Politecnico di Milano

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Mathematical Engineering

Reinforcement Learning For The Debt Value Adjustment Hedging

Relatore:
Prof. Marcello Restelli

Correlatori:
Prof. Carlo Sgarra
Dott. Matteo Pirotta

Tesi di Laurea Magistrale di:
Marco Giovanni Crotti
Matr. 837190

Anno Accademico 2015-2016

Abstract

Le condizioni dei mercati finanziari dopo la crisi economica e l'introduzione di nuovi standard internazionali nell'accounting hanno portato alla luce la necessità di considerare il rischio di credito e più genericamente il rischio di controparte nella valutazione dei derivati finanziari.

Se prima della crisi infatti veniva considerato nullo il rischio di default di una banca, a seguito del 2008 e in particolare a seguito della crisi dei debiti sovrani, si è percepito come questo potesse essere il driver di un rischio finanziario sistemico.

Gli istituti finanziari dunque non sono più solamente interessati a creare delle strategie di copertura contro un eventuale rischio di default della controparte, ma hanno la necessità di coprire anche il rischio derivante da cambiamenti nella proprià qualità di credito.

La scarsità di tecniche di trading indirizzate alla copertura dell'Own Credit Risk è il motivo alla base del seguente lavoro. Questa tesi infatti nasce da una collaborazione tra il Politecnico di Milano e Banca IMI, volta alla ricerca di tecniche di reinforcement learning che permettano al Credit Treasury Desk di automatizzare il processo di copertura. La presente tesi vuole quindi formalizzare il problema dell'hedging del Debt Value Adjustment. In particolare nella prima parte si analizza la questione dal punto di vista finanziario, analizzando i derivati utilizzati per la copertura e i modelli utilizzati per il pricing di questi ultimi. Nella seconda parte invece il problema viene formalizzato sotto una prospettiva basata sul reinforcement learning, e dunque più in generale basata sulle teorie del controllo. L'implementazione degli algoritmi per la rappresentazione dell'environment sono stati svolti utilizzando linguaggio C++ e sono parte integrante del seguente lavoro.

Chapter 1

Introduction

Market conditions after the economic crisis and the introduction of IFRS 13 Fair Value Measurement have brought to light the need to consider credit risk properly in the fair value of derivative contracts.

Questions and issues are likely to be raised in the future since new istituions are continuing to apply IFRS. In addition, various groups as the International Valuation Standards Council, are developing guidance in respect of credit and debit valuation adjustments. Until now, there is no specific guidance on processes used to calculate CVA and DVA, which creates challenges in evaluation.

After major bank defaults during the financial crisis, it was clear the need to take into account the counterparty credit risk into the valuation method. Thus, most market partecipants began to apply a CVA adjustment on their derivative assets. However, number of reasons have cited from financial institutions for not incorporating a DVA in their derivative liability position. First of all, observe a gain in profit or loss as their own creditworthiness deteriorates was considered counterintuitive and secondly there could have been an increase in systemic risk from hedging the DVA.

Anyway, IFRS 13 established that own credit risk must be contemplated into the fair value measurement of a derivative liability under the approach of an exit price. [1] The transfer notion is essential for measuring fair value, bacause it "capture market partecipants" expectations about uncertainty, but also the liquidity, and other associated factors.

IFRS 13 requires that techniques used for the valutation are built in such a way that they minimise the use of unobservable inputs and maximise the use of relevant

[1] "Entry price represents the perspective of buy-side: what a company would pay to acquire an asset or pay to settle a liability. Fair value was previously viewed as entry price. It is now synonym with an exit price (sell-side). Exit price reflects the standpoint of sell-side: what a company would receive if it were to sell the asset in the marketplace or paid if it were to transfer the liability."

observable inputs. This concept reflects the need that the fair value is a market-based measurement. Therefore, if available market-observable credit spreads are used to measure the fair value of an OTC derivative under IFRS 13.

1.1 Motivation and objectives

Machine learning and disciplines that are close to it have seen a great raise in interest. From the use of artificial neural networks in pattern recognition, to more recent methods such as Bayesian probability methods and reinforcement learning, machine learning techniques have been explored in a large number of applications, and finance is surely one of the fields where these techniques are applied.

In the quest for a trading algorithm, artificial intelligence methods have been emplyoed to construct systems that are better than traders in timing trade entry and exit opportunities. During last years, different attempts have been made in order to create consistently profitable system and ideas came from different fields ranging from fundamental analysis to machine learning. Only few of them were actually successful and the most promising could not be used to trade on market because of practical disadvateges. For example they led to large draw-down in profits and the trading strategy was characterized by excessive switching (long to short or viceversa) that implied huge transaction costs.

Since CVA and DVA are used for the valuation of an Over The Counter derivative, techniques in order to hedge them are now essentials. And if on the one hand there is an extensive literature on reinforcement learning techniques applied to financial problem as algorithmic trading, on the other hand there is no literature regarding the use of these techniques for hedging purposes.

Reinforcement Learning (RL) has essentially its roots in control theory, and as specified before it falls under the discipline of Machine Learning. The aim of Reinforcement Learning is to make a system learn a how to behave. The milestones of RL application are:

- The definition of the state space. The elements of the state space will be the variable on which the agent will take an action.
- The definition of a set of possible actions from each state.
- The definition of the impact that some actions can have on the systmes. In our specific case, the impact that our trading can influence the market.
- The choice of an algorithm that can learn an optimal policy, considering computational costs.

- The evaluation of the policy given by the previous point, by valuating its out-of-sample performance.

Despite the fact that Reinforcement Learning technique have been widely applied in different situations to solve also very complex problems, a practical application, like the one that is discussed on this thesis, requires a long process of experimentation that cannot be fully covered in a period of 6 months.

The present work, that is the result of an intese cooperation with BANCA IMI, therefore does not aim to cover all the elements of a RL Application, but it aims to set the basis of a project that will last years, formalizing essentially the whole problem that must be solved, and fixing the first bullet points of the list provided above. The coding part of the present work has been developed in C++.

Whereas actually the final aim of the whole project is to exploit reinforcement learning techniques for the DVA hedging.

1.2 Outline of Contents

The present work is structured as follows:

- In Chapter 2 will be introduced the concept of Counterparty Risk, considering how it is related with other financial risks. It will be also introduced the concept of credit exposure and expected exposure.

- In Chapter 3 some finance tools that will be used during the problem formalization are introduced. In particular the differences between real-world and risk-neutral probabilities are highlighted. Then the chapter deals with credit derivatives in particular focusing on the structure of Credit Default Swaps on single names, Contigent Credit Default Swaps and Credit Default Swap on Indices.

- In Chapter 4 the definition of Poisson Process is given, highlighting its main properties. Then the derivation of a deterministic intensity based model for single names is done. This model will be used many times during the present work, for different aims. At the end of the chapter is briefly and qualitatively described the Curve Mapping method for deriving credit spreads for those counterparties that are not traded on the market.

- In Chapter 5 the Counterparty Risk will be defined in quantitative terms, thus defining and discussing the Credit Value Adjustment and the Debt Value Adjustment. In particular it will be given a general definition of the two quantities also highlighting some hypothesis that can be taken in order to simplify their computation but also the description of their main properties.

- In Chapter 6 is introduced and formalized the core-problem of this thesis. In the first section is highlighted the concept of DVA hedging, and which instruments can be used for hedging purposes, but also what critical issues they bear. Then the variables involved in the hedging strategy and the hedging strategy itself are formalized. As last also a proper formalization of the objective is given.

- In Chapter 7 the mathemtical presentation of Sequential Decision Problem in provided. In particular some definitions are given in a rigorouw way in order to properly fix the basis of the mathematical framework in which the problem will be solved. In the second part of the chapter the same problem will be discussed under the Reinforcement Learning Perspective, focusing on Markov Decision Process, and deriving milestones results of the control theory as the Bellman Equation.

- In Chapter 8 some implementative aspects are outlined in order to give a semplified version of the problem for a first implementation. In particular some hypothesis are taken with respect to the general problem discussed in chapter 3, and the objectives are specified in view of the definitions in chapter 7.

- In Chapter 9 is described the algorithm called as 'Simulator'. The simulator is necessary to describe the environment in which the agent moves. The implementation of the simulator is a core-part of the present work, and it was done in C++. The implementation of the simulator requires a deep comprehension of the whole problem, because it merges notions of finance, and notions of reinforcement learning; Thus also the ability to switch perspective, from finance to RL or viceversa, was necessary.

- In Chapter 10 will be discussed two Reinforcement Algorithms respectively for the policy valuation and policy improvement. This chapter does not aim to discuss RL techniques that will actually be applied on this project, bacause this is not the objective of the present work; but for completeness, I think that they should be in this thesis too.

- In Chapter 11 I have done a summary of the present work, outlining the objectives reached, but also which aspects should be revised in order to have more performing algorithm in the "Learning Phase" of the project. In particular I gave an idea about how this project will develop, and hpw different techniques will be used in the light of the first results that we will obtain.

Chapter 2

Counterparty Risk

In this chapter, a broad overview of the counterparty risk will be provided. In particular in the first section the counterparty risk will be defined qualitatively underlining also how it is related with other financial risks. In the second section the counterparty risk will be further analyzed introducing the concept of credit exposure and expected exposure. These two quantities will be useful to define the credit value adjustment and the debt value adjustment.

2.1 Definition and Relations with other Financial Risks

Definition: Counterparty credit risk (or counterparty risk) is the risk that the counterpart (thus the entity that has entered with me into a financial contract) will not be able to fulfill its side of the contractual agreement because it defaults.

Counterparty risk is typically defined as arising from Over The Counter derivatives. Over The Counter (OTC) derivatives are usually more exotic, i.e. non-standards. They are traded between two parties and usually they are not protected by any government insurance programme. In other words, each party takes the counterparty risk deriving from the contract with the other party. Some parties in the market may have a deteriorated credit quality and are not even able to post collateral for reducing counterparty risk. Thus a general increase of interest around the counterparty risk raised as the Over The Counter derivatives market grew. Moreover, if before the financial crises banks and more in general institutions did not consider the counterparty risk with high quality (e.g., Triple A) rated institutions, after it, it was obvious how these entities were representing the most counterparty risk. Finally, also the regulatory pressure gave a boost to this interest.

This risk is obviously related to the financial risk's components:

- **Market Risk.** The Market risk is simply the risk deriving from the movement of market prices. Thus, it can be caused by some changes in underlying variables as stock prices, foreign exchange rates, credit spreads, interest rates or commodity prices, but it can be caused also by strong changes in market volatility. It has been widely studied, and it has driven the evolution of some mathematical models for assessing the risk. The Market risk can be trivially eliminated trough an offsetting contract, thus with a so called "back-to-back" position, and therefore assuming an opposite position in regards to the original opening position. But, if the offsetting contract is done with another counterpart, then a counterparty risk will arise. Indeed, the position is no longer neutral, considering that the two counterparties differ, and one of the two may fail. In conclusion, Market risk contributes to the counterparty risk.

- **Credit Risk.** The Credit risk is the risk that a debtor will not repay the debt or in general fulfill an obligation. Thus, it is strictly linked with the counterparty risk to default. So, it is important to study the probability to default during the lifetime of the contract. Under the Credit risk falls also the risk migration risk. Therefore a deterioration in credit quality, that will be reflected in a mark-to-market loss, must be taken into account. So as expected understanding the term structure of the counterparty's default probability is essential for analyzing the counterparty risk.

- **Liquidity Risk.** It can happen that a transaction cannot be executed for example because of the size of the position. But it can also happen to be not able to fulfill collateral requirements. All these fall under the concept of Liquidity risk. Therefore, reducing counterparty risk through collateralization or central clearing, is reflected in a liquidity risk growth.

- **Operation Risk.** The operation risk includes risks resulting from breakdowns in internal procedures, people and systems. The operation risk, therefore, is caused by human error, model risk (think about inaccurate models), failed processes, legal risk and fraud. Mechanism as the collateralization to mitigate counterparty risk give rise to operational risks.

Thus the counterparty risk is just a combination of market risk, which reflects the exposure, and credit risk, which reflects the counterparty credit quality. It is quite hard to assess in quantitative terms the counterparty risk, but the Credit Value Adjustment (that will be introduced later) puts a precise value on it and potentially allows it to be traded and also hedged (but as noticed before mitigating counterparty risk creates other financial risks).

Remark: : traditional credit risk is different from the counterparty risk, because:

- The value of the contract in the future is not deterministic but stochastic in most cases.
- The counterparty risk is undertaken by both the two parties, so it is bilateral and can be positive or negative.

The counterparty risk is reducible in different ways. Collateral agreements are commonly exploited at this aim. Usually they are bilateral agreement, thus they reduce the risk of both the two parties. As observed before, the collateral is able also to eliminate this risk, but at the same time it creates other risks as liquidity risk, but it also increases operational costs. The hedging of counterparty risk is a viable option, in particular after credit derivatives market grew. However hedging can be expensive. In general, the mitigation of counterparty risk is a double-edge sword, because it is not said, that it will completely eliminate it, and it can potentially allow financial markets to reach a dangerous size.

2.2 The Exposure and The Expected Exposure

In order to define properly the Credit Value Adjustment and the Debt Value Adjustment the definition of credit exposure must be given.

An institution may liquidate the contract and stop paying any future cash flow, if the counterparty has defaulted. Thus, the two counterparts must evaluate the net amount owing between them (also considering posted collateral).

The definition of credit exposure depends on whether the value of the contract is positive or negative. In particular:

- **Negative Value** As the value of the contract is negative for the institution, it is in debt with the counterpart, and therefore it is obliged to settle the amount. Therefore, the institution position is not changed, and neither losses nor gains are recorded from the counterparty's default.
- **Positive Value** In this case if the counterparty defaults, the institution will have a claim on the positive value of the contract, because the counterpart will not be able to fulfill the future payments. Just as bondholders, the institution expect to receive a fraction of the claim. Since it is unknown it is not considered in the definition of exposure.

