Lecture Notes on Topics in Derivatives Pricing, Imperial

Vladimir V. Piterbarg

February 24, 2025

Contents

Topic 1

Overview and Introduction to Markets

1.1 Course Overview

The main goal of this module is to demonstrate, and teach, how sophisticated mathematical models are applied in practice by various participants in derivatives markets. We explore how the Quantitative Finance theory is translated into practical applications of making money by different market participants, such as banks' trading desks/dealers, hedge funds, active investment managers, and others. We review various tools used by different players and gain familiarity with the basics of real-world hedging, relative value strategies, positioning investment portfolios for specific views, and (quasi) arbitrage strategies.

We will strive to maintain a fair balance between the mathematical framework and specific tools (stochastic analysis), the numerical aspects (actual implementation of the models, optimization routines), the data (calibration on real data, backward testing of hedging strategies), and practical applications.

We will

- Understand different derivatives market participants – trading desks, hedge funds, investment managers – and the role of, and the requirements, on quantitative models in their activities
- Outline the precise interaction between pricing and hedging of options, and apply these ideas to volatility arbitrage, including realized vs. implied volatility strategies, vol carry and gamma scalping
- Link derivatives theory to the practice of structuring, pricing and risk managing derivatives portfolios and generating above-market risk-adjusted returns (”alpha”)
- Give mathematical definitions, build intuition, and look at practical examples of such important, for active investors, concepts as carry, term premia and convexity
- Understand the types of systematic strategies based on the notions of relative value trading, rich/cheap analysis, momentum, mean reversion, etc.
- Systematically examine the central role that the differences between the real-world, the risk-neutral, and the subjective measures (“P vs. Q vs. S”) play in designing trading strategies and risk premia extraction
- Understand the types of true and quasi-arbitrage that arise from static and semi-static approaches to hedging and pricing options in a model-independent framework

After a brief review of practical aspects of using Black-Scholes (how do we make money in options?) we will mostly focus on Fixed Income markets as they are the largest, and arguably the richest in terms of macro/RV strategies available, markets in the world. We will look at

- What drives interest rate markets and how various players take advantage of that
- The intricacies of building interest rate curves and how various choices affect the ability of dealers to make, or not lose, money
- Trading curve shapes
- Common Relative Value (RV) and systematic strategies in linear rates markets
- Static arbitrage with (interest rate) options

- Bring to life the theoretical concept of measure change and what it means in terms of trading convexity
- Vanilla interest rate options and various quasi-arbitrage strategies
- Positioning for macro views using options

Depending on the pace we may not be able to cover it all!

1.2 Market Participants

A vast market in derivatives products linked to all asset classes exist: equity, foreign exchange, commodities, interest rates/fixed income, credit, etc. Basic instruments include "linear" products such as futures and forward, "non-linear" products such as European calls and puts, and various gradations of ever more exotic options, all the way to derivatives linked to multiple asset classes at the same time (hybrids).

The derivatives market can be broadly split into two (overlapping) segments: the exchange market and the over-the-counter, or OTC, market. Securities exchanges are trading platforms facilitating trading between different types of participants such as market makers, hedgers and speculators; see Hull [2006] for details on all. The OTC market, especially in fixed income markets, is vastly larger. In OTC markets a critical role is played by dealers, which are typically trading desks of investment banks.

1.2.1 Dealers

Dealers are known as the "sell side". Their primary role is to supply liquidity to other participants in the markets. You can think of them as supermarkets of derivatives, where buyers come to buy specific products. Less charitably they can also be thought as the "casino" that facilitates gambling by clients. Hence

- Dealers generally do not take directional views, i.e.
- They always try to offset exposures obtained from clients either with other clients or other dealers
    - Basic example. When a hedge fund, say, buys a call option on, say, an equity index, they are taking a "long" position, i.e. that the equity index will go up in price. Correspondingly, the dealer is left with a "short" position – they will lose money if the equity index goes up on this position. A dealer is typically not supposed to take directional positions, hence she will hedge the exposure, in this case possibly buying the underlying index in certain proportion (delta) that would offset the losses (and gains) from the client position. This hedging trade could be done on an exchange, or with other dealer, or there may be an offset with another position against another client

- Dealers (are supposed to) make money from charging fees for facilitating trading, not betting on the underlying markets
- Products sold by dealers often look like insurance vs. adverse (to clients) market moves
  - Unlike insurance, dealers generally cannot reply on a portfolio effect (like, say, car insurers do)
- Sophisticated hedging is a must, and building tools for hedging is the primary job of quants in banks
- Regulation and capital requirements are the main mechanisms limiting directional view taking by the dealers

While dealers are called "sell side", most of the other market participants are called "buy side" as they are the "buyers" of derivatives for their own purposes, be that hedging their economic activity, support their investment objectives, or just make money through arbitrage, relative value strategies or taking directional views on the markets.

1.2.2 Corporates

This is a generic name for companies whose primary activity is not linked to derivatives markets, but whose operational results are affected by the financial markets. A classic example is a company that makes cars in the US (hence spends US dollars to make them) but sells in Europe (hence makes profits denominated in Euro). Fluctuations in the EURUSD exchange rate will affect its profits. Such a company may, and often will, enter into transactions that would either lock the exchange rate in the future (forwards) or protect it from the downside while allowing for the upside (options).

Hedging strategies by corporates can become a lot more sophisticated than simple forwards and options as they often are incentivized to find a "cheaper" hedge (e.g. adding knock-out barriers to an option), or that one specifically tailored to its particular circumstances (e.g. an airline may decide to hedge jet fuel price increases and FX rate movements not separately but in one package).

Dealers spend considerable effort designing custom-made hedging instruments for corporates and educating them on ever more sophisticated (and more profitable to dealers) instruments.

1.2.3 "Real Money"

"Real money" is a commonly used term in the financial markets to denote a fully funded, long-only traditional asset manager. Real money managers are often referred to as institutional investors. The term real money means the money is managed on an unlevered basis (i.e. investing "real", fully funded, not borrowed, money). Typically these are mutual funds managing investor money on their behalf, pension funds, insurance companies, and the like. While their primary activity often takes place in the underlying markets (e.g. they buy equities or bonds), they have risk-management needs that sometimes require trading in derivatives.

"Real money" is a significant force in the markets (e.g. BlackRock has $> 10^{12}$ USD) and understanding their positioning/views on the economy is a key part of dealers' market making activities as well as hedge fund macro views.

1.2.4 Issuers and Retail Investors

Retail investors can invest into equities and bonds via mutual funds. They can also buy structured notes with payouts linked to performance of equity indices or interest rates or sometimes FX, that provide enhanced returns in certain future scenarios (but, inevitably, less-than-market return in some other scenarios). Structured notes are a mechanism for investors to gain exposure to future market states that they deem more likely. Structured notes are in their essence options, often quite complicated, on the underlying risk factors, with liability to investors limited to their initial cash outlay to buy the note.

Structured notes are often issues by private banks or specialized issuers, who do not take the other side of the bet but usually pass it to the dealers.

We will not be spending any time on structured notes/exotic derivatives in this course so this is just for completeness.

1.2.5 "Fast Money" aka Hedge Funds

Hedge funds are sophisticated investment companies that use investors money, often highly leveraged, with a mandate to deliver above-market returns by engaging in pretty much any type of trading activity, from equities and bonds to exotic derivatives. There are many types of hedge funds. Our main interest in this course are hedge funds that engage in relative value (RV) strategies, arbitrage, and macro trading in Fixed Income markets.

- RV: seeking to exploit temporary differences in prices of related securities. For example, in interest rate markets that could be two slightly different bonds whose difference in market prices is perceived to be bigger than justified by the differences in the bonds. RV strategies often deliver small steady returns until the temporal difference is gone, but are subject to potential significant losses if there is an underlying under-appreciated factor that drives the difference even further (e.g. for bonds could be liquidity or central bank targeted intervention). Quantitative tools are essential for finding RV opportunities

- Arbitrage, true (rare) or quasi (more common) – trading strategies that allow you to lock in a certain (true) or highly-likely (quasi) profit with no downside risk by trading in combinations of instruments. These often require sophisticated quantitative tools

- Macro: Positioning the portfolio to take advantage of potential future macro-economic changes, such as high inflation or interest rate hikes. Requires managers to have a strong view on the future economy that is sufficiently different from what is priced in the markets (market consensus). May deliver significant returns but often require a long-term view. May also lead to significant losses. Main returns here come from the skills of the HF manager and the strength of their economic research team. However, quantitative tools are used to find the right (most beneficial, cheapest, most certain, least downside, etc.) expression of the macro view, as well as to "size the bet" i.e. figure out what is the optimal amount of capital that should be committed to each particular opportunity. Often RV and macro are combined where RV strategies are used to "build a war chest" which is then employed for setting macro positions of strong conviction.

1.3 Foundations of Fixed Income Markets

At the most fundamental level, interest rates determine the economic cost of borrowing and lending, and as such define present values of future cash flows. In general, cash flows occurring at different times are discounted at different rates, reflecting market fluctuations in demand for money and risk preferences of market participants. The dependence of interest rates on time is described by the so-called term structure of interest rates, easily visualized as a curve that assigns a particular interest rate (or, equivalently, a discount factor) to each future date.

Interest rates change day-to-day in response to changing macroeconomic and market conditions. With the cost of borrowing and lending money affecting all aspects of the economy, it is no surprise that a vast market in derivatives on interest rates has developed. Motivations of participants are diverse, ranging from locking in the cost of financing to pure speculation.

1.3.1 Market Participants

Fixed income market participant taxonomy essentially mimics that of the derivatives markets overall, with some notable differences

- The role of dealers ("sell" side) is the same as in other markets.
- For corporates, arguably, the interest rate market is the most important one as most corporates have financing needs and issue debt. They typically issue fixed-rate bonds but often want to pay a floating rate, creating ongoing demand for fixed-floating swaps. As an aside, they sometimes issue bonds in one currency (where issuance conditions are the most favourable, for example) but want to have exposure in a different (say, their "home") currency, creating demand for cross-currency basis swaps. But we will not look into cross-currency swaps in any detail as it is a whole other topic.

- “Real money” investors are as active, if not more active, in debt instruments than in others.
- Notes issued for retail investors more often then not have an interest rate component, either explicitly linking to future interest rates, or just by the nature of then being long-dated enough so that discounting is an important component of the pricing of notes, hence requiring interest rate management and hedging.
- For hedge funds, macro strategies invariably involve an interest rate component.

There is also a number of fixed-income-specific market participants, such as mortgage originators and servicers, governments and government agencies, municipalities that issue their own debt, and the like. We will generally not focus too much on these types of entities.

It is probably also worth mentioning the role of intermediaries. Exchanges of course are the venues where price discovery happens for exchange-traded instruments. OTC markets for the same largely rely on

- Inter-dealer brokers

Inter-dealer brokers (ICAP, BGC, Tullett Prebon) connect dealers and arrange transactions between them in OTC derivatives such as swaps. This happens both on electronic platforms that are gaining more and more market share, but also by voice, especially for large transactions.

Brokers see a lot of activity in the OTC markets, more than any one individual dealer. They continuously stream current market prices of benchmark, i.e. standard, instruments such as 2y, 5y, 10y, 30y swap rates. These are often considered the best available and most accurate estimates of benchmark market prices, and most active market participants use these feeds as primary inputs into their trading systems and interest rate curves.

Among all the market participants there is one category that stands out. Arguably, the most important fixed income “market participants” are central banks. While their objective is not to make money in the markets, their influence on the markets is of utmost importance.

1.3.2 The Role of Central Banks

Central banks of many countries (including all major economies) play a crucial role in ensuring economic and financial stability. They implement monetary policy to achieve low and stable inflation, economy growth, and (often) target low unemployment. Their primary instrument of influence over the economies of respective countries is controlling the short-term interest rates, by setting rates at which qualified market participants receive or pay interest on their balances on mandatory deposits with the relevant central banks. (Since GFC other tools have been developed such as quantitative easing, or QE, which is central banks buying government bonds to control the long end of the interest rate curve). Hence, the role of central banks is absolutely critical to fixed income markets. “Reading the Fed” is a fundamental function of interest rate dealers and macro-oriented hedge funds. We will shortly see how to incorporate expected/predicted central bank actions in building interest rate curves and conducting trading strategies.

Topic 2

Making Money with Options

2.4 Discrete Case

2.4.1 The Parable of the Bookmaker

This is adapted from the Baxter and Rennie book – highly recommended!

- Bookmaker taking bets in a two-horse rate
- After doing much research he calculates (correct) odds of one horse winning with 25% and the other with 75%
- In bookmaker language the odds are set to 3-1 against and 3-1 on (”$n-m$ odds against” is you pay \$m and get back your stake plus \$n. So implied fair probability is $m/(m+n)$)
- Bets are \$5,000 for the first and \$10,000 for the second – total takings of \$15,000
    - First horse wins: Pays out \$20,000 so a loss of \$5,000 (to the house)
    - Second horse wins, pays out \$10,000×4/3 =\$10,000+\$3,334 so a win of \$1,667 (to the house)
- Good trade?
    - On average $-0.25 \times 5000 + 0.75 \times 1667 = 0$
    - But significant risk
- Better trade? Set the odds based on the money received
- 2-1 against and 2-1 on for the two horses
    - If horse A wins give the money that was bet on horse B to them
    - And vice versa
    - Make money on the extra spread
    - Riskless!

- The ”better” odds look like probabilities: 1/3 on the first horse and 2/3 on the second
- Use these ”risk-neutral” odds/probabilities for risk-less profit
- Same with options
  - Think of how you can weight different outcomes to immunize yourself
  - Use them as probabilities
  - Carry on as usual

2.4.2 My Standard Interview Question

- Reminder: Call option is the right to buy stock in the future at a certain price called strike.
  - So payoff is $\max(S_T - K, 0)$
- Question
  - One-period economy
  - Stock is at 100
  - Probability of going to 150 is 1/3
  - Probability of going to 50 is 2/3
  - What is the price of a call option struck at 100?
- Naive answer
  - $50 \times 1/3 + 0 \times 2/3 = 16.6$
  - Follow up question − call option with the strike 0
  - By previous logic $150 \times 1/3 + 50 \times 2/3 = 83.3$
  - But has to be equal to the stock price today. Which is 100. So something is wrong
- Right answer
  - Probabilities irrelevant
  - Can replicate the payoff of the call option with stock and bond (zero interest rate account in this case)
  - Just like the horse story
- Replicate the payoff
  - Buy half the stock so can go from 50 to 75 or 25
  - With some of the borrowed money in the amount of 25

- So the portfolio can go 25 to 50 or 0
    - Which is the same payoff as the call option
    - And hence the right price is 25

- Risk neutral probability
    - $100 = 150 \times p + 50 \times (1 - p)$
    - Implies $p = 1/2$
    - Then for call option $50 \times p + 0 \times (1 - p) = 25$
    - Which is the replication price for the call option!
    - Risk-neutral probabilities!

2.4.3 A More Extreme Example – an Intro into Risk Preferences

Consider the following example, courtesy of Keith Lewis XXX https://keithalewis.github.io/math/123.html.

> Suppose a one-period market has a bond with price 1 at the beginning of the period that goes to price 2 at the end of the period, and a stock with price 1 at the beginning of the period that goes to price 1 with probability 0.1 or price 3 with probability 0.9 at the end of the period. What is the value of a call with strike 2?

- Call pays 1 in the up scenario and 0 in the down scenario
- Naive answer: $(0 \times 0.1 + 1 \times 0.9)/2 = 0.45$
    - We know that does not work since for the stock $(1 \times 0.1 + 3 \times 0.9)/2 = 1.4 \neq 1$
- Using risk-neutral probabilities to match the stock we get $p = 0.5$ so that $(1 \times 0.5 + 3 \times 0.5)/2 = 1$
- So that "correct" value of the call option is $(0 \times 0.5 + 1 \times 0.5)/2 = 0.25$

Quoting verbatim from Lewis,

> When I was proudly showing off this mathematically correct analysis to a trader he looked at me as if I had lost my mind. "Wait, what? I can give you 0.25 to get back a dollar 90% of the time? If I have to borrow at 100% interest that is still a quarter to get half a buck. I'll take that all day long!

Here is an important point that we will elaborate over and over in the future. The risk-neutral price of the call option in this example looks very attractive to this trader. It could be based on his experience, or historical analysis where doubling your money with 90% probability is a good trade.

How can we quantify this "goodness" ? Risk-neutral/no-arbitrage approach is fundamental for pricing derivatives by decomposing them into a (sequence of) locally risk-less trades, as we will demonstrate shortly. It does not, however, give traders the tools to assess what is a "good" vs. a "bad" trade. This determination is subjective and is based on the trader's, or their employer's, tolerance to risk (risk aversion) and their perceived benefits from profits vs. losses (their utility function). Additionally, as we mentioned, historical/statistical (vs. risk-neutral) probabilities, and specifically the divergence between the two, often leads to Relative Value trading strategies. Developing tools grounded in theory that can answer some of these questions, and understanding the interplay between risk-neutral, statistical, and subjective probabilities, is the theme for much of this set of lectures.

2.5 Continuous Case

2.5.1 Black-Scholes-Merton Options Pricing

Let us quickly review BS formula derivation so we can see the risk-neutral arguments in action.

Asset $S(t)$, no dividends, risk free rate r. Asset dynamics in real world measure are given by

$$dS(t)/S(t) = \mu_S \, dt + \sigma_S \, dW(t)$$

Let $V\,(t, S)$ be a derivative security

- Pays $V(T, S(T))$ at time T – fully determined by the value of $S(T)$ at time T
  - E.g. $V(T, S) = \max(S - K, 0)$ paid at T

By Ito's lemma

$$dV(t) = (\mathcal{L}V(t)) \;\; dt + \Delta(t) \, dS(t),$$

$$\mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{2}\sigma_S^2 S^2 \frac{\partial^2}{\partial S^2}$$

where Δ is the option's delta,

$$\Delta(t) = \frac{\partial V(t)}{\partial S}$$

To replicate the derivative, at time t we hold $\Delta(t)$ units of stock and $\phi(t)$ bonds $B(t)$. Then the value of the replication portfolio, which we denote by $\Pi(t)$, is equal to

$$V(t) = \Pi(t) = \Delta(t)S(t) + \phi(t)B(t) = \Delta(t)S(t) + \gamma(t) \tag{1}$$

where $\gamma(t) \triangleq \phi(t)B(t)$ is the value of the cash position.

Bonds grow at the risk-free rate

$$dB(t) = r\ B(t)\,dt$$

By self-financing condition

$$dV(t) = d\Pi(t) = \Delta(t)dS(t) + \phi(t)dB(t) = \Delta(t)dS(t) + r\gamma(t)dt \tag{2}$$

From (2)

$$r\gamma(t)dt = dV(t) - \Delta(t)\,dS(t)$$

which is, by Ito's lemma,

$$dV(t) - \Delta(t)\,dS(t) = (\mathcal{L}V(t))\ \ dt = \left(\frac{\partial}{\partial t} + \frac{\sigma_S^2}{2}S^2\frac{\partial^2}{\partial S^2}\right)V(t)\,dt$$

Thus we have

$$\left(\frac{\partial}{\partial t} + \frac{\sigma_S^2}{2}S^2\frac{\partial^2}{\partial S^2}\right)V = r\ \gamma(t) = r\ \left(V - \frac{\partial V}{\partial S}S\right)$$

which, after some rearrangement, yields

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{\sigma_S^2}{2}S^2\frac{\partial^2 V}{\partial S^2} = rV$$

Note that there is no dependence on the real-world drift.

2.6 Risk-Neutral Probabilities

Let us recall the derived BS PDE

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma_S^2 S^2\frac{\partial^2 V}{\partial S^2} = rV$$

The solution, by Feynman-Kac formula, is given by

$$V(t) = \mathrm{E}_t\left(e^{-\int_t^T r\ du}V(T)\right)$$

in measure in which the stock grows at rate r, i.e.

$$dS(t)/S(t) = r\,dt + \sigma_S\,dW_S(t).$$

- Risk-neutral measure – unreal measure to get the real result
- Real-world drift was taken out by hedging
    - Just like in the binomial model and the horses example

2.7 The Balance Equation

The BS equation also sometimes called the Balance Equation, is given by

Recall
$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma_S^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV$$

Recall some jargon

- $\frac{\partial V}{\partial t}$: Theta
- $\frac{\partial V}{\partial S}$: Delta
- $\frac{\partial^2 V}{\partial S^2}$: Gamma
- (Not in the equation) $\frac{\partial V}{\partial \sigma}$: Vega

So
$$\text{Theta} + rS \times \text{Delta} + \frac{1}{2}\sigma_S^2 S^2 \times \text{Gamma} = r \times \text{Value}$$

2.8 Cost of Hedging

We can now calculate the cost of hedging. Notations: option price $C(t)$ underlying $S(t)$, delta $\Delta(t)$, cash account $\gamma(t)$. Recall (assume $r = 0$ and use self-financing)

$$\begin{aligned} C(t) &= \Pi(t) = \Delta(t)S(t) + \gamma(t) \\ dC(t) &= \Delta(t)\, dS(t) \end{aligned}$$

so
$$\Delta(t)\, dS(t) = dC(t) = d(\Delta(t)S(t) + \gamma(t))$$

In discrete-time notation with $0 = t_0 < t_1 < \cdots < t_N = T$,

$$\Delta(t_i)(S(t_{i+1}) - S(t_i)) = \Delta(t_{i+1})S(t_{i+1}) - \Delta(t_i)S(t_i) + \gamma(t_{i+1}) - \gamma(t_i)$$

hence after some simplifications

$$\gamma(t_{i+1}) - \gamma(t_i) = -(\Delta(t_{i+1}) - \Delta(t_i))S(t_{i+1})$$

Here
$$-\left(\Delta(t_{i+1}) - \Delta(t_i)\right) S(t_{i+1})$$
is the cost of re-hedging, i.e. the cost of the shares to transact at t_{i+1} to maintain the hedge

Now let us recall

$$\gamma(t_{i+1}) - \gamma(t_i) = -(\Delta(t_{i+1}) - \Delta(t_i))S(t_{i+1})$$

Thus

$$\gamma(T) - \gamma(0) = -\sum_i (\Delta(t_{i+1}) - \Delta(t_i))S(t_{i+1})$$

Also

$$C(t) = \Pi(t) = \Delta(t)S(t) + \gamma(t)$$

so that

$$\begin{aligned} C(T) - C(0) &= \Delta(T)S(T) - \Delta(0)S(0) + \gamma(T) - \gamma(0) \\ &= \Delta(T)S(T) - \Delta(0)S(0) - \sum_i (\Delta(t_{i+1}) - \Delta(t_i))S(t_{i+1}) \end{aligned}$$

Here

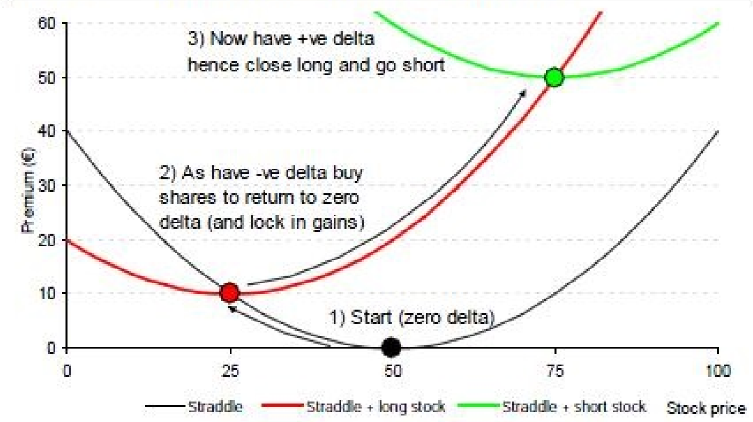
- $C(T) - C(0)$ is the change in option value
- $\Delta(T)S(T)$ is the proceeds from selling delta hedge at the end
- $-\Delta(0)S(0)$ is the cost of buying the delta hedge initially
- $-\sum \left(\Delta(t_{i+1}) - \Delta(t_i)\right) S(t_{i+1})$ is the cost of re-hedging throughout life

Therefore, a critical observation:

- PnL of option position is exactly equal to the cost of setting/unwinding the hedge plus cost of re-hedging throughout life
  - True path by path, not on average! I.e. for any realization of the market evolution in the future
  - Obviously somewhat idealized

2.9 Gamma Trading

Figure 25. Locking in Gains through Delta Hedging



Source: Santander Investment Bolsa.

Figure 1: Long gamma hedging

A worked out example is available in T2_Gamma_Scalping_01.ipynb.

Long gamma position makes money on each move/re-hedging
Who would want to be on the other side?

- One has to pay to get into a positive gamma position
- Also lose on Theta (time decay)

So when do we make money?

- When the scale of underlying moves is larger than embedded in the price

For the short gamma position it is the opposite

- We get paid to be in one – so money up-front
- But we lose money on each re-hedging

The main dilemma with the short gamma positions – hedge or not

- Yes – then definitely lose some money
- No – then possibly lose a lot of money

This activity is sometimes also called *gamma scalping*.

2.10 Fundamental Theorem of Derivatives Trading

Let us assume the real-world measure (P) volatility is $\sigma(t)$, and the underlying process is given by

$$dS/S = O(dt) + \sigma(t)\ dW^P(t)$$

Consider a hedged portfolio $V = V(S)$ that we value/hedge using Black-Scholes. On one hand,

$$\begin{aligned} dV(t) &= \left(\frac{\partial V(t)}{\partial t} + \frac{1}{2}\sigma(t)^2 S(t)^2 \frac{\partial^2 V(t)}{\partial S^2} \right)\ dt + \frac{\partial V(t)}{\partial S}\ dS(t) \\ &= \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma(t)^2 S(t)^2 \frac{\partial^2 V(t)}{\partial S^2} \right)\ dt + 0 \text{ (hedged!)} \end{aligned}$$

On the other hand, if we price at the implied σ_{impl}, (with $r = 0$) by Black-Scholes,

$$\frac{\partial V(t)}{\partial t} = -\frac{1}{2}\sigma_{\text{impl}}^2 S(t)^2 \frac{\partial^2 V(t)}{\partial S^2}$$

Hence we obtain the main result,

$$dV = \frac{1}{2}\left(\sigma(t)^2 - \sigma_{\text{impl}}^2\right)\tilde{\Gamma}(t)dt \tag{3}$$

where the so-called "cash gamma" is defined as

$$\tilde{\Gamma}(t) \triangleq S(t)^2 \frac{\partial^2 V}{\partial S^2}(t)$$

What this formula tells us is that "PnL on a hedged portfolio is given by its cash gamma times the difference between the realized and the implied volatility". So we should buy at the implied volatility that is lower than expected realized volatility. This is the main job of options traders. While the conclusion seems rather obvious in hindsight, it highlights the fact that trading options is basically taking views on where the realized volatility is going to go. Next section explores this aspect.

2.11 Implied vs Realized Volatility

Apart from helping us understand where portfolio PnL is coming from, the equation (3) is useful for understanding the relationship between the realized and the implied volatilities. Let $V(t)$ be a particular option, say a call option $C(t, T, K)$ with expiry T and strike K, hedged with the underlying stock. If the interest rates are zero (our assumption here), then we would expect

$$\mathrm{E}(V(T) - V(0)) = 0.$$

As

$$V(T) - V(0) = \int_0^T dV(t),$$

using (3) we have

$$\mathrm{E}\left(\int_0^T \left(\sigma(t)^2 - \sigma_{\mathrm{impl}}(T,K)^2\right)\tilde{\Gamma}(t)\ dt\right) = 0$$

or, rewriting slightly

$$\sigma_{\mathrm{impl}}(T,K)^2\mathrm{E}\left(\int_0^T \tilde{\Gamma}(t)\ dt\right) = \mathrm{E}\left(\int_0^T \sigma(t)^2\tilde{\Gamma}(t)\ dt\right). \tag{4}$$

On one hand, as $\sigma_{\mathrm{impl}}(T,K)$ is observable, it gives us a way to estimate what the market thinks the realized volatility $\sigma(t)$ would be. Clearly we need to estimate $\tilde{\Gamma}(t)$ but that is based on where the underlying $S(t)$ is going, and here you can do a number of scenarios.

On the other hand, we can use (4) to come up with our own view on the implied volatility, based on our estimate on future values of the realized volatility:

$$\hat{\sigma}_{\mathrm{impl}}(T,K)^2 = \frac{\mathrm{E}\left(\int_0^T \sigma(t)^2\tilde{\Gamma}(t)\ dt\right)}{\mathrm{E}\left(\int_0^T \tilde{\Gamma}(t)\ dt\right)}. \tag{5}$$

We used the hat in the notation $\hat{\sigma}_{\mathrm{impl}}(T,K)$ to emphasize that this is our own estimate of the implied volatility that is "fair", not what you observe on the market. One use of the formula (5) is to compare our views vs. the market. If we have strong views on where $\sigma(\cdot)$ is going, we can calculate our estimate of the implied volatility $\hat{\sigma}_{\mathrm{impl}}(T,K)$ and compare it to the market one, $\sigma_{\mathrm{impl}}(T,K)$, and depending on the sign of the difference, either buy or sell the option in question.

Of course it is impossible to know future realized volatility. But one can take a view that the history is a reasonable estimate of the future. So yet another way to use (5) is to estimate *historical* realized volatility, assume it will be the same, on average, in the future, and trade options based on that. This is not an uncommon strategy.

2.12 The Fundamental Role of Funding

Optional section

Classic BS is derived assuming "risk free rate" r used as the rate on bank deposits, as the rate at which we can borrow money, and the rate at which we can finance stock purchases. Post financial crisis these idealized set up became no longer tenable. Let us review how the BS formalism is adapted when we make more realistic assumptions about the *cost of funding*. This is adapted from Piterbarg (2010). The cost of funding is fundamental to understanding the PnL of trading desks and, to some extent, hedge funds and other buy side firms.

A derivatives trading desk:

- Trades derivatives
- Hedges with underlying assets
- Lends/borrows money

Cash for this activity is required and it comes from different sources:

- Unsecured
- Secured by assets
- Collateral (daily MTM paid/received in cash or high quality bonds)

As we said, historically, we did not differentiate between different rates. But since the GFC, there are large differences and volatility in the *funding spread*

- How does it affect pricing?
- We adopt funding prospective, not credit risk prospective

2.12.1 Black-Scholes with Funding

Let us derive the BS extension with funding. We have an asset $S(t)$, and various rates

- Risk-free overnight rate $r_C(t)$, paid on collateral accounts (under so-called CSA)
- Asset dividend rate $r_D(t)$
- Rate on borrowing secured by asset $r_R(t)$
- Unsecured bank funding rate $r_F(t)$
- funding spread $s_F(t) \triangleq r_F(t) - r_C(t)$

Asset dynamics in the real-world measure are given by

$$dS(t)/S(t) = \mu_S(t)\,dt + \sigma_S(t)\,dW(t)$$

Let $V(t, S)$ be a derivative security. By Ito's lemma

$$dV(t) = (\mathcal{L}V(t))\,dt + \Delta(t)\,dS(t),$$

where Δ is the option's delta,

$$\Delta(t) = \frac{\partial V(t)}{\partial S}$$

and $\mathcal{L}$ is the usual BS operator that we have seen before,

$$\mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{2}\sigma_S^2 S^2 \frac{\partial^2}{\partial S^2}.$$

Let $C(t)$ be the collateral held at time t against this derivative, allowed to be different from $V(t)$

- At time t we hold $\Delta(t)$ units of stock and $\gamma(t)$ cash
- The cash amount $\gamma(t)$ is split among a number of accounts:

1. Amount $C(t)$ is in collateral
2. Amount $V(t) - C(t)$ needs to be borrowed/lent unsecured from the treasury desk
3. Amount $\Delta(t)S(t)$ is borrowed to finance the purchase of $\Delta(t)$ stocks. It is secured by stock purchased
4. Stock is paying dividends

The growth of all cash accounts is given by

$$d\gamma(t) = \left[r_C(t)C(t) + r_F(t)\left(V(t) - C(t)\right) - r_R(t)\Delta(t)S(t) + r_D(t)\Delta(t)S(t)\right] dt$$

To replicate the derivative, at time t we hold $\Delta(t)$ units of stock and $\gamma(t)$ cash. Then the value of the replication portfolio, which we denote by $\Pi(t)$, is equal to

$$V(t) = \Pi(t) = \Delta(t)S(t) + \gamma(t). \tag{6}$$

On the other hand, from (6), by self-financing condition,

$$d\gamma(t) = dV(t) - \Delta(t)\, dS(t)$$

which is, by Ito's lemma,

$$dV(t) - \Delta(t)\, dS(t) = (\mathcal{L}V(t))\, dt = \left(\frac{\partial}{\partial t} + \frac{\sigma_S(t)^2}{2} S^2 \frac{\partial^2}{\partial S^2}\right) V(t)\, dt$$

Thus we have

$$\left(\frac{\partial}{\partial t} + \frac{\sigma_S(t)^2}{2} S^2 \frac{\partial^2}{\partial S^2}\right) V = r_C(t)C(t) + r_F(t)\left(V(t) - C(t)\right) + \left(r_D(t) - r_R(t)\right)\frac{\partial V}{\partial S} S$$

After some rearrangement

$$\frac{\partial V}{\partial t}+\left(r_R(t)-r_D(t)\right)\frac{\partial V}{\partial S}S+\frac{\sigma_S(t)^2}{2}S^2\frac{\partial^2 V}{\partial S^2}=r_F(t)V(t)-\left(r_F(t)-r_C(t)\right)C(t).$$

The solution, by Feynman-Kac formula, is given by

$$V(t)=\mathrm{E}_t\left(e^{-\int_t^T r_F(u)du}V(T)+\int_t^T e^{-\int_t^u r_F(v)dv}\left(r_F(u)-r_C(u)\right)C(u)\,du\right) \tag{7}$$

in measure in which the stock grows at rate $r_R(t)-r_D(t)$, i.e.

$$dS(t)/S(t)=\left(r_R(t)-r_D(t)\right)\,dt+\sigma_S(t)\,dW_S(t). \tag{8}$$

2.12.2 Valuation with Collateral

Recall that the solution is given by

$$V(t)=\mathrm{E}_t\left(e^{-\int_t^T r_F(u)du}V(T)+\int_t^T e^{-\int_t^u r_F(v)dv}\left(r_F(u)-r_C(u)\right)C(u)\,du\right) \tag{9}$$

in measure in which the stock grows at rate $r_R(t)-r_D(t)$, i.e.

$$dS(t)/S(t)=\left(r_R(t)-r_D(t)\right)\,dt+\sigma_S(t)\,dW_S(t) \tag{10}$$

There is another useful formula obtained by simple manipulations,

$$\begin{aligned}V(t)=&\ \mathrm{E}_t\left(e^{-\int_t^T r_C(u)\,du}V(T)\right)\\&-\mathrm{E}_t\left(\int_t^T e^{-\int_t^u r_C(u)du}\left(r_F(u)-r_C(u)\right)\left(V(u)-C(u)\right)\,du\right)\end{aligned} \tag{11}$$

Clearly

$$\begin{aligned}\mathrm{E}_t\left(dV(t)\right)&=\left(r_F(t)V(t)-\left(r_F(t)-r_C(t)\right)C(t)\right)dt\\&=\left(r_F(t)V(t)-s_F(t)C(t)\right)\,dt\end{aligned} \tag{12}$$

If the collateral is equal to the value V then

$$\mathrm{E}_t\left(dV(t)\right)=r_C(t)V(t)\,dt,\quad V(t)=\mathrm{E}_t\left(e^{-\int_t^T r_C(u)\,du}V(T)\right),$$

and the derivative should be discounted at risk-free rate

If the collateral is zero, then

$$\mathrm{E}_t\left(dV(t)\right)=r_F(t)V(t)\,dt \tag{13}$$

and the rate of growth is equal to the bank's unsecured funding rate.

First (rather trivial) conclusion that CSA and non-CSA trades should be discounted off different curves. This of course should be applied on portfolio level

2.12.3 Example: Zero-Strike Call Option

Probably the simplest derivative contract: promise to deliver this asset at a given future time T. It is called a zero-strike call option with expiry T. In the standard theory the value of is equal to the value of the asset itself (in the absence of dividends). But it is different in our set up

- The payoff is given by $V(T) = S(T)$ and the value, at time t, assuming no CSA, is given by

$$V_{\text{zsc}}(t) = \mathrm{E}_t \left(e^{-\int_t^T r_F(u)du} S(T) \right)$$

- On the other hand, if $r_D(t) = 0$, then

$$S(t) = \mathrm{E}_t \left(e^{-\int_t^T r_R(u)du} S(T) \right)$$

Clearly, $S(t) \neq V_{\text{zsc}}(t)$. Why?

In the language of funding, the asset $S(\cdot)$ can be used to secure funding — which is reflected in the discount rate applied — while V_{zsc} cannot be.

Topic 3

Fixed Income Instruments pre- and post-Libor reform

3.13 A Brief Review of Notations

3.13.1 Zero-Coupon Bonds and Forward Rates

Let $P(t,T)$ denote the time t price of a zero-coupon bond (also known as a *discount bond*) delivering for certain \$1 at maturity $T \geq t$. Suppose we are interested in purchasing at some future time T a zero-coupon bond maturing at $T+\tau$, $\tau > 0$. At time $t < T$, the price of such a bond can be locked in by

1. purchasing at time t one $(T+\tau)$-maturity zero-coupon bond; and
2. selling short ("shorting") $P(t,T+\tau)/P(t,T)$ $T-$maturity zero-coupon bonds.

The time t cost of executing this strategy is zero,

$$-1 \cdot P(t,T+\tau) + P(t,T+\tau)/P(t,T) \cdot P(t,T) = 0,$$

but a flow of

$$-P(t,T+\tau)/P(t,T)$$

will take place at time T as the $T-$maturity short position matures. This is compensated by an inflow of \$1 at time $T+\tau$. In other words, our trading strategy effectively fixes the time T purchase price of the $(T+\tau)$-maturity bond at

$$P(t,T,T+\tau) \triangleq P(t,T+\tau)/P(t,T), \quad \tau > 0,$$

a quantity known as the time t *forward price* for the zero-coupon bond spanning $[T, T+\tau]$.

It is often convenient to characterize a forward bond price by a discount rate. One such rate is the *continuously compounded forward yield* $y(t,T,T+\tau)$, defined by

$$e^{-y(t,T,T+\tau)\tau} = P(t,T,T+\tau). \tag{14}$$

The time between the maturity of the forward bond and the expiry of the forward contract, i.e. τ, is often called the *tenor* of the forward bond or the forward yield. In the definition of the continuously compounded yield lies an implicit, and idealized, assumption of continuous reinvestment of investment proceeds. Most actual market quotes, however, are based on discrete-time compounding of proceeds. Accordingly, we define a *simple forward rate* $L(t,T,T+\tau)$ as

$$1 + \tau L(t,T,T+\tau) = 1/P(t,T,T+\tau). \tag{15}$$

Again, τ is the *tenor* of the forward rate. For an arbitrary set of dates $T = T_0 < T_1 < T_2 < \ldots < T_n$, notice that forward bond prices can be recovered from forward rates by simple compounding,

$$P\left(t,T_n\right)/P(t,T) = \prod_{i=1}^{n} \frac{1}{1 + \left(T_i - T_{i-1}\right) L(t,T_{i-1},T_i)}.$$

In the limit $\tau \downarrow 0$,

$$L(t,T,T+\tau) \to f(t,T),$$

where the quantity $f(t,T)$ is the time t *instantaneous forward rate* to time T. We think of $f(t,T)$ as the forward rate spanning $[T, T+dT]$, observed at time t. The relation between instantaneous forward rates and forward bond prices is given by the continuous compounding formula

$$P(t,T,T+\tau) = \exp\left(-\int_T^{T+\tau} f(t,u)\,du\right), \tag{16}$$

such that

$$f(t,T) = -\frac{\partial \ln P(t,T)}{\partial T}, \tag{17}$$

and, from (14),

$$y(t,T,T+\tau) = \tau^{-1}\int_T^{T+\tau} f(t,u)\,du, \quad f(t,T) = \lim_{\tau\downarrow 0} y(t,T,T+\tau).$$

We also notice the relationship

$$f(t,T) = \frac{\partial\left(y(t,t,T)(T-t)\right)}{\partial T} = y(t,t,T) + (T-t)\frac{\partial y(t,t,T)}{\partial T}.$$

The quantity

$$r(t) \triangleq f(t,t) \tag{18}$$

is an $\mathcal{F}_t$-measurable random variable known as the *short rate* or sometimes the *spot rate*. Loosely speaking, we can think of $r(t)$ as the overnight rate in effect at time t.

Finally, let us note that interest rates of various flavours and related quantities are typically quoted in percentage points or sometimes in *basis points*, where 1 basis point $= 1/100$ of one percent.

3.13.2 Annuity Factors and Par Rates

Most fixed income securities involve multiple cash flows taking place on a pre-set schedule of dates, often referred to as a *tenor structure*,

$$0 \le T_0 < T_1 < \ldots < T_N.$$

Given a tenor structure, for any two integers k, m satisfying $0 \le k < N$, $m > 0$, and $k + m \le N$, we can define an *annuity factor* $A_{k,m}$ by

$$A_{k,m}(t) = \sum_{n=k}^{k+m-1} P(t, T_{n+1})\tau_n, \quad \tau_n = T_{n+1} - T_n. \tag{19}$$

Annuity factors provide for compact notation when pricing coupon-bearing securities. For instance, a security making m coupon payments of $c\tau_n$ at all $T_{n+1}, n = k, \ldots, k+m-1$, is easily seen to have time t value of

$$cA_{k,m}(t), \quad t \le T_k.$$

If the security also involves a back-end return of notional at time T_{k+m} (as is the case for a regular coupon-bearing bond), the pricing expression is

$$cA_{k,m}(t) + P(t, T_{k+m}), \tag{20}$$

where we assume that the bond has been normalized to have a unit notional. The time t forward price to T_k of the security (20) is

$$cA_{k,m}(t)/P(t,T_k) + P(t,T_{k+m})/P(t,T_k);$$

the value of the coupon c for which this expression equals 1 is known as the *forward par rate* or, when used in the context of swap pricing, as the *forward swap rate*. With $S_{k,m}(t)$ denoting the time t swap rate, we apparently have

$$S_{k,m}(t) = \frac{P(t,T_k) - P(t,T_{k+m})}{A_{k,m}(t)}, \quad t \le T_k. \tag{21}$$

From the definition of $L(t, T_n, T_{n+1})$ in (15), a little thought shows that the numerator of the expression for $S_{k,m}(t)$ can be expanded into a weighted sum of forward rates, leading to the alternative expression

$$S_{k,m}(t) = \frac{\sum_{n=k}^{k+m-1} \tau_n P(t, T_{n+1}) L_n(t)}{A_{k,m}(t)}, \quad t \leq T_k, \tag{22}$$

where we have introduced the useful shorthand

$$L_n(t) \triangleq L(t, T_n, T_{n+1}).$$

It follows that the forward swap rate can be loosely interpreted as a weighted average of simple forward rates on the specified tenor structure. We note for the future that the time T_k is sometimes referred to as the *fixing date*, or *expiry*, of the swap rate $S_{k,m}$, while the length of the corresponding swap, $T_{k+m} - T_k$, is sometimes called the *tenor* of the swap rate.

