limit $T \gg \tau$ is exponentially small in T.[5] This result is to be compared to that of the naive constant execution rate strategy $\phi_n = \phi_0 = \mathcal{G}/(mT)$ for all n, for which the variance $\mathcal{R}^2 \sim 1/T$ (according to the CLT) which is much larger than that of the optimal strategy at large T, see Fig. 9.3.

Two remarks are in order. The first one is that the optimal strategy allows for huge losses at intermediate times (which will on average be compensated by the positive trend), but this is not acceptable for several obvious reasons. The second is that all this story relies – in addition to the returns not being Gaussian, iid etc. – on the assumption that the drift m is perfectly known, which is also not the case in practice.[6] Notwithstanding, this simple model shows that trading strategies can be monitored to modify the gains distribution, a useful concept for the following sections.

[4] Assuming that $\{\phi_\ell^\star\}_{\ell > n}$ have been determined, one can write:

$$\mathcal{R}^2 = \left\langle (g_n + \phi_n r_n + \textstyle\sum_{\ell=n+1}^{N-1} \phi_\ell^\star r_\ell - \mathcal{G})^2 \right\rangle = \phi_n^2 \langle r_n^2 \rangle + 2\phi_n \left(\langle r_n \rangle (g_n - \mathcal{G}) + \textstyle\sum_{\ell=n+1}^{N-1} \phi_\ell^\star \langle r_n r_\ell \rangle \right) + O(\phi_n^0) .$$

Recalling that $\langle r_n r_\ell \rangle \propto \delta_{n,\ell}$ and taking $\partial_{\phi_n} \mathcal{R}^2 = 0$ one obtains Eq. (9.11).

[5] Taking the continuous time limit $\tau \to 0$ and assuming the price $X(t)$ to be a continuous time Brownian motion one obtains $\mathrm{d}g = \phi^\star \mathrm{d}X$ and $dX = m\mathrm{d}t + \sigma \mathrm{d}W$ with $W(t)$ a Brownian motion of zero drift and unit volatility, to be interpreted in the Ito sense since the price increment is posterior to the determination of the optimal strategy $\phi^\star$. Noting that $\mathrm{d}g = \mathrm{d}(\mathcal{G} - g)$ one obtains:

$$\mathrm{d}(\mathcal{G} - g) = (\mathcal{G} - g) \left(-\tfrac{m^2}{\sigma^2} \mathrm{d}t - \tfrac{m}{\sigma} \mathrm{d}W \right) .$$

Integrating, one obtains a geometric Brownian walk for $\mathcal{G} - g$ of the form:

$$\mathcal{G} - g = \mathcal{G} \exp\left(-\tfrac{3}{2} \tfrac{m^2}{\sigma^2} t + \tfrac{m}{\sigma} W(t) \right) ,$$

where we have used $g(t=0) = 0$. Using that $W(t) \sim \sqrt{t}$, this equation shows that provided $T \gg \sigma^2/m^2 := t^\star$ (see Chap. 2) the gain converges almost surely to the target. In particular the variance reads $\mathcal{R}^2 = \left\langle (g - \mathcal{G})^2 \right\rangle = \mathcal{G}^2 \exp(-T/t^\star)$.

[6] Drift uncertainty could be taken into account by replacing σ^2 by $\sigma^2 + (\Delta m)^2 T$ with Δm the standard deviation of m, which reveals that for large T drift uncertainty becomes the major source of risk.

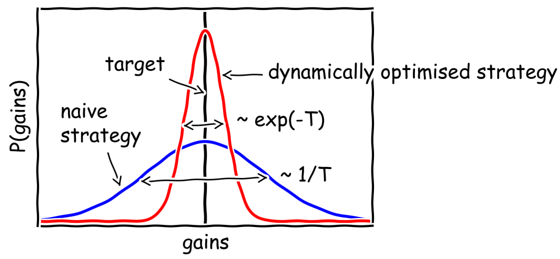


Figure 9.3: Dynamical optimisation.

9.3 Options

Options are another way to chisel one's PnL distribution.

9.3.1 Bachelier's fair price

Options are insurance policies which cover losses beyond a certain threshold. They are of two sorts:

- A *put* option protects the owner against the potential drawdown of the price of a given asset (called the *underlying*). More precisely, the put covers losses larger than $\mathscr{L} = \mathscr{W}_0 \frac{p_0 - p_<}{p_0}$ where $p_<$ is called the strike.
- Symmetrically, a *call* option protects the owner against the increase of the price of the underlying asset that he will need to buy in the future, by warranting a maximum buy price $p_>$, also called the strike (e.g. you need to buy wheat in a year but you are worried that the price will go up due to global warming; if you are ready to pay up to \$10, but not more, you can buy a one-year maturity call option with a strike at $p_> = \$10$).

We here restrict to the analysis of *European options* (or *plain vanilla*) which have a well defined *maturity* (or *expiry date*) T at which the option can be exercised.

Let us take the example of a put option. The owner of the option has clearly shaped his PnL distribution to cover losses larger than $\mathscr{L}$ but one must not forget that he also had to buy the option. Denoting $\mathscr{C}_<$ the cost of the option, his actual losses covered are those beyond $\mathscr{L} + \mathscr{C}_<$ (see the left panel of Fig. 9.4). The natural question is thus, what should be the price of the option contract? This problem, at the very origin of the *derivatives pricing* science, was first solved by Bachelier in 1900 with a fair game argument: the cost of the option contract should be such that on average no party is favoured. Noting that a put option pays $X := (p_< - p_T)\mathbb{1}_{p_T < p_<}$ per underlying asset, this is $\mathscr{C}_< = \langle X | (p_T, p_0) \rangle$, and assuming that prices follow additive continuous time random walks:

$$\mathscr{C}_< = \int_0^{p_<} (p_< - p)\mathbb{P}(p, t = T | p_0, t = 0)\, \mathrm{d}p\,. \tag{9.12}$$

Similarly, noting that a call option pays $Y := (p_T - p_>)\mathbb{1}_{p_T > p_>}$, one has $\mathscr{C}_> = \langle Y | (p_T, p_0) \rangle$, or:

$$\mathscr{C}_> = \int_{p_>}^{\infty} (p - p_>)\mathbb{P}(p, t = T | p_0, t = 0)\, \mathrm{d}p\,. \tag{9.13}$$

Further, in the Gaussian case $\mathbb{P}(p, t = T | p_0, t = 0) = \frac{1}{\sqrt{2\pi\sigma^2 T}} \exp\left(-\frac{(p-p_0)^2}{2\sigma^2 T}\right)$.[7]

There exists a variety of different option contracts. The common *American options* are similar to the European kind with the difference that they can be exercised at any time before the maturity. Other kinds can become very complex, and therefore hard to price, in particular for the buyer who usually gets scammed. *Barrier put options* cover losses only between $\mathscr{L} = \mathscr{W}_0 \frac{p_0 - p_<}{p_0}$ and $\mathscr{L}_e = \mathscr{W}_0 \frac{p_0 - p_\ll}{p_0}$ with

[7] Note that the price of an *at-the-money* (ATM) option ($p_0 = p_>$) is simply given by $\mathscr{C}_>^{\mathrm{ATM}} = \sqrt{\frac{\sigma^2 T}{2\pi}}$.

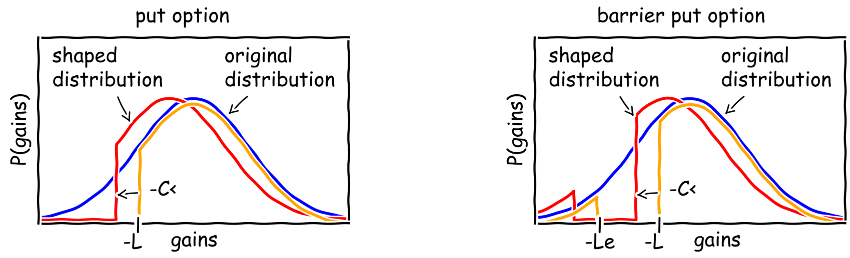


Figure 9.4: Shaped PnL distributions (put options).

$p_{\ll} < p_{<}$, such that if the price swings below $p_{\ll}$ extreme losses are left unhedged (see the right panel of Fig. 9.4). These contracts are particularly toxic, they are tantamount to a property insurance policy covering flooding but only if the water level does not exceed 10cm below the ceiling. The thing is that, even if the insurer tells you that is very unlikely, chances are that if the water reaches such a level it will likely go all the way to the top. In the same vein, all products for which there is a large probability of making a small profit and a very small probability of a huge loss, (say 99.99% chances to make \$1 and 0.01% chances to loose \$10,000) are highly vicious, as estimating such small probabilities often comes from unreliable models that most often underestimate the occurrence of rare events (for one thing, because they did not occur in the past). Complex, unrealistic and hardly calibratable models are often used to bury risk.

9.3.2 Black and Scholes' extravaganza

Let us now take the perspective of the insurer who writes the option contract. Can he modify his own PnL distribution expected from selling an option? The idea of hedging relies on the following remark. Say the insurer sells a call option; if the option is exercised, it means that the price of the underlying went up by quite a bit ($p_T > p_> > p_0$) and so that he could have benefitted from buying a certain amount of shares at $t = 0$ to be sold at maturity making a profit. Conversely, for a put option, the hedging would consist in shorting and buying back the underlying. Black and Scholes addressed this problem in 1972 [116], just before an exchange opened in Chicago in 1973, where standardised option contracts could be bought and sold anonymously, much like stocks, removing the distinction between insurers and customers. In all the following we consider a call option but of course the same could be done with a put.

Denote by ϕ_0 the number of underlying shares bought by the insurer at $t = 0$.[8] His PnL writes:

$$g = \mathscr{C}_> - Y + \phi_0(p_T - p_0)\,. \tag{9.14}$$

Finding the optimal hedging strategy amounts to finding $\phi_0^\star$ which minimises variance and remains consistent with Bachelier's argument $\langle g\rangle = 0$. In the zero drift case $m = 0$ – to which we shall stick in the following – Bachelier's argument yields again Eq. (9.13), regardless of ϕ_0.[9] The variance writes:

$$\mathscr{R}^2 = \left\langle g^2\right\rangle = \mathscr{R}_0^2 + \phi_0^2\left\langle (p_T - p_0)^2\right\rangle - 2\phi_0\left\langle Y(p_T - p_0)\right\rangle\,, \tag{9.15}$$

where $\mathscr{R}_0^2 = \left\langle (\mathscr{C}_> - Y)^2\right\rangle$ is the *unhedged* risk of the option. $\partial_{\phi_0}\mathscr{R}^2 = 0$ yields:

$$\phi_0^\star = \frac{\langle Y(p_T - p_0)\rangle}{\sigma^2 T}\,. \tag{9.16}$$

This shows that there is an optimal number of underlying shares that one should hold to minimise the risk. On the one hand holding shares reduces the risk because part of the potential loss at maturity due

[8]For the sake of simplicity, we restrict to the simplest case when the hedging strategy is constrained to be static, but one can naturally do better with dynamical optimisation $\{\phi_n\}_{n\in[0,N-1]}$ (see Sect. 9.2).

[9]In full generality $\mathscr{C}_> = \langle Y|(p_T, p_0)\rangle - \phi_0 m T$.

to the option being exercised is covered, but on the other hand holding too many shares is bad because one gets exposed to the fluctuations of the underlying's price. In the Gaussian case a simplification appears:

$$\phi_0^\star = \frac{1}{\sigma^2 T}\int_{p_>}^{\infty} \frac{(p-p_>)(p-p_0)}{\sqrt{2\pi\sigma^2 T}} \exp\left(-\frac{(p-p_0)^2}{2\sigma^2 T}\right) \mathrm{d}p = \partial_{p_0}\mathscr{C}_>\,, \tag{9.17}$$

often called *Black-Scholes Delta hedge*. The derivatives of $\mathscr{C}_>$ are called the *Greeks* because they are denoted by Greek letters; in particular $\Delta := \partial_{p_0}\mathscr{C}_>$, hence the name of the hedging strategy.[10] Injecting $\phi_0^\star$ in Eq. (9.15) and using Eq. (9.16) yields $\mathscr{R}^2 = \mathscr{R}_0^2 - \sigma^2 T \Delta^2$.

Black and Scholes considered the continuous time limit where the price follows a continuous time random walk.[11] From here on we no longer make the assumption of zero drift and allow for $m \neq 0$. Still from the perspective of the insurer, consider a portfolio short an option and long ϕ_t stocks. Its value Π follows $\mathrm{d}\Pi = -\mathrm{d}\mathscr{C}_> + \phi_t \mathrm{d}p$.[12] Further, using Ito's formula yields $\mathrm{d}\mathscr{C}_> = \partial_t\mathscr{C}_>\mathrm{d}t + \partial_p\mathscr{C}_>\mathrm{d}p + \frac{\sigma^2}{2}\partial_{pp}\mathscr{C}_>\mathrm{d}t$ which leads to:

$$\mathrm{d}\Pi = -\left(\partial_t\mathscr{C}_> + \tfrac{\sigma^2}{2}\partial_{pp}\mathscr{C}_>\right)\mathrm{d}t + \left(\phi_t - \partial_p\mathscr{C}_>\right)\mathrm{d}p\,. \tag{9.18}$$

Black-Scholes Delta hedge follows, quite miraculously, by noting that the only source of uncertainty in the value of the portfolio being $\mathrm{d}p$, the evolution of Π becomes purely deterministic (zero risk portfolio) if one chooses $\phi_t^\star = \partial_p\mathscr{C}_>$. Black-Scholes solution is thus the perfect hedging strategy: the PnL distribution is, regardless of the risk criterion, a δ distribution at $g = 0$ (impossible to do better, see Fig. 9.5). As we shall see below, this strange (and dangerous) property does not survive in real life conditions.

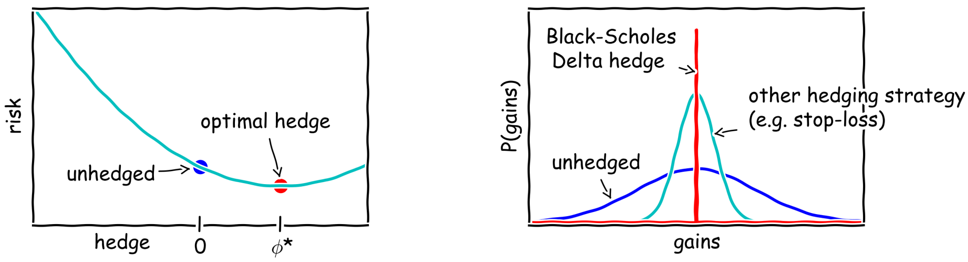


Figure 9.5: (left) $\mathscr{R}^2$ as function of ϕ. (right) Shaped PnL distributions of the insurer.

For the model to be arbitrage free requires $\mathrm{d}\Pi = 0$ which leads to a backwards diffusion equation,[13] sometimes coined the *Black-Scholes PDE*:

$$\partial_t\mathscr{C}_> + \tfrac{\sigma^2}{2}\partial_{pp}\mathscr{C}_> = 0\,. \tag{9.19}$$

Note that the average return m does not appear in Eq. (9.19) and will thus not condition the solution either, or in other words, the cost of insuring a security in a bear market is, oddly enough, the same

[10] Note that the option's cost depends on the maturity T, the strike $p_>$, p_0 and p_T, and the way it varies with these parameters is the main focus of the options trading community. The Δ of the option is positive given that the higher the p_0, the more likely it is to reach the strike at maturity, and the more expensive the option. The other greeks are the *Gamma* $\Gamma := \partial_{p_0}\Delta$, the *Theta* $\Theta := -\partial_T\mathscr{C}_> < 0$, and the *Vega* $V := \partial_\sigma\mathscr{C}_> > 0$. For $m = 0$, σ and T only appear through the combination $\sigma^2 T$, and one has $V = -2T/(\sigma\Theta)$.

[11] While the results derived here can also be painfully derived within the previous framework, the formalism of stochastic differential calculus wedded to the unrealistic continuous limit (on which most of mathematical finance relies) appears to be extremely convenient.

[12] The sign of the first term on the RHS is negative because the insurer has sold the option, and thus he will loose gains if the option's price increases.

[13] A backwards diffusion equation is a diffusion equation with the wrong sign. The 'proper' diffusion equation is recovered by letting $t \to -t$, which allows the same physical interpretation only *backwards* in time.

as in a bull market! Because time flows backwards, one needs a boundary condition in the future (instead of an initial condition). Here it is given by the price of the option *at maturity*, obvious in all circumstances, $\mathscr{C}_>(t=T) = Y = (p_T - p_>)\mathbb{1}_{p_T > p_>}$ – to be propagated from $t = T$ to $t = 0$. It is then easy to solve Eq. (9.19) and show that it leads again to Bachelier's fair price, Eq. (9.13).

9.3.3 Residual risk beyond Black-Scholes

As mentioned above the Black-Scholes model is flawed and not suited to describe the real world. In spite of what one may think, the major flaw does not come from the world not being Gaussian, or at least not only. Several corrections/extensions exist for non-Gaussian (but still iid) returns.

- In such a case, the optimal strategy is not simply given by Black-Scholes Delta hedge. Corrections can be computed as an expansion involving the cumulants of the returns and the higher order derivatives $\partial_{p^n}\mathscr{C}_>$. However, the resulting $\phi^\star$ is still an increasing function of p which varies from 0 to 1. Most importantly the difference with Δ is actually numerically quite small, and so is the resulting increase of residual risk.
- Other risk objectives might be more suited to non Gaussian statistics. Optimal strategies in terms of VaR display similar features with a slower variation of $\phi^\star$ with p, leading to less portfolio re-balancing and thus less trading costs.
- The independence of the Δ hedge with m is at first sight counterintuitive, to say the least. With m very large and positive, one would expect an increased average pay-off of the option, while for m large and negative, the option should be worthless. However, such a reasoning does not take into account the impact of the hedging strategy on the global wealth balance, which is itself proportional to m. In other words, the average PnL associated to trading the underlying partly compensates for the modified option's pay-off due to m in the fair game argument; the compensation is perfect in the Gaussian case. In the general case, the corrections associated to $m \neq 0$ remain in practice numerically quite small, at least for maturities up to a few months.