Therefore, we define the exposure simply as:

$$Exposure = max(contract\, value, 0)$$

Thus, summing up, a counterparty's default, is reflected in an $Exposure$, which actual impact is a loss equal to :

$$R_c * Exposure$$

where R_c is the counterparty recovery rate.

As noticed also before, the counterparty risk is bilateral, thus it characterizes both two parties. Therefore, to be precise, we must take into account the losses derived by both two defaults.

The institution default will lead to a loss to all the counterparties it is in debt with. Therefore we must define also a negative exposure. The definition of negative exposure depends on whether the value of the contract is positive or negative. In particular by simmetry it must be that:

$$NegativeExposure = min(contract\, value, 0)$$

Therefore a negative exposure is reflected in a gain. This gain arises from the symmetry effect: where one party loss the other must gain. It seems to be also counterintuitive define a gain with the own default. This will be discussed deeper introducing the concept of Debt Value Adjustment.

As specified by the two formulas above, the valuation of the contracts at the default time is needed. This procedure is not easy, because it is essential to be in agreement with the counterpart.

In order to deeply understand the asymmetric risk profile that arises from the counterparty risk a comparison can be useful. Indeed, the fact that the institution loses if the value is positive and does not gain if the value is negative can be seen as short option position. Thus:

- As the exposure is comparable with a short option position, the volatility plays a key role.
- Compute the exposure can be hard, exactly as options pricing can be complex.

By symmetry, an institution has long optionality from its own default.

Observe that for risk management purposes, and in particular for the computation of the credit value adjustment and its symmetric part it is more important to define the exposure that there might be in the future, than the current exposure. If on the one hand the current exposure is certain, on the other hand the future exposure is stochastic, since for example it is influenced by future market movements that are obviously unknown. Valuating the future exposure therefore can be very complex, in particular if long periods are involved.

It is important to define the expected exposure and the expected negative exposure, also in order to give a proper definition of Credit Value Adjustment and Debt Value Adjustment. The expected exposure (EE) at a given future time is just the average of all exposures value. In line with the previous remarks negative values will give a zero contribution and will influence the exposure only through their probability, and only positive value will rise the expected exposure. By symmetry, the expected negative exposure at a given future time is just the average of all exposures value. In line with the previous remarks positive values will give a zero contribution and will influence the exposure only through their probability, and only negative value will rise the expected negative exposure.

Chapter 3

Default Probability and Credit Default Swap

In this chapter is discussed the concept of default probability, highlighting the differences between real-world probabilities and risk-neutral probabilities, and which one will be used for our purposes. Then it is introduced the concept of credit derivative and how the Credit Default Swap works.

3.1 Real-World Probability and Risk-Neutral Probability

As always in finance there is a huge difference between real-world parameter and the risk-neutral one: think for example to the real world rate of return for studying the asset evolution and the risk neutral rate of return used in the black schools framework. In particular:

- The **real-world parameters** should reflect the true value of a financial quantity.

- The **risk-neutral parameters** are directly derived from market prices, therefore they reflect the market perception of a financial quantity.

For our purpose in particular it is fundamental to understand the differences between risk-neutral default probability and the real- world default probability, indeed default probability is a key aspect for assessing and valuating the counterparty risk.

It is easy to understand how the two probabilities can be also significantly different through an example.

$\Downarrow$

Example: Consider a Zero Coupon Bond, that has a nominal value of 100€. Suppose that the Zero Coupon Bond was issued by a corporate having 3 % of probability to default. Supposing the absence of interest rates and recovery, then the price of the Zero Coupon should be 97 €. However, no one would invest, since there is the possibility to lose everything. Therefore we can suppose that in order to consider the uncertainty of the return there will be a default risk premium, and therefore a reduction in the bond price (suppose 3 € of reduction). Moreover, the investor can also worried about the liquidity of the bond, because it is not sure that he will be able to sell it in the future. Thus, another reduction (let's say of 1 is applied). Adding the default risk premium and the liquidity premium, the price of the Bond on the market would be 93 €, which would correspond to a default probability equal to 7%, that is not the actual default probability but rather constructed risk-neutral probability to make the numbers balance.

Summing up, we can say that the risk-neutral probability to default is derived from the market, and does not represent the actual (real-world) probability to default. Thus, real-world and risk-neutral default probabilities are not in conflict, but they just represent two different things.

It is also quite natural to understand how real world probabilities are in line with risk management purposes, while risk neutral probabilities should be used for hedging purposes. Thus, since the aim of this thesis is hedging the Debt Value Adjustment, risk-neutral probabilities will be used. Thus, also models to derive them from market prices are necessary.

In general the Risk-netrual probabilities are derived from credit spreads observed in the market. Different financial instruments bear with them different spreads. A common way to assess the credit spread is trough the premiums of single-name (referencing a single component such as corporate) Credit Default Swaps.

3.2 Credit Default Swap

In recent years there has been an exponential growth of the credit derivatives market. After the financial crisis, it was even more fundamental to transfer credit risk efficiently, therefore ad-hoc financial products were developed for investors. A credit derivative, is an agreement between two parties that is created in order to shift the credit risk, and its value is linked with the credit performance of a name or more names. Therefore we can breakdown them into single-name credit derivatives (referencing a single component such as corporate) and portfolio credit derivatives (referencing lot of components). Since credit derivatives are designed to shift credit risks, it is natural that they represents an opportunity for the trading, the hedging and the diversification of the counterparty risk.

The Credit Default Swap (CDS) was invented by Blythe Masters from JP Morgan in 1994. It is a financial swap agreement where the protection seller will compensate the buyer (usually the creditor of the reference loan) in the event of a loan default (by the debtor) or another credit event. In other words the seller of the CDS insures the buyer against a default event. The contract provides for the following cash flows:

- At the issue time, the protection buyer can enter at par. So no cash flows are provided.
- The buyer of the CDS makes a series of payments (the CDS spread) to the seller. Usually the buyer pays the CDS spread on a semester basis.
- The seller pays to the buyer a quantity of cash equal to the loss given default (of the reference loan) if the loan defaults.

If the CDS is issued at par, then the Net Present Value of the contract at the issue time, must be 0. So the following must hold:

$$s \sum_{i=1}^{N} (t_i - t_{i-1})\, \mathbb{E}_0[D(t_0, t_i)\, 1_{\{\tau > t_i\}}] - (1 - \pi) \sum_{i=1}^{N} e(0, t_i, t_{i-1}) = 0$$

where:

- s is the spread paid by the buyer to the seller.
- $1 - \pi$ is the loss given default of the reference loan.
- τ is the time to default of the reference loan.
- $e(t_1, t_2, t_3) = \mathbb{E}_{t_1}[D(t_1, t_3)\, (1_{\{\tau > t_2\}} - 1_{\{\tau > t_3\}})]$ and it is the value of a unitary cash in t_1 if the default occurs between t_2 and t_3.

Therefore, solving the above equation with respect to the spread s, we obtain the spread/premium that the buyer must pay, if the CDS is issued at par.

Remark: sometimes it can happen that the CDS spread is fixed in advance. Indeed, the standard is that a credit default swap built on investment-grade reference (thus with a credit quality superior than BBB) is traded with a fixed premium of 100 basis points, whereas a credit default swap built on reference with lower credit quality (and thus higher spread) is traded with a fixed premium of 500 basis points. In this case the CDS is also governed by an upfront, so that the NPV, considering the upfront, is equal to 0.

3.3 Contingent Credit Default Swap

As observed before the CDS built on a single-name provides protection to its buyer on a fixed notional amount. Thus, the notional amount is chosen by the buyer in order to cover its credit exposures arising from instruments such as bonds. For example, if an investor have invested 10 billion € on a corporate bond, then to hedge its position he will buy a CDS built on the same name with a 10 billion € notional. However, a characterizing aspect of the counterparty risk is that the actual loss at the credit event time is not known.

A Contingent Credit Default Swap is built to to overcome this problem. Indeed it works like a standard Credit Default Swap but at the same time the Notional Amount of protection is linked to another transaction. This transaction can be any cash flow linked to any financial product of any asset class. Therefore, a contingent credit default swap provides the perfect hedge against the counterparty risk arising from a derivative, just linking its notional amount with transactions provided by the same derivative. If on the one hand they represent a perfect hedge, on the other hand they are tailor-made products, therefore they are not flexible as standard CDSs. Although it has been discussed how contingent credit default swaps are able to eliminate counterparty risk, they are not popular in the financial sector.

The unpopularity of the CCDS arises from:

- **Documentation:** The contract of a CCDS is a an embedded termsheet since it must contains also information regarding the transaction to which is linked the national amount.
- **Privacy:** The counterparty of the CCDS will have also information regarding the investor's trades that have been hedged trough the CCDS.
- **No recognition of netting:** No recognition of netting: usually the CCDS is linked to a single transaction, not to a netting set, which would be useful. This because of the complexity that would arise in designing a CCDS covering different trades related to a netting set.
- **No recognition of collateral:** The CCDS does not account a potential collateralization of the credit exposure.

- **Credit Quality of the CCDS provider**:The CCDS seller, must have a high credit quality, possibly not correlated with the counterpart to which the notional amount is linked. So that, the investor must not incur in a double default. (the default of the original counterpart, and the default of the CCDS provider). For our purposes in particular, in order to hedge a large component of the DVA with such an entity, we should be very confident with the entity's ability to combat a high default rate situation.

Therefore, we will not use this tailor-made products four our hedging purposes, but in each case it was worth mentioning them.

3.4 CDS on Indices

In this section Credit Default Swap on indices will be introduced, and it will be also used in our DVA hedging strategy.

A Credit Default Swap in index can be seen as a convex combination on single-name Credit Default swaps, where each CDS has the same weight. Thus, also the spread in CDS on index, can be approximated just doing an average of the CDS premium within the index. Quite known credit indices are:

- **iTraxx Europe index.** It is also known as "The Main" is composed by 125 equally weighted European corporate investment-grade reference Entities (thus having a credit quality superior to BBB).
- **DJ CDX NA IG.** It composed by 125 equally weighted North American corporate investment-grade reference Entities (thus having a credit quality superior to BBB).

These two indices are the most common credit indices and therefore also the more liquid, but there are also other indices containing names on the same sector (e.g. the senior financial) or for example names operating in the same geographical region.

A typical technical feature of credit indices is the *roll*, which happen every 6 months. The roll consists of 3 operations:

- **Changing names within the index:** The aim of this operation is to keep an homogenous credit quality from series to series, replacing defaulted names and removing from the index other names because of credit events or migration events (thus rating downgrades), and therefore maintaining the same level of premium in the CDS on index.

- **An adjustment of the maturity:** In general the CDS on indices have a maturity of 5, 7 or 10 years. When the new series is issued the initial maturities are 5.25, 7.25 and 10.25 years. Hence, after 6 months they will be 4.75, 6,75 and 9.75, so another new series will be issued resetting the maturities to the original value.

- **Fix the index premium:** As already noticed standard CDS are not issued at par but it is with a fixed premium of 100 basis points or 500 basis points depending on the references' credit quality, and therefore the CDS is governed by an upfront. Hence, in the period before a roll, the premium is fixed at a level of 100 basis points or 500 basis points. Anyway, in general the fixed premium does not change from series to series also in order to reflect what mentioned in the 1st point, and therefore in order to maintain stable the credit quality of the index.

Chapter 4

The Poisson Process and the Jarrow & Turnbull Model

In this chapter is defined the Poisson Process and the Jarrow and Turnbull is introduced providing therefore a method to obtain risk-neautral default probabilities from CDS prices.

4.1 The Poisson Process

It is useful to recall the concept of Homogeneous Poisson Process and of Memoryless Random Variables:

Definition (Homogeneous Poisson Process): Let $(\tau_i)_{i\geq 1}$ be a sequence of random variables independent and identically distributed, such that $\tau_i \sim \varepsilon(\lambda)$. Define $T_n = \sum_{i=1}^{N} \tau_i$. Then the process $(N_t, t \geq 0)$ defined as:

$$N_t = \sum_{n\geq 1} 1_{\{t\geq T_n\}}$$

is said to be a Homogeneous Poisson Process with intensity λ.

These random variables $T_n = \sum_{i=1}^{N} \tau_i$ are called $jump - times$ and represent the times at which some repeating phenomenon that we will call *jump* occurs.

The Homogeneous Poisson Process is therefore a counting process, so it counts the number of random times T_n, and therefore of jumps, that occured between 0 and t. It is obvious that:

$$(T_n \leq t) = (N(t) \geq n)$$

In order to further analyze the properties of the Poisson Process, it is useful to introduce the concept of Memoryless Property for a random variable.

Definition (Memoryless Random Variables): a random variable X possesses the memoryless property if $\mathbb{P}(X > 0) = 1$ (i.e. X is a positive random variable), and for every $x \geq 0$ and $t \geq 0$,

$$\mathbb{P}(X > t + x) = \mathbb{P}(X > x)\, \mathbb{P}(X > t) \tag{4.1}$$

Observe that if X is an exponential random variable with intensity $\lambda > 0$, then: $\mathbb{P}(X > x) = e^{-\lambda x}$ for $x \geq 0$. X therefore satisfies the above equation for all $x \geq 0$, $t \geq 0$, so X is memoryless. Conversely, it is easy to prove that an arbitrary random variable X is memoryless only if it is exponential.

The memoryless property of exponential random variables is useful to find the distribution of the first jump in a Poisson Process after an arbitrary given time $t > 0$. In particular it can be shown that the first jump after $t > 0$ is indipendent of all jumps occured before (and also including) t. To be more precise, the following theorem holds.