3.14 Fixed Income Measures

3.14.1 Risk Neutral Measure

We use $V(t)$ to denote the time t price of a derivative security making an $\mathcal{F}_T$-measurable payment of $V(T)$. The numeraire defining the risk-neutral measure Q is the continuously compounded money market account $\beta(t)$, satisfying the locally deterministic SDE

$$d\beta(t) = r(t)\beta(t)\, dt, \quad \beta(0) = 1, \tag{23}$$

where $r(t)$ is the short rate, $r(t) = f(t, t)$. Solving this equation yields

$$\beta(t) = e^{\int_0^t r(u)\, du}.$$

In the absence of arbitrage the numeraire-deflated process $V(t)/\beta(t)$ must be a martingale, implying the derivative security valuation formula

$$V(t)/\beta(t) = \mathrm{E}_t^{\mathrm{Q}} \left(V(T)/\beta(T) \right), \quad t \leq T, \tag{24}$$

or equivalently

$$V(t) = \mathrm{E}_t^{\mathrm{Q}} \left(e^{-\int_t^T r(u)\, du} V(T) \right). \tag{25}$$

If we apply (25) to the special case of $V(T) = 1$, we obtain a fundamental bond pricing formula. In the absence of arbitrage, the time t price $P(t, T)$ of a $T-$maturity zero-coupon bond is

$$P(t, T) = \mathrm{E}_t^{\mathrm{Q}} \left(e^{-\int_t^T r(u)\, du} \right). \tag{26}$$

It follows from (26) that specification of the dynamics of $r(t)$ under Q suffices to determine the prices of discount bonds at all times and maturities. Models that are based on such a direct specification of $r(t)$ dynamics are known as *short rate models*. Notice the resemblance between expressions (16) and (26); if $r(t)$ is deterministic, the two expressions will agree as $r(u) = f(t,u)$, $u \geq t$. If $r(t)$ is random, one may wonder whether this result will hold in expectation. The answer to this is negative, i.e.

$$f(t,u) \neq \mathrm{E}_t^{\mathrm{Q}}\left(r(u)\right), \tag{27}$$

provided r is random.

3.14.2 T-Forward Measure

Other traded instruments can be used as a numeraire. Recall that the so-called $T-$forward measure Q^T uses the discount bond $P(\cdot, T)$ as the numeraire. Then, with $\mathrm{E}_t^{\mathrm{T}}$ denoting the expected value operator in this measure,

$$V(t)/P(t,T) = \mathrm{E}_t^T\left(V(T)/P(T,T)\right), \quad t \leq T.$$

As, obviously, $P(T,T) = 1$, the expression simplifies to the convenient form

$$V(t) = P(t,T)\mathrm{E}_t^T\left(V(T)\right).$$

Comparison of the risk-neutral and $T-$forward valuation formulas shows that shifting to the $T-$forward measure in a sense decouples the expectation of the terminal payout $V(T)$ from that of the numeraire. This is often very convenient when we attempt to construct analytical formulas for prices of certain simple interest rate derivatives.

3.15 Libor Rates and Related Instruments

3.15.1 Libor Rates

The so-called Libor rates, and Libor-based instruments, used to be the dominant underlying instruments for a vast majority of interest rate derivatives. After the Libor reform of the early 2020's they have been replaced by the so-called risk free rates (RFR). We will briefly discuss the Libor reform later in the course, but understanding Libor-based instruments is still essential to having a good grasp of the functioning of interest rate markets, in parts because the market features of RFR markets often mimic their Libor predecessors.

For a given entity, the cost of borrowing money will depend on its credit quality. Governments of developed countries, perceived to have virtually no possibility of default, issue bonds at comparatively low interest rates that reflect this perception. While the market in government debt is vast, corporations typically find it more convenient to use and originate fixed income instruments linked to rates that are more reflective of their own financing costs (i.e., credit quality). By far the most common of such reference rates used to be, pre-Libor reform, the London Interbank Offered rate, commonly known as the *Libor rate*. The Libor rate is a filtered average of bank estimates of rates at which they can borrow for a given term in the *interbank money market*, i.e. the wholesale market in which banks provide unsecured short-term credit to each other. Libor rates were quoted for multiple deposit maturities ranging from one day to one year, and were set every business day by averaging polling results from a number of large banks. Libor rates were available for deposits in different currencies, so that there is a USD-Libor rate, a EUR-Libor rate (EURIBOR), and so on.

Most Libor rates have ceased as the result of the Libor reform. USD-:Libor rate is set to be decommissioned in June 2023, but the trading in USD-Libor linked instruments is currently restricted (hedging only). EURIBOR is deemed to continue to be a representative benchmark and currently there are no plans to kill it.

While Libors are (mostly) gone, understanding them is still important. We will discuss later in the course that the Libor replacements, and new benchmarks, Overnight Index Rates (OIS), also known as Risk-Free Rates (RFR), when compounded over a term, look very much like forward Libor rates, so much of our understanding of (forward) Libor rates can be directly translated into the post-Libor reform world.

3.15.2 Certificates of Deposit

We now proceed to define the universe of securities of interest to us. For technical precision, we shall occasionally need to refer to the risk-neutral measure Q, as well as its associated expectation operator $\mathrm{E} = \mathrm{E}^{\mathrm{Q}}$ and its numeraire $\beta(t)$.

We start with the *certificate of deposit* (or *CD*), a deposit of money for a pre-specified term at a pre-specified interest rate. Terms may range from one week to one year or more, with the most popular being a 3 month or a 6 month term, depending on the currency of the deposit. If 1 (dollar) is deposited at time T for a period of τ years, then the amount of capital to be returned at time $T + \tau$ is given by

$$1 + \tau L,$$

where L is, by definition, the interest rate for the CD. The rate is quoted as a simple rate, i.e. a rate with the compounding frequency equal to the term of the deposit. Notice that the average value of L for CDs quoted in the interbank market will, by definition, be equal to the (spot) Libor rate for tenor τ. Spot Libor rates for various tenors used to be calculated daily and published by major news services such as Bloomberg or Reuters. As mentioned above, Libor used to serve as the primary reference rate in fixed income markets.

If $P(T, T+\tau)$ is the (Libor-based) discount factor to date $T+\tau$ as observed at T, then the discounted value of receiving $1 + \tau L$ at time $T + \tau$ should be equal to 1 at time T, i.e.

$$1 = P\left(T, T+\tau\right)\left(1 + \tau L\right).$$

In particular, recalling the definition (15) of $L(t,T,T+\tau)$, the rate L paid on the CD is a simple spot rate

$$L = L\left(T,T,T+\tau\right) = \frac{1}{\tau}\left(\frac{1}{P\left(T,T+\tau\right)} - 1\right). \tag{28}$$

3.15.3 Forward Rate Agreements (FRA)

A certificate of deposit allows a market participant to lock in an interest rate for a given period of time, effective immediately. Many market participants, however, find it convenient to lock in interest rates for a given period of time that starts in the future. Contracts that provide such a rate guarantee are known as *forward contracts* or, in a fixed income context, *forward rate agreements* (FRAs). An FRA for the period $[T,T+\tau]$ is a contract to exchange fixed rate payment (agreed at the initiation of the contract) against a payment based on the time T spot Libor rate of tenor τ. While all payments on an FRA are exchanged at, or near[1], time T, the contract is structured so that the payments are made in $T+\tau$ money.

Formally, consider the origination at time t, $t \le T$, of a unit notional FRA contract with a rate of k. Ignoring payment delays, from the perspective of the fixed rate payer the net payment at time T will be

$$V_{\mathrm{FRA}}(T) = \tau\left(L\left(T,T,T+\tau\right) - k\right)/\left(1+\tau L(T,T,T+\tau)\right),$$

with the (contractually specified) factor $1/(1+\tau L(T,T,T+\tau))$ applied to roll the payment to the future date $T+\tau$. We note that

$$1/\left(1+\tau L(T,T,T+\tau)\right) = P(T,T+\tau)$$

so, by the fundamental pricing result (24), the value of this contract at time t is equal to

$$V_{\mathrm{FRA}}(t) = \beta(t)\mathrm{E}_t\left(\beta(T)^{-1}\tau\left(L\left(T,T,T+\tau\right)-k\right)P\left(T,T+\tau\right)\right)$$

(recall that $\beta(\cdot)$ is the money market account). Substituting (28) we obtain

$$V_{\mathrm{FRA}}(t) = \beta(t)\mathrm{E}_t\left(\beta(T)^{-1}\left(1 - P\left(T,T+\tau\right) - \tau k P\left(T,T+\tau\right)\right)\right).$$

Since $P(\cdot,T+\tau)$ is a traded asset, its price deflated by the numeraire $\beta(\cdot)$ is a martingale. Thus

$$\begin{aligned} V_{\mathrm{FRA}}(t) &= P\left(t,T\right) - P\left(t,T+\tau\right) - \tau k P\left(t,T+\tau\right) \\ &= \tau P\left(t,T+\tau\right)\left(\frac{P\left(t,T\right) - P\left(t,T+\tau\right)}{\tau P\left(t,T+\tau\right)} - k\right). \end{aligned} \tag{29}$$

[1] Typical market conventions call for a two business day payment delay.

Most often, FRAs are issued at no cost to either party at the time of origination. The value of k that makes the FRA contract have value 0 at the contract initiation time t is given by the *forward Libor rate* (see (15)),

$$k = L\left(t, T, T+\tau\right) = \frac{P\left(t, T\right) - P\left(t, T+\tau\right)}{\tau P\left(t, T+\tau\right)}.$$

Thus, a forward Libor rate has the financial interpretation of being a break-even rate on an FRA contract in interbank markets.

We also note an important result linking Libor rates and forward measures:

$$L(t, T, T+\tau) = \mathrm{E}_t^{T+\tau} L(T, T, T+\tau), \tag{30}$$

so that the forward Libor rate $L(t, T, T+\tau)$ is the expected value, in the $(T+\tau)-$forward measure, of the Libor fixing (spot Libor rate) $L(T, T, T+\tau)$.

3.15.4 Eurodollar Futures

FRAs, being forward contracts on Libor rates, allow market participants to either lock in favourable rates for future periods, or to speculate on the future direction of rates. FRAs trade in the OTC market, and are open only to institutions that participate in this market. Alternatively, *futures contracts* on Libor rates used to be available on a number of international exchanges, including the Chicago Mercantile Exchange (CME), London International Financial Futures and Options Exchange (LIFFE), and Marché à Terme International de France (MATIF). The CME interest rate futures contract on a three-month spot Libor rate on US dollar denominated deposits is called the *Eurodollar futures* or, simply, *ED futures* contract.

At maturity T, an ED futures contract is settled at

$$100 \times \left(1 - L\left(T, T, T+\tau\right)\right).$$

The *futures rate* $F(t, T, T+\tau)$ at time t is defined to be the rate such that the *quoted futures price* at time t is equal to[2]

$$100 \times \left(1 - F\left(t, T, T+\tau\right)\right).$$

As is the case for all futures contracts, ED futures are settled (marked to market) daily. Confusing matters somewhat, the actual amount of money that is settled between holders of the long and the short positions in an ED future is determined by the daily change in the *actual futures price* defined by

$$N_{\mathrm{ED}} \times \left[1 - \frac{1}{4} F\left(t, T, T+\tau\right)\right],$$

where N_{ED} is the notional principal of the contract ($\$1,000,000$ for the CME's ED futures). In particular, for 1 basis point (0.01%) increase in the rate

[2] So, if the futures rate is 5%, the quoted futures price is 95.

$F(t, T, T+\tau)$, the CME contract buyer pays $1,000,000 \times 0.25 \times 0.0001 = 25$ dollars to the seller.

As explained later in the course, futures rates $F(t, T, T+\tau)$ are generally different from forward Libor rates $L(t, T, T+\tau)$. The problem of computing the difference, the *ED convexity adjustment*, is considered later.

Unlike FRAs, for which the deposit period is negotiated between two parties, ED futures are standardized. Available contracts expire on four specific dates, one each in March, June, September and December, over the next ten years. Such standardization increases liquidity in each particular contract.

3.15.5 Fixed-for-Floating Swaps

A *swap* is a generic term for an OTC derivative in which two counterparties agree to exchange one stream of cash flows against another stream. These streams are called the *legs* of the swap. A *plain vanilla fixed-for-floating interest rate swap* (a *plain vanilla swap*, or just a *swap* if there is no confusion) is a swap in which one leg is a stream of fixed rate payments and the other a stream of payments based on a floating rate, used to be Libor but now are compounded RFR rates. The legs are denominated in the same currency, have the same notional, and expire on the same date. Payment streams are made on a pre-defined schedule of contiguous time intervals, known as *periods*. Pre-Libor reform, the floating rate would be observed (or *fixed*) at the beginning of each period, with both fixed and floating rate coupons being paid out at the end of the period. With RFRs the rate is only known at the end of the period but, as we shall see later, that does not change the formulas below. We will stick to the Libor view of the world for now.

A plain-vanilla swap is economically equivalent to a multi-period FRA, and serves the same purpose in the market as regular FRAs. Between interest rate dealers and financial institutions, swaps of different maturities are often traded to adjust interest risk positions of the parties involved, or to simply make bets on future direction of interest rates. Swaps are also used by corporates, often in conjunction with bond or note issuance, to transform fixed rate obligations into floating ones, or vice versa.

To formally define a fixed-floating swap, one specifies a tenor structure, i.e. an increasing sequence of maturity times, normally spaced roughly equidistantly

$$0 \leq T_0 < T_1 < T_2 < \ldots < T_N, \quad \tau_n = T_{n+1} - T_n. \tag{31}$$

In a fixed-floating swap with fixed rate k, one party (the fixed rate payer) pays simple interest based on the rate k in return for simple interest payments computed from the Libor rate fixing on date T_n, for each period $[T_n, T_{n+1}]$, $n = 0, \ldots, N-1$. The payments are exchanged at the end of each period, i.e. at time T_{n+1}. In practice, the payments are netted, and only their difference changes hands. From the perspective of the fixed rate payer, the net cash flow of the swap at time T_{n+1} is therefore given by (on a unit notional)

$$\tau_n \left(L_n(T_n) - k \right), \quad L_n(t) = L\left(t, T_n, T_{n+1}\right),$$

for $n = 0, \ldots, N-1$. Dates when the Libor rates are observed are typically called *fixing dates*; dates when payments occur are called *payment dates*. By the fundamental valuation result (24), the value of a swap is equal to the expected discounted value of its (netted) payments. Specifically, the value to the fixed rate payer of a unit notional fixed-floating swap at time t, $0 \leq t \leq T_0$, is given by[3]

$$\begin{aligned} V_{\text{swap}}(t) &= \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta\left(T_{n+1}\right)^{-1} \left(L_n(T_n) - k\right) \right) \\ &= \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta(T_n)^{-1} \left(L_n(T_n) - k\right) P\left(T_n, T_{n+1}\right) \right). \end{aligned}$$

Using the definition of Libor rates $L_n(T_n)$,

$$V_{\text{swap}}(t) = \beta(t) \sum_{n=0}^{N-1} \mathrm{E}_t(\beta(T_n)^{-1}(1 - P(T_n, T_{n+1}) - \tau_n k P(T_n, T_{n+1}))).$$

For each n, $P(\cdot, T_n)$ is a traded asset, so its price deflated by the numeraire $\beta(\cdot)$ is a martingale. Hence

$$V_{\text{swap}}(t) = \sum_{n=0}^{N-1} \left(P\left(t, T_n\right) - P\left(t, T_{n+1}\right) - \tau_n k P\left(t, T_{n+1}\right) \right).$$

Recalling the definition of $L_n(t)$, this can be rewritten as

$$V_{\text{swap}}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \left(L_n(t) - k\right).$$

An important observation is that a vanilla fixed-floating swap can be valued on date t using only the term structure of interest rates observed on that date. In particular, swap values are not affected by the *dynamics* of interest rates, only their current levels.

The swap valuation formula above can be rewritten as follows,

$$V_{\text{swap}}(t) = \left(\sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) \right) \left(\frac{\sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) L_n(t)}{\sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right)} - k \right).$$

Using the definitions (19), (21) and (22),

$$A(t) \triangleq A_{0,N}(t) = \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right), \tag{32}$$

$$S(t) \triangleq S_{0,N}(t) = \frac{\sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right) L_n(t)}{\sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right)}, \tag{33}$$

[3] This is a somewhat idealized expression. See Andersen & Piterbarg (2010) for more details on market day counting conventions and related topics.

we obtain the convenient formula

$$V_{\text{swap}}(t) = A(t)\,(S(t) - k)\,. \tag{34}$$

The quantity $A(\cdot)$ is the *annuity* of the swap (or its $PVBP$, for Present Value of a Basis Point), and the quantity $S(t)$ is the *forward swap rate.* Clearly, $S(t)$ is the value of the fixed rate that makes the swap have value 0 to both parties at time t; S is consequently often referred to as a *par* or *break-even* rate.

3.16 The Libor Reform

See https://ssrn.com/abstract=3684535 and references there for more details.

Libor rates, first introduced in 1980s, over time became the benchmark rates for 100's of trillions of notional of interest rate derivatives. During and post GFC it became clear that they are open to manipulation. Recall that a Libor rate (one for each of 7 tenors for 5 currencies) is determined daily by a poll of participating banks on their estimate of the rate at which they can borrow for a given term in the interbank money market. It came to light in the early 2010s that during GFC, some banks were submitting estimates that were intentionally, and materially, lower than their true cost of borrowing to signal to market that they were not in financial difficulties.

It further became known that banks were making Libor submissions that directly benefited their trading positions. For example if they were "long" the next fixing on their derivatives (the simplest example being a portfolio with a single swap paying fixed, receiving Libor, with the next Libor payment fixing on the day), they would submit a higher estimate for Libor calculation thus influencing the average. Moreover, groups of banks (so-called "cartels") were coordinating their submissions to mutual benefit. These revelations lead to Libor scandals of 2012 and beyond.

After the scandals the rules for Libor submissions were considerably tightened by the regulators. Chinese walls were erected, and Libor submission methodology had to be developed and validated independently, leaving little room, in theory, to submitters "judgement".

Incidentally, this codifying of Libor submission methodologies lead to a new trading strategy by some hedge funds who reverse-engineered banks' methodologies and used that to try to predict upcoming Libor fixings to trade them.

The problem with this solution was that the regulators wanted banks to use rates on actual loans in the interbank cash market. This market however by 2010s had very low liquidity, in low \$ billions, compared to the market that the Libor rates were driving (essentially all interest rate derivatives) in the 100s of \$ trillions. Eventually, in 2017, FCA (regulator responsible for Libor) made a bold move and announced that Libor rates must be phased out by 2021 and replaced by *representative* benchmark rates.

Much industry and regulatory activity ensued (as well as hedge fund activity in predicting, and betting on, the parameters of the future alternative benchmarks). Consensus eventually coalesced on using the so-called Risk-Free Rates (RFR): rates with very short, daily, tenors that were credit-risk free (not because their are short-tenor, but because they are based on transactions secured by collateral – see the earlier course discussion about collateral).

3.16.1 Legacy Overnight Index Rates

Overnight rates have existed long before the Libor reform, alongside Libor rates. In the United States, banks are required to hold certain balances ("Federal funds") with the Federal Reserve, the central bank of the US. If a bank does not have sufficient balances, it can borrow them from another bank that has an excess on its account. The overnight interest rate charged in this case is called the *(effective) Federal funds rate*[4]. Instruments linked to averages of the (effective) Fed fund rate over different terms were (and still are) actively traded, giving rise to a term structure of Fed funds linked rates. The Euro and GBP markets did not have the same mechanism as the US did/does for Federal funds, but overnight rates in those currencies also existed. They are called *Eonia* (Euro Overnight Index Average) in the Eurozone and *Sonia* (Sterling Overnight Index Average) in Great Britain, and are computed as averages of all *actual* overnight lending/borrowing transactions by qualifying banks weighted by the size of the transactions. We emphasize that these rates reflect the actual transactions that have happened, in contrast to Libor which reflect banks' estimates of rates at which borrowing (for a given term) might take place. The rates were the underlying of the *overnight index swaps*, or OIS, of various maturities.

3.16.2 New Overnight Index Rates

While the OIS rates and trading in OIS swaps existed pre-Libor reform, not all of them were deemed representative enough to serve as *the* benchmarks for the whole of interest rate market. New RFRs have been developed and by now adopted in all major economies.

- the US selected a new Treasuries repo financing rate called SOFR (Secured Overnight Funding Rate)
- the UK selected the reformed Sonia (Sterling Overnight Index Average)
- the Euro zone selected a new unsecured overnight rate called ESTR (Euro Short-Term Rate)
- Also Switzerland selected SARON (Swiss Average Rate Overnight) and Japan selected TONA (Tokyo Overnight Average Rate)

In the UK, Sonia completely replaced GBP-Libor as the benchmark – all new trades are linked to Sonia and legacy Libor trades were replaced by (somewhat)

[4]The target rate, set by the Federal reserve, is aptly called the *target* Fed funds rate, or sometimes simply the Fed funds rate

equivalent Sonia trades. In the US, USD-Libor has not yet been discontinued (slated for June 2023) and only new trades that are allowed for USD-Libor are for hedging purposes only. The liquidity in SOFR, however, has been growing steadily. In the Eurozone, EURIBOR has been reformed and is not going away any time soon, although ESTR liquidity has also been steadily growing.

3.17 RFR-Linked Instruments

3.17.1 Daily-Compounded RFR

Pre-Libor reform, OIS instruments such as swaps, even if linked to a daily (overnight, or ON) rate, exchanged payments on a much coarser term structure, 3M or 6M or 1Y depending on currency conventions. This has always been the case to simplify settlements and payments. ON rates can be converted to a term rate that determines a payment for a 3M period (say) in a variety of different ways. Arguably the most natural is compounding. It corresponds to the idea of rolling the initial deposit of \$1 daily for 3 months; each day the deposit grows at the then-daily rate.

In mathematical notations, an ON rate is (somewhat idealistically) identified with the short rate $r(t)$. For a tenor structure $0 \leq T_0 < T_1 < \ldots < T_N$, the *daily-compounding setting-in-arrears* rate for the period $[T_n, T_{n+1})$ is defined as

$$R(T_n, T_{n+1}) \triangleq \frac{1}{\tau_n} \left(\prod_{i=1}^{k} (1 + r(t_i)\delta_i) - 1 \right),$$

where the product is over the business days t_i in $[T_n, T_{n+1})$, $r(t_i)$ is the daily fixing of the RFR (same as short rate in our math notations) at time t_i, and δ_i is the associated day-count fraction (roughly 1/252). The *in-arrears* part of the RFR name specifies that the observation period aligns with the interest period – this is a rather technical point that we will not dwell over.

For modelling purposes, intuition, and notational clarity, it is often convenient to use an approximation formula

$$R(T_n, T_{n+1}) = \frac{1}{\tau_n} \left(e^{\int_{T_n}^{T_{n+1}} r(u)\ du} - 1 \right)$$

which can also be written in terms of the continuously compounded money market account $\beta(\cdot)$ as

$$R(T_n, T_{n+1}) = \frac{1}{\tau_n} \left(\frac{\beta(T_{n+1})}{\beta(T_n)} - 1 \right). \tag{35}$$

As forward Libor rates are expected values of spot Libor rates in the appropriate forward measure, see (30), it is natural to define a time$-t$ forward term daily-compounded RFR, for $t \leq T_n$, by

$$R(t, T_n, T_{n+1}) \triangleq \mathrm{E}_t^{T_{n+1}} R(T_n, T_{n+1}). \tag{36}$$

It turns out that

$$R(t, T_n, T_{n+1}) = \frac{1}{\tau_n}\left(\frac{P(t, T_n)}{P(t, T_{n+1})} - 1\right) = L(t, T_n, T_{n+1}), \quad t < T_n, \tag{37}$$

so that the forward term daily-compounded rate is equal to the simply compounded forward rate. This justifies us discussing Libor-linked instruments earlier despite Libor demise – much of the machinery developed for Libor rates can be applied to term RFR rates with little or no modification. Another useful, for modelling, formula expresses R in terms of the instantaneous forward rates,

$$R(t, T_n, T_{n+1}) = \frac{1}{\tau_n}\left(e^{\int_{T_n}^{T_{n+1}} f(t,u)\ du} - 1\right).$$

However, it is worth pointing out one critical difference. Unlike for Libor,

$$R(T_n, T_n, T_{n+1}) \neq R(T_n, T_{n+1}),$$

which can be readily observed by comparing (35) and (37). The *in-arrears compounded* rate $R(T_n, T_{n+1})$ is only known at T_{n+1}, when all the daily fixings of the ON rate have been observed. The *term compounded* T_n-forward rate $R(T_n, T_n, T_{n+1})$ is the expected value, observed at time T_n, of the rate calculated by compounding future daily fixings.

In fact, for RFRs we have a different relation. Note that here $t = T_{n+1}$ in the rate on the left-hand side:

$$R(T_{n+1}, T_n, T_{n+1}) = R(T_n, T_{n+1})$$

where we extended (36) for $t \in [T_n, T_{n+1})$. For such t, (37) no longer holds; in particular we have

$$R(t, T_n, T_{n+1}) \triangleq \mathrm{E}_t^{T_{n+1}} R(T_n, T_{n+1}) = \frac{1}{\tau_n}\left(\frac{\beta(t)}{\beta(T_n) P(t, T_{n+1})} - 1\right), \quad t \in [T_n, T_{n+1}).$$

(HW: Prove this).

We can combine different formulas for $R(t, T_n, T_{n+1})$ for the three different regimes $t < T_{n+1}$; $T_n \leq t < T_{n+1}$; $T_{n+1} < t$ if we adopt the convention that

$$P(t, T) = 1 \text{ for } t > T.$$

Then the combined formula that works for any $t \geq 0$ is given by

$$R(t, T_n, T_{n+1}) = \frac{1}{\tau_n}\left(\frac{\beta(t)}{\beta(\min(t, T_n))} \times \frac{P(t, T_n)}{P(t, T_{n+1})} - 1\right), \quad \text{any } t > 0. \tag{38}$$

3.17.2 Fixed-Floating OIS Swaps

In a fixed-floating OIS swap for the tenor structure $0 \leq T_0 < T_1 < T_2 < \ldots < T_N, \quad \tau_n = T_{n+1} - T_n$.with fixed rate k, one party (the fixed rate payer) pays simple interest based on the rate k in return for simple interest payments based on the daily compounded RFR rate $R(T_n, T_{n+1})$ for each period $[T_n, T_{n+1}]$, $n = 0, \ldots, N-1$. The payments are exchanged at the end of each period, i.e. at time T_{n+1}. Note that this is in agreement with Libor swaps and also guarantees that all the daily RFR fixings going into $R(T_n, T_{n+1})$ calculation have been observed. From the perspective of the fixed rate payer, the net cash flow of the swap at time T_{n+1} is therefore given, on a unit notional, by $\tau_n \left(R(T_n, T_{n+1}) - k\right)$.

The value of the swap (to the fixed-rate payer) at time t before the start of the first observation period $0 \leq t \leq T_0$, is given by

$$\begin{aligned} V_{\text{swap}}(t) &= \beta(t) \sum_{n=0}^{N-1} \tau_n \mathrm{E}_t \left(\beta\left(T_{n+1}\right)^{-1} \left(R(T_n, T_{n+1}) - k\right) \right) \\ &= \sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) \mathrm{E}_t^{T_{n+1}} \left(\left(R(T_n, T_{n+1}) - k\right) \right) \\ &= \sum_{n=0}^{N-1} \tau_n P(t, T_{n+1}) \left(R(t, T_n, T_{n+1}) - k\right), \end{aligned}$$

where we switched to the appropriate forward measure for each term and used (36). We note that the expression for the value of the swap exactly matches that for a Libor-based swap. This, clearly, would not be the case for t's in the middle of an observation period.

3.17.3 OIS FRAs

Unlike swaps, RFR FRAs work slightly differently than Libor FRAs. Recall that a Libor RFR pays at the time of the Libor fixing, i.e. at the beginning of the observation period. This is not possible for RFR as the term compounded rate is only known at the end of the period. Hence an RFR FRA pays at the end. For valuation purposes, however, the two can use the same formulas before the start of the period.

3.17.4 OIS Futures

Futures on in-arrears compounded RFRs started trading on various exchanges (e.g. CME) in the run up to Libor cessation. They are similar to FRAs (albeit exchange-traded), so the same considerations apply here. One small difference is that 1M futures settle into a weighted average, not a compounded, of overnight rates. 3M versions compound 1M average rates. While important for actual valuation, the difference between averaging and compounding is immaterial for conceptual understanding, They settle into an average RFR on, or actually just after, the last observation date of the corresponding period; again this is in contrast to Libor futures that settle at the beginning (fixing date) of the Libor rate. Common examples here include 30D Fed Fund futures in USD, 1M and 3M SOFR futures in USD, and 1M Sonia futures in GBP.

3.17.5 OIS Deposits

As RFR rates are 1D, there are only deposits for 1D, with the obvious formulas linking the two. The market is still in flux and there could be deposits linked to term RFR rates which are, in our notations, $R(0, 0, \tau)$ for various τ such as 3M or 6M or 1Y depending on currency conventions. Some market data providers are publishing these rates which are implied from other market data inputs such as RFR futures. It is still, at the time of writing, unclear whether these will be used in anger, both from the client perspective and the regulatory perspective.

3.17.6 Instruments Linked to Central Bank Meeting Dates

We have already commented on the central role of central banks in Fixed Income markets. CB actions determine the short-end rates more or less directly. Let us consider the UK market as an example. The Bank of England's (CB of UK) Monetary Policy Committee (MPC) is responsible for making decisions about the Bank Rate. MPC meets 8 times a year, and at each meeting they essentially set the Sonia ON rate for the period from that date until the next MPC date. They could either lower the rate (as repeatedly happened at the onset of the pandemic), keep it the same (quite often), or increase it, as has been happening recently (from Q3 2021) to fight inflation, the main mandate for BoE.

Since the MPC decisions (and the equivalents from ECB and the Fed) are so fundamental to the interest rates, it is no surprise that a market in instruments that allow direct exposure to these decisions has developed. These are known, in the UK, as MPC FRAs. They are FRAs on the compounded in-arrears Sonia rate that cover periods between MPC dates.

Let M_1 be the next MPC date, and M_n, $n > 1$, be the subsequent MPC dates. At the next meeting date, MPC will essentially tell the market what $r(M_1)$ is set to. It is generally expected, barring extraordinary circumstances, that the Sonia rate fixings will stay at, or very close, to the rate fixed on M_1 until the next MPC date,

$$r(t) = r(M_1),\ t \in [M_1, M_2).$$

Hence, the FRA rate for the period $[M_1, M_2)$ is expected to be equal to whatever the MPC declares to be the value of $r(M_1)$,

$$R(M_1, M_2) = r(M_1).$$

Therefore, entering a FRA covering MPC dates $[M_1, M_2)$ is essentially betting on what MPC is going to set the policy rate to at their meeting date M_1. The market in MPC FRAs is liquid and very active especially since Q3 2021 through 2022 during an active hiking cycle.

3.18 Bonds, Quick Into

Government bond markets (Treasuries in the US; EGBs in Eurozone, Gilts in the UK) are some of the largest and most liquid markets in the world. Governments issue bonds to raise money for its expenditures on a periodic schedule, covering a range of maturities from a couple of weeks to thirty years (in the US), and lately even longer to 50 (Gilts) and, sometimes, 100 years (Austria, Argentina).

Depending on the country of origin and maturity, government bonds go by different names. In the US, there are T-bills, T-notes and T-bonds. In the UK all bonds are called gilts. Germany bonds (key benchmarks for Eurozone) are called Shatz, Bobls and Bunds, etc.

A bond is a security that pays a coupon on a regular schedule such as every 6 months, and repay the notional at its maturity. They are usually priced at par at issuance, i.e. one pays \$100 to receive a stream of coupons and a repayment of \$100 at maturity. As interest rates change, a bond can trade above or below par. Bonds are described in their annual coupon, and the maturity date. For an annual coupon c and coupon schedule $\{T_i\}_{i=1}^N$ (and hence maturity T_N) the price at time t (for a unit notional) is given by

$$
\begin{aligned}
V_{\text{bond}}(t) &= \beta(t)\mathrm{E}_t\left(\sum_{T_i>t}\beta(T_i)^{-1}\times c/f + \beta(T_N)^{-1}\times 1\right) \qquad (39)\\
&= \sum_{T_i>t}(c/f)\,P_{\text{gov}}(t,T_i) + P_{\text{gov}}(t,T_N),
\end{aligned}
$$

where $P_{\text{gov}}(t,\cdot)$ are the zero-coupon discount bonds/factors suitable for discounting payments made by the government, and f is the number of coupons in one year (e.g. $f=2$ for semi-annual coupon payments).

Bonds are quoted in price terms, often as "clean" prices, as opposed to "dirty" prices. A dirty price is the PV of future expected cashflows as in (39). A clean price strips out the accrued value of the next coupon, when one is in the middle of a coupon period. Details can be found in Andersen & Piterbarg (2010).

3.19 On Idiosyncrasy of Bonds

On any given day, a large number of different bonds from the same family (such as Treasuries) are available for trading. The most recently issued bonds are called "on-the-run", while the rest – the old issuance – are called "off-the-run". On-the-runs are typically the most liquid and the easiest to trade. Other, seemingly almost equivalent, bonds may trade at prices that are not entirely consistent with the on-the-runs for a variety of reasons mostly having to do with liquidity and availability of certain bonds. Unlike OTC swaps, that are fully fungible – a 10y swap is the same instrument no matter who traded it – a 30y bond issued 20y ago is significantly different from a freshly issued 10y note in terms of how it is traded in the market.

Because of this idiosyncrasy, it is not possible, as a rule, to fit all bonds with a unique discount curve as in (39). Market prices of certain hard-to-find bonds may not be in sync with the on-the-runs. Some bonds may be in special demand[5] ("trading on special") so that their prices are inconsistent with the others. There are a number of implications to this fundamental feature of the bond market that we explore later.

3.20 Measuring Bonds

Using prices to compare different bonds is not always useful, as bonds with different annual coupons are hard to compare in price terms. To correct for that, it is conventional to express bond values in terms of their YTM, or *yield-to-maturity*. A bond YTM is a single rate that, when applied to all coupons and the notional, recovers its (dirty) price. It serves as a (rough) indication of the rate of return on the investment into the bond if held to maturity. It is a rough equivalent of a swap rate for swaps we introduced earlier, but defined in a different way (bond markets existed way before the swap markets so the conventions are different, if somewhat antiquated).

Specifically, given the market price $V_{\text{bond}}(t)$ (for unit notional!) the *coupon-bearing bond yield* $y_{\text{bond}} = y_{\text{bond}}(t)$ is defined by the relation

$$V_{\text{bond}}(t) = \sum_{T_i > t} \frac{c/f}{\left(1 + y_{\text{bond}}/f\right)^i} + \frac{1}{\left(1 + y_{\text{bond}}/f\right)^N}. \tag{40}$$

Of course in practice we need to be careful with day counting rules, date adjustments etc.

Government bonds reflect the credit of the issuing government and as such trade at somewhat different interest rates than OTC instruments such as swaps. Benchmark swaps are often quoted as a spread off a corresponding government bond, e.g. a 10y USD swap would often be quoted as a spread to the YTM on a 10 year Treasury note.

Bonds also serve as an underlying for various OTC derivatives. The most common, arguably, is an *asset swap*.

3.21 Asset swaps and ASW

In an asset swap one exchanges the coupons of a government bond for floating payments based on a benchmark rate such as compounded Sonia (historically, a relevant Libor rate), plus a spread. The spread, called the asset swap spread (ASW), is another common measure of value of a bond, and expressed bond values as an extra yield/spread over and above (or below, as the case might be) of a par swap of the same maturity. If $V_{\text{bond}}(t)$ is the (dirty) price of a given

[5]Special demand could be from the fact that this particular bond is the cheapest to deliver on a bond futures contract, a subject we do not have time to explore. See Andersen & Piterbarg (2010) for details.

bond with coupon c and maturity T_N, then the *asset swap spread* s_{asw} is defined so that the swap that

- Pays bond coupon c on the bond schedule $\{T_i\}_{i=1}^N$;
- Receives $R(T_i, T_{i+1}) + s_{\text{asw}}$ on the same (or different-frequency but same final maturity, depending on the currency conventions for a swap) schedule $\{T_i\}_{i=1}^N$, where $R(T_i, T_{i+1})$ is the benchmark rate for the period $[T_i, T_{i+1}]$;
- Compensates for the bond dirty price not equal to par via various mechanisms;

has value 0 at the time of entering the asset swap.

3.21.1 Technical diversion about asset swaps

There are various flavours and subtleties having to do with $V_{\text{bond}}(t)$ not being equal to par normally etc. Typically there are three types of asset swaps:

- *Par/Par asset swap.* The notional of the floating leg is equal to the nominal of the bond (i.e. 1 for unit nominal of the bond). There is an initial payment to enter into this type of swap equal to $1 - V_{\text{bond}}(t)$, i.e. this asset swap does not have value of 0 at inception.
- *Market asset swap.* Here the notional of the *floating leg* is adjusted for the dirty price of the swap and is set to be equal to the dirty price $V_{\text{bond}}(t)$. This type of swap has value 0 at inception i.e. no initial payment is required.
- *Matched maturity swap.* Technically not an asset swap as it is just a market-rate (i.e. swap rate) receiving fixed for floating swap of the same maturity as the bond.

For example, you can see the function `Bond.asset_swap_spread()` in FinancePy:

- *par/par asset swap* compensates for the dirty bond price not being equal to par
- `asw = (pvIbor - bondPrice / par) / pv01` where `pvIbor` is the value of coupons/redemption of a bond off a Libor (or, these days, an RFR) curve, `bondPrice` is the dirty price, `par` is par (bond notional), `pv01` is the pv01 of the floating leg
- Formula for ASW is thus

$$\sum_{T_i > t} (c/f)\, P(t, T_i) + P(t, T_N) - V_{\text{bond}}(t) = -1 + P(t, T_N) + \sum_{T_i > t} \tau_{i-1} P(t, T_i)\, (R(T_{i-1}, T_i) + s_{\text{asw}}) \quad (41)$$

which, after some simplifications, reads

$$V_{\text{bond}}(t) - 1 = \sum_{T_i > t} (c/f)\, P(t, T_i) - \sum_{T_i > t} \tau_{i-1} P(t, T_i)\, (R(T_{i-1}, T_i) + s_{\text{asw}})\,.$$

(End of the technical diversion)

The *asset swap spread* s_{asw} measures the extra return (that could be negative) one gets by buying the bond vs. entering a standard fixed-for-floating swap paying bond coupons on the fixed side of the same maturity. Very roughly, the asset swap spread is equal to the bond YTM minus the par swap rate for the swap of the same maturity as the bond.

Asset swaps are commonly used to express and trade views on the relative value between bonds and swaps.

Expressing bond value as a spread over the benchmark swaps has an advantage over the outright YTM measure as it splits the latter into a contribution from the overall level of interest rates in a particular currency (as measured by the relevant par rate on the standard swap) vs. the extra difference specific to a particular bond.

3.22 Z-spreads

Another, somewhat similar, measure that is even more convenient when comparing different bonds is the so-called *z-spread*.

The z-spread z_{bond} is defined as a constant such that, when used on top of the standard (these days RFR; historically, Libor) discounting, recovers the bond's market price. Specifically, let $P(t, T)$, $T \geq t$, be the OIS discount curve (e.g. Sonia or SOFR). Then the z-spread z_{bond} is given by the equation

$$V_{\text{bond}}(t) = \sum_{T_i > t} (c/f)\, e^{-z_{\text{bond}}(T_i - t)} P(t, T_i) + e^{-z_{\text{bond}}(T_N - t)} P(t, T_N) \tag{42}$$

$$= \sum_{T_i > t} (c/f)\, e^{-\int_t^{T_i} (f(t,u) + z_{\text{bond}}) du} + e^{-\int_t^{T_N} (f(t,u) + z_{\text{bond}}) du}. \tag{43}$$

Here, as we recall, $f(t, u)$, $u \geq t$, are the instantaneous forward rates observed at time t.

One can think of the z-spread as a credit spread, and clearly the two are closely related. However, the z-spread captures additional characteristics of a bond such as its liquidity, “specialness’, etc. If one's goal is to really imply market probabilities of a sovereign default from bond prices, these extra factors should be accounted for. Specifically, it is not uncommon for bonds to trade at a basis (i.e. a difference) to sovereign CDSs, the latter being more “pure” instruments to extract probabilities of default from. We do not pursue this line of inquiry here.

Topic 4

Building Libor/OIS Rate Curves

4.23 Basics of Curve Building

4.23.1 Motivation and Basic Setup

An interest rate curve, or a *discount curve* observed at time 0 is a mapping between time T and discount factors $P(T) = P(0,T)$. It is fundamental to interest rate modelling and trading. We start with the classical construction of a single (Libor) discount curve from observed swap, deposits, and Eurodollar futures. As we explained previously, the ideas extend fairly seamlessly to OIS curves.

The objective of building a discount curve is to match the prices $V_i = V_i(0)$, $i = 1, 2, \ldots, N$, of N linear *benchmark securities*, all of which are assumed to be of the form (many $c_{i,j}$ are zero)

$$V_i = \sum_{j=1}^{M} c_{i,j} P(t_j), \quad i = 1, \ldots, N.$$

A standard (idealized) swaps paying a coupon $c\tau$ at times τ, 2τ, 3τ,..., $n\tau$ can be written this way, since

$$V_{\text{swap}} = 1 - P(t_n) - \sum_{j=1}^{n} c\tau P(j\tau) \quad \Rightarrow \quad 1 - V_{\text{swap}} = P(t_n) + \sum_{j=1}^{n} c\tau P(j\tau).$$

FRAs and deposits also fit this standard form.

Therefore the problem of curve construction can be stated as estimating $\mathbf{P} = (P(t_1), \ldots, P(t_M))^\top$ as

$$\mathbf{V} = \mathbf{c}\mathbf{P} \tag{44}$$

where $\mathbf{V} = (V_1, \ldots, V_N)^\top$ and $\mathbf{c} = \{c_{i,j}\}$ is an $(N \times M)$-dimensional coupon matrix.

In its basic formulation the curve construction problem appears to be rather simple; however, each active market participant, especially dealers and hedge funds, spends an enormous amount of resources on refining and improving curve building methods and adapting them to ever-changing market developments.

Benchmark securities are usually nicely staggered i.e. we can naturally order them in the increasing order of their *last* maturity dates T_n, the last date for which the discount factor is needed to value the n-th benchmark security. Throughout, we assume that we can select and arrange our benchmark set of securities to guarantee that the maturities of the benchmark securities satisfy

$$T_i > T_{i-1}, \quad i = 2, 3, \ldots, N, \tag{45}$$

where the inequality is strict. Equation (45) constitutes a "spanning" condition and allows us to select the N maturities as distinct knots in our yield curve splines.

On the other hand, typically $M > N$, so there are fewer benchmark securities than the number of dates for which we need discount factors to value them, and the linear system (44) is under-determined. In practical terms that means that there is an infinity of discount curves that would fit all benchmarks.

4.23.2 Linear Swap Rate Bootstrap

We will discuss more modern methods of regularizing (44) (i.e. imposing additional constraints to choose the "best" solution out of the infinity of options) shortly. But for building intuition it is worth reviewing probably *the* earliest curve fitting algorithm which is the linear swap rate bootstrap.

Recall that we denoted by $\{T_n\}$ the collection of final maturity dates and by $\{t_j\}$ the payment dates of the coupons in underlying instruments, i.e. all the dates for which we need discount factors for to be able to value benchmark securities.. It is common that the coupon payment dates for different swaps in the benchmark set are shared, i.e. for the benchmark swap n, the coupon dates are all $\{t_i : t_i \leq T_n\}$. The linear swap rate bootstrap utilizes this observation by first assigning "market" values to all swaps V_j', $j = 1, \ldots, M$, where swap V_j' has the last payment date t_j and coupon payment dates $\{t_i, i \leq j\}$. This is done by linear interpolation of the actual benchmark swap breakeven rates. To assign a (fake) market breakeven rate to swap V_j', we first determine n such that $T_{n-1} < t_j \leq T_n$, and then set i_i and i_2 by $t_{i_1} = T_{n-1}$, $t_{i_2} = T_n$. Let the market breakeven rate for swap V_i' be given by c_i. Then we set c_j by linearly interpolating (actual) market-observed breakeven rates c_{i_1} and c_{i_2}:

$$c_j = c_{i_1} \frac{T_n - t_j}{T_n - T_{n-1}} + c_{i_2} \frac{t_j - T_{n-1}}{T_n - T_{n-1}}, \quad j = 1, \ldots, M.$$

Once all $\{c_j\}_{j=1}^M$ are thus determined, the linear system (44) has as many equations as there are unknowns and could be solved in closed form. In fact it is solved sequentially for each $P(t_j)$ from V_j' and c_j, hence the name "bootstrap".