Black-Scholes zero risk miracle actually comes from the continuous time hypothesis, combined with Gaussian statistics. Indeed, writing $\mathscr{R}^2$ in discrete time, one finds that the Δ hedge only perfectly compensates for $\mathscr{R}_0^2$ in the limit $\tau \to 0$, see e.g. [2]. In real life the limit $\tau \to 0$ is unachievable, first because re-hedging in continuous time is physically impossible, and because trading costs increase with re-hedging frequency![14] For τ finite, the residual risk is *not* negligible (even with Gaussian increments). At the leading order in τ the residual risk is given by:

$$\mathscr{R}^{\star 2} = \frac{\sigma^2\tau}{2}\rho(1-\rho) + O\left(\tau^{3/2}\right), \tag{9.20}$$

with ρ the probability at $t = 0$ that the option is exercised at expiry $t = T$. To evaluate the quality of a hedging strategy one commonly defines a *quality ratio* $\mathscr{Q} := \mathscr{C}_>/\mathscr{R}^\star$. Here, in the case of an *at-the-money* option (for which $\rho = 1/2$) with maturity $T = N\tau$ and assuming Gaussian fluctuations, one has $\mathscr{Q} \approx \sqrt{4N/\pi}$, which increases only slowly with N. An option of maturity $T = 1$ month, re-hedged daily ($N \approx 25$) yields $\mathscr{Q} \approx 5$, which means that the residual risk is one-fifth of the price of the option itself. Even if one re-hedges every 30 min the quality ratio is only $\mathscr{Q} \approx 20$.

9.3.4 The volatility smile

In most cases, options are not appropriately priced by Black-Scholes' formula. This is in particular the case for short maturities, or equivalently small N, for which fat tail effects are most important (large deviations from the CLT). Interestingly enough, markets self-correct for such non Gaussian effects empirically, by effectively substituting in Black-Scholes formula the true historical volatility of the underlying by an *implied volatility* Σ which depends on both the strike $p_>$ and the maturity T. The

[14] Re-hedging frequency results for a trade-off between the performance of the hedging strategy and the associated trading costs. Let us stress that within the Black-Scholes framework there are no trading costs in buying or selling the stock or the option, there is no market impact and short selling is assumed to be possible without financial penalty.

manifold $\Sigma(p_>, T)$, commonly coined the *volatility surface*, appears to increase with the difference between p_0 and $p_>$, which is called the *smile effect*, and flattens with increasing maturity (as the Gaussian approximation improves), see Fig. 9.6.

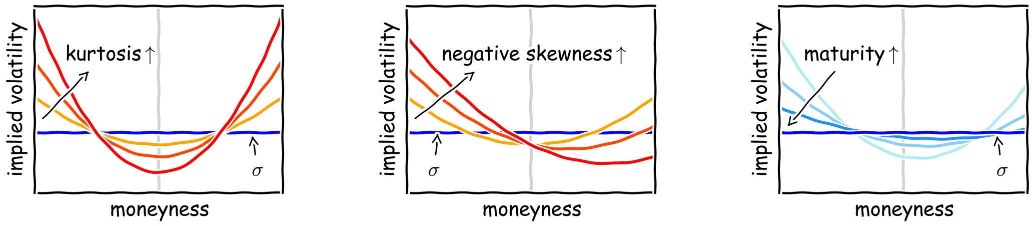


Figure 9.6: Volatility smile for increasing kurtosis, absolute skewness and maturity.

To understand this and in particular the shape of the volatility surface, a cumulant expansion of Eq. (9.13) can be performed to account for non-zero skewness and kurtosis. Denoting by $\zeta_T = \zeta/\sqrt{N}$ and $\kappa_T = \kappa/N$ respectively the skewness and kurtosis of the price increments on the scale of the maturity of the option, one has:

$$\mathscr{C}_> = \mathscr{C}_>^{\mathrm{G}} + \sigma\sqrt{\frac{T}{2\pi}}e^{-\mathscr{M}^2/2}\left[\frac{\zeta_T}{6}\mathscr{M} + \frac{\kappa_T}{24}\left(\mathscr{M}^2 - 1\right) + O\left(\mathscr{M}^3\right)\right], \tag{9.21}$$

with $\mathscr{C}_>^{\mathrm{G}}$ the Gaussian pricing formula, as given by Eq. (9.13), and $\mathscr{M} := (p_> - p_0)/(\sigma\sqrt{T})$ the *rescaled moneyness*. On the other hand, the variation of the Gaussian pricing corresponding to a variation in volatility $\delta\sigma^2 = 2\sigma\delta\sigma$ can be easily computed and writes:

$$\mathscr{C}_>^{\mathrm{G}}(\sigma + \delta\sigma) = \mathscr{C}_>^{\mathrm{G}}(\sigma) + \delta\mathscr{C}_>^{\mathrm{G}}, \quad \text{with} \quad \delta\mathscr{C}_>^{\mathrm{G}} = \delta\sigma\sqrt{\frac{T}{2\pi}}e^{-\mathscr{M}^2/2}, \tag{9.22}$$

Identifying Eqs (9.21) and (9.22) to set $\delta\sigma$, shows that the effect of non-zero skewness and kurtosis can thus be reproduced (to first order) by a Gaussian pricing formula with an effective volatility given by $\Sigma = \sigma + \delta\sigma$:

$$\Sigma(p_>, T) = \sigma\left[1 + \frac{\zeta_T}{6}\mathscr{M} + \frac{\kappa_T}{24}\left(\mathscr{M}^2 - 1\right) + O\left(\mathscr{M}^3\right)\right], \tag{9.23}$$

consistent with the shape and amplitude of the smile, see Fig. 9.6. Kurtosis reduces the effective at-the-money volatility, and increase it out-of-the-money.[15] For $\kappa_T = 1$ (typical of one month options), the kurtosis correction to strongly out-of-the money options, say $\mathscr{M} = 3$, is $\approx 30\%$. Negative skew shifts the minimum to the right such that, around the money, the implied volatility follows a straight line with negative slope (well observed for options on stock indices, which have a large historical skew). Finally, note that the heteroskedasticity of financial time series (see Chapter 2) leads to an anomalously slow decay of kurtosis (less than $1/N$), and thus a very slow flattening of the smile with maturity.

9.3.5 Model-generated crashes

While they are often formidably comfortable from the mathematical perspective, continuous time descriptions of reality should always be considered carefully. Discrete time (even with small time steps) combined with the presence of market jumps (equivalently fat-tailed returns or positive kurtosis), reveals that options are always risky, in contradiction with Black and Scholes predictions.

In the early fall of 1987, the unwarranted use of the Black-Scholes model led to the famous world-wide crash known as *Black Monday*, during which the Dow Jones index dropped a memorable 23% in

[15] An option *out-of-the-money* is one that is worthless if exercised today. For a call, it this is the case if the current price of the underlying is less than the strike, equivalently $\mathscr{M} > 0$. Conversely, $\mathscr{M} < 0$ is referred to as *in-the-money*.

a single day. Ironically enough, it was the very use of a crash-free model (Gaussian tails, no extreme events) that helped to trigger a crash.[16] Let us quickly go through this instructive event which should have been a lesson to remember better.

The idea of *portfolio insurance*, principally due to Leland, O'Brien and Rubinstein (LOR) [118, 119], consisted in using Black-Scholes theory to guide trading such as to set a floor below which the value of an investment portfolio cannot fall. One may wonder, if the objective was to ensure the possibility to sell the assets at a guaranteed price level, why not just buy a put option? While as mentioned above, organised options exchanges existed since 1973, they were actually quite limited [120]. Chiefly because the SEC was back then rather suspicious of derivatives, they were restricted to short-term, only on individual stocks (no stock indices), and there were limits on the size of positions that could be accumulated, making them unsuitable for the insurance of large diversified portfolios. Black and Scholes had showed that it was possible to mirror, or *replicate*, perfectly the returns on an option by continuously adjusting a position on the underlying. While BS had used the idea to compute options' costs, LOR focused on the interest of the replicating portfolio itself to manufacture a synthetic put, and implementing large scale portfolio insurance.

Such products were proposed to investors as a substitute for genuine put options, and in the mid-80s portfolio insurance became a big business. Theoretically the idea was a good one, but that was without counting on Black-Scholes flaws, the fact that liquidity is finite and that market impact is real. In 1987 LOR managed over 50B$ when the daily liquidity back then was of the order of 5B$ only. With such orders of magnitude, it is fairly easy to understand that with abnormal price swings, rebalancing the replicating portfolio without totally swamping the market would take a long time. This sets the real system quite far from Black-Scholes continuous time hypothesis... Further, the strategy, which basically consisted in buying stocks as prices rose and selling them as the value of the portfolio fell towards its floor, was a clear unstable feedback loop which would do nothing but amplify the swings. And this is precisely what happened on October 19, 1987.[17] See e.g. [120–124].

9.4 The Financial Modelers' Manifesto

Financial engineering provides a slew of products that allow one to act on the PnL distribution of investors, regardless of whether this is done to responsibly manage risk or smuggle it. The most important message, illustrated notably by the Black-Scholes example, is that relying on models based on incorrect assumptions can have dramatic effects. One should be extremely cautious with the use of quantitative models in finance, and always carefully question the validity of all the explicit or implicit hypotheses of the model by comparing them, as much as possible, with empirical data. One should also attempt to anticipate the feedback loops that the use of a model can create in markets. Trading, even in limited amounts, always impacts the market and can initiate model-generated crashes [123].

A sort of Hippocratic oath, called the *Financial Modelers' Manifesto* [125], was proposed by famous financial engineers Derman and Wilmott after the financial crisis of 2007-2008 to promote more responsibility in risk management and quant finance in general:

- *I will remember that I didn't make the world, and it doesn't satisfy my equations.*
- *Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.*
- *I will never sacrifice reality for elegance without explaining why I have done so.*
- *Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.*
- *I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.*

[16] A very similar story took place in 2008. Pricing models were again fundamentally flawed as they underestimated the probability of global systemic risk (multiple borrowers defaulting on their loans simultaneously). By neglecting the very possibility of a global crisis, they contributed to triggering one [117].

[17] While some may argue that portfolio insurance did not trigger the crash, it is clear that it was the catalyst, responsible for the snowball effect that exacerbated it.

Or as they had already said in previous occasions:

> *"There are always implicit assumptions behind a model and its solution method. But human beings have limited foresight and great imagination, so that, inevitably, a model will be used in ways its creator never intended. This is especially true in trading environments [...] but it's also a matter of principle: you just cannot foresee everything. So, even a correct model, correctly solved, can lead to problems. The more complex the model, the greater this possibility."*
>
> – Emanuel Derman, 1996

> *"Unfortunately, as the mathematics of finance reaches higher levels so the level of common sense seems to drop. There have been some well publicised cases of large losses sustained by companies because of their lack of understanding of financial instruments [...]. It is clear that a major rethink is desperately required if the world is to avoid a mathematician-led market meltdown."*
>
> – Paul Wilmott, 2000

Appendices

Appendix A

Choice theory and decision rules

Here we introduce some important ideas on classical choice theory and discuss its grounds.

A.1 The logit rule

A.1.1 What is it?

Classical choice theory assumes that the probability p_α to choose alternative α in a set $\mathscr{A}$ of different possibilities is given by the *logit rule* or *quantal response* [48]:

$$p_\alpha = \frac{1}{Z} e^{\beta u_\alpha}, \quad \text{with} \quad Z = \sum_{\gamma \in \mathscr{A}} e^{\beta u_\gamma}, \tag{A.1}$$

where u_α denotes the *utility* of alternative α,[1] and β is a parameter that allows to interpolate between deterministic utility maximization ($\beta \to \infty$) and equiprobable choices or full indifference ($\beta = 0$).[2] Indeed, in the $\beta = 0$ limit one obtains $\forall \alpha,\ p_\alpha = 1/N$ with $N = \text{card}(\mathscr{A})$, regardless of the utilities. In the $\beta \to \infty$ limit, one obtains that all p_α are zero except for $p_{\alpha_{\max}} = 1$ where $\alpha_{\max} = \text{argmax}_{\alpha \in \mathscr{A}}(u_\alpha)$. In analogy with statistical physics, $\beta := 1/T$ is often called *inverse temperature*. In the 2D case $\mathscr{A} = \{1, 2\}$ one has:

$$p_1 = \frac{e^{\beta u_1}}{e^{\beta u_1} + e^{\beta u_2}} = \frac{e^{\beta \Delta u}}{1 + e^{\beta \Delta u}} = \frac{1}{2}\left[1 + \tanh\left(\tfrac{\beta \Delta u}{2}\right)\right], \quad \text{with} \quad \Delta u = u_1 - u_2, \tag{A.2}$$

and $p_2 = 1 - p_1$. See Fig. A.1 for an illustration of the limit cases discussed above.

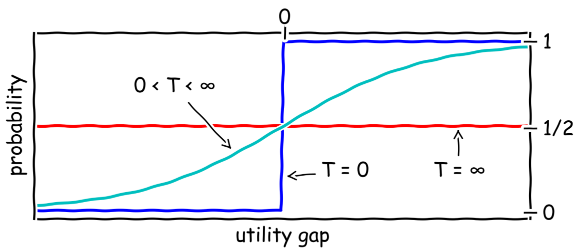


Figure A.1: Probability p_1 as function $\Delta u = u_1 - u_2$ for three different values of the *temperature* $T = 1/\beta$.

[1] Utility maximisation in economics is tantamount to energy minimisation in physics; one might define an energy scale for the alternatives as $e_\gamma = -u_\gamma$.

[2] There is no reason to think that the inverse temperature β should be the same for all agents. Some may be more/less rational than others, such there could be a distribution of β's. This however falls beyond the scope of this course.

A.1.2 How should I interpret it?

There are actually two possible interpretations of the logit rule.

- The first one is *static*: any given agent i always makes the same choice, based on the maximisation of his perceived utility $\hat{u}^i_\alpha = u_\alpha + \epsilon^i_\alpha$ where the ϵ^i_α are random variables fixed in time. Agents' estimates of the utilities are *noisy* [126]; such noise varies across agents because of irrational effects such as illusions and intuitions, or cognitive limitations. In such a case and for a large population, p_α is the fraction of agents having made choice α.
- The second one is *dynamic*: agents continuously flip between different alternatives, because of a truly time evolving context or environment, or just because people change their minds all the time even when confronted to the very same information. The perceived utility now writes $\hat{u}^i_\alpha = u_\alpha + \epsilon^i_\alpha(t)$ where $\epsilon^i_\alpha(t)$ are time dependent random variables. Here, p_α is the probability for a given agent to choose α at a given instant in time.

The second interpretation is more sound, especially when considering interacting agents.

A.1.3 Is it justified?

Well, not really. Typical justifications found in the literature are given below.

- *Axiomatic* [48].
- If one considers that the ϵ^i_α are iid random variables distributed according to a Gumbel law (double-exponential), it is possible to show that the probability is indeed given by Eq. (A.1) [127]. The deep reason for choosing a Gumbel distribution remains rather elusive.
- Rational choice theory goes with the assumption that the agent considers *all* available choices presented to him, weighs their utilities against one another, and then makes his choice. A number of criticisms to this view of human behaviour have emerged, with e.g. Simon [128] as a key figure. Simon highlighted that individuals may be "satisfiers" rather than pure optimisers, in the sense that there is both a computational cost and a cognitive bias related to considering the universe of available choices. This led to the idea of bounded rationality as a way to model real agents [129–132]. Also Schwartz [133] observed that while standard economic theory advocates for the largest number of options possible, "more can be less" due to both cognitive and processing limitations. In other words, agents maximizing their utility take into account the information cost (or the entropy) which according to Shannon [134] writes $S = -\sum_\alpha p_\alpha \log p_\alpha$. In this *exploration-exploitation* setting, maximizing $F = U + TS$ with $U = \sum_\alpha p_\alpha u_\alpha$ naturally yields $p_\alpha \sim \exp(\beta u_\alpha)$, see [135]. Although clearly more sound than the previous argument, there are no solid behavioral grounds supporting it either.
- The ultimate (and honest) argument is that the logit rule is mathematically very convenient. Indeed, it is non other than the Boltzmann-Gibbs distribution used in statistical physics for which a number of analytical results are known.

A.2 Master equation

Provided there is no learning or feedback inducing some systematic evolution of the utilities, one can think of a dynamical model of decisions as the following *Markov chain*. At each time step, agent i reviews his alternatives $\gamma \neq \alpha$ sequentially. He decides to go for alternative γ with probability equal to the probability that $\hat{u}^i_\gamma > \hat{u}^i_\alpha$, equivalently $\epsilon^i_\gamma(t) - \epsilon^i_\alpha(t) > u_\alpha - u_\gamma$. In other words, alternative γ is adopted with probability $R_>(u_\alpha - u_\gamma)$ where $R_>$ is the complementary cumulative distribution function (ccdf) of noise differences, see Fig. A.2.[3]

The *Master equation* is the equation for the evolution of the probabilities p_α, which here thus writes $p_\alpha(t+1) - p_\alpha(t) = \mathrm{d}p_\alpha$ where:

$$\mathrm{d}p_\alpha = \sum_{\gamma \neq \alpha} p_\gamma(t) R_>(u_\gamma - u_\alpha) - \sum_{\gamma \neq \alpha} p_\alpha(t) R_>(u_\alpha - u_\gamma). \tag{A.3}$$

[3] Denoting $r(\Delta\epsilon)$ the probability distribution function of noise difference, the ccdf writes $R_>(\Delta\epsilon) = \int_{\Delta\epsilon}^{\infty} r(x)\mathrm{d}x$.