Theorem 1. *Consider an Homogeneous Poisson Process with intensity λ and a given $t > 0$. Denote with W the random variable representing the time-interval from t until the first jump after t. Then W is a nonnegative random variable with the distribution function $1 - e^{-\lambda, w}$ for $w \geq 0$. Moreover W is independent of all the jump-times before time t.*

The theorem can be easily proved. Here, I give just an idea about the indipendence:

Proof. As regards the indipendence the basic idea behind the proof is to note that W conditional on the time of the last jump-time ϑ before t is just the remaining time until the next jump.

Considering now that the jump-times were defined as $T_n = \sum_{i=1}^{N} \tau_i$ and therefore as a sum of random variables independent and identically distributed, such that $\tau_i \sim \varepsilon(\lambda)$, then the interval time starting at ϑ is exponential and thus memoryless. Then W is indipendent of $\vartheta \leq t$ and of all earlier jump-times. □

Notice also that:

- for any interval of size t, $\lambda\, t$ is the expected number of arrivals in that interval.
- if we consider a time interval $\Delta\, t$ sufficiently small then the probability to have a jump in $(t, t + \Delta\, t)$ is $\lambda \Delta\, t$.

For our aims, we consider the time to default τ of a single name, as the first jump of a poisson process. Thus, we need to compute the survival probability $P(0,T)$, that is equal to $P(N_T = 0)$. In order to do this, divide the interval $[0,T]$ in N intervals of size $\Delta\, t = T/N$. Now, using the two properties above, it holds:

$$P(0,T) = \prod_{i=1}^{N} (1 - \lambda(t_i)\,\Delta\, t) = e^{\sum_{i=1}^{N} ln(1-\lambda(t_i)\,\Delta\, t)}$$

now considering $\lambda(t)$ sufficiently regular in t, and applying the first order taylor expansion, it holds:

$$e^{\sum_{i=1}^{N} ln(1-\lambda(t_i)\,\Delta\, t)} = e^{-\sum_{i=1}^{N},\lambda(t_i)\,\Delta\, t}$$

Therefore:

$$P(0,T) = e^{-\sum_{i=1}^{N},\lambda(t_i)\,\Delta\, t}$$

Computing the limit $\Delta\, t \to 0$, we get:

$$P(0,T) = e^{-\int_0^T \lambda(s)\,ds} \tag{4.2}$$

Remark: Since we have considered an Homogeneous Poisson Process, $\lambda(t)$ is actually constant, so sufficiently regular.

4.2 Jarrow and Turnbull Model

The Jarrow and Turnbull Model provides a method to obtain risk-neautral default probabilities from CDS prices. The assumptions underlying the Jarrow and Turnbull model are the following:

- It considers a CDS with a spread that is paid continuously.
- $\lambda(t)$ is constant in t.

With these hypothesis, a relation between the spread in CDS and the intesity λ is obtained. Indeed, as the NPV is 0, it must be that:

$$s\int_0^T dt\,\overline{B_0(0,t+dt)} = (1-\pi)\int_0^T e(0,t,t+dt)$$

where: $\overline{B_0(0,t+dt)} = \mathbb{E}_0[D(0,t_t+dt)\,1_{\{\tau>t+dt\}}]$.

Considering that:

$$e(0,t,t+dt) = \mathbb{E}_0[D(0,t+dt)\,(1_{\{\tau>t\}} - 1_{\{\tau>t+dt\}})]$$

and supposing that rates are indipendent of the default:

$$\mathbb{E}_0[D(0,t+dt)\,(1_{\{\tau>t\}}-1_{\{\tau>t+dt\}})] = B(0,t+dt)\,(P(0,T)-P(0,T+dt)) = B(0,t+dt)\,e^{-\lambda\,(t+dt)}\,(e^{\lambda\,d}$$

$$\Rightarrow e(0,t,t+dt) = B(0,t+dt)\,e^{-\lambda\,(t+dt)}\,(e^{\lambda\,dt}-1)$$

and just applying the first order Taylor expansion series of $e^{\lambda\,dt}$, we get:

$$B(0,t+dt)\,e^{-\lambda\,(t+dt)}\,(e^{\lambda\,dt}-1) = B(0,t+dt)\,e^{-\lambda\,(t+dt)}\,\lambda\,dt$$

thus:

$$e(0,t,t+dt) = B(0,t+dt)\,e^{-\lambda\,(t+dt)}\,\lambda\,dt$$

where:

- $B(t_0,t_1) = \mathbb{E}_{t_0}[D(t_0,t_1)]$ is the expected value of the stochastic discount, so it is the price of a default free zero coupon bond (it is also called "initial discount").

Thus, we get:

$$s\int_0^T dt\,\overline{B_0(0,t+dt)} = (1-\pi)\,\lambda\int_0^T B(0,t,t+dt)\,P(0,t+dt)\,dt$$

$$\Rightarrow s\int_0^T dt\,\overline{B_0(0,t+dt)} = (1-\pi)\,\lambda\int_0^T \overline{B_0(0,t+dt)}\,dt$$

thus it holds:

$$s = (1-\pi)\,\lambda$$

The equation above, put in relation a market quantity (the spread), with λ which is strictly related to the default.

Observe that we will use this model to obtain risk-neutral probabiilities that must be computed for the DVA hedging.

4.3 Curve Mapping

In the previous section we analyzed the Jarrow and Turnbull model for the quantification of risk-neutral default probabilities from the credit spread in CDSs. At the same time, for assessing the counterparty risk, and in particular in order to price the Credit Value Adjustment and the Debt Value Adjustment, and therefore also the strategies to hedge them, is important to obtain credit spreads for non-observable names. Thus, in this section, it will be given an idea about how deriving credit spreads for those counterparties which credit spread is not traded in the market.

The regulation defined by Basel III says: "Whenever such a CDS spread is not available, the bank must use a proxy spread that is appropriate based on the rating, industry and region of the counterparty."

The technique used to obtain a general curve based on observable credit spreads is called *creditcurvemapping*. The idea behind this method is to consider credit instruments referencing names in the same class, then use some relevant pillars, so some relevant maturities, and from these points obtain an entire curve, that in a certain way will also characterize names whose spread is not traded in the market.

The given class can be broad, for example we can derive a single curve that describe all names with the same rating, or can be more granular, for example we can derive a single curve that describe all names with the same rating, operating in the same sector and in the same geographical region. Obviously in the first case, there will be more data to fit but counterparties are maybe breakdown too broadly, while in second case the calibration is harder since there are less data.

Chapter 5

CVA and DVA

Previously we have analyzed the concepts of credit exposure and default probabilities, also providing a model to obtain risk-neutral default probabilities from credit spreads. In this section the two quantities will be combined for defining the counterparty risk in quantitative terms, thus we will define the Credit Value Adjustment.

5.1 The Credit Value Adjustment

The Credit Value Adjustment (CVA) is an adjustment to the fair value price of derivative instruments in order to consider Counterparty Credit Risk (CCR). Thus, CVA can easily be viewed as the price of the counterparty credit risk. This quantity obviously depends on the market risk factors that can influence derivatives' values (exposure) as well as on counterparty credit spreads.

The size of the credit risk depends on the exposure's size that you have with your counterparty, e.g., a corporate. If the derivative position with a corporate is in the money (so the derivative's value is positive) then it means that, given current market expectations of future market conditions, the future cashflows is likely to be valuable to you. But you could potentially lose the value of the derivative, if the counterparty were to default.

Suppose now that the derivative's value became negative to you and so you are out of the money on this contract, because the current markets rates and expectations of future market rates changed. Now, the counterparty defaulting would not be a true concern, indeed the credit risk is very little as you are not expected to receive money from the derivative. Obviously, as there is always the possibility that market factors change (in a such a way that the derivative becomes in the money again), there is always a credit risk until the contract's maturity.

The CVA can be computed as:

$$\mathsf{CVA}(t) = \mathbb{E}_t[\mathsf{LGD}_C \, 1_{\{\tau_C \leq T\}} \, 1_{\{\tau_C < \tau_I\}} \, D(t, \tau_C) \, (V_0(\tau_C))^+]. \tag{5.1}$$

where:

- LGD_C: is the counterparty Loss Given Default, so the share of an asset that is lost if the counterparty defaults. In particular the loss given default can be also written as $1 - \pi$, where π is the recovery rate, so the proportion of a bad debt that can be recovered. It can be very arduous to model/estimate the recovery rate and therefore the loss given default.

- τ_C: is the random variable representing the counterparty time to default. Usually a Poisson process is used to describe the default time of a company. The default time can be viewed as the first jump of a Poisson process. A Poisson process can be classified in different ways, based on the nature of the intensity function.

- τ_I: is the random variable representing the investor time to default. So, for our purposes, it will represent the Intesa San Paolo time to default.

- T: is the maturity of the contract.

- $D(t, .)$: is the risk free discount, used to discount future cash flows at a risk-free rate. In the formula, it is evaluated in τ_C.

- $(V_0(\tau_C))^+$: is the positive part of the derivative's value evaluated at the counterparty time to default. So the contribution given to the CVA by a negative exposure at a given time is equal to zero.

Observe also that:

- The indicator function $1_{\{\tau_C \leq T\}}$ incorporates only the counterparty time to default occuring before the contract's maturity. Indeed for the bank it does not matter if the counterparty will default after the conclusion of the contract.

- The indicator function $1_{\{\tau_C < \tau_I\}}$ incorporates only the counterparty time to default occuring before the bank time to default. Indeed for the bank it does not matter if the counterparty will default after itself.

Pricing the CVA and therefore assessing the counterparty risk in quantitative terms in not easy. In particular the difficulty arises from the bilateral nature of the Counterparty risk. If it is unilateral it is much easier. How the situation can get worse, from the computational point of view, with a bilateral risk, can be easily understood through an example.

Example: First, consider a bond. In this case to assess the risk we have to consider default in the discounting procedure and add the default payments. This is quite trivial.

If we consider another simple derivative but bilateral the quantification is really more difficult. For example, consider a swap, then to assess the counterparty risk, it must be taken into account that not all payments are at risks, because there is a partial cancellation with my own payments. Thus, the counterparty risk in a swap is much smaller thanks to this phenomena, but at the same time to determine which cash flows are at risks we need take into account lot of factors as forward rates or volatilities.

In order to greatly facilitate the initial exposition of the credit value adjustment and to easily explain the key feuatures of the CVA, we consider the following three assumptions:

- **The institution itself cannot default**. This assumption is equivalent to set $\tau_i = \inf$ and therefore (as it will be clearer later) to set the Debt Value Adjustment to 0.
- **The Risk-free valuation can be performed**,and therefore we suppose to be able to calculate the value of the contract ignoring the counterparty risk.
- **The independence between the credit exposure and default probability**.

After have computed the CVA, and considering $DVA = 0$, we can write:

$$V(t) = V_{RiskFree}(t) - CVA(t)$$

where $V_{RiskFree}(t)$ is the value of the derivative at time t, under the default-risk free assumption. Whereas $V(t)$ is the value of the derivative without this last assumption.

If on the one hand we have simply highlighted that the CVA is a negative correction to the contract value, on the other one this formula highlight also that the risk-free valuation and the computation of the CVA are two completely different problems.

Therefore, contracts and the related counterparty risk may then be priced separately. Thus, the CVA can be traded separately from the originating contract. This is also why in credit institutions one desk is responsible for risk-free valuation and one for assessing the counterparty risk.

Considering the formula (5.1) and the assumption $\tau_i = \infty$, it holds that:

$$CVA(t) = \mathbb{E}_t[\mathsf{LGD}_C\, 1_{\{\tau_C \leq T\}}\, D(t, \tau_C)\, (V_0(\tau_C))^+]$$

therefore considering that the probability to default is indipendent of the interest rates, we obtain:

$$\mathbb{E}_t[\mathsf{LGD}_C\, 1_{\{\tau_C \leq T\}}\, D(t, \tau_C)\, (V_0(\tau_C))^+] = \mathsf{LGD}_C \int_t^T B(t, s)\, EE(s)\, dPD(s)$$

where:

- EE(s) is the expected exposure at time s.
- PD(s) is the probability to default at time s.
- $B(t, s) = \mathbb{E}_t[D(t, s)]$ is the expected value of the stochastic discount, so it is the price of a default free zero coupon bond (it is also called "initial discount").

thus, we have:

$$CVA(t) = \mathsf{LGD}_C \int_t^T B(t, s)\, EE(s)\, dPD(s)$$

and supposing to know the PD in a finite number of interval $[t_{i-1}; t_i)$, we get:

$$\mathsf{LGD}_C \int_t^T B(t, s)\, EE(s)\, dPD(s) \approx \mathsf{LGD}_C \sum_{i=1}^{m} B(t, t_i)\, EE(t_i)\, PD(t_{i-1}, t_i)$$

$$\Downarrow$$

$$CVA(t) \approx \mathsf{LGD}_C \sum_{i=1}^{m} B(t, t_i)\, EE(t_i)\, PD(t_{i-1}, t_i)$$

that is surely easier to compute with respect to (5.1).

Remark: An increase in the credit spread is obviously reflected in a CVA increase, but the effect is not linear since probabilities are always lower than 1.

Remark: An increase in the recovery has a net impact on the CVA of the second order. Indeed, increasing the recovery will increase also the default probability as it can be noticed thanks to the relation obtained with the Jarrow and Turnbull model, but at the same time it reduces the resulting loss. The net impact will be negative on the CVA, because the risk-neutral probabilities increase sub-linearly, while the effect on the resulting loss is linear.

5.2 The Debt Value Adjustment

In the previous section the Credit Value adjustment was introduced, and in order to simplify the exposition the assumption that the institution itself could not default was taken. But as already said in the introduction the international accountancy standards let the institutions consider also their own default for valuation purposes.