We have glossed over a number of details on how to deal with deposits and FRAs, and how to deal with dates that do not quite line up. This method does rely on a particular structure in the benchmark set, but it is a structure that was typical in the early days of swaps markets. In fact this classic bootstrap can be taken quite far and deal with less idealized, more realistic benchmark sets. While every bank probably has an implementation of this classic method, over time it has been supplanted by more modern approaches.

A demonstration notebook is mentioned on BlackBoard and is available in `FinancePy` library at `./notebooks/products/rates/FINIBORSINGLECURVE_BuildingIborCurveInterpolationSIMPLE.ipynb`.

4.23.3 Curve Bootstrap with Interpolation Control

Let us now develop the tools needed for a slightly more modern version of the classic bootstrap. Before we start, let us note that a discount curve needs to value interest rate instruments that require discount factors that do not necessarily fall on the discount factor dates $t_1, \ldots, t_M$. Hence, we need to couple (44) with an interpolation scheme, i.e. how to determine $P(t)$ from $(P(t_1), \ldots, P(t_M))^\top$ for $t \notin \{t_1, \ldots, t_M\}$. In fact, as will be clear momentarily, we need to specify the interpolation scheme even earlier, in order for the curve building to work. Hence let us assume that we have selected an interpolation scheme that, given $P(T_1), \ldots, P(T_k)$, can return a discount factor $P(t)$ for any $t \leq T_k$.

The basic idea of bootstrap is encapsulated in the following iteration:

1. Let $P(t_j)$ be known for $t_j \leq T_{i-1}$, such that prices for benchmark securities $1, \ldots, i-1$ are matched.
2. Make a guess for $P(T_i)$.
3. Use an interpolation rule to fill in $P(t_j)$, $T_{i-1} < t_j < T_i$.
4. Compute V_i from the now-known values of $P(t_j)$, $t_j \leq T_i$.
5. If V_i equals the value observed in the market, stop. Otherwise return to Step 2.
6. If $i < N$, set $i = i + 1$ and repeat.

The updating of guesses when iterating over Steps 2 through 5 can be handled by any standard one-dimensional root-search algorithm (e.g., the Newton-Raphson or secant methods).

As mentioned if the interpolation is particularly simple, e.g. flat ON forward rate between $[T_{n-1}, T_n]$ (see details later) the process of recovering $P(T_n)$ from V_n can be carried out analytically without a solver.

A demonstration notebook is mentioned on BlackBoard and is available in `FinancePy` library at `./notebooks/products/rates/FINIBORSINGLECURVE_BuildingIborCurveInterpolationCOMPLEX.ipynb`.

In general, there are strong limitations on what kind of interpolation rule can be applied in Step 3. The main requirement is locality, meaning that as we move $P(T_n)$ around and interpolate the curve for $t \in [0, T_n]$, the part of the curve that is already fixed, namely $t \in [0, T_{n-1}]$, must not change. Again we will see in a bit that many (in fact most) interpolation scheme do not have this property. Other considerations are more subtle. For instance, one might consider using a representation in terms of instantaneous forwards $f(T)$, with the assumption that $f(T)$ is a continuous piecewise linear function on the maturity grid $\{T_i\}_{i=1}^N$. While based on seemingly natural assumptions, this interpolation rule can, however, be shown to be a particularly *bad* choice, as it is numerically unstable and prone to oscillations (we will see this later). Some stable, and standard, choices for interpolation rules are covered later; common for both is that the resulting yield curve is continuous, but non-differentiable. This, in turn, implies that the instantaneous forward curve is discontinuous.

Some interpolation schemes are less local that just defined but are still local, i.e. $P(t)$ for $t \in [T_{n-1}, T_n]$ depend on $P(T_n)$ *and* $P(T_{n+1})$ but not beyond T_{n+1}. Then a suitable modification of the bootstrap algorithm can still be devised.

The impact of different interpolation schemes is shown in `T4_Curve_Ringing_01.ipynb` linked on BlackBoard.

4.24 Parametrization for Interpolation

To interpolate a curve we need to specify which values we should actually interpolate. Interpolating discount factors directly is by far not the best choice (and is not normally done, as far as I am aware). There are quite significant restrictions on $P(\cdot)$ such as being being 0 and 1, (typically) monotonically decreasing, etc. that are just hard to enforce in an interpolation. Luckily there are many equivalent ways to parametrize a discount curve, and different choices lead to different underlying functions to interpolate. Let us list some of the common choices.

4.24.1 Instantaneous (ON) Forward Rates

Recall that $f(t) = f(0, t)$ is defined by

$$f(t) = -\partial \ln P(t)/\partial t$$

so that

$$P(t) = \exp\left(-\int_0^t f(u)\ du\right).$$

Hence one can recover discount factors $P(\cdot)$ from the overnight (ON) forward rates $f(\cdot)$ and the latter is a much more suitable object for interpolation – as there are few, if any, economic restrictions on the values of $f(\cdot)$. (It used to be that one would require $f(\cdot) \geq 0$ but even that stopped being a restriction in the last decade).

The downside of this parametrization is that calculating discount factors from ON forward rates requires integration, which could be somewhat slow.

4.24.2 Yields or Zero Rates

Another popular choice for interpolation is the *yield curve* $y(T)$, defined by the one-to-one transformation $e^{-y(T)T} = P(T)$. The yield curve is flatter than the discount curve, and easier to interpolate. Sometimes $y(T)$ is called a zero rate to T. Clearly also

$$y(T) = \frac{1}{T}\int_0^T f(u)\ du$$

so a zero rate is an average of ON forward rates between today and the maturity date T.

4.24.2.1 Log-Discount Factors

Log-discount factor, as the name implies, is defined by

$$l(T) = \ln P(T).$$

Clearly,

$$l(T) = y(T)T.$$

The downside of using $l(\cdot)$ is that the scales of $l(t)$'s are quite different for significantly different t's such as say 1y and 50y. So this may look like an odd choice, but using log discount factors for interpolation (coupled with cubic splines, see below) was chosen by LCH as their method for constructing their own curves, and all banks who clear on LCH need to be able to replicate LCH methodology exactly to understand the margin calls from LCH. This of course does not mean that this is the curve used for active pricing or risk management; so this is yet another demonstration why multiple curves using different methodologies from the exact same market data are often built by sophisticated market participants.

4.24.3 Term Forward Rates

Half-way between ON (instantaneous) forward rates $f(t)$ and zero rates $y(t)$ are term forward rates, i.e. forward rates for periods longer than 1 day. While it might be tempting to use, say 3M term forward rates fixing at each t as the parametrization (this was particularly relevant when the underlying main rates for 3M Libor curve were 3M rates), a moment's reflection reveals that the discount curve $P(\cdot)$ cannot, in general, be recovered from the $\tau-$tenor forward rate curve

$$\{y(T, T+\tau),\ T \geq 0\}$$

where $y(T, T+\tau) = y(0, T, T+\tau),$. for any τ that is not one day (i.e. ON forward rates).

Yet forward rates offer advantages over zero rates as they isolate specific parts of the discount curve better than zero rates:

$$y(T, T+\tau) = \frac{1}{\tau}\int_T^{T+\tau} f(u)\ du.$$

Hence it is easier to impose conditions that control the relationships between different parts of the discount curve, which is important for devising successful RV strategies (for HFs) or prevent being picked off (for dealers). More on this later.

Given a set of benchmark maturities $\{T_n\}_{n=1}^N$, and setting $T_0 = 0$, a convenient collection of term forward rates that fully and uniquely determines the discount curve $P(\cdot)$ is given by

$$\left\{y(T_{\lfloor t\rfloor}, t),\ t \geq 0\right\}$$

where we define

$$\lfloor t \rfloor \triangleq \max_{n=0,\dots,N} \{n : T_n < t\} \,.$$

In essence, for each t, the term forward rate $y(T_{\lfloor t \rfloor}, t)$ covers the period from the last benchmark maturity date before t, i.e. $T_{\lfloor t \rfloor}$, to t.

4.25 Yield Curve Fitting with N-Knot Splines

In this section we discuss a number of well-known yield curve algorithms based on polynomial and exponential (tension) splines of various degrees of differentiability.

4.25.1 Matrix Formulation

Define the M-dimensional discount bond vector[6]

$$\mathbf{P} = (P(t_1), \dots, P(t_M))^\top \,,$$

and let $\mathbf{V} = (V_1, \dots, V_N)^\top$ be the vector of observable security prices. Also let $\mathbf{c} = \{c_{i,j}\}$ be an $(N \times M)$-dimensional matrix containing all the cash flows produced by the chosen set of securities. The matrix $\mathbf{c}$ would typically be quite sparse, with many rows containing only a few non-zero entries. A typical, albeit simplified, form of the matrix $\mathbf{c}$ might be ($\times$ marks a non-zero element)

$$\mathbf{c} = \begin{pmatrix}
\times & & & & & & & & & & & & & & \\
& \times & & & & & & & & & & & & & \\
& & \times & \times & & & & & & & & & & & \\
& & & \times & \times & & & & & & & & & & \\
& & & & \times & \times & & & & & & & & & \\
& & & & & \times & \times & & & & & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & & & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & & & & \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times & \times
\end{pmatrix},$$

corresponding to two certificates of deposit (first two rows); four FRAs or Eurodollar futures (next four rows); and three swaps (last three rows).

In a consistent, friction-free market without arbitrage opportunities, the fundamental relation

$$\mathbf{V} = \mathbf{c}\mathbf{P} \tag{46}$$

must be satisfied, giving us a starting point to determine $\mathbf{P}$.

[6]For extra clarity, throughout this chapter we use boldface type for vectors and matrices.

4.26 C^0 Yield Curves

4.26.1 Piecewise Linear Yields

The most common discount curve bootstrap algorithm assumes that the continuously compounded yield $y(T)$ is a continuous piecewise linear function on $\{T_i\}_{i=1}^N$. Formally, the interpolation rule in Step 3 of the algorithm in Section 4.26 writes $P(T) = e^{-y(T)T}$, where

$$y(T) = y(T_i)\frac{T_{i+1} - T}{T_{i+1} - T_i} + y(T_{i+1})\frac{T - T_i}{T_{i+1} - T_i}, \quad T \in [T_i, T_{i+1}]. \tag{47}$$

To initiate the iterative bootstrap algorithm, we note that the interpolation rule (47) may require us to provide an equation for $y(t)$, $t < T_1$. There are a number of ways to do this; one common choice is to simply set $y(t) = y(T_1)$, $t < T_1$.

To give a feel for the types of yield curves produced by linear yield bootstrapping, let us consider a simple example with a benchmark set of $N = 10$ swaps, with maturities and quoted par swap rates as given in Table 1[7].

Maturity (Years)	Swap Par Rate
1	4.20%
2	4.30%
3	4.70%
5	5.40%
7	5.70%
10	6.00%
12	6.10%
15	5.90%
20	5.60%
25	5.55%

Table 1: Swap Benchmark Set for Numerical Tests

The swaps are assumed to pay on a semi-annual basis,

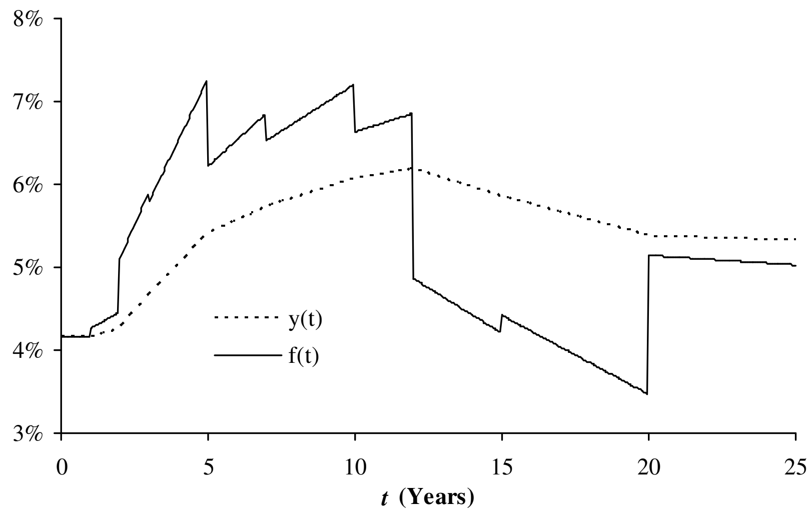
$$t_j = j \cdot 0.5, \quad j = 1, 2, \ldots, 50.$$

Setting $y(t) = y(1)$, $t < 1$, and then running the bootstrap procedure on the swap price expression results in the yield shown in Figure 2. The same figure also shows the continuously compounded forward curve. The discontinuous "saw-tooth" shape of the forward curve is characteristic for bootstrapped yield curves with piecewise linear yield.

[7] In actual markets, swap yields are most often increasing functions of the swap maturity, rather than humped as in Table 1. The data in Table 1 was picked to stress the curve construction algorithms, in order to emphasize their strengths and weaknesses.

The notebook T4_Replicate_Lecture_Curves_01.ipynb shows its construction.

Figure 2: Yield and Forward Curve



Notes: Yield curve is constructed by bootstrapping, assuming piecewise linear yields. Swap data is in Table 1.

4.26.2 Piecewise Flat Forward Rates

Assume now that the instantaneous forward curve is piecewise flat, switching to a new level at each point in $\{T_i\}$, i.e.

$$f(T) = f(T_i), \quad T \in [T_i, T_{i+1}), \tag{48}$$

with $T_0 \triangleq 0$. This corresponds to an interpolation rule where $\ln P(T)$ is linear in T, or

$$P(T) = P(T_i)e^{-f(T_i)(T-T_i)}, \quad T \in [T_i, T_{i+1}),$$

where a bootstrap algorithm can be used to establish the values of the N unknown constants $f(T_0)$, $f(T_1)$,..., $f(T_{N-1})$. From the equation

$$y(T)T = \int_0^T f(u)\,du,$$

we see that the assumption of piecewise flat forwards gives, for $T \in [T_i, T_{i+1})$,

$$y(T) = \frac{y(T_i)T_i + f(T_i)(T-T_i)}{T} = f(T_i) + \frac{(y(T_i) - f(T_i))\,T_i}{T},$$

or

$$y(T) = \frac{1}{T}\left(T_i y(T_i)\frac{T_{i+1}-T}{T_{i+1}-T_i} + T_{i+1}y\left(T_{i+1}\right)\frac{T-T_i}{T_{i+1}-T_i}\right).$$

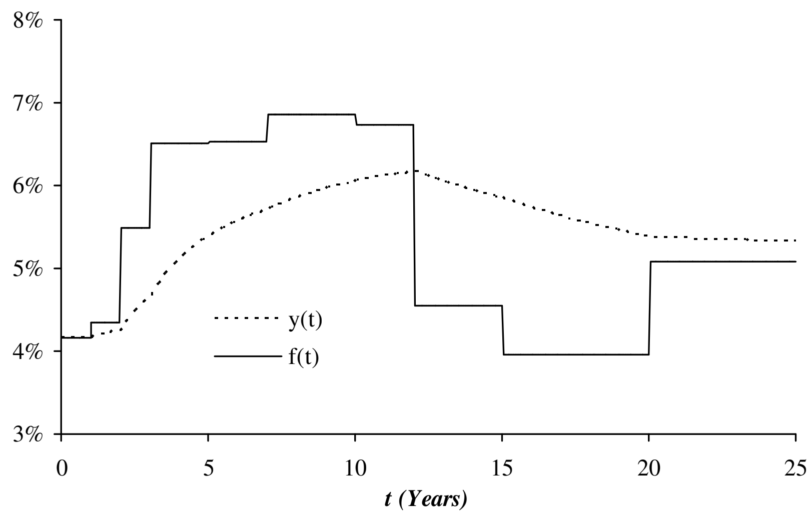
The yield curve will remain continuous.

Figure 3 below shows the results of applying (48) to the swap data in Table 1. Notice the non-linear behaviour of yields between maturity dates and the staircase shape of the forward curve.

The notebook `T4_Replicate_Lecture_Curves_01.ipynb` shows its construction.

We note that linear interpolation on log discount factors is, of course, exactly the same as the piecewise flat interpolation on forward rates, and the former is often used in actual implementation due to some performance advantages.

Figure 3: Yield and Forward Curve



Notes: Yield curve is constructed by bootstrapping, assuming piecewise flat forward rates. Swap data is in Table 1.

4.27 C^1 Yield Curves: Hermite Splines

As we have seen, simply bootstrapped curves generally result in a discontinuous forward curve. From an empirical/economic perspective, such discontinuities are often unrealistic (unless the jumps happen on Central Bank meeting dates, where it is indeed realistic, in the front of the curve, typically), and may also result in distortions of derivative prices and opening a dealer to the possibility of being "picked off" by a hedge fund (more on such strategies later). In this section, we consider a simple scheme to extend the bootstrapping technique to produce a once-differentiable yield curve and a continuous forward curve. Our scheme relies on *Hermite*[8] *cubic splines*, where we write

$$y(T) = a_{3,i}(T - T_i)^3 + a_{2,i}(T - T_i)^2 + a_{1,i}(T - T_i) + a_{0,i}, \quad T \in [T_i, T_{i+1}], \tag{49}$$

for a series of constants $a_{3,i}$, $a_{2,i}$, $a_{1,i}$, $a_{0,i}$ to be determined from given values of $y(T_i)$, $y(T_{i+1})$, $y'(T_i)$, and $y'(T_{i+1})$. Andersen & Piterbarg (2010) contains a review of Hermite spline theory.

A particularly popular choice among Hermite splines is the *Catmull-Rom spline*, where derivatives $y'(T_i)$, $i = 1, \ldots, N$, are constructed by finite differences, relieving the user from directly specifying them.

4.27.1 Technical Details

Optional section

As shown in Andersen & Piterbarg (2010), for the Catmull-Rom spline we can organize (49) in a vector/matrix form as

$$y(T) = \mathbf{D}_i(T)^\top \mathbf{A}_i \begin{pmatrix} y_{i-1} \\ y_i \\ y_{i+1} \\ y_{i+2} \end{pmatrix}, \quad T \in [T_i, T_{i+1}], \quad i = 1, \ldots, N-1, \tag{50}$$

where, adapting as necessary the notation,

$$\mathbf{D}_i(T) = \begin{pmatrix} d_i^3 \\ d_i^2 \\ d_i \\ 1 \end{pmatrix}, \quad d_i = \frac{T - T_i}{h_i}, \quad y_i = y(T_i), \quad h_i = T_{i+1} - T_i,$$

and the matrix $\mathbf{A}_i$ is not shown here but is given in (Andersen & Piterbarg, 2010, (6.54)–(6.56)). While nominally (50) involves the values y_{N+1} and y_0, the matrices $\mathbf{A}_{N-1}$ and $\mathbf{A}_1$ are such that these values are irrelevant.

The Catmull-Rom spline prescription (50) completely specifies the yield curve on the interval $[T_1, T_N]$, given the N constants $y_1, \ldots, y_N$. To extend the yield curve to cover the interval $[0, T_1)$, we need to supply additional extrapolation assumptions. As in bootstrapping, possible choices for this additional equation is $y_0 = y(0) = y_1$, or perhaps the slope condition

$$\frac{y_1 - y_0}{h_0} = \frac{y_2 - y_1}{h_1}. \tag{51}$$

Away from the boundaries, we notice that the price of security i depends only

[8]So called by virtue of being parametrized by the values of the function and its first derivative at the edges of each segment.

on $y_1, \ldots, y_{i+1}$, as the pricing equations take the diagonal form

$$\begin{aligned}
V_1 &= F_1\left(y_1, y_2, y_3\right), \\
V_2 &= F_2\left(y_1, y_2, y_3\right), \\
V_3 &= F_3\left(y_1, y_2, y_3, y_4\right) \\
&\vdots \\
V_{N-1} &= F_{N-1}\left(y_1, \ldots, y_N\right), \\
V_N &= F_N\left(y_1, \ldots, y_N\right),
\end{aligned}$$

for non-linear functions F_i. Here F_i is typically only mildly sensitive to y_{i+1}, so the system of equations is nearly, but not quite, in bootstrap form. This makes solving for the y_i's an easy fare for a standard non-linear root-search algorithm (see Press et al. (1992) for several algorithms). We can also consider an iteration on a series of bootstrap procedures. To describe this idea, let $y_i^{(k)}$ be the value for y_i found in the k-th iteration, and consider then the following algorithm:

1. Let $y_j^{(k)}$, $j = 1, \ldots, i-1$, and $y_{i+1}^{(k-1)}$ all be known.
2. Make a guess for $y_i^{(k)}$.
3. Compute $V_i = F_i(y_1^{(k)}, \ldots, y_i^{(k)}, y_{i+1}^{(k-1)})$.
4. If V_i equals the market value stop. Otherwise return to Step 2.
5. If $i < N$, set $i = i + 1$ and repeat.

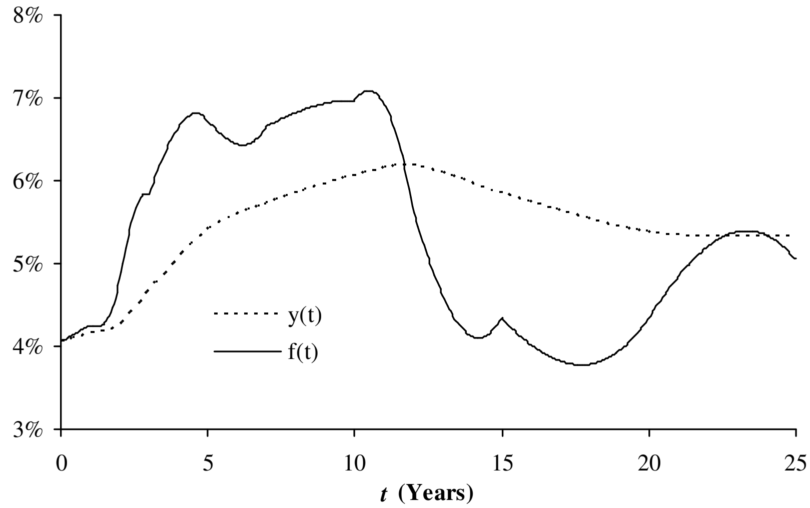
We emphasize that the iteration over Steps 2–4 is still only one-dimensional, as in the bootstrapping algorithm of Section 4.26. Upon completion, the algorithm above yields $y_1^{(k)}, \ldots, y_N^{(k)}$. Iterating over k, we repeat the algorithm until the differences between the yields found at the k-th and $(k+1)$-th iteration are sufficiently small, say when

$$N^{-1} \sum_{i=1}^{N} \left(y_i^{(k+1)} - y_i^{(k)}\right)^2 < \varepsilon^2,$$

where ε is a given tolerance. To initialize the iteration over k, we need a starting guess $y_1^{(0)}, \ldots, y_N^{(0)}$; a good choice is the yield curve constructed by regular bootstrapping.

In Figure 4, we show the results of applying the algorithm above (using the boundary choice (51)) to the numerical example of Sections 4.26.1 and 4.26.2. We see that, as desired, the yield curve is smooth and the instantaneous forward curve is continuous. As the yield curve by construction is only once differentiable, the forward curve is not differentiable; this is obvious from the figure.

Figure 4: Yield and Forward Curve



Notes: Yield curve is assumed to be a Catmull-Rom cubic spline. Swap data is in Table 1.

We can easily extend the procedure above beyond Catmull-Rom splines to more complicated C^1 cubic splines in the Hermite class, using results from Andersen & Piterbarg (2010). For instance, it is relatively straightforward to add *tension* to the Catmull-Rom spline. We cover twice-differentiable tension splines in a bit.

4.28 C^2 Yield Curves: Twice Differentiable Cubic Splines

While the spline method introduced in the previous section often produces acceptable yield curves, the method is heuristic in nature and ultimately does not produce a smooth forward curve. To improve on the latter, one alternative is to remain in the realm of cubic splines, but now insist that the curve is twice differentiable everywhere on $[T_1, T_N]$. We then write (see Andersen & Piterbarg (2010))

$$y(T) = \frac{(T_{i+1} - T)^3}{6h_i} y_i'' + \frac{(T - T_i)^3}{6h_i} y_{i+1}'' + (T_{i+1} - T)\left(\frac{y_i}{h_i} - \frac{h_i}{6} y_i''\right) + (T - T_i)\left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6} y_{i+1}''\right), \quad T \in [T_i, T_{i+1}], \quad (52)$$

where $y_i'' = d^2 y(T_i)/dT^2$, $y_i = y(T_i)$, and $h_i = T_{i+1} - T_i$. Andersen & Piterbarg (2010) demonstrates that continuity of the second derivative across the $\{T_i\}$

knots requires that the y_i'' and y_i are connected through a tri-diagonal linear system of equations.

4.28.1 Technical Details

Optional section

To state the expressions explicitly in matrix format, let $\mathbf{y}'' = (y_2'', y_3'', \ldots, y_{N-2}'', y_{N-1}'')^\top$ and $\mathbf{y} = (y_2, y_3, \ldots, y_{N-2}, y_{N-1})^\top$ such that

$$\mathbf{B}\mathbf{y}'' = \mathbf{C}\mathbf{y} + \mathbf{M}\left(y_1, y_N, y_1'', y_N''\right), \tag{53}$$

where the matrices $\mathbf{B}$ and $\mathbf{C}$ are both $(N-2) \times (N-2)$ tri-diagonal, with elements given by

$$B_{i,i} = \frac{h_i + h_{i+1}}{3}, \quad B_{i,i+1} = \frac{h_{i+1}}{6}, \quad B_{i,i-1} = \frac{h_i}{6},$$

and

$$C_{i,i} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right), \quad C_{i,i+1} = \frac{1}{h_{i+1}}, \quad C_{i,i-1} = \frac{1}{h_i}.$$

The $(N-2)$-dimensional vector $\mathbf{M}(y_1, y_N, y_1'', y_N'')$ captures boundary terms at T_1 and T_N. The most important — and, as discussed later, in a sense *best* — boundary specification is that of the *natural spline*, where we set $y_1'' = y_N'' = 0$. In this case, we have

$$\mathbf{M}(y_1, y_N, y_1'', y_N'') = \mathbf{M}(y_1, y_N) = \left(\frac{y_1}{h_1}, 0, 0, \ldots, 0, 0, \frac{y_N}{h_{N-1}}\right)^\top.$$

Notice that application of a natural boundary condition at time T_1 allows us to recover yields inside the time period $[0, T_1]$ by linear interpolation, using the gradient $y'(T_1)$ at time T_1 (which can easily be found by differentiating (52)).

We notice that (52) combined with (53) allows us to turn any guess of $y_1, y_2, \ldots, y_N$ into a guess for the vector $\mathbf{P}$ in (46) in Andersen & Piterbarg (2010). Specifically, we perform the following steps:

1. Compute the right-hand side of (53).

2. Use a standard tri-diagonal LU solver (see Press et al. (1992)) to invert (53) and recover $\mathbf{y}''$.

3. Apply (52) to determine[9] all values of $y(t_j)$, $j = 1, \ldots, M$, extrapolating as necessary when $t_j < T_1$.

4. Establish $\mathbf{P}$.

[9] For computational reasons, the terms multiplying the various y and y'' in (52) should be pre-cached, to avoid wasting effort when we ultimately perform an iteration.

The computational effort of Steps 1 through 4 are $O(N)$, $O(N-2)$, $O(M)$, and $O(M)$, respectively.

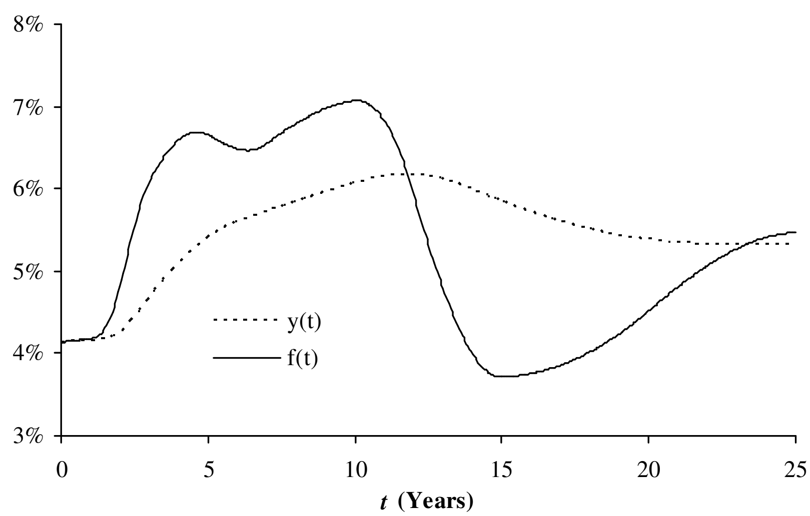
To solve for the correct values of $y_1, y_2, \ldots, y_N$, we iterate on Steps 1–4 using a non-linear root-search algorithm, terminating when the stop criteria is satisfied to within acceptable tolerances. The fitting problem is typically good-natured, and virtually all standard root-search packages (see Press et al. (1992)) can tackle it successfully. Tanggaard (1997), for instance, uses a simple Gauss-Newton scheme with good results. Whatever root-search algorithm is selected, a good first guess can always be found by simple bootstrapping.

4.28.2 Results and Discussion

In Figure 5, we show the results of applying the algorithm above to a natural cubic spline representation of the yield curve example used in earlier sections. The yield curve is smooth and, unlike the Hermite spline case in Figure 4, the instantaneous forward curve is now differentiable, as desired.

The notebook `T4_Replicate_Lecture_Curves_01.ipynb` shows its construction.

Figure 5: Yield and Forward Curve



Notes: Yield curve is assumed to be a C^2 natural cubic spline. Swap data is in Table 1.

While the C^2 cubic spline discussed here has attractive smoothness, it is not necessarily an ideal representation of the yield curve. As discussed in Andersen (2005) and Hagan & West (2004), among others, twice differentiable cubic spline yield curves are often subject to oscillatory behaviour, spurious inflection points, poor extrapolatory behaviour, and non-local behaviour when prices in the benchmark set are perturbed. We shall return to the concept of non-local perturbation effects later, but for now just note that perturbation of a single benchmark price can cause a slow-decaying "ringing" effect on the C^2 cubic yield curve, with the effect of the perturbation of the benchmark instrument price spilling into the entire yield curve. This behaviour is not surprising, given that the spline is constructed through a full $(N-2) \times (N-2)$ matrix system, where interpolation behaviour on the interval $[T_i, T_{i+1}]$ depends on *all* values y_j, $j = 1, \ldots, N$. In contrast, the simple linear-yield bootstrapping method in Section 4.26 interpolation on the interval $[T_i, T_{i+1}]$ involves only the two points y_i and y_{i+1}, and the Hermite spline approach involves only the four points $y_{i-1}, y_i, y_{i+1}, y_{i+2}$.

4.29 C^2 Yield Curves: Twice Differentiable Tension Splines

C^1 Hermite cubic splines are less prone to non-local behaviour than C^2 cubic splines, but accomplish this in a somewhat ad-hoc fashion by giving up one degree of differentiability. Rather than taking such a draconian step, one wonders whether there may be a way to retain the C^2 feature of the cubic spline in Section 4.28, yet still allow control of curve locality and "stiffness". As it turns out, an attractive remedy to the shortcomings of the pure C^2 cubic spline is to insert some tension in the spline, that is, to apply a tensile force to the end-points of the spline. Andersen & Piterbarg (2010) lists the necessary details of this approach, using the classical *exponential tension spline* construction[10] in Schweikert (1966). When applied to the yield-curve setting, the construction involves a modification of the cubic equation (52) for $y(T)$, $T \in [T_i, T_{i+1}]$, to

$$\begin{aligned} y(T) = &\left(\frac{\sinh\left(\sigma\left(T_{i+1} - T\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{T_{i+1} - T}{h_i} \right) \frac{y_i''}{\sigma^2} \\ &+ \left(\frac{\sinh\left(\sigma\left(T - T_i\right)\right)}{\sinh\left(\sigma h_i\right)} - \frac{T - T_i}{h_i} \right) \frac{y_{i+1}''}{\sigma^2} \\ &+ y_i \frac{T_{i+1} - T}{h_i} + y_{i+1} \frac{T - T_i}{h_i}, \end{aligned} \tag{54}$$

where $\sigma \geq 0$ is the *tension factor*, and where we recall the definition $h_i = T_{i+1} - T_i$.

Andersen & Piterbarg (2010) discusses a number of properties of tension splines, the most important perhaps being the fact that setting $\sigma = 0$ will recover the ordinary C^2 cubic spline, whereas letting $\sigma \to \infty$ will make the tension spline uniformly approach a linear spline (i.e. the spline we used in Section 4.26.1). Loosely, we can thus think of a tension spline as a twice differentiable hybrid

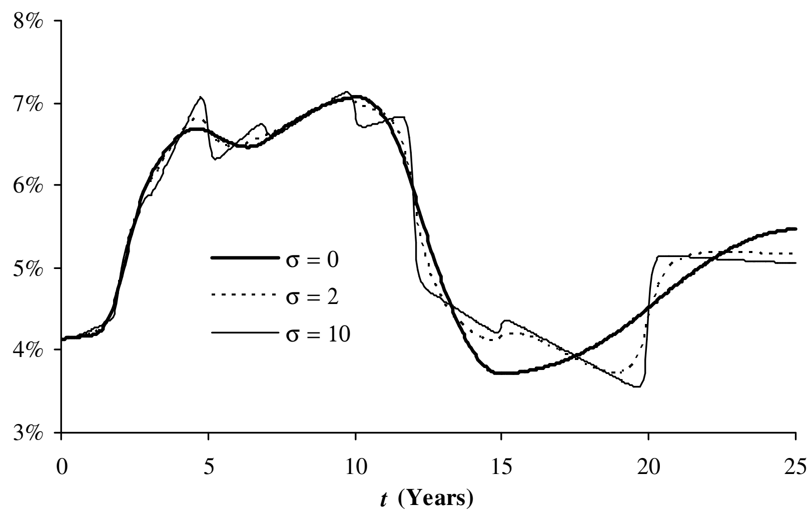
[10] The exponential tension spline is not the only class of twice differentiable tension splines, but is probably the most common. Other classes are discussed in Kvasov (2000) and Andersen (2005).

between a cubic spline and a linear spline. Equally loosely: as we increase σ, spurious inflections and ringing in the cubic spline are gradually "stretched" out of the curve, accompanied by rising (the absolute values of) second derivatives at the knot points. More details on tension splines can be found in Andersen (2005), who also discusses application of computationally efficient local spline bases and the usage of a $T-$dependent tension factor to gain further control of the curve.

We observe that (54) is structurally similar to (52), and allows for a matrix representation of the same form, albeit with suitably modified definitions of the vector $\mathbf{M}$ and the matrices $\mathbf{B}$ and $\mathbf{C}$; we leave these modifications as an exercise to the reader. Suffice to say that once a value of σ has been decided upon, the numerical search for the unknown levels y_i, $i = 1, \ldots, N$, can proceed along the same principles as in Section 4.28 above. Figure 6 below shows an example; notice how increasing the tension parameter gradually moves us from cubic spline behaviour to bootstrap behaviour.

The notebook `T4_Replicate_Lecture_Curves_01.ipynb` shows the construction of these curves.

Figure 6: Forward Curve



Notes: The yield curve is constructed as a C^2 natural tension spline, with tension parameters as given in the graph (only the forward curve $f(t)$ is shown). Swap data is in Table 1.

Remark 1. *If the tension spline is applied not to yields, but to the logarithm of discount factors* $\ln P(t)$*, the limit of* $\sigma \to \infty$ *will produce a piecewise flat forward curve, as in Figure 3.*

One may at this point wonder whether there are any firm rules as to what σ should be. We have no definitive answers to this question, and we do not try to determine σ automatically (although such routines do exist, see Renka (1987)). Instead, we normally treat σ as an "extra knob" that allows users to balance curve smoothness, shape preservation, and perturbation locality to their particular tastes. Inevitably some element of experimentation is required here.

4.30 Other Interpolators

Many other types of interpolators of the yield curve have been developed over the years, emphasizing this or that aspect of the output (the curve). We are not in a position to review them all, but let us briefly mention a couple that are relatively common.

4.30.1 PCHIP

PCHIP stands for *Piecewise Cubic Hermite Interpolating Polynomial*, and is an interpolation scheme based on Hermite splines that preserves monotonicity in the interpolation data and does not overshoot if the data is not smooth. It is common enough to be part of Python's `Scipy` package, see `https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.PchipInterpolator.html`, and is implemented in `FinancePy`, see class `PchipInterpolator`. As mentioned, a specific version of a cubic Hermite polynomial spline is determined by the conditions on the first and, sometimes, second derivative at the knot dates. PCHIP uses a special way of specifying these conditions to ensure monotonicity of the interpolated values if the input points are monotone. Details are available in `Scipy` documentation mentioned above and in the original paper Fritsch & Butland (1984). This is a C^1 spline, i.e. the first derivative is continuous but the second jumps at the knot points. For completeness, we reproduce the main formulas here. Let $h_k = T_{k+1} - T_k$ and $d_k = (y_{k+1} - y_k)/h_k$ are the slopes at the knot points T_k. If the signs of d_k and d_{k-1} are different or either of them are zero, then $y'_k = 0$. Otherwise, it is given by the weighted harmonic mean

$$\frac{w_1 + w_2}{y'_k} = \frac{w_1}{d_{k-1}} + \frac{w_2}{d_k},$$

where $w_1 = 2h_k - h_{k-1}$ and $w_2 = h_k + 2h_{k-1}$. The end slopes are set using a one-sided scheme.

4.30.2 Hagan West

Hagan and West in their influential paper Hagan & West (2004) developed a so-called monotone convex interpolation method that is still used to this day in some institutions. It is designed to preserve both the monotonicity, and the convexity, properties of the input rates. This is a local cubic spline method that is fairly elaborate to implement and we refer those interested to the original paper Hagan & West (2004). In the words of the authors,

> Very simply, none of the methods mentioned so far are aware that they are trying to solve a financial problem—indeed, the breeding ground for these methods is typically engineering or physics. As such, there is no mechanism which ensures that the forward rates generated by the method are positive, and some simple experimentation will uncover a set of inputs to a yield curve which give some negative forward rates under all of the methods mentioned here, as seen in Hagan and West [2006]. Thus, in introducing the monotone convex method, we use the ideas of Hyman [1983], but explicitly ensure that the continuous forward rates are positive (whenever the discrete forward rates are themselves positive).

At the time, negative interest rates seemed like an aberration – of course 10 years after the GFC showed them anything but. However, the method does generate reasonably-looking yield curves preserving monotonicity and convexity in the input data, and is based on a fair amount of financial intuition, unlike the methods we discussed previously that are "one size fits all" interpolators.

The next Topic will introduce methods where various constraints based on financial considerations can be directly specified, and is our preferred way to build curves, notwithstanding important advances made by Hagan and West.

Topic 5
Advanced Curve Construction Topics

5.31 Non-Parametric Optimal Yield Curve Fitting

The techniques we have outlined so far generally suffice for constructing a discount curve from a "clean" set of non-duplicate benchmark securities, including the carefully selected set of liquid staggered-maturity deposits, futures, and swaps most banks assemble for the purpose of constructing a Libor (or RFR) yield curve. In some settings, however, the benchmark set may be significantly less well-structured, involving illiquid securities with little order in their cash flow timing and considerable noise in their prices. This situation may, say, arise when one attempts to construct a yield curve from corporate bonds. While construction of a Libor/RFR curve(s) is the most important task for most market participants, we nevertheless wish to say a few words about techniques suitable for less cooperative benchmark security sets. These techniques can also be applied to Libor/RFR curve construction, of course, and are particularly relevant for applications where we are willing to sacrifice some precision in the fit to benchmark prices in return for a smoother yield curve.

A demonstration notebook ./notebooks/products/rates/FINIBORSINGLECURVE_IborCurveCalibrateSmooth.ipynb is available from FinancePy library and is mentioned on BlackBoard.

5.31.1 Norm Specification and Optimization

When the input benchmark set is noisy, a direct solution may be erratic or may not exist. To overcome this, and to reflect that noise in the input data may make us content to solve the equation only to within certain error bounds, we now proceed to replace this equation with a problem of minimization of a penalized least-squares norm. Specifically, define the space $\mathcal{A} = C^2([t_1, t_M])$ of all functions $[t_1, t_M] \to \mathbb{R}$ that are twice differentiable with continuous second derivative, and introduce the M-dimensional discount vector

$$\mathbf{P}(y) = \left(e^{-y(t_1)t_1}, \ldots, e^{-y(t_M)t_M}\right)^\top.$$

Also, let $\mathbf{W}$ be a diagonal $N \times N$ weighting matrix. Then, as our best estimate $\widehat{y}$ of the yield curve we will here use

$$\widehat{y} = \operatorname*{argmin}_{y \in \mathcal{A}} \mathcal{I}(y), \tag{55}$$

with

$$\mathcal{I}(y) \triangleq \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(y)\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(y)\right) + \lambda \left(\int_{t_1}^{t_M} \left[y''(t)^2 + \sigma^2 y'(t)^2\right] \, dt\right), \tag{56}$$

where λ and σ^2 are positive constants. The norm $\mathcal{I}(y)$ consists of three separate terms:

- A least-squares penalty term

  $$\frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(y)\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(y)\right) = \frac{1}{N} \sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{i,j} e^{-y(t_j)t_j}\right)^2,$$

  where W_i is the i-th diagonal element of $\mathbf{W}$. This term is an outright precision-of-fit norm and measures the degree to which the constructed discount curve can replicate input security prices. The weights W_i can be used to assign different importance to the various securities in the benchmark set, and/or to translate the precision of the fit from raw dollar amounts into more intuitive quantities, such as security-specific quoted yields[11]. Clearly, if (46) can be satisfied, then the least-squares penalty term will attain its minimum (of zero) for all yield curves that satisfy (46).

- A weighted smoothness term $\lambda \int_{t_1}^{t_M} y''(t)^2\, dt$, penalizing high second-order derivatives of y to avoid kinks and discontinuities.
- A weighted curve-length term $\lambda\sigma^2 \int_{t_1}^{t_M} y'(t)^2\, dt$, penalizing oscillations and excess convexity/concavity.

Our choice of calibration norm is, we believe, an attractive one, but other choices obviously are available as well. For instance, in Adams & van Deventer (1994) the norm contains no curve-length term and the smoothing norm is expressed on the forward curve, rather than on the yield curve.

The following result is discussed in Andersen & Piterbarg (2010).

Proposition 2. *The curve $\widehat{y}$ that satisfies (55) is a natural exponential tension spline with tension factor σ and knots at* all *cash flow dates $t_1, t_2, \ldots, t_M$.*

Remark 3. *If we let $\sigma = 0$, the solution to the optimization problem becomes a* cubic smoothing spline*; see Tanggaard (1997) for more details on this case.*

Remark 4. *If we let $\lambda \downarrow 0$, the resulting spline will often end up hitting all benchmark prices exactly, i.e. will satisfy (46) in the limit. The resulting spline is then the optimal* interpolating *curve, in the sense that of all twice differentiable yield curves that match the benchmark prices, the spline is the minimizer of the regularity term $\int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2]\, dt$. If, for $\lambda \downarrow 0$, we do not satisfy (46), then the resulting spline can be considered a* least-squares regression spline *solution.*

Remark 5. *The choice of the smoothing parameter λ is discussed in Andersen & Piterbarg (2010). The main idea is to fix the target RMS (root-mean-square) error for fitting benchmark securities and vary λ until the actual RMS is just below the target.*

5.31.2 Discrete Smoothness Objective

The formulation (56) leads to a nice theoretical result on the optimal-smoothness curve that is also practical to implement under the conditions it is derived. It is sometimes the case, however, that we want to build a non-parametric smooth curve under the constraints that are more general than those that led to Proposition 2. In this case the standard procedure is to write down a version of the least-squares problem (55)–(56) and solve it numerically using an optimizer.

There are many ways that smoothness constraints could be imposed. They could be a first-order, second-order or a combination of both. They could be imposed on yields, instantaneous forward rates, term forward rates, etc. Different

[11] Most fixed-income securities are quoted through some type of yield, e.g. $V_i = g_i(r_i)$ where r_i is the quoted yield and g_i is a function that encapsulates the quoting convention. The quantity $D_i = -(dg_i/dr_i)/g_i$ is known as the *duration* of V_i. Setting $W_i = 1/D_i$ in the least-squares norm will turn price deviations into yield deviations.

terms in the smoothness objectives (see below) can be weighted individually, e.g. putting higher smoothness penalties on the curve further in the future vs. the short end. Also, any kind of interpolation can be used for the curve between the knot dates.