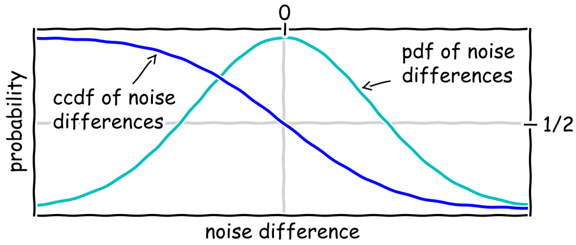


Figure A.2: Probability distribution function (pdf) of noise differences $\Delta\epsilon$ and corresponding ccdf.

Note that the idea of locality would translate in the restriction of the sums to a subset of "neighbouring" choices only.

A.3 Detailed balance

A sufficient condition for equilibrium ($dp_\alpha = 0$) is that each term in the sum of Eq. (A.3) disappears independently:

$$p_\gamma(t)R_>(u_\gamma - u_\alpha) = p_\alpha(t)R_>(u_\alpha - u_\gamma)\,, \tag{A.4}$$

for all α, γ. Equation (A.4) is called the *detailed balance*; it translates the idea in statistical physics that, at equilibrium, each microscopic process should be balanced by its reverse process (microscopic reversibility) [136]. For Eq. (A.1) to be the equilibrium solution of Eq. (A.3), it thus suffices that $R_>(-\Delta u) = e^{\beta\Delta u}R_>(\Delta u)$. While the logistic function $R_>(\Delta u) = 1/[1+\exp(\beta\Delta u)]$ is a natural solution, there might very well be other possibilities. In particular, provided the distribution of noise differences is symmetric, there exists a function $F(\Delta u) = 1/2 - R_>(\Delta u)$ such that $F(0) = 0$, $F(-\Delta u) = -F(\Delta u)$ and $F'(0) \geq 0$. A Taylor expansion at small utility differences Δu reveals:

$$\log\left[\frac{R_>(-\Delta u)}{R_>(\Delta u)}\right] = 4F'(0)\Delta u + O(\Delta u^3)\,, \tag{A.5}$$

such that the detailed balance is always satisfied to first order with $\beta = 4F'(0)$,[4] but may be violated for higher orders in utility differences.

Very little is known about detailed-balance-violating models. In statistical physics this question is relevant for the dynamics of out-of-equilibrium systems. But when it comes to the decisions of people, it is a highly relevant question even for equilibrium systems since there are no solid grounds to support the detailed balance (and the logit rule). In all the following we will assume the logit rule as a sound and simple model for decision making, but one should always remain critical and refrain from drawing quantitative conclusions.

[4]Interestingly enough, this argument allows to relate the temperature to the distribution of noise differences. If $r(\Delta\epsilon)$ is Gaussian with standard deviation σ, one finds $\beta \sim 1/\sigma$.

Appendix B

Imitation of the past

In the previous Chapter we focused on the effects of interactions with the peers. Here, we explore interactions with the past and their consequences.

B.1 Memory effects

As we have seen in Appendix A, the key assumption in rational choice theory is that individuals set their preferences according to an utility maximization principle. Each choice an individual can make is assigned an utility, measuring the satisfaction it provides to the agent and frequently related to the dispassionate forecast of a related payoff. As mentioned in Chapter A, a number of criticisms to this view exist [128–132], with in particular Kahneman [137, 138] who pointed at significant divergences between economics and psychology in their assumptions of human behaviour, with a special emphasis on the empirical evidence of the cognitive biases and the irrationality that guide individual behaviour.

> *"There is a steadily accumulating body of evidence that people, even in carefully set up experimental conditions, do not behave as they are supposed to do in theory. Heaven alone knows what they do when they are let loose in the real world with all its distractions. (...) This said, it seems reasonable to assume that people are inclined to move towards preferable situations in some more limited sense and not to perversely choose outcomes which make them feel worse off. But, one can think of many ways in which this could be expressed and one does not need to impose the formal structure on preferences that we have become used to. People may use simple rules of thumb and may learn what it is that makes them feel better off, they may have thresholds which when attained, push them to react."*
>
> – Alan Kirman

In this line of thought, an interesting idea is that the utility, or well-being, associated to a certain decision may depend on our memory if it has already been made in the past, see e.g. [139].

B.1.1 Estimation error and learning

Realistically, the utility of a given choice cannot be thought of as time-independent. It is hard to know how satisfying or "useful" choice α is without having tried it at least once (think of having to choose among restaurants). The simplest way to model this effect is to write as in Chapter A the effective or *perceived* utility as:

$$\hat{u}_\alpha(t) = u_\alpha + \epsilon_\alpha(t), \quad \text{where now } \epsilon_\alpha(t) = \epsilon_\alpha(0)e^{-\Gamma_\alpha t}. \tag{B.1}$$

This encodes that the perceived utility is initially blurred by some estimation error that decays to zero as the time standing by α grows; Γ_α^{-1} is the *typical learning time*. Note that this is somewhat tantamount to having an inverse temperature β which increases with time.

B.1.2 Habit formation

Another more interesting and intricate effect is *habit formation*. As agents explore and learn the utilities with time, choices become more valuable only because they happened to be chosen. This can be related to the formation of loyalty relationships with, say, your doctor (who knows best your medical history) or your fishmonger (see Chapter C), or because of risk aversion (another choice might be better, but also much worse) or simple intellectual laziness.[1] In order to model such effects in a simple way, one can write:

$$\hat{u}_\alpha(t) = u_\alpha + \sum_{t'=0}^{t} \phi(t-t')\mathbf{1}_{\gamma(t')=\alpha}, \tag{B.2}$$

where the first term on the RHS is the intrinsic utility of choice α, while the second accounts for memory effects. In other terms, the utility assigned to a choice is the sum of a "bare" component indicating a choice's objective worth plus a memory term affecting the utility of that choice whenever the individual has picked it in the past (see Fig. B.1).[2] ϕ is a decaying memory kernel encoding that more recent choices have a stronger effect, and $\gamma(t)$ indicates the choice of the individual at time t.

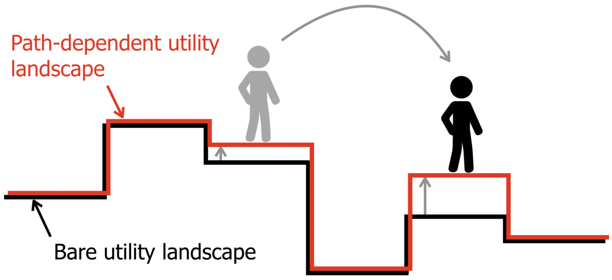


Figure B.1: Illustration of habit formation in an exploration-exploitation setup.

The sign of the kernel separates two different cases: $\phi < 0$ indicates a situation where an individual grows weary of his past choices, while $\phi > 0$ corresponds to the case where an individual becomes increasingly endeared with them. The former case leads to an exploration of all the choices. The latter presents an interesting transition to a *self-trapping* regime when feedback is strong enough and memory decays slowly enough. Memory effects hinder the exploration of all choices by the agent, and may even cause him to leave the optimal choices unexplored: the agent remains stuck with an *a priori* suboptimal choice (out of equilibirum), see [140].

B.2 Self-fulfilling prophecies

When both imitation of the past and imitation of the peers are present, they can reinforce each other leading to *self-fulfilling prophecies*. More precisely, for the latter to arise one generally needs:

- A context of repeated decisions (such as investing in financial markets).
- A common temptation to compare the present situation with similar situations from the past (people tend to think that what already happened is more likely to happen again).
- Some plausible story or *narrative* made up to explain the observed or imagined patterns.

Such a *narrative* then convinces more agents that the effect is real, and their resulting behaviour creates and reinforces the effect. In other terms, a large consensus among agents about the correlations between a piece of information and the system's reaction can be enough to establish these correlations. Keynes called such a commonly shared representation of the world on which uncertain agents can rely a *convention*, he noted that [141]:

[1] Hotel and restaurant chains rely on this strong universal principle: people often tend to prefer things they know.

[2] One may also think, in the physicist's language, of an energy landscape (akin to minus the utility) where the energy of a given site or configuration increases or decreases if the system has already visited that site.

> *"A conventional valuation which is established as the outcome of the mass psychology of a large number of ignorant individuals is liable to change violently as the result of a sudden fluctuation of opinion due to factors which do not really make much difference."*
>
> – John Maynard Keynes

A concrete example of self-fulfilling prophecy with a sudden change of convention is that of the correlation between bond and stock markets as a function of time. The sign of this correlation has switched many times in the past. The last one was took place during the 1997 Asian crisis. Before 1997, the correlation was positive, consistent with the belief that low long term interest rates should favor stocks (since bonds move opposite to rates, an increase in bond prices should trigger an increase in stock prices). But another story suddenly took over when a fall in stock markets triggered an increased anxiety of the operators who sold their risky equity to replace it with non-risky government bonds (Flight to Quality), which then became the dominant pattern.

Appendix C

Fish markets

Financial markets are among the most sophisticated and scrutinised markets. They are different from other markets on many grounds. To highlight their peculiarities, in this appendix we present a very different kind of markets: fish markets.

On a general note, interactions and transactions most often take place in markets. While this might seem rather obvious once said, it is surprising to see how more often than not markets are completely ignored in standard models.

> *"It is a peculiar fact that the literature on economics and economic history contains so little discussion of the central institution that underlies neoclassical economics – the market."*
>
> – Douglass North, 1977

C.1 Why fish markets?

A priori, one expects that different markets with different rules will display different features and yield different outcomes. Stock markets, commodities markets, auction markets or art auctions to name a few differ on many grounds. Here we focus on a simple market that exists since the dawn of time: fish markets. Such markets are particularly adapted to academic analysis for several reasons.

- Transactions are public and available to us, in particular thanks to famous economist Alan Kirman [142–147] who thoroughly recorded them in different market places.
- Fish markets are simple notably because Fish is a perishable good and as a consequence there is no stock management to deal with from one day to the next.
- They can be of different sorts, from peer-to-peer (P2P) such as Marseille, to centralised auctions such as Ancona, Tokyo or Sydney. We can thus expect to learn something from the differences and similarities of their outcomes.

As we shall see, perhaps surprisingly, from all the *a priori* disorder (several buyers and sellers with different needs and prices), aggregate coordination emerges making the whole thing rather efficient. Such aggregate behaviour displays a number a regularities, but the latter are clearly not the result of isolated optimisers, they cannot be attributed to individual rationality, nor can they be accounted for in the standard competitive framework. The aim of this Chapter is precisely to understand and model aggregate coordination from the perspective of agents who learn from their past experience, rather than optimise estimates of their future utilities.

> *"What is the meaning of having preferences over future bundles of goods? How do I know what my preferences will be when I arrive at a future point in time? In particular, if my experiences influence my tastes how can I know what I will turn out to prefer. [...] There was an advertisement for Guinness which said, 'I don't like Guinness. That's why I have never tried it'. This seems absurd to most people but is perfectly consistent with an economist's view of preferences. Since*

my preferences are well defined I do, in fact, know whether I like Guinness or not. Therefore there is no reason for me to try it if I happen not to like it."

– Alan Kirman

C.2 The Marseille fish market

The Marseille fish market runs from 2 to 6am. A few hundreds of buyers interact with a few tens of sellers to exchange a little bit more than a hundred kinds of fish. In contrast with auction markets, prices are not publicly displayed in real time. There is no or little room for negotiation (take or leave prices). Most importantly, the uncertainty about the quantity of available fish to buy at the beginning of the day is very strong.

According to competitive equilibrium theory, and naively, one would expect that (i) after a short relaxation time, the price of a given fish kind converges to some fair value, and (ii) prices decrease with time in order to boost sales and reduce stock losses. Actually, neither is supported by empirical data. Instead, there is no equilibrium price (large inter-agent and inter-temporal fluctuations persist). One can nonetheless define an equilibrium distribution of prices which is stable at time scales larger than a month. While difficult to measure, empirical data shows signs of learning (see supporting documents).

"The habits and relationships that people have developed over time seem to correspond much more to things learnt by the force of experience rather than to conscious calculation. "

– Alan Kirman

C.3 Trading relationships and loyalty formation

C.3.1 A simple model

Following the work of Weisbuch and Kirman [148], we present a simple framework to understand the impact of loyalty on the outcomes of markets such as Marseille. The main idea is that instead of anticipating the future utility of choosing one seller or another, buyers develop affinities based on past experience. Consider:

- N buyers $i \in [1, N]$ and M sellers $j \in [1, M]$.
- Buyer i updates his probability $p_{ij}(t)$ of choosing seller j according to a matrix of "preferences" $J_{ij}(t) \geq 0$, which depends on the accumulated profits resulting from $i - j$ transactions until time t. In particular, one can choose:

$$J_{ij}(t) = \pi_{ij}(t) + (1-\gamma)J_{ij}(t-1), \tag{C.1}$$

  where $\pi_{ij}(t)$ denotes the accumulated profits and γ is a discount factor accounting for finite memory effects (the typical memory time scale is γ^{-1}).
- We choose logit (or Boltzman) statistics:

$$p_{ij}(t) = \frac{1}{Z_i} e^{\beta J_{ij}(t)}, \quad \text{with } Z_i = \sum_k e^{\beta J_{ik}(t)}. \tag{C.2}$$

C.3.2 Mean Field Analysis

We start by looking at the mean field approximation. Here, it amounts to replacing random variables by the mean values $\pi_{ij} \to \langle \pi_{ij} \rangle$ and taking the continuous time limit:

$$\partial_t J_{ij} = -\gamma J_{ij} + \langle \pi_{ij} \rangle. \tag{C.3}$$

We further note that $\langle \pi_{ij} \rangle$ relies on (i) seller j still having some quantity q_j of fish on his stand, *and* (ii) i visiting j, such that one can write:

$$\langle \pi_{ij} \rangle = \mathbb{P}(q_j > 0) p_{ij} \bar{\pi}\,, \tag{C.4}$$

with $\bar{\pi}$ the average profit obtained from getting a quantity q_j.[1]

Let us consider a simple case with only $M = 2$ sellers who always have fish available $\mathbb{P}(q_j > 0) = 1$. Since in such a case buyers do not interact,[2] we leave the i index out and focus on the one buyer having to choose between the two sellers. If there is an equilibrium, it satisfies in particular $\partial_t J_1 = 0$, which combined with Eqs. (C.2), (C.3) and (C.4) yields $\gamma J_1 = \bar{\pi} Z^{-1} e^{\beta J_1}$, and symmetrically $\gamma J_2 = \bar{\pi} Z^{-1} e^{\beta J_2}$ with $Z = e^{\beta J_1} + e^{\beta J_2}$. Defining $\Delta = J_1 - J_2$, one obtains:

$$\Delta = \frac{\bar{\pi}}{\gamma} \frac{e^{\beta\Delta} - 1}{e^{\beta\Delta} + 1} = \frac{\bar{\pi}}{\gamma} \tanh\left(\frac{\beta\Delta}{2}\right). \tag{C.5}$$

As in the RFIM, Eq. (C.5) can be solved graphically to obtain the results displayed in Fig. C.1.

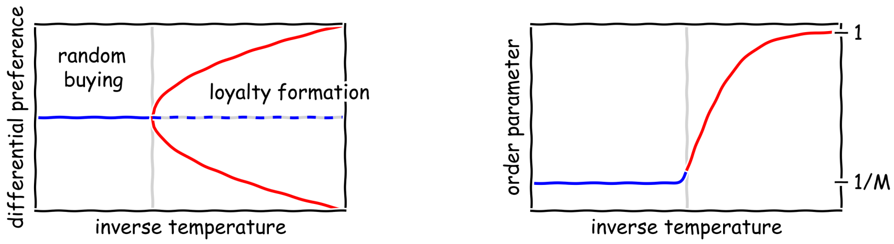


Figure C.1: Loyalty formation bifurcation and order parameter in the Weisbuch-Kirman model [148].

For $\beta < \beta_c := 2\gamma/\pi$, there is only one solution $\Delta = 0$ corresponding to $J_1 = J_2 = \pi/2\gamma$, the buyer visits both sellers with equal probability. For $\beta > \beta_c$, the symmetric solution becomes unstable and two nonzero symmetric solutions $\Delta_\pm$ appear. In particular for $\beta \gtrsim \beta_c$, $\Delta_\pm \sim \pm\sqrt{\beta - \beta_c}$. In this regime, the buyer effectively prefers one of the two sellers and visits him more frequently, he becomes *loyal*. Loyalty formation is more likely when memory is deep: $\beta_c \downarrow$ when $\gamma^{-1} \uparrow$. Let us stress that loyalty emerges spontaneously and is accompanied with symmetry breaking: none of the two sellers is objectively better than the other.

C.3.3 Beyond Mean Field

Simulating the stochastic process described above for many buyers and sellers allows to obtain a large amount of surrogate data to play with. A good order parameter to describe the state of the system, inspired from [149], writes:

$$y_i = \frac{\sum_j J_{ij}^2}{\left(\sum_j J_{ij}\right)^2}, \tag{C.6}$$

In the disorganised phase in which buyer i chooses among the M sellers with equal probability ($\partial_j J_{ij} = 0$), one has $y_i = M J_{ij}^2 / M^2 J_{ij}^2 = 1/M$. In the fully organised phase in which buyer i is loyal to seller k only ($J_{ij} \sim \delta_{jk}$), one finds $y_i = 1$. In particular $1/y_i$ represents the number of sellers visited by buyer i. The simulated data is consistent with a transition from $y_i \approx 1/M$ to 1 when β is increased beyond a certain β_c (see Fig. C.1 and supporting documents).

[1] In full generality $\bar{\pi}_{ij}$ depends on both i and j, is the fish going to be sold at a mall's food court or in fancy restaurant? For the sake of simplicity, we here focus on a symmetric case $\bar{\pi}_{ij} = \bar{\pi}$ for all i, j.

[2] In the general case buyers interact since $\mathbb{P}(q_j > 0)$ depends on other buyers having visited seller j before i.