Therefore, we can consider also the negative exposure that was defined in chapter two, and therefore we can consider this component for assessing the counterparty risk, generating the Debt Value Adjustment (DVA), which is a very controversial component.

CVA was considered by banks for assessing the counterparty risk arising from a contract with a corporate and therefore this charge fueled the tendency to have contract with high credit quality counterparty. Moving the perspective to the corporate point of view the CVA was not considered, indeed it was deemed impossible before the financial crisis that a bank could fail. After the financial crisis there is no more the concept of a default-free counterparty, and the Debt Value Adjustment arises from taking into account a bilateral Credit Value Adjustment.

Thus, defining the Bilateral Credit Value Adjustment (BCVA), the definition of the DVA comes by itself. The definition of BCVA is just as the definition of CVA taking into account that the institution itself can default. For a simpler exposition, we assume that the two defaults are independent.

Thus, we define the BCVA as:

$$\begin{aligned} BCVA(t) \approx \mathsf{LGD}_C \sum_{i=1}^{m} B(t,t_i)\, EE(t_i)\, PD_c(t_{i-1},t_i)\, [1 - PD_I(0,t_i-1)] \\ + \mathsf{LGD}_I \sum_{i=1}^{m} B(t,t_i)\, ENE(t_i)\, PD_I(t_{i-1},t_i)\, [1 - PD_c(0,t_i-1)] \end{aligned} \tag{5.2}$$

where the approximation (≈) is due the discretization of the time. (the precise definition would involve an integral from t to T, as we have done before) and where:

- $EE(t_i)$ is the expected exposure at time t_i.
- $ENE(t_i)$ is the expected negative exposure at time t_i.
- $1-PD_I(0, t_i-1)$ is the probability that the institution itself has not defaulted before t_{i-1}
- $1 - PD_c(0, t_i - 1)$ is the probability that the counterparty has not defaulted before t_{i-1}.
- $PD_c(t_{i-1}, t_i)$ is the probability that the counterparty defaults in the interval $[t_{i-1}, ti)$.
- $PD_I(t_{i-1}, t_i)$ is the probability that the institution itself defaults in the interval $[t_{i-1}, ti)$.

As we can notice the 1st part of the bilateral credit value adjustment is just the CVA introduced in the previous section. The second term is symmetric to the first, and gives a negative contribution to the BCVA since the Negative Expected Exposure is negative. A negative term is reflected in a gain, indeed if the institution itself defaults it will not pay all its negative exposure but just a fraction (the recovery) of it.

The Debt Value Adjustment is a controversial topic, indeed it is counterintuitive to have a gain at our own default. In particular using the Debt Value Adjustment implies that:

- A risk derivative can be evaluated more than a risk-free one. Indeed a great DVA corresponds to a great negative contributes to the BCVA, that can be negative if the 1st term reflecting the unilateral CVA is not that great. Thus, a negative BCVA is reflected in a greater value for the risk-derivative with respect to the same derivative but without considering the counterparty risk.
- Since everything is symmetric, if the all the parties agree for the calculation of the Bilateral Credit Value Adjustment, then the total amount of counterparty risk traded in the market is zero.

Now the general definition of Debt Value Adjustment will be given, obtaining something which is symmetric to the formula (5.1) for the CVA.

Debt Value Adjustment (DVA), similarly, is defined as the difference between the value of the derivative assuming the investor (the bank) is default-risk free and the value taking into account default risk of the investor. Changes in a bank's own credit risk that is reflected in changes in its credit spread therefore result in changes in the DVA component.

As for CVA, DVAs depend also on changes in all key factors that influence the expected exposures and not only to changes in the own creditworthiness (thus in credit spreads or probabilities of default).

The DVA can be computed as:

$$\mathsf{DVA}(t) = \mathbb{E}_t[\mathsf{LGD}_I \, 1_{\{\tau_I \leq T\}} \, 1_{\{\tau_I < \tau_C\}} \, D(t, \tau_I) \, (V_0(\tau_I))^-] \tag{5.3}$$

observe that now:

- The risk-free discount function is evaluated in τ_I.

- $V_0(\tau_I))^-$ is the negative part of the derivative's value evaluated at the investor time to default. So, as discussed so far, the contribution given to the DVA by a positive exposure at a given time is equal to zero.

- The indicator function $1_{\{\tau_I \leq T\}}$ incorporates only the investor/bank time to default occuring before the contract's maturity. This is coherent with the DVA definition.

- The indicator function $1_{\{\tau_I < \tau_C\}}$ incorporates only the investor/bank time to default occuring before the counterparty time to default. Thus, it is symmetric with what was appearing in the CVA formula.

Consider the counterparty and the investor/bank to be default-risk free. Let $V_{RiskFree}(t)$ be the value of the derivative at time t, under the default-risk free assumption. Then the value of the derivative without this last assumption $V(t)$, can be obtained using the CVA and DVA, indeed it holds that:

$$V(t) = V_{RiskFree}(t) - \mathsf{CVA}(t) + \mathsf{DVA}(t)$$

Therefore as noticed before the fair value price of the derivative instrument is adjusted in order to consider the counterparty credit risk (as a negative adjustment), and the bank own credit risk (as a positive adjustment).

Observe also that these two adjustments are going to 0 as the maturity of the contract is approaching, and they are exactly 0 at maturity.

Chapter 6

Problem Formalitazion

Hedge the DVA usually requires very complex hedging strategies. The goal of this project is to apply Reinforcement Learning techniques for the DVA hedging, transferring therefore the issue from the trader to an algorithm.

6.1 DVA Hedging

In order to hedge the CVA an institution should buy CDS built on the counterparty name, or shorting the counterparty bonds, therefore in general the institution have to short the counterparty credit. Thus hedging the CVA is theoretically simple. For DVA things are more complex.

In this first section we highlight the concept of DVA hedging, and which instruments can be used to hedge, but also what critical issues they bear.

As a preliminary case we consider the one-sided (unilateral) DVA, i.e. the case where the counterparty default-risk is negligible. Thus, in this particular case, we set:

$$\tau_C \equiv \infty \Rightarrow 1_{\{\tau_C \leq T\}} \equiv 0$$

Thus, also the Credit Value Adjustment is 0, and the formula for computing the DVA in this setting is:

$$\mathsf{DVA}(t) = \mathbb{E}_t[\mathsf{LGD}_I\, 1_{\{\tau_I \leq T\}}\, D(t, \tau_I)\, (V_0(\tau_I))^-] \tag{6.1}$$

Remark: The DVA adjustment as already remarked is going to 0 as the maturity of the contract is approaching. And is exactly 0 at maturity. In addition to mitigating the risk, an hedging strategy for the DVA must also soften this reduction.

In equation (3.1), is easy to see how the DVA variability depends on:

- credit risk component: changes in the own creditworthiness, reflected by $\mathbb{E}_t[1_{\{\tau_I \leq T\}}]$.
- market risk component: changes in the negative exposure, reflected by $\mathbb{E}_t[(V_0(\tau_I))^-]$.

remark: The two components above are not indipendent. Indeed, the negative exposure is evaluated in τ_I.

The second component, is easily manageble, building hedging portfolios with non-linear payoff instruments. Whereas, to hedge the first component is very arduous.

Indeed, to completely cover the credit risk component, and therefore to be protected by changes in the own creditworthiness, I should sell protection on my own name through a Credit Default Swap. This is obviously not possible, since no one would believe in my promise to pay them in case of my own default.

Alternative hedging instruments are:

- selling protection through CDS on correlated entities. A particularly liquid hedge is the CDS 5y iTraxx Senior Financial on-the-run series, made-up by 30 financial entities from the Markit iTraxx Europe index referencing senior debt. The drawback of this hedge, is that it implies to pay a protection as a member of the index defaults. Moreover this obligation is also what generates correlation with the CDS on our own name and therefore with our default probability. This "technique" maybe is the most common one, and is relatively successfull. It is obvious that the hedging will not be complete since no other CDS will be 100% correlated to the CDS on my own name.
- Buying bonds (or stocks) on our own name. This would be a buy-back operation, that cannot be done so easily or without constraints. Indeed, the buy back operation is usually done by institution when they have had a strong performance and are therefore cash rich. But typically a bank having an increase in its DVA is in the opposite situation, as they must have an increasing credit spread. Moreover, buy-back operations are also management tools to give signals to investors and therefore they cannot be done easily and quickly.
- Buying Government securities of the Italian Treasury, that are obviously correlated with the Intesa San Paolo credit risk. In particular the BTP future is very liquid. Also options on the BTP future are available.
- Buying futures and options on the Eurostoxx Banks(SX7E). In this particular case there is correlation as the corporate equity value should decrease as the default probability increase.

objective: The objective of the project is to apply reinforcemente learning techniques for:

- Finding out what is the optimal hedging allocation given the financial instruments to use.

- Generating reallocation signals, as the market prices move.

remark: The hedging strategy should maximize a profit, but at the same time it should minimize a risk measure, and the use of financial resources.

It necessary therefore to investigate how the hedging instruments influence the regulatory capital. Following the Basel III Framework, there are two different approaches for the CVA and the DVA:

- Banks are permitted to hedge their CVAs by entering into certain defined credit default swaps. To be precise, banks may enter into single name CDS, single name contingent CDS, other equivalent hedging instruments which reference the counterparty directly and index CDS. Other categories of counterparty risk hedges must not be incorporated within the CVA calculation and must be considered as any other instrument in the bank's inventory for regulatory capital purposes.

- The DVA exposure does not fall within the regulatory perimeter, therefore all hedging instruments must be treated as any other instrument for the calculation of the regulatory capital as if they were not used for hedging purposes.

Other constraints that could be considered are the transaction costs that should be kept lows taking under control the rebalancing frequency, and the hedging istruments' size standard.

6.2 Formalization of the variables involved in the hedging strategy

Variables involved in the definition of a DVA hedging strategy for Banca IMI are:

$\mathsf{DVA}(t)$ DVA in t

$\pi_I(t,T)$ Spread in the Intesa San Paolo CDS in t with maturity T; The CDS is listed on a daily basis for more or less 10 maturities T_j called pillar.

$s_I(t,T)$ Spread characterizing Intesa San Paolo bond prices in t with maturity T

$h(t)$ Spread in the CDS 5y iTraxx Senior Financial on-the-run series; The index is made-up by 30 financial entities from the Markit iTraxx Europe index referencing senior debt, and Intesa San Paolo is in the current series (S25) representing the 3.33% of the index.

$s_{\mathsf{BTP}}(t,T)$ Spread in the ”Buoni del Tesoro Pluriennali” (BTP) in t with maturity T; for $T = 10$ the BTP future, is particularly liquid.

$b(t)$ Price in t of the index capitalization-based Eurostoxx Banks (SX7E); Intesa San Paolo is an index member, with a weight equal to the 9.17% of the index's notional.

Some technical aspects for the hedging instruments must be introduced, so that they can be properly managed during the implementation:

- iTraxx FinSen: each CDSindex series has a fixed maturity. When a new series is generated it has a time to maturity of 5 years and 3 months, and it is ”on-the-run” for 6 months. For example the series S26 is generated the 20 September 2016 with maturity 20 December 2021 and will be ”on-the-run” until the 20 March 2017. Each year is characterized by 2 ”index roll date”: 20 September and 20 March. The new series can be slightly different from the previous one.

- BTP/Bund Future: Bund and BTP Futures are outlined by a liquidity concentrated on 4 maturities, typically on the eighth day of March, June, Septmber and December (”futures roll dates”). The closest of these 4 dates (”front contract”), is the one with more liquidity.

- SX7E index: The index cannot be traded directly; thus we will use futures on this index. As for the BTP futures,they are outlined by a liquidity concentrated on 4 maturities, typically on the third friday of March, June, Septmber and December (”futures roll dates”). The closest of these 4 dates (”front contract”), is the one with more liquidity.

The Intesa San Paolo CDS has the same roll dates as the Itraxx FinSen. When a new series is generated it has a time to maturity of 5 years and 3 months, and it is "on-the-run" for 6 months. Notice that, this convention have been used since December 2015; previously a new series was generated every 3 months.

Other relevant quantities that can be relevant are the DVA's sensitivities with respect to the own CDS pillar, i.e.:

$$0.0001 \cdot \frac{\partial \mathsf{DVA}(t)}{\partial \pi_I(t, T_j)} \tag{6.2}$$

DVA depends also on changes in all underlying factors $x_k(t)$ that influence the negative exposures and not only on changes in the own creditworthiness. A tipical example are interest rates in one or more currencies. The DVA has tipically non-trivial cross-gamma:

$$\frac{\partial^2 \mathsf{DVA}(t)}{\partial x_k(t)\, \partial \pi_I(t, T_j)} \tag{6.3}$$

making the rebalancing very frequent.

6.3 Formalization of the hedging strategy

In this section the hedging strategy will be defined. Suppose that the hedging portfolio is composed by K securities, and in these K securities there are also those that underlie the variables defined above. The securities have a price process X and a dividend process Y. The hedging strategy is represented by a $K+1$-dimensional process ψ_t.

The components ψ_t^k for $k = 2, \ldots, K$ represent the amount of the security k defining the hedging portfolio at time t, while ψ^0 represents the cash in the bank account, and ψ^1 represents the cash in the collateral account. We will denote with C the set of instruments that implies a variation margin exchange. The strategy is linked to a *gain process*:

$$\begin{aligned} G(t, T, \psi) &= \int_t^T \psi_u \, dX_u + \int_t^T \psi_u \, dY_u \\ &= \sum_{k=0}^{K} \left(\int_t^T \psi_u^k \, dX_u^k + \int_t^T \psi_u^k \, dY_u^k \right) \end{aligned} \tag{6.4}$$

In order to complete the problem formalitazion, we need to define the price process X and the dividend process Y for each security.