Without trying to list all possible combinations, we present a simple discrete version of (56) suitable for a numerical optimization. Specifically, we replace the first-order term $\int_{t_1}^{t_M} y'(t)^2\,dt$ with a discrete version

$$\sum_{m=1}^{M-1} y(t_m, t_{m+1})^2,$$

where we recall $y(t_m, t_{m+1})$ to be the term forward rates. Likewise, we replace the second-order term $\int_{t_1}^{t_M} y'(t)^2\,dt$ with

$$\sum_{m=1}^{M-2} \left(\frac{y(t_{m+1}, t_{m+2}) - y(t_m, t_{m+1})}{t_{m+1} - t_m} \right)^2,$$

where each term is now the discrete slope of the (term) forward rate curve (which is roughly equivalent to the curvature of the yield curve).

One of the advantages of this formulation of smoothness constraints is that they can be expressed in terms of the same discount factors at the knot dates $\{P(t_m), m = 1, \ldots, M\}$ as the fitness objectives in (56), as

$$y(t_m, t_{m+1}) = -\frac{1}{t_{m+1} - t_m} \ln \frac{P(t_{m+1})}{P(t_m)}.$$

The full objective function then can be written, using only the second-order smoothness conditions for brevity, but allowing for individual fit W_i and smoothness λ_m weights, as (see (55))

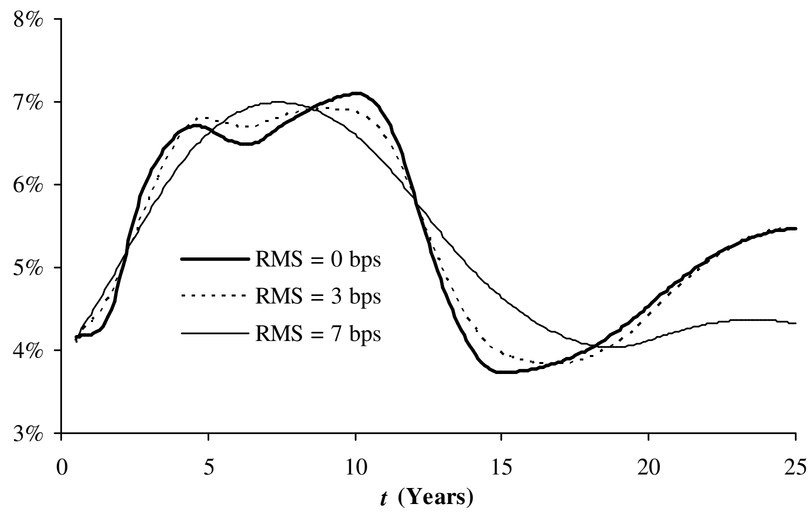
$$\mathcal{I}(y) = \frac{1}{N} \sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{i,j} P(t_j) \right)^2 + \sum_{m=1}^{M-2} \lambda_m^2 \left(\frac{y(t_{m+1}, t_{m+2}) - y(t_m, t_{m+1})}{t_{m+1} - t_m} \right)^2. \quad (57)$$

Since the objective function is given as a sum of squares of individual terms, it is easily amendable to a least-squares numerical optimization such as implemented by `scipy.optimize.least_squares`. Input variables for the optimization are discount factors at all payment dates $\{P(t_j)\}_{j=1}^{M}$ or, alternatively, yields $\{y(t_j)\}_{j=1}^{M}$. We demonstrate this approach in the notebook `./notebooks/products/rates/FINIBORSINGLECURVE_IborCurveCalibrateSmooth.ipynb` which is available from `FinancePy` library and is mentioned on BlackBoard.

5.31.3 Example

Check out the notebook that sits in FinancyPy library at `notebooks/products/rates/FINIBORSINGLECURVE_IborCurveCalibrateSmooth.ipynb`

Figure 7: Forward Curve



Notes: The yield curve is constructed as an optimal C^2 natural tension spline, with an RMS yield error constraint as listed in the graph (only the forward curve $f(t)$ is shown). The tension parameter is set to $\sigma = 0$ for all curves. Swap data is in Table 1.

For our test case, the zero-RMS optimal ($M-$knot) forward curve in Figure 7 is virtually identical to the $N-$knot cubic spline solution in Figure 5. In general, the $N-$knot interpolating curve can be interpreted as a constrained solution to (55) with $\lambda = 0$, with the constraint requiring that knots be placed only at benchmark maturities $\{T_i\}_{i=1}^N$, rather than at all cash flow dates $\{t_j\}_{j=1}^M$. The effect of enforcing this additional constraint is often rather small, at least for the purposes of constructing a Libor/RFR curve.

5.32 Advanced Curve Building Features

What we have covered so far provides a solid base for good understanding of (single) curve construction. In this section we briefly cover some of the more advanced topics.

5.32.1 Mixed Interpolation

Different macro-economic and technical considerations are important for different parts of a yield curve. The short end of the curve (first few years) is primarily driven by the expectations of the actions of the relevant central banks and their monetary policy. The "belly" of the curve is often associated with the levels of perceived risk premia (more on that later), and the long end of the curve responds to the supply and demand of long-dated government and corporate bond issuance, as well as term premia. It also reflects a "convexity premia" that we will discuss in due course. Of course special policies such as QE or Japan's "zero 10 year rate" affect those parts of the yield curve that they target. As different market forces affect different parts of the curve differently, we often need to use somewhat different approaches to building the curve at different time scales. The granularity of available market information, and hence the granularity of the features of the curve, is much higher at the short end than in the long end. This is often reflected in curve construction by using different interpolations for different segments. It is not uncommon, for example, to use piecewise flat forward rate interpolation in the short end taking into account all the micro features of the market (more on that in a bit), and a smooth (cubic spline or similar) interpolation for the rest of the curve, a technique known as mixed interpolation. Adapting the methods we described earlier to use mixed interpolation is usually fairly straightforward.

5.32.2 Turn-of-Year Effects

Optional section

Many of the curve construction algorithms so far have been designed around the implicit idea that the forward curve should ideally be *smooth*. While this is, indeed, generally a sound principle, exceptions do exist. For instance, it may be reasonable to expect instantaneous forwards to jump on or around meetings of monetary authorities, such as the Federal Reserve in the US. In addition, other "special" situations may exist that might warrant introduction of discontinuities into the forward curve. A well-known example is the turn-of-year (TOY) effect where short-dated loan premiums spike for loans between the last business day of the year and the first business day of the following calendar year.

One common way of incorporating TOY-type effects is to exogenously specify an *overlay curve* $\varepsilon_f(t)$ on the instantaneous forward curve. Specifically, the forward curve $f(t) = f(0, t)$ is written as

$$f(t) = \varepsilon_f(t) + f^*(t), \tag{58}$$

where $\varepsilon_f(t)$ is user-specified — and most likely contains discontinuities around special event dates — and $f^*(t)$ is unknown. The yield curve algorithm is then subsequently applied to the construction of $f^*(t)$. That is, rather than solving $\mathbf{cP} = \mathbf{V}$ (see equation (46)), we instead write

$$P(T) = e^{-\int_0^T \varepsilon_f(t)dt} e^{-\int_0^T f^*(t)dt} \triangleq P_\varepsilon(T)P^*(T) \tag{59}$$

and solve

$$\mathbf{c}_\varepsilon \mathbf{P}^* = \mathbf{V}, \tag{60}$$

where $\mathbf{P}^* = (P^*(t_1), \ldots, P^*(t_M))^\top$, and $\mathbf{c}_\varepsilon$ is a modified $N \times M$ coupon matrix, with elements

$$(\mathbf{c}_\varepsilon)_{i,j} = c_{i,j} P_\varepsilon(t_j).$$

Construction of $\mathbf{c}_\varepsilon$ can be done as a pre-processing step, after which any of the algorithms discussed earlier in this chapter can be applied to attack (60). Once the curve $P^*(t)$ (or, equivalently, the yield curve $y^*(t) = -t^{-1} \ln P^*(t)$) has been constructed, any subsequent use of the curve for cash flow discounting requires, according to (59), a multiplicative adjustment of time t discount factors by the quantity $P_\varepsilon(t)$.

5.32.3 Central Bank Policy Effects

Optional section

As mentioned a number of times, CBs have direct influence on the short end of yield curves by essentially setting overnight rates. CB meetings happen on a predetermined schedule and, barring extraordinary meetings (that do happen), it is not unreasonable to expect the instantaneous forward rates to potentially jump on these dates, and stay roughly constant in between. If the future hikes/rate cuts are known with a high degree of certainty, these can be incorporated into the curve construction using the overlay method described above. If however the rate changes are less certain (in terms of their probability or size), then one would want to include CB meeting dates into the curve as the dates when the instantaneous forward rate can jump, but let the curve construction algorithm itself figure out the market-expected size of these rate jumps. In some sense this is a more sophisticated version of determining the market-implied probability of the next CB rate decision that we covered earlier in the course. This is particularly relevant for CB meeting dates further into the future, as the rate change expectations are relatively firm only for the first one or two meetings.

The expected shape of the yield curve in the front end that is aware of CB meetings can be achieved by specifying piecewise flat interpolation of instantaneous forward rates and setting CB meeting dates as knot dates where the ON forward rate is allowed to change. The most appropriate calibration framework here is the non-parametric curve fitting as in Section 5.31. Since we are concerned about ON forward rates specifically, we may want to tweak the objective function (57) to impose second-order smoothness conditions on the *forward* *rates* not yields. For example, we could use this,

$$\mathcal{I}(y) = \frac{1}{N}\sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{i,j}P(t_j)\right)^2$$
$$+ \sum_{m=2}^{M-2} \lambda_m^2 \left(\frac{y(t_{m+1}, t_{m+2}) - y(t_m, t_{m+1})}{t_{m+1} - t_m} - \frac{y(t_m, t_{m+1}) - y(t_{m-1}, t_m)}{t_m - t_{m-1}}\right)^2.$$

As we said, piecewise flat ON forward rate interpolation (between knot dates $\{t_j\}$) works best over the short end of the curve (next 2-3 years, say) where the impact of future CB decisions is most pronounced. It is customary to switch to a smoother interpolation for longer times.

5.33 Single Curve to Multi-Curve and Back (well... maybe)

Once upon a time, there was a single interest rate/discount curve (per currency). The same curve was used to price swaps of different frequencies (linked to 1M,3M,6M Libors) and for discounting pretty much all the cash flows that a bank needed to discount. 1M Libor swaps had market prices/breakeven rates that were barely different from 3M Libor swaps, and even the OIS curve was pretty much in line with Libor. This all changed during GFC when the *basis*, i.e. the difference between say a 1M Libor and a 3M Libor widened from single basis points to 100s of bps (due to credit risk, liquidity and other factors). The approach of using the same curve for everything was no longer valid.

So, instead of building a single generic “Libor” curve, market participants switched to building multiple related interest rate curves – one for OIS, one for 1M Libor, one for 3M Libor, etc. Some even started referring to these collections of curves as “curve surfaces”, with one dimension being the usual maturity and the other being the tenor of the underlying rate. As these multiple curves were intrinsically interlinked (via a tenor basis swap market, as well as OIS used for discounting for Libor swaps), the curve building technology had to be significantly upgraded. Flow rate quant departments went from having single digit number of quants to multiple 10s.

With the Libor reform, the landscape is shifting yet again. With all the Libor rates of all tenors going away and being replaced by a single RFR curve, some hope that the multi-curve technology is no longer needed as the market is moving to a single-curve paradigm. This is not, unfortunately (or fortunately for quants?) entirely true. Indeed many curves are disappearing but it is unlikely we will have a single curve per currency ever again. In EUR, EURIBOR is not going away any time soon while ESTR is the new RFR rate. In USD, Libor rates are going to be with us at least until mid-2023 but, more importantly, FedFunds curve will exist alongside SOFR, as well as possibly some other versions of overnight rate curves. Perhaps in the UK we are closest to having a single Sonia curve for projecting rates but discounting between counterparties depends on the details of the collateral agreements between them so there will always be multiple discount curves. Cross-currency markets will still require their own versions, and tough Legacy contracts are/will be linked to Libor rate replacements that look different than OIS. Also, we have seen cases where the exact same swap has different break-even rates on different clearing houses (LCH vs CME) for various technical (market positioning, liquidity) reasons.

Hence, it is unlikely we will ever be able to retire a multi-curve framework and it is important to understand the concepts. We here however focus on a single RFR curve that typically covers a good majority of use cases.

Topic 6
Bond Curves

6.34 Bonds Are Different

As we mentioned earlier, bonds are idiosyncratic and not fungible. Multiple factors affect their value (on-the-run vs. off-the-run, being "on special", how much liquidity is available in a particular bond, etc.)

To build a "government bond curve", many choices need to be made. One is that a selection of bonds needs to be made to define a consistent government curve. These bonds are often chosen to be on-the-runs due to their liquidity. However, this is not quite as straightforward as just looking at the most-recently issued bonds, as there are typically large gaps between maturities of freshly-issued bonds; on-the-runs give us a 10y point and a 30y point, and possibly a 20y point on the curve, but nothing in between. Hence, to build a proper and useful curve to sufficient granularity, one needs to complement the on-the-runs with some other information.

Another implication of the idiosyncrasy of the government bond market is the need to assess, and quantify, the value of each one of them relative to others. As we said, it is generally impossible to build a curve that reprices every bond exactly. It is, however, possible, and is often the approach employed by market participants, to build a curve that fits all bonds in the best-fit, but not exact, sense, such as by minimizing the root mean square (RMS) error. Then, the degree of deviation for each particular bond from this best-fit curve serves as a measure of its value – being either relatively rich, or cheap, vs. the curve.

6.35 Best-Fit Curves

Bonds, as discussed, provide a rich universe of tradeable securities that, in general, cannot be fit with a unique curve. Here the approach of a best-fit, rather than exact, curve is really beneficial. Bond prices, YTM and ASW spreads can be used as inputs to build a particular curve, but $z-$spreads are often chosen as the most accurate measure of value of one bond vs. the other.

Notebooks `./notebooks/products/bonds/FINBONDYIELDCURVE_FittingToBondMarketPrices.ipynb`, `./notebooks/products/bonds/FINBONDYIELDCURVE_FittingToAswAndZSpreads.ipynb` in `FinancePy` have various examples, fitting various types of curves to bond yields, ASW and z-spreads.

The framework we developed for non-exact best-fit of curves earlier in the course is particularly well suited for building bond yield curves. The approach, as we recall, is a non-parametric one as it uses splines for the fit. The splines are constrained via the smoothness penalty, with the degree of smoothness controlled by the builder. Sometimes, however, it is helpful to specify a particular *functional* form for the curve, with some of the parameters determined via a best-fit procedure. Using a particular functional form aids in interpreting a best-fit curve in financial terms. It is not uncommon, for example, to have parameters that are responsible for the overall level of the curve, its slope and curvature. This is particularly helpful in historical analysis of curve shapes as the historical time series of the parameters of a thoughtfully-parametrized curve could give us insights into the specific changes in the shape of the yield curve through history.

Clearly this idea has many parallels to the PCA method that should be familiar to the reader and will also be discussed later, where factors associated with particular loadings tell us the historical changes in the level, slope, curvature, etc. of the curve. In the PCA method, the loadings themselves are determined by the historical data. We can go even further and specify parametric forms for them.

6.36 Nelson-Siegel-Svensson

The Nelson-Siegel (NS) parametrization, and its extension Nelson-Siegel-Svensson (NSS), is a particularly popular approach for doing so that goes back some ways (the NS paper came out in 1987). Let us review the approach, due to its historical significance and continuing popularity.

In the original NS paper (see Nelson & Siegel (1987) or `https://www.researchgate.net/publication/24103017_Parsimonious_Modeling_of_Yield_Curves)`) the instantaneous forward rate curve is parametrized as follows (time now is $t = 0$, and $\tau = T - t$ so in our notations $f(0, T) = f(\tau)$),

$$f(\tau) = \beta_1 f_1(\tau) + \beta_2 f_2(\tau) + \beta_3 f_3(\tau),$$

where

$$\begin{aligned} f_1(\tau) &\equiv 1, \\ f_2(\tau) &= e^{-\tau/\lambda}, \\ f_3(\tau) &= (\tau/\lambda)\, e^{-\tau/\lambda}. \end{aligned}$$

When averaged ($y(\tau) = \tau^{-1} \int_0^\tau f(u)du$) to describe zero rates (yields), we have, with the same conventions in terms of the calendar time and maturity time,

$$y(\tau) = \beta_1 y_1(\tau) + \beta_2 y_2(\tau) + \beta_3 y_3(\tau), \tag{61}$$

$$y_1(\tau) \equiv 1, \tag{62}$$

$$y_2(\tau) = \frac{1 - e^{-\tau/\lambda}}{\tau/\lambda}, \tag{63}$$

$$y_3(\tau) = \frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda}. \tag{64}$$

The parameters are interpreted as follows in the PCA vernacular: λ is a constant that controls the decay rate and $\beta_1, \beta_2, \beta_3$ are "factors". For the yield curve, β_1 is the level factor and is often interpreted as the long-term average; β_2 is the short-term factor and is interpreted as the slope of the curve, and β_3 is the medium-term factor that is related to the curvature of the yield curve.

We can see clear parallels between (61) and the PCA. The loadings postulated by the NS model are a constant $y_1(\tau) \equiv 1$ multiplied by the level parameter β_1, a "slope" loading $y_2(\tau) = \frac{1-e^{-\tau/\lambda}}{\tau/\lambda}$ (whose rate of decrease is controlled by λ) multiplied by the slope "factor" β_2, and a "curvature" loading $y_3(\tau) = \frac{1-e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda}$ multiplied by the "curvature" factor β_3.

HW: plot the loadings

The NS parametrization is sufficiently rich to describe a shape of the yield curve with a single "hump". Historically, yield curve sometimes exhibit a second hump and that is where the NSS extension helps to capture it. The NSS parametrization specifies

$$y(\tau) = \beta_1 y_1(\tau) + \beta_2 y_2(\tau) + \beta_3 y_3(\tau) + \beta_4 y_4(\tau), \tag{65}$$

where the addition term $y_4(\tau)$, compared to (61), is given by

$$y_4(\tau) = \frac{1 - e^{-\tau/\kappa}}{\tau/\kappa} - e^{-\tau/\kappa} \tag{66}$$

for $\kappa \neq \lambda$.

There are a few different objectives for bond curve building. Let us consider them in turn.

6.36.1 Building a Stand-Alone Bond Curve

Given a collection of coupon-bearing bonds, such as US Treasuries, to be our benchmarks, one objective could be to build a zero-coupon bond curve such that it fits – in the best-fit sense – all of them. We recall the bond-pricing formula (39) where we consider everything at time $t = 0$ (and thus drop it from the formula) and use index $i = 1, \ldots, N$ to denote the i-th benchmark and its coupon paying schedule,

$$V_{\text{bond},i} = \sum_{m=1}^{M_i} (c_i/f)\, P_{\text{gov}}(T_{m,i}) + P_{\text{gov}}(T_{N,i}), \quad i = 1, \ldots, N. \tag{67}$$

The bond prices $V_{\text{bond},i}$, $i = 1, \ldots, N$, are known, and we want to recover the government bond curve $P_{\text{gov}}(T)$ for all $T > 0$.

This is a straightforward extension of our task of building non-parametric curves of Section 5.31 combined with the parametrization of the zero-coupon bond curve that follows from the NS(S) parametrization of the ON forward and zero rate curves (61)–(66). For a given set of NSS parameters $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, \lambda, \kappa)$ we denote a candidate bond curve by

$$P(T;\theta) = \exp\left(-y(T;\theta)T\right), \quad T \geq 0,$$

where $y(T;\theta)$ is given by the NS(S) parametrization (61)–(66). Then, we use the analogue to (67) we denote the values of the benchmarks calculated using the candidate bond curve by $V_{\text{bond},i}(\theta)$. Finally, we set up and solve the following non-linear optimization problem,

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \left(V_{\text{bond},i} - V_{\text{bond},i}(\theta)\right)^2 .$$

The Fed published zero-coupon bond curves calculated using a similar methodology described in the paper `https://www.federalreserve.gov/data/nominal-yield-curve.htm`. From the paper one can download a history of these curves going back many decades.

6.36.2 Interpolating Bond Metrics

The same idea, and the same functional form, could be used to interpolate various metrics associated with individual bonds. This by itself does not produce a curve usable for discounting, as will be clear momentarily. However, this can be used for other purposes, chiefly the rich/cheap analysis and devising RV strategies.

Given a collection of bonds with market coupon-bearing bond yields (or other measures; a bit more on that later) $y^1_{\text{mkt}}, \ldots, y^N_{\text{mkt}}$ with maturities $T_1, \ldots, T_N$, the coefficients of the NSS parametrization are recovered as a solution to

$$\theta^* = \arg\min_{\theta} \sum_{n=1}^{N} \left(y(T_n;\theta) - y^n_{\text{mkt}}\right)^2$$

where $y(\tau;\theta)$ are model-given coupon-bearing bond yields as in (61)–(66) as functions of the model parameters $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, \lambda, \kappa)$, and θ^* is the result of the best fit. Note that in this application $y(\tau;\theta)$ are *not* zero-coupon zero rates – we essentially fit a functional form to a collection of *coupon-bearing* bond yields that are defined by (40),

$$V_{\text{bond}} = \sum_i \frac{c/f}{\left(1 + y_{\text{bond}}/f\right)^i} + \frac{1}{\left(1 + y_{\text{bond}}/f\right)^N}.$$

Instead of the NS(S) parametrization, many other functional forms could be used such as polynomials of various degrees, splines, etc.

Likewise, the same idea could be applied to asset-swap spreads (ASW) or z-spreads, that are defined in (41) and (43). Each metric has its own advantages and disadvantages and the one to use depends on what one is trying to achieve.

The various β parameters enter the objective function linearly and thus are generally well-behaved for fitting, as the problem can be seen as basically a linear regression. The parameters λ and κ affect the fit in non-linear ways – some care needs to be taken to make sure the optimization converges such as, for example, imposing sensible bounds on these parameters and starting from a good initial guess.

6.37 Use of NSS and Other Cross-Sectional Models

Being a quite parsimonious functional form, the NS (or NSS) curve cannot exactly fit market prices/yields of more than a couple of securities. As such, it cannot really be used as "the" yield curve for, say, pricing or risk-management purposes. They are best utilized as "snapshot" models, cross-sectional devices to interpolate, primarily in the best-fit, sense, prices or yields of bonds. These can be used to e.g. determine our best estimates of the "market" yields of the bonds we cannot observe (due to liquidity or timing of market updates, for example) based on those bonds that we do observe.

As we discussed, NSS curve and its many flavours (that we do not discuss much but briefly mention later) are capable of roughly capturing the overall shape of many typical bond curves, with parameters that are easy to interpret. For example the various β's, as we have seen, can be interpreted as factors that capture the overall level, slope and curvature of yield curves. Hence we can perform statistical analysis on these factors that are fitted every day to gain deeper insights on how the yield curve changed throughout history (and, possibly, to even use as the basis of our deductions of how it could move in the future). This is a "macro" picture for which NSS and related methods are sometimes used.

Coupled with a common (but somewhat ad-hoc) assumption that the residuals, or deviations between specific bonds and the overall shape of the curve, are mean reverting, gives practitioners some insights into whether individual securities are priced "out of line" with the rest of the cross-sectional data. E.g. some bonds could be perceived as "cheap" or "dear" based on how far they are out of line with the best-fitted shape, with implications for relative value strategies. For example, the bigger the deviation of a given security from the curve, the higher, potentially, the "value" in that particular bond. This forms the basis for "micro" usage of these types of curves for RV.

Bond yields are not the only parameters that can be fitted with these types of curves, and possibly not the best (depending on the task in hand). Bond yields encapsulate a combination of various market factors, namely, to first order

- The interest rate curve of the economy as measured by, say, a curve build from OIS swaps and other OTC instruments;

- The term structure, or overall shape, of differences (spreads) between bonds and OTC derivatives;
- Bond-specific idiosyncratic circumstances (liquidity, demand, etc.) for each particular bond.

An NSS curve built from bond yields then reflects all of these, and the changes to the yield-based NSS curve are driven by the combination of all three. With a view to capture bond-specific drivers of value, or market drivers, one can use measures other than bond yields. For example, asset swap spreads or $z-$spreads isolate bond-market specific factors from the overall changes in the term structure of interest rates, the latter as measured by the OIS (or Libor) curve. Hence, for understanding either bond vs. swap macro dynamics, or for finding bond-specific rich/cheap bonds (*relative to the OIS curve*), it is often convenient to fit the NSS curve to these measures rather than bond yields.

The NSS parametrization is reasonably intuitive but may not be flexible enough for reflecting more complex shapes of, say, $z-$spreads. Then, one can eschew simplicity and fit more flexible forms such as polynomials or splines, where the degree of resolution of features of bond curves could be controlled via parameters such as the degree of the polynomial used or the number of knots in a spline. The explainability of various parameters in this functional form is then, however, generally lost.

Some examples of these types of curves are presented in notebook `./notebooks/products/bonds/FINBONDYIELDCURVE_FittingToAswAndZSpreads.ipynb` in `FinancePy` library.

Topic 7
Expectations, Term Premia, Convexity

A significant amount of material in this topic is taken from Rebonato (2018).

7.38 What Determines the Yield Curve Shape

Investors, given a choice of two investments with the same expected return but different uncertainty of outcomes will universally prefer investments with lower uncertainty. This "risk aversion" is fundamental to understanding investors' behaviour. Put another way, investors require compensation for the risk they take and demand higher expected return from investments that are more uncertain. This excess return is called *risk premium* or, in the context of yield curve investing, *term premium* because it is typically associated, and increases with, the "term", or the maturity, of bonds that one invests in.

In the context of fixed income investments, risk premium is one of key determinants of the yield curve shape and its changes, but not the only one. It is common to distinguish three factors that drive the shape of the yield curve:

$$\text{market yield}=\text{expectations}+\text{risk premium}+\text{convexity}. \tag{68}$$

Let us briefly review them here.

The first one, commonly called "expectation", is based on the market expectations as to what market participants will do in the future. Of primary importance here are the central banks and the expectations of near-term monetary policy, as we already discussed previously. Expectations generally drive the short end of the yield curve which can be thought of as the expectation, in mathematical (if somewhat loose) sense of what are the next actions that a relevant CB would take. These could be rate hikes/cuts, or simply statements that the current monetary policy (be that accommodative or restrictive) will continue for some time. CBs generally try to avoid surprising the markets and signal their intentions well in advance, unless a rapidly changing economic environment warrants a quick action. We have seen this in the earlier part of 2022 and, in GBP, in September 2022. Hence there is little uncertainty, relatively speaking, in the near-term path of interest rates and, thus, little risk premia in shorter-term rates, roughly up to 2 years.

As the rate duration increases, the uncertainty in the future monetary policy, as well as the economic situation that drives it, of course, grows. The risk/term premia kicks in, with market participants requiring higher compensation, in terms of higher rates paid on government debt and, by extension, on swaps and other relevant securities. The yield curve typically slopes upward because of this. It is important to realize that the steepness of the yield curve in the medium range (around 10y point, where expectations are often non-existent or too vague to matter) *can come only from the term premium.* We will put some math around this shortly.

We shall also explain that the so-called "roll-down" is the way to estimate one-period excess return, and thus an important metric to an investor.

"Risk premium" is a fairly generic term that described higher compensation that investors demand for higher risk. There are many separate risk sources that feed the overall risk premia. Specialized economics literature goes into a great level of detail trying to disentangle various components of it, something that we shall not do here. But the factors of importance to be briefly mentioned are inflation expectations, the level of issuance of bonds, liquidity, and so on. Decomposition of the risk premium by the risk source is by no means unique, and can be represented in different "coordinates" depending on the need.

For our needs it suffices to think of risk premium as the difference between market yields and the expectations of the path of the future short rates.

Further down the maturity spectrum another mechanism kicks in, the so-called *convexity.* This is a somewhat technical factor that we will discuss in more details later in the notes. Being more technical it is sometimes neglected and omitted from introductory texts on yield curves but is nonetheless quite important. As we shall see, the generally downward sloping shape of the yield curve for long maturities (30 years+) is down to convexity.

As an aside, convexity is an over-used term in quant finance. You may have head of bond convexity, funding convexity, Eurodollar future convexity, Libor-in-arrears convexity, CMS convexity, quanto convexity etc. So one has to be quite careful which convexity is being talked about.

The three factors work in unison to determine to shape the yield curve, and changing perceptions of expectations, amount of risk premia investor demand, and convexity determine changes in this shape. There is generally not a strong delineation of where the influence of one factor stops and another begins – they bleed into each other gradually as we go down the maturity spectrum.

It is often said that the extraction of risk premia is the primary goal of hedge funds.

7.38.1 Sources of Alpha

Each factor in "equation" (68) offers its own opportunities for earning excess returns/alpha.

First consider expectations. Having a more "correct" view on the future path of short rates (or economy in general) clearly offers opportunities for making money. These superior views can be developed by better analysis/understanding of the economic situation, understanding inter-dependencies between various parts better than the market does, and structuring trades that maximize the payoff if one's view are correct/minimize the downside. This is the essence of macro trading.

Risk premium, in many of its forms, offers rich opportunities for extraction for those willing to engage in active strategies to do so. Many sources of risk premia in the markets are created by more "traditional" real-money investors who follow various flavours of buy and hold strategies. Even simple roll-down strategies here could be profitable. To extract basic term premium one only needs to enter a long-dated received fixed swap/bond and wait. On average, this strategy will realize the term premium, i.e. above-market return. Of course, one has to be able to endure potentially very significant mark-to-market volatility while holding such a position. This is rarely feasible for banks whose risk (and hence risk capital) is tightly constrained by capital requirements and regulation. Hedge funds are relatively free from such constraints. Of course they also need to be able to control volatility of their portfolio and that is where their skills comes in, both in risk managing these types of positions, and also structuring their risk premia extraction positions to give them the best possible advantage.

Convexity, again, offers opportunities to those willing and able to work to extract them. Just buying a long-dated bond does not realize the value of convexity embedded in it – one need to essentially "gamma scalp" it. This is not what typical investors in long-term bonds such as pension funds do, thus leaving the money on the table, so to speak, for those who can/are willing to execute delta hedging strategies.

Of course, apart from fundamentally aiming to profit from risk premium and convexity, HFs seek to capitalize on

- market frictions and imperfections
- segmentation of markets
- dislocations in the markets brought about by divergent needs and constraints of other market players such as banks, institutional investors, pension funds, debt issuers and the like

7.39 P vs Q vs S, Intro

It is not so easy to link market's "lore" as relevant to yield curve investing/relative value/risk premia extraction with some solid mathematical theory. The goal of this part is to try to attempt this. Of primary interest to us is the concept of risk/term premium and strategies for extracting it. By necessity the math here is not exact and is used more as a language to describe market phenomena rather than a way to rigorously proof theorems. We closely follow Rebonato (2018) which is one of the very few books that has the same goals as this lecture.

7.40 The Origins of Risk Premia

(Rebonato, 2018, p241)

Derivatives pricing theory, fundamentally, answers the following question – if an apple is worth x and an orange is worth y, what is the price of a basket of one apple and one orange? This is done via replication. A rather different strand of economics theory tries to answer a different question – what *is* the price of an apple? Let us have a quick foray into the latter to gain some insights into risk premium. Risk premium is an economic concept that is not important for derivatives pricing where replication via "perfect hedging" rules the day; but it does matter, a lot! for investing, where it is a key consideration.

7.40.1 Utility Functions and Risk Premia

A utility function is an economic concept that measures the degree of satisfaction of an economic agent with his consumption/wealth. It is generally believed that the utility function is not (in fact, sub) linear. It is also not unreasonable to assume that it is separable in time i.e. the satisfaction of consuming X at time t and consuming Y at time $t+1$ is the sum of the satisfactions of the two.

It is also not unreasonable to assume that consuming today is preferable to consuming the same amount tomorrow – this is expressed by a "discount factor"

$$\beta = \exp(-\delta \Delta t)$$

that relates the utility of consuming now vs. later.

Consider an investment opportunity (say, stock) of time-t price S_t, and a utility function that is a function of C, the "consumption", over the two-time period $\{t, t+1\}$. One unit of stock gives a stochastic payoff x_{t+1} at time $t+1$. The investor can consume at time t and $t+1$. She has total wealth today of W and has to decide how much to consume at time t and $t+1$, to derive maximum satisfaction. Her total utility over two times is given by

$$u(C_t, C_{t+1}) = u(C_t) + \beta \mathrm{E}_t \left(u(C_{t+1}) \right).$$

The investor invests a part of its wealth into the stock at time t and "consumes" the result at time $t+1$. The utility is then given by

$$u(C_t, C_{t+1}) = u(W - aS_t) + \beta \mathrm{E}_t \left(u(a x_{t+1}) \right),$$

where we have

$$C_t = W - aS_t, \quad C_{t+1} = a x_{t+1}.$$

Here a is the amount invested in the stock. To maximize the total utility she would maximize over a, or basically solve

$$\frac{\partial}{\partial a} u(C_t, C_{t+1}) = 0.$$

This amounts to

$$\frac{\partial u}{\partial C_t} \frac{\partial C_t}{\partial a} + \beta \mathrm{E}_t \left(\frac{\partial u}{\partial C_{t+1}} \frac{\partial C_{t+1}}{\partial a} \right) = 0$$

which we simplify as

$$-u'(C_t) S_t + \beta \mathrm{E}_t \left(u'(C_{t+1}) x_{t+1} \right) = 0.$$

Thus, the price of the asset (the "apple") is given by

$$S_t = \beta \frac{\mathrm{E}_t \left(u'(C_{t+1}) x_{t+1} \right)}{u'(C_t)}$$

or, defining the *marginal rate of substitution* m_{t+1} by

$$m_{t+1} \triangleq \beta \frac{u'(C_{t+1})}{u'(C_t)},$$

we get

$$S_t = \mathrm{E}_t \left(m_{t+1} x_{t+1} \right). \tag{69}$$

7.40.2 Risk Premia is a Covariance

This can be further rewritten as

$$S_t = \mathrm{E}_t \left(m_{t+1} \right) \mathrm{E}_t \left(x_{t+1} \right) + \mathrm{cov}_t \left(m_{t+1}, x_{t+1} \right). \tag{70}$$

Moreover, $\mathrm{E}_t\left(m_{t+1}\right)$ is just the value of a risk-less investment today ($x_{t+1} \equiv 1$) that we can denote as $P(t, t+1)$, the discount bond value at time t. Thus,

$$S_t = P(t, t+1)\mathrm{E}_t\left(x_{t+1}\right) + \mathrm{cov}_t\left(m_{t+1}, x_{t+1}\right). \tag{71}$$

This can be interpreted as

> Price of a security today is given by the expectation of the payoff discounted at the risk-free rate plus a term that is a covariance between the payoff of the security, and the stochastic discount factor that measures the preferences of the investor's consumption later vs now.

The second term on the right is, clearly, just the risk premium; should investors be insensitive to risk, the first term would be the correct price of the asset.

7.40.3 Risk Premia for Returns

The equation (71) can be re-written in terms of the rates of returns rather than the levels of assets. Let

$$r(t, t+1) \triangleq \frac{1}{P(t, t+1)} - 1$$

be the one-period risk-free rate,

$$R_{t+1} \triangleq \left(\frac{x_{t+1}}{S_t} - 1\right) - r(t, t+1) = \frac{x_{t+1}}{S_t} - \frac{1}{P(t, t+1)}$$

the asset's excess return rate, and

$$M_{t+1} = \frac{m_{t+1}}{P(t, t+1)} - 1$$

the stochastic discount rate. Then

$$\begin{aligned}
P(t, t+1) &= \frac{1}{1 + r(t, t+1)}, \\
x_{t+1} &= R_{t+1} S_t + \frac{S_t}{P(t, t+1)}, \\
m_{t+1} &= \left(1 + M_{t+1}\right) P(t, t+1),
\end{aligned}$$

so that

$$\begin{aligned}
S_t = {} & P(t, t+1)\mathrm{E}_t\left(R_{t+1} S_t + \frac{S_t}{P(t, t+1)}\right) \\
& + \mathrm{cov}_t\left(\left(1 + M_{t+1}\right) P(t, t+1), R_{t+1} S_t + \frac{S_t}{P(t, t+1)}\right), \\
1 = {} & P(t, t+1)\mathrm{E}_t\left(R_{t+1}\right) + 1 + P(t, t+1)\mathrm{cov}_t\left(M_{t+1}, R_{t+1}\right), \\
\mathrm{E}_t\left(R_{t+1}\right) = {} & -\mathrm{cov}_t\left(M_{t+1}, R_{t+1}\right).
\end{aligned}$$

We can write this as follows,

$$
\begin{aligned}
\mathrm{E}_t\left(R_{t+1}\right) &= \beta_t \lambda_t, \\
\beta_t &= \frac{\operatorname{cov}_t\left(M_{t+1}, R_{t+1}\right)}{\operatorname{Var}_t(M_{t+1})}, \\
\lambda_t &= -\operatorname{Var}_t(M_{t+1}),
\end{aligned} \tag{72}
$$

where β_t is the so-called "beta" of the asset's excess return with respect to the stochastic discount rate (essentially a regression coefficient) and λ_t is called *the price of risk*, and is universal (only depends on the marginal rate of substitution m_t so the same for all assets).

7.40.4 Risk Premia, Conclusions

Fundamentally and generically (but not necessarily very usefully in practice), risk premia is driven by the correlation between our desire to consume later vs. now, and the payoff of the security.

An important corollary. If a security pays well in future states of the world where the investor expects to be able to consume little, she will pay a lot for it, driving down the return from this security. Conversely, investors will pay relatively little (and hence drive up expected returns) for those securities that pay well when the investor is already feeling rich. The point about expected returns is most clearly seen in (72): positive correlation between M and R leads to negative expected excess rate of return $\mathrm{E}_t\left(R_{t+1}\right)$ as λ is negative, and the opposite for negative correlation.

Going back to our (contrived) food analogy, you would pay a lot more for an apple that will feed you in the "famine" state of the world than for an apple that would just add to an existing glut of food during harvest.

We translate these ideas into insights on term premia for yield curves in sections to come.

7.41 The Expectation Hypothesis and The Subjective Measure

(Rebonato, 2018, p418)

7.41.1 Definitions

Expectation hypothesis (sometimes called Local EH) is an economic theory that takes many equivalent forms:

1. the expected returns from systematically investing into a long-maturity bond financed by short maturity one is zero;
2. the yield term premia is zero (see later for the definition of the yield term premia);

3. the forward curve is an unbiased predictor of future rates.

The theory has been overwhelmingly disproved empirically. In fact many, if not most, yield curve RV strategies rely on it not being true. Yet it gives us a certain framework to look at excess return, risk premium and so on.

So far we have introduced the real-world measure P that describes the probabilities as statistically observed, and the risk-neutral measure Q that is used for pricing derivatives by replication. Let us introduce yet another measure, S, which is a "subjective" probability measure. It represents the investor's beliefs on the probabilities of the future states of the world.

7.41.2 Applications to Bonds

Let us derive some heuristics that aid our intuition, without being too exact. Let $P_\tau(t) = P(t, t+\tau)$ be a τ-maturity (relative to t) bond. Suppose the (L)EH holds true. Then we can apply (69) to obtain, under the real-world measure,

$$\mathrm{E}_t^{\mathrm{P}}\left(m_{t+1} P_\tau(t+1) - P_\tau(t)\right) = 0, \tag{73}$$

as the expression above is the expected return on holding this bond and by LEH it is zero. Using the definition of the covariance we obtain that

$$\mathrm{E}_t^{\mathrm{P}}\left(m_{t+1} P_\tau(t+1)\right) = \mathrm{cov}_t^{\mathrm{P}}\left(m_{t+1}, P_\tau(t+1)\right) + \mathrm{E}_t^{\mathrm{P}}\left(P_\tau(t+1)\right) \mathrm{E}_t^{\mathrm{P}}\left(m_{t+1}\right)$$

so from (73)

$$0 = \mathrm{cov}_t^{\mathrm{P}}\left(m_{t+1}, P_\tau(t+1)\right) + \mathrm{E}_t^{\mathrm{P}}\left(P_\tau(t+1)\right) \mathrm{E}_t^{\mathrm{P}}\left(m_{t+1}\right) - P_\tau(t)$$

or

$$0 = \left(\mathrm{E}_t^{\mathrm{P}}\left(P_\tau(t+1)\right) - \frac{P_\tau(t)}{\mathrm{E}_t^{\mathrm{P}}\left(m_{t+1}\right)}\right) + \frac{\mathrm{cov}_t^{\mathrm{P}}\left(m_{t+1}, P_\tau(t+1)\right)}{\mathrm{E}_t^{\mathrm{P}}\left(m_{t+1}\right)} \tag{74}$$

For bonds we measure excess returns in yield, let us denote it $y_\tau(\cdot)$. Approximately,

$$P_\tau(t+1) = \frac{P_\tau(t)}{\mathrm{E}_t^{\mathrm{P}}\left(m_{t+1}\right)}(1 - \tau y_\tau(t+1)),$$

the first term being the $(t+1)$-forward value of the bond P_τ.Thus (74) can be re-written as

$$0 = -\mathrm{E}_t^{\mathrm{P}}\left(y_\tau(t+1)\right) + \frac{\mathrm{cov}_t^{\mathrm{P}}\left(m_{t+1}, P_\tau(t+1)\right)}{\tau P_\tau(t)}.$$

Adding and subtracting the expectation of the yield under the *subjective* measure S we obtain

$$\begin{aligned}\mathrm{E}_t^{\mathrm{S}}\left(y_\tau(t+1)\right) = \big(&\mathrm{E}_t^{\mathrm{S}}\left(y_\tau(t+1)\right) - \mathrm{E}_t^{\mathrm{P}}\left(y_\tau(t+1)\right)\big) \\ &+ \frac{\mathrm{cov}_t^{\mathrm{P}}\left(m_{t+1}, P_\tau(t+1)\right)}{\tau P_\tau(t)} = I_1 + I_2,\end{aligned}$$

where on the left-hand side we now have the expected yield, which we associate with the expected excess return, under the subjective measure.

7.41.3 Conclusions

Here comes the crux of the point of this section. The expected excess return of a (skilful) manager is equal to the difference between the returns realized under his (presumed more correct) subjective measure and the objective measure assumed by the rest of the market, plus the risk premium. This crystallizes HF claimed contributions to alpha-generation:

- Having a more informed view on the markets, i.e. their subjective probability being "better" than the objective, available to all, information/probability. This is the I_1 term and is the essence of the "macro" HF;
- Having the ability to "extract" the risk premium, which is the I_2 term. This is the essence of the "RV" HF.
- (The convexity term, which is the third pillar, is covered later).

7.42 Market Price of Risk

(Rebonato, 2018, pp197–200)

Let us now start to specialize our general concept/understanding of the risk/term premium to the case of interest rates (this will take a bit of work). Consider real-world dynamics of the interest rate curve expressed in terms of discount factors

$$dP(t,T)/P(t,T) = \mu_P(t,T)\ dt + \sigma_P(t,T)\ dW(t), \tag{75}$$

where μ_P and σ_P are sufficiently regular but otherwise arbitrary processes, and for simplicity we assume that the source of randomness dW is one-dimensional. The equation (75) is parametrized by $T > 0$.