Interestingly, fluctuations (e.g. of the number of visitors per seller) vanish in the fully organised phase ($\beta \gg \beta_c$), the buyer-seller network becomes deterministic. One might thus argue that the loyal phase is "Pareto superior" given that the deterministic character of the interactions allows sellers to estimate better the quantity of fish that they will sell, avoid unsold fish wasted at the end of the day, which translates into higher profits than in the disorganised phase ($\beta < \beta_c$) for both customers and sellers. This provides a nice example in which spontaneous aggregate coordination is beneficial.

C.3.4 Heterogeneities and real data

So far we have assumed that all buyers are equally rational or irrational and have equal memory (same γ and thus same β_c). In a real market, one expects highly heterogeneous agents, some being naturally more loyal than others. This means that there are different $\beta_c^i = 2\gamma_i/\bar{\pi}_i$; average profit $\bar{\pi}_i$ may vary depending on what the buyer does with the product and γ_i may vary with the number of visits to the market (every day, every week etc.).

Taking the simple case of a population with two sorts of agents $\beta_c^{\mathrm{a}} \ll \beta_c^{\mathrm{b}}$ shows that, in the regime $\beta_c^{\mathrm{a}} < \beta < \beta_c^{\mathrm{b}}$, the presence of noisy b-agents does not prevent loyalty formation. This is not a trivial result: one could have expected that sellers, less capable to properly anticipate their fish sales, could end up out of fish when loyal customers arrive, by that degrading the loyalty relationship.

Actually, real data from the Marseille fish market is very well fitted by the simple case we just discussed of a population with two sorts of agents. The *fidelity histogram*, namely the number of buyers against the number of visited sellers, is bimodal with a large peak at $x = 1$ seller per buyer and a second mode at $x \approx 4$ to 5 sellers per customer (see supporting documents). Interestingly, there are very few buyers in the partially loyal regime, people seem to be either loyal or "persistent explorers" indicating that the loyalty transition is rather sharp.

As a conclusion, this very simple model does a good job at reproducing robust stylised facts in the Marseille P2P fish market. Naturally, the model can be extended at will to account for more complexity (e.g. sellers could propose different prices to different buyers).

C.4 The impact of market organisation

As mentioned above, other markets used different mechanisms. Here we explore how aggregate properties are affected by market organisation [147].

C.4.1 The Ancona fish market

The Ancona fish market, or MERITAN for 'MERcato ITtico ANcona', operates 4 days a week from 3:30 to 7:30am. It is organised as a *Dutch auction* (high initial price which decreases with time until a buyer manifests himself). A few tends of boats present their fish to a few hundreds of buyers which results in ≈ 15 transaction per minute and 25 million euros a year.

In contrast with Marseille, buyers and sellers interact through a centralised system. The buyer-seller network counts $N + M$ links, instead of $N \times M$. Another notable difference with Marseille is that here everyone can see who buys what to whom.

C.4.2 Similarities and differences

Similar to Marseille, signs of learning are distinguishable with in particular a decay of price fluctuations with time. The most interesting question is probably: does the central auction mechanism destroy the loyalty arising from buyer-seller relationships? Data reveals that loyalty also emerges in Ancona; its structure is nonetheless quite different.

To evaluate the degree of loyalty quantitatively, one can compute the *Gini index* which is a good proxy of how much the purchases of a buyer are distributed among the different sellers. It is computed as follows:

- The *Lorentz curve* is computed for one buyer and M sellers.
- The M sellers are ranked on the x-axis from the least visited to the most visited.
- The cumulative number of visits per seller (y-axis) is plotted against x.
- The axes are normalised to $[0, 1]$ and the Gini index $G \in [0, 1]$ is computed as twice the area separating the Lorentz curve and the bisectrix. If all sellers are equally visited by the buyer $G = 0$ (no loyalty), and if the buyer visited only one seller $G = 1$ (fully loyal).

Looking at the pdf of Gini indices over all buyers reveal that loyalty also exists in Ancona ($G \approx 0.4 > 0$), but the distribution is unimodal, in contrast with Marseille where it is bimodal. One can argue that the central auction mechanism erases the behavioral binarity.

Concluding remarks

Markets are a fundamental element of every economy. Fish markets show that the aggregate regularities and coordination arise from the interactions between highly heterogeneous agents. While in Marseille nothing prevents buyers to wander around and just pick the cheapest seller as would be required by the standard model, this is not what happens. Most stylised facts revealed by data cannot be reasonably accounted for with isolated representative agents; they can instead be reproduced in simple agent-based models with memory but limited intelligence.[3] Finally, differences in market organisation can lead to differences in the aggregate results.

> *"Aggregate regularity should not be considered as corresponding to individual rationality. (...) The fact that we observe an aggregate result which conforms to the predictions of a particular model is not enough to justify the conclusion that individuals are actually behaving as they are assumed to do in that model.*
>
> – Alan Kirman

[3]While the gap between micro and macro behavior is not straightforward, one does not need to take into account all the complexity at the individual scale to understand aggregate behavior.

Bibliography

[1] L. Bachelier, *Théorie de la Spéculation*, Ph.D. thesis, Ecole Normale Supérieure (1900).

[2] J.-P. Bouchaud and M. Potters, *Theory of financial risk and derivative pricing: from statistical physics to risk management* (Cambridge university press, 2003).

[3] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, quotes and prices: financial markets under the microscope* (Cambridge University Press, 2018).

[4] R. N. Mantegna and H. E. Stanley, *Introduction to econophysics: correlations and complexity in finance* (Cambridge university press, 1999).

[5] B. Sharma, S. Agrawal, M. Sharma, D. Bisen, and R. Sharma, arXiv preprint arXiv:1108.0977 (2011).

[6] J. Hasbrouck, *Empirical market microstructure: The institutions, economics, and econometrics of securities trading* (Oxford University Press, 2007).

[7] P. Wilmott, *Paul Wilmott on Quantitative Finance-3 Volume Set* (John Wiley & Sons, 2006).

[8] L. Sophie and L. Charles-albert, *Market microstructure in practice* (World Scientific, 2018).

[9] R. Cont, *Encyclopedia of quantitative finance* (Wiley, 2010).

[10] P. Wilmott *et al.*, *Derivatives: The theory and practice of financial engineering* (Wiley Chichester, 1998).

[11] E. Sinclair, *Volatility trading* (John Wiley & Sons, 2013).

[12] M. Levinson *et al.*, *The economist guide to financial markets: Why they exist and how they work* (The Economist, 2014).

[13] A. Ilmanen, *Expected returns: An investor's guide to harvesting market rewards* (John Wiley & Sons, 2011).

[14] I. Tulchinsky, *Finding Alphas: A quantitative approach to building trading strategies* (John Wiley & Sons, 2019).

[15] O. Vasicek, Journal of financial economics **5**, 177 (1977).

[16] B. B. Mandelbrot, Journal of Fluid Mechanics **62**, 331 (1974).

[17] J.-P. Bouchaud and J.-F. Muzy, in *The Kolmogorov Legacy in Physics* (Springer, 2003) pp. 229–246.

[18] A. Einstein, Annalen der Physik **14**, 182 (2005).

[19] J. Gatheral, T. Jaisson, and M. Rosenbaum, Quantitative Finance **18**, 933 (2018).

[20] A. Joulin, A. Lefevre, D. Grunberg, and J.-P. Bouchaud, Wilmott Magazine **September-October**, 1 (2008).

[21] M. Wyart, J.-P. Bouchaud, J. Kockelkoren, M. Potters, and M. Vettorazzo, Quantitative finance **8**, 41 (2008).

[22] J. Hasbrouck, *Empirical market microstructure: The institutions, economics, and econometrics of securities trading.* (Oxford University Press, 2007).

[23] J.-P. Bouchaud, J. D. Farmer, and F. Lillo, *How markets slowly digest changes in supply and demand* (Elsevier: Academic Press, 2008).

[24] P. Weber and B. Rosenow, Quantitative Finance **5**, 357 (2005).

[25] J.-P. Bouchaud, *Price impact. Encyclopedia of quantitative finance.* (Wiley, 2010).

[26] F. Lillo and J. D. Farmer, Studies in Nonlinear Dynamics & Econometrics **8**, 1 (2004).

[27] Y. Lempérière, C. Deremble, P. Seager, M. Potters, and J.-P. Bouchaud, arXiv preprint arXiv:1404.3274 (2014).

[28] A. Majewski, S. Ciliberti, and J.-P. Bouchaud, Journal of Economic Dynamics and Control , 103791 (2019).

[29] J.-P. Bouchaud, S. Ciliberti, Y. Lemperiere, A. Majewski, P. Seager, and K. Sin Ronia, Available at SSRN 3070850 (2017).

[30] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, Quantitative Finance **4**, 176 (2004).

[31] M. Benzaquen, I. Mastromatteo, Z. Eisler, and J.-P. Bouchaud, J. Stat. Mech. **2017**, 023406 (2017).

[32] I. Mastromatteo, M. Benzaquen, Z. Eisler, and J.-P. Bouchaud, Risk **July 2017** (2017).

[33] L. C. G. del Molino, I. Mastromatteo, M. Benzaquen, and J.-P. Bouchaud, arXiv:1806.07791 (2018).

[34] J. Hasbrouck and D. J. Seppi, Journal of financial Economics **59**, 383 (2001).

[35] A. Boulatov, T. Hendershott, and D. Livdan, The Review of Economic Studies **80**, 35 (2013).

[36] P. Pasquariello and C. Vega, Review of Finance **19**, 229 (2013).

[37] S. Wang, R. Schäfer, and T. Guhr, arXiv:1603.01586 (2016).

[38] S. Wang, R. Schäfer, and T. Guhr, Eur. Phys. J. B **89**, 1 (2016).

[39] R. Almgren and N. Chriss, Journal of Risk **3**, 5 (2001).

[40] J.-F. Muzy, J. Delour, and E. Bacry, The European Physical Journal B-Condensed Matter and Complex Systems **17**, 537 (2000).

[41] P. W. Anderson, Science **177**, 393 (1972).

[42] S. H. Strogatz, *Sync: How order emerges from chaos in the universe, nature, and daily life* (Hachette UK, 2012).

[43] A. P. Kirman, Journal of economic perspectives **6**, 117 (1992).

[44] D. Stauffer and A. Aharony, *Introduction to percolation theory* (Taylor & Francis, 2018).

[45] J.-P. Eckmann and E. Moses, Proceedings of the national academy of sciences **99**, 5825 (2002).

[46] J. P. Sethna, K. Dahmen, S. Kartha, J. A. Krumhansl, B. W. Roberts, and J. D. Shore, Physical Review Letters **70**, 3347 (1993).

[47] S. Galam, Y. Gefen, and Y. Shapir, Journal of Mathematical Sociology **9**, 1 (1982).

[48] S. P. Anderson, A. De Palma, and J.-F. Thisse, *Discrete choice theory of product differentiation* (MIT press, 1992).

[49] Q. Michard and J.-P. Bouchaud, The European Physical Journal B-Condensed Matter and Complex Systems **47**, 151 (2005).

[50] J. Bongaarts and S. C. Watkins, Population and development review , 639 (1996).

[51] E. L. Glaeser, B. Sacerdote, and J. A. Scheinkman, The Quarterly Journal of Economics **111**, 507 (1996).

[52] P. Curty and M. Marsili, Journal of Statistical Mechanics: Theory and Experiment **2006**, P03013 (2006).

[53] G. S. Becker, Journal of political economy **99**, 1109 (1991).

[54] D. S. Scharfstein, J. C. Stein, *et al.*, American Economic Review **80**, 465 (1990).

[55] R. J. Shiller, J. Pound, *et al.*, *Survey evidence on diffusion of interest among institutional investors*, Tech. Rep. (Cowles Foundation for Research in Economics, Yale University, 1986).

[56] A. Kirman, The Quarterly Journal of Economics **108**, 137 (1993).

[57] P. Ehrenfest and T. Ehrenfest-Afanassjewa, *Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem* (Hirzel, 1907).

[58] A. Fosset, J.-P. Bouchaud, and M. Benzaquen, Available at SSRN 3496148 (2019).

[59] D. Challet and Y.-C. Zhang, Physica A: Statistical Mechanics and its Applications **246**, 407 (1997).

[60] D. Challet, M. Marsili, and Y.-C. Zhang, Physica A: Statistical Mechanics and its Applications **294**, 514 (2001).

[61] D. Challet, A. Chessa, M. Marsili, and Y.-C. Zhang, (2001).

[62] B. B. Mandelbrot, *Fractals and scaling in finance: Discontinuity, concentration, risk* (Springer, 1997).

[63] X. Gabaix, Annu. Rev. Econom. **1**, 255 (2009).

[64] P. K. Clark, Econometrica **1**, 135 (41).

[65] C. M. Jones, G. Kaul, and M. L. Lipson, Review of financial studies **7**, 631 (1994).

[66] T. Bollerslev and D. Jubinski, JBES **17**, 9 (1999).

[67] T. Ané and H. Geman, J. of Finance **55**, 2259 (2000).

[68] R. Liesenfeld, J. of Econometrics **104**, 141 (2001).

[69] G. E. Tauchen and M. Pitts, Econometrica **51**, 485 (1983).

[70] R. Engle, Engle, R. F. **68**, 1 (2000).

[71] G. Zumbach, Quantitative finance **4**, 441 (2004).

[72] Z. Eisler and J. Kertecz, Eur. J. Phys. **B51**, 145 (2006).

[73] A. S. Kyle and A. A. Obizhaeva, Econometrica **84**, 1345 (2016).

[74] T. G. Andersen, O. Bondarenko, A. S. Kyle, and A. A. Obizhaeva, Unpublished (2015).

[75] M. Benzaquen, J. Donier, and J.-P. Bouchaud, Market Microstructure and Liquidity **2**, 1650009 (2016).

[76] F. Bucci, F. Lillo, J.-P. Bouchaud, and M. Benzaquen, arXiv:1902.03457 (2019).

[77] A. S. Kyle, Econometrica: Journal of the Econometric Society , 1315 (1985).

[78] M. G. Daniels, J. D. Farmer, L. Gillemot, G. Iori, and E. Smith, Physical review letters **90**, 108102 (2003).

[79] R. C. Grinold and R. N. Kahn, *Active Portfolio Management* (McGraw Hill New York, 2000).

[80] R. Almgren, C. Thum, E. Hauptmann, and H. Li, Risk **18**, 58 (2005).

[81] B. Tóth, Y. Lemperiere, C. Deremble, J. De Lataillade, J. Kockelkoren, and J. P. Bouchaud, Phys. Rev. X **1**, 021006 (2011).

[82] J. Donier and J. Bonart, Market Microstructure and Liquidity **1**, 1550008 (2015).

[83] N. Torre, BARRA Inc., Berkeley (1997).

[84] R. Engle, R. Ferstenberg, and J. Russell, (2006).

[85] I. Mastromatteo, B. Tóth, and J.-P. Bouchaud, Phys. Rev. E **89**, 042805 (2014).

[86] X. Brokmann, E. Serie, J. Kockelkoren, and J.-P. Bouchaud, Market Microstructure and Liquidity **1**, 1550007 (2015).

[87] E. Bacry, A. Iuga, M. Lasnier, and C.-A. Lehalle, Market Microstructure and Liquidity **1**, 1550009 (2015).

[88] N. Bershova and D. Rakhlin, Quantitative finance **13**, 1759 (2013).

[89] J. Farmer, A. Gerig, F. Lillo, and H. Waelbroeck, Quant. Finance **13**, 1743 (2013).

[90] E. Zarinelli, M. Treccani, J. D. Farmer, and F. Lillo, Market Microstructure and Liquidity **1**, 1550004 (2015).

[91] C. Gomes and H. Waelbroeck, Quantitative Finance **15**, 773 (2015).

[92] E. Said, A. B. H. Ayed, A. Husson, F. Abergel, and B. Paribas, arXiv:1802.08502 (2018).

[93] G. Huberman and W. Stanzl, Econometrica **72**, 1247 (2004).

[94] J. Gatheral, Quantitative finance **10**, 749 (2010).

[95] F. Bucci, M. Benzaquen, F. Lillo, and J.-P. Bouchaud, arXiv:1901.05332 (2019).

[96] C.-A. Lehalle, O. Guéant, and J. Razafinimanana, in *Econophysics of Order-driven Markets* (Springer, 2011).

[97] J. Donier, J. F. Bonart, I. Mastromatteo, and J.-P. Bouchaud, Quantitative Finance **15**, 1109 (2015).

[98] M. Benzaquen and J.-P. Bouchaud, Quantitative Finance **18**, 1781 (2018).

[99] M. Benzaquen and J.-P. Bouchaud, Eur. Phys. J. B **91**, 23 (2018).

[100] W. R. Schneider, *Fractional diffusion. In Dynamics and Stochastic Processes Theory and Applications* (Springer Berlin Heidelberg, 1990) pp. 276–286.

[101] R. Metzler and T. F. Nonnenmacher, Chemical Physics **284**, 67 (2002).

[102] H. Markowitz, The Journal of Finance **7**, 77 (1952).

[103] J.-P. Bouchaud and M. Potters, (Oxford: Oxford University Press) (2011).

[104] J. Bun, J.-P. Bouchaud, and M. Potters, Physics Reports **666**, 1 (2017).

[105] V. A. Marchenko and L. A. Pastur, Matematicheskii Sbornik **114**, 507 (1967).

[106] J. Bun, J.-P. Bouchaud, and M. Potters, Physics Reports **666**, 1 (2017).

[107] R. Cont, Quantitative Finance **1**, 223 (2001).

[108] G. Bekaert and G. Wu, The review of financial studies **13**, 1 (2000).

[109] J.-P. Bouchaud, A. Matacz, and M. Potters, Phys. Rev. Lett. **87**, 228701 (2001).