For $k > 1$ the processes can be obtained from the hedging instruments' typical features and from market prices. In particular for some instruments the price is obtained from a market quotation Q_t^k (for example the spread in CDS) through an evaluation function $X_t^k\,(Q_t^k)$. Since this is just the "kick-off" of a huge project, we began by using the 3 securities mentioned above, so we must define their price process and dividend process.

6.3.1 iTraxx FinSen: Price and Dividend Process

As mentioned before the CDS 5y iTraxx Senior Financial on-the-run series, is made-up by 30 financial entities from the Markit iTraxx Europe index referencing senior debt. In order to price the CDS from the spread (that is what actually we find on the market), some assumptions regarding the default intensity λ_i associated to the index are needed. In particular following the Jarrow and Turnbull model discussed above, we can suppose λ_i to be constant and set it to:

$$\lambda_i = \frac{Q_i}{60\%} \tag{6.5}$$

where 60% is the market-standard loss given default, and i is an index denoting the estimation day. Thus we can compute the survival probability on the interval [0,T], using the formula (4.2) introduced in section (4),obtaining:

$$S_i(t) = e^{-\lambda_i\,(T-t_i)} \tag{6.6}$$

remark: the index has a 1% fixed coupon, and the difference with the spread is regulated by upfront.

Let T_{ij} be the quarterly payment dates of the index fixed coupon after the time t_i, with $j = 2, \ldots, J$. For the counterparty receiving the spread (so for the protection seller), the approximate price is computed as discussed in chapter 3 for pricing a CDS. We obtain:

$$\begin{aligned} X_I = {} & 0.01\,[B(t_i,T_{i0})\,S_i(T_{i0})\,y(T_{i0},t_i) + \sum_{j=1}^{J} B(t_i,T_{ij})\,S_i(T_{ij})\,y(T_{ij},T_{i(j-1)})] \\ & - 0.6\sum_{j=0}^{J} B(t_i,T_{ij})\,(S_i(T_{i(j-1)}) - S_i(T_{ij})) \end{aligned} \tag{6.7}$$

where $y(t_1,t_2)$ is the fraction of year using the convention Act/360 between t_1 and t_2. The dividend process is:

$$Y_i = 0.01\,y(T_{(i-1)(-1)},T_{(i-1)0}) \qquad for\;\; t_i = T_{(i-1)0}$$

otherwise the dividend process is 0.

6.3.2 BTP and Bund Futures: Price and Dividend Process

For the BTP and Bund futures, the price process $X(Q)$ is exactly what we can find on the market. Thus $X_i(Q_i) = Q_i$. Moreover, the dividend process is trivial (so equal to 0), for both of the two instruments.

6.3.3 SX7E: Price and Dividend Process

Also for the future on this index the price process $X(Q)$ coincides with what we can find on the market. Thus $X_i(Q_i) = Q_i$. Moreover, the dividend process also for this future is equal to 0.

6.3.4 Collateral Account: Value and Dividend Process

The collateral account value denoted with ψ_t^1 can be computed as the sum of the collateralised hedging instruments' present values:

$$d\psi_t^1 = \sum_{k \in C} dX_t^k$$

where C is the set of instruments that implies a variation margin exchange (so the set of collateralised instruments). The dividend process can be easily described by the following ordinary differential equation:

$$dY_t^1 = r_t^1 \, dt$$

where r_t^1 is the collateral rate of return, that can be approximate with the risk free rate r^*.

6.3.5 Bank Account: Value and Dividend Process

The bank account value can be obtained using the autofinancing property of the strategy. Indeed, it holds that:

$$\psi_T \, X_T = \psi_t \, X_t + G(t, T, \psi)$$

Since there is a Debt Value Adjustment we cannot assume that the dividend process of the bank account is drifted by the risk free rate, it would be nonsense. It must therefore grow at a rate r_t^0 consistent with the Intesa San Paolo Unsecured financing rate. Moreover since the collateral account is interchangeable with the bank account, then only the **net cash position** between ψ_0 and ψ_1 is actually remunerated in the bank account dividend process. *Thus what we will actually indicate sometimes with ψ^0 with an abuse of notation will not be the bank account,* *but the net position on which the dividend is actually computed.*In particular during the implementation r^0 is set as follows:

$$r_i^0 = r^* + \pi_i^{1y}$$

where i is an index to denote the estimation day, and π_i^{1y} is the 1y-spread in the Intesa San Paolo CDS, and the correspondent dividend process is given by:

$$dY_t^0 = r_t^0 \, dt$$

To clarify the dynamic of the cash account an example can be useful.

Example: Suppose that at the time t_i the agent buys a CDS index contract paying the upfront X_i. This amount is taken from the bank account:

$$\psi_i^0 - \psi_{i-1}^0 = -X_i(*)$$

At the same time the agent will receive a collateral equal to X_i, thus:

$$\psi_i^1 - \psi_{i-1}^1 = -X_i(*)$$

Therefore there is no amount to be remunerated at trade time, since the net position between the two is 0. Thanks to the autofinancing relation at time t_{i+1} the bank account must remunarate the variation recorded in the collateral account, thus the following variation:

$$\psi_{i+1}^1 - \psi_i^1 = X_{i+1} - X_i^1$$

6.4 Formalization of the Objectives

The gain process can lay the base for what will represent the value function for our problem, so what the reinforcement learning algorithm aims to optimize. At the same time the trader must also be compensated for a gain obtained in shorter horizon, for example on a daily basis. Indeed, even if the trader is operating continually, there is a natural and relevant discretization of time: the one represented by the closing time of the market during working days.

Thus, we define:

$$\begin{aligned}\overline{G_i}(\psi) &= \int_{t_{i-1}}^{t_i} \psi_u \, dX_u + \int_{t_{i-1}}^{t_i} \psi_u \, dY_u \qquad (6.8)\\ &= \sum_{k=0}^{K} \left(\int_{t_{i-1}}^{t_i} \psi_u^k \, dX_u^k + \int_{t_{i-1}}^{t_i} \psi_u^k \, dY_u^k \right)\end{aligned}$$

Where the bar denotes that it is a quantity obtained on a daily interval, while the index i is introduced for indexing days. Therefore $\overline{G_i}(\psi)$ is the daily gain obtained thanks to the hedging strategy, and therefore thanks to the hedging portfolio.

So, denoting with DVA_i the value of DVA after the closing time of the market, then the total $P\&L$ (Profit and Loss) is:

$$U_i = DVA_i - DVA_{i-1} + \overline{G_i}(\psi)$$

Remark: Notice that the total $P\&L$ does not include also the gain/loss coming from the derivative generating the DVA. Indeed, as observed in chapter 5 the CVA desk is separate from other desks. So the CVA desk is responsible only for the quantity defined above.

Remark: Observe also that as the Debt Value Adjustment at the end of the period will be 0, thus DVA(T)=0 and at the end of the hedging strategy the total P&L over the whole period will be equal to the total gain process minus the staring CVA.

The Profit and Loss is also the quantity used in order to compute Risk Measures as the Daily value at Risk (daily VaR). These risk measures obviously define constraints on the whole position in terms of market risk. These risk measures lead to the quantification of a capital that the bank must hold in order to be able to face losses in a given time horizon. The capital can be either computed on merely economic consideration obtaining the economic capital eK, or computed following the Basilea Regulation obtaining the regulatory capital rK.

As already mentioned the DVA exposure does not fall within the regulatory perimeter, therefore all hedging instruments must be treated as any other instrument for the calculation of the regulatory capital as if they were not used for hedging purposes, whereas the DVA exposure fall within the economic capital perimeter, thus all hedging instruments must be treated considering their hedging purposes.

Suppose that both $eK(t)$ and $rK(t)$ can be computed starting from the sensitivity of the whole position with respect to market risk. Since investors are asking for a return on this capital, the remuneration can be seen as a cost for the hedging strategy. Thus we define the cost of the economic capital:

$$eKVA(t) = c \int_t^T D_{eK}(t,s)\, eK(s)\, ds$$

and the cost of the regulatory capital:

$$rKVA(t) = c \int_t^T D_{rK}(t,s)\, rK(s)\, ds$$

Where we assume that both the economic and regulatory capital are remunerated at a fixed rated c and $D_{eK}(t,s)$ and $D_{rK}(t,s)$ are appropriate discount functions. For example, we could define:

$$D_{\mathsf{eK}}(t,T) = D_{\mathsf{rK}}(t,T) = e^{-c\,(T-t)}$$

coherently, with the interpretation of the KVA as an expected value. Otherwise the discount factor can be modeled giving more importance to the future capital than to the present capital for the decision-making process of the algorithm.

So now we can define in general the objectives of the trader, that will be also the objectives of the learning agent. They are the following:

- Optimize market risk in view of the mean-variance P&L.
- The trader has an asymmetric preference with respect to U_i: it is particularly adverse to negative U_i and relatively tolerant to higher earnings.
- In the mean time the trader should try to minimize the cost of capital eKVA and the regulatory capital rKVA. (which obviously creates a direct constraint to the size of the hedging strategy).

Chapter 7

Sequential Decision Problem

Sequential decision models are just the mathematical representations of real problems where the decisions are taken in several stages, and during each stage a certain cost or reward must be paid or received. Each decision will also influence the circumstances at which next decisions will be taken. Thus, for example if the objective is to maximize the total reward, one must balance is desire to maximize the reward of the present decision against the desire to avoid future circumstances where low rewards are inevitable. In the financial sector is easy to find out a sequential decision problems, indeed all problems of portfolio management fall under this framework. The asset manager must balance his desire to achieve an immediate return, against his desire to avoid investments where a low long-run yield is probable.

In this chapter the concept of sequential decision problems is introduced. In the first section are given the main definitions also in a rigorous way that are useful to understand an optimal control problem. In the second section, the mathematical accuracy is sidelined to analyze the problem from the reinforcement learning perspective, thus from a more practical point of view.

7.1 Finite Horizon Optimal Control

First of all let's formalize the control problem.

- I denotes the state space. Its elements are the states that the system can reach during its evolution at discrete time moments.

- A is the space of control actions. Indeed the agent or controller, will take an action from A at each time step to modify the evolution of the system.

- S is the set of all possible values of the noise.

Hypothesis:

- Let (I,$\mathcal{I}$), (A, $\mathcal{A}$) and (S,$\mathcal{A}$) be measurable spaces.
- Let ϑ_n be a sequence of i.i.d. random variables, with values in S, defined on a probability space (Ω, $\mathcal{F}$, $\mathbb{P}$).
- Let $F : I \times A \times S \longrightarrow I$ be a measurable function.
- Let $r : I \longrightarrow [0, \infty]$ and $q : I \times A \longrightarrow [0, 1]$ be measurable functions.

In particular:

- r is the final cost/reward function, and $r(x)$ represents the cost/reward incurred when the final state of the system is x.
- q is the running cost/reward function, and $q(x, \alpha)$ is the cost/reward incurred when the system is in state x and the agent select the action α inside the set A.
- ϑ_n are random variables representing the noise acting on the system.
- The function F represents how the system evolves. Thus if the system at a given time is in the state x, the controller selects the action α, and in the subsequent time step the noise acting on the system takes the value s, then the next state will be: $F(x, \alpha, s)$.

Definition: A sequence $(\overline{\alpha_n})_{n\geq 0}$ of random variables with values in A is called *control*.

Defintion: The sequence $(X_n)_{n\geq 0}$ defined by:

$$X_{n+1} = F(X_n, \overline{\alpha_n}, \vartheta_n) \qquad X_0 = x$$

is called the trajectory starting at x and corresponding to the control $(\overline{\alpha_n})$. Obviously $(X_n)_{n\geq 0}$ is a sequence of random variables with values in I, and X_n represents the state at time n of the system and it depends on the values taken by the control variables and by the noise variables.

A particular interesting case is when the choice of control a time n action depends on the sequence of states up to time n.

Definition: An admissible strategy is a sequence $\alpha = (\alpha_n)_{n\geq 0}$ of measurable functions:

$$\alpha_0 : I \longrightarrow A$$

$$\alpha_1 : I \times I \longrightarrow A$$

$$\alpha_2 : I \times I \times I \longrightarrow A$$

$$....$$

$$....$$

$$\alpha_n : \underbrace{I \times .... \times I}_{n+1} \longrightarrow A$$

We will denote with $\mathcal{A}^{ad}$ the set of all admissible strategies.

Definition: The trajectory and the control associated to the strategy $(\alpha_n)_{n\geq 0} \in \mathcal{A}^{ad}$ and to the starting point x are the random variable sequences $(X_n)_{n\geq 0}$ and $(\overline{\alpha_n})$, defined by the equations:

$$\begin{cases} X_0 = x \\ \overline{\alpha_n} = \alpha_n(X_0, X_1, X_2, ..., X_n) \\ X_{n+1} = F(X_n, \overline{\alpha_n}, \vartheta_{n+1}) \end{cases} \tag{7.1}$$

It is obvious that both X_n and $\overline{\alpha_n}$ depends on α and on x.

Definition: Given a time horizon N, a starting point x and an admissible strategy α, then we define the payoff on the time horizon as:

$$J(x, \alpha) = \mathbb{E}[\sum_{n=0}^{N-1} q(X_n, \overline{\alpha_n}) + r(X_N)]$$

where X_n and $\overline{\alpha_n}$ are computed as specified in the previous definition.

Definition: Suppose that J is a gain, then we say that a strategy $\hat{\alpha}$ is optimal if:

$$J(x, \hat{\alpha}) \geq J(x, \alpha) \qquad \forall \alpha \in \mathcal{A}^{ad}$$

If J is a cost, then a strategy $\hat{\alpha}$ is optimal if:

$$J(x, \hat{\alpha}) \leq J(x, \alpha) \qquad \forall \alpha \in \mathcal{A}^{ad}$$

$$\Downarrow$$

The optimal control problem consists in understanding if there exists an optimal strategy, and in case of positive answer in assessing its properties and characterization.