Let us fix two maturities $T_1 \neq T_2$, and consider a portfolio of the two bonds in proportion 1 to w, i.e.

$$\Pi(t) = P(t,T_1) + wP(t,T_2).$$

Then

$$\begin{aligned} d\Pi(t) &= dP(t,T_1) + w\ dP(t,T_2) \\ &= (\mu_P(t,T_1)P(t,T_1) + w\mu_P(t,T_2)P(t,T_2))\,dt \\ &+ (\sigma_P(t,T_1)P(t,T_1) + w\sigma_P(t,T_2)P(t,T_2))\,dW(t). \end{aligned}$$

We can choose w so that the dW term disappears:

$$\sigma_P(t,T_1)P(t,T_1) + w\sigma_P(t,T_2)P(t,T_2) = 0,$$

which resolves to

$$\hat{w} = -\frac{\sigma_P(t,T_1)P(t,T_1)}{\sigma_P(t,T_2)P(t,T_2)}. \tag{76}$$

Then

$$d\Pi(t) = (\mu_P(t,T_1)P(t,T_1) + w\mu_P(t,T_2)P(t,T_2))\,dt + 0 \times dW(t)$$

and Π is (locally) riskless. Hence, *in the absence of arbitrage opportunities*, it must grow at the risk-free rate:

$$\mu_P(t,T_1)P(t,T_1) + \hat{w}\mu_P(t,T_2)P(t,T_2) = r(t)\Pi(t) = r(t)\left(P(t,T_1) + \hat{w}P(t,T_2)\right).$$

Then

$$\left(\mu_P(t,T_1) - r(t)\right)P(t,T_1) = \hat{w}\left(r(t) - \mu_P(t,T_2)\right)P(t,T_2)$$

or, using (76),

$$\frac{\mu_P(t,T_1) - r(t)}{\sigma_P(t,T_1)} = \frac{\mu_P(t,T_2) - r(t)}{\sigma_P(t,T_2)}.$$

Note this holds for any T_1, T_2, and hence we have a universal (same for all T) constant $\lambda(t)$ such that

$$\lambda_r(t) = \frac{\mu_P(t,T) - r(t)}{\sigma_P(t,T)} \text{ for any } T.$$

This quantity is called the *market price of risk*. In a multi-factor/multi-shock model of interest rates, there would be one market price of risk per source of uncertainty (HW: derive).

How do we interpret this quantity(ies)?

> MPR is the extra return over the risk-free rate $\mu_P - r$ per a unit of risk σ_P demanded by a risk-averse investor. Also known as *Sharpe ratio*. This is how much the market compensates the investor for bearing one unit of risk to which interest rates are exposed.

MPR could be positive (as we normally think of it) or negative if bonds, say, provide strong diversification benefits to a bigger portfolio (of equities, commodities, crypto, etc.). Can only assess the sign via empirical studies – no pre-requisite that it has to be positive.

7.43 Excess Return

(Rebonato, 2018, p200)

If the yield curve is driven by a short rate,

$$dr(t) = \mu_r(t)\ dt + \sigma_r(t)\ dW(t)$$

so that bonds are functions of the short rate,

$$P(t,T) = P(r(t),t,T)$$

then clearly

$$dP(t,T) = O(dt) + \frac{\partial P(t,T)}{\partial r(t)}\sigma_r(t)\ dW(t).$$

On the other hand

$$dP(t,T) = O(dt) + \sigma_P(t,T)P(t,T)\ dW(t).$$

Thus

$$\sigma_P = \frac{1}{P}\frac{\partial P}{\partial r}\sigma_r$$

and the *excess return* is given by

$$\mu_P - r = \lambda_r \sigma_P = \frac{1}{P}\frac{\partial P}{\partial r}\lambda_r \sigma_r. \tag{77}$$

Hence, expected excess return is the product of three terms:

- Duration (this is how the delta for IR products sometimes called) $(\partial P/\partial r)/P$ that measures how much the T-maturity bond changes when short rate changes;
- a "compensation term" λ_r that specifies how large is compensation per unit of risk in the shock factor;
- and a "risk magnitude" term σ_r.

This is important – excess return is one of the few empirical bridges we can find between the P and the Q measures. It also strongly constrains the risk compensation that can be assigned to bonds of different maturities.

MPR can depend on time, common state variables, etc., but does not depend on the specific features of a particular bond such as, say, its coupon or maturity. The excess return of course depends on maturity but only through the duration term $(\partial P/\partial r)/P$.

Risk/term premia is one of the three fundamental building blocks in our description of the yield curve. So it is important to understand how we can assign them, and how we cannot. Specifically, we cannot assign preferences to specific bonds without considering others – excess returns from all the bonds must be simultaneously explained by a common set of risk factors, and by a consistent set of (model-dependent) durations. This is not an easy task.

7.44 Excess Returns and Term Premia

(Rebonato, 2018, p433)

Let us now assume that the economy is shocked by n factors. The expected return (in real world measure) earned from entering a T-maturity bond, over the time t to $t + dt$, is given by

$$\mathrm{E}_t^P\left(\frac{dP(t,T)}{P(t,T)}\right) = \left(r(t) + \sum_{i=1}^{n}\frac{1}{P(t,T)}\frac{\partial P(t,T)}{\partial x_i}\lambda_i\sigma_i\right)dt$$

(this is a multi-factor generalization of (77)). Here λ_i and σ_i are the market price of risk and the volatility associated with the i-th factor, $i = 1, \ldots, n$. The term $\frac{1}{P(t,T)} \frac{\partial P(t,T)}{\partial x_i}$ is the "duration" $D_i(t,T)$ associated with the i-th factor.

Let us suppose that after a period dt we continue investing into the bond, which how has maturity $T - dt$. Continuing till the final maturity, we obtain that the total expected return is given by the integral of the expected return over each period of time, i.e.

$$\mathrm{E}_0^P \int_0^T \frac{dP(s,T)}{P(s,T)} = \mathrm{E}_0^P \int_0^T \left(r(s) + \sum_{i=1}^n D(s, T-s)\lambda_i(s)\sigma_i(s) \right) ds.$$

Neglecting convexity (the second-order derivative in Ito's lemma), i.e. writing

$$\frac{dP(s,T)}{P(s,T)} \approx d \log P(s,T)$$

on the left-hand side, integrating, and using $\log P(0,T) = -y(0,T)T$, we obtain

$$y(0,T) = \frac{1}{T} \mathrm{E}_0^P \int_0^T \left(r(s) + \sum_{i=1}^n D(s, T-s)\lambda_i(s)\sigma_i(s) \right) ds.$$

HW: carry out all the steps

The quantity on the left-hand side is the market yield on a T-maturity bond, so let's call it so: $y(0,T) = y_{\mathrm{mkt}}(0,T)$. Let us denote by $y_{\mathrm{eh}}(0,T)$ the following quantity,

$$y_{\mathrm{eh}}(0,T) = \frac{1}{T} \mathrm{E}_0^P \left(\int_0^T r(s) ds \right).$$

Here "EH" stands for "expectations hypothesis", a fancy name for the world where term yields are just expectations of the path of the future short rate. In this world there is no term/risk premium. Also one can think of $y_{\mathrm{eh}}(0,T)$ as average return from a strategy that rolls a very-short maturity bond from 0 to T. Then, finally, we can define the yield term premium $\gamma(0,T)$ to be the difference of the two,

$$\gamma(0,T) \triangleq y_{\mathrm{mkt}}(0,T) - y_{\mathrm{eh}}(0,T) = \frac{1}{T} \mathrm{E}_0^P \int_0^T \left(\sum_{i=1}^n D(s, T-s)\lambda_i(s)\sigma_i(s) \right) ds. \tag{78}$$

The formula (78), the main result of this section, links yield term premium (excess return from investing into long-tenor bonds financing it by a rolling short bond strategy) to market price of risk(s), volatilities of the factors of the economy, and a series of bond durations from 0 to T.

This is the risk premium that can be extracted by investing in long maturity bonds by a real investor, funding it by a rolling sale of short-maturity bonds, whose cost is linked exactly to $y_{\mathrm{eh}}(0,T)$.

7.45 Links to Carry and Roll-Down

(Rebonato, 2018, p425)

Practitioners often talk about excess return in terms of *carry* and *roll-down*, which are important concepts to understand. Let us define these quantities more exactly for bonds.

Consider a strategy of investing, at time t, in a T-maturity bond, financing the purchase at the short at the short rate, i.e. by selling a short-maturity bond. To set up the strategy we buy \$1 worth of the bond $P(t,T)$, which gives us the notional

$$w_1 = \frac{1}{P(t,T)}.$$

To obtain the \$1 we do not have, we sell enough of the bond $P(t,t+dt)$ to raise sufficient money; specifically we sell the notional w_2 which is given by

$$w_2 = \frac{1}{P(t,t+dt)}.$$

The value of the portfolio $K(t,T)$ at time t is zero but can more usefully be written as

$$K(t,T) = w_1 P(t,T) - w_2 P(t,t+dt).$$

We unwind the position at time $t+dt$. We then receive

$$\begin{aligned} K(t+dt,T) &= w_1 P(t+dt,T) - w_2 P(t+dt,t+dt) \\ &= \frac{P(t+dt,T)}{P(t,T)} - \frac{1}{P(t,t+dt)}. \end{aligned}$$

Given the cost of setting up our portfolio was zero, the *return* on this strategy is then given by $K(t+dt,T)$ which is simply

$$K(t+dt,T) = \frac{1}{P(t,T)}\left(P(t+dt,T) - \frac{P(t,T)}{P(t,t+dt)}\right).$$

Given that

$$1/P(t,t+dt) = e^{r(t)\ dt} \approx 1 + r(t)\ dt$$

we have

$$\begin{aligned} K(t+dt,T) &\approx \frac{1}{P(t,T)}\left(P(t+dt,T) - P(t,T) - r(t)P(t,T)\ dt\right) \\ &= \frac{P(t+dt,T)}{P(t,T)} - 1 - r(t)\ dt \\ &= \frac{P(t,T-dt)}{P(t,T)} - 1 + \frac{P(t+dt,T) - P(t,T-dt)}{P(t,T)} - r(t)\ dt. \end{aligned}$$

We would like to express the *expected* return in terms of the bond *yields*, not prices. With a bit of hand-waving (see (Rebonato, 2018, p426) for marginally less hand wavy argument)

$$
\begin{aligned}
&\mathrm{E}_t^P\left(K(t+dt,T)\right) = R(t) + C(t),\\
&\qquad R(t) \triangleq \mathrm{E}_t^P\left(\frac{P(t,T-dt)}{P(t,T)} - 1\right) \approx \left(y(t,T) - y(t,T-dt)\right)(T-t),\\
&\qquad C(t) \triangleq \mathrm{E}_t^P\left(\frac{P(t+dt,T) - P(t,T-dt)}{P(t,T)} - r(t)\ dt\right) \approx \left(y(t,T-dt) - r(t)\right)dt.
\end{aligned}
$$

The term $R(t)$ is known as "roll-down" and $C(t)$ is known as "carry". The carry term $C(t)$ measures the "funding advantage", the fact that we are buying a long-maturity bond that generates a higher yield while financing it with a sale of a short-maturity bond for which we need to pay only the short rate, typically a smaller number. This is somewhat expected. But the other term, the roll-down $R(t)$, is perhaps less expected and quantifies by how much the yield is expected to decline – and hence the price to rise – as the maturity of the bond we are holding is decreased – if the shape of the yield curve does not change (from time t to time $t+dt$). The name "roll-down" of course comes from this scenario where the yield is rolling down the curve as calendar time moves forward (while the curve holds its shape).

7.46 Why Carry and Roll-Down Matter

(Rebonato, 2018, p429)

In the previous section we expressed the expected excess return as a function of carry and roll-down, which is exactly how much of the market thinks of it. It is, however, predicated on a bold assumption that, as the calendar time moves from t to $t+dt$, the shape of the yield curve does not change. Or, less drastically, that it does not change shape *on average* – changes up and down from the shape at time t are balanced so the expected shape at $t + dt$ is more or less the same as the observed shape at time t. One can argue that choosing a single particular scenario to quantify the expected returns from an investment strategy is a foolish idea. And, it is, indeed, so. However, the assumption that the shape of the yield curve does not change on average, absent current known issues with the economy, appears to be borne by the historical data going back many years or, some would say, decades. So, if one were to choose a particular scenario, the "nothing changes" one as expressed by the constant shape of the yield curve is probably the best one can do. It is markedly different from what the risk-neutral measure – or the Expectation Hypothesis – would tell us and this is discussed in more detail later. Suffices to say the discrepancy of the two is what drives the excess returns generated by active investors/HFs engaging in RV strategies based on the yield curve. For them, the carry/roll-down measures, and their difference from the forwards-implied evolution, is of paramount importance.

7.47 Long-End Convexity: Why Long-Term Rates Are Typically Downward Sloping

7.47.1 The Basics of Convexity

(Rebonato, 2018, p148)

Let us understand the convexity in more detail. As

$$P(t,T) = \exp\left(-y(t,T)(T-t)\right),$$

we immediately notice that P is a convex function of y. If we use Ito's lemma we get, for the bond return, that

$$\begin{aligned}
&\frac{\mathrm{E}_t^P(dP(t,T))}{P(t,T)} \\
&= \frac{1}{P(t,T)}\left(\frac{\partial P(t,T)}{\partial t}dt + \frac{\partial P(t,T)}{\partial y(t,T)}\mathrm{E}_t^P\left(dy(t,T)\right) + \frac{1}{2}\frac{\partial^2 P(t,T)}{\partial y(t,T)^2}\mathrm{E}_t^P\left(dy(t,T)^2\right)\right) \\
&= y(t,T)\ dt + \frac{\partial P(t,T)}{\partial y(t,T)}\mathrm{E}_t^P\left(dy(t,T)\right) + \frac{1}{2}\left(T-t\right)^2\mathrm{E}_t^P\left(dy(t,T)^2\right).
\end{aligned}$$

Using a generic Itô process for the yields

$$dy(t,T) = \mu_y(t,T)\ dt + \sigma_y(t,T)\ dW(t),$$

we obtain

$$\frac{\mathrm{E}_t^P(dP(t,T))}{P(t,T)} = \left(y(t,T) + \frac{\partial P(t,T)}{\partial y(t,T)}\mu_y(t,T) + \frac{1}{2}\left(T-t\right)^2\sigma_y(t,T)^2\right)dt.$$

The three terms are

1. The "normal" yield on the bond $y(t,T)$ – that is what we would intuitively expect the return on buying a bond would be;
2. The term from the growth in the yield $\frac{\partial P(t,T)}{\partial y(t,T)}\mu_y(t,T)$ – this one is typically small as we wound not expect the yields to have significant drifts except in specific circumstances;
3. And the last term is the convexity term – this is an extra boost over and above the yield term $y(t,T) + \frac{\partial P(t,T)}{\partial y(t,T)}\mu_y(t,T)$ by the quantity $\left(T-t\right)^2\sigma_y(t,T)^2/2$, which is always positive.

Here $\sigma_y(t,T)$ is the absolute, or basis-point, volatility of the yield. The extra boost – the convexity – is quadratic in the maturity so it dominates the returns balance for longer maturities. This is because our expectation of rates in 30 years time are unlikely to be very precise, and the risk/term premium is more or less linear in maturity.

7.47.2 Jensen's Inequality

Let us unpack this a little bit more. One typically thinks of swaps, and bonds, as linear instruments, whose risk behaviour is fully captured by the first-order risks. While swaps are indeed linear combinations of zero-coupon bonds, and hence linear securities in the context of the HJM framework, say, their prices are actually not linear in terms of the underlying rates. This is, again, readily seen from the formula for the zero-coupon discount bond which is an exponential, and hence convex, function of the underlying forward rate

$$P(t,T) = \exp\left(-y(t,T)(T-t)\right).$$

Clearly, also, the longer the maturity, the more non-linear the relationship is between the rate and the bond value.

What are the implications of this convexity? As we have seen earlier in the notes, having a convex position in a market factor is beneficial to its holder. Specifically, as we have seen, delta-hedging a convex position makes money/delivers positive PnL on each rehedging. The profitability of a convex position depends on the amount of convexity, as well as the volatility of the underlying market variable. Hence, to enter a convex position one typically needs to pay for it; either in the form of an up-front payment or, in the case of a bond or a (receiver) swap, by accepting a lower yield/fixed rate.

To be more precise, let $V(x)$ be the the value of the payoff as a function of the realization $X = x$ of the underlying market factor X. Then, expanding to the second order around $x_0 = \mathrm{E}(X)$,

$$V(x) \approx V\left(\mathrm{E}(X)\right) + V'\left(\mathrm{E}(X)\right)\left(x - \mathrm{E}(X)\right) + \frac{1}{2}V''\left(\mathrm{E}(X)\right)\left(x - \mathrm{E}(X)\right)^2$$

so that

$$\mathrm{E}\left(V(X)\right) \approx V\left(\mathrm{E}(X)\right) + \frac{1}{2}V''\left(\mathrm{E}(X)\right)\mathrm{Var}\left(X\right).$$

We see that if the payoff is convex, $V''(x) \geq 0$, then the expected value $\mathrm{E}\left(V(X)\right)$ goes up as the convexity $V''(x)$ goes up and/or the variance of X, $\mathrm{Var}\left(X\right)$, goes up. Here is a clear link with Jensen's inequality which states that for convex $V(\cdot)$,

$$\mathrm{E}(V(X)) \geq V(\mathrm{E}(X)).$$

7.47.3 Trading Convexity

The payoff of a long-maturity receiver (recall this means receive fixed rate), hedged with a shorter-maturity payer, as a function of the underlying rate, gets more convex as the maturity increases. See notebook `T7_Long_End_Convexity_01.ipynb`. Hence, holding a longer-maturity receiver is preferred to holding a shorter-maturity one. As free lunch is generally not available in the markets, one has to pay in order to be in a favourable position such as being long convexity – in this case, one has to accept a lower fixed rate (to receive) for longer-dated swaps. Hence, it is generally expected that the yield curve is downward sloping in the long end.

Just like a long-option position (see Topic 2), managing convexity requires an active trading strategy. In essence, just like with an option, one needs to re-hedge the position as the curve moves. The ultimate PnL on such a position is given by the difference between the realized volatility and "implied" volatility, implied volatility that is essentially embedded in the (lower) fixed rate that one agrees to receive at the inception of such a position. Ultimately, then, the PnL is a bet on mis-priced "implied" vs. future realized, volatility. As with most strategies, being long convexity is not a slam-dunk investment, but a judgement call on what is priced by the market vs. how the markets will ultimately evolve. And, as we already mentioned, one needs to work to extract the value of convexity.

This creates opportunities for hedge funds. Structural demand for long-dated swaps most often comes from the need of corporates to convert their fixed payments on bonds they issue into floating ones, or from relatively passive investors such as insurance companies who need, by regulation or internal risk policies, to match their long-dated liabilities. It is typically outside of the mandate of these market participants to actively delta-hedge their long-dated swaps. This structural positions create demand for long-dated swaps and hence they, the passive market players, collectively, end up with convexity positions. These convexity positions may be either too convex, or not enough, vs. what is a "fair" value of convexity. Wherever there is a mismatch between the market value of a given "observable" (convexity in our case) and its "fair" value, HFs come into play and, in general, arbitrage it away, pushing the convexity to its fair value, making money on the way. This is just one of the example of HFs (claiming to be) making markets more efficient – without them there would be no mechanism to keep long-dated swap rates to their fair values.

Topic 8
Measuring the Value in IR Instruments

8.48 Managing Yield Curve Risk

Portfolios of interest rate securities are exposed to interest rate risk, i.e. the value of a portfolio changes when the interest rates – as expressed by a yield curve (or curves) changes. Understanding the exposure of a given portfolio to changes in interest rates is of paramount importance. For dealers/sell side, this facilitates hedging, i.e. immunizing the portfolio against interest rate changes (recall that the primary business model of sell side trading desks is to maintain a carefully hedged portfolio while making money on the bid/ask spread on the trades they do with clients). For active money managers such as hedge funds, understanding the exposures allows them to construct positions that reflect their views on the future direction of the interest rates (macro strategies) or create positions that benefit from perceived temporary dislocations in the market (relative value strategies).

Unlike a simple case of an option on a stock, say, where the first-order exposure to the underlying market is a single number, delta, that measures the change in value of an option with respect to a stock, first-order exposures of an interest rate portfolio have more than one factor. As different parts of the yield curve may move (relatively) independently, an exposure to the *whole* of the yield curve constitutes the set of sensitivities to consider.

We continue with the notations from before. Consider a portfolio of securities with value V_0, where V_0 is a function of the yield curve $y(t)$. The securities in V_0 would typically not be in the benchmark set and could contain, say, interest rate options, seasoned swaps, and so forth. As the yield curve is a function of the benchmark set values $\mathbf{V} = (V_1, \ldots, V_N)^\top$, we may write

$$V_0 = V_0\left(V_1, \ldots, V_N\right),$$

where the function $V_0(\cdot)$ is determined from the curve construction algorithm employed. Clearly, then

$$dV_0 = \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i} dV_i,$$

or, for non-infinitesimal moves,

$$\Delta V_0 \approx \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i} \Delta V_i. \tag{79}$$

For the purpose of managing first-order risk exposure to moves in the yield curve, (79) suggests that the collection of derivatives $\partial V_0/\partial V_i$, $i = 1, \ldots, N$ — often called *(bucketed) interest rate deltas* — forms a natural metric for portfolio risk. In particular, if all these derivatives are zero, our portfolio would, to first order, be immunized against any move in the yield curve that is consistent with the chosen curve construction algorithm. On the other hand, if some or all of the derivatives are non-zero, we could manage our risk by setting up a hedge portfolio of benchmark securities, with notional $-\partial V_0/\partial V_i$ on the i-th security.

8.48.1 Par-Point Approach

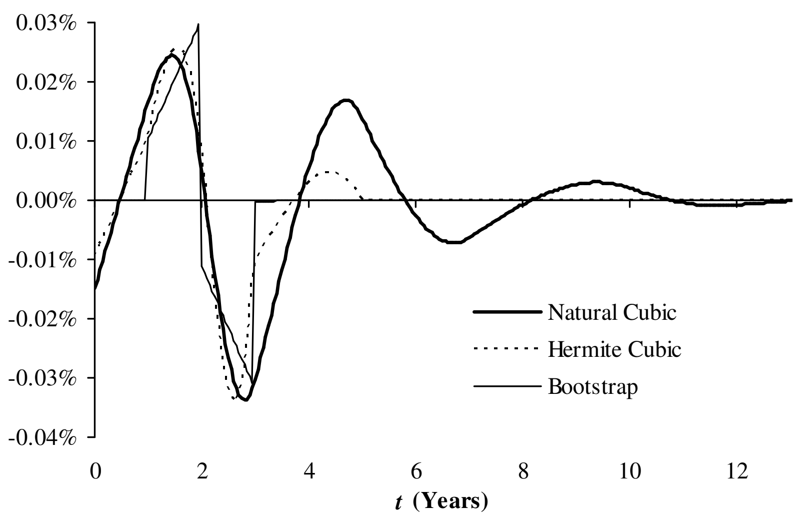
The simplest approach to the calculation of the delta $\partial V_0/\partial V_i$ involves a manual bump[12] to V_i, followed by a reconstruction of the yield curve, and a subsequent repricing of the portfolio V_0. This procedure is sometimes known as

[12] In practice, rather than bumping the price V_i outright, one may instead bump the yield of the i-th benchmark security (typically by 1 basis point).

the *par-point approach* (or *par-rate approach*), and resulting derivatives *par-point (or par-rate) deltas*. For the approach to work properly, it is important that the yield curve construction algorithm is fast and produces clean, local perturbations of the yield curve when benchmark prices are shifted. For instance, perturbing a short-dated FRA price should not cause noticeable movements in long-term yields, lest we reach the erroneous conclusion that we can perfectly hedge a 20 year swap with a 1 month FRA. As we have discussed earlier, Hermite splines and bootstrapped yield curves both exhibit good perturbation locality, but cubic C^2 splines often do not. To illustrate this, consider this Python notebook `./notebooks/products/rates/FINIBORSINGLECURVE_IborCurveApplyParBump.ipynb`. In this notebook, we have followed the recommendations of Section 4.29 and added tension to the C^2 spline, causing a dampening of the perturbation noise. Clearly, the usage of a tension factor can have a beneficial impact on risk reports produced by the par-point approach.

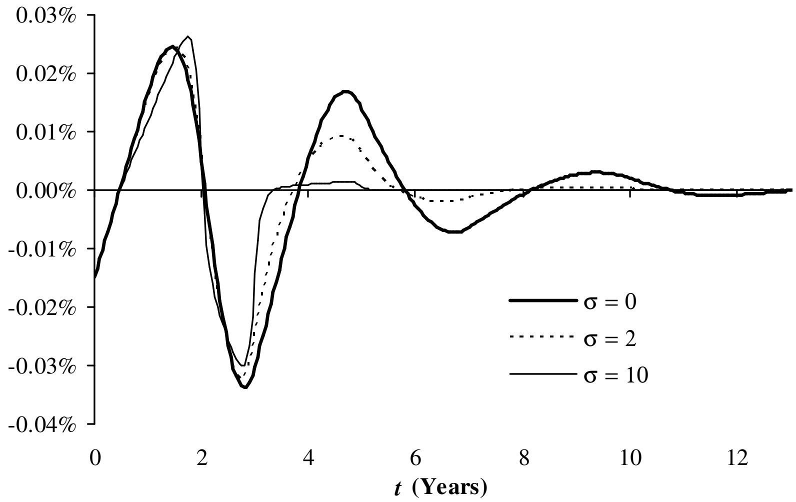
Similar results are shown in figures 8 and 9.

Figure 8: Forward Curve Move



Notes: Change in instantaneous forward curve, from a 1 basis point shift in the 2 year swap yield in Table 1. The curve construction methods tested are: bootstrapping with piecewise linear yields ("Bootstrap"), Hermite C^1 cubic spline ("Hermite"), and C^2 natural cubic spline ("Natural Cubic"). Swap data is in Table 1.

Figure 9: Forward Curve Move



Notes: Change in instantaneous forward curve, from a 1 basis point shift in the 2 year swap yield in Table 1. The yield curve was constructed as a tension spline, with tension factors as given in the graph. Swap data is in Table 1.

8.48.2 Forward Rate Approach

As an alternative to direct perturbation of benchmark security prices, we can consider applying perturbations directly to the discount curve, thereby mostly avoiding the introduction of artefacts specific to the curve construction algorithm. In practice, this technique typically focuses on the forward curve[13] $f(t)$, to which we apply certain functional shifts $\mu_k(t)$, $k = 1, \dots, K$. Writing (loosely) $V_0 = V_0(f)$ to highlight the dependence of V_0 on the forward curve, we then compute functional (Gateaux) derivatives[14] for V_0:

$$\partial_k V_0 = \left.\frac{dV_0\left(f(t) + \varepsilon \mu_k(t)\right)}{d\varepsilon}\right|_{\varepsilon=0}, \quad k = 1, \dots, K. \tag{80}$$

Standard choices for $\mu_k(t)$ are

$$\text{Piecewise Triangular:} \quad \mu_k(t) = \frac{t - t_{k-1}}{t_k - t_{k-1}} 1_{\{t \in [t_{k-1}, t_k)\}} + \frac{t_{k+1} - t}{t_{k+1} - t_k} 1_{\{t \in [t_k, t_{k+1})\}}, \tag{81}$$

$$\text{Piecewise Flat:} \quad \mu_k(t) = 1_{\{t \in [t_k, t_{k+1})\}}, \tag{82}$$

[13]Perturbations may also be performed on discretely, rather than continuously, compounded forward rates.

[14]For a proper definition of the Gâteaux derivative, see Gâteaux (1913).

where $\{t_k\}$ is a user-specified discretization grid. The resulting sensitivities are often called *forward rate deltas*.

It is common practice to use $\{t_k\}$ grids spaced three months apart, with dates on Eurodollar futures maturities. The number of deltas K is thus typically a rather large number, and the K derivatives $\partial_k V_0$ give a detailed picture of where the portfolio risk is concentrated on the forward curve. As forward rate contracts and Eurodollar futures cease to be liquid beyond 4 or 5 year maturities, the forward rate deltas do not directly suggest hedging instruments for the medium and long end of the yield curve exposure; however it is not difficult to translate forward rate deltas into a hedging portfolio using some linear algebra (see Andersen & Piterbarg (2010)). The choice of par point versus forward rate deltas is largely a matter of personal preference, and it is not uncommon for traders to use both at the same time.

It is worth noting that this risk view is often called the forward delta ladder, as it is visualized as a grid ("ladder") of dates spaced 3M apart and forward rate deltas associated with each period. See an example notebook `re_Curve_Strategies_01.ipynb`.

8.48.3 DV01 and Terminology

The sum of all forward rate deltas, normalized to 1bp bump, for one instrument or for a portfolio, is called its *DV01* (Dollar Value of 1 basis point). It is the same as the sensitivity of the portfolio to a parallel shift of the overnight forward curve, also normalized to 1bp bump. It is close, but not exactly equal in some cases, to *PV01* (Present Value of 1 basis point), which is another name for the value of the (unit notional) fixed leg/annuity in a swap.

Positive DV01 means that the instrument/portfolio gains value if the curve moves up, and loses if it moves down. Somewhat confusingly, a long bond position has negative DV01 (prices move in the opposite direction to yields). Hence, even more confusingly, a "long" position in IR markets is the one with a *negative* DV01, and a "short" one with a positive DV01. Likewise, a "rally" in rates markets is when the curve is moving *lower*, and a "sell-off" is when the curve is moving higher. All this comes from thinking about bond prices not rates (i.e. bond prices go up in a rally).

8.48.4 PnL, Basic Notions

The value of a portfolio of derivatives, specifically linear interest rate derivatives such as FRAs, futures, swaps, and bonds fluctuates day to day, as the interest rate curve(s) move in response to changing market conditions. This changes in the portfolio value are customarily called PnL (aka PNL, P&L) which stands for "Profit and Loss". A "realized" PnL is exactly what it sounds like – the actual change in the value of a portfolio between two dates, often one day (daily realized PnL but also over a month or a quarter or a year. Depending on the application this could be narrowly defined as just the change in value of the portfolio, or include other components such as cash paid and received on coupons, funding costs, fees associated with trading, and so on. Quantifying the profitability of any business is, naturally, of key importance, so keeping track and understanding PnL and its drivers is a basic requirement for any trading business, be that on the buy or sell side.

Apart from the headline PnL for the whole portfolio, various granular decompositions of PnL are typically required. A simplest decomposition is trade-by-trade, i.e. change in value of each individual trade. This is often not very useful for dealers due to a large volume of trades. Hedge funds might find it more useful especially when specific trades express specific trading strategies such as various relative value strategies (more on that later). HFs may also aggregate their PnL by "theme" or strategy when multiple individual trades are used to express the same or similar view.

Much more universally useful and, invariably, required in pretty much all settings is the decomposition of the portfolio PnL into its risk drivers. For example, a change in an interest rate curve can often be understood as a change in the overall level of interest rates plus a change in its slope plus a change in its curvature. Which proportion of the PnL should be attributed to each of these changes in the curve?

We will not go into much detail here as the topic of PnL calculation, attribution, explain and prediction is a rich one (see (Andersen & Piterbarg, 2010, Chapter 22)for more details). The first-order risk sensitivities that we just discussed serve as the basis for understanding the PnL In fact, we have already seen the basic PnL equation, see (79). The equation simply states that PnL, to first order, is given by risk sensitivities to certain rates (be that par rates or forward rates) time the change in those rates from one day to the next.

This is indeed but the simplest form of the PnL equation. Second-and higher order terms can be added (relevant even for linear products). Cross-terms are often important in realistic situations when multiple curves are used to value a portfolio. Understanding cash movements, i.e. the actual money going in and out of the door due to coupon payments and the like is an intrinsic, and often fiddly, part of the PnL. Trades can be cancelled, amended, new trades can (and are) added intraday, fees and bid/ask are paid etc.

For our purposes here we will mainly stick to (79) with one important addition that we shall discuss next.

8.49 Theta

8.49.1 Constant Forwards Theta, Motivation

What happens to an interest rate derivative, or a portfolio, over the course of one day (say), if *nothing changes*, i.e. if the interest rates do not move? Clearly we would expect some change. If we buy a zero-coupon discount bond for maturity T we pay $P(0,T)$ which is < 1 (if interest rates are positive!) and if "nothing changes" between now and T we would expect 1 back at T. So there typically should be some PnL associated with purely the passage of time.

In the Black-Scholes option theory, this contribution to PnL is measured by the Greek "theta", which is defined as the change in value of the option V due to the passage of time, all other Black-Scholes parameters (spot, rate, volatility, dividend yield) being the same, i.e. $\partial V/\partial t$. The situation is not so simple for interest rate derivatives as it is not entirely obvious what are the parameters here that should be kept constant when the time t changes. So, for this definition to be properly extended to interest rate derivatives, "nothing changes but the time" needs to be properly defined.

Let us first consider this task from the "quant" point of view. (We shall consider other points of view later on.) We start with a simple discount bond. The bond price for maturity T today is $P(0,T)$ which, in terms of forward rates, is given by

$$P(0,T) = \exp\left(-\int_0^T f(0,u)\ du\right).$$

Now suppose we want to enter into a contract where, at time $T' < T$, we receive $P(T',T)$, i.e. the same discount bond, in exchange for the payment K also at time T'. What value of K makes this contract have zero value at time 0? Of course this is nothing but a forward contract on a discount bond.

We have

$$0 = \mathrm{E}^Q\left(\beta(T')^{-1}\left(P(T',T) - K\right)\right)$$

where E^Q is the risk-neutral measure, i.e. the measure that corresponds to the money-market numeraire $\beta(\cdot)$. As $\beta(T')^{-1}P(T',T)$ is a martingale in this measure, we clearly have

$$0 = P(0,T) - KP(0,T')$$

so that the fair "price" K for this contract is the value of the forward bond

$$K = \frac{P(0,T)}{P(0,T')} = P(0,T',T)$$

which, in terms of the instantaneous forward curve $f(\cdot)$, is given by

$$P(0,T',T) = \exp\left(-\int_{T'}^T f(0,u)\ du\right).$$

Arguably, then, it is reasonable to define "nothing changes" as the forward bond price $P(s,T',T)$ does not change for $s \in [0,T']$.

8.49.2 Constant Forwards Theta, Swaps

Let us calculate forward prices of other linear interest rate derivatives. Consider a fixed rate payer fixed-floating swap at time T', $0 \le T' \le T_0$. Its value at any time T' before the start of the swap, $T' < T_0$ is given by (using notations from before)

$$V_{\mathrm{swap}}(T') = \sum_{n=0}^{N-1} \tau_n P\left(T', T_{n+1}\right)\left(L_n(T') - k\right). \tag{83}$$

Its time$-T'$ forward value, set at time 0, $V_{\text{swap}}(0,T')$ is given then

$$V_{\text{swap}}(0,T') = \frac{1}{P(0,T')}\mathrm{E}^Q\left(\beta(T')^{-1}V_{\text{swap}}(T')\right).$$

However,

$$\begin{aligned}\mathrm{E}^Q\left(\beta(T')^{-1}V_{\text{swap}}(T')\right) &= \mathrm{E}^Q\left(\beta(T')^{-1}\sum_{n=0}^{N-1}\tau_n P\left(T',T_{n+1}\right)\left(L_n(T')-k\right)\right)\\ &= \mathrm{E}^Q\left(\beta(T')^{-1}\sum_{n=0}^{N-1}\tau_n P\left(T',T_{n+1}\right)\left(\frac{P\left(T',T_n\right)-P\left(T',T_{n+1}\right)}{\tau_n P\left(T',T_{n+1}\right)}-k\right)\right),\end{aligned}$$

where in the last equation we used the definition of $L_n(T')$. Thus

$$\begin{aligned}&\mathrm{E}^Q\left(\beta(T')^{-1}V_{\text{swap}}(T')\right)\\ &\qquad = \mathrm{E}^Q\left(\beta(T')^{-1}\sum_{n=0}^{N-1}\left(P\left(T',T_n\right)-P\left(T',T_{n+1}\right)-k\tau_n P\left(T',T_{n+1}\right)\right)\right)\end{aligned}$$

and we see that the forward swap value is just a linear combination of discount bonds (since, no surprise, it is a *linear* derivative). Therefore, using our result for forward bonds,

$$\begin{aligned}V_{\text{swap}}(0,T') &= \frac{1}{P(0,T')}\mathrm{E}^Q\left(\beta(T')^{-1}V_{\text{swap}}(T')\right)\\ &= \sum_{n=0}^{N-1}\left(P\left(0,T',T_n\right)-P\left(0,T',T_{n+1}\right)-k\tau_n P\left(0,T',T_{n+1}\right)\right).\end{aligned}$$

Finally, recalling the definition and after a trivial division by $P(0,T')$ of the numerator and denominator,

$$L_n(0) = \frac{P\left(0,T_n\right)-P\left(0,T_{n+1}\right)}{\tau_n P\left(0,T_{n+1}\right)} = \frac{P\left(0,T',T_n\right)-P\left(0,T',T_{n+1}\right)}{\tau_n P\left(0,T',T_{n+1}\right)}$$

we obtain

$$V_{\text{swap}}(0,T') = \sum_{n=0}^{N-1}\tau_n P\left(0,T',T_{n+1}\right)\left(L_n(0)-k\right). \tag{84}$$

It is instructive to compare (84) to the spot value of the same swap at time 0 (from (83) with $T'=0$),

$$V_{\text{swap}}(0) = \sum_{n=0}^{N-1}\tau_n P\left(0,T_{n+1}\right)\left(L_n(0)-k\right).$$

We see that the Libor rates being used are the same in both expressions, but the spot discount factors for each float-for-fixed rate exchange are replaced by the T'-forward values,

$$P(0,T) \to P(0,T',T), \quad \exp\left(-\int_0^T f(0,u)\ du\right) \to \exp\left(-\int_{T'}^T f(0,u)\ du\right)$$

for all T.

Arguably, again, it is reasonable to define "nothing changes" as the forward bond price $P(s,T',T)$ not changing for $s \in [0,T']$.

8.49.3 Constant Forwards Theta, Definition

These observations lead to the following natural definition of theta for interest rate derivatives, often called the *constant forwards theta* (as there are other kinds). A time-T' theta is defined as the change in value of a derivative when valued by the "along the forward" modification of the yield curve $y^{T'}(\cdot)$ that, in terms of discount factors $\left\{P^{T'}(T), T \geq 0\right\}$ is given by

$$P^{T'}(T) = \frac{P(0,T'+T)}{P(0,T')}, \quad \text{for all } T' \geq 0.$$

In instantaneous forward rate terms, the yield curve $\tilde{y}(\cdot)$ is given by $\left\{f^{T'}(u), u \geq 0\right\}$

$$f^{T'}(u) = f(0,T'+u), \quad \text{for all } u \geq 0.$$

Theta for the derivative with the value V that is a function of a yield curve y, $V = V(y)$, is then defined by

$$\theta_{\text{constfwd}} = \left.\frac{\partial V(y^{T'})}{\partial T'}\right|_{T'=0}$$

or, more practically,

$$\theta_{\text{constfwd}} = V(y^{T'=1\text{day}}) - V(y^{T'=0}).$$

8.49.4 HJM: A Brief Aside

Let us briefly recall the HJM framework. In this framework,

$$dP(t,T)/P(t,T) = r(t)\ ds - \Sigma(t,T)\ dW(t)$$

and

$$df(t,T) = \Sigma(t,T)\sigma(t,T)\ ds + \sigma(t,T)\ dW(t),$$

where

$$\sigma(t,T) = \frac{\partial \Sigma(t,T)}{\partial T}.$$

Calculating the constant forwards theta is then equivalent to setting the forward rate volatility $\sigma(t,T)$ to 0 for $t \in [0, T']$,

$$\sigma(t,T) \equiv 0, t \in [0, T'],$$

and calculating the change in value of our derivatives portfolio in this model between the times $t = 0$ and $t = T'$.

HW: prove this is the case

8.49.5 Theta and PnL

If there is PnL if rates do not move – and there usually is – then the term needs to be included in the PnL equation (79). In fact arguably this is the most important term as, in a sense, it is the zeroth-order term. So formally we need to write

$$\Delta V_0 \approx \frac{\partial V_0}{\partial t} + \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i} \Delta V_i.$$

There are some subtleties here, however. Specifically, the definition of bumps ΔV_i *needs to be consistent* with the definition of theta. Our definition of constant forwards theta is consistent with the use of forward deltas defined in Section 8.48.2. For theta, we keep the forward rates constant, and then we measure the market changes in terms of the changes of these forward rates from their time-0 (i.e. those used in theta) values.

Things get more complicated if we want to use par risk for PnL. Critically, under the constant forwards theta definition, the values of benchmark securities will change. Hence, when we measure the response to the market move, we need to calculate the change in par rates not from day t to day 0, but from day t to their values under the forward curve rolled to t! If we do not do that, some PnL will be double-counted in theta and in deltas.

We do not explore these issues further as this becomes quite specialized and fiddly, but hopefully this brief discussion gives some appreciation of the need for consistency in how various risk measures are calculated.

8.50 Carry for Swaps

8.50.1 General Considerations

Interest rate securities, by their nature, pay coupons periodically. In a large portfolio, there will be payments of coupons (pay or receive) happening every day. These, too, need to be added to the PnL equation, although this is not our main motivation for discussing this topic. Quantifying cash in/out movements is important for properly assessing trading strategies. This is especially true for HFs that need to be careful with cash management (it is less of a concern for dealers where they typically can go to their bank's treasury to borrow/deposit a practically unlimited amount of cash).

The concept of *carry* is one of the metrics used to quantify immediate cash needs (or benefits) of a given position. It is often used as one of the "lenses" to assess a given trading strategy. A position may be attractive from an expectation of the future payout perspective, but if it requires significant cash outlays until it potentially pays out, this may alter the risk/reward considerations.

We have already seen carry for bonds, see Section 7.45 where we defined it, in a somewhat idealized way, as the difference between the yield on the bond and the risk-free rate, identifying it as the funding advantage of holding a long-term bond financed with a short-term one. Let us extend this to swaps.

8.50.2 Libor Swaps

For swaps and similar securities, carry is defined as the value of all *known* cashflows over the next three (or six, or twelve, depending on the convention) months. Consider a Libor swap that starts at T_0 and pays at $T_1, \ldots, T_N$, just like we used before but with notional H. The payment at time T_1 is known at time T_0 and is equal to $H\tau_0 \left(L_0(0) - k\right)$, i.e. the difference between the first Libor fixing (known at time $t = 0$) an the fixed rate, times the notional and the year fraction. This is then the carry of this swap:

$$\text{carry} = H\tau_0 \left(L_0(0) - k\right).$$

The carry of a portfolio is, naturally, the sum of the same over all securities in the portfolio.

As the value of the spot-starting swap is equal to the sum of the values of the first-period cashflows and a swap that starts from the second period, carry can then be also calculated as

$$\text{carry}\left(\text{swap}, 1, N\right) = V_{\text{swap},1,N} - V_{\text{swap},2,N} \tag{85}$$

or, in rate terms,

$$\text{carry}\left(\text{swap}, 1, N\right) = \frac{1}{H\tau_0}\left(V_{\text{swap},1,N} - V_{\text{swap},2,N}\right). \tag{86}$$

Here $V_{\text{swap},n,m}$ is the value of the swap with the first payment date T_n and the last T_m.

8.50.3 Other Securities

The definition of carry is relatively straightforward for Libor swaps but gets a bit more complicated for OIS swaps that are now the most popular type of swaps, post-Libor reform. As one recalls, a compounded in-arrears OIS rate (the standard way to define the floating leg in an OIS swap) is only known at the end of the period, right before the payment date, unlike a Libor rate that fixes in the beginning. To have a compatible notion of carry, its definition for OIS swaps is stretched a bit to include *projected*, rather than certain, payments over the next 3 (6, 12) months. Hence it is calculated using forward term OIS rates ($R_0(0)$ in the notations of before) where the Libor fixing would be used for Libor swaps.

Often carry is expressed in rate terms, and is measured in basis points, not in $'s. Typically, the rate carry is annualized i.e. expressed in bps/year. This facilitates comparing carry across instruments for which different tenors for calculating carry are more natural (e.g. 3M for USD and 6M for GBP). For the swap considered above, the annualized carry in rate terms is then simply $(L_0(0) - k)$.

Carry of futures and FRAs (not spot-starting) is, naturally, 0. We already defined carry for bonds (Section 7.45) but in somewhat simplistic way. When being long a bond, one receives the coupon but seemingly pays nothing – so is the carry of the bond just the coupon rate? Well, it is not. Buying a bond requires an outlay of cash, unlike entering a swap. This cash needs to be obtained from somewhere (or, its other uses need to be abandoned). This costs money – the cost of borrowing money to buy something is called *funding cost*. For bonds specifically, cash for buying a bond can be obtained by borrowing the money secured by the bond one just bought – this is called a repo (re-purchase) transaction and the corresponding rate on the cash loan is called the repo rate. Hence, the carry (in rate terms) of the bond is defined as the (annualized) coupon rate, minus the annualized repo rate.

8.51 Roll-Down

The assumption that when "nothing happens", the yield curve just rolls along the forward curve is the right one for pricing derivatives by linking them to the basic building blocks of the interest rate markets, i.e. for the hedgers – this is how the risk-neutral measure and the HJM framework arise. As discussed, this is not really how the curve in real life evolves and the assumption of evolving along the forward curve, known as the expectation hypothesis, has been soundly rejected by empirical tests. Apart from empirical tests from academics, it is also interesting to look at the mechanics of the market, and the typical behaviour of market participants, to understand why the curve behaves the way it does.

8.51.1 Curve Rolls: Typical Mechanics

Let us first briefly talk about some of the mechanics of the market. Each trading day, in each particular currency, trading in interest rate derivatives largely stops in the late afternoon, around 5-6pm. This is largely driven by exchange open hours as primary hedging and benchmark instruments, such as ED futures and bond futures, are traded on the exchange For example, ICE, where Sonia futures are traded, is open 7:30am to 6pm. While each market participant does it slightly differently, almost always there is a concept of a "closing curve" or "EOD curve" (EOD is End of Day) which is the curve built from the market quotes available just before the markets close. This curve is particularly important as it is used to determine that end of day PnL and feeds various official systems and results such as the ledger, balance sheet, etc.

When the markets open the next day, the calendar time has moved by one day. Not much trading activity, or news arriving, happens overnight. Hence, the SOD curve (SOD stands for "start of day") is a typical example of how the curve rolls (in practice). Inter-dealer brokers, as we recall, are intermediaries between dealers that arrange the bulk of OTC transaction, and they provide what is often considered most accurate market prices for benchmark securities to other market participants that are the main inputs into everyone's curves. Brokers, typically, just *keep the quotes on benchmarks* constant overnight, unless some major news broke. Let us unpack this a little. One of the benchmarks is the 10 year swap rate. On day t, the 10y rate corresponds to a swap that covers the period from t to $t + 10y$. The next day, the 10y swap rate refers to a *different* instrument, a swap that covers the period from $t + 1d$ to $t + 1d + 10y$. It is not hard to see that this is inconsistent with the constant forwards assumption for the curve evolution in the absence of market changes.

8.51.2 Curve Rolls: Market Expectations

The notion that benchmark instruments do not change price (rate) overnight is an intuitively appealing one, and not just to inter-dealer brokers. A typical hedge fund manager, if asked what would happen to the curve if no news arrived over the course of a day, would say that 10y (5y, 30y, etc.) swap rate would stay constant.

The idea that the curve just rolls along its forward does start to look unintuitive (ridiculous?) when we consider longer time shifts. Imagine we move forward 5 years with no news affecting interest rate markets. Then short rates in 5 years time would be what 5y forward rates are today, which are, in a typically upward sloping interest rate environment, significantly higher. The curve would also look remarkably flat.

The idea that a short term lending rate in 5 years time would be significantly higher than today if nothing happens to interest rate markets does indeed stretch credulity. This is especially true if one considers history, and how the curve, mostly, has been upward sloping over long stretches of history.

As we already discussed, forward rates for longer tenors are usually significantly higher than forward rates today because of the risk/term premia – market participants face higher uncertainty over longer period of time and hence require a premium for holding longer-term debt – meaning longer rates are higher than shorter ones. One can argue that "nothing changes" can refer to the amount of risk premia embedded in the curve and if that does not change, we expect the same difference between the short and long term rates to persist in the future.

In this discussion we see another demonstration of what really drives many, if not most, of the trading strategies employed by HFs and, to some extent, other market participants, and that is the discrepancy between the P and the Q measures. The Q measure is, of course, the risk-neutral measure that assigns values to securities based on no-arbitrage/hedging arguments. The P measure is the real-world, or statistical, measure that assigns values to securities based on risk preferences and expectations of market participants that are, often heavily, also influenced by history. At the most fundamental level dealers sell instruments at prices determined by the risk neutral measure as they by and large run well-hedged books. HFs have views on where the market is going and assign potentially different values to the same securities based on historical observations and also their beliefs (recall the subjective measure S).

8.51.3 Roll-Down For Swaps

Constant forwards theta measures the change in value if a position of the curve rolls along its forwards. Constant spot theta (terminology may vary) measures the same under the assumption that the yield curve (or the discount bond curve) remains the same as the time passes. In our notations it corresponds to defining the yield curve $\{y^{T'}(t), t \geq 0\}$ at time T' by

$$y^{T'}(t) = y(t),\ t \geq 0, \tag{87}$$

or

$$P^{T'}(0,t) = P(0,t),\ t \geq 0.$$

It could also mean that benchmark rates that go into the curve construction remain the same for the "same" benchmarks, i.e. 10y rate stays the same. Since some of the benchmarks actually use fixed dates rather than tenors, such as ED futures or MPC FRAs, this definition is not quite identical to our definition above. We do not explore this point further here.

Dealers, being mostly the citizens of the Q world, may still use the spot theta as another view on the characteristics of their portfolio. If it is used as part of PnL explain, then the definitions of market data shocks need to be consistent with the definition of theta, as we already discussed. Par-point risk is mostly consistent with spot theta; if forward deltas are needed, the forward rate bumps need to be measures not from the time-T0 curve, but from time-t curve that was constructed using the assumptions underpinning the spot theta.

Where the spot theta really gets a lot of use is in the realm of HFs and other active investors in fixed income markets. A variant of the measure, called *roll-down*, is a common tool, as well as *lingua fascia*, in that community.

Roll-down measures the change of value in a security/portfolio over a given period of time if the spot curve at the end of that period is the same as the spot curve at the beginning of the period (i.e. now). Consider a $9y$ swap starting in $1y$ time. Then its one-year roll-down is simply given by the difference of today's values of a $1y9y$ forward swap and a $9y$ spot-starting swap. This is easily seen because the $1y9y$ forward-starting swap in one year will become a $9y$ spot starting swap, and if the curve is the same as today, its value will be the same as today's $9y$ spot-starting swap.

In the notations used in (85), we have

$$\text{rolldown}\,(\text{swap}, 2, N) = V_{\text{swap},1,N-1} - V_{\text{swap},2,N}.$$

A spot-starting swap would have both carry and roll-down. The carry has already been defined (essentially the PV of the first-period cashflows). The roll-down is then defined for the rest of the swap, i.e. the forward-starting portion of it:

$$\mathrm{rolldown}\,(\mathrm{swap}, 1, N) = \mathrm{rolldown}\,(\mathrm{swap}, 2, N) = V_{\mathrm{swap},1,N-1} - V_{\mathrm{swap},2,N}.$$

As for the carry, roll-down is often expressed in rate terms and is often annualized. For an individual swap, the rate version of roll-down is obtained by dividing the cash value of roll-down by the parallel DV01 of the swap (or PV01 depending on convention, which is roughly similar) and then by the year fraction for the period over which the roll-down is considered. For a portfolio, the roll-down in rate terms is not well-defined.

HW why is it so?

8.51.4 Roll-Down and Forward Delta Ladders

A forward delta ladder, i.e. a bucketed DV01 of a portfolio, obtained by bumping successive 3M (6M,12M) forward rates, is a convenient tool for estimating roll-down for swaps and other instruments.

Let $\{t_k\}_{k=1}^{K}$ be the grid used to calculate the forward delta ladder, spaced τ years apart (e.g. $\tau = 1/4$ for $3M$ spacing), and $t_0 = 0$. Let $\partial_k V_0$, $k = 1, \ldots, K$, be the forward rate deltas on this grid for the portfolio V_0, where ∂_k, corresponds to the bump

$$\mu_k(t) = 1_{\{t\in[t_{k-1},t_k)\}}$$

to the instantaneous forward curve (see more details in Section 8.48.2). Let f_k, $k = 1, \ldots, K$ be the current forward rate for the period $(t_{k-1}, t_k]$,

$$f_k \triangleq -\frac{1}{t_k - t_{k-1}} \log \frac{P(0, t_k)}{P(0, t_{k-1})} = \frac{1}{t_k - t_{k-1}} \int_{t_{k-1}}^{t_k} f(0, u)\; du.$$

Then the roll-down for V_0 over the period τ is equal to

$$\mathrm{rolldown}\,(V_0) = \sum_{k=2}^{N} (f_{k-1} - f_k)\, \partial_k V_0. \tag{88}$$

Why is this the case? The position V_0 has sensitivity $\partial_k V_0$ to the bump of the forward curve over the period $[t_{k-1}, t_k]$, where the current rate is f_k. Once τ years passes, assuming the spot curve stays the same (in the sense of (87)), the rate over that period will become equal to f_{k-1}, leading to the PnL from that bucket of $(f_{k-1} - f_k)\, \partial_k V_0$. Summing all the buckets give formula (88).

Note that the PnL from the first bucket, $[t_0, t_1]$, is not included in the formula (88). Cashflows paying in this period are accounted separately as carry, as we discussed earlier. It is not uncommon, however, to bundle roll-down and carry into the same number.

See notebook `T8_Explore_Curve_Strategies_01.ipynb`.

8.51.5 Roll-Down and RV Strategies

Roll-down is a measure of profitability of a given position under a very specific scenario of the evolution of future interest rates – when the spot curve does not change. Of course this scenario is unlikely to realize as there will always be market fluctuations of the curve. However, if the unchanging spot curve is indeed the expected value of all possible curves in the future, or close to it (expected value in the P, or real-world, measure!) then the expected PnL is indeed given by roll-down (for linear instruments). Hence it is a popular measure to assess various trading strategies and compare them across different strategies/parts of the curve/currencies/etc.

8.51.6 Roll-Down If Forwards Are Realized

One can look at the roll-down-type measures under other scenarios of future evolution of the yield curve. For example, one can ask how it would look like under the scenario when the yield curve rolls along its forward, the assumption that underpins the constant forwards theta. It turns out that this is actually a not very useful measure because if the forwards are realized, the PnL on the position simply recovers the market price of the instrument.

8.51.7 Words of Wisdom on Roll-Down

- A world in which forwards come true is of relevance to those who hedge, and engage in relative pricing;
- A world in which same-maturity yields do not change is of relevance to those who invest, and who engage in the extraction of risk premia.

Topic 9
Futures Rates, Convexity and Term Premia

Optional topic

The material in this Topic is mostly taken from Morton (2022).

9.52 Predicting Rates in the Future

From a traditional derivatives pricing/derivatives quant perspective, forward rates are the fundamental building blocks of the interest rate curve and are the basic objects to work with. Forward rates, however, are poor predictors of what is going to happen to the rates in the future. They are indeed fundamental building blocks of derivatives pricing but that just assumes that everything is hedged/replicated, and they arise naturally as the "right" rates to use from a derivatives pricing perspective.

An investor in interest rate markets, however, is much more interested in understanding what is the best prediction of where the rates are going to be and, moreover, what does the market prices in as far as future rate expectations are concerned. Having this knowledge allows an investor to compare market-implied expectations – "what is priced in the market" – vs. her own, so she can assess the relative attractiveness of various investment opportunities. As we said before, it is the relative difference between one's own beliefs vs. the market that determines that. One approach, also already discussed, is just to posit that the rates in the future will be just like today (spot), adjusted for any known changes coming from central bank actions and the like. This basically assumes the term premia, and the volatility that drives the convexity term, will remain the same. Let us discuss a more sophisticated, but not that well-known, approach.

We have discussed the central importance of estimating expected (in real measure) rates fixing in the future and a closely linked task of estimating term premia in interest rates earlier in these notes, although in a somewhat abstract world of zero-coupon bonds. Let us see how this can be done in a more realistic context of OTC rates. These tasks could be thought of as calculating P-measure expectations of future Libor rates or, in the new world, of future compounded OIS rates. (As we have seen elsewhere, the distinction makes little to no difference, as far as the math is involved, for rates that start fixing in the future, so we will stick to calling them "Libor rates" from now on.)

To estimate the expected rates fixing far in the future (30y+), and the term premia, we make some assumptions. Before we list them we recall that observed forward rates incorporate three effects: expectation of future rate fixings, term premia (compensation the investors demand for holding higher-risk/longer-dated positions) and convexity.

9.53 Convexity in OTC Derivatives

Convexity is significant at the maturities that we want to look at. Stripping out the convexity effect is task number one for this section as it is significant in maturities of 10 years and above.

The convexity in bonds is easy to understand because the price of the bond is, clearly, a convex function of its yield. So in models, as well as in real life, when yields move up and down, the bond price, as a non-linear function of the yield, responds accordingly and, being convex, generates PnL opportunities to an active (i.e. one who can/want to dynamically hedge) investor to generate money just like in options gamma scalping. This opportunity is of course compensated/paid for by such bonds offering lower yield, as already discussed.

Convexity effects exist, and are significant at longer maturities, in OTC markets but somewhat more subtle, and it is worth to review them. Let us consider a very simple one-period interest rate swap, which is an FRA. A receive-fixed FRA, essentially, pays $c - R(T, T+\tau)$ at time $T+\tau$ where c is the fixed rate on the contract determined at time 0. The PV of this contract at some time t, $0 < t < T$, is then

$$V_{\mathrm{FRA}}(t) = \mathrm{E}_t^{\mathrm{Q}}\left(e^{-\int_t^{T+\tau} r(u)\ du}\left(c - R(T, T+\tau)\right)\right) \tag{89}$$

$$= e^{-\int_t^{T+\tau} f(t,u)\ du}\left(c - \mathrm{E}_t^{T+\tau}\left(R(T, T+\tau)\right)\right) \tag{90}$$

$$= e^{-\int_t^{T+\tau} f(t,u)\ du}\left(c - R(t, T, T+\tau)\right). \tag{91}$$

If we think of how this $V_{\mathrm{FRA}}(t)$ responds to parallel(-ish) shifts of the curve, which are the predominant moves, we have a product of two terms. If the rates go up the FRA fixed rate receiver receives lower payoff *but* also discounted at a higher rate. If the rates go down she receives a higher payoff *but* also discounted at a lower rate. While the payoff is linear in the rate, the PV is not because the discounting is linked/correlated to the payoff. This creates the convexity effect for which, naturally, fixed rate *payers* want to be compensated – by transacting at lower fixed rate.

We want to strip out the convexity effect of trading in OTC flow derivatives (swaps) so we can concentrate on the other two components, the expected future rate and the term premia. Clearly, looking at (89) we want to get rid of the discounting term that creates that nonlinearity. We want something that is linear in the underlying rate and is worth, at time t, something like

$$\mathrm{E}_t^{\mathrm{Q}}\left(c - R(T, T+\tau)\right) = c - \mathrm{E}_t^{\mathrm{Q}}\left(R(T, T+\tau)\right)$$

And of course we recognize $\mathrm{E}^{\mathrm{Q}}\left(R(T, T+\tau)\right)$ as the *futures* rate for the period $[T, T+\tau]$ as opposed to the forward rate $\mathrm{E}_t^{T+\tau}\left(R(T, T+\tau)\right)$ Hence, the main point we make is that

> Stripping out the convexity effect from Libor rates for estimating expected rate/term premia purposes is equivalent to using futures rates, not forward rates, as a starting point.

9.54 Model-Based Futures Convexity

For Libor rates that fix within 5, or perhaps 10, years, one can use ED futures prices to derive the futures rates for further analysis. But we are interested in much more long-dated rates such as those that fix in 30 years and beyond. Here we must use a model. A very sophisticated model is described in Andersen & Piterbarg (2010) or the original paper Piterbarg & Renedo (2006), but for the purposes that we have here, arguably, simpler models are probably better, given the level of uncertainty introduced at other steps. So for the "convexity removing" step we can just use the very basic Ho-Lee (Gaussian without mean reversion short rate) model:

$$dr(t) = O(dt) + \sigma\, dW(t)$$

to write

$$\mathrm{E}^{\mathrm{Q}}\left(R(T,T+\tau)\right) \approx \mathrm{E}^{T+\tau}\left(R(T,T+\tau)\right) + \frac{1}{2}T^2\sigma(T)^2, \tag{92}$$

where $\sigma(T)$ is the implied volatility for options (essentially caplets) on $R(T, T+\tau)$. These can be observed/stripped from the market.

HW: Derive the exact formula in the Ho-Lee model and details the approximations used to come up with (92)

9.55 Term Premia Estimation

With convexity eliminated (which was our first of the two assumptions alluded to earlier), the only remaining drivers of the rates are the expected values of future Libor fixings, and the term premia. We posit that the future expected rate is the same for any T as long as T is far enough in the future. This seems reasonable because it is hard to imagine that market expectations of a 3 month rate in 30 years would be different from a 3 month rate in 35 years, say. Hence we posit that the (annualized!) Sharpe ratio of entering a 30y FRA vs 35y FRA is the same, i.e. for any two times T_1, T_2 sufficiently far in the future, the following holds

$$\frac{\left(\mathrm{E}^{\mathrm{Q}}\left(R(T_1,T_1+\tau)\right) - R^{\infty}\right)/T_1}{\sigma(T_1)} \approx \frac{\left(\mathrm{E}^{\mathrm{Q}}\left(R(T_2,T_2+\tau)\right) - R^{\infty}\right)/T_2}{\sigma(T_2)}.$$

Here R^{∞} is a generic long-maturity τ-tenor expected rate that incorporates all we know about the expected path of interest rates between now and sufficiently far-away T's and $\sigma(T_i)$, as before, is the implied volatility for the option (caplet) on $R(T_i, T_i+\tau)$, $i = 1, 2$. This leads us to the following relationship for the futures (not forward!) rates

$$\mathrm{E}^{\mathrm{Q}}\left(R(T,T+\tau)\right) \approx R^{\infty} + \lambda\sigma(T)T, \tag{93}$$

where λ is the risk premium parameter, a common annualized Sharpe ratio for all far-away rates.

Once $\mathrm{E}^{\mathrm{Q}}\left(R(T,T+\tau)\right)$ are calculated from $\mathrm{E}^{T+\tau}\left(R(T,T+\tau)\right)$ using (92) (or more sophisticated methodology), then it becomes just a matter of linear regression of using (93) to regress $\mathrm{E}^{\mathrm{Q}}\left(R(T,T+\tau)\right)$, as a function of T (for fixed τ) vs. $\sigma(T)T$ This gives both the estimate of the expected fixing in the (far) future R^{∞}, as well as the risk premium parameter λ. As this can be done for each calendar day t (implicitly set to 0 in our formulas above) we can derive the time series of expected long-maturity rates as well as the risk premia. Morton (2022) reports an excellent fit using regression (93) on most days.

Topic 10
Static Curve RV

10.56 Trading Curve Shapes

Let us briefly look at some of the more common curve shape trading strategies. The forward delta ladder for a position is a great tool to determine the impact of a particular curve shape change on the position, as we will demonstrate.

10.57 Outrights

Suppose the trading view is that the overall level of the curve will uniformly increase. As we mentioned before, this is, somewhat counter-intuitively, called (rate) market sell-off (as bonds go down in price if rates/yield go up). A position that benefits in this scenario is a simple payers (pay fixed) swap. For each basis point of increase in the level of the curve, a swap will increase in value by roughly its overall DV01 (see Section 8.48.3). Hence, it is most efficient to use the longest-dated swap that is practical, say 30 years (although up to 50y are reasonably liquid and cleared). A swap rate can be seen as a weighted average of the forward rates over the length of the swap. A look at forward rate DV01s confirms this intuition. If the par rate on the swap is k and the curve moved so that the new swap rate is k', the PnL is approximately equal to

$$\mathrm{PnL} = (k' - k) \sum_{n=1}^{N} \tau_n P(0, T_{n+1}).$$

10.58 Steepeners/Flatteners

10.58.1 Definitions

A steepening of the yield curve is what it sounds like – the slope of the yield curve (whether in par rates, zero rates or forward rates terms) increases. Two kinds of steepening are distinguished. A *bear steepening* corresponds to the long rates going up more than the short rates. It is called a bear steeping because the part of the curve that moves more, the long-term rates, go up implying the corresponding bonds decrease in value (exhibiting bearish behaviour).

A *bull steepening* is when short rates decrease more than the long rates. The name, again comes from what happens to bond prices that correspond to the rates that move the most, in this case the short-term rates. In a bull steepening short-term bonds increase in value (i.e. bullish).

The opposite of the steepening curve is flattening. Again there are two kinds, bull and bear flatteners.

Positions that benefit from steepening are typically flavours of the basic strategy of pairing a long-maturity payer swap with a short-maturity receiver swap. Let us look at the forward delta ladder of an example of such a position in notebook T8_Explore_Curve_Strategies_01.ipynb.

A number of decisions need to be made to fully specify the steepening position. Which maturity swaps are to be used? And in what proportion of notional?

10.58.2 Choosing Notionals

Parking the question of maturities for the moment, the relative notionals are normally determined by the condition that the position is immune to outright changes of the curve level. This is done specifically to isolate the steepening move that the position is intended to capture while not picking up any residual outright exposure. A position would be immune to outright level changes if its DV01 was 0. As the payer swap has positive DV01 and the receiver has negative DV01, this is easy to accomplish – the notionals should be taken in the reverse proportion to the (absolute) DV01s. An example is shown in the notebook mentioned above. We can also use swap PV01s (the values of their annuities) for the same purpose. Let the longer swap have N periods and the shorter M. Then the notionals H_N and H_M should be set so that

$$\frac{H_N}{H_M} = \frac{\sum_{n=1}^{M} \tau_n P(0, T_{n+1})}{\sum_{n=1}^{N} \tau_n P(0, T_{n+1})}.$$

In particular, the shorter swap would have a higher notional.

10.58.3 Choosing Maturities

The choice of swap maturities is largely driven by the expectations of which parts of the curve will move the most. However, if there is no strong conviction here, we can use the roll-down measure to choose the most beneficial position from a set of roughly equivalent ones. This can be seen as prudent risk management – if our view on the curve steepening does not materialize and the curve stays roughly the same, we want to pay the least/make the most of our position on roll-down. (Of course if the curve flattens instead, we will very likely lose money).

HW: Using the notebook mentioned, find the "most beneficial" steepener.

Steepeners could be expressed in terms of forward starting swaps as well. This, arguably, is a more direct expression of the view. For example, suppose we think 15y rate will increase relative to a 5y rate whereas a 10y rate would stay roughly the same. Then we can enter a 5y spot-starting payer and 5y10y receiver, in DV01-proportional notionals as before (here it will be 2 to 1; reader should think why this is the case). The notebook T8_Explore_Curve_Strategies_01.ipynb provides a suitable demonstration.

10.59 Flies

Butterflies, or "flies", are positions whose values are primarily driven by the curvature of the yield curve. For example, a 5y/10y/20y fly would involve a 5y and 20y payer swaps and a 10y receiver swap. Notionals would be chosen to make the overall DV01 equal to 0 (so the position is not sensitive to outright changes in the levels of rates), and also to immunize against steepening/flattening moves.

HW: add this strategy to the notebook T8_Explore_Curve_Strategies_01.ipynb and explore. Convince yourself that it is immunized against steepening/flattening.

The position above benefits when the curve curvature decreases, i.e. when the middle, 10y, rate moves lower than the average of the 5y and 20y rates.

The same view can be expressed using forward swaps, see next section.

Outright level changes in the yield curve and steepening/flattening moves are often directly associated with certain macro-economic developments such as increasing inflation, slowing economy growth, etc. As such, outrights and steepeners are popular for expressing macro views. Increases/decreases in the curvature of the yield curve are harder to link to specific macro-economic scenarios except through convexity, see Section 7.47. Specifically, increased volatility of the long-end part of the curve would normally lead to more curvature in the curve (HW: Why?) so entering a long-curvature trade could be an expression of a macro view of increased long-term curve volatility. RV strategies, however, are a more popular application of flies than for expressing macro views.

10.59.1 Micro RV

10.59.1.1 Motivation

The short end of a yield curve typically enjoys a large number of directly observable market inputs such as ED futures, MPC FRAs, and short-dated swaps with granular maturity grids. Further down the curve, however, observable market inputs become sparser. For example, only 10y, 15y and 20y could be directly observable (i.e. having reliable feeds from brokers). Swaps with maturities between 10 and 15 years are then quoted by dealers on the basis of their house curves that use some interpolation between 10 and 15 years to fill the gaps.

Also relevant is the fact that some benchmarks are updated, by brokers, much more frequently than others. For example, a broker may quote 9y,10y,12y points but the 10y, being a major benchmark, would "tick" much more frequently than 9y and 12y. In a rapid and/or large move it is not uncommon to observe distortions in the curve arising from the fact that the 10y point has been updated but the surrounding points have not been.

Careless interpolation, other sloppy modelling choices, or not being careful about the timing of updates of different benchmarks may result in a dealer having a curve exhibiting some non-economic artefacts. One of the RV strategies employed by HFs, sometimes called "micro RV", involves searching for these artefacts and then trading relevant products with the view that the anomaly would right itself over time, or as the time passes and the position rolls into a more observable part of the curve, the artefact disappears.

10.59.1.2 Execution

Of course HFs cannot observe dealers' curves directly, so they typically resort to asking for specific structures in areas of the curve that are known to be potentially problematic. Here is where flies are often employed. For example, a HF can ask a number of dealers for their market on 9y/10y/11y fly, or 14y/15y/16y fly, and the like. Among the quotes the HF might be able to find one (or more) that is likely to be a result of poor interpolation rather than a "genuine" market price, in which case a HF would put the appropriate position on.

Forward rate flies can also be used to really zoom in on a particular area of the curve. It is not hard to see, and also see the notebook, that a 9y10y11y fly is in fact almost equivalent, in risk terms, to being long (i.e. receive fixed, negative DV01) for the 9y1y forward swap, and being short (pay fixed, positive DV01) for the 10y1y forward swap. Using forward swaps could be more efficient for really isolating a specific area of the curve. However, the more esoteric/leveraged/specific to a particular point on the curve the combination of products that the HF asks for, the more suspicious a dealer would become that HF interest is not driven by genuine market views but by trying to exploit weaknesses of the dealer's curve building technology. Then the dealer would normally quote wide bid/ask, reducing potential trading opportunity. It may also cause the dealer to carefully examine its curve in the area of HF interest to see if it exhibits any undesirable characteristics (jumps, kinks, excessive curvature, etc.)

10.60 PCA

The discussion of simple RV strategies so far has assumed that choosing a position such that its total DV01 is zero immunizes it against parallel moves in the yield curve. This is true insofar as we believe that a move where all the rates move by exactly the same increment is a good definition of a "parallel" move. Empirically, however, this is not how the curve typically moves. Short-term rates usually move by a larger increment than longer-term rates. The question then arises, what are the "most typical" moves of the yield curve? This is a deep subject that, as a starting point, can be tackled by a statistical analysis technique called the Principal Component Analysis (PCA).

10.60.1 Empirical Principal Components Analysis

For some fixed value of τ (e.g. 0.25 or 0.5), let us define "sliding" forward rates. $Y(t, x)$ with tenor x as

$$Y(t, x) = y(t, t + x, t + x + \tau),$$

where $y(t,s,u)$ is a forward rate observed at a (historical) time t for the period $[s,u]$, $t \le s < u$. For a given set of tenors $x_1, \ldots, x_{N_x}$ and a fixed calendar time t, the collection of forward rates $\{Y(t,x_n), n = 1, \ldots, N_x\}$ can be seen as a representation of the yield curve at time t. Given the set of calendar times $t_0, t_1, \ldots, t_{N_t}$, we have a vector $\bar{Y}(t_i) = (Y(t_i, x_n), n = 1, \ldots, N_x)$ for each time $i = 0, \ldots, N_t$.

What are the typical "shapes" of the moves of the yield curve or, more precisely, its vector representation $\bar{Y}(t_i)$? To start answering this question, let us represent the change in the yield curve

$$\Delta\bar{Y}(t_i) \triangleq \bar{Y}(t_{i+1}) - \bar{Y}(t_i)$$

in parametrized form

$$\Delta Y(t, x_n) \approx \sum_{j=1}^{J} l_j(x_n) \Delta f_j(t), \quad n = 1, \ldots, N_x, \quad t \in \{t_1, \ldots, t_{N_t}\}. \tag{94}$$

Here functions $l_j(x)$ are functions of the tenor only, and are called *loadings*, as they define the shapes of the moves. They are scaled by the changes in *factors* $f_j(t)$ (not to be confused with forward rates) that evolve over time. So, in plain English, the change of the yield curve at time t is given by a linear combination of loadings l_j, where weights in the linear combination are given by the "factors" f_j. The number of factors/loadings J is typically chosen to be (much) smaller than the number of rates N_x, e.g. $J \approx 2 \sim 5$ and $N_x \approx 10 \sim 50$.

A question remains, how do we choose the loadings $l_j(\cdot)$ in (94) in such a way that the moves in the yield curve, over the observation period $t \in \{t_1, \ldots, t_{N_t}\}$, are best explained by them? This is where PCA comes in. The next section gives the necessary theoretical background.

10.60.2 Technical Details on Principal Components Analysis (PCA)

Consider a p-dimensional Gaussian variable Z with a given covariance matrix Σ. Assume, with no loss of generality, that the mean of Z is 0 and that Σ has full rank (positive definite). Consider now writing, as an approximation,

$$Z \approx DX, \tag{95}$$

where X is an r-dimensional vector of independent standard Gaussian variables, $r \le p$, and D is a $(p \times r)$-dimensional matrix. How should we choose D in an optimal way?

First, we obviously need to define what constitutes an "optimal" approximation in (95). We here have in mind L^2 closeness of the covariance matrix $DD^\top$ to Σ, so let us define the optimal D^* as the matrix that minimizes the norm $f(D)$, where

$$f(D)^2 \triangleq \operatorname{tr}\left(\left(\Sigma - DD^\top\right)\left(\Sigma - DD^\top\right)^\top\right).$$

This is just the matrix representation of the usual Frobenius norm, the element-wise squared differences between Σ and $DD^\top$,

$$f(D)^2 = \sum_{i,j} \left(\Sigma_{i,j} - \left(DD^\top\right)_{i,j}\right)^2.$$

The value of D that minimizes $f(D)$ can be shown to be

$$D^* = E_r\sqrt{\Lambda_r}, \tag{96}$$

where Λ_r is an $r \times r$ diagonal matrix containing the *largest* r eigenvalues of Σ, and E_r is a $p \times r$ matrix of r p-dimensional eigenvectors corresponding to the eigenvalues in Λ_r.

Equipped with the optimal D, we now go back to the approximation (95) and write

$$Z \approx \widetilde{Z} \triangleq E_r\sqrt{\Lambda_r}X = \sqrt{\lambda_1}e_1X_1 + \sqrt{\lambda_2}e_2X_2 + \ldots + \sqrt{\lambda_r}e_rX_r, \tag{97}$$

where e_i denotes the i-th column of E_r and the λ_i's are the eigenvalues, sorted in decreasing order of magnitude. The (deterministic) vector e_i is known as the i-th *principal component* of Z, and the (random) variable $\sqrt{\lambda_i}X_i$ as the i-th *principal factor*. With (97), we have $\mathrm{tr}(\mathrm{Cov}(Z,Z)) = \mathrm{E}(Z^\top Z) = \sum_{i=1}^p \lambda_i$ and $\mathrm{tr}(\mathrm{Cov}(\widetilde{Z},\widetilde{Z})) = \mathrm{E}(\widetilde{Z}^\top \widetilde{Z}) = \sum_{i=1}^r \lambda_i$, i.e. the first r terms in the decomposition (97) explain a fraction

$$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i}$$

of the sum of the diagonal elements of the covariance matrix of Z. Principal components decomposition will thus result in a loss of total variance, unless the covariance matrix is either rank-deficient (i.e. has eigenvalues that are strictly zero), or we use a full set of principal components ($p = r$). In many cases of interest to us here, the loss of variance can be small, even if r is a modest number, e.g. 2 or 3. We notice that the covariance matrix for Z, as approximated by (97), will be *rank-deficient*, as the number r of non-zero eigenvalues is less than p.

While we have used a setting with Gaussian variables to motivate our treatment of principal components analysis (PCA), it is, in fact, a generically useful tool for uncovering the structure of large-dimensional random vectors, and replacing them with more manageable, lower-dimensional variables; see, e.g., Theil (1971) for more details and an application to empirical non-Gaussian data. Also, PCA identifies which directions of a multi-dimensional random variable are "important", potentially allowing us to allocate computational resources in an intelligent manner.

10.60.3 Example: USD Forward Rates

To give a concrete example of a PCA run, we set $N_x = 9$ and use tenors of $\{x_1, \ldots, x_9\} = \{0.5, 1, 2, 3, 5, 7, 10, 15, 20\}$ years. We fix $\tau = 0.5$ (i.e., all forward rates are 6 months discrete rates) and use 4 years of weekly data from the USD market, spanning January 2003 to January 2007, for a total of $N_t = 203$ curve observations. The eigenvalues of the matrix C are listed in Table 2, along with the percentage of variance that is explained by using only the first m principal components.

m	1	2	3	4	5	6	7	8	9
Eigenvalue	7.0	0.94	0.29	0.064	0.053	0.029	0.016	0.0091	0.0070
% Variance	83.3	94.5	97.9	98.7	99.3	99.6	99.8	99.9	100

Table 2: PCA for USD Rates. All eigenvalues are scaled up by 10^4.

As we see from the table, the first principal component explains about 83% of the observed variance, and the first three principal components together explain nearly 98%. This pattern carries over to most major currencies.

HW: Using the Fed dataset https://www.federalreserve.gov/data/nominal-yield-curve.htm perform PCA on zero-coupon bond yields. Plot the first three loadings and calculate % Variance explained as a function of the number of principal components used.

Having done the homework the reader will see that the first principal component can be interpreted as a near-parallel shift of the forward curve, whereas the second and third principal components correspond to forward curve twists and bends, respectively. This is a very common outcome of PCA on curve data across currencies, curves, and so on.

10.61 PCA-Based Static RV

10.61.1 A Word of Caution

The PCA gives us "Principal Components" (PCs), moves in the yield curve whose combination explain the maximum amount of all changes, in a strictly defined theoretical sense of Section 10.60.2. Specifically, PC1 explains the most variance, PC2 the second most, and so on., as given by the eigenvalues of the covariance matrix. These typically, *but not always*, represent a level, slope and curvature changes of the curve, in that order. However, it is entirely possible during a particular period in history that, for example, the second dominant move (PC2) is in fact a curvature not a slope, change. It is very important to be cognizant of this fact and be careful when mapping PCs to financial concepts such as level or slope or curvature change, especially when creating RV strategies based on PCA, more on which in the moment. A quick visual inspection is really all that's required to ascertain financial meaning of each PC.

Notwithstanding our note of caution, for simplicity of exposition we will identify PCs with their common interpretations as level, slope, curvature.

10.61.2 Strategy Construction

Hence, it makes sense to use the in constructing RV strategies. As we mentioned before, a strictly parallel move of the yield curve is not the most " common" move – as it is in fact given by the first PC. Hence, if we want to create a portfolio that is only sensitive to the changes in the curvature of the yield curve, then we should immunize it not against the (naive) parallel shift and slope changes of the yield curve, but against the first and the second PCs. These two capture the same move (level and slope changes) but account for the fact that different rates move differently in the parallel/steepening moves.

In fact, if we want to put on a position that captures specifically the curvature, or the third component of the PCA (so sensitive to changes in f_3 only and not f_1 or f_2), then all we need to do is to create a position whose risk profile (as measured in bucketed DV01 terms) replicates the third loading $l_3(\cdot)$. This is so because the loadings of PCA, being eigenvectors of the empirical covariance matrix, are orthogonal by construction.

Likewise, if one wants to bet on higher-order PCs, e.g. isolate a "zig-zag" move of the 4th PC, all they need to do is to create a position whose risk profile mimics that higher-order PCA. RV strategies that go beyond the 3d PC are quite rare, however.

10.61.3 Limitations

While the PCA is a powerful method, one needs to keep in mind that it is not "universal" in the sense that the conclusions of the analysis, including the most likely move shapes etc. are dependent on various choices made by the modeller. Specifically, the following non-exhaustive list enumerates considerations that affect PCA conclusions:

- The length of history used;
- The tenors used in analysis;
  - Analysis of a 10y curve moves will give different answers than the analysis of a 40y curve;
  - Even adding extra tenors in between existing tenors affects the results, e.g. doing PCA on 1y,3y,5y,10y tenors is going to be different from 1y, 3y, 5y, **7y**,10y;
- One can perform PCA on forward or spot rates;
- One can perform PCA on changes or outright levels;
- Relative or absolute changes could be used;
- etc.

Topic 11

Historical Analysis

11.62 Historical RV Analysis

Many, if not most, RV strategies are assessed by analysing historical data. A particular fly can exhibit attractive roll-down metrics but more often than not, a trader would want to see how the same position performed in the past. Are the current levels attractive from the historical perspective? There are many aspects to this question that we shall look into in some detail.

Roll-down gives us an "expected", in some sense, PnL of the position – expected in a mathematical sense but under a (strong) assumption that the yield curve stays the same (or its fluctuations are centered around the current curve). The range of possible values of a position in a years' time, even if centered on its expected value, will certainly exhibit random behaviour. The range of possible values and the likelihood of various outcomes is certainly of significant interest to the trader. As the future is unknowable, the answer is often sought in history.

Let $X(t)$, $t \in [-T, 0]$ be the historical realized time series of values of a given strategy. This could be very generic, from a buy-and-hold specific combination of swaps such as a fly to, say, a hedging strategy of buying a longer-dated swap and hedging it with a shorter-dated swap. The current, today's, value is $X(0)$. How do we decide if $X(0)$ represents a "good" opportunity?

Stated in this generality, this question is, of course, at the heart of the whole field of (quantitative) asset management. We cannot hope to answer it completely, but we can offer a few, hopefully useful, observations.

11.62.1 Useful Statistics, Z and P

It is intuitively appealing to buy something when it is at, or near, historical lows and to sell it when it is at, or near, historical highs. Given the observations $X(t)$, $t \in [-T, 0]$, how do we quantify to what extent $X(0)$ is at the edges of its historical range? This is most easily determined by quantifying how far $X(0)$ is from its historical mean, with the deviation expressed in some standard units. The z-score is just such a measure. We define

$$\zeta(X(0)) = \frac{X(0) - \mu_X}{\sigma_X},$$

where we set

$$\mu_X = \frac{1}{T} \sum_{t=-T}^{0} X(t), \quad \sigma_X^2 = \frac{1}{T} \sum_{t=-T}^{0} \left(X(t) - \mu_X\right)^2 .$$

Then $\zeta(X(0))$ represents the deviation of $X(0)$ from its historical mean over the time horizon $[-T, 0]$, measured in the units of the standard deviation σ_X

over the same period. If the realizations $X(\cdot)$ were independent Gaussian, we would consider $|\zeta(X(0))| \geq 2$ to be noteworthy as it would imply the historical probability of being either above 2 or below -2 to be about 2%, and $|\zeta(X(0))| \geq 3$ to be exceptional as the (two-sided) probability of this happening is about 0.3%, so once every year or so.

The so-called p-values formalize this translation of z-scores into probabilities. The p-value p_ζ that corresponds to the z-score ζ is defined simply as the probability of a standard Gaussian random variable exceeding ζ:

$$p_\zeta = \mathrm{P}(|Z| > \zeta) = 2\Phi(-\zeta),$$

where Z is a standard Gaussian random variable and Φ its cumulative distribution function. Sometimes it makes sense to only consider one-sided p-values, i.e. either $\mathrm{P}(Z > \zeta)$ or $\mathrm{P}(Z < \zeta)$.

The p-values show up in many areas of statistics, not just when converting z-scores into probabilities. Formally, a p-value is the level of marginal significance within a statistical hypothesis test, representing the probability of the occurrence of a given event or pattern. Informally, it is the probability of being wrong thinking a particular phenomena is statistically significant. Typical values for statistical significance are 1% or 5%, depending on how "sure" one wants to be – if the calculated p-value is below one of these thresholds, the phenomena is deemed significant.

While it is convenient to represent one's confidence with a single number, there is always a possibility of misuse, inadvertent or intentional. See Section 11.66 later in the notes for a brief discussion.

Financial time series, as a rule, do not fit a convenient mathematical framework of independent Gaussian random variables. They typically exhibit fatter tails (higher probability of extreme events) than the Gaussian distribution would suggest, and also, in many cases, are more correlated across time than the independence assumption would suggest. Hence, in financial time series, having a z-score of 3 is more common that the Gaussian assumption implies. Yet, it is still a signal that a particular trade is interesting/a particular relationship is somewhat out of whack. At the very least, expressing current values of various securities/strategies in terms of their z-score gives the trader/PM an opportunity to quickly scan the universe of possible trades and to zoom-in into the areas that could be potentially interesting. The z-score (or p-value), then, could serve as a fairly coarse filter that could be applied across a multitude of strategies to highlight those that, perhaps, warrant a deeper investigation by the PM.

11.63 Mean Reversion, Basics

The concept that positions with a high absolute z-score are "attractive", because they (statistically/historically) are likely to go back to more "normal" values, is a manifestation of a common concept called *mean reversion*. It is widely believed, and supported by empirical evidence, that many observables meanrevert, i.e. eventually go back to their long-term averages if they deviate too far from them.

From a quantitative perspective, how do we quantify to what extent a particular time series exhibits mean reversion? If one Googles "mean reversion test", they are likely to find multiple references to the *Dickey-Fuller test* or the *augmented DF* test (ADF). Let us briefly describe it. Suppose X_n, $n = 1, \ldots, N$, is a time series of observed values of a particular financial variable. A mean-reverting behaviour of the time series can be described in the following way,

$$X_{n+1} - X_n = \varkappa\,(\theta - X_n) + \varepsilon_n, \tag{98}$$

where ε_n are i.i.d. random increments. This is a a flavour of an ARIMA model, see Theil (1971). If $\varkappa$ is positive then, as X deviates away from the long-term mean θ, there is a "pull" exerted on X to move closer to the mean. A continuous-time equivalent is

$$dX(t) = \varkappa(\theta - X(t))\ dt + \sigma\ dW(t). \tag{99}$$

The ADF test postulates the model (98) and applies statistical methods to determine if $\varkappa$ is non-zero (positive) to a high degree of statistical significance. While theoretically appealing, the test is not particularly useful in practical situations. This is so because the (intuitive) concept of mean reversion is rarely captured by a simplistic model such as (98). Specifically, the conclusions of the ADF test are strongly dependent on the scale of the dynamics it is designed to test – e.g. a time series could exhibit mean reversion on the scale of weeks but when applied with daily increments the DF test may be inconclusive for daily increments as the mean reversion is too weak on the scale of days. And trying to apply it to, say, weekly increments, quickly runs into limitations of the available length of time series and their general non-stationarity.

11.64 Mean Reversion vs. Trend

11.64.1 Motivation

Let us consider three regimes that a time series can exhibit, each one with its own consequences for trading strategies employed. We distinguish the following three regimes/behaviours for a given financial variable:

- Mean reversion – the variable tends to go back to its mean;
- Momentum/trend – the variable tends to continue to go in the direction of its latest move;
- Random walk – the variable is not statistically different from a random walk.

To develop a practically-useful test for distinguishing the three, and by extension to detect the mean reversion regime, let us go back to (99). We want to calculate the variance of the increments of this process *over different time scales*. We shall show that how the variance scales when calculated over different time scales – e.g. daily, weekly and monthly – will help us with our task.

11.64.2 Variance for Different Time Scales

Let us denote

$$\Delta_\tau X(t) \triangleq X(t+\tau) - X(t),$$

where τ represents an offset, i.e. we consider increments of the process $X(\cdot)$ over different time horizons (such as 1 day, 1 week, and so on). In the model (99) we have the following result,

$$\mathrm{Var}_t\left(\Delta_\tau X(t)\right) = \sigma^2 \frac{1-e^{-2\varkappa\tau}}{2\varkappa}.$$

Note that the right-hand side does not depend on t.

HW: Prove the formula.