[110] Q. Li, J. Yang, C. Hsiao, and Y.-J. Chang, Journal of Empirical Finance **12**, 650 (2005).

[111] A. Ang and J. Chen, Journal of financial Economics **63**, 443 (2002).

[112] L. Borland, Quantitative Finance **12**, 1367 (2012).

[113] E. Balogh, I. Simonsen, B. Z. Nagy, and Z. Néda, Physical Review E **82**, 066113 (2010).

[114] M. Wyart and J.-P. Bouchaud, Journal of Economic Behavior & Organization **63**, 1 (2007).

[115] D. G. Baur and B. M. Lucey, Journal of Financial stability **5**, 339 (2009).

[116] F. Black and M. Scholes, Journal of political economy **81**, 637 (1973).

[117] D. Colander, H. Föllmer, A. Haas, M. D. Goldberg, K. Juselius, A. Kirman, T. Lux, and B. Sloth, Univ. of Copenhagen Dept. of Economics Discussion Paper (2009).

[118] M. Rubinstein and H. E. Leland, Financial Analysts Journal **37**, 63 (1981).

[119] H. E. Leland, The Journal of Finance **35**, 581 (1980).

[120] D. MacKenzie, Economy and society **33**, 303 (2004).

[121] D. Bates and R. Craine, *Valuing the futures market clearinghouse's default exposure during the 1987 crash*, Tech. Rep. (National Bureau of Economic Research, 1998).

[122] B. B. Burr, Pensions & Investments (1997).

[123] D. MacKenzie, *An engine, not a camera: How financial models shape markets* (Mit Press, 2008).

[124] J.-P. Bouchaud, Nature **455**, 1181 (2008).

[125] E. Derman and P. Wilmott, Available at SSRN 1324878 (2009).

[126] L. L. Thurstone, Psychometrika **10**, 237 (1945).

[127] D. McFadden and P. Zarembka, Conditional logit analysis of qualitative choice behavior , 105 (1974).

[128] H. A. Simon, The Quarterly Journal of Economics **69**, 99 (1955).

[129] H. A. Simon, Decision and organization **1**, 161 (1972).

[130] R. Selten, Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft **146**, 649 (1990).

[131] W. B. Arthur, The American economic review **84**, 406 (1994).

[132] G. Gigerenzer and R. Selten, *Bounded rationality: The adaptive toolbox* (MIT press, 2002).

[133] S. B., Harper Perennial (2004).

[134] C. E. Shannon, Bell System Tech. J **27**, 379 (1948).

[135] J.-P. Nadal, G. Weisbuch, O. Chenevez, and A. Kirman, Advances in self-organization and evolutionary economics, Economica, London , 149 (1998).

[136] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, Vol. 1 (Elsevier, 1992).

[137] D. Kahneman, American Economic Review **93**, 162 (2003).

[138] D. Kahneman and R. H. Thaler, Journal of Economic Perspectives **20**, 221 (2006).

[139] D. Kahneman, in *Choices, Values, and Frames* (Cambridge University Press, 2000) pp. 673–692.

[140] J. Moran, A. Fosset, D. Luzzati, J. Bouchaud, and M. Benzaquen, (2020).

[141] J. M. Keynes, "The general theory of interest, employment and money," (1936).

[142] A. Kirman and A. Vignes, in *Issues in contemporary economics* (Springer, 1991) pp. 160–185.

[143] W. Härdle and A. Kirman, Journal of Econometrics **67**, 227 (1995).

[144] A. P. Kirman and N. J. Vriend, in *Interaction and Market structure* (Springer, 2000) pp. 33–56.

[145] A. P. Kirman and N. J. Vriend, Journal of Economic Dynamics and Control **25**, 459 (2001).

[146] A. Kirman, Networks and markets , 155 (2001).

[147] M. Gallegati, G. Giulioni, A. Kirman, and A. Palestrini, Journal of Economic Behavior & Organization **80**, 20 (2011).

[148] G. Weisbuch, A. Kirman, and D. Herreiner, The economic journal **110**, 411 (2000).

[149] B. Derrida, Souletie, Vannimenus, Stora (eds.) Chance and Matter, North-Holland (1986).

Tutorial sheets

The following tutorials were designed by teaching assistants Jérôme Garnier-Brun, Ruben Zakine, Théo Dessertaine and José Moran.

Tutorial 1: Time series simulation and analysis

Introduction

Through the course and exercise sessions, you are expected to encounter and analyse numerical time series from financial data. The goal of this session is to get comfortable with the numerical analysis of time series, as well as to give you some basic intuition on what can cause the emergence of certain stylized facts in finance.

Part 1 : Fractional Brownian motion (fBM)

The fractional Brownian Motion, introduced by Mandelbrot & van Ness in 1968, is a Gaussian process satisfying

$$\langle x_H(t)x_H(s)\rangle = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}), \quad t,s > 1\,. \tag{9.7}$$

The parameter H is called the Hurst exponent and controls the "roughness" of the process, $H = 1/2$ corresponding to the standard brownian motion.

1. Let $\mathbf{L}$ be a $T \times T$ matrix. Show that the correlation matrix of the vector $\mathbf{y} = \mathbf{Lx}$, where $\mathbf{x}$ is an iid. gaussian vector of size T, is given by $\mathbf{C} = \mathbf{LL}^\top$.
2. Define a function `C(i,j,H)` giving the C_{ij} term of the correlation matrix of an fBM with exponent H. In Python, compute the matrix for an fBM of length $T = 1000$ and the exponent of your choice. Use the Cholesky decomposition with `np.linalg.cholesky` to generate a fractional brownian motion, rescaled as
$$\tilde{x}_H(t) = 1/T^H x_H(t) \tag{9.8}$$
so that all simulated curves have roughly the same scale.
3. Group everything into a function `gen_fbm(T,H)` that returns an instance of a (rescaled) fBM with exponent H of length T. Plot for $H = 0.25, 0.5, 0.75$. Comments?
4. To generate statistics over the fBM process it is necessary to generate many instances of it. Define now a function `gen_N_fbm(T,N,H)` that returns a $T \times N$ array with N instances of the fractional brownian motion for the same exponent H.
5. Create a DataFrame `df` with 100 columns, each containing a realisation of the fBM with $H = 0.3$ and $T = 1000$.
6. Create a function that computes the value $(x(t) - x(t+\tau))^2$, averages it over t for a given realization, and then averages over all of th realisations in `d` for a single value τ. Use it to compute the variogram,
$$V(\tau) = \left\langle (x(t+\tau) - x(t))^2 \right\rangle \tag{9.9}$$
for $0 \leq \tau < 500$.
7. Plot the results alongside the theoretical value $V(\tau) \propto \tau^{2H}$. Do the same for $H = 0.75$ and compare.

8. Compute the volatility signature plots $V(\tau)/\tau$ for the two fBMs you generated and compare. Comments?

Part 2 : The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process (centered around 0) is defined by the following SDE:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\omega x(t) + \sigma\xi(t), \quad \langle \xi(t)\xi(t')\rangle = \delta(t-t'). \tag{9.10}$$

We remind that the discretization according to the Ito prescription of this SDE is given by:

$$x_{(N+1)\Delta t} = x_{N\Delta t}(1-\omega\Delta t) + \sigma\sqrt{\Delta t}\xi_N \tag{9.11}$$

where the ξ_N variables are normally distributed with variance 1.

1. Write a function `ornstein_uhlenbeck(x_0, omega, sigma, dt, T)` that returns a realisation of the Ornstein-Uhlenbeck process with initial condition x_0 and with parameters ω and σ over T steps with a discretization $\mathrm{d}t$.
2. To generate statistics over the Ornstein-Uhlenbeck process it is necessary to generate many instances of it. Write a new function `N_ornstein_uhlenbeck(x_0, omega, sigma, dt, N, T)` that generates a $T \times N$ array where each column is a realisation of the O-U process with corresponding parameters. The x_0 initial condition is now a vector of size N. We remind that in `Python` `a[i,:]` accesses the i-th row of array `a`, and `a[:,j]` its j-th column.
3. Show that the solution of the Ornstein-Uhlenbeck process is given by

$$x(t) = x(0)e^{-\omega t} + \sigma \int_0^t \mathrm{d}s\; e^{-\omega(t-s)}\xi(s). \tag{9.12}$$

4. Now show that $\langle x(t)\rangle \underset{t\to\infty}{\longrightarrow} 0$. Taking $x(0) = 0$ for simplicity, compute the product $x(t)x(t+\tau)$ and average it over ξ to recover the correlation function:

$$C(t,\tau) = \langle x(t)x(t+\tau)\rangle = \frac{\sigma^2}{2\omega}\left(e^{-\omega|\tau|} - e^{-\omega(2t+\tau)}\right). \tag{9.13}$$

   Comment on its behavior at long times.
5. Create a DataFrame `df` where each row is a time point and with 100 columns corresponding to an instance of an Ornstein-Uhlenbeck process with $\sigma = 1$ and your choice of ω. (hint: use `pd.DataFrame(x)` where `x` is the output of a function you have defined). Take again $T = 10^5$ and $\mathrm{d}t = 10^{-3}$.
6. Create a function that computes the correlation $x(t)x(t+\tau\mathrm{d}t)$, averages it over t for a given realization, and then averages over all the realizations in `df` for a single value τ. Compute this correlation function over a set of points (at least up to $\tau\mathrm{d}t = 2$) and compare this with $\sigma/(2\omega)e^{-\omega\mathrm{d}t\tau}$.
7. Compute the variogram $V(\tau)$ of the Ornstein-Uhlenbeck process, as well as the volatility signature plot $V(\tau)/\tau$. Compare it to the fBM. What do you notice?
8. Recasting the equation in a Langevin form

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -U'(x(t)) + \sigma\xi(t), \tag{9.14}$$

   can you provide an intuitive explanation for the variogram you just obtained? Draw a sketch! Bonus: can you give the steady state probability distribution of the particle's position at a glance?

Part 3 : Scale invariance, monofractality and multifractality

In the first exercise, we studied the fractional Brownian motion though the lens of time-shifted correlations. However, it is also an interesting model due to its *scale invariance* properties, that we will now look into.

1. Let $x(t)$ be a fBM of Hurst exponent H. Show that

$$M_q(\tau) = \langle |x(t+\tau) - x(t)|^q \rangle_t \tag{9.15}$$

   can be rewritten as a function of $\sigma(\tau) = \sqrt{M_2(\tau)}$ in this case where the probability to move by Δx in time τ is given by

$$P_\tau(\Delta x) = \frac{1}{\sigma(\tau)} f\left(\frac{\Delta x}{\sigma(\tau)}\right), \qquad f(u) = \frac{\mathrm{e}^{-\frac{1}{2}u^2}}{\sqrt{2\pi}}. \tag{9.16}$$

   Explain why the fBM is "scale-invariant".
2. Verify your finding numerically.
3. You should have found that for the fBM $M_q(\tau) \propto \sigma(\tau)^q$. Such a scaling is referred to as *monofractal*, while a process with $M_q(\tau) \propto \sigma(\tau)^{\zeta(q)}$ with $\zeta(q)$ a non-linear function of q is called *multifractal*. You should have a file `gen_heliumjet_R89.npy` describing the velocity of a turbulent Helium jet (synthetically generated). Simply looking at the time series (note the file includes 4 times series), do you think it can be adequately described by a monofractal process? What about a multifractal one?
4. Plot $M_q(\tau)$ calculated from the experimental data as a function of τ for different values of q. Overlapping $a\sigma(\tau)^q$ on the data (with a a scaling constant), can you confirm or infirm the monofractal nature of the signal?
5. Motivated by turbulent jets and financial time series (see PC3), Bacry, Muzy & Delour introduced the Multifractal Random Walk (MRW), for which it can be shown that

$$\zeta(q) = q - \lambda^2 q(q-2). \tag{9.17}$$

   Play with the expression on the results from the previous question and comment. What do you think is a good guess of λ?

Tutorial 2: Randomness in complex systems

Introduction

The study of complex systems is intrinsically linked to the limiting statistics of large ensembles of random variables. The goal of this tutorial is to provide reminders on central limit theorem(s) and familiarize ourselves with power laws that are ubiquitous in real data. If time permits, we will also introduce random matrices and their role in the understanding of empirical covariance matrices.

Part 1 : Reminders on the central limit theorem

In this exercise, we will consider the sum

$$Y_N = \sum_{i=1}^{N} X_i, \tag{9.18}$$

where the X_i are iid random variables with mean $\langle X_i \rangle = 0$ and variance $\left\langle X_i^2 \right\rangle = \sigma^2$.

1. State the central limit theorem.
2. Consider the characteristic function $G_X(k) = \left\langle \mathrm{e}^{ikX} \right\rangle$. Show that if the X_i are Gaussian, then Y_N is also Gaussian and of variance $N\sigma^2$.
3. Generalize the result to $Z_N = Y_N/\sqrt{N}$ for all distributions of X_i with bounded variance, thus proving the central limit theorem.
4. Mentally prepare yourself to approximate any large sum of random variables as a constant plus Gaussian white noise until the end of the course.
5. (Bonus) Prove the central limit theorem in a statistical physics style, starting from

$$P_{Y_N}(y) = \int_{-\infty}^{\infty} \left(\prod_{i=1}^{N} \mathrm{d}x_i \right) P_{X_1,\ldots,X_N}(x_1, \ldots, x_N)\, \delta\!\left(y - \sum_i x_i \right). \tag{9.19}$$

Hint: use the integral representation $\delta(u) = \int_{\mathbb{R}} \frac{\mathrm{d}\lambda}{2\pi}\, \mathrm{e}^{i\lambda u}$ and the fact that N is large.

Part 2 : The CLT in the real world

In the previous exercise, we made the assumption that the random variables X_i are independent and identically distributed, and derived results for $N \to \infty$. Unfortunately, all of these assumptions are untenable in the real world. We therefore turn to numerical simulations to assess the robustness of the CLT to more realistic conditions.

1. Let us first take X_i to be iid and simply take N to be finite. Start with Laplace distributed random variables,

$$P_X(x) = \frac{1}{2\Delta} \mathrm{e}^{-\frac{|x|}{\Delta}}, \tag{9.20}$$

giving $\sigma^2 = 2\Delta^2$. Run numerical simulations for $N = \{50, 100, 1000\}$ for a chosen value of Δ and plot the histograms of Z. Compare with the Gaussian prediction from the CLT and comment.
2. Repeat the process for a Student-t distribution (`np.random.standard_t`) of parameter $\nu = 3$. Comment after looking up the pdf of the Student-t and comparing it to the Laplace distribution.
3. Finally repeat for a Pareto II distribution (`np.random.pareto`) with parameter $\alpha = 3$.
4. We now wish to study the case where the random variables are not quite independent. Using the method from PC1, generate a sample of N Gaussian variables correlated as

$$\langle X_i X_j \rangle = \sigma^2\, \mathrm{e}^{-\frac{|i-j|}{N_c}}. \tag{9.21}$$

Does the CLT appear to survive? If so, provide a heuristic expression for the variance from your numerical experiments.

5. Repeat but with variables with longer range correlations, for example

$$\langle X_i X_j \rangle = \frac{\sigma^2}{|i-j|^2}. \tag{9.22}$$

What do you observe?

Part 3 : Heavy tails and the generalized CLT

We have seen that heavy-tailed random variables with finite variance do converge do a Gaussian distribution when summed, but do so very slowly as N increases. What about the case where the variance diverges? To explore such "heavy-tailed" random variables, we will consider the Pareto I distribution, defined through its so-called survival function,

$$S_X(x) = \mathbb{P}(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^\alpha & \text{if } x \geq x_m, \\ 1 & \text{if } x < x_m, \end{cases} \tag{9.23}$$

where x_m is the minimum (necessarily positive) value taken by X, and $\alpha > 0$ is a shape parameter.

1. Calculate the mean and variance of X. What range(s) of values of α should we focus on?
2. Rather than the sum of N random variables, let us now consider the maximum value of the draw, that we will write $M_N = \max\{X_1, \ldots, X_N\}$. Write down the general expression for its PDF, $P_{M_N}(m)$, as a function of the PDF and the cumulative distribution of X and check that it is correctly normalized.
3. Based on $P_{M_N}(m)$, show that the most probable value of the maximum M_N^* satisfies

$$N P_X'(M_N^*) F_X(M_N^*)^{N-1} - N(N-1) P_X(M_N^*)^2 F_X(M_N^*)^{N-2} = 0, \tag{9.24}$$

where F_X is the cumulative distribution of X.

4. Show that in the limit of large N, the above equation can be approximated as

$$S_X(M_N^*) \approx \frac{1}{N}. \tag{9.25}$$

Hint: use l'Hopital's rule.

5. Using this expression, compute M_N^* for the uniform $[0,1]$, Gaussian, and Pareto distributions. Hint: the asymptotic behavior of the complementary error function is

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{d}t\, \mathrm{e}^{-t^2} \sim \frac{\mathrm{e}^{-x^2}}{x\sqrt{\pi}}. \tag{9.26}$$

6. In light of this last result, discuss the contribution of the largest term to the fluctuations of the sum Y_N and relate this to your answer to Question 1.
7. It turns out that the CLT generalizes to heavy-tailed random variables with diverging variance (and average). Based on the result for the maximum, can you hazard a guess at the tail behavior of the empirical mean of $N \to \infty$ heavy-tailed random variables? Verify your intuition with a numerical simulation Student-t case. Hint: the pdf of a Student-t of parameter ν decays as $\sim x^{-(\nu+1)}$.
8. This exercice provided a very coarse introduction to extreme value statistics, which has become a cornerstone of statistical mechanics and complexity science. Based on your very first result in the field, could you explain to a friend why comparing countries normalized by their population sizes at the olympics might not be such a smart idea?