Definition: Suppose that J is a gain, then we define the value function as:

$$V_0(x) = \sup_{\alpha \in \mathcal{A}^{ad}} J(x, \alpha)$$

whereas if J is a cost, the value function is defined as:

$$V_0(x) = \inf_{\alpha \in \mathcal{A}^{ad}} J(x, \alpha)$$

where V_0 can be either finite or ∞. In particular, given a starting point x it holds that:

$$\hat{\alpha} \;\; is \;\; optimal \quad \Leftrightarrow \quad V_0(x) = J(x, \hat{\alpha})$$

Remark: It can be also considered another kind of payoff J of the following nature:

$$J(x, \alpha) = \mathbb{E}[\sum_{n=0}^{N-1} \gamma^n \, q(X_n, \overline{\alpha_n}) + \gamma^N \, r(X_N)]$$

where $\gamma > 0$ is a given number, called $discount\, factor$. In general we have also that $\gamma < 1$ in order to consider very frequent situations where gains and costs delayed in time are reduced by a constant rate.

7.2 Reinforcement Learning Perspective

Surely the feeling that we learn since we were born from the environment is one of the first that comes thinking about the nature of learning. When a child plays does not have a teacher, but he does have a direct interaction with its environment. This interactions produce a set of information and a cause-effect knowledge that will lead to know what to do in order to achieve objectives. Learning from connections is the concept underlying all theories regarding intelligence, also from the philosophical point of view: just think about the $pars destruens$ and the $pars costruens$ discussed so far by the Father of Empiricism Francesco Bacone.

In the previous section, the optimal control problem was introduced in a very rigorous way. This was a necessary step in order to deeply understand the same problem from the reinforcement learning point of view. The reinforcement learning methods do not aim to assess if there exist an optimal strategy from the mathematical point of view, but these methods specify how the agent changes its policy as a result of its experience, and so they need data to be fed and they take account only about the realizations of the random variables introduced in previous section, regardless for example of how are distributed the noises, and if they are actually indipendent; thus, they exploit the data for those that are.

Since under this perspective we are not concerned about all technical aspects introduced before, we change also the notation making it more friendly for the reader.

More specifically, the agent and the environment interact over a discrete sequence of time steps: $t = 0, 1, 2, 3, ...$ For each time step the agent can observe the state s_t of the environment, $s_t \in \mathcal{S}$ where $\mathcal{S}$ is the set of all possible states. Depending on what it observes, the agent will select an action a_t that must belong to $\mathcal{A}(s_t)$, where $\mathcal{A}(s_t)$ is the set of actions that the agent can take at time t if it is in s_t. One step later, as a consequence also of its action, the agent receives a reward. At each time step, the agent observing its state will implement a probability distribution over the possible actions it can take. This mapping is exactly what we have called so far as strategy, that under the reinforcement learning perspective will be called *policy*. We denote the mapping as π_t, where $\pi_t(s, a)$ is the probability that $a_t = a$ if $s_t = s$.

Roughly speaking, actions are decisions that we want to learn how to make, while the states is what is observable and important for the agent in order to learn how to take proper actions.

The goal is to maximize the total reward it receives over the long run. To be precise the objective is to maximize the *expected return*, where if the rewards received after time step t are $r_{t+1}, r_{t+2}, ..., r_T$, then we want to maximize the expectation of:

$$R_t = r_{t+1} + r_{t+2} + ... + r_T$$

or equivalently of:

$$R_t = r_{t+1} + \gamma\, r_{t+2} + ... + \gamma^{N-1}\, r_T$$

if there is the discount factor γ.

Following this notation we will discuss the notion of Markov Property, that will lead to the definition of Markov Decision Process. This is essential, because the problem that we want to solve can be formalized as a Markov Decision Process.

7.3 Markov Property

As discussed so far in the previous section the agent makes its decision on the basis of the environment's state. At the same time we must know what is the proper information that the agent need to know in order to undertake an appropriate policy, and also what kind of information we should not expect it to provide.

Roughly speaking, we say that a state signal has the $MarkovProperty$ when it contains all the relevant information needed by the agent to undertake its action.

Example: Consider the current position, the velocity and acceleration of a ball. These three information summerize everything that is important to understand the sequence of positions that the ball will cover. It doesn't matter how these three came about. Lot of information are actually lost, but in the state there is all I need for understading the future.

This property sometimes is called as "$indipendence\,of\,path$", because knowing the current signal, the history is not taken into account.

We can also formalize this property very easily. In general when the Markov Property does not hold, in order to properly understand the dynamic we need to specify the complete probability distribution, i.e.:

$$\mathbb{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, ...r_1, s_0, a_0)$$

for all s',r and all possible values of the past events: $s_t, a_t, r_t, ..., r_1, s_0, a_0$, thus it is very arduous. But if the Markov Property holds, then:

$$\mathbb{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, ...r_1, s_0, a_0) = \mathbb{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t)$$

because the environment's response in $t+1$ does not depends on actions and states before time t, if we already know the state and action at time t. Since it is true the relation above it holds that:

Result: The best policy for choosing actions as a function of Markov state, coincides with the best policy for choosing actions as a function of the complete history.

It is obvious how the Markov Property is foundamental also from the computational point of view. Thanks to it indeed, we must not keep in memory the whole history, but we just need to know the current state in order to take an action. The reduction of the computation is essentially why, it is even better to approximate a state signal which is Non-Markov, with something that is Markovian.

7.4 Markov Decision Problem: MDP

If the Markov Property holds, then the reinforcement learning task is called a $Markov\,Decision\,Process$ or simply MDP. In particular, if the space of action $\mathcal{A}$, and the space of states $\mathcal{S}$ are finite, then we talk about $finite\,MDP$. The dynamics of a finite MDP can be easily described, through these two definitions:

Definition: We define the transition probabilities as:

$$\mathcal{P}^a_{s\,s'} = \mathbb{P}(s_{t+1} = s'|s_t, a_t)$$

thus they represents the probability for each $s' \in S$ to be the next state, given any a_t and s_t.

Definition: Given any next state s', and any current state s and action a, we define the expected value of the next reward as:

$$R^a_{s\,s'} = \mathbb{E}[r_{t+1}|s_t = s, a_t = a, s_{t+1} = s']$$

As already introduced in the first section in order to evaluate how much is good to be in a given state s under a given policy π, the valuation function $V^\pi(s)$ is used. In particular, it holds that:

$$V^\pi(s) = \mathbb{E}_\pi[R_t|s_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+1}|s_t = s]$$

where $\mathbb{E}_\pi[\quad]$ is the expected value supposing that the agent follows the policy π.

Another quantity that can be defined is the action-value function for policy π, which is denoted with Q^π and defined as:

$$Q^\pi(s,a) = \mathbb{E}_\pi[R_t|s_t = s, a_t = a] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+1}|s_t = s, a_t = a]$$

Thus, it represents the expected return starting from s, taking tha action a, and thereafter following the policy π.

Remark: Observing how Q^π and V^π are defined, it is obvious how they can be easily estimated using Montecarlo Methods.

Moreover, also for the implementation of any algorithm, it is important the following relation, also known as *Bellman Equation*:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'}\, [\mathcal{R}^a_{s\,s'} + \gamma\, V^\pi(s')]$$

indeed:

$$V^\pi(s) = \mathbb{E}_\pi[R_t|s_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+1}|s_t = s] = \mathbb{E}_\pi[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k\, r_{t+k+2}|s_t = s]$$

but:

$$\mathbb{E}_\pi[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k\, r_{t+k+2}|s_t = s] = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'}\, [\mathcal{R}^a_{s\,s'} + \gamma\, \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+2}|s_{t+1} = s']]$$

and:

$$\sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'}+\gamma\,\mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k\,r_{t+k+2}|s_{t+1}=s']] = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'}+\gamma\,V^\pi(s')]$$

Therefore taking the left-hand side of the first equation and the right hand side of the last one, we get:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'} + \gamma\,V^\pi(s')]$$

that is exactly the Bellman equation.

The Bellman equation expresses a mathematical relation between the value of a state, and the values of the next states.

Remark: The structure of the Bellman Equation suggests how fixed point methods can be used for the computation of V^π. In this case if $\gamma < 1$, then the convergence of a fixed point method can be easily proved using the contraction theorem.

As already remarked solving the reinforcement learning task is equivalent to find out a policy that gives an high total reward over the long run. Since we are considering a finite MDP an optimal policy exists. Thus, we can find at least one policy π^*, such that: $V^{\pi^*}(s) \geq V^\pi \quad \forall s \;\; and \;\; \forall \pi$. Thus, we can define the optimal-state value function V^* as:

$$V^*(s) = \max_\pi \; V^\pi(s) \quad \forall s$$

similarly we define:

$$Q^*(s,a) = \max_\pi \; Q^\pi(s,a) \quad \forall s \in \mathcal{S} \;\; and \;\; \forall a \in \mathcal{A}$$

Since by the definition this function gives the expected return under the hyphotesis that the agent is in state s, it takes an action a and thereafter it follows an optimal policy, we can write it in terms of V^*, indeed:

$$Q^*(s,a) = \mathbb{E}[r_{t+1} + \gamma\,V^*(s_{t+1})|s_t = s, a_t = a]$$

and from this formula it is also obvious how the policy that gives $Q*$ is the same that gives $V*$.

Another result that is worth to mention is the Bellman equation for $V*$, also known as Bellman optimality equation. It holds that:

$$V^*(s) = \max_a \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'} + \gamma\, V^\pi(s')]$$

Indeed:

$$V^*(s) = \max_a Q^{\pi^*}(s,a) = \max_a \ \ \mathbb{E}_{\pi^*}[R_t|s_t = s, a_t = a] = \max_a \ \ \mathbb{E}_{\pi^*}[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+1}|s_t = s, a_t = a]$$

but:

$$\max_a \ \ \mathbb{E}_{\pi^*}[\sum_{k=0}^{\infty} \gamma^k\, r_{t+k+1}|s_t = s, a_t = a] = \max_a \ \ \mathbb{E}_{\pi^*}[r_{t+1}{+}\gamma \sum_{k=0}^{\infty} \gamma^k\, r_{t+k+2}|s_t = s, a_t = a]$$

and:

$$\max_a \ \ \mathbb{E}_{\pi^*}[r_{t+1}{+}\gamma \sum_{k=0}^{\infty} \gamma^k\, r_{t+k+2}|s_t = s, a_t = a] = \max_a \ \ \mathbb{E}[r_{t+1}{+}\gamma\, V^*(s_{t+1})|s_t = s, a_t = a]$$

and:

$$\max_a \ \ \mathbb{E}[r_{t+1} + \gamma\, V^*(s_{t+1})|s_t = s, a_t = a] = \max_a \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'} + \gamma\, V^*(s')]$$

Therefore taking the left-hand side of the first equation and the right hand side of the last one, we get:

$$V^*(s) = \max_a \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'} + \gamma\, V^*(s')]$$

that is exactly the equation mentioned above.
Similarly it holds that:

$$Q^*(s,a) = \sum_{s'} \mathcal{P}^a_{s\,s'}\,[\mathcal{R}^a_{s\,s'} + \gamma \max_{a'}\ V^*(s')]$$

Remark: For a finite MDP, the Bellman optimality equation has a unique solution indipendent of policy.

Chapter 8

Simplified Version Of The Problem For A First Implementation

In order to simplify the problem for the first implementation, we have considered the DVA generated by the liability represented by a single cash flow N that the bank must pay at time T.

In this way, we focused on which is the best hedging strategy to adopt against the own credit spread risk, without considering the complexity introduced by the contingent nature of the exposures.

As regard the value of T three different alternatives can be considered:

- 5y maturity rolling each 6 months. So the maturity of the liability would coincide exactly with the maturity of the CDS itraxx 5y.
- 5y "constant maturity". The DVA has each day maturity of 5 years from the reference date.
- The expiry date is fixed, and therefore will not roll or be "updated".

The first alternative was implemented because more in line with the problem that the bank want to analyze.

It must be noticed also that the close-out amount of the cash flow is its value discounted at default time:

$$V_0(\tau_I) = N\, D(\tau_I, T)$$

Thus the DVA can be computed as:

$$DVA(t) = N\, LGD_I\, \mathbb{E}_t[1_{\{\tau_I \leq T\}}\, D(t,\tau_I)\, D(\tau_I, T)]$$

$$\Downarrow$$

$$DVA(t) = N\, LGD_I\, B(t,T)\,(1 - S(T)) \tag{8.1}$$

where $S(T)$ is the survival probability evaluated in T.

Remark: The discount factors are computed as:

$$B(t,T) = e^{-r^*\,(T-t)}$$

This approximation is very useful because it excludes the interest rate as a possible risk factor. For the first implementation it is set to 0.

Remark: In order to obtain the survival probability to compute the DVA, it was used the Jarrow and Turnbull model, exactly as it was done for the computation of the CDS Itraxx upfront, but starting from the CDS 5y spread on Intesa San Paolo. Thus, after have obtained the intensity rate $\lambda_i(T)$ from the spread, we get:

$$S_i(T) = e^{-\lambda_i(T)\,(T-t)}$$

For the hedging strategies the 3 instruments mentioned in the previous section are used. Thus:

- 5y iTraxx Financial Senior (FinSen) CDS index.
- Purchase of 10y BTP futures and simultaneous sale of 10y Bund Futures, i.e. BTP Spread Trade.
- Futures on the Eurostoxx Banks SX7E.