Let us approximate the function on the right-hand side, as a function of τ, with a power function, for reasons to be explained shortly:

$$\frac{1-e^{-2\varkappa\tau}}{2\varkappa} \approx C\tau^{2H}.$$

We do that by expanding both sides around $\tau_0 = 1$. Of course we can use other τ_0 but $\tau_0 = 1$ is convenient to start with. The values $\tau_0 = 1/52$ which is 1 week and 1/12 which is 1 month or 1/4 which is 3 months are all good points to expand around, see later. For now, starting with $\tau_0 = 1$, we have

$$C = \frac{1-e^{-2\varkappa}}{2\varkappa}.$$

To figure out what H is, we match the first derivatives of both sides at $\tau = 1$,

$$e^{-2\varkappa} = 2CH,$$

so

$$H = \frac{e^{-2\varkappa}}{2C} = \varkappa \frac{e^{-2\varkappa}}{1-e^{-2\varkappa}}. \tag{100}$$

For an arbitrary point τ_0 around which we expand, this generalizes to

$$H = \varkappa\tau_0 \frac{e^{-2\varkappa\tau_0}}{1-e^{-2\varkappa\tau_0}}.$$

HW: Derive the expansion for arbitrary τ_0.

11.64.3 Hurst Exponent

The constant H is called the Hurst exponent and is an important quantity in time series analysis. For an arbitrary time series that is not necessarily generated by the model (99) it still makes sense to look at the scaling of the realized variance with respect to the lag. Then H is the constant that gives the best fit across a selection of τ's in

$$\mathrm{Var}_{\mathrm{emp}}\left(\Delta_\tau X(t)\right) \approx C\tau^{2H};$$

Here $\mathrm{Var}_{\mathrm{emp}}$ is the empirical variance of the time series $X(t)$, $t = -T, \ldots, 0$, that is being tested. In practice, the test proceeds as follows. Given a time series of interest $X(t)$, $t = -T, \ldots, 0$,

1. For each chosen lag τ such as 1, 2, 5, 10, 20 days, calculate τ-increments of $X(t)$. This can be done either without overlap (more theoretically correct) or for each day t, so with overlap (more data).

2. For each series of τ-increments (i.e. for each lag τ), calculate its sample variance;

3. Best-fit the log of the sample variance against $\log\tau$;

4. Deduce H from the slope of that regression.

Then

- $H \approx 0.5$ indicates random walk;

- $H < 0.5$ indicates mean reverting behaviour;

- $H > 0.5$ indicates trending/momentum behaviour.

I am not aware of any statistical tests that rigorously define the p-values and such (i.e. by how much does H need to be different from 0.5 to reliably state there is mean reversion), but lots of insight could be obtained from running simulated tests, i.e. simulating various types of time series with different mean reversions many times and understanding what constitutes a reasonable cut off point

A key property of this test is that it can be performed for different lags, i.e. we can look at the scaling of realized variances for lags on the order of days – to detect if there is "fast" mean reversion – or on the order of weeks or even months, to see if there is any "slow" mean reversion. It also has some cool connections to fractal dimensions and some other interesting mathematics.

The estimate of the Hurst exponent can of course be converted into an estimate of the mean reversion using the formula (100) or its extension for an arbitrary expansion point τ_0. Mean reversion is a more intuitive parameter to traders than the Hurst exponent as it basically tells them the half-life of an expected deviation of the underlying time series from its mean, so it gives an indication of how long they have to wait for their mean reversion trade strategy to pay off.

Much of the material in this section is demonstrated in the notebook T11_Mean_Reversion_Trend_01.ipynb.

11.65 Seasonality

Some time series exhibit obvious seasonality. A typical example here is the CPI/RPI which are the inflation indices. Every month a number that measures the price of a pre-defined basket of goods is published by a government agency. As a simple explanation, heating bills are higher in the winter for northern countries, and cooling bills are higher in the summer for southern countries. This type of seasonality is well-understood and is priced in the market, as one can observe in the inflation seasonality marks from the brokers.

Another example is government bond prices before and after an auction. Since auctions happen on a pre-defined schedule, one can have a hypothesis that there is a seasonal pattern to the bond prices anchored on auction dates.

Other variables may exhibit seasonality but in a less obvious way. If one discovers a variable that exhibits seasonality that the other market participants are unaware of, clearly, there is scope for developing a profitable trading strategy. You would buy the "variable" in the months when it is cheap and sell it in the months when it is dear.

Given a time series, how does one determine if it exhibits seasonality in a statistically significant way? Let us say we are looking for monthly seasonality, and there is a monthly time series being tested, $\{X_n\}_{n=1}^N$. Let us denote the seasonality period by p ($p = 12$ for a monthly time series and monthly seasonality). Then we can compute the averages for each period as such,

$$\bar{X}_r = \frac{1}{N_r} \sum_{m:r+pm<N} X_{r+pm}, \ r = 1, \dots, p,$$

where we have denoted

$$N_r = \# \left\{ X_{r+pm}, r + pm < N \right\}.$$

We can also calculate the overall average as

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

Let r_{hi} be the index of the period that has the highest $\bar{X}_r$,

$$r_{\text{hi}} = \arg \max_{r=1,\dots,p} \bar{X}_r.$$

Let r_{lo} be the same but the lowest. Then we want to test the hypothesis that $\bar{X}_{r_{\text{hi}}}$ is statistically significantly different from $\bar{X}$. Alternatively, we may want to test the hypothesis that $\bar{X}_{r_{\mathrm{hi}}}$ is statistically significantly different from $\bar{X}_{r_{\mathrm{lo}}}$. Calculating the z-score for the difference of the two averages being tested is probably the simplest test of statistical significance.

The z-score in the latter case is simply

$$Z_{\mathrm{lo,hi}} = \frac{\bar{X}_{r_{\mathrm{hi}}} - \bar{X}_{r_{\mathrm{lo}}}}{\mathrm{Var}\left(\bar{X}_{r_{\mathrm{hi}}} - \bar{X}_{r_{\mathrm{lo}}}\right)^{1/2}}.$$

The calculation of the variance depends on the properties of the time series. In a simple, and typical, case of independent observations we have

$$\mathrm{Var}\left(\bar{X}_{r_{\mathrm{hi}}} - \bar{X}_{r_{\mathrm{lo}}}\right) = \mathrm{Var}\left(\bar{X}_{r_{\mathrm{hi}}}\right) + \mathrm{Var}\left(\bar{X}_{r_{\mathrm{lo}}}\right)$$

and, for any r,

$$\begin{aligned}\mathrm{Var}\left(\bar{X}_r\right) &= \mathrm{Var}\left(\frac{1}{N_r}\sum_{m=1}^{N_r} X_{r+pm}\right) = \frac{1}{N_r^2}\mathrm{Var}\left(\sum_{m=1}^{N_r} X_{r+pm}\right)\\ &= \frac{1}{N_r^2} N_r \mathrm{Var}\left(\{X_{r+pm}, m = 1, \ldots, N_r\}\right) = \frac{1}{N_r}\sigma_r^2,\end{aligned}$$

where σ_r^2 is the sample variance of the sequence $\{X_{r+pm}, m = 1, \ldots, N_r\}$. Hence

$$Z_{\mathrm{lo,hi}} = \frac{\bar{X}_{r_{\mathrm{hi}}} - \bar{X}_{r_{\mathrm{lo}}}}{\sqrt{\sigma_{r_{\mathrm{hi}}}^2 / N_{r_{\mathrm{hi}}} + \sigma_{r_{\mathrm{lo}}}^2 / N_{r_{\mathrm{lo}}}}}$$

and since $N_r \approx N/p$ we get

$$Z_{\mathrm{lo,hi}} = \sqrt{N/p}\,\frac{\bar{X}_{r_{\mathrm{hi}}} - \bar{X}_{r_{\mathrm{lo}}}}{\sqrt{\sigma_{r_{\mathrm{hi}}}^2 + \sigma_{r_{\mathrm{lo}}}^2}}.$$

If $|Z|$ is sufficiently large, 3 or more say (implying p-value of about 0.3% under Gaussian assumptions), then there is strong evidence of seasonality.

We can likewise test against the overall sample mean $\bar{X}$, and we also do not need to restrict ourselves to the highest and lowest – we can test $\bar{X}_r$ for all r and see which months (say) exhibit strong deviations from the overall mean.

11.66 Backtest Overfitting

If one tests enough time series for a particular trait, such as seasonality, there is high chance that one or more will exhibit this trait with what appears to be statistical significance. This is a classical case of the so-called *backtest overfitting*, also known as data fishing, and sometimes p-hacking. This comic summarizes it well: https://xkcd.com/882/. In rigorous statistical studies one needs to adjust the p-value for the number of experiments performed/time series tested. This is rarely done in finance especially when performed by semi-literate budding statisticians such as sales or structuring, who often troll through mountains of data looking for interesting patters that they can “sell”, as a story, to clients to generate trades exploiting these patterns. This also happens when building trading strategies without having a good economic rationale behind it. The book de Prado (2018) is a good read here.

Topic 12
Allocation of Capital

Optional topic

12.67 Bet Sizing

An investment manager, broadly speaking, has two decisions to make – what to invest in, and what proportion of capital to invest into each opportunity. We have already discussed various ways she may answer the first question, in the realm of interest rate instruments. But if there are, say, 10 different attractive investment opportunities, how does she decide the allocation of capital to each one?

12.68 Coin Flipping

We solve a simple special case later in the notes (see Section 14.77) where we consider a collection of binary bets that are disjoint (only one can pay) and complete (one will always pay). Overall, however, this is a complex question that cannot be fully answered in a general manner. The book Haghani & White (2023) which is yet to be published at the time of writing (February 2023) but can be pre-ordered on Amazon, is a great resource that analyses this question thoroughly. Alternatively some blog posts here `https://elmwealth.com/elm-research/` cover some of the material.

The authors start with a simple experiment that they actually performed on people of varying financial sophistication. Specifically, subjects were given initial capital \$25, and 300 consecutive bets on a coin that is known to be biased 60% heads vs 40% tails. One could bet as much as they liked on each roll from the capital they still had at that point in time. After 300 rolls they would get to keep what is in the account, up to a maximum payout \$250. The question is, what is the optimal strategy?

You can try this yourself at `https://elmwealth.com/coin-flip/`. The optimal strategy will be revealed in the class.

12.69 Merton's Optimal Allocation

12.69.1 The Model

Among the earlier examples of bet sizing is the classical optimal investment allocation problem that was solved by Merton Merton (1991). The Merton investment problem considers dynamic optimal allocation between risky versus riskless securities to optimize a given utility function at a given time horizon. It can be considered a "bet sizing" problem with two types of bets – risk-less and risky. How much of our wealth do we put on each? Let us quickly review the main results here.

We consider a general process for a risky asset in the P, i.e. real-world, measure,

$$dX(t)/X(t) = \mu(t)\ dt + \xi(t)\ dB(t), \tag{101}$$

where both $\mu(t)$ and $\xi(t)$ are deterministic processes (extensions to stochastic processes have been made since the original result). Let $r(t)$ be the risk-free rate, and

$$dR(t)/R(t) = r(t)\ dt$$

be the risk-free asset (bond). Consider a portfolio with $\pi_0(t)$ bonds and $\pi(t)$ stocks in the portfolio at time t. Denote the wealth process by $P(t)$. Under the self-financing condition, $P(t)$ satisfies

$$\begin{aligned} P(t) &= \pi_0(t)R(t) + \pi(t)X(t), \\ dP(t) &= \pi_0\ dR(t) + \pi(t)\ dX(t), \end{aligned}$$

so that

$$\begin{aligned} dP(t) &= \pi_0 r(t)R(t)\ dt + \pi(t)X(t)\left(\mu(t)\ dt + \xi(t)\ dB(t)\right) \\ &= r(t)\pi_0 R(t) + \pi(t)X(t)dt + \pi(t)X(t)\left(\left(\mu(t) - r(t)\right)dt + \xi(t)\ dB(t)\right) \\ &= r(t)P(t)dt + \pi(t)X(t)\left(\mu(t) - r(t)\right)dt + \pi(t)X(t)\xi(t)\ dB(t). \end{aligned}$$

If we define

$$\kappa(t) \triangleq \pi(t)X(t)/P(t)$$

to be the *proportion* of wealth invested into the risky asset, then

$$dP(t)/P(t) = \left(r(t) + \kappa(t)\left(\mu(t) - r(t)\right)\right)dt + \kappa(t)\xi(t)\ dB(t). \tag{102}$$

12.69.2 CRRA Utility

Let $U(w)$ be the utility function. Let us consider the investor's objective to maximize

$$\mathrm{E}\left(U(P(T))\right),$$

the expected utility value under the real-world P-measure, over all admissible choices of $\kappa(\cdot)$. Note that we focus on the problem of maximizing the terminal wealth and not the more general problem of maximizing the utility of consumption for clarity of exposition.

We specialize the utility function to be the family of Constant Relative Risk Aversion (CRRA) utility functions parametrized by the risk aversion parameter $\gamma \geq 1$:

$$U(w) = U(w;\gamma) = \frac{w^{1-\gamma}}{1-\gamma}. \tag{103}$$

The family of the CRRA utility functions for different γ's is shown in Figure 10, as a function of terminal wealth (relative to current wealth). Intuitively, higher risk aversion penalizes lower terminal wealth more.

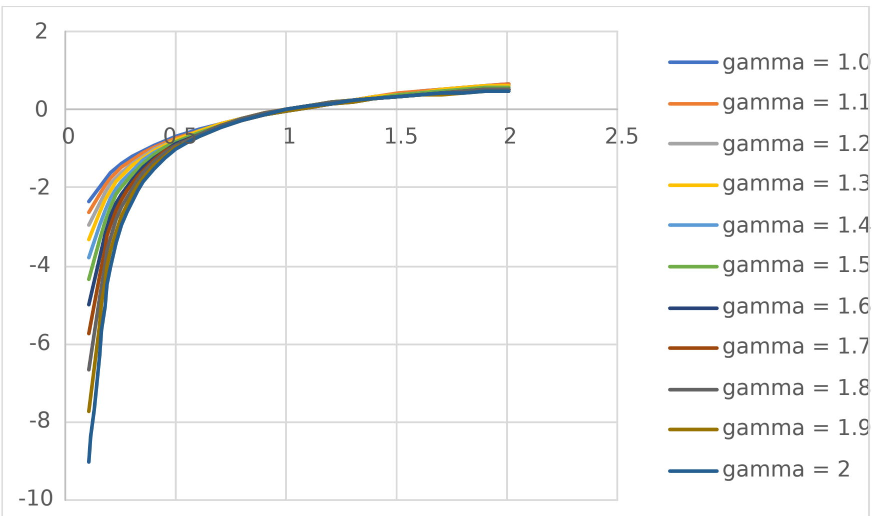


Figure 10: CRRA utility functions for different risk aversion γ

12.69.3 Log-Utility and Merton Ratio

The case of $\gamma \to 1$ is illuminating and corresponds to the utility function being log-utility:

$$U(w;1) = \ln(w).$$

In this case we have

$$\begin{aligned} \mathrm{E}\left(U(P(T);1)\right) &= \mathrm{E}\left(\ln(P(T))\right) \\ &= \int_0^T \left(r(t) + \kappa(t)\left(\mu(t) - r(t)\right) - \frac{1}{2}\kappa(t)^2 \xi(t)^2 \right) dt, \end{aligned}$$

and the optimal $\kappa(t)$ is particularly easy to determine, by setting the first derivative of the integrand with respect to κ to zero, to be

$$\kappa^*(t) = \frac{\mu(t) - r(t)}{\xi(t)^2}. \tag{104}$$

This allocation is know as the *Merton ratio* (for $\gamma = 1$).

12.69.4 Grown-Optimal and General Case

The portfolio $P^*_{\log}(\cdot)$ is called the *growth-optimal portfolio* as it maximizes the expected growth rate (log-utility maximization) of the portfolio in question. This is a continuous-time equivalent of the well-known *Kelly criterion*, see Kelly (1956), which is a method for determining the optimal fraction of wealth to bet on a series of gambles. This criterion aims to maximize the expected rate of growth of wealth. It is generally accepted that the growth-optimal portfolio is too risky for most individuals, i.e. the utility function for most investors is more conservative than $U(w; 1)$. In fact, it is estimated that γ for most investors is about 2 to 3.

It turns out that for general $\gamma > 1$, the optimal allocation is a proportion of the growth-optimal ratio (104):

$$\kappa^*(t) = \frac{\mu(t) - r(t)}{\gamma \xi(t)^2}, \tag{105}$$

where we now have the risk aversion parameter γ in the denominator. It is pretty straightforward, albeit somewhat tedious, to derive this result. HW: Derive this result.

Topic 13
Static Arbitrage in Volatility Markets

13.70 Static Replication of European Style Payoffs

Let us for the moment step away from interest rate derivatives and consider a general problem of valuing European-style derivatives with arbitrary payoffs. Specifically, let $S(t)$, $t \in [0, T]$, be the price process of some asset in the risk-neutral measure, where T is the maturity of a European-style payoff. The payoff is given by a function $f(x)$, so that the payoff of this security is given by

$$f(S(T))$$

at time T. It follows that the value of the security at time 0 is given by

$$V = \mathrm{E}\left(f(S(T))\right), \tag{106}$$

where we ignore discounting for now and going forward.

The question is, how do we value this option? It turns out it can be statically replicated by European calls and/or puts. Let us demonstrate how.

Let $c(T,K)$ be the value, at time 0, of a call option on $S(T)$ with strike K at time T:

$$c(T,K) = \mathrm{E}\left((S(T)-K)^+\right). \tag{107}$$

Consider the payoff of the call option as a function of the strike K,

$$C(K) = (s-K)^+.$$

(Note: plot here). Let us differentiate it with respect to K:

$$\frac{\partial C(K)}{\partial K} = \frac{\partial}{\partial K}\max(s-K,0).$$

Note that

$$\max(s-K,0) = (s-K)\times 1_{\{K<s\}} + 0\times 1_{\{K>s\}}.$$

The derivative of the first term with respect to K is -1, of the second is 0. There is one point where the function is not differentiable but it is still absolutely continuous so we have

$$\frac{\partial C(K)}{\partial K} = -1_{\{K<s\}}.$$

Applying this to (107) we obtain

$$\begin{aligned}
\frac{\partial}{\partial K}c(T,K) &= \frac{\partial}{\partial K}\mathrm{E}\left((S(T)-K)^+\right)\\
&= \mathrm{E}\left(\frac{\partial}{\partial K}\left((S(T)-K)^+\right)\right)\\
&= -\mathrm{E}\left(1_{\{K<S(T)\}}\right)\\
&= -\mathrm{P}\left(S(T)>K\right)\\
&= \Psi(K)-1,
\end{aligned}$$

where $\Psi(x)$ is the cumulative distribution function (CDF) of $S(T)$.

Differentiating the last equality again, we obtain

$$\frac{\partial^2}{\partial K^2}c(T,K) = \psi(K)$$

where ψ is the probability density function (PDF) of the distribution of $S(T)$. A similar result holds for the puts:

$$\begin{aligned}
\frac{\partial}{\partial K}\max(K-s,0) &= 1_{\{K>s\}},\\
\frac{\partial}{\partial K}p(T,K) &= \mathrm{P}\left(S(T)<K\right) = \Psi(K),\\
\frac{\partial^2}{\partial K^2}p(T,K) &= \psi(K).
\end{aligned}$$

This classical result (see Breeden & Litzenberger (1978)) means that the probability density of $S(T)$ can be reconstructed from the prices of calls (or puts) for a continuum of strikes on $S(T)$.

Now let us go back to the problem of evaluating (106). The expected value can be calculated by integrating the payoff against the density of $S(T)$, so that

$$V = \mathrm{E}\left(f(S(T)\right) = \int_{-\infty}^{\infty} f(K)\psi(K)\ dK.$$

The density, as we have just shown, is the second derivative of call (or put) prices, so we can write

$$\mathrm{E}\left(f(S(T)\right) = \int_{-\infty}^{\infty} f(K)\frac{\partial^2}{\partial K^2}c(T,K)\ dK = \int_{-\infty}^{\infty} f(K)\frac{\partial^2}{\partial K^2}p(T,K)\ dK.$$

This can be simplified further. Splitting the integral at the point $S(0)$, the current value of the spot, and integrating by parts twice we obtain

$$V = \mathrm{E}\left(f(S(T)\right) = f(S(0)) + \int_{-\infty}^{S(0)} f''(K)p(T,K)\ dK + \int_{S(0)}^{\infty} f''(K)c(T,K)\ dK. \tag{108}$$

HW: Derive this

In words, a European-style payoff can be exactly replicated by a portfolio of puts and calls, where the weights in the replicating portfolio are given by the second derivative of the payoff. Conversely, if the market price of a European-style option does not satisfy (108), there is an opportunity for *static arbitrage*.

13.71 Libor-In-Arrears

A rich variety of options with European-style payoffs is traded in interest rate markets, providing potential opportunities for static arbitrage. Let us start with a classical example of *Libor-in-arrears* (LIA). Libor-in-arrears pays the Libor rate at the beginning of the period (when it fixes), not at the end as customary for caplets (recall that a caplet is a call option on a Libor rate that pays at the end of the period, and a floorlet is a put option on the same). While most often a whole strip of such cash flows is used as a leg in a Libor-in-arrears swap, we focus our attention on a single cash flow; the valuation of a full strip follows by additivity. Let $T > 0$ be the start date, and M the end date of the period covered by a Libor rate. The forward Libor rate is given, for t such that $0 \le t \le T$, by

$$L\left(t,T,M\right) = \frac{P\left(t,T\right) - P\left(t,M\right)}{\tau P\left(t,M\right)}, \quad \tau = M - T;$$

we use simplified notation $L(t) = L(t,T,M)$ when there is no chance of confusion. The value, at time 0, of a Libor-in-arrears cash flow is then given by

$$V_{\mathrm{LIA}}(0) = \beta(0)\mathrm{E}\left(\beta(T)^{-1}L(T)\right),$$

where $\beta(t)$ is the continuously compounded money market account, and the expected value is taken under the risk-neutral measure Q. The standard approach to valuing payoffs that pay at time T would involve a switch to the T-forward measure, as then the expression under the expected value operator simplifies accordingly. Unfortunately, this is not convenient for LIA cash flows as traded caplets provide information about the distribution of the Libor rate in the M-forward measure, *not* the T-forward measure. We shall apply the M-forward measure in a moment; but for now, using the T-forward measure, we obtain

$$V_{\mathrm{LIA}}(0) = P\left(0, T\right) \mathrm{E}^{T}\left(L(T)\right).$$

While the expression looks rather simple, our progress along this route is hampered by the fact that $L(t) = L(t, T, M)$ is a martingale in the M-forward measure, not the T-forward measure. Thus,

$$\mathrm{E}^{T}\left(L(T)\right) \neq L(0).$$

To characterize this situation, let us define the concept of a *Libor-in-arrears convexity adjustment*, defined by the difference

$$D_{\mathrm{LIA}}(0) \triangleq \mathrm{E}^{T}\left(L(T)\right) - L(0).$$

This adjustment arises from the mismatch between the measure appropriate for the given payment date and the measure in which the market rate is a martingale.

Returning to the issue of valuing an LIA cash flow, we now write the valuation formula in the M-forward measure to obtain that

$$V_{\mathrm{LIA}}(0) = P\left(0, M\right) \mathrm{E}^{M}\left(\frac{1}{P\left(T, M\right)} L(T)\right).$$

Fortunately, the factor $\frac{1}{P(T,M)}$ can be rewritten in terms of the Libor rate,

$$\frac{1}{P\left(T, M\right)} = 1 + \tau L(T),$$

so that

$$V_{\mathrm{LIA}}(0) = P\left(0, M\right) \mathrm{E}^{M}\left(\left(1 + \tau L(T)\right) L(T)\right). \tag{109}$$

The rate $L(t)$ is a martingale in the M-forward measure, i.e. it has no drift,

$$\mathrm{E}^{M}\left(L(T)\right) = L(0).$$

In particular,

$$V_{\mathrm{LIA}}(0) = P\left(0, M\right) \left(L(0) + \tau \mathrm{E}^{M}\left(L(T)^2\right)\right). \tag{110}$$

We observe that LIA is just a European-style payoff as in Section 13.70 with $f(x) = x^2$. Using (108). we obtain

$$\mathrm{E}^{M}\left(L(T)^2\right) = L(0)^2 + 2\int_{-\infty}^{L(0)} p\left(T, K\right)\, dK + 2\int_{L(0)}^{\infty} c\left(T, K\right)\, dK, \tag{111}$$

where $p(T, K)$ $(c(T, K))$ are un-discounted values of put (call) options on the rate $L(T)$ with strike K, i.e. simple un-discounted floorlets (caplets). The values of such options are available from the market, and the value of the Libor-in-arrears cash flow is computed by integrating them up.

The power of the replication method goes beyond a mere calculation of the convexity value. As should be clear from the formula above, it also provides a way to hedge the Libor-in-arrears cash flow with standard puts and calls in a *model-independent* way. In particular, to hedge a contract with the payoff

$$(1 + \tau L(T))\, L(T),$$

we would

- Enter a short FRA (forward rate agreement, see Section 3.15.3) on $L(T)$.
- Put $\tau P(0, M) L(0)^2$ dollars into a money market account.
- Sell $2\tau \cdot (dK)$ K-strike puts for all $K \in (-\infty, L(0)]$.
- Sell $2\tau \cdot (dK)$ K-strike calls for all $K \in [L(0), \infty)$.

The hedge is static, i.e. it never needs adjustment throughout the life of the trade. And, as mentioned earlier, the hedge is model-independent, as it does not rely on any modelling assumptions.

This is how dealers would price LIA and, in principle, how they would hedge it. However, should LIA prices get our of sync with the caplets/floorlets for whatever reason, such as liquidity or directional flows, this would create "true" arbitrage for HFs that can be locked in by trading LIA vs portfolios of caplets/floorlets. Admittedly, it rarely happens for such simple products but hopefully it demonstrates the point.

13.72 European Swaptions

Let us quickly recall how swaptions work as it will be important for future considerations. A swaption is the right to enter a swap on its first fixing date. We work with the tenor structure

$$0 < T = T_0 < T_1 < \ldots < T_N, \quad \tau_n = T_{n+1} - T_n, \tag{112}$$

where $T = T_0$ is the start date of a swap, which we assume has the final payment date at T_N. We define the annuity and the swap rate as previously,

$$A(t) \triangleq \sum_{n=0}^{N-1} \tau_n P\left(t, T_{n+1}\right), \tag{113}$$

$$S(t) \triangleq \frac{P\left(t, T\right) - P\left(t, T_N\right)}{A(t)}. \tag{114}$$

We denote by Q^A the annuity measure, i.e. the measure for which $A(t)$ is the numeraire. The (payer) swap value at time $t < T$ with fixed rate k is given by

$$V_{\mathrm{swap}}(t) = P(t,T) - P(t,T_N) - k \sum_{n=0}^{N-1} \tau_n P(t,T_{n+1}),$$

which can be written as

$$V_{\mathrm{swap}}(t) = A(t)\left(S(t) - k\right).$$

We denote the swaption value by $V_{\mathrm{swaption}}(t)$. Then

$$\begin{aligned} V_{\mathrm{swaption}}(t) &= \beta(t)\mathrm{E}\left(\beta(T)^{-1}\left(V_{\mathrm{swap}}(T)\right)^+\right) \\ &= \beta(t)\mathrm{E}\left(\beta(T)^{-1}A(T)\left(S(T)-k\right)^+\right) \\ &= A(t)\mathrm{E}^A\left(\left(S(T)-k\right)^+\right). \end{aligned}$$

Hence, swaptions can be treated as (call) options on the swap rate. Moreover, the swap rate is a martingale in this annuity measure and, in particular, $\mathrm{E}^A(S(T)) = S(0)$.

Importantly, a continuum of swaption prices for different fixed rate k determines the full distribution of the swap rate $S(T)$ *in the annuity measure*.

13.73 CMS

CMS, or constant-maturity swaps, are a popular interest rate product. Instead of paying a Libor rate as in a floating leg of a standard Libor swap, these instruments pay a *swap rate*, such as a 10y swap rate. Since the rate always has the same tenor (in this example 10y), such a swap is called a constant maturity, or CMS, swap.

Valuing a CMS swap, by linearity, is reduced to valuing individual CMS-linked payments. In the notations of Section 13.72, the value of a single CMS flow is given by

$$V_{\mathrm{CMS}}(t) = \beta(t)\mathrm{E}\left(\beta(T)^{-1}S(T)\right).$$

Switching to the T-forward measure (as the CMS is paying at T although payment delays are possible, a fact we ignore here), we get

$$V_{\mathrm{CMS}}(t) = P(t,T)\mathrm{E}^T\left(S(T)\right).$$

It appears that it should be easy to calculate this as we just need the expected value of the swap rate; but we need the expected value in the "wrong" measure! We know the expected value in the annuity measure – this is just the forward swap rate as we discussed earlier in the notes, but

$$\mathrm{E}^T\left(S(T)\right) \neq S(0).$$

Let us then switch to the annuity measure. We have

$$V_{\text{CMS}}(t) = A(t)\mathrm{E}_t^A\left(\frac{1}{A(T)}S(T)\right). \tag{115}$$

Now we are in the "right" measure but we have an extra term $1/A(T)$ multiplying $S(T)$.

Many approaches have been developed to calculate/approximate (115). We do not have time to consider most of them here; see Andersen & Piterbarg (2010) if interested. Let us demonstrate what probably is the simplest (yet sufficiently accurate!) approach. We condition the expression inside the expected value operator on $S(T)$ using the tower rule, and then approximate the conditional expected value by a linear function:

$$\begin{aligned} V_{\text{CMS}}(t) &= \mathrm{E}_t^A\left(\frac{A(t)}{A(T)}S(T)\right) = \mathrm{E}_t^A\left(\mathrm{E}_t^A\left(\frac{A(t)}{A(T)}S(T)\middle| S(T)\right)\right) \\ &= \mathrm{E}_t^A\left(\mathrm{E}_t^A\left(\frac{A(t)}{A(T)}\middle| S(T)\right)S(T)\right) \\ &\approx \mathrm{E}_t^A\left(\left(1+\gamma\left(S(T)-S(t)\right)\right)S(T)\right) \\ &= S(t) + \gamma\mathrm{E}_t^A\left(\left(S(T)-S(t)\right)^2\right). \end{aligned} \tag{116}$$

HW: justify each step.

We do not discuss how to calculate γ here, see Andersen & Piterbarg (2010). What we have obtained is the following result:

$$\mathrm{E}^T\left(S(T)\right) = S(0) + \gamma\mathrm{E}^A\left(\left(S(T)-S(0)\right)^2\right),$$

where the first term on the rhs is the forward swap rate, and the second is called the *CMS convexity adjustment*. Note that it is in the right (annuity) measure and is given by a quadratic function of the swap rate, just like for Libor-in-arrears. Hence we can apply the replication method (108) to calculate the convexity adjustment in terms of swaptions on $S(T)$ across all strikes.

Alternatively, this result can provide arbitrage opportunities should CMS swaps deviate significantly from the prices of the replicating portfolio for technical or flow reasons. Admittedly it is not as "clean" as for LIA as we did make an approximation (linear dependence of inverse annuity on the swap rate) in (116), so one has to evaluate potential arbitrage using a few alternatives for that step.

13.74 IRR Swaptions

There are a few types of swaptions traded in the market and that is what makes this an interesting topic for arbitrage considerations. The type of a swaption described in Section 13.72 is called a *physically-settled swaption*, where the long swaption party has an option to enter into an actual swap (typically cleared on

CCP). Another type is a *cash-settled swaption*, sometimes called *collateralized cash price* swaption. Here, the swaption holder, rather than entering into a swap upon exercise (if in the money), receives a *cash payment*. The cash payment is calculated by looking up the value of the underlying swap on CCP. Mathematically, there is no difference between the two, and the type of swaptions traded in different currencies is a matter of convention[15].

There is a third type, the IRR (Internal Rate of Return) swaption, and is different from the other two. Just like the collateralized cash price swaption, it pays in cash upon exercise, but the amount of cash paid is determined by discounting swap cashflows at the rate equal to the *swap rate observed at the expiry* of the swaption, rather than using a proper discount curve. We will present formulas shortly. This is the type of swaption that has been around the longest and was the standard in GBP and EUR markets until relatively recently. To calculate the actual value of a swap, one needs to discount its cash flows using a proper discount curve. Back in the old days counterparties often disagreed on the curve as there were no CCPs, bid/ask spreads were wide, swaps were not very liquid, and so on. So not being able to agree on the value of the underlying swap was a major challenge. Hence, the IRR swaption format was used as it did not depend on the whole curve – counterparties only needed to agree on a fixing of a swap rate which, even back then, was easy to do as it was published by benchmark providers.

IRR swaptions continue to be traded in EUR markets alongside with the more modern (collateralized cash price) versions of swaptions.

An IRR swaption pays, at time T, the cash amount equal to

$$A_{\mathrm{IRR}}(S(T))\,(S(T)-k)^{+},$$

where $A_{\mathrm{IRR}}(S)$, the *cash annuity*, is a deterministic function of the swap rate S and is given by (for the tenor structure (112))

$$A_{\mathrm{IRR}}(S)=\sum_{n=0}^{N-1}\frac{\tau_n}{(1+\tau S)^{n+1}}$$

(here τ is a nominal day count fraction equal to $1/m$ where m is the number of periods in a year, i.e. $m=4$ for a 3M swap). The value of the IRR swaption is then given by

$$V_{\mathrm{IRR}}(t)=P(t,T)\mathrm{E}_t^T\left(A_{\mathrm{IRR}}(S(T))\,(S(T)-k)^{+}\right).$$

Switching to the annuity measure (where the distribution of $S(T)$ is known) we obtain

$$V_{\mathrm{IRR}}(t)=A(t)\mathrm{E}_t^A\left(\frac{A_{\mathrm{IRR}}(S(T))}{A(T)}\,(S(T)-k)^{+}\right). \tag{117}$$

[15] In practice the convention is baked into the ISDA Master Agreement, the main document regulating bilateral OTC trading between two counterparties, as a default choice.

Historically, the approximation

$$\frac{A_{\mathrm{IRR}}(S(T))}{A(T)} \approx 1 \tag{118}$$

was made by dealers, which would price IRR swaptions just like the physical ones. This was exploited/arbitraged by those sophisticated HFs that understood the difference. Over time it was realized that (118) is not a good approximation and the pricing methodology was adjusted.

We can apply the same idea to (117) as we used for CMS in (116). Using the conditional expected value we obtain a more accurate pricing formula

$$V_{\mathrm{IRR}}(t) = A(t)\mathrm{E}_t^A \left(A_{\mathrm{IRR}}(S(T))\left(1 + \gamma\left(S(T) - S(t)\right)\right)\left(S(T) - k\right)^+ \right).$$

Now the expression under the expected value operator is a deterministic function of $S(T)$, and the replication method can be applied to either price IRR swaptions in line with physically (or collateralized cash price) settled ones, or to find arbitrage between the former and the latter.

To learn more about the interplay between different types of swaptions see Tee & Kerkhof (2019).

13.75 Zero-Wide Collars

The difference between a payer (physically settled) swaption (call on the swap rate) and a receiver swaption (put on the swap rate) is, naturally, just the underlying swap, since we have for the payoff:

$$A(T)\left(S(T) - k\right)^+ - A(T)\left(k - S(T)\right)^+ = A(T)\left(S(T) - k\right) = V_{\mathrm{swap}}(T).$$

This is not the case for IRR swaptions where we have

$$\begin{aligned} A_{\mathrm{IRR}}(S(T))\left(S(T) - k\right)^+ &- A_{\mathrm{IRR}}(S(T))\left(k - S(T)\right)^+ \\ &= A_{\mathrm{IRR}}(S(T))\left(S(T) - k\right) \neq V_{\mathrm{swap}}(T). \end{aligned}$$

If $k = S(0)$, this combination is called a *zero-wide collar*. They have been trading since recently as a standalone instrument. If a dealer is pricing a zero-wide collar at 0 – like he would for the similar structure in physically-settled swaptions where it would be just an ATM swap – arbitrage opportunities would arise (and did exist until the differences became to be appreciated by all dealers).

There are more subtle ways in which ZWCs can be used to construct arbitrage strategies – see Lutz (2015).

Topic 14
Expressing Views with Options

14.76 One-by-Two, an Option Strategy Example

Static arbitrage opportunities in the options markets are the surest way to make money, but they are relatively rare, and generally cannot be executed in sufficient size to make enough money for a large HF. Options, however, can be used for other purposes, chief among them is efficiently expressing views on macro-economic developments (also of course for RV; we will look at a few examples shortly).

There are many option-based strategies employed by HFs and we will not be able to cover all (or even a small fraction) of them. However, to get a taste for what they are, let us consider one of the strategies, called 1x2 ("one-by-two").

Let us consider a situation where a CB, at the next meeting, is going to hike rates by either 25, 50, or 75 bps (for simplicity assume the rates now are zero). The forward rate is traded at 60. The market assigns non-zero probabilities to all three scenarios. We, however, have a strong conviction that it is most likely going to be 50. Specifically, our subjective probability of 25 is below the implied, higher for 50, and significantly below for 75. What trade should we put on?

We of course can always enter a receive-fixed FRA fixing right after the CB meeting. We would get the current market as the fixed rate (60) and pay 50 if our expectations realize, making a profit of 10 bps. However, in an FRA, we are essentially paying for the opportunity to make money in the 25 state, for which we are paying market price but we think it is not worth it. Here we can then employ the 1x2 strategy. This consists of buying 1 call option on the rate, and selling 2 call options on the same rate at a higher strike. For the long option we will choose a strike between 25 and 50, and for the short option the strike will be 50 – so the combined payoff has a peak at 50, our central scenario. The strike of the short option is at 50, and the strike of the long option is chosen so that the market value (i.e. the expected value of the payoff under market probabilities) of 1x2 is zero, so it is a zero-cost strategy just like an FRA. See `T14_1x2_01.ipynb` for a demonstration.

What we have done in essence, vs. a forward (FRA) on the rate, is that we gave up gains for the 25 bp hike that we deem unlikely, but improved our payoff for our central case of 50 bp hike.

HW: Modify `T14_1x2_01.ipynb` to use a 1x3 strategy, i.e. selling 3 options at the higher strike, to see if you can improve the expected payout even more.

14.77 Subjective Measure and Optimal Positioning

14.77.1 Problem Statement

There is a certain element of art/experience that is required to find "the best" way to express subjective views. In some idealized cases, however, this can be done systematically. Let us look at one of these situations.

Consider a real-valued market variable X and a fine grid on the real line $x_1 < \cdots < x_{N+1}$. An investor believes in the subjective probability distribution P^S of X being in each one of the intervals $[x_n, x_{n+1}]$. Let us denote it by b_n,

$$b_n = \mathrm{P}^S \left(X \in [x_n, x_{n+1}] \right), \quad n = 1, \ldots, N.$$

For each interval n there is also a market price for an option (digital call spread) that pays 1 if X ends up in $[x_n, x_{n+1}]$, and 0 otherwise, which we can express as

$$q_n = \mathrm{P}^Q \left(X \in [x_n, x_{n+1}] \right), \quad n = 1, \ldots, N,$$

where P^Q is the market-implied distribution. Given this information, what is the optimal payoff for this investor? I.e., what values f_n, $n = 1, \ldots, N$, should she choose so that the European-style option with the payoff

$$\pi(x) = \sum_{n=1}^{N} f_n 1_{[x_n, x_{n+1}]}(x) \tag{119}$$

is optimal for her?

14.77.2 Objective

The question of optimality requires a definition of what it means to be optimal. As a simple case, let us consider a growth-optimizing investor who optimizes the logarithm of the return (see also Section 12.67).

The cost of setting up an option with the payoff (119) is given by

$$\mathrm{E}^Q(\pi(X)) = \sum_{n=1}^{N} f_n q_n.$$

Let us assume the investor has a budget of 1 so the cost of setting up the portfolio should be 1,

$$\sum_{n=1}^{N} f_n q_n = 1. \tag{120}$$

The payoff of the option is given by

$$\sum_{n=1}^{N} f_n 1_{[x_n, x_{n+1}]}(X).$$

Then log-return is given, since the cost is 1, by

$$\ln(\pi(X)) = \ln\left(\sum_{n=1}^{N} f_n 1_{[x_n,x_{n+1}]}(X)\right) = \sum_{n=1}^{N} \ln(f_n) 1_{[x_n,x_{n+1}]}(X)$$

(the indicator functions are non-overlapping so can come out of the log). The expected log-return, under the subjective measure, is then

$$\mathrm{E}^S\left(\ln(\pi(X))\right) = \sum_{n=1}^{N} b_n \ln(f_n). \tag{121}$$

Hence the optimization problem for f_n's is to maximize (121) subject to (120),

$$\sum_{n=1}^{N} b_n \ln(f_n) \to \max \ \text{ subject to } \ \sum_{n=1}^{N} f_n q_n = 1.$$

14.77.3 Solution

The Lagrangian for this problem is given by

$$L(f) = \sum_{n=1}^{N} b_n \ln(f_n) + \lambda\left(\sum_{n=1}^{N} f_n q_n - 1\right).$$

Then

$$\frac{\partial L(f)}{\partial f_i} = \frac{b_i}{f_i} + \lambda q_i, \quad i = 1, \ldots, N.$$

Setting these to zero we obtain

$$f_i q_i = -b_i/\lambda, \quad i = 1, \ldots, N. \tag{122}$$

Summing these up over i, using the constraint (120) and the fact that $\sum b_i = 1$, we obtain

$$1 = \sum_{n=1}^{N} f_n q_n = -\frac{1}{\lambda}\sum_{n=1}^{N} b_n = -\frac{1}{\lambda}$$

so that finally, substituting $\lambda = -1$ into (122), we obtain our main result: the payoff that delivers the highest growth rate is given by

$$f_n = \frac{b_n}{q_n}, \quad n = 1, \ldots, N.$$

In the continuous limit, i.e. in the limit of the distance between grid points x_n going to 0, we obtain

$$f(x) = \frac{b(x)}{q(x)},$$

so the growth-optimal payoff is just the ratio of the subjective and implied probability density functions.

Note strong similarities with the Kelly criterion that we already mentioned in Section 12.67. Also this problem can be considered to be a bet sizing problem, so has strong links with what we discussed in Section 12.67. It solves a particular case when bets are disjoint, so that only one can pay, and complete, so that one of them will always pay.

Growth-optimal investing is considered too aggressive for most. The results above, and their extensions for more risk-averse investors, are obtained in Soklakov (2011) and other papers by the same author.

14.78 Structured Notes

Strong convictions (and being right more often than not), as expressed by the subjective measure, are a significant driving force for HF strategies. Differences between the statistical (P) and risk-neutral (Q) also drive interest rate markets in a meaningful way. In options markets, this is, arguably, most clearly seen in structured notes. Structured notes are issued by banks and sold to investors (retail or institutional) and typically offer attractive, above-market, coupons, usually financed by an investor implicitly selling options to the issuer. These are designed to make them attractive to investors who look at historical data to evaluate their worth (P-measure), whereas issuers are happy to pay market prices (Q-measure) for them since they are hedgers.

The universe of structured notes is rather vast and we will not be able to give it justice in these notes. Instead, we consider a simple example that, hopefully, conveys the essence of this market.

Looking at historical data, one notices that the yield curve is upward sloping pretty much all the time (recall our discussion on term premium). For example, a USD 10y swap rate is higher than the 2y swap rate during most of the historical period (but not at the time of writing in Q1 2023 – why?). Hence, an investor is likely to assign very low probability to the event "the 2y rate is higher than the 10y rate", a belief that manifests itself in him assigning very low value to digital options with the payoff

$$1\{S_{10y} - S_{2y} < 0\}.$$

On the other hand, the market price of such an option (Q-measure) is often quite a bit higher. So a structured note could be constructed where, for example, the issuer is paying above market coupon in exchange for the investor selling (usually implicitly, so that it is embedded in the coupon paid by the issuer) such "curve inversion digitals". In practice this may look like a so-called *CMS spread range accrual* note that pays a coupon, for the period $[T_n, T_{n+1}]$, $n = 0, \ldots, N-1$, of the form

$$c \times \frac{1}{M} \sum_{t=1}^{M} 1\{S_{10y}(t_i) - S_{2y}(t_i) > 0\},$$

where $\{t_i\}_{i=1}^{M}$ are the days between T_n and T_{n+1}. Here c would be higher than on a standard fixed-floating swap of the same maturity. The name comes from the fact that the coupon c "accrues" every day when the CMS spread $S_{10y} - S_{2y}$ stays above zero, i.e. while the curve is upward-sloping.

There are multiple variations on this theme of enhancing coupons with (implicitly) sold options, and part of the art of creating structured notes is in finding those options for which the difference in prices in the risk-neutral measure and the statistical measure is as high as possible.

Topic 15
Volatility Markets RV

15.79 Opportunities in Volatility Markets

Apart from macro positioning, interest rate options are used for all sorts of RV strategies. Many are "borrowed" from other markets such as equity derivatives, where volatility smiles play such an important role. The ideas behind the concept of carry and roll-down are extended from yield curves to volatility curves as well. In this section we will look at some RV-related concepts that are somewhat unique to interest rate markets as they arise from a very high-dimensional nature of the market.

There is a large variety of relatively simple volatility products that are traded in OTC interest rate markets. They all depend in some shape or form on the volatilities and correlations of various rates in a given currency. While different types of volatility products depend on the dynamics of the same curve, market segregation and variable demand for different types of volatility instruments create opportunities for those who can link different types of volatility products into a cohesive framework.

As a start, let us briefly review the main types of vanilla volatility instruments. To set the terminology, for a given rate (Libor or swap) we call the volatility between today and the fixing of that rate a *term* volatility. Some particular portion of that volatility is called a *forward* volatility (note there is no assumption that the volatility of a given rate is constant between now and the fixing of the rate – in fact most models apportion it unequally over time). There are types of volatility products, to be described later, that depend on forward volatilities. Volatilities in interest rate markets are typically measured in basis points per year, which are Normal volatilities (to be used in the Bachelier model, not in the Black-Scholes model, to convert between volatilities and prices). For example, volatility 120 means the rate's Normal volatility is 0.012. Correlations between different rates, Libor or swap, are also often important. Correlations can also be time-dependent in principle, i.e. two rates may experience different correlations at different times during their evolution.

We ignore volatility smiles in this section, but of course in reality, volatility products on the same rate are traded with different strikes, exhibiting volatility smile. So when talking about term and forward volatilities we typically mean at-the-money (ATM) volatilities.

15.80 Volatility Products

15.80.1 Caps and Floors

Caps and floors are collections of caplets and floorlets which are European options (calls and puts) on individual forward Libor rates. They naturally depend on term volatilities of forward Libor rates.

In the short end of the curve (up to 5 years, more or less) options on Libor futures (aka Eurodollar, or ED, futures) are also traded. These are essentially equivalent to caplets/floorlets but trade on exchanges, rather than OTC. Incidentally, the dislocations between options on futures and caps/floors do happen as they trade on different markets (exchange vs. OTC); this can lead to some arbitrage (or RV) opportunities.

15.80.2 European Swaptions

European swaptions are European options on swap rates and as such provide information about their term volatilities. A European swaption is defined by essentially two parameters (ignoring the strike, i.e. looking at ATM swaptions only) – in how many years does it expire (expiry; same as the fixing date for the underlying swap rate, or the beginning of the underlying swap), and how many years does the underlying swap span (tenor, or tail).

Term volatilities are naturally organized into a grid. The rows are expiry (in months/years), and the columns are the swap tenors or tails, also in months/years. A typical grid can be found in the Bloomberg historical data spreadsheet.

15.80.3 Midcurve Swaptions

A *midcurve swaption* is a European option on a forward starting swap. The option expires at, say, time T_e, but the swap does not start until some later time T_f. The gap between the two dates could be quite significant, on the order of years.

This is the first example of a vanilla instrument that depends on forward volatility. The term volatility of the T_f-swap rate is naturally given by the standard European swaption expiring at T_f. The midcurve is an option on the *same* swap rate that expires early, essentially giving us information on the portion of the term volatility that is assigned to the time interval $[0, T_s]$.

Note: a drawing here.

As we will see later, this forward volatility is in fact closely linked to a correlation of some swap rates, so an alternative viewpoint is that one can extract this correlation from midcurves.

There are also midcurve ED (Libor, and now OIS) futures – the expiry of the option is some time before the fixing of the relevant Libor rate.

15.80.4 Forward-Starting Swaptions

Closely linked to midcurve swaptions are *forward-starting swaptions.* For a swap that starts at T_f, and a time T_s prior to it, a forward-starting swaption pays

$$A(T_f)(S(T_f) - S(T_s))^+$$

at T_f. What this says is that the strike of the swaption is set at time T_s to the then-value of the underlying swap rate (rather than at time 0), and then it behaves like a normal swaption.

It is not hard to see that a forward-starting swaption isolates the portion of the volatility of the T_f-swap rate between T_s and T_f. In that sense, it complements the midcurve as this forward volatility can be derived from the term volatility and the midcurve volatility.

15.80.5 Amortizing/Accreting/Roller-Coaster European Swaptions

Swaps do not need to have constant notional throughout their life. So-called amortizing swaps have the notional for each period decreasing over time. Notional for each period in an accreting swap increases over time. There are also roller-coaster swaps where the notional moves up and down on a pre-determined schedule. Naturally, European options on these swaps are called *amortizing, accreting*, or *roller coaster* European swaptions.

Note: a drawing here, also for the decomposition mentioned below.

Swaps like this can be de-constructed into baskets of standard (constant notional) swaps. An option on one of these non-constant notional swaps is then an option on a basket of standard swaps. Therefore, its value depends on correlations between standard swap rates, so their valuation requires correlations or, alternatively, their market prices if available could be used as a source of market-implied correlations.

15.80.6 CMS Spread Options

For two rates, say S_{10y} and S_{2y}, fixing at the same time T, a *CMS spread option* pays

$$(S_{10y}(T) - S_{2y}(T) - k)^+$$

at time T, where k is the strike of the option. We have already encountered CMS spread in the section on structured notes.

The option being a option on a spread, it provides the most direct information on market-implied correlations between the two rates involved.

15.80.7 Bermudan Swaptions

A *Bermudan swaption* is an American-style option to enter a swap. Given a swap with a tenor structure

$$0 = T_0 < T_1 < \ldots < T_N,$$

a Bermudan swaption gives the holder the right to exercise on any of the dates T_n, $n = 1, \ldots, N-1$. If exercised on T_n, she receives a swap that goes from T_n to T_N.

Bermudan swaptions are not per se a vanilla product, but they are sufficiently liquid to be considered when trying to come up with a unified picture of the volatilities/correlations for a given market. Bermudans depend on the term volatilities of the so-called co-terminal swap rates with different fixing dates, i.e. swap rates covering periods $[T_n, T_N]$ for $n = 1, \ldots, N-1$. They also depend on forward volatilities of said rates or, alternatively, correlations among them.

15.81 RV Opportunities

Building a picture of all forward and term volatilities and correlations that are consistent with all these types of products is the Holy Grail of volatility modelling in interest rate markets, but it is very hard. It is more typical for market participants to look at the subsets of the types of these products and try to look for RV opportunities among, say, pairs of product types. Let us consider a couple of typical examples.

15.81.1 Caps vs. Swaptions

Looking for RV opportunities between caps and swaptions is perhaps the oldest RV strategy in IR volatility markets, since these product types have been around the longest and are the most liquid. Caplets of course could be seen as just single-period swaptions, i.e. the first column in the swaption volatility grid.

A model that is the most successful in being able to calibrate to the whole volatility grid is, arguably, the so called Libor Market (LM) model, sometimes simply called LMM (or BGM for the names of the authors who were among the first to introduce it). We do not have space here to describe it thoroughly; if interested, see Andersen & Piterbarg (2010) for an extended discussion. But in short, the model parametrizes forward volatilities of forward Libor rates, and correlations among them. There are formulas to calculate swaption volatilities from these model primitives. For valuation purposes of more exotic interest rate products, one tries to fit the whole volatility grid by having maximum flexibility in the model parameters. For RV purposes, one typically constrains these parameters by using functional forms for forward volatilities and correlations with a relatively small number of parameters. Then one does a best fit to the volatility grid and finds those segments of the grid that are above or below the fitted model. This is akin to how the NSS curve model is used to find cheap/rich bonds, but is higher-dimensional, both in the input parameters and the output targets sense.

A very basic approximation of a swap rate in terms of forward Libor rates for the purposes of volatility calculations (much more accurate are possible and normally used) is obtained by a technique called “freezing the annuity”. Recall

($T_0 = T > 0$ is the expiry of the swaption/fixing of the swap rate)

$$\begin{aligned} S(T_0) &= \frac{1 - P(T_0, T_N)}{A(T_0)} \\ &= \frac{\sum_{n=0}^{N-1} \tau_n L_n(T_0) P(T_0, T_{n+1})}{A_n(T_0)} \\ &= \sum_{n=0}^{N-1} w_n(T_0) L_n(T_0), \end{aligned} \tag{123}$$

where we have denoted

$$w_n(t) \triangleq \frac{\tau_n P(t, T_n)}{A_n(t)}, \quad n = 0, \ldots, N-1,$$

and L_n is the Libor rate for the period $[T_n, T_{n+1}]$. The freezing trick replaces $w_n(T_0)$ in (123) with their initial values $w_n(0)$. The rationale is that their volatility is much lower than the Libor rates volatilities (e.g. the ratios are nearly constant under parallel shifts of the yield curve, the predominant mode of yield curve dynamics). This leads to the expression

$$S(T_0) \approx \sum_{n=0}^{N-1} w_n(0) L_n(T_0)$$

where $w_n(0)'s$ are deterministic. The volatility of $S(T_0)$ is then obtained from (forward) volatilities of Libor rates (which are LM model's primitives) and their correlations (also part of the LMM specification).

15.81.2 European vs. Midcurve Swaptions

Continuing with the tenor structure as above, let T_n be the expiry of the midcurve swaption and T_m the start of the underlying swap, $n < m < N$. We will decorate relevant quantities with the start and the end date index, e.g. the underlying swap for the midcurve has value $V_{m,N}$, swap rate $S_{m,N}$, annuity $A_{m,N}$, etc.

We have the following identity,

$$V_{m,N}(T_n) = V_{n,N}(T_n) - V_{n,m}(T_n). \tag{124}$$

This just says that on date T_n, a "long" spot-starting swap $V_{n,N}(T_n)$ can be split into two parts, a "short" spot-starting swap $V_{n,m}(T_n)$ and a forward starting swap $V_{m,N}(T_n)$. We can rewrite (124) using annuities and swap rates, dividing by the annuity of the forward-starting swap:

$$S_{m,N}(T_n) = \frac{A_{n,N}(T_0)}{A_{m,N}(T_0)} S_{n,N}(T_n) - \frac{A_{n,m}(T_0)}{A_{m,N}(T_0)} S_{n,m}(T_n). \tag{125}$$

Then we employ the same freezing technique as in the previous section to write

$$S_{m,N}(T_n) \approx r_1 S_{n,N}(T_n) - r_2 S_{n,m}(T_n), \tag{126}$$

where

$$r_1 = \frac{A_{n,N}(0)}{A_{m,N}(0)}, \quad r_2 = \frac{A_{n,m}(0)}{A_{m,N}(0)}.$$

The rationale for this approximation is as before, the variability in the ratios of annuities is much smaller than in the swap rates.

From (126) we can derive the volatility for the midcurve, i.e. the forward volatility, from the term volatilities of $S_{n,N}$ and $S_{n,m}$, and the correlation between these rates. The correlation can be parametrized directly (e.g. estimated from historical data) or derived from a calibrated LM model as in the previous section.

15.81.3 Bermudan Swaptions

All dealers that are active in this market are structurally long Bermudan swaptions. Callable (by dealers) structured notes are a popular product, leaving dealers long Bermudans. Mortgage providers sell Bermudans to hedge their early prepayment exposure to dealers. There are no natural buyers of Bermudans on the "buy" side. Because of the one-way flow, Bermudans are traded at a significant discount to their theoretical model value under any reasonable model. Hedge funds that are sophisticated (and brave!) enough, can buy Bermudans from dealers at discounted prices, as dealers are only happy to off-load parts of their Bermudan books. The strategy here is to buy and hold Bermudans and either wait till expiry or exercise if the opportunity arises, monetizing the discount. There also a potential to link Bermudans to CMS spread options and look to monetize Bermudans by clever hedging, but this is not trivial.

15.81.4 Conclusions

These are just a couple of typical RV-type strategies involving vanilla volatility instruments. Many more exist, as the dimensions of the input space (primitives of the volatility/correlation structure) and the output space (products/product types traded) is large. In addition to relative value, carry and roll-down are the concepts that are applicable to volatilities in a similar way as they are applicable to yield curves. Of course, there is also the realized vs. implied volatility strategies, that have some subtleties because of the structure of interest rate markets. Finally, the skew dimension, the volatility smile, is another rich source of risk premia extraction.

Topic 16
Arbitrage with Spread Options

16.82 Multi-Underlying Options

Options on individual underlyings are very liquid in a variety of markets. In many markets, additionally, options on *linear combinations of underlyings* are also reasonably liquid. A common type of multi-underlying options is the so-called spread options (i.e. options on the difference of two underlyings), such as CMS spread options in the interest rate markets. While CMS spread options (which we already briefly encountered in Section 15.80.6) are of primary interest to us in this Topic, the discussion here is applicable to generic spread options, and this is how we will present the ideas here. We will be looking for ways to identify arbitrage among single- and multi-underlying options, with spread options being the main focus.

This is not a theoretical exercise – there have been times in history when arbitrage opportunities involving spread options existed, and savvy market participants "in the know" took advantage of them.

We fix a time and only consider options that expire at that time. We will be considering a market where there are two underlyings with liquid European options on each, and also European options on their spread (difference) with the same expiry. Because of call-put parity we will only be looking at calls. We denote the observable traded prices of these call options by

$$C_X(K), \quad C_Y(K), \quad C_S(K) \quad \text{for all strikes } K,$$

which are the market prices of options on the underlyings X, Y, and their spread S.

16.83 Fundamental Theorem of Asset Pricing

Let us recall the Fundamental Theorem of Asset Pricing (FTAP) stating, in our simplified single-period economy, that

> *The absence of arbitrage in a market with multiple underlyings is equivalent to the existence of a joint distribution of these underlyings such that the values of all derivatives (call options in our case) are given by the expected values of their payoffs under this distribution.*

FTAP gives us a way to detect arbitrage opportunities. If we can construct a joint distribution of X and Y such that

$$C_X(K) = \mathrm{E}\left((X-K)^+\right),$$
$$C_Y(K) = \mathrm{E}\left((Y-K)^+\right),$$
$$C_S(K) = \mathrm{E}\left((X-Y-K)^+\right)$$

for all K, where $\mathrm{E}(\cdot)$ is the expected value of this joint distribution, then there is no arbitrage. But if we can prove that no such joint distribution exists, then there is arbitrage, and the question of identifying the best trade to realize it becomes interesting and important.

16.84 Some Special Cases

In this section we consider a few examples to build intuition.

16.84.1 One-Dimensional Case

To warm up, let us consider a market where only $\{C_X(K), \text{ all } K\}$ are observable. How can we detect arbitrage in this price system?

We have already established the connection between call prices and the PDF of the underlying, see Section 13.70. Recall that the PDF of X is the second derivative of the call option price with respect to the strike, i.e.

$$p_X(K) = \frac{\partial^2}{\partial K^2} C_X(K). \tag{127}$$

If $p_X(K)$ thus defined is a proper PDF, i.e. it is non-negative and integrates to one, then we found our distribution, because then (127) could be inverted so that

$$C_X(K) = \int (x-K)^+ p_X(x)\; dK = \mathrm{E}\left((X-K)^+\right) \text{ for any } K,$$

where $\mathrm{E}(\cdot)$ is the expected value operator corresponding to the PDF $p_X(K)$.

What if $p_X(K)$ defined by (127) is not a proper PDF, e.g. there exists K^* such that $p_X(K^*) < 0$? What is the arbitrage? We have from (127) that

$$\frac{\partial^2}{\partial K^2} C_X(K^*) < 0.$$

Hence there exists $\varepsilon > 0$ such that

$$B(K, \varepsilon) \triangleq \frac{1}{\varepsilon^2}\left(C_X(K^* - \varepsilon) - 2C_X(K^*) + C_X(K^* + \varepsilon)\right) < 0,$$

because the left-hand side is a converging (in $\varepsilon \to 0$) approximation to $\partial^2 C_X(K)/\partial K^2$. Here $B(K, \varepsilon)$ is the price of the *butterfly* payoff

$$\frac{1}{\varepsilon^2}\left((X - (K^* - \varepsilon))^+ - 2(X - K^*)^+ + (X - (K^* + \varepsilon))^+\right)$$

which is always non-negative, and strictly positive for $X \in [K^* - \varepsilon, K^* + \varepsilon]$ (HW: convince yourself that this is true by drawing the payoff diagram). Negative price for this positive payoff implies the existence of the so-called *butterfly arbitrage*.

HW: What if the other condition for a valid PDF, $\int p_X(K)\ dK = 1$, is violated? What is the arbitrage then?

16.84.2 Two-Dimensional Case

Let us now add call options on the second underlying $\{C_Y(K),\ \text{all } K\}$ to the economy. If $\{C_X(K)\}_K$ are arbitrage-free amongst themselves, and so are $\{C_Y(K)\}_K$, it is intuitively obvious that the whole economy is arbitrage-free, as there are no securities that link X to Y. How can we prove it? Referring to FTAP, we need a joint density of X and Y such that all derivatives (calls on X and Y) are priced as expectations against this joint density. The easiest construction is obtained by just assuming independence of X and Y; namely defining the joint density as a product of marginal (i.e. of X and Y) densities,

$$p_{X,Y}^{\text{indep}}(x, y) \triangleq p_X(x)p_Y(y).$$

HW: convince yourself that this works, i.e. it is a proper density and calls on X and Y for all strikes are obtained by integrating the appropriate payoff against $p_{X,Y}^{\text{indep}}(x, y)$.

As an aside, one can also use *copulas* to create any possible joint distribution consistent with the marginals. We do not have the time to explore this in detail here, but see Andersen & Piterbarg (2010) (and many other books) if interested.

16.84.3 2D+Spread Case

Finally, let us add spread options $\{C_S(K),\ \text{all } K\}$ to the economy. The situation becomes radically more difficult and is the subject of the rest of the Topic. This is so because now we have securities that provide us with information on the joint behaviour of X and Y, as clearly an option on $X - Y$ tells us something on the dependence between X and Y. Not the complete information on the joint distribution of X and Y as we will demonstrate in due course, but partial information that needs to be incorporated into the joint distribution of X and Y.

There are, however, some simple necessary conditions for the absence of arbitrage. From the densities p_X, p_Y and p_S we can calculate the means of these variables μ_X, μ_Y and μ_S via the usual formula

$$\mu_X = \int x p_X(x)\ dx$$

and the same for others. Clearly if

$$\mu_S \neq \mu_X - \mu_Y \tag{128}$$

we cannot have a valid joint distribution of X and Y because for any joint distribution (measure) we must have

$$\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y).$$

Clearly this is not a sufficient condition.

16.84.4 2D+Spread: Gaussian Case

We can tackle the general problem in some specific cases. Let us assume that the market prices of all options are such that $p_X(x)$, $p_Y(y)$ and $p_S(s)$ are (one-dimensional) Gaussian, so that

$$X \sim \mathcal{N}(0, \sigma_X^2), \quad Y \sim \mathcal{N}(0, \sigma_Y^2), \quad S \sim \mathcal{N}(0, \sigma_S^2). \tag{129}$$

Here for simplicity we assumed zero means for all three variables. Then we can ask ourselves, does there exist a *joint Gaussian* distribution for (X, Y) that is consistent with (129)?

A 2D Gaussian distribution is parametrized by two standard deviations, one for each element of the Gaussian vector, and a correlation ρ between them. The standard deviations must clearly be σ_X and σ_Y. Given some correlation ρ, we have

$$\mathrm{Var}\,(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y.$$

From (129), $\mathrm{Var}\,(X - Y)$ must be equal to σ_S^2 for market values of options on the spread to be consistent with our Gaussian model. Thus we have the following expression for the *implied correlation* ρ_{impl},

$$\rho_{\text{impl}} = \frac{\sigma_S^2 - \sigma_X^2 - \sigma_Y^2}{2\sigma_X\sigma_Y}.$$

For the joint Gaussian distribution to exist, the correlation must be between -1 and 1. Hence the necessary condition for no arbitrage is given by the following condition on the standard deviations of X, Y and S:

$$\begin{aligned} -1 &\le \rho_{\text{impl}} \le 1, \\ -1 &\le \frac{\sigma_S^2 - \sigma_X^2 - \sigma_Y^2}{2\sigma_X\sigma_Y} \le 1. \end{aligned} \tag{130}$$

So if this condition is satisfied there is no arbitrage in this economy.

The converse, interestingly, does not hold. Even if (130) is not satisfied, there may well be a valid joint distribution of X, Y that reprices all the options. Of course, it cannot be a Gaussian distribution, but FTAP does not require it to be Gaussian, just its existence.

16.84.5 2D+Spread: Triangle Arbitrage

Recall the necessary condition (128) which is obviously pretty basic, and is almost never violated. Stronger necessary (but not sufficient) no-arbitrage conditions have been developed over the years. One of the best-known is the so-called *triangle arbitrage*.

Triangle arbitrage, well-covered in McCloud (2011), is based on the observation that

$$(X-K_X)^+-(Y-K_Y)^+ \leq (X-Y-(K_X-K_Y))^+ \leq (X-K_X)^+ + (K_Y-Y)^+ \tag{131}$$

for any strikes K_X, K_Y, so if a joint distribution for (X,Y) exists, then we must have that

$$C_X(K_X)-C_Y(K_Y) \leq C_S(K_X-K_Y) \leq C_X(K_X)+P_Y(K_Y), \tag{132}$$

where $P_Y(K) = C_Y(K) - \mathrm{E}(Y) + K_Y$ is the put on Y.

This is the triangle arbitrage condition stating that options on the spread must be within the lower and upper bounds that depend only on the marginal distributions. Conversely, if one of these bounds is violated, there is an arbitrage opportunity.

This is not a sufficient condition i.e. even if (131) holds true for all strikes, there could still be arbitrage, as we will show later, see Section 16.86.3.

The bounds in (131) follow from a simple observation that for any a, b,

$$\max(a+b,0) \leq \max(a,0)+\max(b,0). \tag{133}$$

The most direct proof uses simple properties of the function $x \longmapsto \max(x,0)$, as it is convex and homogeneous:

$$\begin{aligned}\max(a+b,0) &= 2\max\left(\frac{a+b}{2},0\right)\\ &\leq 2\left(\max\left(\frac{a}{2},0\right)+\max\left(\frac{b}{2},0\right)\right)\\ &= \max(a,0)+\max(b,0).\end{aligned}$$

Alternatively one can consider a few cases such as $a \geq 0$, $b \geq 0$; $a \geq 0, b \leq -a$; etc.

To prove the rhs of (131), we use (133) with $a = X - K_X$ and $b = K_Y - Y$. To prove the lhs, we use $a = Y - K_Y$ and $b = X - Y - (K_X - K_Y)$.

Some refinements of these bounds are possible (see McCloud (2011)) but they still do not give us sufficient conditions for no arbitrage.

16.85 Existence of a Joint Distribution

In this section we present the main theoretical result of this Topic that links the existence of a joint distribution (i.e. no arbitrage) to a certain condition on

European style options with general payoffs on X, Y and S. This is a generalization of the triangle arbitrage (132) that, critically, is a necessary *and a sufficient* condition.

As before, let X and Y be two random variables representing two underlyings, and let S represent their spread (difference). Recall that their distributions (densities) are known from the options markets on X, Y, and S.

Let $f(x)$, $g(y)$ be two functions; we think of them as defining payoffs on X and Y correspondingly. In particular the value of an option on X with payoff $f(x)$ is $\mathrm{E}(f(X))$.

Let us define the *spread envelope* of f, g by

$$\mathcal{E}_{f,g}(z) = \max_{x-y=z} \{f(x) + g(y)\}.$$

The meaning of the spread envelope should be clear – it is the smallest function of $x - y$ only that dominates the function $f(x) + g(y)$ for all x, y:

$$\mathcal{E}_{f,g}(\cdot) = \min_{h(\cdot)} \{f(x) + g(y) \leq h(x - y) \text{ for all } x, y\}.$$

Figure 11 shows an example of a spread envelope for $f(x) = 1\{x > 0\}$ and $g(y) = 1\{y < 0\}$. To better understand the concept, the reader is encouraged to use the notebook `T16_Spread_Envelope_01.ipynb`.

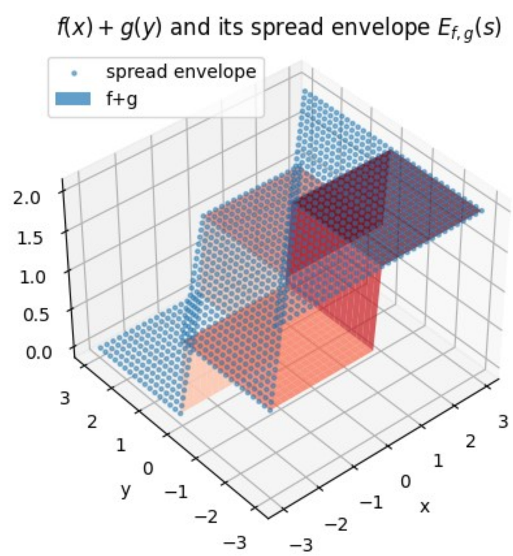


Figure 11: An example of the spread envelope for $f(x) = 1\{x > 0\}$ and $g(y) = 1\{y < 0\}$.

If (X, Y) have a joint distribution such that $X-Y$ has the same distribution as S, then we would clearly have

$$\mathrm{E}\left(f(X)\right) + \mathrm{E}\left(g(Y)\right) \leq \mathrm{E}\left(\mathcal{E}_{f,g}(X-Y)\right),$$

since the absence of arbitrage implies that, if one payoff is dominated by another, the value of the corresponding option is lower. It turns out that the reverse is also true, and that is the main theoretical result of this Topic.

Theorem 6. *Let three random variables X, Y, S be given. If for any continuous, bounded payoffs $f(\cdot)$, $g(\cdot)$ we have that*

$$\mathrm{E}\left(f(X)\right) + \mathrm{E}\left(g(Y)\right) \leq \mathrm{E}\left(\mathcal{E}_{f,g}(S)\right),$$

then, and only then, there exists a joint distribution of (X, Y) such that it has marginals X, Y and the distribution of $X - Y$ is the same as S.

The result has clear financial interpretation. It states that if we cannot construct an arbitrage that involves an arbitrary payoff f of X only, g of Y only and their spread envelope (a function of S only), then a joint distribution exists. This is a version of FTAP specialized to our particular problem. The proof is given in Piterbarg (2011).

Recall the triangle arbitrage (132). The connection with this result is seen from the fact that the triangle inequality (131) is a special case of a spread envelope construction (for the lower bound we would take $f(x) = (x - K_S)^+$ and $g(y) = -(y - K_Y)^+$ and similar for the upper bound). As we pointed out before, the absence of triangle arbitrage is not sufficient for the existence of the joint distribution. In other words, we may have a situation where triangle arbitrage is absent but more general *spread envelope arbitrage* still exists, see Section 16.86.3 below. It is an open question whether it is sufficient to check the no-arbitrage conditions in Theorem 6 for only a subset of functions f, g.

16.86 Spread Options by Linear Programming

Theorem 6 gives us a nice theoretical result with strong financial interpretation, but from a practical prospective it is not very useful as it would be impossible to check all possible payoffs for spread envelope arbitrage. In this section we develop a practical approach to the existence problem and related questions. Given that in actual computer calculations we always work with discretized quantities, we assume that X, Y are discrete, with the support on the common grid $\{i/N, i = 0, \ldots, N\}$. We assume that S is on the grid $\{(i-N)/N, i = 0, \ldots, 2N\}$. Let us denote the distributions of the three random variables by:

$$r_i = \mathrm{P}(Y = (N-i)/N),\ c_j = \mathrm{P}(X = j/N),\ d_k = \mathrm{P}(S = (k-N)/N),$$

with $i, j = 0, \ldots, N$ and $k = 0, \ldots, 2N$. Vectors r, c, d are non-negative with the elements summing up to 1. Note that we label "rows" r_i in the reverse order (i.e.

r_0 corresponds to the maximum value of Y) for convenience and to reconcile the Descartes view of the world (with the y-axis pointing upward) with the matrix indexing on computers where the row index increases in the downward direction.

16.86.1 Existence

We can reformulate the existence question in the discrete setting as follows. Under what conditions on r, c, d does there exist a matrix $p = \{p_{i,j}\}_{i,j=0}^N$ such that

$$p_{i,j} \geq 0, \quad i, j = 0, \ldots, N, \tag{134}$$

$$\sum_{j=0}^{N} p_{i,j} = r_i, \quad i = 0, \ldots, N, \tag{135}$$

$$\sum_{i=0}^{N} p_{i,j} = c_j, \quad j = 0, \ldots, N, \tag{136}$$

$$\sum_{(i,j)\in D_k} p_{i,j} = d_k, \quad k = 0, \ldots, 2N, \tag{137}$$

where D_k is the k-th diagonal,

$$D_k = \{(i, j) : i + j = k, 0 \leq i, j \leq N\}, \; k = 0, \ldots, 2N.$$

While a theoretical answer is given by a discrete version of Theorem 6, ideas behind *Linear Programming* (LP), see Gass (2010), give us a way to find a practical solution. Recall that LP is concerned with optimizing a linear function given a set of linear equality and inequality constraints. Since (134)–(137) is nothing but a set of linear equality and inequality constraints, the problem of finding p is just a problem of finding a *feasible* solution to an LP problem, i.e. some point in the solution set that is not necessarily optimal. To find a feasible solution, another LP problem can be set up using so-called *slack* variables q_i, $i = 0, \ldots, 4N + 2$, where we have one for each row, column and diagonal constraint in (135)–(137). So we look for $\{p_{i,j}\}$, $\{q_i\}$ such that

$$\sum_{i=0}^{4N+2} q_i \to \min, \tag{138}$$

$$p_{i,j} \geq 0, \quad q_i \geq 0,$$

$$\sum_{j=0}^{N} p_{i,j} \underline{+q_i} = r_i, \quad \sum_{i=0}^{N} p_{i,j} \underline{+q_{N+1+j}} = c_j, \quad \sum_{(i,j)\in D_k} p_{i,j} \underline{+q_{2N+2+k}} = d_k$$

(we underline slack variables). LP problems can be solved by the simplex method that, basically, goes vertex to vertex improving the objective function.

A starting point to (138) is easy to find – we can just take $p_{i,j} = 0$ and

$$q_i = \begin{cases} r_i, & i \le N, \\ c_{i-N-1}, & N+1 \le i \le 2N+1, \\ d_{i-2N-2}, & 2N+2 \le i. \end{cases}$$

Then running the simplex method we will find the (globally) optimal solution to (138). If it satisfies $q_i = 0$ for all i, then the original problem (134)–(137) has a solution given by $\{p_{i,j}\}$. If not, the solution to (134)–(137) does not exist. The simplex algorithm is very efficient and can easily handle the case of $N = 100$, say (implying about 10,000 variables to optimize over).

16.86.2 Extreme Solutions

Having constructed *a* solution, the question then becomes how we could find *all* solutions or, failing that, solutions with special properties. Of particular interest in that regard are solutions that maximize or minimize the value of some options on X, Y that are not locked in by the construction, i.e. options other than on X, Y or $X - Y$. For example, consider options on the index, or basket, $X + Y$. An important practical question is what are the potential bounds on the value of such an option given that the marginals and all the spread options (options on $X - Y$) are fixed. Here, too, LP gives us a practical answer.

The value of an option with the payoff $(X + Y - K_{\text{idx}})^+$ is a linear function in p:

$$T(p) = \sum_{i,j=0}^{N} \left(\frac{j}{N} + \frac{N-i}{N} - K_{idx} \right)^+ p_{i,j}.$$

Hence, we can find the joint density p that satisfies constraints (134)–(137) and maximizes the value of the index option (for a given strike K_{idx}) by solving the following LP problem:

$$T(p) \to \max, \text{ subject to constraints (134)-(137).} \tag{139}$$

Likewise, the density that gives the lowest value to the index option is found by replacing max with min in (139). In both cases the simplex algorithm can start from the feasible solution found in Section 16.86.1. The difference in values of the index option under the two densities gives a measure of model uncertainty for the case when only options on marginals and the spread are traded.

To consider an example, we look at a test case where both X and Y are Gaussian with zero mean and volatility 1%, and S is Gaussian with volatility 0.77% (implying correlation of 70% between X and Y). The values of the ATM index option in the two cases correspond to the implied Gaussian copula correlation of 9% (low option on index) and above 100% (high option on index). The densities of the index $X + Y$ in the two cases are shown in Figure 12.

Figure 12 also demonstrates an undesirable property of the simplex method in that it would set as many elements of p to zero as it can (normally # of variables minus # of constraints). The resulting density can be quite "spiky"

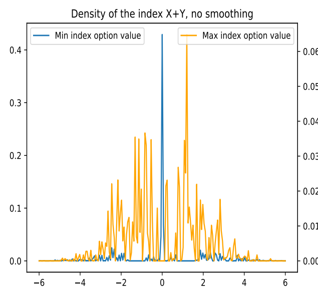


Figure 12: Densities of the index $X + Y$ using the simplex method. Note two different scales used for clarity.

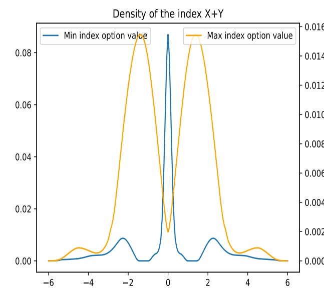


Figure 13: Densities of the index $X + Y$ using the QP method with smoothness. Note two different scales used for clarity.

and unrealistic from the financial prospective, see Figures 14, 15 for the view of the joint density of X and Y “from the top” (like a contour plot).

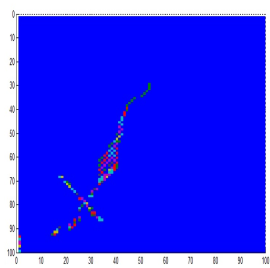


Figure 14: Joint density of X, Y using the simplex method for min index. Viewing from top.

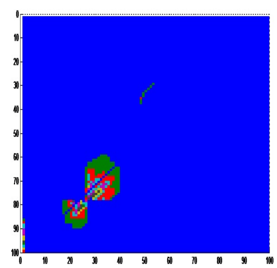


Figure 15: Joint density of X, Y using the simplex method for max index. View from top.

Another method for solving LP problems, the so-called “interior point” method, usually results in smoother solutions; however, there is no guarantee as to how smooth the solution would be. To obtain densities with a desired level of smoothness, we can modify our method and instead of the linear optimization use Quadratic Programming (QP), where the objective function is quadratic while the constraints are still linear. Efficient numerical methods for QP also exist. To put this into practice, we would replace the objective function in (139) with something like

$$T(p) + w_{\text{smooth}} \sum_i \sum_j \left((p_{i,j} - p_{i,j+1})^2 + (p_{i,j} - p_{i+1,j})^2 \right) \to \min,$$

where the smoothing terms penalize high local variations in the density. Figures 16 and 17 show two extreme densities with the smoothness constraints, and

Figure 13 shows the two corresponding densities of the index $X+Y$. The ATM index option value corresponds to the implied Gaussian copula correlation of 49% (low index option value) and 82% (high index option value).

Adding a modest amount of smoothness reduces the gap between the maximum and minimum index option values achieved but only marginally, while generating a much more plausible looking joint distributions.

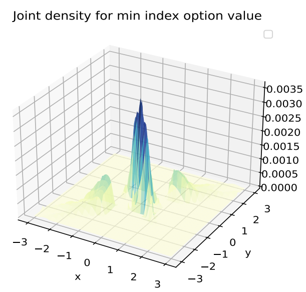


Figure 16: Joint density for min index value

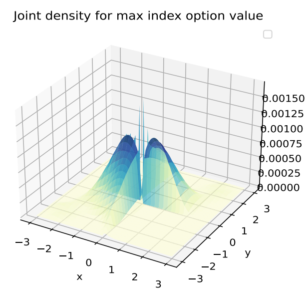


Figure 17: Joint density for min index value

16.86.3 Finding Arbitrage

If no solution to (134)–(137) exists, Theorem 6 implies that there is arbitrage that can be realized by trading in options on X, Y and $X-Y$. How do we find what positions we need to put on? Again, LP gives us a solution. We set up the following optimization problem (a problem *dual* to (138)): find vectors f, g, h such that

$$\sum_j f_j c_j + \sum_i g_i r_i - \sum_k h_k d_k \to \max \tag{140}$$

subject to

$$h_k \geq f_{i'} + g_{j'} \text{ for all } (i', j') \in D_k \tag{141}$$

for all k and

$$|f_i|, |g_j| \leq 1 \text{ for all } i, j. \tag{142}$$

The constraints (141) require the function $h(S)$ to dominate $f(X)+g(Y)$ for all values of the underlying random variables, and the constraints (142) prevent infinite solutions (for numerical robustness we may also impose boundness conditions on h such as $|h_k| \leq H$ for some H). If the optimal value of the objective function in (140) is positive, then we have found arbitrage – we sell options with payoffs $f(X)$ and $g(Y)$, buy the one with the payoff $h(S)$, get positive cash upfront (from (140)), and then have a position that can never result in a negative payoff (from (141)). Note that by the optimality property, the solution h will be the spread envelope of f and g.

As an example, consider CMS spread option prices observed in November 2010 in Euro, which happened to present an arbitrage opportunity. Using the procedure outlined above, we obtain the following results. Figures 18, 19, 20 show the market-implied densities of X, Y and $X - Y$, together with the functions $f(x)$, $g(y)$ and $h(s)$.

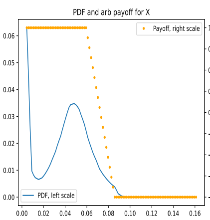


Figure 18: PDF of X and spread envelope arbitrage function $f(x)$

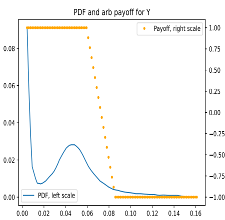


Figure 19: PDF of Y and spread envelope arbitrage function $g(x)$

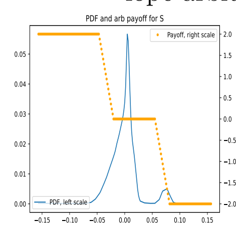


Figure 20: PDF of $X - Y$ and spread envelope arbitrage function $h(s)$

In Figure 21 we present the optimal spread arbitrage payoff achieved with the marginals $f(x) + g(y)$, and the payoff of the spread $h(s) = h(x - y)$, as a 3D plot. We can see $h(x - y)$ dominates $f(x) + g(y)$ (and is, naturally, the spread envelope of f, g), yet, in this case,

$$\mathrm{E}(f(X)) + \mathrm{E}(f(Y))) = 1.59 > 0.08 = \mathrm{E}(h(S)).$$

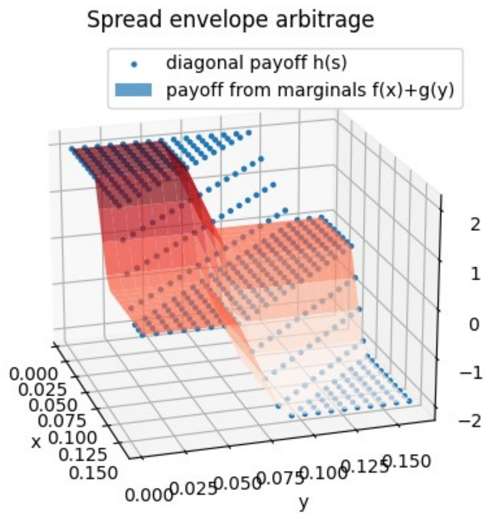


Figure 21: Spread arbitrage functions $f(x) + g(y)$ and $h(x - y)$.

It is easy to check that the marginal/spread densities used in this example do not admit triangle arbitrage (or, equivalently, spread option prices satisfy lower/upper Frechet bounds on copulas), but clearly exhibit the more general spread envelope arbitrage.

References

Adams, K. J., & van Deventer, D. R. (1994). Fitting yield curves and forward rate curves with maximum smoothness. *Journal of Fixed Income*, *4*, 52 – 62.
67

Andersen, L. B. (2005). Yield curve construction with tension splines. *Review of Derivatives Research*, *10*(3), 227–267.
61, 62, 63

Andersen, L. B., & Piterbarg, V. V. (2010). *Interest rate modeling, in three volumes*. Atlantic Financial Press.
37, 44, 45, 57, 59, 60, 62, 67, 98, 99, 111, 137, 147, 152

Breeden, D. T., & Litzenberger, R. H. (1978). Price of state-contingent claims implicit in option prices. *Journal of Business*, *51*(4), 621–651.
133

de Prado, M. L. (2018). *Advances in financial machine learning*. Wiley.

128

Fritsch, F. N., & Butland, J. (1984). A method for constructing local monotone piecewise cubic interpolants. *SIAM J. Sci. Comput.*, *2*(5), 300–304.
64

Gass, S. I. (2010). *Linear programming: Methods and applications* (5th ed.). Dover Publications.
157

Gâteaux, R. (1913). Sur les fonctionnelles continues et les fonctionnelles analytiques. *Comptes rendus de l'academie des sciences (Paris)*, *157*, 325–327.
97

Hagan, P. S., & West, G. (2004). *Interpolation methods for yield curve construction.* Retrieved from https://www.deriscope.com/docs/Hagan_West_curves_AMF.pdf (Working paper)
61, 64

Haghani, V., & White, J. (2023). *The missing billionaires: A guide to better financial decisions.* Wiley.
128

Kelly, J., John L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926.
131

Kvasov, B. (2000). *Methods of shape-preserving spline approximation.* World Scientific.
62

Lutz, M. (2015). Two collars and a free lunch. *SSRN eLibrary*.
139

McCloud, P. (2011). The CMS triangle arbitrage. *Risk*, *1*, 126–131.
154

Merton, R. C. (199l). *Continuous-time finance.* Blackwell Publishing Press.
129

Morton, A. (2022). *Inferring an expected rate and a term premium from long maturity swaps and options* [Working Paper]. (Citigroup working paper)
109, 112

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, *60*, 473-489.
74

Piterbarg, V. V. (2010). Funding beyond discounting: Collateral agreements and derivatives pricing. *Risk*, *2*, 97–102.
25

Piterbarg, V. V. (2011). Spread options, Farkas lemma and linear programming. *Risk*, *September*.
156

Piterbarg, V. V., & Renedo, M. A. (2006). Eurodollar futures convexity adjustments in stochastic volatility models. *Journal of Computational Finance*, *9*(3), 71–94.
111

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in c: The art of scientific computing*. Cambridge University Press. Hardcover.
58, 60, 61

Rebonato, R. (2018). *Bond pricing and yield curve modeling: A structural approach*. Cambridge University Press.
78, 81, 84, 86, 87, 88, 90, 91, 92

Renka, R. J. (1987). Interpolatory tension splines with automatic selection of tension factors. *SIAM Journal of Scientific and Statistical Computing*, *8*(3), 393-415.
64

Schweikert, D. G. (1966). An interpolating curve using a spline in tension. *Journal of Mathematics and Physics*, *45*, 312-317.
62

Soklakov, A. N. (2011). *Learning, investments and derivatives.* (https://arxiv.org/abs/1106.2882) doi: 10.48550/ARXIV.1106.2882
143

Tanggaard, C. (1997). Nonparametric smoothing of yield curves. *Review of Quantitative Finance and Accounting*, *9*, 251-267.
61, 67

Tee, C. W., & Kerkhof, J. (2019). A unified market model for swaptions and constant maturity swaps. *SSRN eLibrary*.
139

Theil, H. (1971). *Principles of econometrics*. Amsterdam: Wiley.
118, 123