Part 4 : An introduction to random matrices

Having considered sequences of scalar random variables, we will now take a look at random matrices. In the context of complex systems, these can represent interactions between agents (i.e. a network), but may also be an essential tool to clean noisy real-world data. In this exercise, we will focus on the latter and more specifically how random matrix theory may help us to estimate the covariance matrix between N time-varying variables $X_1,\ldots,X_N$ from simultaneous realizations (e.g. stocks). Assuming these variables are centered and of unit variance, a standard estimator for the covariance matrix is given by

$$E_{ij} = \frac{1}{T}\sum_{t=1}^{T} X_i^t X_j^t, \tag{9.27}$$

where T is the number of timesteps we have access to.

1. Show that $\mathbf{E}$ is symmetric and positive semi-definite by rewriting it with the $N \times T$ data matrix $\mathbf{H}$.
2. Let $q = N/T$. What is an immediate consequence of trying to estimate the matrix when $q > 1$, i.e. when there are more variables than samples in time?
3. In the context of random matrices, the natural extension of the expected value is the normalized trace operator,

$$\tau(\mathbf{A}) = \frac{1}{N}\langle \mathrm{Tr}\,\mathbf{A}\rangle\,, \tag{9.28}$$

such that the k^{th} moment of $\mathbf{A}$ is $\tau(\mathbf{A}^k)$. To convince ourselves that the empirical covariance matrix will be distorted by noise whenever $q > 0$, calculate the first two moments of $\mathbf{E}$, assuming that the data is uncorrelated in time i.e. $\langle X_i^t X_j^s\rangle = C_{ij}\delta_{t,s}$. Hint: use Wick's theorem,

$$\langle X_1 X_2 \ldots X_{2n}\rangle = \sum_{\text{pairings}}\prod_{\text{pairs}}\langle X_i X_j\rangle, \tag{9.29}$$

where we sum over all distinct pairings of $\{X_1,\ldots,X_{2n}\}$ and each summand is the product of the n pairs.

4. A common approach to compute the eigenvalue distribution of a random matrix is to go through the Stieltjes transform,

$$g_{\mathbf{A}}(z) = \tau\Big([z\mathbf{1}-\mathbf{A}]^{-1}\Big), \tag{9.30}$$

from which the statistics of the spectrum can be recovered as

$$\rho(x) = \frac{1}{\pi}\lim_{\eta\to 0^+}\mathrm{Im}\, g_{\mathbf{A}}(x - i\eta), \tag{9.31}$$

where $\rho(x)$ is the probability density function of the eigenvalues when $N\to\infty$. In the simplest possible case $\mathbf{C} = \mathbf{1}$, $\mathbf{E}$ is referred to as the Wishart matrix $\mathbf{W}$, the Stieltjes transform of which can be shown to be given by

$$g_{\mathbf{W}}(z) = \frac{z+q-1 \pm \sqrt{(z+q-1)^2 - 4qz}}{2qz}. \tag{9.32}$$

Show that the zeros of the argument under the square root are given by $\lambda_\pm = (1\pm\sqrt{q})^2$, and subsequently choose the correct branch of $g_{\mathbf{W}}(z)$ in order to satisfy $g_{\mathbf{W}}(z) \to \frac{1}{z}$ as $z\to\infty$.

5. Given that only the square root has a contribution in the imaginary part of the Stieltjes transform when $\eta \to 0^+$, express the eigenvalue density as a function of $\lambda_\pm$. This density is known as the Marčenko-Pastur law. For what values of q is it correctly normalized?
6. Run numerical simulations to test the eigenvalue density that you have found for $q < 1$ and $N = \{20, 50, 100\}$. What do you observe when the density is no longer normalized analytically?
7. How do you think one could use the Marčenko-Pastur density to improve the measurement of empirical covariance matrices, for example on the stock market?

Tutorial 3: Stylized facts in financial time series

Introduction

We strongly believe that a good knowledge of empirical facts is paramount to, and should in fact precede, establishing good theory. In this session you will get acquainted with financial data from the real world and study the stylized facts of the log-returns of the S&P 500 Index downloaded from Yahoo! Finance. A critical discussion of models "à la Black-Scholes" will follow.

Part 1 : Download and understand the data

1. Use the code provided to download the S&P 500 data from Yahoo! Finance using the `pandas_datareader` module. The ticker `^GSPC` is the ticker symbol for the S&P 500, but the code you are given can be used to look up any ticker.
2. "High" and "Low" represent the highest and lowest prices of the stock during the day. Look up the meaning of "Adj Close" in Yahoo! Finance. Why should we use this instead of "Close"? From now on, p_t refers to this variable for the day t.
3. Plot the "Adj Close" variable. Comments?

Part 2 : Log-return statistics

1. Define a column `df[‘‘a_returns’’]` for the daily additive returns, defined as $d_t^1 = p_t - p_{t-1}$, and do the same for a column `df[‘‘l_returns_1’’]` containing the log returns, defined as $r_t^1 = \log\left(\frac{p_t}{p_{t-1}}\right)$. Plot both returns through time. What do you notice? Which variable makes more sense to you and why?
2. Write a function to compute the variogram of the log-price, $V(\tau) = \left\langle \left(\log \frac{p_t}{p_{t-\tau}}\right)^2 \right\rangle$. Plot $V(\tau)/V(1)$ for $0 \leq \tau < 500$ in log-log scale. What slope do you see?
3. Define the return at scale Δ (in days) as $r_t^\Delta = \log\left(\frac{p_t}{p_{t-\Delta}}\right)$ (for simplicity, $r_t^1 := r_t$). Noting that $r_t^\Delta = \sum_{i=1}^{\Delta} r_{t-i}$, what simple model would explain the previous plot?
4. Write a function that computes the series for r_t^Δ .Compute the mean and std. of r_t^1, as well as that of $r_t^{250} = \log\left(\frac{p_t}{p_{t-250}}\right)$, the yearly returns. Comments?
5. In mathematical finance, people often resort to the following model, "à la Black-Scholes " to study price dynamics:
$$p_t = (1+\mu)p_{t-1} + \sigma \eta_t p_{t-1} \tag{9.33}$$
where η_t is a gaussian random variable with $\langle \eta_t \eta_{t'} \rangle = \delta(t-t')$ and $\langle \eta_t \rangle = 0$ (white noise), and with $\sigma, \mu \ll 1$. Interpret the terms μ and σ. How does one write r_t within this framework?
6. Use `np.random.randn` to draw a random gaussian numbers $\tilde{r}_t$ of the same length as r_t and multiply them by your estimation of σ to have the correct variance. Plot $\tilde{r}_t$. Comments?
7. We define the survival function (recall PC2) of the returns as
$$F(x) := \int_x^\infty \mathrm{d}r\, \rho(r) \tag{9.34}$$
where ρ is the density function associated to the returns. Plot the survival functions of the daily returns r_t on the right (use r_t on positive-valued bins) and left (-r_t on positive-valued bins) tails. Hint: look up how to do this on the **Numpy Cheatsheet**.
8. After importing the correct module with `import scipy.stats`, use the functions
```
y_normal = scipy.stats.norm.sf(x = x, scale = sigma)
y_student = scipy.stats.t.sf(x = x, scale = sigma, df = nu)
```
to compute the survival functions of a normal distribution of std. σ and of a Student-t distribution of std. σ and tail parameter ν at x (can be a `numpy` array). Estimate σ from data, but play with ν. Which parameter fits best the right and left tails (try $\nu \in [2,6]$)?

9. Do the same as the two previous functions, but for $\Delta \in \{30, 60, 90\}$. Qualitatively, what is happening? (Remember to adjust with the corresponding value of σ_Δ).
10. What can you comment about the model proposed in Eq. (9.33)?

Part 3 : Correlations

1. Before beginning, center the daily returns by removing the mean, i.e. $r_t \leftarrow r_t - \langle r_t \rangle$. This is standard when working with correlations.
2. Write a function to compute the correlation function

$$C_{r,r}(\tau) = \frac{\langle r_t r_{t+\tau} \rangle}{\sqrt{\langle r_t^2 \rangle}} \tag{9.35}$$

Compute it for $-200 \leq \tau < 200$. What do you expect to see? Same question but for

$$C_{|r|,|r|}(\tau) = \frac{\langle |r_t| \, |r_{t+\tau}| \rangle}{\sqrt{\langle r_t^2 \rangle}} \qquad \text{and} \qquad C_{r^2,r^2}(\tau) = \frac{\left\langle r_t^2 r_{t+\tau}^2 \right\rangle}{\sqrt{\langle r_t^4 \rangle}}.$$

3. Compute and plot $C_{|r|,|r|}(\tau)$ and $C_{r^2,r^2}(\tau)$ for $1 \leq \tau < 1000$. Choose a log-log scale when plotting. Comments?
4. What does this imply for the proper modeling of return dynamics?

Part 4 : Volatility and conclusion

1. Define now the volatility as $\sigma_t^2 = r_t^2$. Use a qualitative argument to predict the behaviour of $C_{\sigma^2,r} = \left\langle r_t \sigma_{t+\tau}^2 \right\rangle$ (use your intuition), then compute it and see for yourself by choosing $-100 \leq \tau < 200$. How can you interpret this? This is called the leverage effect.
2. In light of what you have seen, what arguments can you find against the modelling proposed in Eq. (9.33)?
3. You can re-execute all the cells by picking a different ticker. Try for instance the Euro Stoxx, with the ticker `^STOXX50E` or the CAC40 with ticker `^FCHI`. What do you notice?

Tutorial 4: Mesoscopic models in finance

Introduction

In this exercise session, we will first implement a propagator model with real trade data on the E-mini S&P 500 futures contract. We will then study a simple model for the formation and collapse of financial bubbles.

The propagator model

Part 1 : Reminders on the model

We recall the propagator model, where the return dynamics are given by

$$r_t = \sum_{t' \le t} \mathcal{G}(t-t')\varepsilon_{t'} + \eta_t, \tag{9.36}$$

where η_t is a white noise of 0 mean and variance $\langle \eta_t \eta_{t'} \rangle = 2\sigma_0^2 \delta(t-t')$ and where $\mathcal{G}(t) = G(t+1) - G(t)$ is the discrete derivative of the propagator. In this case, the mid price $m_t = m_{t-1} + r_{t-1}$ reads

$$m_t = m_0 + \sum_{t'<t} G(t-t')\varepsilon_{t'} + \sum_{t'<t} \eta_{t'}. \tag{9.37}$$

1. We are interested first in the response function of the returns to sign fluctuations, namely

   $$S(\ell) := \langle r_{t+\ell}\varepsilon_t \rangle, \tag{9.38}$$

   show that it verifies

   $$S(\ell) = \sum_{n \ge 0} \mathcal{G}(n) C(n-\ell), \tag{9.39}$$

   where C is the sign correlation function,

   $$C(\ell) = \langle \varepsilon_t \varepsilon_{t+\ell} \rangle. \tag{9.40}$$

   (Note: the average must be understood to run over t).
2. Another correlation function with a "physical meaning" is the function R defined by

   $$R(\ell) := \langle (m_{t+\ell} - m_t) \cdot \varepsilon_t \rangle. \tag{9.41}$$

   Show that

   $$R(\ell) = \sum_{0 \le i < \ell} S(i). \tag{9.42}$$

Part 2 : Empirical mid prices, signs and returns

The tedious process of cleaning the data has been done for you, although if you are curious there is a notebook available detailing the steps for the E-mini S&P 500 futures contracts. In this exercise, we will extract useful information from the cleaned data.

1. Load the file `spmini_trades_cleaned.pkl` with `pd.read_pickle` and inspect how the data is organized using the `df.head()` command. What time period is covered? What information is contained in the file?
2. Define the sign of a transaction: create a column `df['sign']` equal to $\varepsilon_t = +1$ for an ask order, and to $\varepsilon_t = -1$ for a bid order.
3. Compute the previously defined sign correlation function, where the average is first computed for each day (use `df.groupby(df.index.day).apply(function)` for a suitably defined `function`) in Jan. 2018 for $0 \le l < 2000$, obtaining thus an array of size 2000 per day, and then averaged over all days to obtain a single array. Plot it in log-log scale. What do you notice?

4. Next, compute the price variogram

$$V(\tau) = \left\langle (m_{t+\tau} - m_t)^2 \right\rangle_t, \tag{9.43}$$

again computing first for each day and then averaging over all days in January. Compute for $0 \leq \tau < 2000$ and plot in log-log scale, then plot the volatility signature plot $V(\tau)/\tau$.

5. How can you interpret previous figures?

Part 3 : Calibrating the propagator

Having looked at the data and extracted the correlations of interest, we can finally empirically calibrate the propagator model.

1. Compute the response function of returns to sign fluctuations S for $0 \leq \ell < 2000$, first over each day and then averaging over the whole month. Plot it with a log-log scale.
2. Compute $R(\ell)$, first using its definition from S and then directly from the correlations for $0 \leq \ell < 5000$.
3. Solve for the discrete derivative $\mathcal{G}$ using eq. (9.39). (Hint: define a matrix $A_{\ell n} = C(\ell - n)$ minding the fact that C is symmetric). Compute G from $\mathcal{G}$ and plot it. Comments?

A simple model for bubble formation

Part 1 : Reminder on Langevin dynamics

The aim of this exercise is to provide some intuition for overdamped Langevin dynamics, defined as

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -U'(x(t)) + \sqrt{2T}\xi(t), \tag{9.44}$$

where T corresponds to the temperature in physics, associated to the thermal fluctuations ξ that are modelled as a Gaussian white noise, and U is the potential (recall PC1 and the Ornstein-Uhlenbeck process).

1. Consider the potentials $U_1(x) = -x^2$ and $U_2(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. Discuss what you think will happen as a function of temperature in both cases.
2. Now consider the general cubic potential

$$U(x) = \frac{a}{3}x^3 + \frac{b}{2}x^2 + cx.$$

Discuss the influence of the parameters on the number and position of the extrema of U.

3. In field theory, the simplest model to study critical phenomena for is to consider the "mexican hat" or φ^4 free energy functional at a relative "distance" t from the critical temperature T_c,

$$\mathscr{F}[\varphi(\mathbf{x})] = \int_\Omega \left(\frac{t}{2}\varphi^2(\mathbf{x}) + \varphi^4(\mathbf{x}) \right) \mathrm{d}\mathbf{x}, \qquad t = \frac{T - T_c}{T_c} \tag{9.45}$$

that is quartic and not cubic. Do you have any idea why?

Part 2 : A Langevin pricing model for bubbles

Having gained some intuition on how the shape of the potential affects the stability of the steady state Langevin dynamics, we can now delve into the relevance of the model in finance.

The central assumption is that at time t the return r_t is directly proportional to the demand-supply imbalance ϕ_t

$$r_t = \frac{\phi_t}{\lambda}$$

where λ represents the "market depth" (the larger the market, the smaller the effect of inbalance on the return). We can then write a generic model for the dynamics of the demand-supply imbalance:

$$\phi_{t+1} - \phi_t = ar_t - b{r_t}^2 - a' r_t - k(p_t - p_F) + \chi \xi_t \tag{9.46}$$

with ξ_t a Gaussian white noise (zero mean, unit variance), and p_F the fundamental value assuming it exists (or the market's belief of the fundamental value if it doesn't).

1. Discuss the mechanisms that the different terms in equation (9.46) are attempting to model.

   We now take the continuous limit such as to approximate the return as $r_t \simeq u = \partial_t p_t$.
2. Show (with a physicist's level of rigor), that equation (9.46) may be rewritten as

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -U'(u(t)) + \tilde{\xi}_t \tag{9.47}$$

   with the potential

$$U(u) = \kappa(p_t - p_F)u + \frac{\alpha}{2}u^2 + \frac{\beta}{3}u^3.$$

   What are the expressions of κ, α, β and $\tilde{\xi}_t$ as a function of the initial parameters?
3. Based on our previous experience with a cubic potential, how do you think the different parameters will affect the evolution of the price? **Notice that the price alters the potential as it evolves!**

Part 3 : Get your hands dirty

1. Start by considering the case where $\beta = 0$. What is the underlying assumption on the trader's reasoning? What is the requirement on α for this?
2. Notice that keeping $\beta = 0$, substituting $u = \partial_t p_t$ reads

$$\frac{\mathrm{d}^2 p_t}{\mathrm{d}t^2} = -\alpha \frac{\mathrm{d}p_t}{\mathrm{d}t} - \kappa(p_t - p_F) + \tilde{\xi}_t. \tag{9.48}$$

   If we are close to the fundamental value of the price, can you say anything on the fluctuations of the price 'velocity'? Think of the other SDEs we have seen so far!
3. Suppose now that $\beta > 0$, $\alpha > 0$ and $p = p_F$ $(t = 0)$. Locate the system's equilibria, study their stability and calculate the height of the potential barrier ΔU between the equilibria. How do you think this quantity impacts the system's evolution?

Part 4 : Simulating bubbles

We now wish to study the formation of bubbles and crises in this simple model, meaning we take $\beta > 0$, $\alpha < 0$. This is difficult analytically, so we will turn to numerics.

1. Take $\kappa = 1$, $\beta = 0.5$, $\alpha = -2$ and fix $p_F = 0$. Plot the potential for p varying from -1 to 1 and comment the results.
2. Write a function to simulate the system's evolution

$$u_{t+\delta t} = u_t - U'(u_t)\delta t + \chi \xi_t \sqrt{\delta t}$$

   with, at this stage, $p_t = p$ fixed and independent of u_t. Take $\chi = 1$ and try $p = \pm 1$, what difference do you observe?
3. Now introduce the feedback by updating p_t as

$$p_{t+\delta t} = p_t + r_t \delta t$$

   and give an intuitive explanation for the results you observe.

Tutorial 5: The Random Field Ising model

Introduction

In this exercise session, we will consider a model from physics that has gained quite a lot of traction in the modelling of socioeconomic systems: the Random Field Ising Model (RFIM). Originally devised to understand the magnetization of imperfect metallic alloys under the presence of external fields, we will see that the RFIM displays many properties that may be very interesting to model collective human behavior.

Part 1 : The model

In the socioeconomic context, the RFIM describes N agents, labeled by $i = 1, \dots, N$, making a binary decision $S_i = \pm 1$. While seemingly minimalistic, such a binary decision is actually interesting to model a wide variety of phenomena: bipartisan elections, to sell or buy a stock, to commit tax evasion or not...

We postulate that the choice of agent i will evolve as

$$S_i(t+1) = \mathrm{sign}\left(h_i + F(t) + \sum_{j(\neq i)} J_{ij} S_j(t) \right), \tag{9.49}$$

where h_i is an idiosyncratic bias proper to each agent. We suppose that the h_i are random iid variables with a density $\rho(h)$ centered at zero and of characteristic width σ.

1. Interpret the different terms in the equation.
2. From now on, we take the mean-field limit $J_{ij} = J_0/N \ \forall (i,j)$, $J_0 > 0$. What does Eq. (9.49) look like now? Rewrite it in terms of the average opinion $m(t) = \frac{1}{N}\sum_i S_i(t)$.
3. What are the values of the S_i and m variables in the limits $F = \pm\infty$? How about when $J_0 = F = 0$?
4. How do you think the average opinion evolves with $F(t)$ for $J_0/\sigma = 0$? Same question but now for J_0/σ positive but small, and finally for J_0/σ large.

Part 2 : Simulating the system

For this exercise, a Python *class* is provided for you to complete. If you are not familiar with object-oriented programming, you can look at the very simple `Rectangle` class included. We will consider a Gaussian external field,

$$\rho(h) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{h}{\sigma}\right)^2}. \tag{9.50}$$

1. Fill out everything in the class, except for the `equilibrate` method. You can also test things incrementally, but you are mostly independent here.
2. How can we equilibrate the system? An equilibrium is a configuration of the S_i variables such that none can be "flipped" according to the decision rule of Eq. (9.49).
3. Create an instance of the `RFIM` with the value $J_0 = 1$ and $\sigma = 0.1$. Start with $F = -10$ for the first one, and compute the equilibrium values $S_i(F = -10)$ and therefore the average opinion $m(F = -10) = \langle S_i \rangle$. Increase the value of F progressively up to $F = 10$ (you can use `np.linspace(-10,10,100)` to do this in 100 points), and store the values of m you obtained for each value of F. (It's important you do this for the **same instance of the class**, changing F and re-equilibrating successively).
4. Do the same but going from $F = 10$ to $F = -10$. Plot the two curves you obtained for m. What do you notice? Repeat the experiment with $\sigma = 1$ and going from $F = -10$ to $F = 10$. Are there any differences? Finally try out the case $J_0 = 0$ for any σ of your choice.

Part 3 : Solving the model

1. Suppose first that $\rho(h) = \delta(h)$. Starting from $F = -\infty$ and increasing F to ∞, when do all opinions change?

2. Now take the generic case where $\rho(h)$ is a unimodal distribution with a finite characteristic width. We introduce ϕ the fraction of agents with opinion $+1$. Show that in $N \to \infty$ limit, we can rewrite the dynamics as

$$\phi(t+1) = P_>(-F(t) + J_0 - 2J_0\phi(t)), \tag{9.51}$$

where $P_>$ is the survival function of h. Using this expression, write the equation for a fixed point to the dynamics in terms of the average opinion.
3. Writing m^* the average opinion at the fixed point, study its equation graphically to determine the number of solutions. (Hint: you may fix $F = 0$ and consider the case where h is Gaussian.) Can you now see why J_0/σ is the relevant parameter to characterize the system?
4. Show that the number of solution changes at the critical point reached for

$$2J_0\rho(-F^* - J_0 m^*) = 1. \tag{9.52}$$

(Hint: notice $P'_>(x) = -\rho(x)$.)
5. A different way to see this result is to consider the system through the contagion – or avalanche – of opinion changes. Suppose agent i changes its opinion. What is the condition for this opinion change to lead agent j to also change their mind?
6. Using this result, show that the probability that *any* agent changes opinion following i is given by

$$\mathbb{P}(\text{any flip}) = 2J_0\rho(-F - J_0 m). \tag{9.53}$$

Can you see why this recovers the condition for the critical point?

Part 4 : Criticality and avalanche sizes

We now wish to study the distribution of avalanche size in the system. In the following, an avalanche of size s will refer to an opinion change of $s-1$ agents following on opinion flip.

1. We have seen that the probability that an agent flips following the flip of another agent is given by

$$\mathbb{P}(1 \text{ flip}) = \mathbb{P}\left(h \in \left[-F - J_0 m - \frac{2J_0}{N}, -F - J_0 m\right]\right). \tag{9.54}$$

Give (in words) the expression for the probability of an avalanche of total size s.
2. Show that the avalanche size probability density is given by

$$P(s) = \frac{\mathrm{e}^{-\lambda(s)}\lambda(s)^{s-1}}{(s-1)!}, \tag{9.55}$$

with $\lambda(s) = 2J_0 s\rho(-F - J_0 m)$.
3. We admit that among all these configurations, there is only a fraction c/s that in fact lead to a single avalanche. Using the Stirling approximation $n! \approx \sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n$, show that in the limit $s \to \infty$

$$P(s) \sim s^{-3/2}\mathrm{e}^{s(1-\tilde{\lambda})+s\log\tilde{\lambda}}, \tag{9.56}$$

where we have taken the rescaling $\lambda = s\tilde{\lambda}$.
4. Comment on what happens at the critical point in light of we have seen on power laws. How does the typical avalanche size diverge? What about the mean avalanche size? (Hint: the critical point corresponds to $\tilde{\lambda}_c = 1$ from the previous exercise.)

Tutorial 6: Herd behavior and aggregate fluctuations

Introduction

In 1998, Rama Cont and Jean-Philippe Bouchaud proposed a simple model of returns that could account for the excess kurtosis of their distribution. As we empirically observe, returns (whether absolute or relative) are heavy-tailed and seem to be properly fitted by a student law. Furthermore, they display a shear asymmetry - evidence of the difference of perception of positive or negative returns - along with non-trivial correlation structures. The model proposed here will only account for the heavy-tailed nature of returns (but we can imagine way to introduce skewness and long-range correlations) by introducing herding with a network structure.

The model

We consider N agents trading a single asset. Each agent i has a trading intention $\phi_i(t) \in \{-1, 0, 1\}$, where -1 represents a sell intention, 1 a buy and 0 an agent doing nothing. At each time t, these intentions are chosen randomly with probability a for ± 1 and $1-2a$ for 0. The underlying process of why an agent choses either one of the intentions is not specified.

Furthermore, we assume that agents display herding behavior: (1) they cluster with each other and share a common trading intention within one cluster (2) clusters are independent. The clustering procedure is as follows. Each agent may form a link with the $N-1$ others. A link is formed with probability p independent of the link. The average number of links around one agent is $p(N-1)$. For this number to remain finite in the limit $N \gg 1$, we choose $p = c/N$. Such procedure for constructing the graph has first been brought up by Erdo s and Renyi. Interestingly enough, these graphs display a phase transition for the size of their largest connected component S. If $c > 1$ then $f := S/N > 0$, otherwise $f \to 0$. On can show that the size W of one randomly chosen connected component of the graph is such that

$$\mathbb{P}(W = w) \underset{w\to\infty}{\sim} \frac{A}{|w|^{5/2}} \exp\left(-(c-1)\frac{w}{w_0}\right), \tag{9.57}$$

which becomes a pure power-law for $c = 1$.

Finally, as a first approximation, we assume that the price variation associated with the trading intentions $\{\phi_\alpha\}$ depends linearly on the intentions. As a result, we model the price change of the asset as follows

$$\Delta p(t) = p(t+1) - p(t) = \frac{1}{\lambda}\sum_{\alpha=1}^{n_c} W_\alpha \phi_\alpha(t) := \frac{1}{\lambda}\sum_{\alpha=1}^{n_c} X_\alpha(t), \tag{9.58}$$

with n_c the (random) number of clusters, W_α the size of cluster α and ϕ_α the associated trading intention. The parameter λ is called *market depth* and measures how much trading intention is needed if one wants to move the price by one unit.

Part 1 : Analysis of the model

1. Argue as to why the X_α's are independent identically distributed random variables. Generically, how is the distribution of the sum N i.i.d. random variables related to the distribution of the a single draw?
2. Denoting by $F(x) = \mathbb{P}(X_\alpha \leq x|\phi_\alpha \neq 0)$, show that

$$\mathbb{P}(X_\alpha \leq x) = (1-2a)\Theta(x) + 2aF(x),$$

   with $\Theta(\bullet)$ the step function.
3. We assume that F is associated to a density $f(x) := \mathbb{P}(X_\alpha = x|\phi_\alpha \neq 0)$. Explain why f is even and why it displays the same asymptotic behavior as (9.57) when $|x| \to \infty$.

4. Show that the probability for a return to be equal to ξ at time t is given by

$$\mathbb{P}(\Delta p = \xi) = \sum_{k=1}^{N} \mathbb{P}(n_c = k) \sum_{j=0}^{k} \binom{k}{j} (2a)^j (1-2a)^{k-j} f^{*j}(\lambda \xi)$$

with $(\bullet)^{*j}$ the j-fold convolution product.
5. We introduce the function $\tilde{f}(z) = \mathbb{E}[e^{X'_\alpha z}]$ and $\mathscr{F}(z) = \mathbb{E}[e^{\lambda \Delta p z}]$ (i.e the generating functions) where $X'_\alpha = \pm W_\alpha$. Show that

$$\mathscr{F}(z) = \sum_{k=1}^{N} \mathbb{P}(n_c = k)[1-2a+2a\tilde{f}(z)]^k = \exp \psi \left(\log \left(1-2a+2a\tilde{f}(z) \right) \right)$$

with ψ the cumulant generating function of n_c. We denote by $\gamma(z)$ the quantity $1-2a+2a\tilde{f}(z)$. (Hint: the cumulant generating function of a random variable X is given by $\log \mathbb{E}\left[e^{zX}\right]$.)
6. Using graph theory, one can show that

$$\psi(z) = Nz + \frac{Nc}{2}(e^{-z} - 1).$$

Deduce

$$\mathscr{F}(z) = \gamma^N \exp\left[\frac{Nc}{2} \left(\frac{1}{\gamma} - 1 \right) \right].$$

7. What is the average number n_o of trading agents in the market? How can it remain finite in the limit $N \to \infty$?
8. With this prescription, show that in the limit $N \to \infty$, one gets

$$\mathscr{F}(z) \approx \exp\left[n_o \left(1 - \frac{c}{2} \right) (\tilde{f}(z) - 1) \right].$$

9. We give the second moment $\mu_2(\alpha) = A(c)(1-c)^{-1}$ and the fourth moment $\mu_4(\alpha) = A(c)(1+2c)(1-c)^{-5}$ of W_α. Deduce the excess kurtosis of the distribution of the price variation $\lambda \Delta p$

$$\kappa(\lambda \Delta p) = \frac{1+2c}{n_o \left(1 - \frac{c}{2}\right) A(c)(1-c)^3}.$$

10. Comments?

Part 2 : Simulations

Using the `networkx` package (imported as `nx`), one can easily generate random Erdo s-Renyi graphs using

```
nx.fast_gnp_random_graph(N,p)
```

1. For $N = 10^4$, $n_o = 2000$, $\lambda = 10$, generate a series of returns on $T = 10^4$ and for $c = 1$.
2. Plot the histogram of positive and negative log-returns. Compare it to a Gaussian distribution. What is the exponent of the tail?

Part 3 : Critical discussion

What are the main features of financial returns that are not reproduced by this simple model? Can you propose some mechanisms to recover these features?

Tutorial 7: The ant recruitment model

Introduction

In a series of experiments carried on harvester's ants, entomologists Deneubourg and Pasteels stumbled upon a curious phenomenon. Provided with two *a priori* identical food sources, ants tend to prefer one over the other but sometimes switch without any apparent reason. Ants display an asymmetrical behavior in a symmetrical situation. In the realm of social sciences, we can relate this behavior to the phenomenon known as "herding". Herd instinct in finance is the phenomenon where investors follow what they perceive other investors are doing rather than their own analysis. Herd instinct has a history of starting large, unfounded market rallies and sell-offs that are often based on a lack of fundamental support to justify either. In his article *Ants, rationality and recruitment*, economist Alan Kirman proposes a simple model to explain ants' behaviors and by extension "herding".

The model

Consider N ants divided between identical and always-full food sources denoted by F_1 and F_2. We denote by $k(t)$ the number of ants in zone F_1 at time t. Between time t and time $t+1$, one ant is randomly drawn and can switch food source subjected to two influences

- spontaneous switch to the other food source with probability $\varepsilon \in]0,1]$. We exclude 0 to avoid all the ants being stuck in one food source.
- The first ant follows another randomly chosen ant with probability $\frac{\mu}{N} \in [0,1]$.

We denote by $\mathbb{P}(k,t)$ the probability to find k ants in F_1 at time t; and by $W(\ell \to k;t) = \mathbb{P}(k(t+1) = k|k(t) = \ell)$, the transition rate between the state of ℓ to k ants in F_1.

Part 1 : Discrete setup

1. Justify that the transition rates do not depend on the entire history of the system but only on the present time-step.
2. Let $k \in \{0,\dots,N\}$.
   (a) Show that, whenever $\ell \notin \{k-1,k,k+1\}$ then $W(\ell \to k) = W(k \to \ell) = 0$.
   (b) Show that the transition rates from k to $k+1$ and from k to $k-1$ read

$$W(k \to k+1) = \left(1 - \frac{k}{N}\right)\left(\varepsilon + \frac{\mu}{N}\frac{k}{N-1}\right) \tag{9.59}$$

$$W(k \to k-1) = \frac{k}{N}\left(\varepsilon + \frac{\mu}{N}\frac{N-k}{N-1}\right) \tag{9.60}$$

3. (a) Show that the probability to have k ants in F_1 at time t obeys the so-called Master Equation

$$\mathbb{P}(k,t+1) = \mathbb{P}(k-1,t)W(k-1 \to k) + \mathbb{P}(k+1,t)W(k+1 \to k) + \mathbb{P}(k,t)W(k \to k) \tag{9.61}$$

   (b) Denoting by $\mathbf{P}(t)$ the vector $\mathbf{P}(t)_k = \mathbb{P}(k,t)$, show that there exists a stochastic matrix $\mathbb{T}$ such that

$$\mathbf{P}(t+1) = \mathbb{T}\mathbf{P}(t)$$

   (c) Deduce the existence of a stationary probability measure $\mathbf{P}_s$ for the repartition of ants between the two food sources.

4. If you had to guess, how do you think the stationary distribution looks when switches or following dominate?
5. For the stationary probability measure $\mathbf{P}_s$, establish the *global balance* condition

$$\sum_{\ell \neq k} W(\ell \to k)\mathbb{P}_s(\ell) = \sum_{\ell \neq k} W(k \to \ell)\mathbb{P}_s(k) \tag{9.62}$$

and provide an interpretation. What further condition on the stationary distribution directly satisfies this global balance condition?

Part 2 : Continuous limit

In order to find the stationary probability measure, we will take a continuous time and "space" limit as follows. In the master equation, we replace $t+1$ by $t+\mathrm{d}t$ with $\mathrm{d}t = \frac{1}{N} \ll 1$. We also replace μ by μN^2, ε by εN, and $\mathbb{P}(k,t)$ by the probability density function $f(x,t)$ with $x = k/N$. Finally, we take the limit $N \to \infty$.

1. Show that, with these replacements and new scalings, the master equation takes the form of a so-called Fokker-Planck equation in the limit $N \to \infty$

$$\partial_t f = \partial_x \left[-\varepsilon(1-2x)f(x,t) + \mu\partial_x \left[x(1-x)f(x,t)\right]\right]$$

2. Deduce that the stationary density function $f_s(x)$ is given by

$$f_s(x) = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)}\left[x(1-x)\right]^{\alpha-1}, \tag{9.63}$$

with $\alpha = \varepsilon/\mu$ and $\Gamma(x) = \int_0^\infty \mathrm{d}t\, t^{x-1}e^{-x}$. We give the identity $B(\alpha,\beta) := \int_0^1 \mathrm{d}x\, x^{\beta-1}(1-x)^{\alpha-1} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
3. Sketch this density and interpret it for different values of α.

Part 3 : Simulating the dynamics

Using the correspondence between a Fokker-Planck equation and a Langevin equation, one can show that the Fokker-Planck equation obtained in the previous section is associated to a random rocess $x(t)$ following a Langevin equation

$$\dot{x} = \varepsilon(1-2x) + \sqrt{2\mu x(1-x)}\eta(t) \tag{9.64}$$

where η is a Gaussian white noise of unit variance. We can get the discretized version of this dynamics with Ito rules

$$x_{t+dt} - x_t = \varepsilon(1-2x_t)dt + \sqrt{2\mu x_t(1-x_t)}\sqrt{dt}\eta_t \tag{9.65}$$

with dt the step size. We will fix $\Delta t = 10^{-3}$.

1. Why do you think we choose to simulate the model in continuous time rather than with the discrete time transitions studied in the first part?
2. (a) Write a function that simulates the above dynamics for given values of ε and μ.
   (b) Plot the fraction of ants in F_1 for $T = 10^5$ iterations, $\varepsilon = 0.15$ and $\mu = 0.3$. What do you observe? What is the associated stationary density?
   (c) Same question for $\varepsilon = 0.002$ and $\mu = 0.01$.
3. (a) Plot the empirical stationary distributions for the following values:
   - $\varepsilon = 0.1$, $\mu = 0.01$,
   - $\varepsilon = 0.1$, $\mu = 1$,
   - $\varepsilon = 0.1$, $\mu = 0.3$.

   (b) Plot the associated theoretical stationary distribution.

Tutorial 8: The latent order book model

Introduction

For a long-time, people believed that the price impact of a meta-order of size Q was linear with Q. Many simple models (Kyle model or Santa Fe model for instance), would predict such a linear impact. However, empirical studies showed that this linear dependency was spurious and that a square-root impact was a better fit with reality. In this tutorial, we will model the latent order book with a so-called "reaction-diffusion" dynamics and retrieve the square root law.

The model

In this model, we assume that orders may be place for any value of the price therefore taking a continuous limit. We denote by $\rho_A(x,t)$ and $\rho_B(t)$ the latent volume densities in the Ask/Bid sides. These quantities evolves according to a set of rules

- Latent orders diffuse with diffusivity constant D;
- Latent orders are canceled with multiplicative rate ν;
- New intentions are deposited with additive rate λ;
- When a buy intention meets a sell intention they are instantaneously matched and are thus removed from the LOB. We implicitly assume that latent orders are revealed in the vicinity of the trade price p_t.
- The trade price p_t is conventionally defined through the equation $\rho_B(p_t,t) = \rho_A(p_t,t)$.

We finally define the reduced latent order book density by $\phi(x,t) = \rho_B(x,t) - \rho_A(x,t)$.

Part 1 : Diffusive orders in the absence of cancelations or depositions

In this first part, we provide a microscopic derivation for the diffusion part of the orders. We assume that the agent i placing its order revise its reservation price $p_{i,t}$ between time t and $t+\delta t$ according to

$$p_{i_t+\delta t} = p_{i,t} + f_t \beta_i + \eta_{i,t} \tag{9.66}$$

where both β_i and $\eta_{i,t}$ are random variable with densities P_β and P_η.

1. Interpret the different terms of the update. What can you say about the distributions of β and η?
2. Show that the density of latent orders $\rho(x,t)$ at price x at time t, follows the evolution

$$\rho(x,t+\delta t) = \int \mathrm{d}\beta\, \mathrm{d}\eta\, \mathrm{d}x'\, P_\beta(\beta) P_\eta(\eta) \rho(x',t) \delta(x - x' - \beta f_t - \eta),$$

   with $\delta(\bullet)$ the dirac distribution.
3. Deduce from the previous equation

$$\rho(x,t+\delta t) = \sum_{n=0}^{+\infty} \frac{(-1)^n}{n!} \partial_x^n \rho \sum_{k=0}^{n} \binom{n}{k} f_t^k \langle \beta^k \rangle \langle \eta^{n-k} \rangle.$$

   This equation is called Kramers-Moyal expansion.
4. We formally write $f_t = V_t \delta t$ and $\langle \eta^2 \rangle = 2D\delta t$, with η Gaussian. Show that in the limit $\delta t \to 0$, keeping leading order terms in the Kramers-Moyal equation yields

$$\partial_t \rho = -V_t \partial_x \rho + D \partial_x^2 \rho.$$

5. Show that, with the appropriate change of the variable x, we get the *diffusion equation* with diffusion constant D
$$\partial_t \rho = D \partial_x^2 \rho. \tag{9.67}$$
In the following, we will consider the case of *uninformed impact* by setting $V_t = 0$.
6. Show that
$$G(x,t) = \frac{1}{\sqrt{4\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right),$$
is a solution to the above equation. Hint: $\forall a > 0, \forall b \in \mathbb{C}, \int_{-\infty}^{+\infty} e^{-\frac{ax^2}{2}+bx} \mathrm{d}x = \sqrt{\frac{2\pi}{a}} e^{\frac{b^2}{2a}}$.

Part 2 : Cancelations, depositions and reactions

The previous simple microscopic model provides an explanation for the diffusive features of orders. We will add the possibility to place and cancel orders with rates λ and ν.

1. (a) Justify why the cancelation is modelled by a multiplicative action on $\rho_{B/A}$.
   (b) Deduce that (9.67) is modified as follows
$$\partial_t \rho_{B/A} = D \partial_x^2 \rho_{B/A} - \nu \rho_{B/A}. \tag{9.68}$$
   (c) Show that the solution to the previous equation is given by
$$G_\nu(x,t) = e^{-\nu t} G(x,t).$$
2. (a) Justify why the deposition rate can be modelled by an additive action. If the execution price is p_t, on average, are buy (resp. sell) orders place above or below p_t?
   (b) Deduce that (9.68) is modified as follows
$$\begin{aligned} \partial_t \rho_A &= D \partial_x^2 \rho_A - \nu \rho_A + \lambda \Theta(x - p_t) \\ \partial_t \rho_B &= D \partial_x^2 \rho_B - \nu \rho_B + \lambda \Theta(p_t - x), \end{aligned} \tag{9.69}$$
   with Θ the step function.
3. Deduce, from all the above discussion, the following equation of ϕ
$$\partial_t \phi = D \partial_x^2 \phi - \nu \phi + \lambda \mathrm{sign}(p_t - x). \tag{9.70}$$

Part 3 : Stationary order book and market-impact

In this section, we will briefly discuss the reaction-diffusion equation obtained above. We then take the *infinite memory limit* $\lambda, \nu \to 0$ while keeping the ratio $\mathscr{L} \sim \lambda \nu^{-1/2}$ constant.

1. Show that the stationary shape of the order book is given by
$$\phi^{st}(\xi) = -\frac{\lambda}{\nu} \mathrm{sign}(\xi) \left[1 - e^{-|\xi|/\xi_c}\right]$$
with $\xi = x - p_\infty$ and $\xi_c := \sqrt{D \nu^{-1}}$.
2. Draw the shape of the stationary order book and exhibit a quantity below which the order book can be considered linear. Take the infinite memory limit. What is the shape of $\phi^{st}(\xi)$?

   Since meta-order of size Q are too big to be executed at once, they are usually broken down and placed within a time-span of T with a rate m_t such that
$$Q = \int_0^T \mathrm{d}s \, m_s.$$

We can model this as an extra source term $m_t\delta(x-p_t)\cdot\mathbb{1}_{[0,T]}$ in the reaction-diffusion equation. Using results from partial differential equations, one can show that the solution of this equation in the infinite memory limit reads

$$\phi(x,t)=[G(x,t)\star(\phi_0(x)\delta(t)+m_t\delta(x-p_t)\cdot\mathbb{1}_{[0,T]})](x,t),$$

where $\star$ denotes the space-time convolution, $\phi_0(x)=\phi(x,0):=\phi^{st}(x)$ and $p_0=p_\infty=0$. Computing this convolution, one can show that

$$\phi(x,t)=-\mathscr{L}x+\int_0^{t\wedge T}\frac{\mathrm{d}s}{\sqrt{4\pi D(t-s)}}\,m_s\exp\left(-\frac{(x-p_s)^2}{4D(t-s)}\right),$$

with $t\wedge T=\min(t,T)$.

3. We assume that the execution rate is constant ($m_t:=m_0$) and we place ourselves at the end of the execution ($t=T$). Show that

$$p_T=\frac{m_0}{\mathscr{L}}\int_0^T\frac{\mathrm{d}s}{\sqrt{4\pi D(T-s)}}\exp\left(-\frac{(p_T-p_s)^2}{4D(T-s)}\right).$$

4. It is straightforward to check that $p_s=A\sqrt{Ds}$ is a solution provided A is solution to a particular algebraic equation.
   (a) Show that
   $$A=\frac{m_0}{\mathscr{L}D}\int_0^1\frac{\mathrm{d}u}{\sqrt{4\pi(1-u)}}\exp\left(-\frac{A^2(1-\sqrt{u})}{4(1+\sqrt{u})}\right).$$
   (b) In the limit $m_0\ll\mathscr{L}D$, derive $A\approx\frac{m_0}{\mathscr{L}D\sqrt{\pi}}$.
   (c) In the limit $m_0\gg\mathscr{L}D$, derive $A\propto\sqrt{\frac{m_0}{\mathscr{L}D}}$.
5. Recalling that p_T is the impact of the meta-order when assuming $p_0=0$, show that $I(Q)\propto\sqrt{Q}$.

Part 4 : An algorithm to simulate the Latent Order Book

In this last section, we suggest a way of simulating the latent order book introduced above. The model falls into the "reaction-diffusion" type models, where orders can be seen as particles diffusing on a lattice, and where opposite orders react and cancel each other immediately when they meet on the same lattice site. Simulating the dynamics then simply relies on our ability to simulate jump processes, that are ubiquitous in mathematics, physics, chemistry, biology, and quantitative finance. We propose a continuous-time and discrete-space implementation of the dynamics of orders. This avoids the consideration of time steps that must be chosen smaller than any other characteristic time appearing in the system.

1. Consider a Poisson process of rate λ.
   (a) Show that the inter-arrival time of events is exponentially distributed with rate λ.
   (b) From algorithmic theory, any computer can draw uniformly distributed random numbers on $[0,1]$. Assuming that you have such random number generator at your disposal, show that if you set
   $$Y_i=-\frac{1}{\lambda}\log(U_i), \tag{9.71}$$
   with $U_i\sim\mathrm{Unif}[0,1]$, then the Y_i are exponentially distributed with rate λ. Hint: use the cumulative distribution function.
2. Consider a particle on a lattice that jumps to the right with rate D_r and jumps to the left with rate D_ℓ. How do you select the next particle action and how do you find the time of the jump?

3. Consider N particles on the same site. For a single particle, the rate to jump to the right is D_r, to jump to the left is D_ℓ. Particles disappear with multiplicative rate ν. In addition, a new particle can be deposited on the site with rate λ. What is the total event rate on the site? How do you choose the upcoming event?
4. Consider a discrete LOB with a large number of sites S. Each site k gathers n_k orders placed at a given price (on the US market, intersite spacing is \$0.01). How do you determine the next event in the whole LOB?
5. Consider now two types of orders (or particles): buy limit orders (piling up at the left of the mid-price), and sell limit orders (piling up at the right the the mid-price). List all events that can change the mid-price.
6. Using the previous questions, suggest an algorithm that simulates the LOB dynamics.

Part 5 : Implementing the Latent Order Book

We assume we have implemented the LOB dynamics.

1. Given the shape of the order book predicted in Part 3, suggest reasonable values for the parameters λ, ν and D for an LOB of $S \simeq 100$ sites.
2. Plot the stationary profile of the simulated order book. Check that it matches the profile obtained in Part 3. Vary the different parameters and observe the change in the density profile.
3. We now place a meta-order of size Q and we assume that the order is split into individual orders executed with a constant rate m_0. By considering an "additional clock" for meta-orders, implement the action of sending buy (or sell) meta-orders in the LOB.
4. Plot the impact of the meta-order on the mid-price. Do you recover that $I(Q) \propto \sqrt{Q}$?
5. The prefactor of the "square-root law" depends on the execution rate m_0. Compute $I(Q)/\sqrt{Q}$ for different values of m_0. What do you observe?

Tutorial 9: Optimal portfolios

Introduction

Constructing portfolios with the smallest possible level of risk (or the highest returns) is at the heart of asset management. Since the celebrated analytical solution provided by Markowitz in 1952, portfolio optimization has been a very successful playing field for physicists. In this tutorial, we will first look at the Markowitz solution and the efficient frontier, before considering the *long-only* case and the parallels that can be drawn with spin-glasses.

Part 1 : The Markowitz optimal portfolio

Suppose an investor has access to N financial assets (think stocks, bonds etc.) and wishes to construct the portfolio with the smallest possible risk for a given expected return. We write $\mathbf{w}$ the portfolio, represented by a vector with entries w_i corresponding to the fraction of the investor's total wealth allocated to asset i. The portfolio is assumed to be fully invested, meaning $\sum_i w_i = 1$.

1. We assume that the assets are fully characterized by their expected returns μ_i and the return correlation matrix $\mathbf{C}$. What do you think of such an assumption?
2. In such a Gaussian case, it turns out that all measures of risk are equivalent to the total portfolio variance σ_p^2. Write this total portfolio variance as well as the total portfolio expected return.
3. Show that the portfolio with the smallest variance for a given expected return μ_p is given by

$$\mathbf{w}^* = \frac{1}{2}\mathbf{C}^{-1}(\lambda\boldsymbol{\mu}+\gamma\mathbf{1}), \tag{9.72}$$

   where λ and γ are Lagrange multipliers, who's values are fixed by solving the system

$$\frac{1}{2}\begin{bmatrix} \boldsymbol{\mu}^\top\mathbf{C}^{-1}\boldsymbol{\mu} & \boldsymbol{\mu}^\top\mathbf{C}^{-1}\mathbf{1} \\ \mathbf{1}^\top\mathbf{C}^{-1}\boldsymbol{\mu} & \mathbf{1}^\top\mathbf{C}^{-1}\mathbf{1} \end{bmatrix}\begin{bmatrix} \lambda \\ \gamma \end{bmatrix} = \begin{bmatrix} \mu_p \\ 1 \end{bmatrix}. \tag{9.73}$$

   In what case is this system not solvable?
4. Plotting the optimal solution in the (σ_p^2, μ_p) plane for a given $(\mathbf{C}, \boldsymbol{\mu})$, we find a line known as the *efficient frontier*. Propose a sketch of what it looks like without computing it exactly.
5. The optimal portfolio will most likely include some negative weights. What sort of positions are these negative weights associated to?

Part 2 : Random portfolios and the efficient frontier

We now move on to numerics to gain a better understanding of the efficient frontier.

1. Write a function `gen_efficient_frontier(C,mu,mu_p)` that, given a target portfolio return and couple $(\mathbf{C}, \boldsymbol{\mu})$ will output the minimal portfolio variance.
2. Write a function `gen_random_portfolios(C,mu,M)` that, given a couple $(\mathbf{C}, \boldsymbol{\mu})$, will generate M randomly generated portfolios (say with uniformly distributed weights) satisfying the fully invested constraint and will output their total expected returns and variance.
3. Using these two functions, plot all the random portfolios' expected returns and variance in the (σ_p^2, μ_p) plane, as well as the theoretical efficient frontier. You can take $N = 4$ assets, a randomly drawn μ and a covariance matrix of the form

$$\mathbf{C} = \mathbf{I}_N + \boldsymbol{\beta}\boldsymbol{\beta}^\top \tag{9.74}$$

   with $\beta_i \sim \mathcal{N}(1, \sigma^2)$, $\sigma \approx 0.15$. What do you observe?
4. The above covariance matrix is known is a the "one factor" or "single index" risk model. Can you provide an interpretation for this structure and the vector $\boldsymbol{\beta}$? Why are the β_i centered at 1?
5. Modify the function generating the random portfolios such that it only includes positive weights and repeat the draw. Comment.

Part 3 : Long-only portfolios and spin-glasses

We now consider the case where investors can no longer take *short* positions associated to negative weights. This constraint could be due to a variety of reasons, ranging from investment mandates (e.g. in ESG portfolios) to regulation. For simplicity, we will take $\boldsymbol{\mu} = \mathbf{1}$ and $\mu_p = 1$. In such a case, one can show that

$$w_i^* \propto \sum_j (\mathbf{C}^{-1})_{ij} \tag{9.75}$$

1. Now that riskier assets can no longer be balanced with both long and short positions, do you think the optimal portfolio should still have all nonzero weights?
2. We introduce a vector of "spins" $\theta_i = \{0, 1\}$, $i = 1, \ldots, N$, representing the exclusion or inclusion of an asset in the portfolio. A heuristic algorithm to construct a long-only optimal portfolio is as follows:
   1. Start from $\boldsymbol{\theta} = \mathbf{1}$,
   2. Compute the Markowitz optimal portfolio $\mathbf{w}^*$,
   3. Whenever $w_i^* < 0$, set $\theta_i = 0$,
   4. Go back to step 2, computing $\mathbf{w}^* \to \tilde{\mathbf{w}}^*$ with a *reduced* matrix $\tilde{\mathbf{C}}$ including only entries with $\theta_i = 1$,

   and iterate until $\tilde{w}_i^* > 0$ $\forall i$ with $\theta_i = 1$. Code this procedure in a function `gen_Markowitz_LO(C)`.
3. We introduce the average "magnetization"

$$m = \left\langle \frac{1}{N} \sum_i \theta_i \right\rangle. \tag{9.76}$$

   Provide an interpretation for this quantity in the portfolio context. What should we take the average on?
4. Going back to the single-index risk model, compute the average magnetization of the long-only Markowitz portfolios for fixed values of σ (for example $\sigma = \{0.01, 0.05, 0.1\}$) and N ranging from 100 to 10^4. Plot m as a function of N. How about as a function of σN? Provide an interpretation for the effect of σ.
5. What do you think is the consequence for an investor of having a very sparse portfolio? Keep in mind that in reality the vector $\boldsymbol{\beta}$ will vary in time!

Part 4 : An analytical description

We now provide a coarse analytical description for the behavior observed numerically.

1. Within the single-index risk model, show that the Markowitz optimal portfolio has entries scaling as

$$w_i^* \propto 1 - \beta_i \frac{\sum_j \beta_j}{1 + \sum_j \beta_j^2}. \tag{9.77}$$

   (Hint: use the Sherman-Morrison formula $(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}$.)
2. Under the long-only constraint, this formula clearly shows that larger β_i will be removed first, consistent with the intuition that riskier assets must now be avoided. Show that we can therefore establish a threshold β^+, corresponding to the largest allowable risk in the portfolio, given by

$$\beta^+ = \frac{\sum_j \beta_j^2 \theta_j + 1}{\sum_j \beta_j \theta_j}. \tag{9.78}$$

   How can you relate the magnetization m to the mean threshold when $N \to \infty$?

3. Using a (loose) CLT argument, show that the mean threshold should satisfy

$$\beta^+ = \frac{\langle \beta^2 \rangle_c}{\langle \beta \rangle_c} + \frac{1}{N}\frac{1}{\langle \beta \rangle_c}, \tag{9.79}$$

with the notation $\langle g(\beta) \rangle_c = \int_{-\infty}^{\beta^+} \mathrm{d}\beta\, g(\beta)\rho(\beta)$. Can you guess why the scaling in σN found earlier requires us to keep the $\mathcal{O}(1/N)$ term? What next steps would you take to try to go further?