8.1 Objectives of the strategy

Also in this simplified version the first two objecives are in line with the objectives outlined in the section (6.4). Thus, we want to optimize market risk in view of the Mean-Variance $P\&L$ on the fixed horizon, considering that on the total daily $P\&L$ denoted as U_i, there is an asymmetric preference.

Remark: It is important to remember that the DVA is computed on a rolling expiry. Thus, we need to fix a time horizon on which we want to do the Mean-Variance analysis.

Observe that our ultimate objective therefore does not consist in developing a investment strategy which aim is to maximize the expected total $P\&L$ over a given time horizon.

This reflects the interest of the investor in minimizing also some form of *risk* of the policy. Typical performance criteria that are considered to this purpose involves the variance of the cumulative reward. Thus, if J is the cumulative expected reward, and V the variance of the cumulative reward, then the performance can be:

- Maximize J such that $V \leq c$
- Minimize V such that $J \geq c$
- Maximize the Sharpe Ratio: $J/\sqrt{V}$
- Maximize $J - c\sqrt{V}$

Remark: Observe that the problem of maximizing a return under a risk measure has been widely studied. In particular in 1952 Harry Markowitz studied the problem of maximizing the return supposing to be under a determined value of Variance, and therefore under a determined risk-profile. In this way a curve called efficient frontier is obtained, and it represents the maximum expected return that I can get under a specified profile allocating the instruments in the portfolio in a proper way (optimal way actually). Observe that if on the one hand the Markowitz problem has an objective that is close to our objective, on the other hand the initial framework is completely different. Indeed in the Markowitz problem the idea is to maximize the return of a linear portfolio, thus composed essentially by stocks and eventually by a risk-free asset. Our portfolio is really more complex, and there is also and uncontrollable variable, the DVA, to take into account in the reward.

Surely the performance measure that is most used in practice is the Sharp Ratio. However, in the first implementation we will not consider the measures above, but we will follow the following procedure:

- A utility function $f(u)$ is introduced to consider the asymmetric preference. In particular the function must be concave for positive values of U_i (so for profits), and must be convex for negative values of U_i (so for losses).
- Then what we will actually maximize is not the sum of U_i but the sum of $f(U_i)$.

As function f, the Kahneman and Tverky curve can be used, but considering a great slope for losses. Since the slope of the curve is quite great, even a little loss is very penalized. This framework aims also to keep the variance low. Indeed if little losses lead to a great penalization the agent will be more prudent, reducing therefore the variance of its actions and as consequence the variance of the rewards.

For simplicity we substituted the third objective, which was based on KVA measures with a constraint on the position' sign of the instruments inside the hedging portfolio. In particular we imposed that the positions on the hedging instruments have a sign that is in the direction of mitigating the own credit risk. In particular the portfolio must be:

- Short on the CDS FinSen. Since I cannot be short on the CDS on my own name, I need to be short on CDS built on names positively correlated with it.
- Long on BTP Futures. Thus, short on BTP Spread. Indeed if be Long on this instruments,
- long on the SX7E index futures.

Indeed if the DVA increases in one time step, the contribution given by it to the total daily gain U_i is positive. But if the DVA increases then it means that there has been an increase in the own credit spread. An increase in the own credit spread is reflected also in:

- An increase in the spread in the CDS FinSen, because they are positively correlated. Thus the CDS FinSec Upfront increases too and so if the portfolio position is negative on this instrument, the contribution given to the total daily gain is negative and therefore opposite to the contribution given by the variation of DVA.
- An increase in the spread of the BTP bonds underlying the futures. And therefore also an increase in the BTP/BUND spread. So, I need to be short of spread, if we want to go in the opposite direction of the DVA.
- A decrease in the value of futures SX7E, and therefore to go in the opposite direction of the DVA we need to be long on these instruments.

8.1.1 Value Function

For the value function we consider a time horizon of 3 months, that in the code is approximate with 100 days.

8.1.2 Rebalancing frequency

In the first implementation, a single daily rebalancing is assumed, even if at the end of the project the aim is to generate also intraday signals. In other words, the time is discretized in business days t_i. The interval time between business days is not constant, and for convention we have considered the market condition at close-time. Similarly discretized quantities ψ_i^k, X_i^k, Y_i^k are considered.

Thus, under this framework the gain process between t_{i-1} and t_i is:

$$G_i = \sum_{k=0}^{K}(\psi_{i-1}^k[(X_i^k - X_{i-1}^k) + (Y_i^k - Y_{i-1}^k)])$$

Chapter 9

The Simulator

In this section, the algorithm called as ‘simulator’ is described.

In reinforcement learning problems, the simulator is necessary to describe the environment in which the agent moves. As discussed before reinforcement learning is the art of learning from interaction with an environment, so defining the environment is a crucial aspect. As before, we will call the learner and decision-maker the *agent*, while the thing it interact with, is called the *environment*.

The idea is that the agent continually interact with the environment. Broadly we can say that the agent selects an action, the environment responds to this action presenting a new situation to the agent. Thus the agent is the controller, the environment is the controlled system, and the action is the control signal. The environment also gives rise to rewards, that the agent want to maximize over the time.

To a very high-level, we formalize it, considering a discrete sequence of time steps: $t = 0, 1, 2, 3....$ At each time step, the agent has a full representation of the environment's state $s_t \in S$, where S is a set of possible state, and considering this state, it will take a decision; thus it will select an action $a_t \in A(s_t)$, where $A(s_t)$ is the set of action available in state S_t. One time step later the agent receives a reward r_{t+1} and it will be also in a new state.
Thus, as introduced before, the simulator aims to describe the environment, therefore it is an algorithm that given the state and given the action, it will move the agent to another state compensating it with a reward.

Let's analyze the simulator in our specific case, by delineating the features that appear also in the C++ code. In our specific case the trader is the agent, that have to take a decision regarding the positions on its hedging portfolio, in particular as already discussed in the previous section the trader wants to maximize a value function over a time horizon of 3 months. Therefore given a starting time, the simulator will move the agent from state to state for 3 months, compensating him

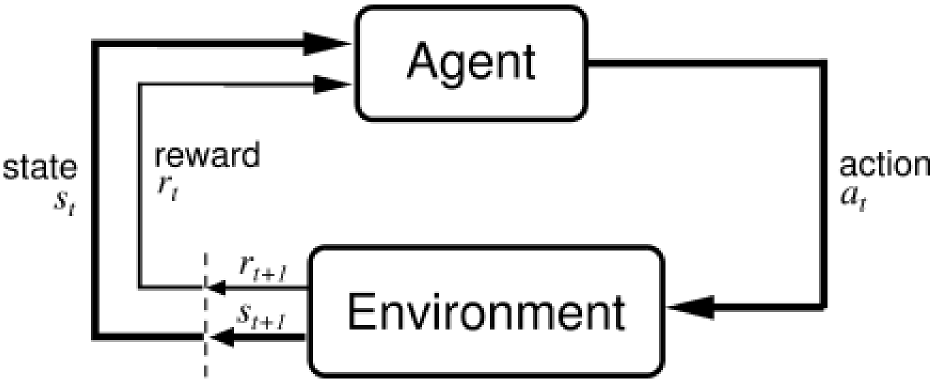

Figure 9.1: Interaction between the agent and the environment

with a reward after every transaction; this simulation over the 3 months will be called from now on: *episode*. Obviously we will run a great number of episodes, starting from a different time step and therefore also a different state, in order to properly train the agent.

First of all, we introduced a discrete sequence of time steps, where each time step represents a trading day. Indeed, as mentioned previously intraday trading is not considered in the first phase of the project. For each time step inside the dataset we can find:

- The reference date.
- The price of the future on SX7E index, considering all the technicalities involved to manage the roll.
- The price of the future on BTP and BUND, considering all the technicalities involved to manage the roll.
- The CDS 5y iTraxx Senior Financial upfront, obtained through the *J&T* model, and the correspondent dividend.
- The 1y spread in Intesa CDS, useful to define the interest rate on the bank account.
- The DVA.

9.1 The State

In order to take a proper action the agent must know its current state, that should contain all the observable information that can provide help to the agent in order to take its action. It is therefore of primary importance to find out which processes must be considered so that the trader can capture the expectations of the market on the hedging instruments.

Remark: The state signal that will be introduced is actually markovian. At the same time it does not fully represent the environment. This state is just a first trial to see if the algorithm just having this description of the environment is able to learn. It is obvious that a development of the state must be done, for example considering the sensitivies discussed on the formalization of the problem, which actually are essential for hedging problems.

9.1.1 Process for the BTP/BUND future

As regard the future on the BTP/BUND the price process X defined in (6.3.2) is sufficient in order to capture the variation that the trader need to know in order to fully understand the expectations of the market on it, because it does not provide any dividend or coupon. So the process V that we will consider in the state is exactly X.

9.1.2 Process for the SX7E Future

Since it is a future, the same reasoning used for BTP/BUND futures can be used. Thus also for SX7E future the process $V \equiv X$.

9.1.3 Process for the CDS itraxx

As regard the CDS itraxx, the situation is more delicate. Indeed the process X is not sufficient to describe completely what the agent needs to know in order to undertake an appropriate decision.

If I would consider only the process X, then the day before the dividend payment, the price of the CDS would be given by the following formula:

$$\begin{aligned} X_i(Q) = 0.01\,[B(t_i, T_{i0})\,S_i(T_{i0})\,y(T_{i0}, t_i) + \sum_{j=1}^{J} B(t_i, T_{ij})\,S_i(T_{ij})\,y(T_{ij}, T_{i(j-1)})] \\ - 0.6 \sum_{j=0}^{J} B(t_i, T_{ij})\,(S_i(T_{i(j-1)}) - S_i(T_{ij})) \end{aligned} \tag{9.1}$$

where all the terms in the formula were specified in the section (6.3.1).

The day after, when the dividend is paid, then the upfront (or price) is given by:

$$X_i(Q) = 0.01\,[\sum_{j=1}^{J} B(t_i, T_{ij})\, S_i(T_{ij})\, y(T_{ij}, T_{i(j-1)})] - 0.6 \sum_{j=0}^{J} B(t_i, T_{ij})\, (S_i(T_{i(j-1)}) - S_i(T_{ij})) \tag{9.2}$$

Where therefore the contribute given by the first term does not appear anymore, because it was the contribute given by the paid coupon. So the agent looking only at X would see a decrease on the value of the contract, without considering that it would have also received the premium spread.

So an approach that can be considered is to take into account also the contribute given by dividend in the day that it is paid. Thus, the process V regarding the CDS itraxx can be simply the sum of X and Y, where X is defined as above and Y is:

$$Y_i = 0.01\, y(T_{(i-1)(-1)}, T_{(i-1)0}) \qquad for\ \ t_i = T_{(i-1)0}$$

So, considering the process $X + Y$ the day before the payment of the dividend we just would have exactly what is in formula (9.2), and the day that the dividend is paid the contribute given by the first term in the formula is absorbed by Y, so as aimed I will only capture the net variation of the price, without changing the datum because of the dividend.

But what will happen the day after the dividend payment? Simply there will be a drawdown of the price due to the fact that a portion of the price was absorbed by the dividend. This drawdown therefore is because of the nature of the contract and not on the some changes in the market perception of the contract. This situation must be avoided, because the agent is not able to understand the contract nature, and thus its action must be influenced only by true changes in the market perception of the contract.

The two situations above can be easily described through an example.

Example: Suppose that the value of the CDS itraxx is 100 €. And that the market does not change during this period. Thus, since the perception of the market over this contract is the same, our agent should not see any jump in the process considered for its decision, even if the dividend of 1 € is paid. But a jump is actually recorded, with the two processes previously discussed, indeed:

Process	Before dividend	Dividend day	After dividend
X	100	99	99
X+Y	100	99+1=100	99

Roughly speaking, the agent must not see any jump because of the dividends. Jumps can happen only if the market perception on the contract is changed. Therefore what can be considered as process for managing the CDS itraxx is the process given by:

$$V_i = X_i + \sum_{j=0}^{i} Y_i$$

In partular:

- j=0 is the index representing the issuing date of the CDS. Thus when the CDS "was born".
- the process V_i consider therefore all the dividend paid before i (including i). Thus, it is like considering the value of the contract in the case that no dividends are paid.

In this way, the agent is actually able to capture the market perception on the contract without being hampered by the nature of the contract itself. Following the previous example in this case we would get:

Process	Before dividend	Dividend day	After dividend
V	100	99+1=100	99+1

that is actually what we want.

Remark: In this way in the state we lose the information of the dividend. But this is not a true problem. Indeed it is coherent with the reward, where we do not consider "owning cash" to be better than "owning a financial instrument" with the same value.

9.2 Information inside the state

The state in this particular problem is composed by:

- Prices:
  - The prices of the future on SX7E index on a given window.

  - The spreads on BTP-BUND future on a given window.
  - The values of the process V for the CDS Itraxx introduced in the previous section on a given window.
  - The spreads on CDS 1y Intesa San Paolo on a given window.

- Variation on the Allocation:

  - The difference between the number of futures on SX7E owned in t and in $t-1$. These differences are inside the state on a given window.
  - The difference between "the number of spread" owned in t and in $t-1$. These differences are inside the state on a given window.
  - The difference between the number of futures on CDS 5y iTraxx Senior Financial owned in t and in $t-1$. These differences are inside the state on a given window.
  - The difference between the bank account in t and in $t-1$. These differences are inside the state on a given window.

- Total Allocation of the 3 hedging instruments and of the bank account in the last day.

Remark: It must be observed that inside the state there are prices on a given window, thus also in preceding days with respect to today. This because the agent in order to take a proper action will consider the evolution of the prices in the preceding days, and not just the last level of prices.

9.3 The action

It is not the purpose of the simulator to choose which type of action the agent have to undertake. At the same time the simulator will use these actions, in order to move the agent from the current state to the following one, and to compute the reward. In order to define the set $A(s_t)$ of action available in state S_t, we need to understand:

- How the hedging instruments are tradeable. It is important to understand therefore number of underlying that there are in a future contract. In general the value of the single contract is indeed the number of underlying multiplied by the price of the underlying.

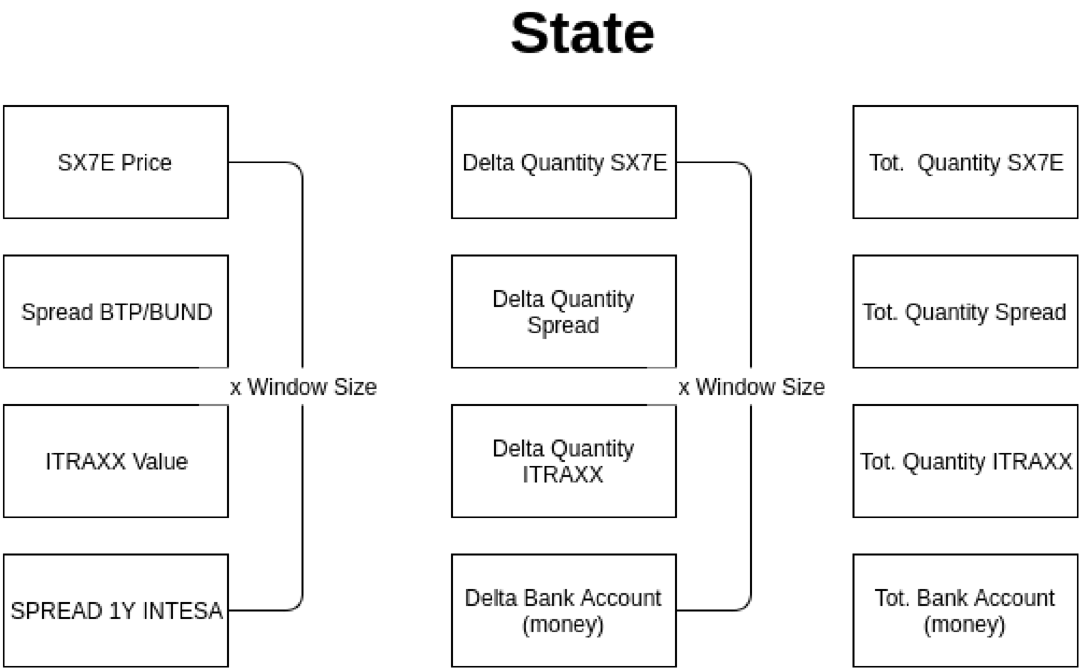

Figure 9.2: Representation of the State Vector

- What is the magnitude of the Debt Value Adjustment that we want to hedge. Indeed, we need to buy/sell amounts of the 3 instruments in such a way that the Gain given by the portfolio is comparable in terms of magnitude to the daily DVA variation.

- What is the amount that I have to buy/sell for each hedging instrument, so that as said before the total daily gain is comparable to the daily variation of the DVA, but also so that the contribute from each single instrument is comparable in terms of magnitude with the contribute given by the others.

All these information are useful to describe $A(s_t)$, and therefore to understand which is the space of action that the agent can cover given a particular state. This is also done in order to have a reduction of the space of action, thus the algorithm will not explore actions that a priori we know are not coherent with the formalization of the problem. It is obvious how the reduction will increase the efficiency of the algorithm.

The problem above can be solved following this procedure:

- Fix the Nominal Value on which is calculated the Debt Value Adjustment. This corresponds in fixing the N in the formula (8.1).

- Looking at the medium absolute daily variation of an instrument, I determine how much of that instrument I should take to hedge the DVA not considering the presence also of the other instruments. Thus what I will define is a “cap” on how much of each instruments I can buy. It will be the reinforcement learning algorithm, to understand how much of each instrument I should actually buy/sell, considering the presence of the others.

In our particular case:

- We considered a nominal value for the Debt Value Adjustment equal to 400 Milions €.
- The CDS itraxx can be sold on a single basis. Applying the method described above, we obtain that to hedge 400 millions €of DVA, up 700 millions of CDS itraxx are needed. This corresponds in having a single CDS with a nominal value up to 700 millions €. Since the magnitude of the nominal is so high, we rescaled the CDS nominal of 1 million. Thus, if the agent buys 3 CDS itraxx, actually it is buying 3 millions CDS with nominal 1€, or 1 CDS with a nominal value of 3 millions €. So the agent will act as if the CDS itraxx contract contains 1 million of CDS.
- The SX7E Future contracts actually contains 50 underlying instruments. Applying the method described above we get that up to 11500 future contracts on SX7E are needed. Thus actually they are 11500×50 single futures.
- The BTP/BUND Future contracts actually contain 1000 underlying instruments. We get that up to 2100 of them are needed. Thus actually they are 2100×1000 single futures.

Remark: Having defined in this way the space of actions $\mathcal{A}$, and the set of the states $\mathcal{S}$, we can notice that even if they are actually continous, it is easy to consider them as discrete and in particular finite space. For example as regard actions we can consider the case where the agent can buy/sell a quantity of instruments $\in \mathbb{N}$, then considering that its actions are bounded thanks to what we have just outlined, we can conclude that $\mathcal{A}$ is actually finite. This is important because we can actually model our problem as a finite Markov Decision Process in first approximation.

9.4 How the simulator works

As already remarked before it is not the purpose of the simulator to choose which type of action the agent have to undertake. Thus, to understand how the simulator works the policy of the agent is considered to be fixed a priori.

The simulator act as follows:

- The simulator chooses the starting date randomly. This date will be the starting point of the episode.
- Given the date by the previous point, the agent observes the associated state, and implement an action following the fixed policy. During the 1st day the net position between bank account and collateral account is zero, thus the process on which I will pay dividend at a rate equal to the 1y Intesa San Paolo Spread is 0.
- The prices have evolved thus there is a daily gain. And a correspondent change is recorded in the collateral account and in the net position on which dividend will be paid.
- Since the day changed and the prices have evolved, the agent will observe a different state. It will implement an action following the fixed police.
- The prices evolves again, and there is a daily gain because of their variations but also because of the dividend process computed on the net position of the previous day. And so on.

Remark: At the end of the 3 months a total reward on the episode is given.

Remark: In the reward is also considered the Debt Value Adjustment in line with the problem formalization.

Remark: Also in the simulator, in line with the assumptions, the risk free rate is set to 0. Thus, the dividend process on the collateral account is 0. And as written above the dividend paid on the net position between collateral and bank account is drifted just by the 1y Spread on Intesa San Paolo, and not by the risk free rate.

Chapter 10

Reinforcement Learning Algorithms

In this chapter will be discussed two Reinforcement Algorithm respectively for the policy valuation and policy improvement. This chapter does not aim to discuss RL techniques that will actually be applied on this project, bacause this is not the objective of the present work, and in particular further considerations must be done before passing to the "learning phase"; but for completeness, I think that they should be in this thesis too.

Remark: If the underlying system is a finite Markov Decision Process, the two methods that will be discussed can be actually used.

10.1 Dynamic Programming

Dynamic Programming stands for a collection of tools that can be used in order to compute the optimal policies.

The main characteristic of this algorithms is that they compute optimal policies given a perfect model of the environment. So for example the environment should be a Markov Decision Process. Where I want to underline that for Markov Decision Process is intended a process which state signal is markovian, but also the state signal assumed to describe the environment actually can completely describe it without losing any information. Thus, in order to apply these techniques to our problem we must be confident that our state signal can fully represent the environment, otherwise the agent will not learn properly. They are not very used in reinforcement learning because they are very expensive from the computational point of view, and as noticed before the perfect knowledge of the model is an assumption that is really hard to satisfy. Anyway they are important from a theoretical point of view, and they put also the basis for methods applicable to continous problem, or methods where the knowledge of a perfect model for the environment is not necessary.

The main idea underlying Dynamic Programming, is the use of the value functions in order to find good policies, in particular it exploits the Bellman Equations derived in chapter 7.

10.1.1 Iterative Policy Evaluation

The objective of this algorithm is to evaluate the value function V^π introduced in chapter 7, given a policy π. From chapter 7 we know that:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'} \left[\mathcal{R}^a_{s\,s'} + \gamma\, V^\pi(s')\right]$$

for all $s \in \mathcal{S}$ where we have already specified what each term represents. The idea is to exploit fixed point methods. Thus, suppose to choose V_0 arbitrarily, then we can create a sequence of approximatin function V_k, through the recursive formula:

$$V_{k+1}(s) = \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{s\,s'} \left[\mathcal{R}^a_{s\,s'} + \gamma\, V_k(s')\right] \quad \forall s \in \mathcal{S}$$

It is obvious how $V_k = V^\pi$ is a fixed point for the recursive formula, thanks to the Bellman Equation. Moreover it can be shown that the sequence V_k converge to V^π under the same hypothesis that ensure the existence of V^π.

Thus, the algorithm to use for an iterative policy evalutation is quite natural.
1) Take a policy π to evaluate.
2) Initialize V(s)=0, $\forall s \in \mathcal{S}$.
3) Apply the recursive formula above, obtaining a new $V(s) \forall s \in \mathcal{S}$.
4) Repeat the point 3), with the new $V(s)$, until a convergence criteria is not satisfied.

Remark: The convergence criteria can be for example:

$$\max_{s \in S} |V_{k+1} - V_k| < \varepsilon$$

where therefore we stop when the difference between 2 consecutive approximation of V^π is very small.

10.1.2 Policy Improvement

Hypothesis: In this chapter we consider determistic policies. Thus given a state s, the agent knows which action adopt. This obviously does not imply to know also the following state.

If the valuation of a given deterministic policy is done, then it also easier to find another policy that is better than this one. For a given state s, we should inquire if we should change the policy to deterministically take an action $a \neq \pi(s)$. In order to have an answer to this question, we have just to consider the formula in chapter 7.

$$Q^{\pi}(s,a) = \sum_{s'} \mathcal{P}^{a}_{s\,s'} \left[\mathcal{R}^{a}_{s\,s'} + \gamma\, V^{\pi}(s')\right]$$

Indeed, a way to answer to that formula is to select a in the state s at the beginning of the episode, and thereafter follow the existing policy. It is then obvious that if it is better to select a in s and thereafter follow the policy π, then it is better to select a everytime the agent in s, defining therefore a new policy π'. The result just mentioned can be easily proved. Thanks to this observation is quite natural to introduce the following algorithm which consider also the iterative policy evaluation considered in the previous section:
1) Initialize $V(s)$ and $\pi(s) \forall s \in \mathcal{S}$.
2) Do the policy evaluation as outlined in the previous section.
3) Do the policy Improvement. Thus, for each $s \in \mathcal{S}$:

- Set a variable $b(s) = \pi(s)$
- Update $\pi(s) = \text{argmax}_a \sum_{s'} \mathcal{P}^{a}_{s\,s'} \left[\mathcal{R}^{a}_{s\,s'} + \gamma\, V^{\pi}(s')\right]$
- if there is an $s \in \mathcal{S}$ such that $b(s) \neq \pi(s)$, then the policy is actually improved, thus we need to go back, and repeat from point 2). If such an s does not exist, then the policy is not changed, thus it cannot be improved anymore, therefore it is optimal.

Remark: The policy iteration involves policy evaluation each step. But policy evaluation is itself iterative and can requires lot of iterations before occuring to convergence. Thus, it can be be improved, but also extended considering also non-deterministic policies.

Chapter 11

Conclusions

As outlined in the first chapter the main objective of the present work was to formalize the problem and to implement the simulator in order to represent the environment. Regarding the formalization we are satisfied, since we have been actually able to define all features, considering also the problem from the Reinforcement Learning perspective, and to collect all financial data that were necessary. The simulator has already been tested by the CVA desk of Banca IMI, and it works as we expected.

It is obvious how in the current version of the simulator optimal result cannot be obtained; indeed we are not considering quantities as sensitivities that could be even crucial to determine a good hedging policy. In general we can say that the agent will never learn without informative features, and if these features are not expressed properly Learning can also fail.

Thus the project will proceed in the following way:

- Some reinforcement learning algorithms will be implemented in a basic version in order to understand if the agent can actually learn with the current representation of the environment.

- If the results from the previous point are satisfactory then the idea is to extend the basic version implemented to the original one presented in chapter 6.

- If the results from the algorithms are not satisfactory then two different ways can be followed:

    - Enrich the current state considering also sensitivities and other features that can be essential for the agent to learn.

- The features that we have considered until now, are features that were provided by experts (i.e. they were domain features). It would be also interesting besides considering the enrichment of these features to synthesize some of them through *Deep Learning* algorithms. Indeed, one of the key points of deep learning is replacing handcrafted features with feature learning algorithms, that as the name suggests are able to synthesize features from which the agent will learn.

Thus, with regard to the objectives of this thesis I can say that they were fully achieved, at the same time there is a lot of work to do and challenges to face in order to complete the whole project.

Bibliography

[1] Basel Committee on Banking Supervision (2016). Minimum capital requirements for market risk. `www.bis.org/bcbs/publ/d352.pdf`

[2] Basel Committee on Banking Supervision (2016). Instructions: CVA QIS. Annex 1. Draft minimum capital requirements for CVAs. `www.bis.org/bcbs/qis/instructions_CVA_QIS.pdf`

[3] Brigo D. e Mercurio F. (2001) Interest rate models theory and practice. Springer-Verlag

[4] Morini M. and Prampolini A. (2010). Risky funding with counterparty and liquidity charges. *Risk* March

[5] Jon Gregory. Counterparty Credit Risk.

[6] Martin Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming.

[7] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction.