HEC MONTRÉAL
École affiliée à l'Université de Montréal

Equal-risk pricing, hedging, and portfolio management using dynamic risk measures and deep reinforcement learning methods

par
Saeed Marzban

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(spécialisation Financial engineering)

Août 2021

HEC MONTRÉAL
École affiliée à l'Université de Montréal

Cette thèse intitulée :

Equal-risk pricing, hedging, and portfolio management using dynamic risk measures and deep reinforcement learning methods

Présentée par :

Saeed Marzban

a été évaluée par un jury composé des personnes suivantes :

Michel Denault
HEC Montréal
Président-rapporteur

Erick Delage
HEC Montréal
Directrice de recherche

Jonathan Yu-Meng Li
Telfer School of Management, University of Ottawa
Codirecteur de recherche

Alexandre F Roch
Université du Québec à Montréal (UQAM)
Membre du jury

Matt Davison
University of Western Ontario
Examinateur externe

Chantal Labbé
HEC Montréal
Représentante du directeur de HEC Montréal

Résumé

Dans cette thèse, nous explorons deux problèmes fondamentaux en ingénierie financière, à savoir la tarification d'options et la gestion de portefeuille. Dans le premier article de la thèse, nous adaptons les fondements théoriques d'un nouveau cadre de tarification d'options nommé "Equal Risk Pricing" (ERP) pour tenir compte des mesures de risque dynamiques convexes. Le cadre ERP suggère de tarifer un contrat d'option sur la base de l'idée que si le vendeur et l'acheteur de l'option couvrent tous deux leurs positions en investissant activement dans les actifs sous-jacents de l'option, le risque global auquel ils sont exposés devrait être égal. Cela contraste avec les méthodes conventionnelles d'évaluation des options où le prix suggéré ne prend en compte que du risque auquel le vendeur est confronté. Nous exploitons les propriétés des mesures de risque dynamiques convexes pour formuler les équations d'un programme dynamique averse au risque (DP) qui identifie les politiques de couverture optimales. En nous concentrant sur les mesures de risque du pire cas, nous montrons les avantages de l'ERP par rapport aux modèles conventionnels en termes de tarification et de couverture pour les options vanille.

Sur la base des reformulations théoriques obtenues dans le premier article, le deuxième article étend le cadre ERP à un cadre plus pratique en tirant parti du "Deep Reinforcement Learning" (DRL). En particulier, nous sommes en mesure de surmonter les difficultés de calcul des modèles DP conventionnels causées par la haute dimensionnalité de l'espace d'état, en nous appuyant sur la puissance d'approximation des réseaux neuronneaux profonds. Bien que nous ne soyons pas les premiers à explorer l'application du DRL à l'ERP, notre travail se différencie des études précédentes en étant le premier à employer une mesure de risque cohérente dans le temps. Nous expliquons d'abord dans un cadre simple les inconvénients de considérer une mesure de risque statique (c'est-à-dire incohérente dans le temps), qui

néglige l'aversion au risque de l'investisseur à des moments futurs. À l'aide d'un processus de mouvement brownien géométrique, nous comparons ensuite l'ERP sous des modèles de tarification cohérents et non-cohérents dans le temps pour les cas d'une option vanille et d'une option panier et discutons des avantages d'y employer une politique cohérentes dans le temps.

Bien que les deux premiers articles démontrent empiriquement des avantages de l'ERP, leurs résultats numériques sont entièrement basés sur des données synthétiques. Or, dans les marchés réels l'évolution des valeurs des titres se comporte différemment des environnements de simulation. Par conséquent, dans le dernier article de cette thèse, nous explorons l'application du DRL à l'analyse de séries chronologiques de données de marché réels. Pour ce faire, nous avons choisi le problème de la gestion de portefeuille car il est suffisamment général pour couvrir le problème de couverture qui est au coeur du calcul de l'ERP. Pour résoudre ce problème, nous proposons un modèle innovateur de réseau de neurones profonds basé sur des réseaux de neurones convolutifs (CNN) dilatés qui, lorsqu'ils sont utilisés dans un cadre DRL, surpassent l'état de l'art en termes de rendements ajustés au risque. En particulier, nous démontrons que notre modèle tire meilleur parti des informations de corrélation entre actifs et qu'il possède une propriété spéciale appelée invariance au permutation. Notre approche est la première à satisfaire cette propriété tout en prenant des corrélations entre actifs. Notre cadre DRL est finalement mis à l'épreuve sur des ensembles de données des marchés boursiers canadiens et américains pour confirmer la précision et la fiabilité de notre approche.

Mots-clés

Tarification des options, couverture du risque, mesures de risque convexes, marché incomplet, programmation dynamique, optimisation numérique, apprentissage par renforcement, expectiles, gradient de politique, options panier

Méthodes de recherche

Analyse numérique, intelligence artificielle et heuristique, programmation mathématique, recherche quantitative, simulation

Abstract

In this thesis, we consider two fundamental problems in finance, namely option pricing and portfolio management. In particular, we investigate a new mechanism for option pricing and develop new machine learning methodologies for solving this problem and a general portfolio management problem in large scale. This thesis consists of three articles. The first one presents the theory of equal risk pricing, a new derivative pricing framework, that builds upon the notion of monotone risk measures. The second one develops a novel reinforcement-learning methodology that allows the equal risk pricing problems to be solved when the number of underlying assets is large or when there is no model that can describe the underlying stochastic dynamics. Lastly, the third one proposes a new design of neural networks in deep reinforcement-learning for solving a general class of dynamic portfolio management problems.

The Equal Risk Pricing (ERP) framework addressed in the first article is different from the conventional methods of option pricing in that ERP seeks a price that takes into account both the perspective of the buyer and the writer, and ensures the risk exposure of both parties are the same, whereas the conventional methods only take into account a single trader's perspective (i.e. the writer). Motivated by the modern theory of convex risk measures, we study equal risk pricing problems where both parties' risk preferences are captured by some monotone risk measures. As the main results, we characterize the equal risk price by exploiting the properties of convex risk measures and we show how the hedging problems of both parties can be solved by risk-averse dynamic programming equations. We demonstrate the effectiveness of our framework by implementing it for European and American options and by considering worst-case risk measures motivated by the literature on robust optimization.

The second article addresses the practical issues arising from consideration of options with a large number of assets, and the cases where a stochastic model is hard to identify for the underlying markets, in which case Dynamic Programming (DP) is no longer applicable. We consider ERP problems formulated based on dynamic expectile risk measures and present a novel Deep Reinforcement-Learning (DRL) approach that can solve ERP problems by exploiting the properties of expectile risk measures. While we are not the first to explore ERP using DRL, our work differentiates itself from previous studies in the literature by being the first to employ a risk measure that is time consistent. We provide an illustrative example of how hedging policies that violate the time consistency conditions can be problematic for practical implementation. We then develop the first off-policy actor-critic Reinforcement-Learning (RL) algorithm that can generate time-consistent hedging policies for both the writer and the buyer. Through extensive experiments, we show for problems of pricing first a European option and then a basket option how the algorithm allows for solving ERP problems with a large number of underlying assets and how the time consistent policies generated from the algorithm can be practically more useful than the policies generated based on static risk measures.

In the third article, we pay particular attention to the development of a DRL method for solving portfolio management problems that entail the use of long time series data. Motivated by the nature of time series data, we propose a new design of the neural network for a portfolio policy that can exploit cross-asset dependency information from the time series data. In particular, we identify and define for the first time an essential property, namely the property of asset-permutation invariance, that a portfolio policy network should satisfy so as to have a stable performance (i.e. keeping everything else unchanged, modifying the indexing of the assets in the input matrix of the policy network before training does not significantly affect the resulting performance). While several other state-of-the-art neural network architectures fail to satisfy this property, we propose an innovative dilated Convolutional Neural Network (CNN), called WaveCorr, that not only enjoys the strength of CNN in that it is parsimoniously parameterized but also satisfies the asset-permutation invariance property. In addition to respecting this property, our proposed architecture also succeeds to take the cross-asset correlation information into account. We demonstrate through testing on data sets from both Canadian and American stock markets

the effectiveness of our design to utilize the correlation information and achieve superior (and more stable) risk-adjusted performances. The superiority of our network in solving portfolio management problems implies also the possibility of applying it to improve the DRL approach developed in the second article for solving the hedging problems embedded in the ERP problems.

Keywords

Option pricing, risk hedging, convex risk measures, incomplete market, dynamic programming, numerical optimization, risk-averse reinforcement-learning, expectile risk measure, policy gradient, basket options

Research methods

Numerical analysis, artificial intelligence and heuristics, mathematical programming, quantitative research, simulation.

Contents

List of Tables

List of Figures

List of acronyms

ACRL Actor-critic reinforcement learning

AORL Actor-only reinforcement learning

ATM At-the-money

CVaR Conditional Value-at-Risk

CNN Convolutional neural network

CORL Critic-only reinforcement learning

DRL Deep reinforcement learning

DP Dynamic programming

DRM Dynamic risk model

ERP Equal risk pricing

FPI Fair price interval

ITM In-the-money

LSTM Long-Short-Term-Memory

ML Machine learning

MDP Markov decision process

NN Neural network

OTM Out-of-the-money

RL Reinforcement learning

SRM Static risk model

VaR Value-at-Risk

Acknowledgements

Throughout the writing of this thesis I have received invaluable support and assistance, and I would like to thank those who helped me succeed in my PhD.

I would like to first and foremost thank my first co-supervisor, Professor Erick Delage, who supported me during all stages of my PhD and writing of this thesis. His expertise and the invaluable efforts that he devoted to my work along with his patience made producing such quality work possible.

Second, I would like to thank my second co-supervisor, Professor Jonathan Yumeng Li, who was guiding me in all years of the PhD by providing different perspectives and insightful feedback into my work. His support always pushed me to sharpen my thoughts and to see through the problems differently.

I would also like to thank Professor Alexandre F Roch, who was accompanying my journey as the committee member in all phases of my PhD, and providing precious comments at different points of this work.

Finally, I would like to thank my parents, my sister, my brother, and their families for their emotional support in my difficult times during the PhD. I would not have been able to complete this stage of my life without their wise counsel and sympathetic ears.

Preface

This thesis includes three articles listed as follows:

- Saeed Marzban, Erick Delage, and Jonathan Yumeng Li. Equal risk pricing and hedging of financial derivatives with convex risk measures. Accepted in the journal of Quantitative Finance, 2021.

- Saeed Marzban, Erick Delage, and Jonathan Yumeng Li. Deep Reinforcement Learning for Equal Risk Pricing and Hedging under Dynamic Expectile Risk Measures. Submitted to International Conference on Learning Representations (ICLR).

- Saeed Marzban, Erick Delage, Jonathan Yumeng Li, Jeremie Desgagne-Bouchard, Carl Dussault. WaveCorr: Correlation-savvy Deep Reinforcement Learning for Portfolio Management. Submitted to the International Conference on Learning Representations (ICLR).

General Introduction

Derivative pricing and portfolio management are two fundamental problems in finance. The former is closely related to the latter in that the price of a derivative can often be interpreted as the initial capital required for building a self-financing portfolio that replicates the payoff of the derivative. From this perspective, these two problems share the similarity of dynamically managing a portfolio over time. It is known that derivative pricing is particularly challenging when the markets are incomplete, in which case there does not exist a self-financing portfolio that can perfectly replicate the payoff of the derivative. The first chapter presented in this thesis presents an article that seeks to address the challenge of derivative pricing in incomplete markets by providing a new pricing framework, known as Equal Risk Pricing (ERP), that takes into account both the writer and the buyer's hedging decisions and risk exposures. Central to the development of ERP are the modelling of risk preferences of both the writer and the buyer in a dynamic setting and the formulation of corresponding risk-averse hedging problems. Motivated by the recent advances of risk theory, we propose in the first chapter to study ERP problems by exploring the use of convex risk measures to model the risk preferences, and provide detailed derivations for solving ERP problems using Dynamic Programming (DP). The use of DP, however, has its practical limitations; namely it can only be applied when the number of state variables is small and the dynamics of the underlying stochastic systems can be perfectly modelled. As another main goal of this thesis, we seek to take advantage of the new advances in Machine Learning (ML) and develop new numerical solutions that can solve dynamic problems in large scale, i.e. with a large number of state variables, and in a data-driven fashion, i.e. relying only on sample data. In particular, in the second chapter of this thesis, we present an article that considers ERP problems when the number of underlying assets is large and

only sample data may be available for describing the underlying stochastic systems. We present a novel deep reinforcement learning approach for solving the hedging problems in ERP. Throughout our development, we discover there is a deep connection between the structure of a risk measure, namely the property of elicitability, and the design of a reinforcement learning approach that can be used to solve the corresponding risk-averse dynamic programs. Leveraging our new deep reinforcement learning approach, we are able to further demonstrate at large scale how the hedging policies generated based on dynamic risk measures can benefit from the property of time-consistency compared to the policies generated based on static time inconsistent risk measures. In the last chapter of this thesis, a third article seeks to develop a deep reinforcement learning approach for solving a general portfolio management problem. In particular, the focus is on designing a policy neural network that can better exploit the cross-asset correlation information embedded in time series data. We identify the property of permutation invariance that can be used to guide the design of a policy network and present the first convolutional neutral network (CNN) infrastructure that satisfies this property.

Overall, the thesis provides solid theoretical grounds and solution methods that are implementable by practitioners for solving the equal risk pricing problem and general portfolio management problem. In what follows, we briefly introduce the work that is done in each of the three chapters along with previous studies in the literature:

Chapter 1: Equal Risk Pricing and Hedging of Financial Derivatives with Convex Risk Measures

Contrary to complete markets, financial derivatives in an incomplete market cannot be priced solely according to non-arbitrage theory as traditionally exploited in Black and Scholes (1973); Merton (1973); Cox et al. (1979); King (2002). The risk premium for the unhedgable risk needs to be calculated in pricing financial derivatives. Following this, most researchers tackled this problem by using two different approaches. The first one is to exploit a fixed risk-neutral martingale measure, including for example Hull and White (1987); Heston (1993); Amin (1993); Delbaen and Schachermayer (1995), and Brennan (1979), whereas the second set of approaches rely on identifying the indifference price of a risk-averse hedging problem, including for example Jaschke and Küchler (2001); Carr et al. (2001); Föllmer et al. (1985); Schweizer (1996); Gourieroux et al. (1998), and Bertsimas

et al. (2001).

In this first article, we investigate a recently developed option pricing framework named Equal Risk Pricing (ERP) (Guo and Zhu, 2017) that is more closely related to the second aforementioned category in that it also involves the formulation of risk-averse hedging problems. However, unlike most pricing schemes in this category that only consider a single trader, namely the writer, in the formulation of risk-averse hedging problem, the ERP framework is formulated based on two separate risk-averse hedging problems, one for the writer and another for the buyer. The minimum price that a writer is willing to take according to a writer's hedging problem is generally higher than the highest price that a buyer is willing to pay according to a buyer's hedging problem. The novelty of ERP lies in providing a mechanism to determine a "fair" price that can be acceptable to both parties, namely a price that leaves both the writer and buyer with equivalent risk exposure.

While the initial ERP framework proposed in Guo and Zhu (2017) focuses on incomplete markets where the risk is measured according to expected utility, in this first article we extend the definition of ERP to the set of all monotone risk measures that can be interpreted as certainty equivalent measures. This class of risk measures include (dynamic) convex risk measures developed in the modern risk theory. We establish for the first time that based on this class of risk measures, ERP is arbitrage-free under weak conditions and actually reduces to computing the center of a so-called Fair Price Interval (FPI). In comparison to the work of Guo and Zhu (2017) which focused on an expected disutility framework that employs a fixed martingale measure, our generalized framework allows an arbitrary, and possibly different, probability measure to be used by each party, and corrects for the fact that the expected disutilities experienced by the two different parties are intrinsically incomparable. In the case of discrete-time hedging, we show how the boundaries of such a fair price interval can be obtained using dynamic programming for both European and American options as long as the convex risk measures employed by the two parties are one-step decomposable and satisfy a Markovian property. These dynamic programs are amenable to numerical computation given that they employ a finite-dimensional state space. In the case of American options, they will also provide a different price depending on whether the buyer is willing to commit up front to an exercise strategy.

In our implementation of ERP, we consider a form of worst-case risk measures motivated

by the literature of robust optimization that considers only a subset of the outcome space, also known as the uncertainty set, in calculating the largest possible risk. Such risk measures are highly interpretable and, as shown in our work, can be well incorporated into ERP based on the analysis we established. Our numerical experiments indicate that the fair price interval might converge, as the number of rebalancing periods increases, to the Black-Scholes price when an uncertainty set inspired by the work of Bernhard (2003) is properly calibrated in a market driven by a geometric Brownian motion. This makes the connection between the common risk-neutral pricing scheme and the equal risk price generated based on worst-case risk measures. Finally, we present the first numerical study that provides evidence that equal risk prices allow both the writer and the buyer to be exposed to risks that are more similar and on average smaller than what they might experience with a risk neutral probability measure or quadratic hedging prices. In particular, when a worst-case risk measure is used, the risk inequity for the higher quantiles of each party's final loss will be reduced by a factor between 2 and 10 (depending on the type of option) compared to ϵ-arbitrage and Black-Scholes prices. This is done while keeping the average risk among the two parties to a similar or better level.

Chapter 2: Deep Reinforcement Learning for Equal Risk Pricing and Hedging under Dynamic Expectile Risk Measures

The ERP method developed in our first article provides a theoretical basis for pricing options, but it relies on the use of DP to solve the problems. It is known that DP becomes computationally intractable when applied to options that are written on assets with complicated price processes or on multiple underlying assets. In this second article, we extend the scalability of the ERP model by developing a Deep Reinforcement Learning (DRL) approach for solving hedging problems in high dimension, i.e. with a large number of state variables, so that the ERP model can be used in practice to price high dimensional option contracts.

The recent works of Carbonneau and Godin (2020) and Carbonneau and Godin (2021) are the first that apply DRL to solve ERP problems. While the DRL approaches developed in Carbonneau and Godin (2020) and Carbonneau and Godin (2021) are applicable only to ERP problems formulated based on static risk measures, we develop in this second article a DRL approach that can solve ERP problems formulated based on a class of dynamic risk

measures. The consideration of dynamic risk measures is critical in ensuring the resulting hedging policies are operationally meaningful. Namely, it is known that hedging policies generated based on static risk measures could violate the property of time-consistency, meaning that the hedging policies may not be considered optimal once a downstream state is visited, whereas hedging policies generated based on dynamic risk measures would naturally be time-consistent. The consideration of dynamic risk measures also has an implication on the type of DRL approach that needs to be developed. In particular, while the DRL approach developed in Carbonneau and Godin (2020) and Carbonneau and Godin (2021) is limited to a policy optimization scheme, a.k.a. Actor-Only Reinforcement Learning (AORL) algorithm (see Williams (1992)) due to the lack of dynamic optimality principle in the use of static risk measures, we are able to explore the design of an alternative type of DRL approach, namely Actor-Critic Reinforcement Learning (ACRL) algorithms (see Lillicrap et al. (2015)) by leveraging the DP equations that are available in the use of dynamic risk measures.

In particular, we propose the use of dynamic expectile risk measures to formulate time-consistent ERP problems. This is motivated by the theory of coherent risk measures, which identifies the advantages of expectile risk measures over two other popular risk measures, namely Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), particularly that expectile risk measures are the only elicitable (coherent) risk measures. Moreover, we discover how the property of elicitability can actually facilitate the design of a model-free actor-critic algorithm, i.e. allowing a policy to be updated per sample data. Our deep reinforcement learning architecture is built upon the formulation of Q-value dynamic equations for the expectile-based hedging problems, and it consists of two networks, a policy network (actor) and a Q network (critic). The novelty lies in the design of an algorithm used to update the two networks based on stochastic gradients. Our algorithm may be considered as an extension of the off-policy deterministic actor-critic method. Following similar arguments made in Degris et al (2014), the algorithm updates the policy network based on an approximate stochastic gradient descent algorithm. Updating the Q network also requires care, as the hedging problem is evaluated in terms of risk rather than expected value commonly applied in a RL problem. By leveraging on the elicitability property of expectile risk measures, which implies that the risk measure can always be calculated as the

optimal solution with respect to a certain score function, we show that the algorithm can update the Q network using stochastic gradients calculated based on the score function.

We observe that this ACRL has a particularly good convergence behaviour when there is an immediate reward following each action. While the hedging problem in ERP is only concerned about the cumulative wealth and the payoff of a derivative that occurs at the very end, the translation invariance property of expectile risk measures allows the cumulative wealth to be re-expressed as a sequence of immediate rewards that need to be optimized over time. Indeed, throughout our experiments, noticeably stronger convergence behaviour is found for the hedging problem formulated in terms of immediate rewards than the one formulated based on cumulative wealth. The effectiveness of our ACRL for finding a good policy is further demonstrated in the numerical section.

In the numerical section we first investigate a vanilla option where our purpose is twofold. First, we want to show the Q-function is precise enough to be used for the sake of pricing based on the ERP framework. This is performed by comparing the results of this model with a grid-based DP model that can be trusted to provide a baseline for the value function. The results show that the Q-function can approximate well the true value function when the policies are coming from the ACRL model. Second, we demonstrate the benefits of having a model that provides a time-consistent solution in practice. The numerical results support our claim that a time-consistent model can be used for training a model over options with long maturities and then using the trained model to hedge and price options with shorter maturities. We also show the benefits of formulating the option pricing problem such that the immediate rewards are included in the RL setting. More precisely, we show this transformation greatly improves the convergence properties of the model. Finally, we focus on basket options where we demonstrate the main purpose of extending the ERP model to using RL for pricing and hedging, which is improving the scalability. The results in this section follow our previous results in the case of vanilla option where the time-consistent solution is able of outperforming the time inconsistent solution as the risk is measured at later points of time.

Chapter 3: WaveCorr: Correlation-savvy Deep Reinforcement Learning for Portfolio Management

In this last article, we turn our attention to a more general dynamic portfolio management

problem and study the potential of DRL for generating portfolio policies that can effectively exploit the information embedded in the time series of historical asset prices. Although implemented for a portfolio optimization problem, the work presented in this papaer is closely related to the work presented in the second paper in that both seek to develop DRL approaches for solving dynamic problems in large scale and in a data-driven fashion. While the second paper focuses on developing a general DRL approach for addressing hedging problems defined based on dynamic risk measures, the last paper pays particular attention to the design of the architecture of investment policy neural network (embedded in DRL) motivated by the structure of time series data used in portfolio management problems. In particular, we introduce an innovative type of neural network architecture that can replace the simple, yet heavily parametrized, fully connected architecture commonly applied in DRL (e.g. in the actor network of our ACRL framework developed in the second paper). This new architecture is parsimonously paramterized but still has the ability to capture sophisticated forms of dependency from historical time series data, namely both the cross-asset and cross-time dependencies of the assets, and use them to improve portfolio decisions.

Most of the papers in the literature that study portfolio management problem using DRL, such as Moody et al. (1998); He et al. (2016); Liang et al. (2018) among others, focus on the task of prediction, rather than the problem of making allocation decisions. Training a DRL model for the purpose of portfolio allocation has requirements that time series prediction models may not necessarily satisfy. For example, it is known that extracting and exploiting cross-asset dependencies over time is crucial to the performance of portfolio management. However, the neural network architectures adopted in most existing works, such as Long-Short-Term-Memory (LSTM) or Convolutional Neutral Network (CNN), only process input data on an asset-by-asset basis and thus lack a mechanism to capture cross-asset dependency information. Also, several works that tried to modify these models in order to consider cross-asset correlation end up having trouble in satisfying a crucial property that we introduced in this paper as “asset-permutation invariance". Asset-permutation invariance is critical in that it ensures that the model performance is insensitive to the permutation of the assets in the input tensor.

The architecture presented in this paper, named as WaveCorr, is built upon the WaveNet

(Oord et al., 2016), which uses dilated causal convolutions at its core, and a new design of correlation block that can process and extract cross-asset information. We show that WaveCorr, despite being parsimoniously parameterized, can satisfy the property of asset-permutation invariance, whereas a naive extension of CNN can fail to satisfy this property. Closer to our work is the recent works of Zhang et al. (2020) and Xu et al. (2020), both seek to extract cross-asset dependency information. As Zhang et al.'s work shares a similar architecture as ours, we demonstrate in details why their architecture fails to satisfy the property of asset-permutation invariance and how this can lead to unstable performances. In the numerical section, we test the performance of WaveCorr using data from both Canadian and American stock markets. The experiments demonstrate that WaveCorr consistently outperforms our benchmarks under a number of variations of the model: including the number of available assets, transaction costs, etc., which makes it a reliable model to be incorporated into the ERP framework.

Chapter 1

Equal Risk Pricing and Hedging of Financial Derivatives with Convex Risk Measures

Chapter information

This article is a joint work with my supervisors, Erick Delage, and Jonathan Yu-Meng Li. It is accepted in the journal of Quantitative Finance. The preprint is available at `https://arxiv.org/abs/2002.02876`

Abstract

In this article, we consider the problem of equal risk pricing and hedging in which the fair price of an option is the price that exposes both sides of the contract to the same level of risk. Focusing for the first time on the context where risk is measured according to convex risk measures, we establish that the problem reduces to solving independently the writer and the buyer's hedging problem with zero initial capital. By further imposing that the risk measures decompose in a way that satisfies a Markovian property, we provide dynamic programming equations that can be used to solve the hedging problems for both the case of European and American options. All of our results are general enough to accommodate situations where the risk is measured according to a worst-case risk measure as is typically

done in robust optimization. Our numerical study illustrates the advantages of equal risk pricing over schemes that only account for a single party, pricing based on quadratic hedging (i.e. ϵ-arbitrage pricing), or pricing based on a fixed equivalent martingale measure (i.e. Black–Scholes pricing). In particular, the numerical results confirm that when employing an equal risk price both the writer and the buyer end up being exposed to risks that are more similar and on average smaller than those they would experience with the other approaches.

1.1 Introduction

One of the main challenges in pricing and hedging financial derivatives is that the market is often incomplete and thus there exists unhedgeable risk that must be further considered in pricing. In such a market, the price of a financial derivative cannot be set according to non-arbitrage theory as traditionally exploited in Black and Scholes (1973); Merton (1973); Cox et al. (1979); King (2002). Modern approaches to incomplete market pricing can be broadly divided into two main categories. The first one involves pricing a derivative based on a fixed "risk-neutral" martingale measure, either obtained from calibrating against market data (Hull and White, 1987; Heston, 1993; Amin, 1993), by minimizing the distance to a physical measure (Delbaen and Schachermayer, 1995), or by marginal indifference pricing (Brennan, 1979). The second category involves methods that rely on identifying the indifference price of a risk-averse hedging problem, including for example good deal bounds (Jaschke and Küchler, 2001), expected utility indifference pricing (Carr et al., 2001), or the quadratic hedging models (Föllmer et al., 1985; Schweizer, 1996; Gourieroux et al., 1998; Bertsimas et al., 2001). We refer readers to Schweizer (1999) and Staum (2007) for comprehensive surveys of these methods.

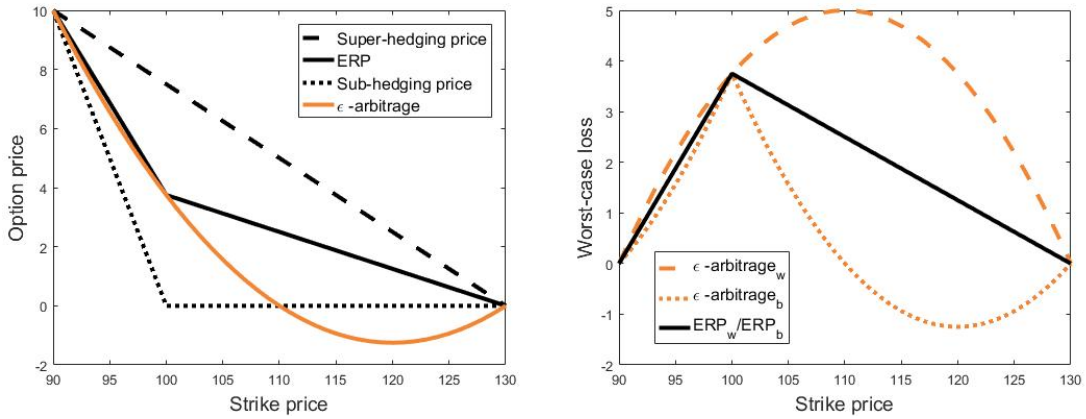
In this article, our focus is on studying a pricing method known as equal risk pricing (ERP), which was first introduced in the recent work of Guo and Zhu (2017). The method can be considered to fall into the second category mentioned above in that it involves the formulation of risk-averse hedging problems. In particular, it takes into account the risk preferences of both sides of a contract and seeks a fair unique transaction price that would ensure the minimal risk exposures (according to the formulated risk-averse hedging problems) of both sides of a contract are equal. In Guo and Zhu (2017), special attention

was paid to the case where risk is measured based on an expected disutility framework and where the market is incomplete due to no-short-selling constraints on the hedging positions. They proved the existence and the uniqueness of the equal risk price and provided pricing formulas for European and American options with payoffs that are monotonic in the underlying asset price. In the case where the constraints are lifted, they showed that the equal risk price coincides with the price resulting from a complete market model.

To put into perspective the strength of ERP, we should emphasize that most pricing methods focus only on a single side of the contract when formulating risk-averse hedging problems. The minimum price that a writer is willing to take according to a writer's hedging problem is, however generally higher than the highest price that a buyer is willing to pay according to a buyer's hedging problem. Hence, there is a lack of mechanism to suggest a "transaction" price, i.e. acceptable to both the writer and the buyer. ERP provides such a mechanism by suggesting that a transaction should occur at a price which leaves both the writer and buyer with equivalent risk exposure. To better illustrate this point, one can consider the example of pricing an European call option in the context where hedging can only occur at time zero. We further assume that the risk-free rate is zero, and that the underlying stock price starts at a value of 100$ while its value at exercise time is known to be uniformly distributed over [90, 130]. In this context, a risk-averse writer might require that the price of an at-the-money option be set as high as 7.5$ to fully cover her risk while the buyer can use the same argument to require a price of 0$. When a worst-case risk measure is used for both parties, one can show that the ERP allows the two parties to settle for the price of 3.75$ which exposes both of them to the same risk, i.e. 3.75$. Alternatively, one could suggest a transaction price based on a quadratic hedging scheme such as ϵ-arbitrage pricing (see its application with worst-case risk measure in Bandi and Bertsimas (2014)), yet as shown in Figure 1.1, such paradigms can propose prices that leaves both parties with surprisingly uneven risk, giving in some case even rise to arbitrage opportunities (c.f. the negative price for strike prices between 110 and 130). We refer the reader to Appendix 1.6.1 for details of the analysis presented in this figure.

The contribution of the article can be summarized as follows:

- We extend the definition of ERP to the set of all monotone risk measures that can be



(a) The option prices

(b) The hedging loss results

Figure 1.1 – Comparison of prices and hedging loss in the simple one-period European call option pricing example. (a) shows the upper and lower bound of the no-arbitrage interval, together with the equal risk and ϵ-arbitrage prices. (b) shows the worst-case loss incurred by each party of the contract under their respective optimal hedging strategies.

interpreted as certainty equivalent measures (i.e. $\rho(t) = t$ for all $t \in \mathbb{R}$). This class of risk measure includes the set of convex risk measures for which we establish for the first time that ERP is arbitrage-free under weak conditions and actually reduces to computing the center of a so-called fair price interval (FPI). In comparison to the work of Guo and Zhu (2017) which focused on an expected disutility framework that employs a fixed equivalent martingale measure, our generalized framework allows an arbitrary, and possibly different, probability measure to be used by each party, and corrects for the fact that the expected disutilities experienced by the two different parties are intrinsically non comparable.

- In the case of discrete-time hedging, we show how the boundaries of such a fair price interval can be obtained by using dynamic programming for both European and American options as long as the convex risk measures employed by the two parties are one-step decomposable and satisfy a Markovian property (see Section 1.2.1 for proper definitions). These dynamic programs are amenable to numerical computation given that they employ a finite-dimensional state space. In the case of American options, they will also provide a different price depending on whether the buyer is willing to commit up front to an exercise strategy.

- In the context where the underlying asset follows a geometric Brownian motion, we show for the first time how robust optimization can motivate the use of worst-case risk measures that only consider a subset of the outcome space. Similar to risk neutral risk measures (i.e. that measure risk using expected value), these worst-case risk measures are easily interpretable and will satisfy the properties needed for dynamic programming to be used. On the other hand, unlike risk neutral measures, they also provide risk-aware hedging policies. Our numerical experiments also indicate that the fair price interval might converge, as the number of rebalancing periods increases, to the Black–Scholes price when an uncertainty set inspired by the work of Bernhard (2003) is properly calibrated in a market driven by a geometric Brownian motion. If supported theoretically, such a property would close the gap between risk neutral pricing and risk-averse discrete-time hedging using worst-case risk measures.

- We present the first numerical study that provides evidence that equal risk prices allow both the writer and the buyer to be exposed to risks that are more similar and on average smaller than what they might experience with risk neutral or quadratic hedging prices. In particular, when a worst-case risk measure is used, the risk inequity for the higher quantiles of each party's final loss will be reduced by a factor between 2 and 10 (depending on the type of option) compared to ϵ-arbitrage and Black–Scholes prices. This is done while keeping the average risk among the two parties to a similar or better level.

The article is organized as follows. In Section 1.2 we formally define the equal risk pricing framework and demonstrate that this price coincides with the mid point of the fair price interval when risk is captured using convex risk measures. In Section 1.3, we focus on the context of discrete-time option pricing and derive the dynamic programming equations that can be used to compute the equal risk price for European and American contingent claims. Next, in Section 1.4, an application of equal risk pricing is presented where the risk attitude of both writer and buyer is captured by so-called worst-case risk measures. A numerical study is also presented to validate the quality of prices obtained using the equal risk pricing paradigm both from the point of view of risk exposition for the parties and fairness. Finally, we conclude the article in Section 1.5. We further refer the reader to an

extensive set of Appendices describing detailed arguments supporting all propositions and lemmas presented in this article.

1.2 The Equal Risk Pricing Framework

This section presents equal risk pricing framework and provides an interpretation of the price resulting from this model. In particular, we introduce the use of risk measures in pricing and hedging options based on this framework.

1.2.1 The Equal Risk Pricing Model

To present the equal risk framework, we consider a model of the market proposed by Xu (2006). Namely, we assume that the market is frictionless, i.e. there is no transaction cost, tax, etc. The filtered probability space is defined as $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$ and there is a money market account with zero interest rate, for simplicity and a risky asset S_t, with $0 \leq t \leq T$, which is $\mathcal{F}_t$-measurable. As in Xu (2006), we assume that the risky asset S_t is a locally bounded real-valued semi-martingale process. Furthermore, the set of equivalent local martingale measures for S_t is assumed non-empty to exclude arbitrage opportunity.

The set of all admissible self-financing hedging strategies with the initial capital p_0 is shown by $\mathcal{X}(p_0)$:

$$\mathcal{X}(p_0) = \left\{ X_t \,\middle|\, \exists \xi_s, \exists c \in \mathbb{R}, \quad X_t = p_0 + \int_0^t \xi_s dS_s \geq c, \quad \forall t \in [0,T] \right\},$$

in which, for each t, the decision ξ_t is $\mathcal{F}_t$-measurable and represents the number of shares of the risky asset in the portfolio, X_t is the accumulated wealth, and for simplicity, we assume without loss of generality that the risk-free rate is zero. Although we impose very few restrictions on the hedging strategies in the set $\mathcal{X}(p_0)$, as mentioned in Xu (2006), the assumption of locally bounded real-valued semi-martingale S_t allows many jump-diffusion and pure-jump models to be considered for the price process thus already giving rise to the possibility of an incomplete market. Alternatively, other definitions of $\mathcal{X}(p_0)$ could be used here to model different characteristics of the market, e.g. discrete trading times (see Section 1.3), or transaction costs, etc., without affecting the nature of our discussion.

We consider in this article a class of payoff functions $F(\{S_t\}_{0\leq t\leq T})$ that admit the formulation of $F(S_T, Y_T)$ where Y_t is an auxiliary fixed-dimensional stochastic process that is $\mathcal{F}_t$-measurable. This class of payoff functions is common in the literature, for example in Bertsimas et al. (2001) and is more easily amenable to numerical methods (see Section 3 for more detailed discussions in a Markovian setting). Here are two examples, where we denote by $\{t_k\}_{k=1}^N$ a set of discrete-time points.

1. **Options on the maximum value reached by a stock.** The option pays off the maximum stock price reached during $\{t_k\}_{k=1}^N$ and is defined by:

$$F\left(\max_{k=1,\dots,N} S_{t_k}\right),$$

which is a function of the whole history of the stock price. In order to define the payoff as a function of some variables at the current time, Y_t can be defined as:

$$Y_t = \max_{k:t_k\leq t} S_{t_k}.$$

Using this definition, the payoff of the option at the maturity can be written as a function of Y_T.

2. **Asian options.** The payoff of an Asian option is a function of the average stock price during $\{t_k\}_{k=1}^N$:

$$F\left(\frac{1}{N}\sum_{k=1}^{N} S_{t_k}\right).$$

Letting Y_t have the following form:

$$Y_t = \frac{1}{N}\sum_{k:t_k\leq t} S_{t_k}.$$

Again the payoff of the option at the maturity can be written as a function of Y_T.

We refer the reader to Bertsimas et al. (2001) for a broader range of options that can be recast into such a general form of payoff function.

Knowing the option payoff at maturity and assuming that the risk aversion of both participants are respectively characterized by two risk measures ρ^w and ρ^b, i.e. that map any random liability in $\mathcal{L}_p(\Omega, \mathcal{F}_T, \mathbb{P})$, which one wishes to minimize, to the set of real numbers (or infinity) and capture the amount of risk that is perceived by the participants,

it is possible to define the minimal risk achievable by the option writer and the buyer as follows:

$$\varrho^w(p_0) = \inf_{X \in \mathcal{X}(p_0)} \rho^w(F(S_T, Y_T) - X_T) \tag{1.1a}$$

$$\varrho^b(p_0) = \inf_{X \in \mathcal{X}(-p_0)} \rho^b(-F(S_T, Y_T) - X_T)\,, \tag{1.1b}$$

where $p_0 \in \mathbb{R}$ represents the price that is charged by the writer to the buyer for committing to pay an amount $F(S_T, Y_T)$ to the buyer at time T. The quantities $\varrho^w(p_0), \varrho^b(p_0)$ are the minimal risks associated to the optimal hedging of the writer and the buyer, respectively. In equation (1.1a), the writer is receiving p_0 as the initial payment and implements an optimal hedging strategy for the liability captured by $F(S_T, Y_T)$. On the other hand, in (1.1b), the buyer is assumed to borrow p_0 in order to pay for the option and then to manage a portfolio that will minimize the risks associated to his final wealth $F(S_T, Y_T) + X_T$. Note that while in practice the buyer usually might not buy an option by short-selling the risk-free asset and might not optimize a portfolio with the intent of hedging the option, equation (1.1b) identifies the minimal risk that he could achieve by doing so, which can certainly serve as an argument in negotiating the price of the option given that this is always a possibility for him.

Following the work of Guo and Zhu (2017), the notion of minimal risk achievable for both participants can in turn be used to define an equal risk price as follows.

Definition 1. (Equal risk price) Given that both the writer's risk measure, ρ^w, and buyer's risk measure ρ^b are interpretable as certainty equivalents, i.e.:

$$\forall c \in \mathbb{R}, \quad \rho(c) = c\,, \tag{1.2}$$

and are monotone[1], i.e. having X, Y representing costs,

$$\forall\, X,\, Y, \quad X \geq Y \text{a.s.} \Rightarrow \rho(X) \geq \rho(Y)\,, \tag{1.3}$$

then the equal risk price is defined as the unique p_0^* that satisfies

$$\varrho^w(p_0^*) = \varrho^b(p_0^*) \in \mathbb{R}\,, \tag{1.4}$$

[1]Technically speaking, we also require both risk measures to satisfy Fatou's property and to satisfy $\rho(X) = \lim_{m\to\infty} \rho(\min(X, m))$ when X is uniformly bounded from below while it should satisfy $\rho(X) = \lim_{m\to\infty} \rho(\max(X, -m))$ when X is uniformly bounded above if the limit is finite otherwise be considered undefined (see Xu (2006) for details).

when such a unique price exists.

The reason for imposing equation (1.2) is to make sure that the units of ϱ^w and ϱ^b are comparable, i.e. that risk is expressed in the units of equivalent certain payoffs. Note that this assumption is not imposed in Guo and Zhu (2017) where the notion of equal risk price can become arbitrary, e.g. when ρ measures expected utility since utility functions are defined only up to positive affine transformations. Note also that this definition of equal risk price holds for general European options yet we will later present a similar definition for American options as well.

Besides the equal risk price, in an incomplete market with two risk-averse market participants, another relevant and closely related concept takes the form of the following "fair price interval" (c.f. Bernhard et al. (2013)).

Definition 2. (Fair price interval) Given a writer's risk measure ρ^w and buyer's risk measure ρ^b, the fair price interval is defined as the interval of prices for which both the writer and the buyer are unable to exploit the market to completely hedge the risk of the contract they have agreed upon. Mathematically, the fair price interval takes the form $[p_0^b, p_0^w]$, where $p_0^b = \sup\{p_0 | \varrho^b(p_0) \leq 0\}$ and $p_0^w = \inf\{p_0 | \varrho^w(p_0) \leq 0\}$.

It is worth differentiating the FPI from the no-arbitrage interval. In particular, the latter is defined as the interval $[\bar{p}_0^b, \bar{p}_0^w]$, such that:

$$\bar{p}_0^b := \sup\{p_0 | \exists X \in \mathcal{X}(-p_0), F(S_T, Y_T) + X_T \geq 0 \text{ a.s.}\}$$

and

$$\bar{p}_0^w := \inf\{p_0 | \exists X \in \mathcal{X}(p_0), F(S_T, Y_T) - X_T \leq 0 \text{ a.s.}\}.$$

One can easily exploit the fact that the risk measures are interpretable as certainty equivalents and monotone to show that the FPI, which accounts for the fact that the two parties are not arbitrarily risk-averse, is necessarily a subset of the no-arbitrage interval. While the no-arbitrage price interval is always guaranteed to be non-empty, this is not necessarily the case for the FPI. An empty FPI captures the existence of a price for which both the writer and buyer end up being exposed to a negative risk thus making the ERP paradigm less relevant.

Note also that both the equal risk price and fair price interval can only be measured if the risk measures ρ^w and ρ^b are known. In practice, this might require both parties involved to provide supporting evidence for their respective choice of risk measure in the form of historical decisions that were taken using such measures. In the rest of the article, we make the assumption that the true risk measures are known by each party.

1.2.2 Equal Risk Pricing with Convex Risk Measures

Since the work of Artzner et al. (1999), it is now common to define coherent risk measures as risk measures that satisfy the following properties, where X and Z represent two random liabilities:

- Monotonicity: if $X \leq Z \quad a.s.$ then $\rho(X) \leq \rho(Z)$
- Subadditivity: $\rho(X + Z) \leq \rho(X) + \rho(Z)$
- Positive homogeneity: If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda\rho(X)$
- Translation invariance: If $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) + m$
- Normalized risk: $\rho(0) = 0$.

The first property naturally applies because if at any possible state that may happen the amount of liability X is less than the liability Z, then the risk of X is less than the risk of Z. The second property specifies that diversification does not increase the risk, and may decrease it. The third property implies that the risk of a position is linearly proportional to its size. Finally, the fourth property implies that the addition of a sure amount to a random liability will decrease the risk by the same amount. By relaxing positive homogeneity and subadditivity with the following convexity property, the family of risk measures becomes known as the larger family of “convex risk measures" (Föllmer and Schied (2011)):

- Convexity: $\rho(\lambda X + (1 - \lambda)Z) \leq \lambda\rho(X) + (1 - \lambda)\rho(Z)$, for $0 \leq \lambda \leq 1$.

Without loss of generality, in order to ensure that an equal risk price exists, we impose that participants are unable to design self-financing hedging strategies that reach arbitrarily low risks.

Assumption 1.2.1. *We assume that the risk measures, ρ^w and ρ^b satisfy a "bounded market risk" assumption, i.e.*

$$0 \geq \inf_{X \in \mathcal{X}(0)} \rho^w(-X_T) > -\infty, \quad 0 \geq \inf_{X \in \mathcal{X}(0)} \rho^b(-X_T) > -\infty\,.$$

In particular, if the risk measures are coherent, then this assumption implies that[2]

$$\inf_{X \in \mathcal{X}(0)} \rho^w(-X_T) = 0, \quad \inf_{X \in \mathcal{X}(0)} \rho^b(-X_T) = 0\,.$$

Note that this assumption was also made in Xu (2006) (see Assumption 2.3) and reflects the fact that a participant believes that he cannot make an arbitrarily large risk-adjusted profit from trading in this market. We argue that in the context of equal risk price, it is made without loss of generality since if either risk measure violates the condition, then one should simply conclude that there exists no equal risk price as defined in Definition 1. This is due to the fact that for all p_0, we would have for example that

$$\begin{aligned}\varrho^w(p_0) &= \inf_{X \in \mathcal{X}(p_0)} \rho^w(F(S_T, Y_T) - X_T) \leq \inf_{X \in \mathcal{X}(p_0)} (1/2)\rho^w(2F(S_T, Y_T)) + (1/2)\rho^w(-2X_T) \\ &= (1/2)\rho^w(2F(S_T, Y_T)) - p_0 + \inf_{X \in \mathcal{X}(0)} (1/2)\rho^w(-2X_T) = -\infty \notin \mathbb{R}\,.\end{aligned}$$

An interesting conclusion can be drawn regarding the relation between the equal risk price and the fair price interval when both risk measures are convex risk measures.

Proposition 1.2.1. *Given that both ρ^w and ρ^b are convex risk measures, an equal risk price exists if and only if the fair price interval is bounded. Moreover, when it exists it is equal to:*

$$p_0^* = (\varrho^w(0) - \varrho^b(0))/2\,,$$

which is the center of the fair price interval if the latter is non-empty.

Based on the Proposition 1.2.1, when using convex risk measures, the equal risk price can simply be found by evaluating the two boundaries of the fair price interval.

Following up on an important concern raised about the ϵ-arbitrage pricing approach, based on a result from Xu (2006), we can actually confirm the fact that for convex risk measures that satisfy the bounded market risk property, the equal risk price is arbitrage-free under weak conditions.

[2]Indeed, for a coherent risk measure, we have that $\inf_{X \in \mathcal{X}(0)} \rho^w(-X_T) < 0$ implies that $\inf_{X \in \mathcal{X}(0)} \rho^w(-X_T) = -\infty$ because of positive homogeneity.

Lemma 1.2.2. *If the fair price interval exists and is non-empty and both ρ^w and ρ^b are convex risk measures, then the equal risk price lies in the no-arbitrage price interval.*

In what follows, we will show how the result of Proposition 1.2.1 can be further exploited to identify the equal risk price of both European and American-style options using dynamic programming in a context where hedging is implemented at discrete-time points.

Remark 1. We should note here that in the case where the risk measures do not satisfy the translation invariance property, one can still exploit the above observation that the equal risk price falls within the fair price interval and is therefore arbitrage-free assuming non-emptiness of this interval. Namely, if such a price exists, one can identify it by employing a bisection algorithm that can establish $\Delta(p_0) := \varrho^w(p_0) - \varrho^b(p_0) = 0$. The convergence of a bisection method can rely on the fact that $\Delta(p_0)$ is non-increasing and that it is greater or equal to zero at p_0^b and lower or equal to zero at p_0^w. Finally, some guidance regarding the derivation of dynamic programming equations for this more general context can be found in Appendix 1.6.5.

1.3 Discrete Dynamic Formulations for Equal Risk Pricing Framework

In contexts where trading can only occur at specific periods of time $\{t_k\}_{k=0}^{K-1} \subset [0, T[$, one typically redefines the set of all admissible self-financing hedging strategies in terms of the wealth accumulated at each period:

$$\bar{\mathcal{X}}(p_0) = \left\{ X : \Omega \to \mathbb{R}^K \,\middle|\, \exists \{\xi_k\}_{k=0}^{K-1}, \quad X_k = p_0 + \sum_{k'=0}^{k-1} \xi_{k'} \Delta S_{k'+1}, \quad \forall k = 1, \ldots, K \right\},$$

where $\Delta S_{k+1} = S_{t_{k+1}} - S_{t_k}$ and where, for each $k = 0, \ldots, K-1$, the hedging strategy ξ_k is a $\mathcal{F}_{t_k}$-measureable random variable and captures the number of shares of the risky assets held in the portfolio during the period $[t_k,\ t_{k+1}]$. Finally, we assume that all random variables of interest in the discrete hedging problem are well behaved, and for simplicity will refer to $\mathcal{F}_{t_k}$ as $\mathcal{F}_k$.

Assumption 1.3.1. *There exists some* $p \in [1, \infty]$ *such that all* $X \in \bar{\mathcal{X}}(p_0)$ *is such that for all* $k = 1, \ldots, K$ *we have that* $X_k \in \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$ *and that the payoff function* $F(S_T, Y_T) \in \mathcal{L}_p(\Omega, \mathcal{F}_T, \mathbb{P})$.

In particular, this assumption allows us to make use of a decomposability property of risk measures, which as shown in Ruszczynski and Shapiro (2006); Ruszczyński (2010); Pichler and Shapiro (2018) is a key concept for producing a dynamic formulation for problems (1.1a) and (1.1b).

Definition 3. (One-step decomposable risk measures) The measure $\rho : \mathcal{L}_p(\Omega, \mathcal{F}_T, \mathbb{P}) \to \mathbb{R}$ is "one-step decomposable" if there exists a set of risk measures $\{\rho_k\}_{k=0}^{K-1}$ such that $\rho(X) = \rho_0(\rho_1(\cdots \rho_{K-2}(\rho_{K-1}(X)) \cdots)$ and where each measure $\rho_k : \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P}) \to \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$ is a conditional risk mapping (as defined in Ruszczynski and Shapiro (2006)), i.e. it satisfies the following properties:

- Conditional convexity : $\forall\, \theta \in [0,\, 1]$, $\forall\, X, Y \in \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P})$, $\rho_k(\theta Y + (1-\theta)X) \leq \theta\rho_k(Y) + (1-\theta)\rho_k(X)$ a. s.
- Conditional monotonicity : $\forall\, X, Y \in \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P})$, $Y \geq X$ a. s. $\Rightarrow$ $\rho_k(Y) \geq \rho_k(X)$ a. s.
- Conditional translation invariance : $\forall\, X \in \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$, $Y \in \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P})$, $\rho_k(X + Y) = X + \rho_k(Y)$ a. s.

Additionally, a coherent risk measure is said to be "one-step coherently decomposable" if each measure ρ_k also satisfies

- Conditional scale invariance : $\forall\, \alpha \geq 0$, $\forall\, X \in \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P})$, $\rho_k(\alpha X) = \alpha\rho_k(X)$ a. s.

Among all risk measures that are one-step decomposable, a special class of risk measures can be shown to be especially attractive from a computational point of view. We will refer to these measures as Markovian risk measures which can be used when the filtered measurable space $(\Omega, \mathcal{F})$ is a progressively revealed product space.

Definition 4. The filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is said to be supported on a progressively revealed product space if there exists a sequence $\{(\Omega_k, \Sigma_k)\}_{k=1}^{K}$ such that

$(\Omega, \mathcal{F})$ is the product space, i.e. $\Omega := \times_{k=1}^{K} \Omega_k$ and $\mathcal{F} := \otimes_{k=1}^{K} \Sigma_k$, and $\mathbb{F}$ is the natural filtration in this space, i.e. $\mathcal{F}_k := \sigma(\pi_{k'} : k' \leq k)$, where $\pi_k(\omega) := \omega_k$.

With this definition in hand, we are now ready to define the class of Markovian risk measures. We note that a similar class of risk measure was proposed in Ruszczyński (2010) in the context of a Markov decision processes. We, however, simplify the definition by exploiting the fact that conditional risk mappings are unaffected by decisions.

Definition 5. (Markovian risk measure) Given that $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is supported on a progressively revealed product space defined through some $\{(\Omega_k, \Sigma_k)\}_{k=1}^{K}$, a one-step decomposable risk measure is said to be Markovian if there exists an $\mathbb{F}$ measurable stochastic process $\theta_k : \Omega \to \mathbb{R}^m$, with $k = 0, \ldots, K$, that follows some dynamics $\theta_{k+1}(\omega) = f(\theta_k(\omega), \omega_k)$ for some $f : \mathbb{R}^m \times \Omega_k \to \mathbb{R}^m$, and some $\bar{\rho}_k : \mathcal{L}_p(\Omega_{k+1}, \Sigma_{k+1}, \mathbb{P}_{k+1}) \times \mathbb{R}^m \to \mathbb{R}$, with $\mathbb{P}_{k+1}$ the marginalization of $\mathbb{P}$ on Ω_{k+1} such that:

$$\rho_k(X, \omega) = \bar{\rho}_k(\bar{\Pi}_k(X, \omega), \theta_k(\omega)),$$

where $\bar{\Pi}_k(X, \omega)$ is a random variable in $\mathcal{L}_p(\Omega_{k+1}, \Sigma_{k+1}, \mathbb{P}_{k+1})$ defined as $\bar{\Pi}_k(X, \omega, \bar{\omega}_{k+1}) = X(\omega_{1:k}, \bar{\omega}_{k+1}, \omega_{k+1:K})$.

Finally, given that the decomposable risk measure in this paper will be used in an arbitrage-free financial market, one can formulate an assumption that imposes the bounded market assumption 1.2.1 on each conditional risk mapping.

Assumption 1.3.2. *(Bounded conditional market risk) A one-step decomposable risk measure ρ is said to express "bounded conditional market risk" if each conditional risk mapping ρ_k satisfies the following properties:*

$$0 \geq \inf_{\xi_k, \ldots, \xi_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \xi_\ell \Delta S_{\ell+1}) > -\infty \ a.\ s.,$$

where $\rho_{k,K}(X) := \rho_k(\rho_{k+1}(\cdots \rho_{K-1}(X) \cdots)$. Furthermore, if it is conditionally scale invariant then $\inf_{\xi_k, \ldots, \xi_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \xi_\ell \Delta S_{\ell+1}) = 0$.

In what follows, we derive dynamic equations that can be used to compute the equal risk price of European and American-style options in discrete-time trading. We further

exploit the translation invariance and Markovian properties to reduce the dimension of the state space required to formulate the Bellman equations. We also conclude this section with an example of such equations when employing a recursive conditional value-at-risk measure.

1.3.1 European Style Options

In order to evaluate the equal risk price of options of the form of $F(S_T, Y_T)$ in discrete-time with convex risk measures, as described in Proposition 1.2.1 one should solve problems (1.1a) and (1.1b) under the feasible set of strategies $\bar{\mathcal{X}}(0)$. Interestingly, the work of Pichler and Shapiro (2018) provide simple arguments for deriving useful dynamic equations in contexts where the risk measures are one-step decomposable risk measures.

Proposition 1.3.1. *Given that ρ^w and ρ^b are one-step decomposable risk measures as defined in Definition 3, then $\varrho^w(0) = V_0^w$ and $\varrho^b(0) = V_0^b$, where each V_k^w and V_k^b for $k = 0, \ldots, K$ are defined as follow:*

Writer's model:

$$V_k^w(\omega) := \inf_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w, \omega)\,,\ k = 0, \ldots, K-1 \tag{1.5a}$$

$$V_K^w(\omega) := F(S_K(\omega), Y_K(\omega))\,. \tag{1.5b}$$

Buyer's model:

$$V_k^b(\omega) := \inf_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b, \omega)\,,\ k = 0, \ldots, K-1 \tag{1.6a}$$

$$V_K^b(\omega) := -F(S_K(\omega), Y_K(\omega))\,, \tag{1.6b}$$

and assuming that each $V_k^w \in \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$ and $V_k^b \in \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$. Furthermore, the minimal risk hedging policy for both the writer and the buyer can be described respectively as:

$$\xi_k^{w*}(\omega) \in \arg\min_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w, \omega),\ \forall\, k = 1, \ldots, K-1$$

$$\xi_k^{b*}(\omega) \in \arg\min_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b, \omega),\ \forall\, k = 1, \ldots, K-1\,.$$

We then get that if the filtered probability space is supported on a progressively revealed product space, if both ΔS_k and $\Delta Y_k := Y_k - Y_{k-1}$ are measurable on Σ_k, such that they co-exist in both $\mathcal{L}_p(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{L}_p(\Omega_k, \Sigma_k, \mathbb{P}_k)$, and if both ρ^w and ρ^b satisfy the Markovian assumption with respect to θ^w and θ^b respectively, then we can derive finite-dimensional Bellman equations that allow us to compute the equal risk price. These can be defined as follows:

$$\tilde{V}_K^w(S_K, Y_K, \theta_K^w) := F(S_K, Y_K)\,,$$

and recursively

$$\tilde{V}_k^w(S_k, Y_k, \theta_k^w) := \inf_{\xi_k} \bar{\rho}_k(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(S_k + \Delta S_{k+1}, Y_k + \Delta Y_{k+1}, f_k(\theta_k^w)), \theta_k^w)\,,$$

where each $\tilde{f}_k(\theta_k^w)$ can be considered a random variable in $\mathcal{L}_p(\Omega_{k+1}, \Sigma_{k+1}, \mathbb{P}_{k+1})$. These equations have the property that:

$$V_K^w(\omega) = \tilde{V}_K^w(S_K(\omega), Y_K(\omega), \theta_K^w(\omega))\,,$$

and recursively that if $V_{k+1}^w(\omega) = \tilde{V}_{k+1}^w(S_{k+1}(\omega), Y_{k+1}(\omega), \theta_{k+1}^w(\omega))$, then we have that:

$$\begin{aligned}
V_k^w(\omega) &= \inf_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w, \omega) \\
&= \inf_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(S_{k+1}, Y_{k+1}, \theta_{k+1}^w), \omega) \\
&= \inf_{\xi_k} \bar{\rho}_k^w(\bar{\Pi}_k(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(S_{k+1}, Y_{k+1}, \theta_{k+1}^w)), \theta_k^w(\omega)), \\
&= \inf_{\xi_k} \bar{\rho}_k^w(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(S_k(\omega) + \Delta S_{k+1}, Y_k(\omega) + \Delta Y_{k+1}, f(\theta_k^w)), \theta_k^w(\omega)) \\
&= \tilde{V}_k(S_k(\omega), Y_k(\omega), \theta_k(\omega))\,.
\end{aligned}$$

From these derivations we see that $\varrho^w(0) = V_0^w = \tilde{V}_0^w(S_0, Y_0, \theta_0^w)$. In the case of the buyer, similar derivations lead to the Bellman equations:

$$\tilde{V}_K^b(S_K, Y_K, \theta_K^b) := -F(S_K, Y_K)$$

and

$$\tilde{V}_k^b(S_k, Y_k, \theta_k^b) := \inf_{\xi_k} \bar{\rho}_k^b(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^b(S_k + \Delta S_{k+1}, Y_k + \Delta Y_{k+1}, f_k(\theta_k^b)), \theta_k^b)\,,$$

which can be used to compute $\varrho^b(0) = V_0^b = \tilde{V}_0^b(S_0, Y_0, \theta_0^b)$. We can, therefore, conclude that $p_0^* = (\tilde{V}_0^w(S_0, Y_0, \theta_0^w) - \tilde{V}_0^b(S_0, Y_0, \theta_0^b))/2$.

1.3.2 American-style Option

Contrary to European options, the exercise time of American options is flexible and up to the buyer's decision. Therefore, in the equal risk model we need to consider the interaction between an optimal exercise time and one-step decomposable risk measures. Similarly as in Pichler and Shapiro (2018), we will define the exercise time as a "stopping time" with respect to the filtration $\mathbb{F}$, i.e. that it is a random variable $\tau : \Omega \to \{0, \ldots, K\}$, such that $\{\omega : \tau(\omega) = t\} \in \mathcal{F}_t$, for all $\forall t = \{0, ..., T\}$. Considering that the option payoff is now $F_t(S_t, Y_t) \in \mathcal{L}_p(\Omega, \mathcal{F}_t, \mathbb{P})$ at time t if and only if $\tau = t$, we let $F_\tau(S_\tau, Y_\tau)$ capture the new payoff function which is defined as follows:

$$F_\tau(S_\tau, Y_\tau) := \sum_{t=0}^{T} \mathbb{I}_{\{\tau=t\}} F_t(S_t, Y_t)\,,$$

where $\mathbb{I}_{\{\tau=t\}}$ is the indicator function, which is one for $\tau = t$, and zero otherwise. We also redefine the set of self-financing hedging strategy to make its relation to τ explicit:

$$\bar{\mathcal{X}}_\tau(p_0) := \left\{ X^\tau : \mathcal{T} \times \Omega \to \mathbb{R}^K \;\middle|\; \begin{array}{l} \exists X_0^\tau = p_0, \forall k = 1, \ldots, K-1, \exists \xi_k, \{\hat{\xi}_k^i\}_{i=0}^k \\ X_{k+1}^\tau(\tau) = X_k^\tau(\tau) + (\xi_k \mathbf{1}\{\tau > k\} + \sum_{i=0}^k \hat{\xi}_k^i \mathbf{1}\{\tau = i\}) \Delta S_{k+1},\ \forall \tau \in \mathcal{T} \end{array} \right\},$$

where $\mathcal{T}$ is the set of all exercise time process, and where each ξ_k and $\hat{\xi}_k^i$ are $\mathcal{F}_k$-measurable. Specifically, ξ_k models the hedging strategy that is implemented at time k when exercise has not occurred yet while $\hat{\xi}_k^i$ models the hedging strategy that is implemented at time k when exercise occurred in period $k' = i$.

Remark 2. We need to emphasize the fact that in most of the recent literature, hedging is considered to stop once the option is exercised. We intentionally omit making this assumption up front and choose to model the possibility of hedging for both the writer and the buyer throughout the horizon. We will later show that when ρ^w and ρ^b are one-step coherently decomposable, the two approaches become equivalent, i.e. one can consider that $\hat{\xi}_k^i = 0$ for all $k = 0, \ldots, T-1$ and all $i = 0, \ldots, k$ (see Section 1.3.2 for further discussion). In cases where the assumption does not hold, we consider important to model hedging beyond exercise time in the buyer problem in order to avoid having incentives to delay exercise time simply to be able to benefit from later market opportunities. Similarly, in the writer's problem, if hedging stops at exercise time, the worst-case exercise time policy could be biased towards zero in order to prevent the writer from benefiting from later market

conditions.We also note that in any case, the analysis that follows can straightforwardly be adapted to a definition of the set of self-financing hedging strategies that explicitly enforces no hedging beyond exercise time.

In this context, the definition of the equal risk framework needs to be adapted to account for the presence of τ. In what follows, we will consider two formulations.

Definition 6. ERP with Commitment. Given that both the writer's risk measure, ρ^w, and buyer's risk measure ρ^b are interpretable as certainty equivalents and are strictly monotone with respect to certain amounts, then the equal risk price with commitment, when it exists, is defined as the unique p_0^* for which there exists a stopping time policy τ^* that satisfies:

$$\varrho^w(p_0^*, \tau^*) = \varrho^b(p_0^*, \tau^*) \in \mathbb{R} \qquad \& \qquad \tau^* \in \arg\min_{\tau} \varrho^b(p_0^*, \tau)\,,$$

where

$$\varrho^w(p_0, \tau) = \inf_{X^\tau \in \bar{\mathcal{X}}_\tau(p_0)} \rho^w(F(S_\tau, Y_\tau) - X_K^\tau(\tau)) \tag{1.7}$$

$$\varrho^b(p_0, \tau) = \inf_{X^\tau \in \bar{\mathcal{X}}_\tau(-p_0)} \rho^b(-F(S_\tau, Y_\tau) - X_K^\tau(\tau))\,. \tag{1.8}$$

In simple terms, Definition 6 reflects the assumption that the buyer of the option commits to following a risk-minimizing exercise strategy at the moment of purchasing the option. With this information, the writer can be more effective in hedging the option which, as will be shown, has the effect of giving rise to a lower equal risk price then when no commitment is made by the buyer. While we note that in practice, it might not be interesting for a buyer to commit up front to an exercise strategy, the notion of ERP with commitment can serve the purpose of assessing the "cost of non-commitment", which is a concept that is unique to the pricing of American options in an incomplete markets (because of the multiplicity of arbitrage-free prices) and which can help interpreting the ERP without commitment.

Definition 7. ERP without Commitment. Given that both the writer's risk measure, ρ^w, and buyer's risk measure ρ^b are interpretable as certainty equivalents and are strictly monotone with respect to certain amounts, then the equal risk price without commitment,

when it exists, is defined as the unique p_0^* that satisfies:

$$\varrho_\tau^w(p_0^*) = \varrho_\tau^b(p_0^*) \in \mathbb{R}\,,$$

where

$$\varrho_\tau^w(p_0) = \inf_{X^\tau \in \bar{\mathcal{X}}_\tau(p_0)} \sup_\tau \rho^w(F(S_\tau, Y_\tau) - X_K^\tau(\tau)) \tag{1.9}$$

$$\varrho_\tau^b(p_0) = \inf_{X^\tau \in \bar{\mathcal{X}}_\tau(-p_0)} \inf_\tau \rho^b(-F(S_\tau, Y_\tau) - X_K^\tau(\tau))\,. \tag{1.10}$$

Note that in this definition, the writer is unaware of the exercise strategy that will be employed by the buyer. He, therefore, considers the minimal risk of entering into this contract agreement as being the risk achieved by following an optimal hedging strategy that is adapted to both the filtration and the information about τ that is progressively revealed.

We now demonstrate how the equal risk price necessarily increases when passing from the "with commitment" to "without commitment" framework.

Lemma 1.3.2. *Given any American type option, the ERP with commitment p_c^* is always smaller or equal to the ERP without commitment p_{nc}^*.*

In what follows, we derive dynamic programming equations that can be used to compute the ERP in both type of settings. Section 1.3.2 then exploits these equations to establish that when the risk measures are coherently decomposable, risk cannot be reduced by hedging beyond the exercise time.

Bellman Equations for Equal Risk Price with Commitment

We start with a simple lemma that extends the result of Proposition 1.2.1 to the context of an American option with commitment.

Lemma 1.3.3. *Given that both ρ^w and ρ^b are convex risk measures, an equal risk price exists if and only if the fair price interval defined as:*

$$[-\varrho^b(0,\tau_0),\ \varrho^w(0,\tau_0)]\,,$$

where $\tau_0 \in \arg\min_\tau \varrho^b(0,\tau)$, is bounded. Moreover, when it exists it is equal to the center of this interval, which can be calculated as:

$$p_0^* := (\varrho^w(0,\tau_0) - \varrho^b(0,\tau_0))/2\,.$$

Lemma 1.3.3 indicates that, to evaluate the ERP, one needs to be able to compute $\varrho^w(0,\tau)$ and $\varrho^b(0,\tau)$ for any fixed exercise policy τ, and to identify a procedure that can solve the optimal exercise time problem: $\min_\tau \varrho^b(0,\tau)$. As for the case of European options, all these elements can be characterized using dynamic programming equations.

Proposition 1.3.4. *Given that ρ^w and ρ^b are one-step decomposable risk measures as defined in Definition 3, then $\varrho^w(0,\tau_0) = V_0^w(\tau_0)$ and $\varrho^b(0,\tau_0) = V_0^b(0)$, where for any exercise strategy τ, each $V_k^w(\tau)$, $V_k^b(0)$, and $V_k^b(1)$ for $k = 0,\ldots,K$ are defined as follow:*

Writer's model:

$$\begin{aligned} V_k^w(\tau,\omega) &:= \inf_{\xi_k} \rho_k^w(V_{k+1}^w(\tau) - \xi_k \Delta S_{k+1}, \omega) + \mathbf{1}\{\tau(\omega) = k\} F(S_k(\omega), Y_k(\omega)) \\ V_K^w(\tau,\omega) &:= \mathbf{1}\{\tau(\omega) = K\} F(S_K(\omega), Y_K(\omega))\,. \end{aligned} \tag{1.11}$$

Buyer's model:

$$V_k^b(1,\omega) := \inf_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b(1), \omega) \tag{1.12}$$

$$V_k^b(0,\omega) := \min(V_k^b(1,\omega) - F(S_k(\omega), Y_k(\omega)),\ \inf_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b(0), \omega)) \tag{1.13}$$

$$V_K^b(\bar{Z}_K,\omega) := -(1 - \bar{Z}_K(\omega)) F(S_K(\omega), Y_K(\omega))\,. \tag{1.14}$$

and assuming that each $V_k^w(\tau)$, $V_k^b(0)$, and $V_k^b(1)$ are in $\mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$. Furthermore, a feasible candidate for τ_0 can be found using

$$\tau_0(\omega) = \min\{k = 0,\ldots,K | V_k^b(0,\omega) = V_k^b(1,\omega) - F(S_k(\omega), Y_k(\omega))\}\,. \tag{1.15}$$

Finally, given that the option is sold at the equal risk price $p_0^* = (V_0^w(\tau_0) - V_0^b(0))/2$ based on an exercise strategy τ_0, the minimal risk hedging policy for both the writer and the buyer can be described respectively as:

$$\hat{\xi}_k^{i*}(\tau,\omega) \equiv \xi_k^*(\tau,\omega) \in \arg\min_{\xi_k} \rho_k^w(V_{k+1}^w(\tau) - \xi_k \Delta S_{k+1}, \omega), \quad \begin{matrix} \forall\, k = 1,\ldots,K-1 \\ \forall\, i = 0,\ldots,k, \end{matrix}$$

for the writer, and

$$\hat{\xi}_k^{i*}(\omega) \equiv$$

$$\xi_k^*(\omega) \in \arg\min_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b(\mathbf{1}\{\tau_0(\omega) \le k\}), \omega), \quad \begin{matrix} \forall\, k = 1,\ldots,K-1 \\ \forall\, i = 0,\ldots,k, \end{matrix} \tag{1.16}$$

for the buyer.

In order for the evaluation of $\varrho^w(0,\tau_0)$ and $\varrho^w(0,\tau_0)$ to be computable numerically, it becomes essential to identifying Bellman equations on a finite-dimensional state space. When the Markovian assumption holds for both ρ^w and ρ^b with respect to some process θ_k, these can be derived as follows. For the buyer's problem, we have that:

$$\begin{aligned}
&\tilde{V}_k^b(1,S_k,Y_k,\theta_k) := \inf_{\xi_k} \bar{\rho}_k^b(-\xi_k\Delta S_{k+1} + \tilde{V}_{k+1}^b(1,S_k+\Delta S_{k+1},Y_k+\Delta Y_{k+1},f(\theta_k)),\theta_k)\\
&\tilde{V}_k^b(0,S_k,Y_k,\theta_k) := \min(\tilde{V}_k^b(1,S_k,Y_k,\theta_k) - F(S_K,Y_K),\\
&\qquad \inf_{\xi_k} \bar{\rho}_k^b(-\xi_k\Delta S_{k+1} + \tilde{V}_{k+1}^b(0,S_k+\Delta S_{k+1},Y_k+\Delta Y_{k+1},f(\theta_k)),\theta_k))\\
&\tilde{V}_K^b(\bar{Z}_K,S_k,Y_k,\theta_k) := -(1-\bar{Z}_K)F(S_K,Y_K)\,.
\end{aligned} \tag{1.17}$$

In order to obtain an optimal exercise policy, one can first observe that with

$$\begin{aligned}
\tau_0(\omega) = \min\Big\{k=0,\dots,K | &\tilde{V}_k^b(0,S_k(\omega),Y_k(\omega),\theta_k(\omega))\\
&= \tilde{V}_k^b(1,S_k(\omega),Y_k(\omega),\theta_k(\omega)) - F(S_k(\omega),Y_k(\omega))\Big\}\,.
\end{aligned}$$

Yet, when letting $Z_k := \mathbf{1}\{\tau_0 = k\}$ and $\bar{Z}_k := \mathbf{1}\{\tau_0 < k\}$, we can define:

$$g_k(\bar{Z}_k,S_k,Y_k,\theta_k) := \mathbf{1}\{(\bar{Z}_k=0)\ \&\ (\tilde{V}_k^b(0,S_k,Y_k,\theta_k) = \tilde{V}_k^b(1,S_k,Y_k,\theta_k) - F(S_K,Y_K))\}$$

so that

$$Z_k(\omega) = g_k(\bar{Z}_k(\omega),S_k(\omega),Y_k(\omega),\theta_k(\omega))\,,$$

and

$$\bar{Z}_{k+1}(\omega) = \bar{Z}_k(\omega) + g_k(\bar{Z}_k(\omega),S_k(\omega),Y_k(\omega),\theta_k(\omega))\,.$$

This implies that $\tau_0(\omega) = \sum_{k=0}^K g_k(\bar{Z}_k(\omega),S_k(\omega),Y_k(\omega),\theta_k(\omega))$ which can be implemented by exploiting the Bellman equations. We can then proceed with describing the reduced equations for the writer's problem:

$$\begin{aligned}
&\tilde{V}_k^w(\bar{Z}_k,S_k,Y_k,\theta_k) :=\\
&\quad \inf_{\xi_k} \bar{\rho}_k^w(-\xi_k\Delta S_{k+1} + \tilde{V}_{k+1}^w(\bar{Z}_k + g_k(\bar{Z}_k,S_k,Y_k,\theta_k),S_k+\Delta S_{k+1},Y_k+\Delta Y_{k+1},f_k(\theta_k)),\theta_k)\\
&\qquad\qquad + g_k(\bar{Z}_k,S_k,Y_k,\theta_k)F(S_k,Y_k)\\
&\tilde{V}_K^w(\bar{Z}_K,S_K,Y_K,\theta_K) := g_K(\bar{Z}_K,S_K,Y_K,\theta_K)F(S_K,Y_K)\,,
\end{aligned}$$

so that

$$V_k^w(\tau_0, \omega) = \tilde{V}_k^w(\bar{Z}(\omega), S_k(\omega), Y_k(\omega), \theta_k(\omega)) .$$

Bellman Equations for Equal Risk Price without Commitment

In the context of a contract where the buyer does not commit to a specific exercise policy, Proposition 1.2.1 extends in a straightforward manner. Nonetheless, for completeness, we provide the details in the following lemma.

Lemma 1.3.5. *Given that both ρ^w and ρ^b are convex risk measures, an equal risk price exists if and only if the fair price interval defined as $[-\varrho_\tau^b(0), \varrho_\tau^w(0)]$, is bounded. Moreover, when it exists it is equal to the center of this interval which can be calculated as $p_0^* := (\varrho_\tau^w(0) - \varrho_\tau^b(0))/2$.*

The main difference between this case and the case with commitment is that in order to compute $\varrho_\tau^w(0)$, now there is a need to further determine the worst-possible exercise policy that the writer would hedge against. Since the whole hedging problem for the writer now takes the form of a minimax optimization problem, additional care has to be taken to ensure the decisions of hedging and exercising (the options) are executed in the right order when formulating the Bellman equations. In particular, we proceed by fixing first the hedging decisions and identifying recursive equations that solve the worst-case exercise time problem (Pichler and Shapiro (2018)). We then use the arguments based on the interchangeability principle in dynamic programming (see Pichler and Shapiro (2018)) to establish that the hedging decisions that minimize the recursive equations globally can be obtained from decisions that minimize the recursive equations stage-wise. The details can be found in the appendix and this leads to the following dynamic programming equations. On the other hand, it is not hard to confirm that the computation of $\varrho_\tau^b(0)$ for the buyer coincides with the computation required in the case of commitment.

Proposition 1.3.6. *Given that ρ^w and ρ^b are one-step decomposable risk measures as defined in Definition 3, then $\varrho_\tau^w(0) = V_0^w(0)$ and $\varrho^b(0) = V_0^b(0)$, each $V_k^w(0)$ and $V_k^w(1)$ for $k = 0, \ldots, K$ are defined as follow:*

Writer's model:

$$V_k^w(1,\omega) := \inf_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w(1),\omega) \tag{1.18}$$

$$V_k^w(0,\omega) := \max(V_k^w(1,\omega) + F(S_k(\omega), Y_k(\omega)),\ \inf_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w(0),\omega)) \tag{1.19}$$

$$V_K^w(\bar{Z}_K,\omega) := (1-\bar{Z}_K(\omega))F(S_K(\omega), Y_K(\omega))\,, \tag{1.20}$$

while $V_k^b(0)$ and $V_k^b(1)$ are defined as in equations (1.12)-(1.14) *and assuming that each $V_k^w(0)$, $V_k^w(1)$, $V_k^b(0)$, and $V_k^b(1)$ are in $\mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$. Furthermore, given that the option is sold at the equal risk price $p_0^* = (V_0^w(0) - V_0^b(0))/2$, a minimal risk hedging policy for the writer can be described as:*

$$\hat{\xi}_k^{i*}(\bar{\omega}) \equiv \xi_k^*(\omega) \in \arg\min_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w(\mathbf{1}\{\tau \le k\}),\omega), \quad \begin{matrix} \forall\, k = 1,\ldots,K-1 \\ \forall\, i = 0,\ldots,k\,, \end{matrix}$$

where τ is the observed exercise strategy. In the case of the buyer, a risk minimizing hedging strategy is as in equation (1.15)*, while a risk minimizing exercise strategy can be found using equation* (1.16)*.*

When the Markovian assumption holds with respect to some process θ_k, we can again derive finite-dimensional equations. In particular, for the buyer's problem, these are exactly as presented in equations (1.17). On the other hand, for the writer's problem, we have that:

$$\begin{aligned}
\tilde{V}_k^w(1, S_k, Y_k, \theta_k) &:= \inf_{\xi_k} \bar{\rho}_k^w(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(1, S_k + \Delta S_{k+1}, Y_k + \Delta Y_{k+1}, f(\theta_k), \theta_k) \\
\tilde{V}_k^w(0, S_k, Y_k, \theta_k) &:= \max(\tilde{V}_k^w(1, S_k, Y_k, \theta_k) + F(S_K, Y_K), \\
&\qquad \inf_{\xi_k} \bar{\rho}_k^w(-\xi_k \Delta S_{k+1} + \tilde{V}_{k+1}^w(0, S_k + \Delta S_{k+1}, Y_k + \Delta Y_{k+1}, f(\theta_k), \theta_k)) \\
\tilde{V}_K^w(\bar{Z}_K, S_k, Y_k, \theta_k) &:= (1-\bar{Z}_K)F(S_K, Y_K)\,.
\end{aligned} \tag{1.21}$$

On the Value of Hedging Beyond the Exercise Time

As pointed out in the beginning of Section 3, our dynamic programming (DP) formulations of the hedging problem are more general in that they allow for the possibility of hedging after the exercise of the options. This in principle provides the opportunities for both the writer and buyer to seek further risk reduction. But at the same time it adds additional

complexity to the DP formulation, which becomes computationally more costly to solve than the DP that assumes no hedging after exercise of the options. In this section, we identify the condition under which hedging beyond the exercise time actually does not reduce risk. In particular, based on our general DP formulation, we find that it is actually optimal to stop hedging after the exercise time if the employed risk measure is coherent.

Corollary 1.3.7. *If ρ^w is one-step coherently decomposable, then it becomes optimal for the writer to terminate the hedging strategy at the exact moment that the American option is exercised. The same applies to the buyer.*

As detailed in Appendix 1.6.11, this observation is closely related to the assumption of bounded market risk, in which case there exists no risk reduction opportunity when measured according to a coherent risk measure. Since in this case hedging beyond exercise time adds no value, one can simply employ a DP formulation that assumes that hedging stops at the exercise time.

In the next section, we elaborate on a specific class of coherently decomposable risk measure, referred to as "worst-case risk measures". We further provide numerical evidence on the quality of prices obtained using such risk measure both from the point of view of risk exposure and fairness.

1.3.3 Recursive Conditional Value-at-Risk Example

In this section, we provide a specific example of the Markovian counterpart of a popular one-step decomposable risk measure. We demonstrate how our results can be applied to this risk measure so as to write the corresponding Bellman equations.

We start by assuming the stochastic processes S_k and Y_k admit the following recursive representation, which is common in many applications:

$$S_{k+1} = f(S_k, \epsilon_{k+1}),\ Y_{k+1} = g(Y_k, \epsilon_{k+1}),$$

for some $f : \mathbb{R} \times \mathbb{R}^{n_\epsilon} \to \mathbb{R}$ and $g : \mathbb{R}^{n_y} \times \mathbb{R}^{n_\epsilon} \to \mathbb{R}^{n_y}$, and $(\epsilon_1, ..., \epsilon_K)$ is a realization of the progressively revealed product space with $\Omega := \times_{k=1}^{K} \mathbb{R}^{n_\epsilon}$ and $\mathcal{F} := \otimes_{k=1}^{K} \mathcal{B}(\mathbb{R}^{n_\epsilon})$, where $\mathcal{B}(\mathbb{R}^{n_\epsilon})$ refers to the Borel σ-algebra, equipped with probability measure $\mathbb{P}$ and natural filtration $\mathbb{F}$.

Definition 8. (Recursive conditional value-at-risk) Given a random variable X and a process $\{\beta_k\}_{k=0}^{K-1}$ which is $\mathcal{F}_k$ measurable, i.e. $\beta_k := \mathbb{R}^{n_\epsilon^k} \to \mathbb{R}$, the recursive conditional value-at-risk measure ρ is a one-step decomposable risk measure obtained using a conditional value-at-risk measure, defined as

$$\rho_k(X, \epsilon_1, \ldots, \epsilon_K) = \inf_t \; t + \frac{1}{1 - \beta_k(\epsilon_1, \ldots, \epsilon_k)} \mathbb{E}\left[(X - t)^+ \middle| \epsilon_1, \ldots, \epsilon_k\right],$$

as the conditional risk mapping.

Note that the recursive conditional value-at-risk measure defined above only qualifies, in its general form, as a Markovian risk measure if one considers $\theta_k := [\epsilon_1^T \quad \ldots \quad \epsilon_k^T]^T$. This can quickly give rise to the curse of dimensionality when constructing and solving the associated DP formulation. To circumvent this issue, a common practice is to assume that the ϵ_k process satisfies the Markov property, i.e. $\mathbb{P}(\epsilon_{k+s} \in \mathcal{A} | \epsilon_1, \ldots, \epsilon_k) = \mathbb{P}(\epsilon_{k+s} \in \mathcal{A} | \epsilon_k)$ for all $s \geq 0$ and all $\mathcal{A} \in \mathcal{B}(\mathbb{R}^{n_\epsilon})$. One however also needs an additional assumption about the β_k process such that $\beta_k = h(\beta_{k-1}, \epsilon_k)$ for some $h : \mathbb{R} \times \mathbb{R}^{n_\epsilon} \to \mathbb{R}$, in order to satisfy the Markovian risk measure assumption under a process $\theta_k := [\beta_{k-1} \;\; \epsilon_{k-1}^T]^T$.

We can now summarize the Bellman equations that can be derived for the case of a recursive conditional value-at-risk that is Markovian with respect to θ_k by following the result and discussions in Section 1.3.1. Namely, the writer problem's Bellman equations for the case of European options can be written as follows:

$$\tilde{V}_0^w(S_0, Y_0) = \inf_{\xi, t} \; t + \frac{1}{1 - \beta_0} \mathbb{E}[(-\xi(f(S_0, \epsilon_1) - S_0) + \tilde{V}_1^w(f(S_0, \epsilon_1), g(Y_0, \epsilon_1), h(\beta_0, \epsilon_1), \epsilon_1) - t)^+],$$

$$\tilde{V}_k^w(S_k, Y_k, \beta_k, \epsilon_k) =$$

$$\inf_{\xi, t} \; t + \frac{1}{1 - \beta_k} \mathbb{E}[-\xi(f(S_k, \epsilon_{k+1}) - S_k) + \tilde{V}_{k+1}^w(f(S_k, \epsilon_{k+1}), g(Y_k, \epsilon_{k+1}), h(\beta_k, \epsilon_{k+1}), \epsilon_{k+1}) - t)^+ | \epsilon_k]$$

and $\tilde{V}_K^w(S_K, Y_K, \beta_K, \epsilon_K) = F(S_K, Y_K)$. Similar Bellman equations can be derived for the buyer and we omit them for brevity.

In the case of American option, we can follow the result and discussions in Section 1.3.2. to write down the following Bellman equations for the buyer:

$$\tilde{V}_k^b(0, S_k, Y_k, \beta_k, \epsilon_k) = \min\Bigg(- F(S_k, Y_k), \inf_{\xi, t} \; t + \frac{1}{1 - \beta_k} \mathbb{E}[(-\xi_k(f(S_k, \epsilon_{k+1}) - S_k) + \tilde{V}_{k+1}^b(0, f(S_k, \epsilon_{k+1}), g(Y_k, \epsilon_{k+1}), h(\beta_k, \epsilon_{k+1}), \epsilon_{k+1}) - t)^+ | \epsilon_k] \Bigg)$$

and

$$\tilde{V}_K^b(\bar{Z}_K, S_K, Y_K, \beta_K, \epsilon_K) = -(1 - \bar{Z}_K)F(S_K, Y_K),$$

where we exploited the fact that $\tilde{V}_k^b(1, S_k, Y_k, \beta_k\epsilon_k) = 0$ since the conditional value-at-risk (CVaR) is coherent and Corollary 1.3.7 applies. We omit the writer's equations for brevity.

The arguments used above can be employed for many other recursive risk measures as long as ρ_k is the conditional analog of a law-invariant coherent risk measure.

Example 1. *One obtains a recursive mean semi-deviation measure when using a conditional risk mapping ρ_k defined as $\rho_k(X) = \mathbb{E}[X|\mathcal{F}_k] + \kappa_k\mathbb{E}[((X - \mathbb{E}[X|\mathcal{F}_k])_+)^r|\mathcal{F}_k]^{\frac{1}{r}}$, where κ_k is $\mathcal{F}_k$-measurable.*

Example 2. *One obtains a recursive mean CVaR measure when using a conditional risk mapping ρ_k defined as $\rho_k(X) = \mathbb{E}[X|\mathcal{F}_k] + \kappa_k \left(\inf_t\, t + \frac{1}{1-\beta_k}\mathbb{E}[(X - t)^+|\mathcal{F}_k]\right)$, where $\kappa_k > 0$ and $\beta_k \in [0, 1)$ are $\mathcal{F}_k$-measurable.*

On the other hand, it is worth emphasizing that a one-step decomposable risk measure that is constructed based on the composition of law-invariant coherent risk measures as suggested above is not law-invariant unless the conditional mappings are expectation or worst-case risk measures (Shapiro, 2012). This motivates us, in the following section, to focus our numerical study on the latter class of risk measures.

1.4 Numerical Study with Worst-case Risk Measures

In this section, we provide necessary details of implementing the equal risk pricing model in the case where the risk measure takes the form of a worst-case risk measure. In particular, such a form of risk measures has been considered in the literature of robust optimization, which requires the specification of an uncertainty set U over which the worst-case loss is calculated. The ϵ-arbitrage pricing model mentioned earlier in the introduction is one example that employs an uncertainty set motivated by central limit theorem. While the ϵ-arbitrage pricing model does not distinguish between the writer's and the buyer's loss, the equal risk pricing model proposed in this paper does, and one of our goals in this section is to demonstrate numerically the strength of the equal risk pricing model over the

ϵ-arbitrage pricing model. We will also benchmark the equal risk pricing model against the Black–Scholes pricing model in the case of European option, and against the binomial pricing model in the case of American option.

To facilitate the comparisons between the aforementioned models, we start by considering a market of assets that are driven by a Geometric Brownian Motion (GBM). We assume that the asset returns can only be observed at a set of uniformly distributed time points on the interval $[0, T]$ such that each time point $t_k := kT/K$, $k = 1, ..., K$. Without loss of generality, we can write $S_{t_k} = S_0 \Pi_{l=1}^{k}(1 + r_k)$ to denote the dynamic of asset price given a random vector of observed returns taking values in $\mathbb{R}^K$ and an initial asset price S_0. In order to formalize worst-case risk measures over such a market, we consider an outcome space $\Omega := \mathbb{R}^K$ and an associated filtered probability space $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K), \bar{\mathbb{F}}, \bar{\mathbb{P}})$, where $\mathcal{B}(\mathbb{R}^K)$ is the Borel σ-algebra on $\mathbb{R}^K$, and $\bar{\mathbb{F}} := \{\sigma(r_{k'} : k' \leq k)\}$ is the natural filtration. We let $\bar{\mathbb{P}}$ be the probability measure that captures

$$(1 + r_k) \sim \text{i.i.d. Lognormal}(\mu T/K, \sigma^2 T/K),\ k = 1, ..., K\,,$$

where μ and σ are the statistics of the GBM per unit of time T. Note that this filtered probability space is supported on a progressively revealed product space as defined in Definition 4. For the sake of convenience, we reformulate the hedging decision problem in terms of how much money is invested in the risky asset at each time point, denoted by $\zeta_0, \ldots, \zeta_{K-1}$, instead of the number of shares of the risky assets, i.e. $\zeta_k = \xi_k S_{t_k}$. This leads to the following equation representing the evolution of wealth:

$$X_k = p_0 + \sum_{k'=0}^{k-1} \zeta_{k'} r_{k'+1}, \forall k = 1, ..., K.$$

In this numerical study, we will assume that the writer and buyer are employing a risk measure that is motivated by robust optimization. In particular, we will assume that they are concerned about the worst-case performance for realizations that arise in a predefined uncertainty set $\mathcal{U}$. We, therefore, define a worst-case risk measure on a random liability $X : \mathbb{R}^K \to \mathbb{R}$ as:

$$\rho(X) = \operatorname*{ess\,sup}_{\mathbb{U}(\mathcal{U})} X\,,$$

where $\mathcal{U} \subset]-1, \infty[^K$ is compact and regular closed, and where $\mathbb{U}(\mathcal{U})$ refers to the uniform distribution over $\mathcal{U}$. We simplify the following presentation by employing standard notation

from robust optimization with $\mathcal{U}$ as the so-called uncertainty set:

$$\rho(X) = \sup_{r \in \mathcal{U}} X(r)\,.$$

Clearly, this risk measure is necessarily monotone, translation invariant, and coherent. Moreover, it is also one-step decomposable using:

$$\rho_k(X, r) := \begin{cases} \sup_{r' \in \mathcal{U}: r'_{1:k} = r_{1:k}} X(r') & \text{if } \exists\, r' \in \mathcal{U},\ r'_{1:k} = r_{1:k} \\ X([r_{1:k}^T \;\; 0_{k+1:K}^T]^T) & \text{otherwise} \end{cases},$$

where $r_{1:k} \in \mathbb{R}^k$ refers to the first k-th first terms of r, and where $X([r_{1:k}^T \;\; 0_{k+1:K}^T]^T)$ is short for

$$\inf_{\epsilon > 0} \operatorname*{ess\,sup}_{r' \in]-1, \infty[^K : r'_{1:k} = r_{1:k}, \|r'_{k+1:K}\|_\infty \leq \epsilon} X\,.$$

Note that the conditional measure that is used for the case where $\nexists\, r' \in \mathcal{U},\ r'_{1:k} = r_{1:k}$ can be arbitrary if one is only interested in calculating $\varrho(0)$ given that the latter is unaffected by the level of loss when $r \notin \mathcal{U}$. In practice however, one might get a "better" hedging policy by employing a more risk-aware measure than $X([r_{1:k}^T \;\; 0_{k+1:K}^T]^T)$. Indeed, one can confirm that ρ can equivalently be described as:

$$\begin{aligned} \rho(X) &= \sup_{r^1 \in \mathcal{U}} \sup_{r^2 \in \mathcal{U}: r_1^1 = r_1^2} \cdots \sup_{r^K \in \mathcal{U}: r_{1:K-1}^{K-1} = r_{1:K-1}^K} X(r^K) \\ &= \rho_0(\rho_1(\cdots \rho_{K-1}(X) \cdots)\,. \end{aligned}$$

In many cases, the one-step decomposable risk measure ρ_k can be further shown to satisfy the Markov property, e.g. with the uncertainty sets presented in the following sections. One can then follow the discussion in the Section 1.3 to write down the dynamic programming equations for both cases of European and American options.

In all of our experiments, we consider an option with maturity $T = 1$ (year) that is written over an asset with $\mu = 0.0718$ (annualized mean), $\sigma = 0.1283$ (annualized volatility), and with an initial price $S_0 = 1000$. Our choices of values for μ and σ come from François et al. (2014) and were calibrated on historical data of the S&P 500 index between Jan-2016 and Jan-2017.

1.4.1 Comparison with ϵ-arbitrage Pricing

We present in this section the results of comparing the equal risk pricing model with the ϵ-arbitrage pricing model proposed in Bandi and Bertsimas (2014). Recall that the uncertainty set $\mathcal{U}$ employed in Bandi and Bertsimas (2014) admits the following form motivated by the central limit theorem:

$$\mathcal{U}_1 = \left\{ r \in \mathbb{R}^K \middle| \left| \frac{\sum_{\ell=1}^{k} \log(1+r_\ell) - \mu k T/K}{\sigma\sqrt{kT/K}} \right| \leq \Gamma, \forall k \in \{1,...,K\} \right\}, \tag{1.22}$$

where K is the number of periods up to the maturity of the option, and Γ denotes the "budget" of uncertainty at each time point t_k. Unfortunately, the above uncertainty set cannot be directly applied in the equal risk pricing model, since its associated worst-case risk measure does not necessarily satisfy the bounded conditional market risk, i.e. Assumption 1.3.2. We show in the following how the set can be slightly modified so that it satisfies Assumption 1.3.2.

Lemma 1.4.1. *Given that $\mu\sqrt{kT/K}/\sigma \leq \Gamma$ for all $k \in \{1,\ldots,K\}$, then the worst-case risk measure ρ that exploits $\mathcal{U}_1' = \mathcal{U}_1 \cap \mathcal{W}$ with*

$$\mathcal{W} = \left\{ r \in \mathbb{R}^K \middle| \max_{k' \in \{k+1,\ldots,K\}} \left| \frac{\sum_{\ell=1}^{k} \log(1+r_\ell) - \mu k' T/K}{\sigma\sqrt{k'T/K}} \right| - \Gamma \leq 0, \forall k \in \{1,...,K\} \right\}$$

satisfies both assumptions 1.2.1 and 1.3.2.

It is worth noting that the above set is smaller than the original set $\mathcal{U}_1$, as it excludes the sample paths that can lead to infinitely small risk. But as shown in the appendix, the above-modified set is in some sense the "largest" subset of $\mathcal{U}_1$ that, makes the worst-case risk measure satisfy assumptions 1.2.1 and 1.3.2. It is not hard to confirm that when using $\mathcal{U}_1'$, the worst-case risk measure is Markovian with respect to $\theta_k := \sum_{\ell=1}^{k} \log(1+r_\ell)$ (see appendices 1.6.13 and 1.6.14 respectively for a proof and details about the implementation of the dynamic program).

The parameter that needs to be further determined in our experiments is the budget parameter Γ. To do so, we start by first sampling 100,000 price paths from the GBM and then calibrating Γ so that the uncertainty set would cover at least 95% of the paths. In

Table 1.1, we present the option prices generated from the equal risk and the ϵ-arbitrage pricing models for various values of K and different types of options: In-The-Money (ITM), At-The-Mone (ATM), and Out-of-The-Money (OTM). The table also presents the fair price intervals.

From Table 1.1, we can make a few observations about the prices generated from the two models. Firstly, in the case of OTM, the prices generated from the ϵ-arbitrage pricing model are consistently lower than the prices generated from the equal risk pricing model. This is consistent with what was observed for the single period example in the introduction. Recall that in the case of single period (see Figure 1.1), the ϵ-arbitrage prices were always smaller or equal to ERP and differed most significantly from ERP when the options were out-of-the-money. Indeed, we see from Table 1.1 that in the case of ITM and ATM, the prices of the two models are more similar (without any clear dominance), but in the case of OTM options, the ERP is always significantly bigger than the ϵ-arbitrage price. This confirms that the ϵ-arbitrage pricing model can generate unrealistically low prices even in a multi-period hedging problem. Secondly, one can notice in Table 1.1 that the FPI-lower bounds always take the value of zero, i.e. the buyer's perception of minimal hedging risk is invariant to the number of rebalancing periods. While this may seem counter-intuitive, we can actually find an explanation by taking a closer look at the structure of the uncertainty set $\mathcal{U}_1'$. Namely, the set only imposes upper bounds on the variations of the underlying asset process. It turns out, however, that for the buyer's optimal hedging strategy, the worst paths are paths where the prices stay constant. These paths remain feasible regardless of the value of Γ. This explains why the lower bounds always reach the lowest possible value, i.e. zero, regardless of the number of hedging periods. Lastly, in Table 1.1, we also provide the prices generated from the Black–Scholes formula, and one can notice that the prices from equal risk pricing are not likely to converge to the Black–Scholes prices. This can also be explained by the conservativeness of the FPI-lower bounds, which drives up the ERP. We will discuss in the next section how such an issue might be resolved with a different choice of uncertainty set.

We compare also the risk exposure and level of fairness achieved by the transaction prices and hedging strategies produced from the two models. In particular, in our experiments we first simulate a set of 100,000 different sample paths for the risky asset and then for each

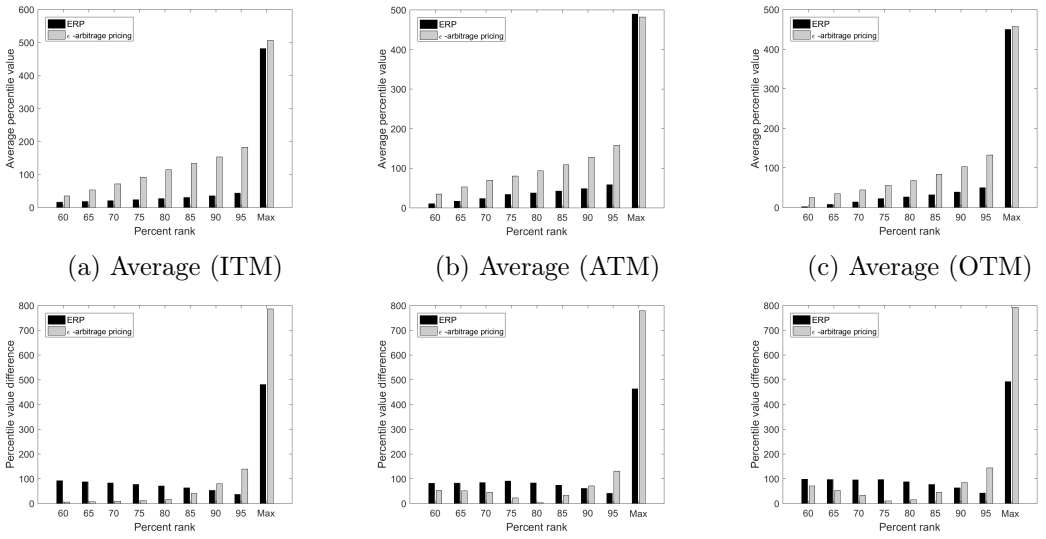
Table 1.1 – The prices resulting from ERP with $\mathcal{U}_1'$, ϵ-arbitrage pricing and the Black–Scholes models for options written on an asset with the initial price of 1000, expected annual return of 0.0718, annual standard deviation of 0.1283, strike prices of 950 (ITM), 1000 (ATM), and 1050 (OTM), and one year until maturity. For ERP, the fair price interval is also presented as FPI-Upper and FPI-Lower.

	ITM				**ATM**				**OTM**			
Periods	16	25	49	100	16	25	49	100	16	25	49	100
Γ	2.63	2.70	2.79	2.87	2.63	2.70	2.79	2.87	2.63	2.70	2.79	2.87
FPI-Upper	159.20	152.95	155.88	173.90	105.72	99.36	100.81	110.90	91.73	85.21	86.69	95.45
ERP	79.60	76.47	77.94	86.95	52.86	49.68	50.41	55.45	45.86	42.61	43.34	47.73
FPI-Lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ϵ-arbitrage price	78.40	73.60	75.20	91.20	57.60	52.80	56.00	62.40	32.00	25.60	27.20	38.40
BS	78.80	78.80	78.80	78.80	51.15	51.15	51.15	51.15	31.17	31.17	31.17	31.17

path we implement the optimal hedging strategy of each model starting with an initial capital that accounts for the transaction price. We record the hedging loss (for both the writer and the buyer) resulting from each sample path and compare different quantiles of the realized losses for both the writer and the buyer. For each quantile level of interest, we compare two different metrics: the average of the quantile value among the writer and buyer's loss, and the absolute difference between each party's quantile value. Figure 1.2 presents these metrics for options with different moneyness levels. As seen in Figure 1.2 (d-f), the hedging strategy and transaction price suggested by the ERP model lead to lower differences between the two parties' losses when considering quantiles above 90%. This is clear evidence that ERP is better at sharing the risks among the two parties. It is worth noting that for lower quantiles, ϵ-arbitrage becomes more attractive in this regard which can be explained by the fact that our worst-case risk measures that are used by ERP are insensitive to the performance achieved at lower quantiles. From Figure 1.2 (a-c), we see another strength of the ERP model, namely that it does have the ambition of producing optimal risk-averse hedging strategies for the two parties together with the ERP. Indeed, this is not the case of the ϵ-arbitrage pricing model, which searches for a single hedging strategy that minimizes the worst-case absolute "deviation" of the cumulated wealth from the payout.

1.4.2 Comparison with Black–Scholes

In the previous section, we highlighted how the FPI-lower bound becomes overly conservative when employing $\mathcal{U}_1$. We believe this explains why the ERP did not show signs it was



(a) Average (ITM) (b) Average (ATM) (c) Average (OTM)

(d) Difference (ITM) (e) Difference (ATM) (f) Difference (OTM)

Figure 1.2 – Comparison of hedging performance achieved under ϵ-arbitrage and equal risk pricing of an European call option with $K = 16$ rebalancing periods under a worst-case risk measure that accounts for $\mathcal{U}_1$. (a),(b), and (c) present for different percentile ranks q, the average among the q-percentile of the loss incurred for the writer and buyer of the ITM, ATM, and OTM options respectively. (d), (e), and (f) present the difference between the same q-percentile losses for different percentile rank. Note that the "Max" rank refers to the worst-case sample path.

converging to the Black–Scholes price even when the number of rebalancing periods became large. Given that such a convergence property is appealing when the market is actually based on a GBM process, in this section, we address this issue by employing a different uncertainty set that is now motivated by the work of Bernhard (2003), namely:

$$\mathcal{U}_2 = \left\{ r \in \mathbb{R}^K \middle| \sum_{\ell=1}^{sN} r_\ell^2 \in [\sigma^2 sT/S - \Gamma\sqrt{sN}, \sigma^2 sT/S + \Gamma\sqrt{sN}], \forall s \in \{1, ..., S\} \right\},$$

with Γ small enough so that $\mathcal{U}_2 \subset [-1, \infty[^K$. Here we consider the time horizon to be partitioned into S intervals of duration T/S, and each interval consists of a set of $N := K/S$ periods at which the portfolio can be rebalanced. Note that unlike for the set $\mathcal{U}_1$, the set $\mathcal{U}_2$ constrains both the maximum and minimum long-term observed deviations. The main motivation behind the above set is that in the case $\Gamma = 0$, we can expect based on Bernhard (2003) that the FPI will converge to Black–Scholes price as both K and N increase to infinity. On the other hand, for finite values of K and N, the "so-called" budget Γ of the set $\mathcal{U}_2$ allows to characterize a meaningful confidence region for the trajectory of the risky

Table 1.2 – The option prices resulting from the equal risk (ERP) and the Black–Scholes (BS) models by using $\mathcal{U}_2$. The table also shows the calibrated Γ and the upper and lower bounds of the fair price interval.

	ITM				**ATM**				**OTM**			
Periods	16	49	100	225	16	49	100	225	16	49	100	225
Γ	0.083	0.022	0.006	0.004	0.083	0.022	0.006	0.004	0.083	0.022	0.006	0.004
FPI-Upper	125.19	113.66	107.04	102.18	99.31	88.61	81.66	75.96	83.00	69.01	61.57	55.77
ERP	87.60	84.10	83.36	83.24	49.65	54.96	55.38	54.99	41.50	36.21	35.25	34.88
FPI-Lower	50.00	54.55	59.67	64.30	0.00	21.30	29.11	34.03	0.00	3.41	8.93	13.98
BS price	78.80	78.80	78.80	78.80	51.15	51.15	51.15	51.15	31.17	31.17	31.17	31.17

asset process. Lastly, as shown in Appendix 1.6.12, one can verify that the worst-case risk measure with $\mathcal{U}_2$ satisfies the bounded conditional market risk property and that is Markovian with respect to $\theta_k := \sum_{\ell=1}^{k} r_\ell^2$ (see Appendix 1.6.13 for details).

In our experiments, we set the number of partitions to the square root of the total number of rebalancing periods, i.e. $S = \sqrt{K}$. We also calibrate Γ so that the set $\mathcal{U}_2$ contains 95% of simulated price paths. Table 1.2 presents the equal risk and the fair price intervals against the Black–Scholes prices. From the table, we now see some evidence that the price generated from equal risk pricing is likely to converge to the Black–Scholes price. In particular, one can notice for each type of option that as the total number of rebalancing periods increases, both the upper and lower bounds evolve monotonically towards the Black–Scholes price, thus driving the equal risk prices closer and closer to it. Unlike with $\mathcal{U}_1$, we see that the FPI-lower bounds are now sensitive to the total number of rebalancing periods. Indeed, the lower bound on the total deviation in $\mathcal{U}_2$ allows the buyer of call options to have a less conservative perception of hedging risk. Finally, it is worth noting that the resulting equal risk prices tend to be slightly higher than the Black–Scholes prices. Indeed, from a practical point of view, this margin can be interpreted as a "risk premium" on the Black–Scholes price that compensates for the uncertainty that is unaccounted for by the Black–Scholes formula.

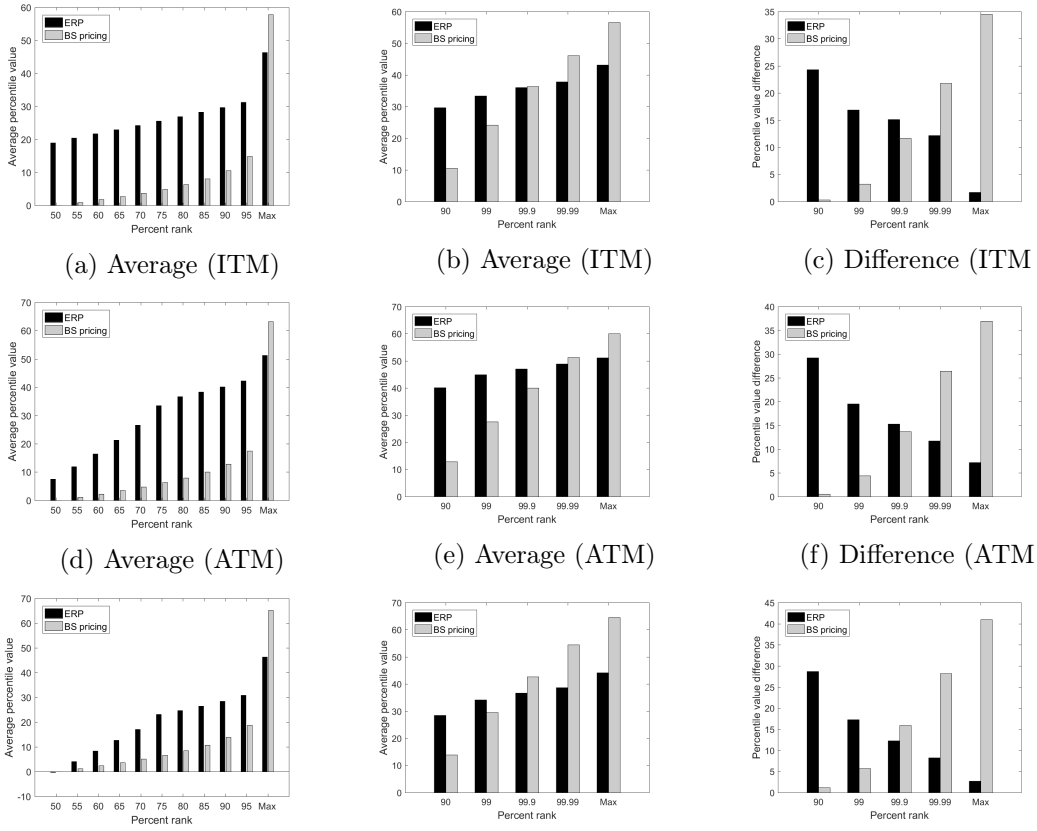
Also depicted in Figure 1.3 is a comparison of hedging performances between the equal risk pricing model and the Black–Scholes model, in the case $K = 16$. As in the case of comparing with ϵ-arbitrage pricing, we report the performances in terms of both the average and the absolute difference of the writer and buyer's quantiles of their realized loss distribution under their respective hedging strategy. In particular, here we provide these

metrics for quantiles starting from 99% in order to emphasize what happens at the tail of the loss distributions. For these figures, the last group of bars is labeled with "Max" to show the worst-case value of the metrics in all samples. This is in line with the type of risk measure that is used in this section. The results for smaller quantile levels are also provided for average losses (see Figure 1.3 (a,d,g)) to present a complete picture. We see that hedging according to the Black–Scholes model actually performs fairly well across a wide range of lower level quantiles, which is not surprising given the market is assumed to follow the GBM assumed by Black–Scholes. Unlike the Black–Scholes model, the ERP model employs a worst-case risk measure that controls the risk in the tail of the loss distributions. As shown in the figures with higher quantile levels, hedging and pricing according to ERP model does indeed become the best scheme when focusing on those regions in terms of both the averages and the differences of risks for the two parties.

We continue with Figure 1.4, which presents the hedging strategies at time $k = 0$ proposed by the equal risk pricing and the Black–Scholes pricing for an ATM call option under 16 and 225 rebalancing periods for different asset prices. The first observation is that as K increases, the hedging strategy seems to more closely resemble the Black Scholes strategy for both the writer and the buyer. This complements the observation that the ERP appeared to converge to the Black–Scholes price. For lower values of K, the hedging strategy for the buyer differs significantly from Black–Scholes hedging because of the larger uncertainty about the risky asset's price process. Specifically, it swings from fully shorting the risky asset to keeping only the risk-free asset. The latter strategy becomes optimal because Γ is large enough to allow the risky asset to evolve exactly as the risk-free one, which also drags the lower bound of FPI to zero as discussed in Section 1.4.1. The second observation is that the hedging strategies of the equal risk model are less sensitive for the writer and more sensitive for the buyer to the variations in the underlying stock price at time $k = 0$. In particular, Figure 1.4(a) shows that the ERP model provides a hedging strategy with a lower slope for the writer of the option compared to the Black–Scholes. On the other side, the opposite is happening for the option buyer. The equal risk hedging strategy is more sensitive to the asset price compared to Black–Scholes strategy.

Finally, we present additional information about the different strategies in Table 1.3. In particular, the table presents the mean of the average portfolio turnover, computed

as $\frac{\sum_{k=1}^{K-1}|\xi_{k+1}-\xi_k|}{K-1}$, over 100,000 sample paths for each strategy. This statistic provides evidence that the writer incurs less rebalancing while the opposite is true for the buyer. For completeness, we also present in the table the mean of the average number of shares of the risky asset that are held by each strategy, together with the mean and standard deviation of the respective portfolio values at the maturity of the option.

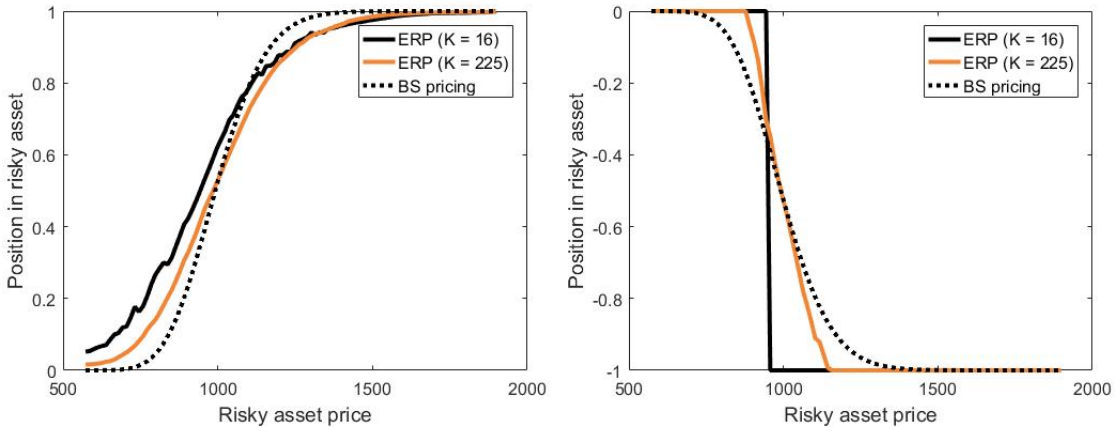


(a) Average (ITM) (b) Average (ITM) (c) Difference (ITM

(d) Average (ATM) (e) Average (ATM) (f) Difference (ATM

(g) Average (OTM) (h) Average (OTM) (i) Difference (OTM

Figure 1.3 – Comparison of hedging performance achieved under the Black–Scholes and the equal risk pricing of a European call option with $K = 16$ rebalancing periods under a worst-case risk measure that accounts for $\mathcal{U}_2$. (a),(d), and (g) present for different percentile ranks q, the average among the q-percentile of the loss incurred for the writer and buyer of the ITM, ATM, and OTM options respectively. (b),(e), and (h) presents similar information but focusing on higher percentiles. (c), (f), and (i) present the difference between the same q-percentile losses. Note that the "Max" rank refers to the worst-case sample path.



(a) Writer (b) Buyer

Figure 1.4 – Comparison of hedging strategies for a European ATM Call option under different number of rebalancing periods. (a) presents the optimal strategies for the writer under the Black–Scholes and the equal risk pricing models for time $k = 0$. (b) presents the same for the buyer.

Table 1.3 – Comparison of the hedging strategies resulting from the equal risk model and the Black–Scholes for $K = 225$

Measure	**ITM**			**ATM**			**OTM**		
	ERP_w	ERP_b	BS pricing	ERP_w	ERP_b	BS pricing	ERP_w	ERP_b	BS pricing
Average position in risky asset - Mean	0.7060	0.7997	0.7501	0.5950	0.6172	0.60954	0.4815	0.4360	0.4606
Average portfolio turnover - Mean	0.0136	0.0200	0.0168	0.0154	0.026	0.0203	0.0160	0.0272	0.0210
Terminal portfolio value - Mean	135.15	142.43	136.68	100.05	102.45	98.88	72.39	70.00	67.91
Terminal portfolio value - STD	113.37	131.87	122.48	101.93	118.98	110.33	89.17	102.29	95.14

1.4.3 The Case of American Options

In this section, we take a further step to benchmark equal risk pricing model against a binomial tree model in the case of American option. For the same reason discussed in the previous section, we assume the worst-case risk measures are defined according to the uncertainty set $\mathcal{U}_2$. Here, we consider put options rather than call options, as the former has attracted more attention in the literature treating American options.

The calibration of the uncertainty set $\mathcal{U}_2$, i.e. Γ is done in the same fashion as in the previous section. We implement the equal risk model for both the case of with commitment (see Definition 6) and without commitment (see Definition 7). We summarize in Table 1.4 all the prices and FPI bounds generated from the model against the option prices generated

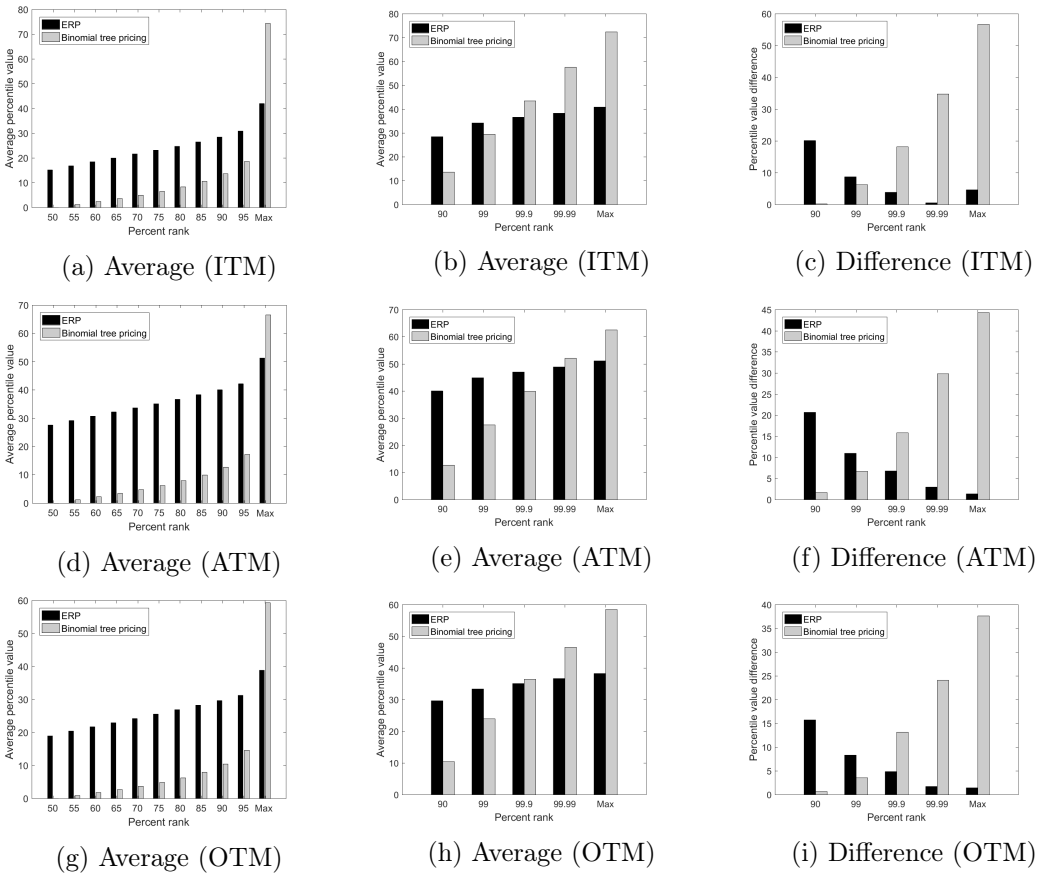
Table 1.4 – The option prices resulting from the equal risk pricing model (ERP) under $\mathcal{U}_2$ compared to the binomial tree model (BTM) for American put options. The models with and without commitment are identified respectively as WC and NC.

	ITM				**ATM**				**OTM**			
Periods	16	49	100	225	16	49	100	225	16	49	100	225
Γ	0.083	0.022	0.006	0.004	0.083	0.022	0.006	0.004	0.083	0.022	0.006	0.004
FPI-Upper-NC	134.18	117.54	114.16	106.79	103.52	87.56	84.06	77.20	75.65	62.72	59.25	53.22
ERP-NC	92.09	84.73	86.85	84.10	51.76	52.95	56.49	54.25	37.83	32.95	34.11	32.16
FPI-Lower-NC	50.00	51.93	59.53	61.42	0.00	18.34	28.92	31.31	0.00	3.18	8.97	11.11
FPI-Upper-WC	134.18	116.14	108.54	100.43	103.52	85.12	81.36	72.95	75.65	60.84	55.76	47.14
ERP-WC	92.09	84.03	84.04	80.92	51.76	51.73	55.14	52.13	37.83	32.01	32.37	29.12
FPI-Lower-WC	50.00	51.93	59.53	61.42	0.00	18.34	28.92	31.31	0.00	3.18	8.97	11.11
BTM price	81.52	81.19	81.13	81.21	50.36	51.41	51.02	51.21	29.00	28.73	28.68	28.85

from the binomial tree model.

As we expected (see Lemma 1.3.2), the equal risk prices with commitment are always smaller or equal to the prices without commitment. We can also confirm that the differences in the prices between the two cases result from the differences in their respective upper bound prices, since their lower bounds are similar. The results also show that as the number of rebalancing periods K increases the equal risk price is getting closer to the binomial tree price. This is happening for all types of options. We see that the equal risk price without commitment is larger than the price with commitment by a factor as large as 4% for ATM and ITM options, and 10% for OTM options. This non-negligible difference between the prices of the two cases highlights the importance of commitment as a factor to be considered in the negotiation between the two parties regarding the transaction price. This also indicates that the value of the buyer's commitment to an exercise policy is particularly high for an OTM option.

In terms of hedging, the equal risk model shows similar results to the case of European options. Figure 1.5 shows that for high quantiles of loss, the equal risk model outperforms the binomial tree model. This is understood from comparing the graphs that focus on the quantiles at the tails of the loss distributions. The higher performance of the equal risk model is specifically more outstanding in terms of the equality of hedging loss for the two sides. However, having a GBM price process prepares the ground for the binomial tree model to perform well in terms of lower quantiles as shown in figure 1.5 (a,d,g).



(a) Average (ITM) (b) Average (ITM) (c) Difference (ITM)

(d) Average (ATM) (e) Average (ATM) (f) Difference (ATM)

(g) Average (OTM) (h) Average (OTM) (i) Difference (OTM)

Figure 1.5 – Comparison of hedging performance achieved under equal risk pricing, with $\mathcal{U}_2$, and a binomial tree model, of an American put option with $K = 16$ rebalancing periods. (a),(d), and (g) present for different percentile ranks q, the average among the q-percentile of the loss incurred for the writer and buyer of the ITM, ATM, and OTM options respectively. (b),(e), and (h) presents similar information but focusing on higher percentiles. (c), (f), and (i) present the difference between the same q-percentile losses.

1.5 Conclusion

In this article, we explore the famous problem of pricing and hedging options in an incomplete market under a recently proposed framework called equal risk pricing. Under this framework, the pricing of an option requires that the risk of both sides of the contract be considered in order to make them equal. We consider for the first time the special case of equal risk pricing under convex risk measures for which we show that ERP conveniently reduces to the center of the fair price interval. This price can thus be obtained by solving two dynamic derivative hedging problems, i.e. for the writer and the buyer. By further imposing that the risk

measures be one-step decomposable, Markovian, and satisfy a bounded conditional market risk condition, we derive finite-dimensional risk-averse dynamic programming equations that can be used to solve the discrete-time hedging problems for both European and American options. With the latter type of option, the resulting Bellman equations further depend on whether the buyer is willing to commit or not to an exercise strategy up front. All of our results are general enough to accommodate situations where the risk is measured using a worst-case risk measure that considers only a subset of realizations from the outcome space, as typically done in robust optimization.

In our numerical study, we compare the performance of using equal risk pricing with a worst-case risk measure to the performance of ϵ-arbitrage pricing and pricing using the Black–Scholes model in a market that is based on a discretized geometric Brownian motion. In particular, the numerical results confirm that, when using the equal risk price, both the writer and the buyer end up having risks that are more similar and on average smaller than the risks that they would experience by the two other approaches. In addition, by proposing a new uncertainty set inspired from the work of Bernhard (2003), we show that the prices generated from equal risk pricing have the potential to converge to Black–Scholes prices as the hedging frequency increases. In the case of pricing American put options, we show how to calculate the value of commitment to an exercise policy, which ranges between 0% and 10% for the instances we considered. The evidence seems to indicate that this relative value decreases as the ERP without commitment increases.

Finally, it is worth mentioning that the results presented in this article have natural extensions to more general settings than the one that is considered, i.e. with a single underlying asset, zero risk-free rate, frictionless market. In some cases, the Bellman equations might need to be extended to account for a larger state space, which is likely to increase the computational efforts needed to identify the equal risk price and the hedging strategies. To circumvent this issue, one might resort to approximate dynamic programming methods. One interesting recent attempt in this direction can be found in Carbonneau and Godin (2020) that proposes a deep reinforcement-learning approach to approximate the equal risk price in a variety of market dynamics (including GARCH and Merton jump-diffusion processes) and exotic options with multiple underlying assets.

1.6 Appendix

1.6.1 Analytical Solutions of One-period Example

We recall from Bandi and Bertsimas (2014) that the ϵ-arbitrage model under a worst-case risk measure can be defined as follows for an European option:

$$\min_{\xi,p_0} \max_{S_1 \in \mathcal{U}} \left| (S_1 - K)^+ - p_0 - \xi(S_1 - S_0) \right| , \tag{1.23}$$

where S_0 is the initial stock price, K is the strike price of the option, S_1 is the price at the next time period, and $\mathcal{U} \subseteq \mathbb{R} = [l, u]$. Without loss of generality, we set $l \leq K \leq u$ and consider the risk-free rate to be zero.

In the framework of equal risk pricing (ERP), we consider modeling separately the hedging problem of the writer and the buyer. When considering a one-period problem, the equal risk model is as follows:

$$\varrho^w(p_0) := \min_{\xi_w} \max_{S_1 \in \mathcal{U}} (S_1 - K)^+ - p_0 - \xi_w(S_1 - S_0)$$
$$\varrho^b(p_0) := \min_{\xi_b} \max_{S_1 \in \mathcal{U}} -(S_1 - K)^+ + p_0 - \xi_b(S_1 - S_0).$$

The equal risk price is set to be the initial wealth p_0 that leads to $\varrho^w(p_0) = \varrho^b(p_0)$.

Analytical Solution for the one-period Equal Risk Model

The analytical solution of the one-period equal risk model is as follows:

$$\xi_w^* = \frac{u - K}{u - l}, \quad \xi_b^* = \begin{cases} 0, & \text{if } S_0 < K \\ -1, & \text{if } S_0 \geq K \end{cases}$$
$$p_0^* = (1/2)(S_0 - l)\frac{u - K}{u - l} + (1/2)(S_0 - K)^+ .$$

Considering the writer's side of the equal risk model, since $(S_1 - K)^+ - \xi_w(S_1 - S_0)$ is a convex function of S_1, the maximum in the interval of $\mathcal{U} = [l, u]$ is at the boundaries, resulting in

$$\begin{aligned} \varrho^w(p_0) &= \min_{\xi_w} \max_{S_1 \in \mathcal{U}} (S_1 - K)^+ - p_0 - \xi_w(S_1 - S_0) \\ &= -p_0 + \min_{\xi_w} \max\{u - K - \xi_w(u - S_0), -\xi_w(l - S_0)\} . \end{aligned}$$

Since the first argument is decreasing in ξ_w and the second one is increasing, the minimum is at the intersection of the two functions, which results in

$$\xi_w^* = \frac{u-K}{u-l}, \quad \varrho^w(p_0) = -p_0 + (S_0 - l)\frac{u-K}{u-l}\,.$$

On the other hand, for the buyer of the option we can show that $\varrho^b(p_0) = -(S_0-K)^+ + p_0$ and is achieved using the described hedging strategy ξ_b^*. In particular, we can first establish that, for all hedging strategies $\xi_b \in \mathbb{R}$

$$\max_{S_1 \in \mathcal{U}} -(S_1-K)^+ + p_0 - \xi_b(S_1 - S_0) \geq -(S_0-K)^+ + p_0\,,$$

where we simply use the fact that $S_0 \in \mathcal{U}$.

Now, since $g(y) = -(y-K)^+$ is a concave function of y, if $\nabla g(S_0)$ is a supergradient of $g(y)$ at S_0 then we have that:

$$g(S_1) \leq g(S_0) + \nabla g(S_0)^T(S1 - S_0)\,,$$

which means that since it can be verified that ξ_b^* is a valid candidate for $\nabla g(S_0)$, we have that:

$$\max_{S_1 \in \mathcal{U}} -(S_1-K)^+ + p_0 - \xi_b^*(S_1 - S_0) \leq -(S_0-K)^+ + p_0\,.$$

This proves that ξ_b^* achieves the minimum value of $-(S_0-K)^+ + p_0$.

We conclude this discussion by verifying that for p_0^* indeed leads to the same risk for both the writer and the buyer:

$$\begin{aligned}
\varrho^w(p_0^*) &= -p_0^* + (S_0 - l)\frac{u-K}{u-l} \\
&= -\left((1/2)(S_0-l)\frac{u-K}{u-l} + (1/2)(S_0-K)^+\right) + \frac{u-K}{u-l}(S_0-l) \\
&= \frac{1}{2}(S_0-l)\frac{u-K}{u-l} - (1/2)(S_0-K)^+ = p_0^* - (S_0-K)^+ = \varrho^b(p_0^*)\,.
\end{aligned}$$

Analytical Solution for the one-period ϵ-arbitrage Model

In this section, we will demonstrate that an optimal solution for the ϵ-arbitrage model takes the following form:

$$\xi^* = \frac{u-K}{u-l}, \quad p_0^* = \frac{u-K}{u-l}\left(S_0 - \frac{1}{2}(K+l)\right).$$

To do so, we will exploit the following lemma, which appears as proposition 3.1.4 in Bertsekas (2015)

Lemma 1.6.1. *A vector x^* minimizes a convex function $f : R^n \to R$ over a convex set $\mathcal{X} \subset R^n$ if and only if there exists a subgradient $\nabla f(x^*)$ of f at x^* such that $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \mathcal{X}$.*

In other words, we will be able to conclude that the (ξ^*, p_0^*) pair is a minimizer of problem (1.23), if we can show that 0 is a subgradient of the objective function at (ξ^*, p_0^*). Based on Section 3.1.1 of Bertsekas (2015), one can show that the set of all subgradients at (ξ^*, p_0^*) include

$$\left\{ \nabla g \in \mathbb{R}^2 \,\middle|\, \exists \lambda \in \mathbb{R}^4,\ \lambda \geq 0,\ \sum_{i=1}^{4} \lambda_i = 1,\ \nabla g = \begin{bmatrix} -u + S_0 & -l + S_0 & K - S_0 & K - S_0 \\ -1 & -1 & 1 & 1 \end{bmatrix} \lambda \right\} \quad (1.24)$$

One can then readily verify that 0 is a member of this set using $\lambda_1 := (1/2)(K - l)/(u - l)$, $\lambda_2 := (1/2)(u - K)/(u - l)$, $\lambda_3 := 0$, and $\lambda_4 := 1/2$ as a certificate.

To provide more details on obtaining the set in (1.24), we start by recalling that when f is the maximum of m subdifferentiable convex functions $\phi_1, ..., \phi_m$:

$$f(x) = \max\{\phi_1(x), ..., \phi_m(x)\}, x \in \mathbb{R}^n\,, \quad (1.25)$$

then a subset of the subdifferential of f can be described as:

$$\partial f(x) = \text{conv}\{\nabla \phi_j(x) | j \in \mathcal{J}(x)\}\,, \quad (1.26)$$

where $\mathcal{J}(x) := \{j \in \{1, \ldots, m\} \,|\, \phi_j(x) = f(x)\}$, and each $\nabla \phi_j(x)$ is a subgradient of $\phi_j(\cdot)$ at x. To obtain the set in (1.24), we first formulate the objective function in the form of equation (1.25), we then identify a subgradient $\nabla \phi_j(x^*)$ of each $j \in \mathcal{J}(x^*)$ at our proposed solution x^* to compose the set described in (1.26).

Step 1. We can rewrite the objective function of problem (1.23) by exploiting a partition of $\mathcal{U}$ as follows:

$$\max_{S_1 \in \mathcal{U}} |(S_1 - K)^+ - p_0 - \xi(S_1 - S_0)| = \max\left\{\phi_1(\xi, p_0), \phi_2(\xi, p_0), \phi_3(\xi, p_0), \phi_4(\xi, p_0)\right\}\,,$$

where

$$\phi_1(\xi, p_0) := \max_{S_1 \in [K,u]} |(S_1 - K)^+ - p_0 - \xi(S_1 - S_0)| = \max_{S_1 \in [K,u]} (S_1 - K) - \xi(S_1 - S_0) - p_0$$
$$\phi_2(\xi, p_0) := \max_{S_1 \in [l,K]} |(S_1 - K)^+ - p_0 - \xi(S_1 - S_0)| = \max_{S_1 \in [l,K]} -\xi(S_1 - S_0) - p_0$$
$$\phi_3(\xi, p_0) := \max_{S_1 \in [K,u]} |(S_1 - K)^+ - p_0 - \xi(S_1 - S_0)| = \max_{S_1 \in [K,u]} -(S_1 - K) + \xi(S_1 - S_0) + p_0$$
$$\phi_4(\xi, p_0) := \max_{S_1 \in [l,K]} |(S_1 - K)^+ - p_0 - \xi(S_1 - S_0)| = \max_{S_1 \in [l,K]} \xi(S_1 - S_0) + p_0 \,.$$

Step 2. In order to find $\mathcal{J}(x^*)$, we study the maximum of all four functions when $\xi = \xi^*$ and $p_0 = p_0^*$. Specifically, we have:

$$\phi_1(\xi^*, p_0^*) = \max_{S_1 \in [K,u]} S_1 - K - \frac{u-K}{u-l}(S_1 - S_0) - \frac{u-K}{u-l}(S_0 - \frac{1}{2}(K+l)) = \frac{u-K}{u-l}\frac{K-l}{2}$$
$$\phi_2(\xi^*, p_0^*) = \max_{S_1 \in [l,K]} -\frac{u-K}{u-l}(S_1 - S_0) - \frac{u-K}{u-l}(S_0 - \frac{1}{2}(K+l)) = \frac{u-K}{u-l}\frac{K-l}{2}$$
$$\phi_3(\xi^*, p_0^*) = \max_{S_1 \in [K,u]} -S_1 + K + \frac{u-K}{u-l}(S_1 - S_0) + \frac{u-K}{u-l}(S_0 - \frac{1}{2}(K+l)) = \frac{u-K}{u-l}\frac{K-l}{2}$$
$$\phi_4(\xi^*, p_0^*) = \max_{S_1 \in [l,K]} \frac{u-K}{u-l}(S_1 - S_0) + \frac{u-K}{u-l}(S_0 - \frac{1}{2}(K+l)) = \frac{u-K}{u-l}\frac{K-l}{2}\,,$$

where we exploited the fact that the functions that are maximized are either non-decreasing for the case of ϕ_1 and ϕ_4 or non-increasing for ϕ_2 and ϕ_3. In each case, the maximum is achieved at $S_1^* = u$ for ϕ_1, $S_1^* = l$ for ϕ_2, and $S_1^* = K$ for ϕ_3 and ϕ_4. Based on this conclusion, we get the following four subgradients:

$$\nabla\phi_1(\xi^*, p_0^*) := \begin{bmatrix} S_0 - u \\ -1 \end{bmatrix} \qquad \nabla\phi_2(\xi^*, p_0^*) := \begin{bmatrix} S_0 - l \\ -1 \end{bmatrix}$$
$$\nabla\phi_3(\xi^*, p_0^*) := \begin{bmatrix} K - S_0 \\ 1 \end{bmatrix} \qquad \nabla\phi_4(\xi^*, p_0^*) := \begin{bmatrix} K - S_0 \\ 1 \end{bmatrix}.$$

This completes our proof.

1.6.2 Proof of Proposition 1.2.1

This proof mainly relies on the translation invariance property together with the following property of $\mathcal{X}(p_0)$:

$$\begin{aligned}\mathcal{X}(p_0) &= \left\{ X \,\middle|\, \exists \xi_s,\, \exists c \in \mathbb{R}, \quad X_t = p_0 + \int_0^t \xi_s dS_s \geq c \,; \forall\, t \in [0,\, T] \right\} \\ &= p_0 + \left\{ X' \,\middle|\, \exists \xi_s,\, \exists c \in \mathbb{R}, \quad X'_t = \int_0^t \xi_s dS_s \geq c \,; \forall\, t \in [0,\, T] \right\} \\ &= p_0 + \mathcal{X}(0)\,,\end{aligned}$$

where $p_0 + \mathcal{X}(0)$ refers to a set addition. These two properties can be used to show that both

$$\begin{aligned}\varrho^w(p_0) &= \inf_{X \in \mathcal{X}(0)+p_0} \rho^w(F(S_T, Y_T) - X_T) \\ &= \inf_{X \in \mathcal{X}(0)} \rho^w(F(S_T, Y_T) - X_T - p_0) \\ &= \inf_{X \in \mathcal{X}(0)} \rho^w(F(S_T, Y_T) - X_T) - p_0 \\ &= \inf\{s | \inf_{X \in \mathcal{X}(0)} \rho^w(F(S_T, Y_T) - X_T) \leq s\} - p_0 \\ &= \inf\{s | \varrho^w(s) \leq 0\} - p_0 = p_0^w - p_0\,, \end{aligned} \tag{1.27}$$

and similarly,

$$\begin{aligned}\varrho^b(p_0) &= \inf_{X \in \mathcal{X}(0)-p_0} \rho^b(-X_T - F(S_T, Y_T)) \\ &= \inf_{X \in \mathcal{X}(0)} \rho^b(-X_T - F(S_T, Y_T)) + p_0 \\ &= \inf\{s | \varrho^b(0) \leq s\} + p_0 \\ &= \inf\{s | \varrho^b(-s) \leq 0\} + p_0 \\ &= -\sup\{-s | \varrho^b(-s) \leq 0\} + p_0 \\ &= -p_0^b + p_0\,. \end{aligned} \tag{1.28}$$

Hence, we can obtain our result by verifying both directions of the biconditional logical connective. First, given that an equal risk price exists, say $p_0^* \in \mathbb{R}$, it must be that both $\varrho^b(p_0^*)$ and $\varrho^b(p_0^*)$ are members of $\mathbb{R}$. This necessarily implies that $p_0^w \in \mathbb{R}$ and $p_0^b \in \mathbb{R}$ and so the fair price interval is bounded, although possibly empty. Conversely, if the fair price

interval is bounded, then one can verify that $p_0^* := (p_0^w + p_0^b)/2$ does satisfy the equal risk price condition:

$$\varrho^w(p_0^*) = p_0^w - p_0^* = p_0^w/2 - p_0^b/2 = -p_0^b + p_0^* = \varrho^b(p_0^*)\,.$$

Furthermore, this p_0^* can be calculated as:

$$\begin{aligned} p_0^* &= (1/2)(\inf\{p_0|\varrho^w(p_0) \leq 0\} - \sup\{-s|\varrho^b(-s) \leq 0\}) \\ &= (1/2)(p_0^w + p_0^b) = (1/2)(\varrho^w(0) - \varrho^b(0)))\,, \end{aligned}$$

following exactly the same arguments as in (1.27) and (1.28). This completes our proof.

1.6.3 Proof of Lemma 1.2.2

The proof follows directly from Property 2 in Xu (2006). In particular, we have that p_0^w is bounded above by the super-hedging price and p_0^b is bounded below by the sub-hedging price. Hence, since $p_0^b \leq p_0^w$, we must have that the fair price interval is a subset of the no-arbitrage interval. This lets us conclude that the equal risk price is also a member of the no-arbitrage interval.

1.6.4 Proof of Proposition 1.3.1

We focus on providing the arguments supporting the claims for the writer model as these are analogous for the buyer model. In doing so, we will closely follow the theory presented in Pichler and Shapiro (2018). We start by constructing the so-called additive preference system $\{\mathcal{R}_{k,l}\}_{(k,l)\in\{0,\ldots,K\}^2:k<l}$ (a.k.a. a dynamic risk measure) based on ρ^w, where each $\mathcal{R}_{k,l} : \mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P}) \times \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P}) \times \cdots \times \mathcal{L}_p(\Omega, \mathcal{F}_l, \mathbb{P})$ takes the form:

$$\mathcal{R}_{k,\ell}(Z_k, Z_{k+1}, \ldots, Z_\ell) := \rho_k^w(\rho_{k+1}^w(\cdots \rho_K^w(\sum_{k'=k}^{\ell} Z_{k'})\cdots)\,,\ \forall\, 0 \leq k < \ell \leq K\,.$$

Based on this definition of $\mathcal{R}_{k,\ell}$ it is easy to see that:

$$\varrho^w(0) = \inf_{X\in\mathcal{X}(0)} \mathcal{R}_{0,K}(0, 0, \ldots, 0, F(S_T, Y_T) - X_K)\,.$$

Given that ρ^w is one-step decomposable, it is easy to show that $\mathcal{R}$ is both "Monotone" and "Recursive" (see definitions 2.3 and 4.1 in Pichler and Shapiro (2018)). In particular, for

monotonicity we have that:

$$\forall (Z_k, \dots, Z_\ell), (Z'_k, \dots, Z'_\ell),\ Z_{k'} \geq Z'_{k'} \text{ a.s. } \forall\, k' = k, \dots, \ell \Rightarrow$$

$$\begin{aligned}
\mathcal{R}_{k,\ell}(Z_k, \dots, Z_\ell) &= \rho_k^w(\rho_{k+1}^w(\cdots \rho_K^w(\sum_{k'=k}^{\ell} Z_{k'}) \cdots) \\
&\geq \rho_k^w(\rho_{k+1}^w(\cdots \rho_K^w(\sum_{k'=k}^{\ell} Z'_{k'}) \cdots) \\
&= \mathcal{R}_{k,\ell}(Z'_k, \dots, Z'_\ell)\,,
\end{aligned}$$

given that each ρ_k^w is monotone. On the other hand, for recursivity, we have

$$\forall Z_k, \dots, Z_\ell,$$

$$\begin{aligned}
\mathcal{R}_{k,\ell}(Z_k, \dots, Z_\ell) &= \rho_k^w(\rho_{k+1}^w(\cdots \rho_{v-1}^w(\rho_v^w(\cdots \rho_K^w(\sum_{k'=k}^{\ell} Z_{k'}) \cdots) \\
&= \rho_k^w(\rho_{k+1}^w(\cdots \rho_{v-1}^w(\sum_{k'=k}^{v-1} Z_{k'} + \rho_v^w(\cdots \rho_K^w(\sum_{k'=v}^{\ell} Z_{k'}) \cdots) \\
&= \rho_k^w(\rho_{k+1}^w(\cdots \rho_{v-1}^w(\sum_{k'=k}^{v-1} Z_{k'} + \mathcal{R}_{v,\ell}(Z_v, \dots, Z_\ell)) \cdots) \\
&= \rho_k^w(\rho_{k+1}^w(\cdots \rho_{v-1}^w(\rho_v^w(\cdots \rho_K^w(\sum_{k'=k}^{v-1} Z_{k'} + \mathcal{R}_{v,\ell}(Z_v, \dots, Z_\ell)) \cdots) \\
&= \mathcal{R}_{k,v}(Z_k, \dots, Z_{v-1}, \mathcal{R}_{v,\ell}(Z_v, \dots, Z_\ell)))\,,
\end{aligned}$$

where we exploited monotonicity, the definition of $\mathcal{R}_{v,\ell}$, and the conditional transition invariance of ρ_k^w for all $k = v, \dots, K$.

Having verified these conditions, Proposition 1.3.1 and the discussion that follows in Section 4 of Pichler and Shapiro (2018) allows us to conclude that:

$$\begin{aligned}
\varrho^w(0) &= \inf_{X \in \mathcal{X}(0)} \mathcal{R}_{0,K}(0, 0, \dots, 0, F(S_T, Y_T) - X_K) \\
&= \inf_{\xi_0} \mathcal{R}_{0,1}(0, \inf_{\xi_1} \mathcal{R}_{1,2}(0, \cdots \inf_{\xi_{K-1}} \mathcal{R}_{K-1,K}(0, F(S_T, Y_T) - X_K) \cdots)\,.
\end{aligned}$$

Hence, $\varrho^w(0) = \bar{V}_0^w(0, \omega)$ where

$$\begin{aligned}
\bar{V}_k^w(X_k, \omega) &:= \inf_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k + \xi_k \Delta S_{k+1}), \omega) \\
\bar{V}_K^w(X_K, \omega) &:= F(S_T(\omega), Y_T(\omega)) - X_K(\omega)\,.
\end{aligned}$$

Furthermore, the set of optimal policies for problem (1.1a) must contain the following hedging policies:

$$\bar{\xi}_k^{w*}(X_k,\omega) \in \arg\min_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1}),\omega),\ \forall\, k=0,\dots,K-1\,.$$

Yet, by conditional translation invariance, we know that:

$$\bar{V}_K^w(X_K,\omega) = V_K^w(\omega) - X_K(\omega)\,,$$

and recursively that:

$$\begin{aligned}
\bar{V}_k^w(X_k,\omega) &= \inf_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1}),\omega)\\
&= \inf_{\xi_k} \rho_k^w(V_{k+1}^w - X_k - \xi_k\Delta S_{k+1},\omega)\\
&= \inf_{\xi_k} \rho_k^w(V_{k+1}^w - \xi_k\Delta S_{k+1},\omega) - X_k(\omega) = V_k^w(\omega) - X_k(\omega)\,.
\end{aligned}$$

Hence, $\varrho^w(0) = \bar{V}_0^w(0,\omega) = V_0^w(\omega)$. A similar argument confirms that the set of optimal policies for problem (1.1a) contains for all $k=0,\dots,K-1$:

$$\begin{aligned}
\bar{\xi}_k^{w*}(X_k,\omega) &\in \arg\min_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1}),\omega)\\
&= \arg\min_{\xi_k} \rho_k^w(V_{k+1}^w - \xi_k\Delta S_{k+1},\omega) - X_k(\omega)\\
&= \arg\min_{\xi_k} \rho_k^w(V_{k+1}^w - \xi_k\Delta S_{k+1},\omega)\,,
\end{aligned}$$

thus equivalent to $\xi_k^{w*}(\omega)$.

1.6.5 Dynamic Programming Equations for the Case of Non-translation Invariant Risk Measures

In the case where ρ^w and ρ^b do not satisfy the translation invariance property, the bisection method described in Remark 1 relies on computing the value of $\varrho^w(p_0)$ and $\varrho^b(p_0)$ for any value of p_0. Similar arguments as used in sections 1.3.1 and 1.3.2 to identify dynamic programming equations, which now depend on accumulated wealth, when the risk measures are one-step decomposable and Markovian.

In particular, focusing on the case of the writer's problem associated to a European option, by following the steps in Section 1.6.4, one can simply work with the following

unreduced value functions:

$$\bar{V}_k^w(X_k,\omega) := \inf_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1}),\omega)\,,\ k=0,\dots,K-1 \tag{1.29a}$$

$$V_K^w(X_K,\omega) := F(S_T(\omega),Y_T(\omega)) - X_K(\omega)\,, \tag{1.29b}$$

serving the purpose of computing $\varrho^w(p_0) = \bar{V}_0^w(p_0)$.

Furthermore, by exploiting the Markovian risk measure assumption, one can easily reduce the representation to the following Bellman equations:

$$\tilde{V}_K^w(X_K,S_K,Y_K,\theta_K^w) := F(S_K,Y_K) - X_K\,,$$

and recursively

$$\tilde{V}_k^w(X_k,S_k,Y_k,\theta_k^w) := \inf_{\xi_k} \bar{\rho}_k(\tilde{V}(X_k+\xi_k\Delta S_{k+1},S_k+\Delta S_{k+1},Y_k+\Delta Y_{k+1},f_k(\theta_k^w)),\theta_k)\,.$$

Then, considering that

$$\bar{V}_K^w(X_K,\omega) = \tilde{V}_K^w(X_K(\omega),S_K(\omega),Y_K(\omega),\theta_K^w(\omega))\,,$$

and recursively that if $\bar{V}_{k+1}^w(X_{k+1},\omega) = \tilde{V}_{k+1}^w(X_{k+1}(\omega),S_{k+1}(\omega),Y_{k+1}(\omega),\theta_{k+1}^w(\omega))$, then we have that:

$$\begin{aligned}
\bar{V}_k^w(X_k,\omega) &= \inf_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1}),\omega)\\
&= \inf_{\xi_k} \rho_k^w(\tilde{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1},S_{k+1},Y_{k+1},\theta_{k+1}^w),\omega)\\
&= \inf_{\xi_k} \bar{\rho}_k^w(\bar{\Pi}_k(\tilde{V}_{k+1}^w(X_k+\xi_k\Delta S_{k+1},S_{k+1},Y_{k+1},\theta_{k+1}^w)),\theta_k^w(\omega))\\
&= \inf_{\xi_k} \bar{\rho}_k^w(\tilde{V}_{k+1}^w(X_k(\omega)+\xi_k\Delta S_{k+1},S_k(\omega)+\Delta S_{k+1},Y_k(\omega)+\Delta Y_{k+1},f(\theta_k^w)),\theta_k^w(\omega))\\
&= \tilde{V}_k(X_k(\omega),S_k(\omega),Y_k(\omega),\theta_k(\omega))\,.
\end{aligned}$$

Now, we see that $\varrho^w(p_0) = \bar{V}_0^w(p_0) = \tilde{V}_0^w(p_0,S_0,Y_0,\theta_0^w)$. In the case of the buyer, similar derivations lead to the Bellman equations:

$$\tilde{V}_K^b(X_K,S_K,Y_K,\theta_K^b) := -F(S_K,Y_K) - X_K$$

and

$$\tilde{V}_k^b(X_k,S_k,Y_k,\theta_k^b) := \inf_{\xi_k} \bar{\rho}_k^b(\tilde{V}_{k+1}^b(X_k-\xi_k\Delta S_{k+1},S_k+\Delta S_{k+1},Y_k+\Delta Y_{k+1},f_k(\theta_k^b)),\theta_k^b)\,,$$

which can be used to compute $\varrho^b(p_0) = \tilde{V}_0^b(p_0,S_0,Y_0,\theta_0^b)$.

1.6.6 Proof of Lemma 1.3.2

This can be shown by contradiction. Let us assume that $p_{\mathrm{c}}^* > p_{\mathrm{nc}}^*$ and denote with τ_{c}^* a risk minimizing stopping time strategy for the buyer when the price of the option is set to p_{c}^*. One can straightforwardly establish that either:

$$\varrho^w(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*) \leq \sup_\tau \varrho^w(p_{\mathrm{c}}^*, \tau) \leq \varrho_\tau^w(p_{\mathrm{c}}^*) \leq \varrho_\tau^w(p_{\mathrm{nc}}^*) = \varrho_\tau^b(p_{\mathrm{nc}}^*) < \varrho_\tau^b(p_{\mathrm{c}}^*) = \varrho^b(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*)\,,$$

or

$$\varrho^w(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*) \leq \sup_\tau \varrho^w(p_{\mathrm{c}}^*, \tau) \leq \varrho_\tau^w(p_{\mathrm{c}}^*) < \varrho_\tau^w(p_{\mathrm{nc}}^*) = \varrho_\tau^b(p_{\mathrm{nc}}^*) \leq \varrho_\tau^b(p_{\mathrm{c}}^*) = \varrho^b(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*)\,,$$

where in the second inequality, we used the fact that the risk can only increase when the supremum over τ is evaluated after the hedging policy has been fixed. We also used in the following two strict inequalities the fact that both ρ^w and ρ^b are monotone and that p_{nc}^* is the unique equal risk price without commitment, which implies that:

$$\varrho_\tau^w(p_{\mathrm{c}}^*) = \varrho_\tau^w(p_{\mathrm{nc}}^*) = \varrho_\tau^b(p_{\mathrm{nc}}^*) = \varrho_\tau^b(p_{\mathrm{c}}^*)$$

is not possible. Our analysis leads to a contradiction since it implies that $\varrho^w(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*) < \varrho^b(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*)$ while by definition $\varrho^w(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*) = \varrho^b(p_{\mathrm{c}}^*, \tau_{\mathrm{c}}^*)$.

1.6.7 Proof of Lemma 1.3.3

This proof follows similar arguments as the proof of Proposition 1.2.1. In particular, one can again demonstrate that for any $p_0 \in \mathbb{R}$ and any τ, the minimal risk achievable are $\varrho^w(p_0, \tau) = \varrho^w(0, \tau) + p_0$ and $\varrho^b(p_0, \tau) = \varrho^w(0, \tau) - p_0$ because of the translation invariance property of ρ^w and ρ^b. We can then prove the two conditional statements.

First, in the case that an equal risk price p_0^* exists, based on the definition of p_0^*, there must also exist a $\tau^* \in \arg\min_\tau \varrho^b(p_0^*, \tau)$. This further implies that $\tau^* \in \arg\min_\tau \varrho^b(0, \tau) - p_0^*$ hence that $\tau^* \in \arg\min_\tau \varrho^b(0, \tau)$ and can, therefore, play the role of τ_0. Next, the definition of p_0^* also ensures that $\varrho^b(p_0^*, \tau^*) = \varrho^w(p_0^*, \tau^*) \in \mathbb{R}$ which implies that both $\varrho^w(0, \tau^*)$ and $\varrho^b(0, \tau^*)$ are finite.

Reversely, in the case that the fair price interval is bounded and $\tau_0 \in \arg\min_\tau \varrho^b(0, \tau)$ exists, then we can construct $p_0^* := (\varrho^w(0, \tau_0) - \varrho^b(0, \tau_0))/2 \in \mathbb{R}$. Necessarily, $\tau_0 \in$

$\arg\min_\tau \varrho^b(0,\tau) - p_0^* = \arg\min_\tau \varrho^b(p_0^*,\tau)$. Finally, we have that:

$$\varrho^w(p_0^*,\tau_0) = \varrho^w(0,\tau_0) - p_0^* = \frac{\varrho^w(0,\tau_0) - \varrho^b(0,\tau_0)}{2} = \varrho^b(0,\tau_0) + p_0^* = \varrho^b(p_0^*,\tau_0)\,.$$

1.6.8 Proof of Proposition 1.3.4

In the case of the writer, the argument are exactly analogous as for the proof of Proposition 1.3.1. In particular, one can use Proposition 1.3.1 and the discussion in Section 4 of Pichler and Shapiro (2018) to conclude that:

$$\varrho^w(0,\tau) = \bar{V}_0^w(0,\tau)\,,$$

with

$$\begin{aligned}
\bar{V}_k^w(X_k,\tau,\omega) &:= \inf_{\xi_k,\{\hat{\xi}^i\}_{i=0}^k} \rho_k^w(\bar{V}_{k+1}^w(X_k + (\mathbf{1}\{\tau > k\}\xi_k + \sum_{\ell=0}^k \mathbf{1}\{\tau=\ell\}\hat{\xi}_k^\ell)\Delta S_{k+1},\tau),\omega) \\
&= \inf_{\bar{\xi}_k} \rho_k^w(\bar{V}_{k+1}^w(X_k + \bar{\xi}_k\Delta S_{k+1},\tau),\omega) \\
\bar{V}_K^w(X_K,\tau,\omega) &:= F(S_{\tau(\omega)}(\omega), Y_{\tau(\omega)}(\omega)) - X_K(\omega)\,,
\end{aligned}$$

where the first argument of each $\bar{V}_k^w$ is a random variable in $\mathcal{L}_p(\Omega,\mathcal{F}_k,\mathbb{P})$. By exploiting conditional translation invariance, one then easily obtains:

$$\bar{V}_K^w(X_K,\tau,\omega) = V_K^w(\tau,\omega) + \sum_{k=0}^{K-1} \mathbf{1}\{\tau(\omega)=k\}F(S_k(\omega),Y_k(\omega)) - X_K(\omega)\,,$$

and recursively that:

$$\begin{aligned}
\bar{V}_k^w(X_k,\tau,\omega) &= \inf_{\bar{\xi}_k} \rho_k^w(\bar{V}_{k+1}^w(X_k + \bar{\xi}_k\Delta S_{k+1},\tau),\omega) \\
&= \inf_{\bar{\xi}_k} \rho_k^w(V_{k+1}^w(\tau) + \sum_{\ell=0}^k \mathbf{1}\{\tau=\ell\}F(S_\ell,Y_\ell) - X_k - \bar{\xi}_k\Delta S_{k+1},\omega) \\
&= \inf_{\bar{\xi}_k} \rho_k^w(V_{k+1}^w(\tau) - \bar{\xi}_k\Delta S_{k+1},\omega) + \sum_{\ell=0}^k \mathbf{1}\{\tau(\omega)=\ell\}F(S_\ell(\omega),Y_\ell(\omega)) - X_k(\omega) \\
&= V_k^w(\tau,\omega) + \sum_{\ell=0}^{k-1} \mathbf{1}\{\tau(\omega)=\ell\}F(S_\ell(\omega),Y_\ell(\omega)) - X_k(\omega)\,.
\end{aligned}$$

Hence, we have that:

$$\varrho^w(0,\tau) = \bar{V}_0^w(0,\tau) = V_0^w(\tau)\,.$$

In the case of the buyer's equations, the proof is more challenging yet follows similar arguments[3]. In particular, we first define an operator that computes the minimal risk under an initial price of p_0, and presents an equivalent reformulation:

$$\begin{aligned}\varrho^b(p_0) &:= \min_{\tau} \inf_{X \in \bar{\mathcal{X}}_\tau(p_0)} \rho^b(F(S_\tau, Y_\tau) - X_K) \\ &= \inf_{Z \in \mathcal{Z},\, X \in \bar{\mathcal{X}}(p_0, Z)} \rho^b(\sum_{k=0}^{K-1} Z_k F(S_K, Y_K) - X_K)\,,\end{aligned}$$

where $\mathcal{Z} := \{Z : \Omega \to \{0,1\}^K \mid \sum_{k=0}^{K-1} Z_k \leq 1\}$ and each Z_k is $\mathbb{F}_k$-adapted and captures $Z_k := \mathbf{1}\{\tau = k\}$, and where

$$\bar{\mathcal{X}}(p_0, Z) := \left\{ X : \Omega \to \mathbb{R}^K \;\middle|\; \begin{array}{l} \exists X_0 = p_0, \forall k = 1, \ldots, K-1, \exists \xi_k, \{\hat{\xi}_k^i\}_{i=0}^k \\ X_{k+1} = X_k + (\xi_k + \sum_{i=0}^k (\hat{\xi}_k^i - \xi_k) Z_i) \Delta S_{k+1} \end{array} \right\}.$$

Once again, we use the arguments in Pichler and Shapiro (2018) to conclude that:

$$\varrho^b(0) = \bar{V}_0^b(0)\,,$$

with

$$\begin{aligned}&\bar{V}_K^b(X_K, Z_{0:K-1}, \omega) \\ &\quad := -\sum_{k=0}^{K-1} Z_k(\omega) F(S_k(\omega), Y_k(\omega)) - (1 - \sum_{k=0}^{K-1} Z_k(\omega)) F(S_K(\omega), Y_K(\omega)) - X_K(\omega) \\ &\bar{V}_k^b(X_k, Z_{0:k-1}, \omega) \\ &\quad := \inf_{Z_k, \xi_k, \{\hat{\xi}^i\}_{i=0}^k : Z_k \leq 1 - \sum_{\ell=0}^{k-1} Z_\ell} \rho_k^b(\bar{V}_{k+1}^b(X_k + ((1 - \sum_{\ell=0}^{k} Z_\ell)\xi_k + \sum_{\ell=0}^{k} Z_k \hat{\xi}_k^\ell)\Delta S_{k+1}, Z_{0:k}), \omega) \\ &\quad = \inf_{Z_k, \bar{\xi}_k : Z_k \leq 1 - \sum_{\ell=0}^{k-1} Z_\ell} \rho_k^b(\bar{V}_{k+1}^b(X_k + \bar{\xi}_k \Delta S_{k+1}, Z_{0:k}), \omega)\,.\end{aligned}$$

By exploiting conditional translation invariance, one then easily obtains:

$$\bar{V}_K^b(X_K, Z_{0:K-1}, \omega) = V_K^b(\sum_{k=0}^{K-1} Z_k, \omega) - \sum_{k=0}^{K-1} Z_k(\omega) F(S_k(\omega), Y_k(\omega)) - X_K(\omega)\,,$$

[3]Note that here we diverge from the arguments used in Section 6.1.2 of Pichler and Shapiro (2018) to simplify exposition

and recursively that:

$$\begin{aligned}
\bar{V}_k^b(X_k, Z_{0:k-1}, \omega) &= \inf_{Z_k, \bar{\xi}_k : Z_k \leq 1 - \sum_{\ell=0}^{k-1} Z_\ell} \rho_k^b(\bar{V}_{k+1}^w(X_k + \bar{\xi}_k \Delta S_{k+1}, Z_{0:k}), \omega) \\
&= \inf_{Z_k, \bar{\xi}_k : Z_k \leq 1 - \sum_{\ell=0}^{k-1} Z_\ell} \rho_k^b(V_{k+1}^b(\sum_{\ell=0}^{k} Z_k) - \sum_{\ell=0}^{k} Z_\ell F(S_\ell, Y_\ell) - X_k - \bar{\xi}_k \Delta S_{k+1}, \omega) \\
&= \inf_{Z_k, \bar{\xi}_k : Z_k \leq 1 - \sum_{\ell=0}^{k-1} Z_\ell} \rho_k^b(V_{k+1}^b(\sum_{\ell=0}^{k} Z_k) - \bar{\xi}_k \Delta S_{k+1}, \omega) - \sum_{\ell=0}^{k} Z_\ell(\omega) F(S_\ell(\omega), Y_\ell(\omega)) - X_k(\omega) \\
&= V_k^b(\sum_{\ell=0}^{k-1} Z_k, \omega) - \sum_{\ell=0}^{k-1} Z_\ell(\omega) F(S_\ell(\omega), Y_\ell(\omega)) - X_k(\omega)\,.
\end{aligned}$$

Hence, we have that $\varrho^b(0) = \bar{V}_0^b(0)$. Optimal policies for each problem can be identified similarly as was done in the proof of Proposition 1.3.1.

1.6.9 Proof of Lemma 1.3.5

This proof mainly relies on the translation invariance property together with the property that once again $\bar{\mathcal{X}}_\tau(p_0) = p_0 + \bar{\mathcal{X}}_\tau(0)$. These two properties can be used to show that:

$$\begin{aligned}
\varrho_\tau^w(p_0) &= \inf_{X \in \bar{\mathcal{X}}_\tau(0) + p_0} \sup_\tau \rho^w(F(S_\tau, Y_\tau) - X_K(\tau)) \\
&= \inf_{X \in \bar{\mathcal{X}}_\tau(0)} \sup_\tau \rho^w(F(S_\tau, Y_\tau) - X_K(\tau) - p_0) \\
&= \inf_{X \in \bar{\mathcal{X}}_\tau(0)} \sup_\tau \rho^w(F(S_\tau, Y_\tau) - X_K(\tau)) - p_0 \\
&= \inf\{s | \inf_{X \in \bar{\mathcal{X}}_\tau(0)} \sup_\tau \rho^w(F(S_\tau, Y_\tau) - X_K(\tau)) \leq s\} - p_0 \\
&= \inf\{s | \varrho_\tau^w(s) \leq 0\} - p_0 = p_0^w - p_0\,. \qquad (1.30)
\end{aligned}$$

Similarly for the buyer, we have that $\varrho_\tau^b(p_0) = -p_0^b + p_0$. The rest follows as in the proof of Proposition 1.2.1.

1.6.10 Proof of Proposition 1.3.6

This proof focuses on the case of the writer given that the buyer's problem was already studied in the proof of Proposition 1.3.4. In particular, the proof follows similar lines as in Section 6.1.1 of Pichler and Shapiro (2018) and in fact extends that result to a case where hedging is allowed to pursue past the exercise time up to the end of the horizon. We start by considering that the self-financing hedging policy described by ξ and $\hat{\xi}$ is fixed

and reformulate the worst-case exercise time problem. We then look into reformulating the optimization of the hedging policy.

Step 1 (Worst-case exercise time problem): For any fixed hedging strategy, one can define the writer's worst-case exercise time problem as:

$$\nu_0 := \sup_{\tau} \rho^w \left(F(S_\tau, Y_\tau) - \sum_{\ell=0}^{\tau-1} \xi_\ell \Delta S_{\ell+1} - \sum_{\ell=\tau}^{K-1} \hat{\xi}_\ell^\tau \Delta S_{\ell+1} \right) .$$

In order to find a dynamic programming formulation for this problem, we start by reformulating it in the form of an "optimal stopping problem" as defined in Pichler and Shapiro (2018). In particular, we can consider that:

$$\nu_0 = \sup_{\tau} \rho^w(E_\tau^1) \,,$$

where $E_k^1(\omega) := \rho_{k,K}^w(F_k(S_k, Y_k) - \sum_{\ell=0}^{k-1} \xi_\ell \Delta S_{\ell+1} - \sum_{\ell=k}^{K-1} \hat{\xi}_\ell^k \Delta S_{\ell+1}, \omega)$ with $\rho_{k,K}^w(X) := \rho_k^w(\rho_{k+1}^w(\cdots \rho_{K-1}(X) \cdots)$. Hence, based on Theorem 6.4 in Pichler and Shapiro (2018), we can conclude that if we define:

$$\begin{aligned} E_K^2(\omega) &:= E_K^1(\omega) \\ E_k^2(\omega) &:= \max(E_k^1(\omega),\, \rho_k(E_{k+1}^2, \omega)) \,, \end{aligned}$$

and

$$\tau_m^*(\omega) := \min\{k \,|\, E_k^2(\omega) = E_k^1(\omega),\, m \le t \le T\} \,,$$

then τ_m^* is an optimal solution to:

$$\sup_{\tau:\tau \ge m} \rho^w(F(S_\tau, Y_\tau) - \sum_{\ell=0}^{\tau-1} \xi_\ell \Delta S_{\ell+1} - \sum_{\ell=\tau}^{K-1} \hat{\xi}_\ell^\tau \Delta S_{\ell+1}) \,,$$

and $\nu_0 = E_0^2$.

Step 2 (Optimal hedging optimization problem): Based on our analysis of the worst-case exercise time problem, we have found that the optimal hedging problem has the following form:

$$\begin{aligned} \varrho_\tau^w(0) &= \inf_{\xi, \hat{\xi}} E_0^2(\xi, \hat{\xi}) \\ &= \inf_{\xi, \hat{\xi}} \mathcal{R}_{0,1}(E_0^1(\hat{\xi}^0), \mathcal{R}_{1,2}(E_1^1(\xi_0, \hat{\xi}^1), \cdots, \mathcal{R}_{K-1,K}(E_{K-1}^1(\xi_{0:K-1}, \hat{\xi}^{K-1}), E_K^1(\xi)) \,, \end{aligned}$$

where $\mathcal{R}_{k,k+1}(X,Y,\omega) := \max(X(\omega), \rho_k^w(Y,\omega))$ and where we made explicit the influence of ξ and $\hat{\xi}$ on each E_k^1. As argued in Pichler and Shapiro (2018), given that each $\mathcal{R}_{t,t+1}(\cdot,\cdot)$ is monotone, one can apply the interchangeability principle to generate the following reformulation:

$$\varrho_\tau^w(0) = \mathcal{R}_{0,1}(\inf_{\hat{\xi}^0} E_0^1(\hat{\xi}^0), \inf_{\xi_0} \mathcal{R}_{1,2}(\inf_{\hat{\xi}^1} E_1^1(\xi_0, \hat{\xi}^1), \cdots \\ \inf_{\xi_{K-1}} \mathcal{R}_{K-1,K}(\inf_{\hat{\xi}^{K-1}} E_{K-1}^1(\xi_{0:K-1}, \hat{\xi}^{K-1}), E_K^1(\xi)) \cdots) .$$

Based on this argument, we create the following operators:

$$\begin{aligned}
\bar{V}_k^w(1, X_k, \omega) &:= \inf_{\hat{\xi}_{k:K-1}^k} \rho_{k,K}^w(F_k(S_k, Y_k) - X_k - \sum_{\ell=k}^{K-1} \hat{\xi}_\ell^k \Delta S_{\ell+1}, \omega) \\
\bar{V}_k^w(0, X_k, \omega) &:= \max(\bar{V}_k^w(1, X_k, \omega), \inf_{\xi_k} \rho_k^w(\bar{V}_{k+1}^w(0, X_k + \xi_k \Delta S_{k+1}), \omega)) \\
\bar{V}_K^w(0, X_K, \omega) &:= F_K(S_K(\omega), Y_K(\omega)) - X_K(\omega) ,
\end{aligned}$$

in order to have that $\varrho_\tau^w(0) = \bar{V}_0^w(0,0)$, and where we again let the second argument of $\bar{V}_k^w$ be a random variable in $\mathcal{L}_p(\Omega, \mathcal{F}_k, \mathbb{P})$.

In the case of $\bar{V}_k^w(1, X_k, \omega)$, we can further apply the interchangeability principle to get that for all $k = 0, \ldots, K-1$,

$$\begin{aligned}
\bar{V}_k^w(1, X_k, \omega) &= \inf_{\hat{\xi}_k^k} \rho_k^w(\inf_{\hat{\xi}_{k+1}^k} \rho_{k+1}^w(\cdots \inf_{\hat{\xi}_{K-1}^k} \rho_{K-1}^w(F_k(S_k, Y_k) - X_k - \sum_{\ell=k}^{K-1} \hat{\xi}_\ell^k \Delta S_{\ell+1}) \cdots), \omega) \\
&= F_k(S_k(\omega), Y_k(\omega)) - X_k(\omega) + \\
&\qquad \inf_{\hat{\xi}_k^k} \rho_k^w(-\hat{\xi}_k^k \Delta S_{k+1} + \inf_{\hat{\xi}_{k+1}^k} \rho_{k+1}^w(-\hat{\xi}_{k+1}^k \Delta S_{k+2} + \cdots \inf_{\hat{\xi}_{K-1}^k} \rho_{K-1}^w(-\hat{\xi}_{K-1}^k \Delta S_K)), \omega) \\
&= F_k(S_k(\omega), Y_k(\omega)) - X_k(\omega) + V_k^w(1, \omega) ,
\end{aligned}$$

where we applied conditional translation invariance. While one can verify that we also have:

$$\begin{aligned}
\bar{V}_K^w(0, X_K, \omega) &= V_K^w(0, \omega) - X_K(\omega) \\
\bar{V}_k^w(0, X_k, \omega) &= \max(V_k^w(1, \omega) + F_k(S_k(\omega), Y_k(\omega)) - X_k(\omega), \inf_{\xi_k} \rho_k^w(V_{k+1}^w(0) - X_k - \xi_k \Delta S_{k+1}, \omega)) \\
&= \max(V_k^w(1, \omega) + F_k(S_k(\omega), Y_k(\omega)), \inf_{\xi_k} \rho_k^w(V_{k+1}^w(0) - \xi_k \Delta S_{k+1}, \omega)) - X_k(\omega) \\
&= V_k^w(0, \omega) - X_k(\omega) .
\end{aligned}$$

Hence, $\varrho_\tau^w(0) = \bar{V}_0^w(0,0) = V_0^w(0)$. An optimal hedging policy can be identified with an optimal solution to the infimum operations in equation (1.18) or (1.19) depending on whether the option was exercised at a period smaller or equal to k.

1.6.11 Proof of Corollary 1.3.7

We start by looking at the case of an American option with commitment. Based on Proposition 1.3.4, we can first prove that, given any exercise policy τ, the writer should stop hedging after exercise by studying, for each k, whether $\xi_k^{i*}(\omega) = 0$ is optimal when $\tau(\omega) \leq k$. Specifically, if $\tau(\omega) \leq k$, then we have that:

$$\begin{aligned}
&\arg\min_{\xi_k} \rho_k^w(V_{k+1}^w(\tau) - \xi_k \Delta S_{k+1}, \omega) \\
&\qquad = \arg\min_{\xi_k} \inf_{\xi_{k+1:K-1}} \rho_{k,K}^w\Big(\sum_{\ell=k+1}^{K} \mathbf{1}\{\tau = \ell\} F(S_\ell, Y_\ell) - \sum_{\ell=k}^{K-1} \xi_\ell \Delta S_{\ell+1}, \omega\Big) \\
&\qquad = \arg\min_{\xi_k} \inf_{\xi_{k+1:K-1}} \rho_{k,K}^w\Big(-\sum_{\ell=k}^{K-1} \xi_\ell \Delta S_{\ell+1}, \omega\Big) \;\supseteq\; \{0\}\,,
\end{aligned}$$

since the ρ_k^w satisfies the bounded conditional market risk assumption and is conditionally coherent, thus $\inf_{\xi_{k:K-1}} \rho_{k,K}^w(-\sum_{\ell=k}^{K-1} \xi_\ell \Delta S_{\ell+1}, \omega) = 0$. This confirms that if $\tau(\omega) \leq k$, then $\hat{\xi}_k^{i*}(\tau, \omega) := 0$ is optimal for all $i \leq k$ so that the number of shares of the risky asset becomes:

$$\xi_k \mathbf{1}\{\tau > k\} + \sum_{i=0}^{k} \hat{\xi}_k^i \mathbf{1}\{\tau = i\} = \xi_k \cdot 0 + \sum_{i=0}^{k} 0 \cdot \mathbf{1}\{\tau = i\} = 0\,.$$

In other words, it is optimal to stop hedging at $\tau(\omega)$.

A similar argument can be used for the buyer. Namely, we can study, for each k, the structure of $\xi_k^{i*}(\omega)$ when $\tau_0(\omega) \leq k$. This is done as follows:

$$\begin{aligned}
&\arg\min_{\xi_k} \rho_k^b(-\xi_k \Delta S_{k+1} + V_{k+1}^b(\mathbf{1}\{\tau_0 \leq k\}), \omega) \\
&\qquad = \arg\min_{\xi_k} \inf_{\xi_{k+1:K-1}} \rho_{k:K}^b\Big(-\sum_{\ell=k}^{K-1} \xi_\ell \Delta S_{\ell+1}, \omega\Big) \;\supseteq\; \{0\}\,,
\end{aligned}$$

which again implies that it is optimal to stop hedging at τ_0.

For the case of an American option without commitment, the same argument as for the with commitment case applies for the buyer. On the other hand, for the writer we can

retrieve the optimal hedging policy from Proposition 1.3.6. Looking carefully, for each k, at the structure of $\xi_k^{i*}(\omega)$ when $\tau(\omega) \leq k$, we realize that the same arguments apply

$$\arg\min_{\xi_k} \rho_k^w(-\xi_k \Delta S_{k+1} + V_{k+1}^w(\mathbf{1}\{\tau \leq k\}), \omega) = \arg\min_{\xi_k} \inf_{\xi_{k+1:K-1}} \rho_{k,K}^w \left(-\sum_{\ell=k}^{K-1} \xi_\ell \Delta S_{\ell+1}, \omega \right) \supseteq \{0\}.$$

Hence, once again it is optimal to stop hedging starting at τ.

1.6.12 Verifying the Bounded (Conditional) Market Risk Property for Worst-case Risk Measures

In this section we identify sufficient conditions under which the one-step decomposition of a worst-case risk measure satisfies the bounded conditional market risk property. However, before studying such conditions we need to first define a useful projection operator.

Definition 9. Given an uncertainty set $\mathcal{U} \subseteq \mathbb{R}^K$, and a history of observations $\hat{r}_{1:k-1} \in \mathbb{R}^{k-1}$, we define the operation of projecting $\mathcal{U}$ over the time interval $\{k, \ldots, k'\}$ with $1 \leq k \leq k' \leq K$ as follows:

$$\mathcal{U}_{k:k'}(\hat{r}_{1:k-1}) := \left\{ r \in \mathbb{R}^{k'-k+1} \middle| \begin{array}{ll} \text{If } k' < K, & \exists \bar{r} \in \mathbb{R}^{K-k'}, [\hat{r}_{1:k-1}^T \;\; r^T \;\; \bar{r}^T]^T \in \mathcal{U} \\ \text{If } k' = K, & [\hat{r}_{1:k-1}^T \;\; r^T]^T \in \mathcal{U} \end{array} \right\}.$$

This definition is helpful in describing, for a given worst-case risk measure that exploits some uncertainty set $\mathcal{U}$, the set of all realizations of the return vector for which the bounded conditional market risk property is satisfied.

Definition 10. Given a worst-case risk measure using an uncertainty set $\mathcal{U} \subseteq \mathbb{R}^K$, we define the set of returns with bounded conditional market risk as follows:

$$\mathcal{A}(\mathcal{U}) := \left\{ r \in \mathbb{R}^K \middle| \forall k \in \{0, ..., K-1\}, \inf_{\zeta_k, \ldots, \zeta_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \zeta_\ell r_{\ell+1}, r) \in]-\infty, \; 0] \right\}. \tag{1.31}$$

In particular, one can also reformulate the definition of $\mathcal{A}(\mathcal{U})$ as follows:

$$\mathcal{A}(\mathcal{U}) := \left\{ r \in \mathbb{R}^K \middle| \begin{array}{c} \forall k \in \{0, ..., K-1\}, \\ \mathcal{U}_{k+1:K}(r_{1:k}) = \emptyset \;\vee\; \inf_{\zeta_k, \ldots, \zeta_{K-1}} \sup_{\bar{r}_{k+1:K} \in \mathcal{U}_{k+1:K}(r_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_\ell \bar{r}_{\ell+1} \in]-\infty, \; 0] \end{array} \right\},$$

where $\bar{r}_{k+1:K}$ refers to a vector in $\mathbb{R}^{K-k-1}$ with indexes in the range $\{k+1,\ldots,K\}$, and where we use ζ_ℓ as shorthand notation for $\zeta_\ell(\bar{r}_{k:\ell})$. We will repeat this abuse of notation throughout the section to simplify the presentation of equations.

Based on Definition 10, it is clear that a worst-case risk measure will satisfy the bounded conditional market risk condition if $\mathcal{U} \subseteq \mathcal{A}(\mathcal{U})$. This is formally stated by the following lemma.

Lemma 1.6.2. *Let ρ be a worst-case risk measure that uses an uncertainty set $\mathcal{U} \subseteq \mathbb{R}^K$ such that $\mathcal{U} \subseteq \mathcal{A}(\mathcal{U})$, then ρ necessarily satisfies the bounded conditional market risk property.*

Proof. This result simply follows from the fact that for any $r \in \mathbb{R}^K$ and any $k \in \{1,\ldots,K\}$, two situation can occur. First, the set $\mathcal{U}_{k+1:K}(r_{1:k})$ might be empty, which leads to $\inf_{\zeta_k,\ldots,\zeta_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \zeta_\ell r_{\ell+1}, r) = 0$ by the definition of $\rho_k(X,r)$ thus the market risk is bounded for this realization. Secondly, one should investigate the case where $\mathcal{U}_{k+1:K}(r)$ is non-empty. In this case, there exists a $\hat{r} \in \mathcal{U} \subseteq \mathcal{A}(\mathcal{U})$ such that $r_{1:k} = \hat{r}_{1:k}$. Hence, one can verify that:

$$\begin{aligned}\inf_{\zeta_k,\ldots,\zeta_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \zeta_\ell r_{\ell+1}, r) &= \inf_{\zeta_k,\ldots,\zeta_{K-1}} \sup_{\bar{r}_{k+1:K} \in \mathcal{U}_{k+1:K}(r_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_\ell \bar{r}_{\ell+1} \\ &= \inf_{\zeta_k,\ldots,\zeta_{K-1}} \sup_{\bar{r}_{k+1:K} \in \mathcal{U}_{k+1:K}(\hat{r}_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_k \bar{r}_{k+1} \in \left]-\infty, 0\right],\end{aligned}$$

based on the fact that $\hat{r} \in \mathcal{A}(\mathcal{U})$. This implies that the conditional market risk is bounded on all of $\mathbb{R}^K$ which is a stronger condition than in Assumption 1.3.2 where the condition is only imposed with probability one. □

Based on the above discussion, given an arbitrary uncertainty set which might not satisfy the condition $\mathcal{U} \subseteq \mathcal{A}(\mathcal{U})$, it, therefore, appears that we are in need of a procedure that would select a subset $\mathcal{U}'$ of $\mathcal{U}$ for which this property is satisfied. One attractive candidate takes the form of the following set which we will call the no-arbitrage subset of $\mathcal{U}$, when it exists.

Definition 11. Given an uncertainty set $\mathcal{U}$, we define the no-arbitrage subset $\mathcal{U}^{na}$ of $\mathcal{U}$ as the largest set $\mathcal{U}' \subseteq \mathcal{U}$ that satisfies $\mathcal{U}' \subseteq \mathcal{A}(\mathcal{U}')$. Mathematically, $\mathcal{U}^{na}$ satisfies the following two properties:

1. $\mathcal{U}^{na} \subseteq \mathcal{A}(\mathcal{U}^{na})$

2. $\forall \mathcal{U}' \subseteq \mathcal{U}, \mathcal{U}' \subseteq \mathcal{A}(\mathcal{U}')$ we have that $\mathcal{U}' \subseteq \mathcal{U}^{na}$.

Considering the previous definitions, one might wonder if such a no-arbitrage subset always exists. The following theorem confirms that it does always exist when $\mathcal{U}$ is both closed and convex.

Theorem 1.6.3. *Given that $\mathcal{U}$ is convex and closed, the no-arbitrage subset of $\mathcal{U}$ is equal to $\mathcal{V}(\mathcal{U}) \cap \mathcal{U}$ where*

$$\mathcal{V}(\mathcal{U}) = \left\{ r \in \mathbb{R}^K \middle| 0 \in \mathcal{U},\ \left(\sum_{j=1}^{k} e_j r_j \in \mathcal{U} \right) \vee \left(r_{1:k} \notin \mathcal{U}_{1:k} \right), \forall k = 1, ..., K-1 \right\},$$

with $e_j \in \mathbb{R}^K$ as the j-th column of identity matrix, and where $\mathcal{V}(\mathcal{U}) = \emptyset$ if $0 \notin \mathcal{U}$.

Proof. The proof of this theorem is divided in four parts. First, we show that $\mathcal{V}(\mathcal{U}) = \mathcal{A}(\mathcal{U})$. This step is itself divided in two parts, namely first that $\mathcal{V}(\mathcal{U}) \subseteq \mathcal{A}(\mathcal{U})$ and then that $\mathcal{V}(\mathcal{U}) \supseteq \mathcal{A}(\mathcal{U})$. The second step consists in proving that $\mathcal{V}(\mathcal{U}) \cap \mathcal{U}$ satisfies the two conditions of the no-arbitrage subset $\mathcal{U}^{na}$.

Step 1.a ($\mathcal{V}(\mathcal{U}) \subseteq \mathcal{A}(\mathcal{U})$). Given any member r of $\mathcal{V}(\mathcal{U})$, we know that for all $k = 1, \ldots, K$, either $r_{1:k} \notin \mathcal{U}_{1:k}$ which leads to:

$$\inf_{\zeta_k, \ldots, \zeta_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \zeta_\ell r_{\ell+1}, r) = 0\,,$$

by definition. Otherwise, the vector $[r_{1:k-1}^T \ \ 0_{1:K-k+1}^T]^T \in \mathcal{U}$ thus we can conclude that:

$$\inf_{\zeta_k, \ldots, \zeta_{K-1}} \sup_{\bar{r}_{k+1:K} \in \mathcal{U}_{k+1:K}(r_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_\ell r_{\ell+1} \geq \inf_{\zeta_k, \ldots, \zeta_{K-1}} -\sum_{\ell=k}^{K-1} \zeta_\ell \cdot 0 = 0 > -\infty\,.$$

From this, we conclude that $\mathcal{V}(\mathcal{U}) \subseteq \mathcal{A}(\mathcal{U})$.

Step 1.b ($\mathcal{A}(\mathcal{U}) \subseteq \mathcal{V}(\mathcal{U})$). Given any member r of $\mathcal{A}(\mathcal{U})$, for any $k = 1, \ldots, K-1$, we have that:

$$\inf_{\zeta_k, \ldots, \zeta_{K-1}} \rho_{k,K}(-\sum_{\ell=k}^{K} \zeta_\ell r_{\ell+1}, r) > -\infty\,.$$

This means that either $r_{1:k} \notin \mathcal{U}_{1:k}$ or

$$\inf_{\zeta_k,\dots,\zeta_{K-1}} \sup_{\bar{r}_{k+1:K} \in \mathcal{U}_{k+1:K}(r_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_\ell \bar{r}_{\ell+1} > -\infty \,.$$

We can further process this second condition by rewriting it as:

$$\inf_{\zeta_k} \sup_{\bar{r}_{k+1} \in \mathcal{U}_{k+1}(r_{1:k})} \zeta_k \bar{r}_{k+1} + \pi_{k+1}([r_{1:k}^T \;\; \bar{r}_{k+1}]^T) > -\infty \,,$$

where $\mathcal{U}_{k+1}(r_{1:k})$ is short for $\mathcal{U}_{k+1:k+1}(r_{1:k})$, and with

$$\pi_k(r_{1:k}) := \inf_{\zeta_k} \sup_{\bar{r}_{k+1} \in \mathcal{U}_{k+1}(r_{1:k})} \zeta_k \bar{r}_{k+1} + \pi_{k+1}([r_{1:k}^T \bar{r}_{k+1}]^T) \,,$$

for all $k = 0, \dots, K-1$ while $\pi_K(r) := 0$. Yet, one quickly realizes that:

$$\begin{aligned}
\pi_{K-1}(r_{1:K-1}) &= \inf_{\zeta_{K-1}} \sup_{\bar{r}_K \in \mathcal{U}_K(r_{1:K-1})} \zeta_{K-1} \bar{r}_K \\
&= \inf_{\zeta_{K-1}} \max \left(\zeta_{K-1} \inf_{\bar{r}_K \in \mathcal{U}_K(r_{1:K-1})} \bar{r}_K,\; \zeta_{K-1} \sup_{\bar{r}_K \in \mathcal{U}_K(r_{1:K-1})} \bar{r}_K \right) \\
&= \begin{cases} 0 & \text{if } 0 \in \mathcal{U}_K(r_{1:K-1}) \\ -\infty & \text{otherwise} \end{cases} ,
\end{aligned}$$

where the second equality follows from the fact that $\mathcal{U}_K(r_{1:K-1})$ is a closed interval given that $\mathcal{U}$ is convex and closed. The third equality comes from the fact if $0 \in \mathcal{U}_K(r_{1:K-1})$ then the infimum over ζ_k is reached by $\zeta_k = 0$, while when it $\mathcal{U}_K(r_{1:K-1})$ does not include zero, then the infimum can reached an arbitrarily low value since the sign of $\bar{r}_K$ is determined. Consequently, by induction, for any $k = 0, \dots, K-1$, it must actually be that:

$$\begin{aligned}
\pi_k(r_{1:k}) &= \inf_{\zeta_k} \sup_{\bar{r}_{k+1} \in \mathcal{U}_{k+1}(r_{1:k})} \zeta_k \bar{r}_{k+1} + \pi_{k+1}([r_{1:k}^T \;\; \bar{r}_{k+1}]^T) \\
&= \inf_{\zeta_k} \max \left(\zeta_k \inf_{\bar{r}_{k+1}:[\bar{r}_{k+1} \;\; 0_{k+2:K}^T]^T \in \mathcal{U}_{k+1:K}(r_{1:k})} \bar{r}_{k+1},\; \zeta_k \sup_{\bar{r}_{k+1}:[\bar{r}_{k+1} \;\; 0_{k+2:K}^T]^T \in \mathcal{U}_{k+1:K}(r_{1:k})} \bar{r}_{k+1} \right) \\
&= \begin{cases} 0 & \text{if } 0 \in \mathcal{U}_{k+1:K}(r_{1:k}) \\ -\infty & \text{otherwise} \end{cases} .
\end{aligned}$$

Based of this argument, we must, therefore, conclude that if $r \in \mathcal{A}(\mathcal{U})$, then for all k, either $r_{1:k} \notin \mathcal{U}_{1:k}$ or $\pi_k(r_{1:k}) > -\infty$ hence that $0 \in \mathcal{U}_{k+1:K}(r_{1:k})$. Overall, this confirms that $r \in \mathcal{V}(\mathcal{U})$.

Step 2.a ($\mathcal{U}^{na} \subseteq \mathcal{A}(\mathcal{U}^{na})$). To prove this property, we need to show that:

$$\mathcal{U} \cap \mathcal{A}(\mathcal{U}) \subseteq \mathcal{A}(\mathcal{U} \cap \mathcal{A}(\mathcal{U}))\,,$$

which is equivalent to showing that:

$$\mathcal{U} \cap \mathcal{V}(\mathcal{U}) \subseteq \mathcal{V}(\mathcal{U} \cap \mathcal{V}(\mathcal{U}))\,,$$

since $\mathcal{U}$ is convex and closed so that $\mathcal{A}(\mathcal{U}) = \mathcal{V}(\mathcal{U})$ and, therefore, $\mathcal{U} \cap \mathcal{A}(\mathcal{U}) = \mathcal{U} \cap \mathcal{V}(\mathcal{U})$ is also convex and closed so that $\mathcal{A}(\mathcal{U} \cap \mathcal{A}(\mathcal{U})) = \mathcal{V}(\mathcal{U} \cap \mathcal{A}(\mathcal{U})) = \mathcal{V}(\mathcal{U} \cap \mathcal{V}(\mathcal{U}))$. We will tackle the second equivalent condition, where we will make use the following representation:

$$\mathcal{V}(\mathcal{V}(\mathcal{U}) \cap \mathcal{U}) = \left\{ r \in \mathbb{R}^K \middle| \sum_{j=1}^{k} e_j r_j \in \mathcal{V}(\mathcal{U}) \cap \mathcal{U} \ \vee \ \left(r_{1:k} \notin (\mathcal{V}(\mathcal{U}) \cap \mathcal{U})_{1:k}\right), \forall k = 1, ..., K-1 \right\}.$$

Specifically, given any $r \in \mathcal{U} \cap \mathcal{V}(\mathcal{U}) \subseteq \mathbb{R}^K$, and for all $k = 1, \ldots, K$, we will confirm that $\sum_{j=1}^{k} e_j r_j \in \mathcal{V}(\mathcal{U}) \cap \mathcal{U}$. We can first check that:

$$\sum_{j=1}^{k} e_j r_j \in \mathcal{U}\,,$$

since $r \in \mathcal{V}(\mathcal{U})$. Furthermore, letting $w := \sum_{j=1}^{k} e_j r_j$, we can further check that for all $\ell = 1, \ldots, K$,

$$\sum_{i=1}^{\ell} e_i w_i = \sum_{i=1}^{\ell} e_i e_i^T \left(\sum_{j=1}^{k} e_j r_j \right) = \sum_{j=1}^{k} \sum_{i=1}^{\ell} e_i e_i^T e_j r_j = \sum_{j=1}^{\min(k,\ell)} e_j r_j \in \mathcal{U}\,,$$

since again $r \in \mathcal{V}(\mathcal{U})$, which implies that $w \in \mathcal{U} \cap \mathcal{V}(\mathcal{U})$. Based on these arguments, we can conclude that $r \in \mathcal{V}(\mathcal{U} \cap \mathcal{V}(\mathcal{U}))$.

Step 2.b ($\mathcal{U}^{na}$ is the largest). The second property is proved as follows:

$$\mathcal{U}' \subseteq \mathcal{U} \Rightarrow \mathcal{U}' \cap \mathcal{V}(\mathcal{U}') \subseteq \mathcal{U} \cap \mathcal{V}(\mathcal{U}) \Rightarrow \mathcal{U}' = \mathcal{U}' \cap \mathcal{A}(\mathcal{U}') = \mathcal{U}' \cap \mathcal{V}(\mathcal{U}') \subseteq \mathcal{U} \cap \mathcal{V}(\mathcal{U})\,,$$

where the first implication comes from the definition of $\mathcal{V}(\mathcal{U}')$ and the fact that $\mathcal{U}' \subseteq \mathcal{U}$. The second implication first exploits the fact that $\mathcal{U}' \subseteq \mathcal{A}(\mathcal{U}')$ and then that $\mathcal{A}(\mathcal{U}') = \mathcal{V}(\mathcal{U}')$. This concludes the proof.

□

Bounded Conditional Market Risk Property for $\mathcal{U}_1$

Exploiting the result of Theorem 1.6.3, we can now provide a proof of Lemma 1.4.1. Specifically, since $\mathcal{U}_1$ is a closed convex set, the theorem provides a recipe to construct the no-arbitrage subset of $\mathcal{U}_1$, i.e. $\mathcal{U}^{na} := \mathcal{U} \cap \mathcal{V}(\mathcal{U})$. In the context that is studied $\mathcal{V}(\mathcal{U})$ reduces to

$$\mathcal{V}(\mathcal{U}) = \left\{ r \in \mathbb{R}^K \;\middle|\; \begin{array}{c} \max_{k\in\{1,\dots,K\}} \mu\sqrt{kT/K}/\sigma - \Gamma \leq 0 \\ \left(\max_{k'\in\{k,\dots,K\}} \left|\frac{\sum_{\ell=1}^{k}\log(1+r_\ell)-\mu k'T/K}{\sigma\sqrt{k'T/K}}\right| - \Gamma \leq 0\right) \vee (r_{1:k} \notin \mathcal{U}_{1:k}), \forall k \in \{1,\dots,K\} \end{array} \right\}$$

One can easily verify that $\mathcal{W} \cap \mathcal{U} = \mathcal{V}(\mathcal{U}) \cap \mathcal{U}$ under the conditions of Lemma 1.4.1.

Bounded Conditional Market Risk Property for $\mathcal{U}_2$

In this section, we show that the worst-case risk measure defined based on the set $\mathcal{U}_2$ satisfies the bounded conditional market risk property by showing that $\mathcal{A}(\mathcal{U}_2) = \mathbb{R}^K$ and exploiting Lemma 1.6.2. Specifically, we can show that for all $r \in \mathbb{R}^K$ and all $k \in \{0,\dots,K-1\}$, either $r_{1:k} \notin \mathcal{U}_{1:k}$ or

$$\pi_k(r_{1:k}) := \inf_{\zeta_k,\dots,\zeta_{K-1}} \sup_{\bar{r}_{k+1:K}\in\mathcal{U}_{k+1:K}(r_{1:k})} -\sum_{\ell=k}^{K-1} \zeta_\ell \bar{r}_{\ell+1} = 0\,.$$

Indeed, when $r_{1:k} \notin \mathcal{U}_{1:k}$, one can rewrite

$$\pi_k(r_{1:k}) = \inf_{\zeta_k} \sup_{\bar{r}_{k+1}\in\mathcal{U}_{k+1}(r_{1:k})} -\zeta_k \bar{r}_{k+1} + \pi_{k+1}([r_{1:k}^T \;\; \bar{r}_{k+1}]^T)\,,$$

where

$$\pi_{K-1}(r_{1:K-1}) = \inf_{\zeta_{K-1}} \sup_{\bar{r}_K\in\mathcal{U}_{K:K}(r_{1:K-1})} -\zeta_{K-1}\bar{r}_K\,.$$

Given that for all k, the function π_k is evaluated with some non-empty and symmetric $\mathcal{U}_{k+1}(r_{1:k})$, we thus have that:

$$\pi_{K-1}(r_{1:K-1}) = \inf_{\zeta_{K-1}} |\zeta_{K-1}| \sup_{\bar{r}_K\in\mathcal{U}_{K:K}(r_{1:K-1})} \bar{r}_K = 0\,,$$

and recursively for $k = K-1,\dots,0$,

$$\begin{aligned}\pi_k(r_{1:k}) &= \inf_{\zeta_k} \sup_{\bar{r}_{k+1}\in\mathcal{U}_{k+1}(r_{1:k})} -\zeta_k\bar{r}_{k+1} + \pi_{k+1}([r_{1:k}^T \;\; \bar{r}_{k+1}]^T) = \inf_{\zeta_k} \sup_{\bar{r}_{k+1}\in\mathcal{U}_{k+1}(r_{1:k})} -\zeta_k\bar{r}_{k+1} + 0 \\ &= \inf_{\zeta_k} |\zeta_k| \sup_{\bar{r}_{k+1}\in\mathcal{U}_{k+1}(r_{1:k})} \bar{r}_{k+1} = 0\,.\end{aligned}$$

This confirms that the worst-case risk measure defined based on the set $\mathcal{U}_2$ satisfies bounded conditional market risk property.

1.6.13 Worst-case Risk Measures with $\mathcal{U}_1$ or $\mathcal{U}_2$ Satisfying the Markov Property

In this section we identify two state processes $\theta_k : \mathbb{R}^K \to \mathbb{R}$ under which the worst-case risk measures with $\mathcal{U}_1$ and $\mathcal{U}_2$ are respectively Markovian.

Starting with the set inspired from Bandi and Bertsimas (2014), we let $\theta_k := \sum_{\ell=1}^{k} \log(1+ r_\ell)$. With this definition in hand, we can demonstrate the properties that are described in Definition 5. First, we have that $\theta_{k+1} = \sum_{\ell=1}^{k+1} \log(1 + r_\ell) = \log(1 + r_{k+1}) + \theta_k$ and hence can be measured directly from $(\theta_k, r_k + 1)$. Second, we can confirm that for all $X \in \mathcal{L}_p(\Omega, \mathcal{F}_{k+1}, \mathbb{P})$, if $r_{1:k} \in \mathcal{U}_{1:k}$, then:

$$\begin{aligned}
\rho_k(X, r) &= \sup_{r' \in \mathcal{U} : r'_{1:k} = r_{1:k}} X(r') \\
&= \sup_{\bar{r}_{k+1} \in \mathcal{U}_k^\theta(\sum_{\ell=1}^{k} \log(1+r_\ell))} X([r_{1:k}^T \;\; \bar{r}_{k+1} \;\; r_{k+2:K}]^T) \\
&= \sup_{\bar{r}_{k+1} \in \mathcal{U}_k^\theta(\theta_k(r))} \Pi_k(X, \bar{r}_{k+1}) = \bar{\rho}_k(\Pi_k(X, r), \theta_k(r)) \,,
\end{aligned}$$

where

$$\mathcal{U}_k^\theta(\theta_k) := \left\{ r \in \mathbb{R} \;\middle|\; \left| \frac{\theta_k + \log(1 + r) - \mu k' T/K}{\sigma \sqrt{k' T/K}} \right| \leq \Gamma, \quad \forall k' \geq k \right\}$$

and

$$\bar{\rho}_k(X, \theta_k) := \begin{cases} \sup_{r_{k+1} \in \mathcal{U}_k^\theta(\theta_k)} X(r_{k+1}) & \text{if } \mathcal{U}_k^\theta(\theta_k) \neq \emptyset \\ X(0) & \text{otherwise} \end{cases} .$$

In the case of the set $\mathcal{U}_2$ inspired from Bernhard (2003), we let instead $\theta_k := \sum_{\ell=1}^{k} r_\ell^2$, with $\theta_{k+1} := \theta_k + r_{k+1}^2$ and $\bar{\rho}_k(X, \theta_k) := \sup_{r_{k+1} \in \mathcal{U}_k^\theta(\theta_k)} X(r_{k+1})$ where

$$\mathcal{U}_k^\theta(\theta_k) := \left\{ r \in \mathbb{R} \;\middle|\; \theta_k + r^2 \in [\sigma i N T/K - \Gamma \sqrt{iN}, \sigma i N T/K + \Gamma \sqrt{iN}] \,, \forall\, i \geq k/N \right\} .$$

The rest of the details are very similar as previously.

1.6.14 Implementation Details Regarding How the Dynamic Program Was Solved

In order to solve the dynamic programs of the models, in the first step, we divide our simulated stock paths into a training, D_{train}, and a test set, D_{test}. Then we calibrate the uncertainty set parameter Γ in a way that 95 percent of the train paths fall into the uncertainty set. Depending on the type of the uncertainty set, $\mathcal{U}_1$ or $\mathcal{U}_2$, we consider either cumulative log returns $\sum_{l=1}^{k} \log(1+r_l)$, or cumulative square returns $\sum_{l=1}^{k} r_l^2$ along with the stock price S_k as state variables of the DP. Next, we generate a two-dimensional grid for the state variables at each time step. The upper and lower bounds of the grid for the stock prices are obtained by simply considering the price bounds in D_{train}. The bounds could be computed in a more conservative way by considering some deviations from the minimum and maximum values of D_{train} so that they contain with a higher probability the paths of the D_{test}, however, we did not observe improvements in terms of the hedging performance by considering such bounds in our setting. Another issue worth mentioning is regarding the time dependency of the grid bounds. In our implementation we consider the same bounds for all time periods. This would allow the model to consider the cases where the whole budget of the uncertainty sets are used up in the very first steps. However, in general, the bounds of the grid could be time dependent and determined at each time by using the price paths in that time. For the other state variable, we first use the D_{train} to compute the associated paths of $\sum_{l=1}^{k} \log(1+r_l)$ or $\sum_{l=1}^{k} r_l^2$. The upper and lower bounds of the grid could be computed from these values.

In order to solve the dynamic model, starting from the last period, for each combination of state variables, and for each side of the contract, the "best hedging risk to go" is computed and assigned to the point. For periods other than the last period, we need to solve an optimization model to obtain the optimal allocation of wealth. Since we discretized the state space, at each point we find the reachable values in the next period for the two state variables (θ_k, S_k). To do so, we start with the reachable values for r_{k+1} using the definition of the projected uncertainty set. Specifically, for $\mathcal{U}_1$ we will have $r_{k+1} \in [\underline{r}_{k+1}, \overline{r}_{k+1}]$, while for $\mathcal{U}_2$ we will have $r_{k+1} \in [\underline{r}^1_{k+1}, \overline{r}^1_{k+1}] \cup [\underline{r}^2_{k+1}, \overline{r}^2_{k+1}]$. When choosing the reachable grid points, we always include up to the first grid point that falls outside the intervals in order

to induce a conservative bias to our approximation. Having these points, the next step is to find the optimal wealth allocation by solving a piece-wise linear convex optimization problem. We pursue recursively until $k = 0$.

References

Amin, K. I. (1993). Jump diffusion option valuation in discrete time. *The Journal of Finance*, 48(5):1833–1863.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Bandi, C. and Bertsimas, D. (2014). Robust option pricing. *European Journal of Operational Research*, 239(3):842–853.

Bernhard, P. (2003). A robust control approach to option pricing. *Applications of Robust Decision Theory and Ambiguity in Finance. City University Press, London.*

Bernhard, P., Engwerda, J., Roorda, B., M. Schumacher, J., Kolokoltsov, V., Saint-Pierre, P., and Aubin, J.-P. (2013). *The Interval Market Model in Mathematical Finance*, pages 293–317.

Bertsekas, D. P. (2015). *Convex optimization algorithms.* Athena Scientific Belmont.

Bertsimas, D., Kogan, L., and Lo, A. W. (2001). Hedging derivative securities and incomplete markets: an ϵ-arbitrage approach. *Operations Research*, 49(3):372–397.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Brennan, M. J. (1979). The pricing of contingent claims in discrete time models. *The Journal of Finance*, 34(1):53–68.

Carbonneau, A. and Godin, F. (2020). Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, pages 1–16.

Carr, P., Geman, H., and Madan, D. B. (2001). Pricing and hedging in incomplete markets. *Journal of Financial Economics*, 62(1):131 – 167.

Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263.

Delbaen, F. and Schachermayer, W. (1995). The variance-optimal martingale measure for continuous processes. *Bernoulli*, 2:81–105.

Föllmer, H. and Schied, A. (2011). *Stochastic finance: an introduction in discrete time*. Walter de Gruyter.

Föllmer, H., Sondermann, H., and Sondermann, D. (1985). *Hedging of non-redundant contingent claims*, pages 205 – 223.

François, P., Gauthier, G., and Godin, F. (2014). Optimal hedging when the underlying asset follows a regime-switching markov process. *European Journal of Operational Research*, 237(1):312–322.

Gourieroux, C., Laurent, J. P., and Pham, H. (1998). Mean-variance hedging and numéraire. *Mathematical Finance*, 8(3):179–200.

Guo, I. and Zhu, S.-P. (2017). Equal risk pricing under convex trading constraints. *Journal of Economic Dynamics and Control*, 76:136–151.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300.

Jaschke, S. and Küchler, U. (2001). Coherent risk measures and good-deal bounds. *Finance and Stochastics*, 5(2):181–200.

King, A. J. (2002). Duality and martingales: a stochastic programming perspective on contingent claims. *Mathematical Programming*, 91(3):543–562.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, pages 141–183.

Pichler, A. and Shapiro, A. (2018). Risk averse stochastic programming: time consistency and optimal stopping. *arXiv preprint arXiv:1808.10807*.

Ruszczyński, A. (2010). Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261.

Ruszczynski, A. and Shapiro, A. (2006). Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561.

Schweizer, M. (1996). Approximation pricing and the variance-optimal martingale measure. *The Annals of Probability*, 24(1):206–236.

Schweizer, M. (1999). A guided tour through quadratic hedging approaches. Technical report, SFB 373 Discussion Paper.

Shapiro, A. (2012). Time consistency of dynamic risk measures. *Operations Research Letters*, 40(6):436 – 439.

Staum, J. (2007). *Chapter 12 Incomplete Markets*, volume 15 of *Handbooks in Operations Research and Management Science*, pages 511 – 563. Elsevier.

Xu, M. (2006). Risk measure pricing and hedging in incomplete markets. *Annals of Finance*, 2(1):51–71.

Chapter 2

Equal Risk Pricing and Hedging Using Deep Reinforcement-Learning under Dynamic Expectile Risk Measures

Chapter information

This article is a joint work with my supervisors, Erick Delage, and Jonathan Yu-Meng Li. It is submitted to the International Conference on Learning Representations (ICLR).

Abstract

Recently equal risk pricing, a framework for fair derivative pricing, was extended to consider dynamic risk measures. However, all current implementations either employ a static risk measure that violates time-consistency, or are based on traditional dynamic programming solution schemes that are impracticable in problems with a large number of underlying assets (due to the curse of dimensionality) or with incomplete asset dynamics information. In this article, we extend for the first time a famous off-policy deterministic actor-critic deep reinforcement-learning (ACRL) algorithm to the problem of solving a risk-averse Markov decision process that models risk using a time-consistent recursive expectile risk

measure. This new ACRL algorithm allows us to identify time-consistent hedging policies (and equal risk prices) for options, such as basket options, that cannot be handled using traditional methods, or in context where only historical trajectories of the underlying assets are available. Our numerical experiments, which involve both a simple vanilla option and a more exotic basket option, confirm that the new ACRL algorithm can produce 1) in simple environments, nearly optimal hedging policies, and highly accurate prices, simultaneously for a range of maturities 2) in complex environments, good quality policies and prices using reasonable amount of computing resources; and 3) overall, hedging strategies that actually outperform the strategies produced using static risk measures when the risk is evaluated at later points of time.

2.1 Introduction

Derivative pricing remains a challenging problem in finance when the markets are incomplete and the derivatives are dependent on multiple underlying assets. The incompleteness of a market implies that the price of some derivatives cannot be uniquely determined by the standard replication argument, as in such a market no self-financing hedging strategy exists that can perfectly replicate the payoffs of some derivatives. Many mechanisms have been proposed for pricing in an incomplete market but most were developed from the perspective of a single trader. Unfortunately, a price that is set only according to one party's interest, e.g. a super-replication price that a seller may wish to charge, may not be acceptable to the buyer and thus does not represent a plausible transaction price. Recently, a new pricing scheme, known as Equal Risk Pricing (ERP), was proposed by Guo and Zhu (2017) and further adapted to convex risk measures in the work of Marzban et al. (2020).

The scheme of ERP is built upon the idea of modelling separately the risk exposure of the buyer and the seller of a derivative, and seeking a price that ensures that the risk exposure of both parties is the same under their respective optimal self-financing hedging strategy. The price generated from ERP thus has the merit of fairness to both parties. While ERP has its conceptual appeal, there remains a gap between its general construct and the actual implementation. In particular, as shown in Marzban et al. (2020), great care must be taken to define properly how risk should be measured in a dynamic hedging setting

in order to obtain hedging problems that are operationally meaningful and computationally solvable. The work of Marzban et al. (2020) provides necessary analysis for solving the equal risk pricing and hedging problem based on traditional dynamic programming (DP). It is known however that DP suffers from the issue of the curse of dimensionality, which restricts the applicability of the results in Marzban et al. (2020). In addition, traditional DP assumes the knowledge of a stochastic model that precisely captures the dynamics of the markets, which may not be available in practice.

In the past decade, Deep Reinforcement-Learning (DRL) has proven to be a powerful tool for solving dynamic optimization problems when the number of state variables is large and/or when no stochastic model is known for the underlying system dynamics. In particular, the recent works of Carbonneau and Godin (2020) and Carbonneau and Godin (2021) are the first that apply DRL to solve ERP problems and they demonstrate the possibility of pricing a broad range of over-the-counter options such as basket options. Unfortunately, the DRL approaches proposed in Carbonneau and Godin (2020) and Carbonneau and Godin (2021) can only be used in settings where the risk is measured according to a static risk measure. This raises the serious issue that the hedging problem exploited by the ERP could be time inconsistent, i.e. the hedging decisions planned for future state of the world may no longer be considered optimal once the state is visited. The violation of time-consistency implies that equal risk prices calculated based on static risk measures will assume a hedging policy that cannot be implemented in practice, and thus are optimistically biased. From a numerical perspective, employing a static risk measure in ERP also limits the type of DRL algorithms that can be used to solve each party's hedging problem. Specifically, Carbonneau and Godin (2020) and Carbonneau and Godin (2021) employ a policy optimization scheme, a.k.a. Actor-Only RL (AORL) algorithm (see Williams (1992) as an example of this method), while other approaches such as critic-only or actor-critic algorithms (such as Mnih et al. (2015) and Lillicrap et al. (2015) respectively) that rely on an equivalent DP formulation remain out of reach.

In this article, we seek to develop a DRL approach for solving a class of time-consistent ERP problems. It is known that, to ensure time-consistency, a dynamic risk measure should be employed to measure risk in a recursive fashion. In particular, motivated by the theory of coherent risk measures, which identifies expectile risk measures as the only elicitable

coherent risk measures, we propose in this article the use of dynamic expectile risk measures to formulate time-consistent ERP problems. The dynamic nature of risk measures suggests the consideration of an Actor-Critic RL (ACRL) algorithm for solving the hedging problem. It turns out that the elicitability property of expectile risk measures facilitates greatly the design of a model-free ACRL algorithm. The convergence of this algorithm is also greatly improved due to the translation invariance property of the risk measures.

Overall, we may summarize the contribution of this paper as follows:

- We present the first model-free DRL based algorithm for computing equal risk prices that rely on option hedging strategies that are time-consistent. To reinforce the importance of this contribution, we in fact demonstrate using a simple single asset two-period horizon option pricing problem how equal risk prices might suffer from an optimistic bias when static risk measures are used (as in Carbonneau and Godin (2020) and Carbonneau and Godin (2021)). A side benefit from pricing an option with maturity T using dynamic risk measures will be that we will easily obtain equal risk prices for any other maturity $T' < T$.

- The ACRL algorithm that we propose is the first model-free DRL algorithm to naturally extend the famous off-policy deterministic actor-critic method presented in Silver et al. (2014) to the risk-averse setting. Unlike the ACRL proposed in Tamar et al. (2015) and Huang et al. (2021) for risk-averse DRL, which can employ up to five neural networks, our algorithm will only require two deep neural networks: a policy network (actor) and a Q network (critic). While our policy network will be trained following a stochastic gradient procedure similar to Silver et al. (2014), to the best of our knowledge we are the first to leverage the elicitability property (i.e. existence of a scoring function) of expectile risk measures and to propose a procedure for training the "risk-to-go" Q network that is also based on stochastic gradient descent.

- We perform a comprehensive evaluation of the training efficiency, quality of option hedging strategies, and quality of equal risk prices obtained using our ACRL algorithm on a synthetic multi-asset geometric Brownian motion market model. In the simple case of vanilla option pricing, we provide empirical evidence that ACRL provides

nearly optimal hedging policies, and highly accurate prices, simultaneously for a range of maturities. The latter is in sharp contrast with approaches, such as in Carbonneau and Godin (2021), that employ time inconsistent risk measures and produce investment strategies that are visibly outperformed by the ACRL strategy in terms of the risk measured as time to maturity reduces. This phenomenon is also observed, although less prominently, in the context of a basket option over 5 underlying assets, where good quality policies and prices are obtained using our ACRL algorithm using a reasonable amount of computing resources.

The rest of this article is organized as follows. Section 2.2 introduces equal risk pricing and illustrates using a simple two-period pricing problem the practical issues related to using static risk measures for option hedging and pricing. Section 2.3 adapts the ERP framework to the case of a dynamic expectile risk measure and proposes the new ACRL algorithm. Finally, Section 2.4 presents and discusses our numerical experiments.

2.2 Equal risk pricing and hedging under coherent risk measures

In this section, we provide a brief overview of ERP under coherent risk measures based on the recent work of Marzban et al. (2020). We pay particular attention to the issue of time (in)consistency by presenting an example that demonstrates numerically that employing a time-inconsistent static risk measure can lead to an under-evaluation of the risk to which each party are actually exposed in practice.

2.2.1 ERP under coherent risk measures

The problem of ERP can be formalized as follows. Consider a frictionless market, i.e. no transaction cost, tax, etc, that contains m risky assets, and a risk-free bank account with zero interest rate. Let $\mathbf{S}_t : \Omega \to \mathbb{R}^m$ denote the values of the risky assets adapted to a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t\}_{t=0}^T, \mathbb{P})$, i.e. each $\mathbf{S}_t$ is $\mathcal{F}_t$ measurable. It is assumed that $\mathbf{S}_t$ is a locally bounded real-valued semi-martingale process and that the set of equivalent local martingale measures is non-empty (i.e. no arbitrage opportunity). The

set of all admissible self-financing hedging strategies with the initial capital $p_0 \in \mathbb{R}$ is shown by $\mathcal{X}(p_0)$:

$$\mathcal{X}(p_0) = \left\{ X : \Omega \to \mathbb{R}^T \,\middle|\, \exists \{\boldsymbol{\xi}_t\}_{t=0}^{T-1}, \quad X_t = p_0 + \sum_{t'=0}^{t-1} \boldsymbol{\xi}_{t'}^\top \Delta \mathbf{S}_{t'+1}, \quad \forall t = 1, \ldots, T \right\},$$

where $\Delta \mathbf{S}_{t+1} := \mathbf{S}_{t+1} - \mathbf{S}_t$, the hedging strategy $\boldsymbol{\xi}_t \in \mathbb{R}^m$ is a vector of random variables adapted to the filtration $\mathbb{F}$ and captures the number of shares of each of the risky assets held in the portfolio during the period $[t,\ t+1]$, $\boldsymbol{\xi}_t^\top \Delta \mathbf{S}_{t'+1}$ is the inner product of the two random vectors, and X_t is the accumulated wealth.

Let $F(\{\mathbf{S}_t\}_{t=1}^T)$ denote the payoff of a derivative. Throughout this paper, we assume $F(\{\mathbf{S}_t\}_{t=1}^T)$ admits the formulation of $F(\mathbf{S}_T, \mathbf{Y}_T)$ where $\mathbf{Y}_t$ is an auxiliary fixed-dimensional stochastic process that is $\mathcal{F}_t$-measurable. This class of payoff functions is common in the literature, (see for example Bertsimas et al. (2001) and Marzban et al. (2020)). The problem of ERP is defined based on the following two hedging problems that seek to minimize the risk of hedging strategies, one is for the writer and the other is for the buyer of the derivative:

$$\text{(Writer)} \qquad \varrho^w(p_0) = \inf_{X \in \mathcal{X}(p_0)} \rho^w(F(\mathbf{S}_T, \mathbf{Y}_T) - X_T) \tag{2.1}$$

$$\text{(Buyer)} \qquad \varrho^b(p_0) = \inf_{X \in \mathcal{X}(-p_0)} \rho^b(-F(\mathbf{S}_T, \mathbf{Y}_T) - X_T)\,, \tag{2.2}$$

where ρ^w and ρ^b are two risk measures that capture respectively the writer and the buyer's risk aversion. In words, equation (2.1) describes a writer that is receiving p_0 as the initial payment and implements an optimal hedging strategy for the liability captured by $F(\mathbf{S}_T, \mathbf{Y}_T)$. On the other hand, in (2.2) the buyer is assumed to borrow p_0 in order to pay for the option and then to manage a portfolio that will minimize the risks associated to his final wealth $F(\mathbf{S}_T, \mathbf{Y}_T) + X_T$. With equations (2.1) and (2.2), ERP defines a fair price p_0^* as the value of an initial capital that leads to the same risk exposure to both parties, i.e.

$$\rho^w(p_0^*) = \rho^b(p_0^*).$$

Motivated by the theory of coherent risk measures (Artzner et al. (1999)), Marzban et al. (2020) study the ERP problem by imposing the property of coherency to the risk measures ρ^w and ρ^b. Namely, a risk measure is said to be coherent if it satisfies the following five conditions:

- Monotonicity: if $X \leq Z \quad a.s.$ then $\rho(X) \leq \rho(Z)$
- Subadditivity: $\rho(X+Z) \leq \rho(X) + \rho(Z)$
- Positive homogeneity: If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda \rho(X)$
- Translation invariance: If $m \in \mathbb{R}$, then $\rho(X+m) = \rho(X) + m$
- Normalized risk: $\rho(0) = 0$.

It is well known that Value-at-Risk (VaR), a risk measure commonly applied in financial risk management, is not coherent, whereas its convex counterpart, namely Conditional Value-at-Risk (CVaR) is coherent. The application of CVaR in ERP can be found for example in Carbonneau and Godin (2020). As one of the key results in ERP, Marzban et al. (2020) establishes that an equal risk price p_0^* can actually be found by solving the writer and buyer's hedging problem with no initial payment, i.e. (2.1) and (2.2), separately. Namely, it can be calculated by the following result.

Theorem 2.2.1. *Let ρ^w and ρ^b be two coherent risk measures. In the case where the equal risk price p_0^* exists, it can be calculated by*

$$p_0^* = (\varrho^w(0) - \varrho^b(0))/2\,,$$

when $\infty > \varrho^w(0) \geq \varrho^b(0) > -\infty$.

2.2.2 The issue of time inconsistency

As briefly mentioned in the introduction, measuring risk in a dynamic setting requires additional care. The use of a coherent risk measure, without any further adaptation to a dynamic setting, can lead to solutions that suffer from the issue of time inconsistency. The goal of this section is to carefully demonstrate this point by presenting a numerical example that quantifies the impact of time inconsistency. Our demonstration is inspired by the work of Rudloff et al. (2014), where the impact of time inconsistency is discussed in a portfolio management problem. Here, we present an example based on a vanilla option hedging problem.

In this example, we consider a stock price process modelled by a simple two-stage trinomial tree. Specifically, the horizon spans $t \in \{0, 1, 2\}$ and the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is such that $\Omega = \{\omega_i\}_{i=1}^{9}$, $\mathcal{F}_1 := \sigma(\{\{\omega_i\}_{i=1}^{3}, \{\omega_i\}_{i=4}^{6}, \{\omega_i\}_{i=7}^{9}\})$, and all outcomes are equiprobable. The market contains a risk-free asset (with a risk-free rate of zero) and a risky asset S which are used to hedge a vanilla at-the-money call option on S_2 with strike price $K := S_0$. The details of the price process is shown in Table 2.1. For simplicity, we set the initial capital for hedging to zero and employ a $\text{CVaR}_{60\%}$ risk measure for hedging [1].

When hedging the call-option using a static CVaR measure, the writer of the option solves the following two-period optimization model:

$$\min_{\xi_0, \xi_1} \text{CVaR}_{60\%}((S_2(\omega) - K)^+ - (S_1(\omega) - S_0)\xi_0 - (S_2(\omega) - S_1(\omega))\xi_1(\omega)) \tag{2.3}$$

where $(y)^+ := \max(y, 0)$ and $K := S_0$. The optimal solution of this problem will prescribe purchasing 0.93 shares of the risky asset at time 0, i.e. $\xi_0 = 0.9341$, using money borrowed at the risk-free rate (see Table 2.1 for the optimal shares to hold at $t = 1$). The resulting $\text{CVaR}_{60\%}$ is 26.36, implying that if the writer charges the buyer with a price above 26.36, the writer would consider the price being sufficient to cover the hedged risk of this call option.

Note that in the risk-averse hedging problem (2.3), it is not clear what motivates the writer of the option to implement the prescribed hedging strategy once new information is revealed at time $t = 1$. In particular, he/she might be curious to compare the prescribed strategy with the strategy that minimizes the CVaR from the new perspective at $t = 1$, i.e., the following hedging problem:

$$\min_{\bar{\xi}_1} \text{CVaR}_{\bar{\alpha}_1}\left((S_2(\omega) - K)^+ - (S_1(\omega) - S_0)\xi_0^* - (S_2(\omega) - S_1(\omega))\bar{\xi}_1(\omega)|\mathcal{F}_1\right), \tag{2.4}$$

where $\bar{\alpha}_1 := 60\%$ and where $\xi_0^* = 0.9341$, i.e. the optimal first stage solution in (2.3).

Table 2.1 presents the optimal conditional hedging strategy $\bar{\xi}_1^*$ as a function of the information revealed by $\mathcal{F}_1$. While it does appear that $\bar{\xi}_1^*$ agrees with ξ_1^* when $\omega \in \{\omega_i\}_{i=1}^{3}$, the investment in the risky asset ends up significantly reduced in the other two sets of outcomes. More importantly, we established numerically[2] that in order to motivate the

[1] The issue could arise for any risk level when the outcome space is large enough. Here 60% CVaR was used to produce a simple example.

[2] `https://github.com/saeedmarzban/ERP-Dynamic-Expectile-RM.git`

prescribed hedging strategy ξ_1^*, the risk aversion level used in problem (2.4) would need to be in the range of $[0.4580,\ 0.4585]$, when $\omega \in \{\omega_i\}_{i=4}^{6}$, or $[0.1992,\ 0.2]$, when $\omega \in \{\omega_i\}_{i=7}^{9}$. This confirms that ξ^* is likely to be perceived as overly risky given the information revealed at time $t = 1$. Ultimately, in the likely case where the writer decides to replace ξ_1^* with $\bar{\xi}_1^*$, one can establish that the overall exposition to risk from the perspective of $t = 0$ should have rather been estimated to 27.94 instead of 26.36. This implies that employing a static risk measure here underestimated the necessary coverage capital by 6%.

While this issue of time-consistency has been discussed significantly in the recent years, a common approach to overcome it is to employ a so-called dynamic risk measure (Detlefsen and Scandolo, 2005) as will be done in the following section. In the context of this example, this would reduce to replacing problem (2.3) with:

$$\min_{\xi_0,\xi_1} \mathrm{CVaR}_\alpha\Big(\mathrm{CVaR}_\alpha((S_2(\omega)-K)^+ - (S_2(\omega)-S_1(\omega))\xi_1(\omega) - (S_1(\omega)-S_0)\xi_0|\mathcal{F}_1(\omega))\Big),$$

where α can be chosen to characterize the right level of risk aversion for the "dynamic conditional value-at-risk measure". This formulation ensures that the prescribed policy at time $t = 1$ remains optimal (according to problem (2.4)) at the moment where it is actually implemented thus preventing the necessary coverage capital from being under estimated.

Table 2.1 – Example of a time inconsistent hedging strategy obtained from employing a static risk measure. ξ^* is obtained by solving problem (2.3), $\bar{\alpha}_1$ is the risk aversion level that motivates ξ_1^* at $t = 1$, $\bar{\xi}_1^*$ is the actual investment prescribed by $\mathrm{CVaR}_{60\%}$ at $t = 1$.

Atoms of $\mathcal{F}_1$	Price process			Time inconsistent hedging strategy			Optimal conditional hedging strategy
	$S_0(\omega)$	$S_1(\omega)$	$S_2(\omega)$	ξ_0^*	$\xi_1^*(\omega)$	$\bar{\alpha}_1(\xi^*)$	$\bar{\xi}_1^*(\omega)$
$\omega \in \{\omega_i\}_{i=1}^{3}$	100	150	{270,150,75}	0.9341	0.8718	[0.4580,1.0000]	0.8718
$\omega \in \{\omega_i\}_{i=4}^{6}$	100	100	{180,100,50}	0.9341	0.7665	[0.4580,0.4585]	0.6154
$\omega \in \{\omega_i\}_{i=7}^{9}$	100	80	{120,80,64}	0.9341	0.5000	[0.1992,0.2000]	0.3571

2.3 ERP under dynamic expectile risk measure and an actor-critic algorithm

While time-consistent ERP problems can be formulated by employing dynamic risk measures and be calculated, in principle, by solving a set of dynamic programming (DP) equations

(Marzban et al. (2020)), there remains the challenge of determining which dynamic risk measure one should employ and how these equations might be solved in high dimension, i.e. multiple underlying assets. In this section, we address the two issues by first motivating the use of dynamic expectile risk measures to formulate time-consistent ERP hedging problems and then presenting a Deep Reinforcement-Learning approach (DRL) for approximately solving this problems.

2.3.1 Dynamic expectile risk measures and DP equations

Expectile has been proposed in the recent literature (see Bellini and Bignozzi (2015)) as a replacement of VaR and CVaR given that it is not only coherent but also elicitable. It is known that VaR is not coherent but is elicitable, whereas CVaR is coherent but is not elicitable. A risk measure is said to be elicitable if it can be expressed as the minimizer of a certain scoring function, and this property is found to be critical in practice due to the need of backtesting (Chen, 2018). In fact, the expectile is the only elicitable coherent risk measure. Recall the following definition of expectile.

Definition 12. (Bellini and Bernardino (2017)) The $\tau-$expectile of a random liability X is defined as:

$$\bar{\rho}(X) := \arg\min_{q} \tau\mathbb{E}\left[(q-X)_+^2\right] + (1-\tau)\mathbb{E}\left[(q-X)_-^2\right] .$$

Like CVaR, expectile covers at one extreme the case of risk-neutrality, i.e. with $\tau = 1/2$, and at the other extreme the case of converging towards the worst-case risk, i.e. as $\tau \to 1$. Thus, expectile also allows for modelling a wide spectrum of risk aversion. Using expectile as the basis, we define its dynamic version as follows.

Definition 13. A dynamic recursive expectile risk measure takes the form:

$$\rho(X) := \bar{\rho}_0(\bar{\rho}_1(\dots\bar{\rho}_{T-1}(X)))\,,$$

where each $\bar{\rho}(\cdot)$ is an expectile risk measure that employs the conditional distribution based on $\mathcal{F}_t$. Namely,

$$\bar{\rho}_t(X_{t+1}) := \arg\min_{q} \tau\mathbb{E}\left[(q-X_{t+1})_+^2|\mathcal{F}_t\right] + (1-\tau)\mathbb{E}\left[(q-X_{t+1})_-^2|\mathcal{F}_t\right]$$

where X_{t+1} a random liability measureable on $\mathcal{F}_{t+1}$.

We apply dynamic expectile risk measures to formulate the two hedging problems in ERP. By further imposing the following assumption that there exists a sufficient statistic process ψ_t such that $\{(\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\}_{t=0}^T$ satisfies the Markov property, we can obtain compact dynamic equations for them.

Assumption 2.3.1. *[Markov property] There exists a sufficient statistic process ψ_t adapted to $\mathbb{F}$ such that $\{(\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\}_{t=0}^T$ is a Markov process relative to the filtration $\mathbb{F}$. Namely, $\mathbb{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s}) \in \mathcal{A}|\mathcal{F}_t) = \mathbb{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s}) \in \mathcal{A}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ for all t, for all $s \geq 0$, and all sets $\mathcal{A}$.*

In particular, based on Proposition 3.1 and the examples presented in section 3.3 of Marzban et al. (2020), together with the fact that both ρ^w and ρ^b are dynamic recursive expectile risk measures, the Markovian assumption allows us to conclude that the ERP can be calculated using two sets of dynamic programming equations. Specifically, on the writer side, we can define

$$V_T^w(\mathbf{S}_T, \mathbf{Y}_T, \psi_T) := F(\mathbf{S}_T, \mathbf{Y}_T)\,,$$

and recursively

$$V_t^w(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_{\boldsymbol{\xi}_t} \bar{\rho}(-\boldsymbol{\xi}_t^\top \Delta\mathbf{S}_{t+1} + V_{t+1}^w(S_t + \Delta\mathbf{S}_{t+1}, \mathbf{Y}_t + \Delta Y_{t+1}, \psi_{t+1})|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\,,$$

where $\bar{\rho}(\cdot|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ is the expectile risk measure that uses $\mathbb{P}(\cdot|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. This leads to considering $\varrho^w(0) = V_0^w(\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$. On the other hand, for the buyer we similarly define:

$$V_T^b(\mathbf{S}_T, \mathbf{Y}_T, \psi_T) := -F(\mathbf{S}_T, \mathbf{Y}_T)\,,$$

and

$$V_t^b(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_{\boldsymbol{\xi}_t} \bar{\rho}(-\boldsymbol{\xi}_t^\top \Delta\mathbf{S}_{t+1} + V_{t+1}^b(\mathbf{S}_t + \Delta\mathbf{S}_{t+1}, \mathbf{Y}_t + \Delta\mathbf{Y}_{t+1}, \psi_{t+1})|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\,,$$

with $\varrho^b(0) = V_0^b(\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$. The following lemma summarizes how DP can be used to compute ERP.

Lemma 2.3.1. *Under Assumption 2.3.1, the ERP that employs dynamic recursive expectile riks measure can be computed as: $p_0^* = (V_0^w(\mathbf{S}_0, \mathbf{Y}_0, \psi_0) - V_0^b(\mathbf{S}_0, \mathbf{Y}_0, \psi_0))/2$.*

2.3.2 A novel Expectile-based actor-critic algorithm for ERP

In this section, we formulate each option hedging problem as a Markov Decision Process (MDP) denoted by $(\mathcal{S}, \mathcal{A}, r, P)$. In this regard, the agent (i.e. the writer or buyer) interacts with a stochastic environment by taking an action $a_t \equiv \boldsymbol{\xi}_t \in [-1, 1]^m$ after observing the state $s_t \in \mathcal{S}$, which includes $\mathbf{S}_t$, $\mathbf{Y}_t$, and ψ_t. Note that to simplify exposition, in this section we drop the reference to the specific identity (i.e. w or b) of the agent in our notation. The action taken at each time t results in the immediate stochastic reward that takes the shape of the immediate hedging portfolio return, i.e. $r_t(s_t, a_t, s_{t+1}) := \boldsymbol{\xi}_t^\top \Delta \mathbf{S}_{t+1}$ when $t < T$ and otherwise of the option liability/payout $r_T(s_T, a_T, s_{T+1}) := F(\mathbf{S}_T, \mathbf{Y}_T)(1 - 2 \cdot \mathbf{1}\{\text{agent=writer}\})$, which is insensitive to s_{T+1}. Finally, the Markovian exogeneous dynamics described in Assumption 2.3.1 are modeled using P as $P(s_{t+1}|s_t, a_t) = \mathbb{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. Overall, each of the two dynamic derivative hedging problems presented in Section 2.3.1 reduce to a version of the following risk-averse reinforcement-learning problem:

$$\varrho(0) = V_0(\mathbf{S}_0, \mathbf{Y}_0, \psi_0) = \min_\pi Q_0^\pi(\bar{s}_0, \pi_0(\bar{s}_0)) \,,$$

where $\bar{s}_0 := (\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$ is the initial state in which the option is priced while

$$Q_t^\pi(s_t, a_t) := \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1}))\,|s_t) \,,$$

and $Q_T^\pi(s_T, a_T) := -r_T(s_T, a_T, s_T)$.

Lemma 2.3.2. *Let $\bar{\pi}$ be an arbitrary reference policy that puts strictly positive probability on all of $\mathcal{A}$ for each state, and has a strictly positive probability of reaching all of $\mathcal{S}$ for all $t \geq 1$ when starting from $\bar{s}_0$.*[3] *For any π^* that satisfies*

$$\pi^* \in \arg\min_\pi \mathbb{E}_{\substack{\tilde{t}\sim\{0,\ldots,T\} \\ s_{t+1}\sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} [Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \pi_{\tilde{t}}(s_{\tilde{t}}))] \tag{2.5}$$

where $s_0 := \bar{s}_0$ and $\tilde{t}$ is uniformly drawn, we necessarily have that π^* minimizes $Q_0^\pi(\bar{s}_0, \pi_0(\bar{s}_0))$.

Proof. We start by proving first that given any π^* that satisfies (2.5), it must also satisfy

$$\pi^* \in \arg\min_\pi \mathbb{E}_{(s,t)\sim\beta}[Q_t^{\pi^*}(s, \pi_t(s))] \,, \tag{2.6}$$

[3] In our option hedging problem, given that s_t is entirely exogenous, the distribution of s_{t+1} is unaffected by $\bar{\pi}$, which can therefore be chosen arbitrarily.

where β captures the distribution of $(\tilde{t}, s_{\tilde{t}})$ used in (2.5). We do so by contradiction. Let's assume that there exists a $\bar{\pi}$ such that

$$\mathbb{E}_{(s,t)\sim\beta}[Q_t^{\pi^*}(s, \bar{\pi}_t(s))] < \mathbb{E}_{(s,t)\sim\beta}[Q_t^{\pi^*}(s, \pi_t^*(s))]\,.$$

Then, one can design the following policy:

$$\bar{\pi}_t^*(s) := \begin{cases} \bar{\pi}_t(s) & \text{if } Q_t^{\pi^*}(s, \bar{\pi}_t(s)) < Q_t^{\pi^*}(s, \pi_t^*(s)) \\ \pi_t^*(s) & \text{otherwise.} \end{cases}$$

Using a recursive argument, one can show that $Q_t^{\bar{\pi}^*}(s_t, a_t) \leq Q_t^{\pi^*}(s_t, a_t)$ for all t and (s_t, a_t) pair. In this recursion, we start with:

$$Q_T^{\bar{\pi}^*}(s_T, a_T) = -r_T(s_T, a_T, s_T) = Q_T^{\pi^*}(s_T, a_T)\,.$$

Moreover, for all t, and (s_t, a_t) pairs, we have that:

$$\begin{aligned} Q_t^{\bar{\pi}^*}(s_t, a_t) &= \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\bar{\pi}^*}(s_{t+1}, \bar{\pi}^*(s_{t+1}))|s_t) \\ &\leq \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \bar{\pi}^*(s_{t+1}))|s_t) \\ &\leq \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \pi^*(s_{t+1}))|s_t) = Q_t^{\pi^*}(s_t, a_t), \end{aligned}$$

where we first used $Q_{t+1}^{\bar{\pi}^*}(s_t, a_t) \leq Q_{t+1}^{\pi^*}(s_t, a_t)$, then exploited the definition of $\bar{\pi}_t^*$. With this result in hand we can obtain that for all t and s_t

$$Q_t^{\bar{\pi}^*}(s_t, \bar{\pi}_t^*(s_t)) \leq Q_t^{\pi^*}(s_t, \bar{\pi}_t^*(s_t)) \leq Q_t^{\pi^*}(s_t, \bar{\pi}_t(s_t)),$$

where we again used the definition of $\bar{\pi}^*$. Finally, we must therefore have that:

$$\mathbb{E}_{(s,t)\sim\beta}[Q_t^{\bar{\pi}^*}(s, \bar{\pi}_t^*(s))] \leq \mathbb{E}_{(s,t)\sim\beta}[Q_t^{\pi^*}(s, \bar{\pi}_t(s))] < \mathbb{E}_{(s,t)\sim\beta}[Q_t^{\pi^*}(s, \pi_t^*(s))]$$

which leads to a contradiction, hence (2.6) must hold.

Next, applying the interchangeability property (see Shapiro (2017)) to equation (2.6) and using the fact that the β distribution puts positive probability on all time periods and all sub-regions of $\mathcal{S} \times \mathcal{A}$, we know that the following necessarily hold:

$$\pi_t^*(s) \in \arg\min_a Q_t^{\pi^*}(s, a),\ \forall s \in \mathcal{S}, \forall t \in \{0, \ldots, T\}\,.$$

Our last step involves using recursion to show that $\pi^* \in \arg\min_\pi Q_t^\pi(s_t, \pi_t(s_t))$ for all t and all s_t. We start once more at $t = T$ where:

$$Q_T^{\pi^*}(s_T, \pi_T^*(s)) = \min_{a_T} Q_T^{\pi^*}(s_T, a_T) = \min_{a_T} -r_T(s_T, a_T, s_T) = Q_T^\pi(s_T, \pi_T(s_T)), \quad \forall \pi.$$

And then recursively,

$$\begin{aligned}
Q_t^{\pi^*}(s_t, \pi_t^*(s_t)) &= \min_{a_t} Q_t^{\pi^*}(s_t, a_t) \\
&= \min_{a_t} \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \pi_{t+1}^*(s_{t+1}))|s_t) \\
&\leq \min_{a_t} \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1}))|s_t) \; \forall \pi \\
&\leq \bar{\rho}(-r_t(s_t, \pi_t(s_t), s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1}))|s_t) \; \forall \pi \\
&\leq \min_{\pi} Q_t^\pi(s_t, \pi_t(s_t)).
\end{aligned}$$

□

In the context of a deep reinforcement-learning approach, we can employ a procedure based on off-policy deterministic policy gradient (Silver et al., 2014) to optimize (2.5). Specifically, given a policy network π^θ, we wish to optimize:

$$\min_\theta \mathbb{E}_{\substack{\tilde{t} \sim \{0,\ldots,T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, \pi_{\tilde{t}}^\theta(s_{\tilde{t}})) \right],$$

using a stochastic gradient algorithm. In doing so, we rely on the fact that:

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{\substack{\tilde{t} \sim \{0,\ldots,T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, \pi^\theta(s_{\tilde{t}})) \right] \\
&= \mathbb{E}_{\substack{\tilde{t} \sim \{0,\ldots,T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[\nabla_\theta Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a)\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} + \nabla_a Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a) \nabla_\theta \pi_{\tilde{t}}^\theta(s_{\tilde{t}})\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} \right] \\
&\approx \mathbb{E}_{\substack{\tilde{t} \sim \{0,\ldots,T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[\nabla_a Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a) \nabla_\theta \pi_{\tilde{t}}^\theta(s_{\tilde{t}})\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} \right].
\end{aligned}$$

Note that in the above equation, we have dropped the the term that depends on $\nabla_\theta Q_{\tilde{t}}^{\pi^\theta}$ as is commonly done in off-policy deterministic gradient methods and usually motivated by a result of Degris et al. (2012), who argue that this approximation preserves the set of local optima in a risk neutral setting, i.e. $\rho(\cdot) := \mathbb{E}[\cdot]$. While we do consider as an important subject of future research to extend this motivation to general recursive risk measures, our

numerical experiments (see Section 2.4.3) will confirm empirically that the quality of this approximation permits the identification of nearly optimal hedging policies.

Given that we do not have access to an exact expression for $Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a)$, this operator needs to be estimated directly from the training data. Exploiting the fact that ρ is a utility-based shortfall risk measure, we get that:

$$Q_t^\pi(s_t, a_t) \in \arg\min_q \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[\ell(q + r(s_t, a_t, s_{t+1}) - Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1})))]$$

where $\ell(y) := (\tau \mathbf{1}\{y > 0\} - (1-\tau)\mathbf{1}\{y \leq 0\})y^2$ is the score function associated to the τ-expectile risk measure (see Definition 12). As explained in Shen et al. (2014) for the case of a tabular MDP, this suggests using the following stochastic gradient step to improve each expectile estimators:

$$Q_t^\pi(s_t, a_t) \leftarrow Q_t^\pi(s_t, a_t) - \alpha\partial\ell(Q_t^\pi(s_t, a_t) + r(s_t, a_t, s_{t+1}) - Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1})))\,,$$

where $\partial\ell(y) := 2(\tau \max(0, y) - (1-\tau)\max(0, -y))$ is the derivative of $\ell(y)$.

In the non-tabular setting, we replace $Q_t^\pi(s_t, a_t)$ with two estimators: i.e. the "main" network $Q_t^\pi(s_t, a_t|\theta^Q)$ for the immediate conditional risk and the "target" network $Q_t^\pi(s_t, a_t|\theta^{Q'})$ for the next period's conditional risk. The procedure consists in iterating between a step that attempts to make the main network $Q_t^\pi(s_t, a_t|\theta^Q)$ a good estimator of $\rho(-r(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, a_{t+1}|\theta^{Q'}))$ and a step that replaces the target network $Q_t^\pi(s_t, a_t|\theta^{Q'})$ with a network more similar to the main one $Q_t^\pi(s_t, a_t|\theta^Q)$. The former is achieved, similarly as with the policy network, by searching for the optimal θ^Q according to:

$$\min_{\theta^Q} \mathbb{E}_{\substack{\tilde{t}\sim\{0,\ldots,T-1\} \\ s_{t+1}\sim P(\cdot|s_t,\bar{\pi}_t(s_t))}} [\ell(Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q) + r(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}}), s_{\tilde{t}+1}) - Q_{\tilde{t}+1}^\pi(s_{\tilde{t}+1}, \pi_{\tilde{t}+1}(s_{\tilde{t}+1})|\theta^{Q'}))]\,,$$

which suggests a stochastic gradient update of the form:

$$\theta^Q \leftarrow \theta^Q - \alpha\partial\ell(Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q) + r(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}}), s_{\tilde{t}+1}) - Q_{\tilde{t}+1}^\pi(s_{\tilde{t}+1}, \pi_{\tilde{t}+1}(s_{\tilde{t}+1})|\theta^{Q'}))\nabla_{\theta^Q} Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q)\,.$$

These two types of updates are integrated in our proposed expectile-based actor-critic deep RL (a.k.a. ACRL) algorithm described in Algorithm 1. One may note that in each episode, the reference policy $\bar{\pi}_t$ is updated to be a perturbed version of the main policy network in order to focus the accuracy of the main critic network $Q(s, a|\theta^Q)$ value and derivatives on actions that are more likely to be produced by the main policy network. We also choose to

update the target networks using convex combinations operations as is done in Lillicrap et al. (2015) in order to improve stability of learning.

Algorithm 1 : The actor-critic RL algorithm for the dynamic recursive expectile option hedging problem (ACRL)

Randomly initialize the main actor and critic networks' parameters θ^{π} and θ^{Q};

Initialize the target actor, $\theta^{\pi'} \leftarrow \theta^{\pi}$, and critic, $\theta^{Q'} \leftarrow \theta^{Q}$, networks;

for $j = 1 : \#Episodes$ **do**

Randomly select $t \in \{0, 1, ..., T-1\}$;

Sample a minibatch of N triplets $\{(s_t^i, a_t^i, s_{t+1}^i)\}_{i=1}^N$ from $P(\cdot|s_t, \bar{\pi}_t(s_t))$, where $\bar{\pi}_t(s_t) := \pi_t(s_t|\theta^{\pi}) + \mathcal{N}(0, \sigma)$;

Update the main critic network:

$$\begin{aligned}\theta^{Q} \leftarrow \theta^{Q} - \alpha \frac{1}{N} \sum_{i=1}^{N} \partial \ell [Q_t(s_t^i, a_t^i|\theta^{Q}) + r(s_t^i, a_t^i, s_{t+1}^i) - \\ Q_{t+1}(s_{t+1}^i, \pi_{t+1}(s_{t+1}^i|\theta^{\pi'})|\theta^{Q'})] \nabla_{\theta^{Q}} Q_t(s_t^i, a_t^i|\theta^{Q})\end{aligned} \tag{2.7}$$

Update the main actor network:

$$\theta^{\pi} \leftarrow \theta^{\pi} - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_t(s_t^i, a|\theta^{Q})|_{a=\pi_t(s_t^i|\theta^{\pi})} \nabla_{\theta^{\pi}} \pi_t(s_t^i|\theta^{\pi})$$

Update the target networks:

$$\begin{aligned}\theta^{Q'} &\leftarrow \alpha\theta^{Q} + (1-\alpha)\theta^{Q'} \\ \theta^{\pi'} &\leftarrow \alpha\theta^{\pi} + (1-\alpha)\theta^{\pi'}\end{aligned} \tag{2.8}$$

end

Remark 3. We note that in our problem,

$$P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t') = \mathbb{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$$

meaning that the action is not affecting the distribution of state in the next period. This is a direct consequence of using a translation invariant risk measure, which eliminates the need to keep track of the accumulated wealth in the set of state variables as explained in Marzban et al. (2020) and allows the reward function to provide an immediate signal regarding the quality of implemented actions. In the context of our deep reinforcement-learning approach, we observed that convergence speed is improved in training due to this property.

Furthermore, the fact that this property makes the dynamics exogenous lifts the need for keeping a replay buffer, which is also known to affect negatively convergence speed.

Remark 4. It is worth noting that there has been a large number of DRL approaches recently proposed to address risk-averse MDP using coherent risk measures. However, to the best of our knowledge, all of those that are model-free, except for two exceptions, consider a law invariant risk measure (i.e. a static risk measure) applied on the discounted sum of total rewards (see Castro et al. (2019); Singh et al. (2020); Urpí et al. (2021)). Such methods therefore suffer from the issues identified in Section 2.2.2. The two exceptions consist of Tamar et al. (2015) and Huang et al. (2021) who propose ACRL algorithms to deal with general dynamic law-invariant coherent risk measures. While being applicable to a wider range of dynamic risk measures, the two algorithms either assume that it is possible to generate samples from a perturbed version of the underlying dynamics, or rely on training three additional neural networks (namely a state distribution reweighting network, a transition perturbation network, and a Lagrangean penalisation network) concurrently with the actor and critic networks. Furthermore, only Huang et al. (2021) actually implemented their method. This was done on a toy tabular problem involving 12 states and 4 actions where it produced questionable performances[4]. While our approach can only be used with the dynamic expectile risk measure, it offers a much simpler implementation that naturally extends DDPG to the risk-averse setting. Section 2.4 will present a real application of this approach on an option hedging problem involving a portfolio of 6 different assets.

2.4 Experimental results

In this section we provide two different sets of experiments that are run over one vanilla and one basket option. We will compare both algorithmic efficiency and quality, in terms of pricing and hedging strategies, of the dynamic risk model (DRM), which employs a dynamic expectile risk measure and is solved using our new ACRL algorithm, and the static risk model (SRM), which employs a static expectile measure and is solved using an AORL algorithm similar to Carbonneau and Godin (2021). All experiments are done using

[4]At the time of writing this paper, the risk-averse implementation of this algorithm in Huang et al. (2021) was unable to recommend an optimal risk neutral policy in a deterministic setting, while the risk neutral implementation produced policies that were outperformed by risk-averse ones in a stochastic setting.

simulated price processes of five risky assets: AAPL, AMZN, FB, JPM, GOOGL. The price paths are simulated using correlated Brownian motions considering the empirical mean, variance, and the correlation matrix of five reference stocks (AAPL, AMZN, FB, JPM, and GOOGL) over the period between January 2019 to January 2021. The vanilla option will be over AAPL while the basket option will contain all five stocks. In both cases, the maturity of the option will be one year and the hedging portfolios will be rebalanced on a monthly basis. Table 2.2 provides the descriptive statistics of our underlying stochastic process.

Table 2.2 – Stock data including the mean, standard deviation, and the correlation matrix

	AAPL	AMZN	FB	JPM	GOOGL
S_0	78.81	1877.94	221.77	137.25	1450.16
μ	-0.0015	-0.0017	-0.0001	0.0006	-0.0004
σ	0.0298	0.0243	0.0295	0.0345	0.0246
AAPL	1.0000	0.7133	0.7744	0.5383	0.7680
AMZN	0.7133	1.0000	0.6903	0.2685	0.6837
FB	0.7744	0.6903	1.0000	0.4807	0.8054
JPM	0.5383	0.2685	0.4807	1.0000	0.6060
GOOGL	0.7680	0.6837	0.8054	0.6060	1.0000

In what follows, we first explain the network architecture of our ACRL model, which is composed of an actor and a critic network. Then, the training procedure of the network under the conditional risk measurement using unconditional assessment of risk is elaborated. We also numerically demonstrate the benefit of exploiting translation invariance in an option hedging problem using RL, which is for a different purpose than what is previously shown by Marzban et al. (2020) in a DP setting. Finally, the main numerical results of the paper is presented for pricing and hedging a vanilla and a basket option. This example illustrates and quantifies the advantages of a time-consistent risk measurement over a time inconsistent approach. In particular, we first focus on the vanilla option to show the precision of our approach by bench-marking its results against a discretized DP model and then extend the results to the case of basket options.

2.4.1 Actor and critic network architecture

Our implementation of the ACRL algorithm involves two networks, one for the actor and one for the critic, both of which are presented in Figure 2.1. Since the numerical experiments assume that the underlying assets of the options follow a Brownian motion process, the model only needs to consider the most recent price for each asset to make investment decisions and the time to maturity. Consequently, the input state to each of the actor and critic networks includes the logarithm of each asset's cumulative return, and the time remaining until maturity, which together correspond to an input vector of dimension $m + 1$.

The actor network is composed of three fully connected layers where the number of neurons are chosen to be $k = 32$ in the first two layers and then maps back to the number of assets in the last layer so that the model generates the investment policy accordingly for each asset. The activation functions in our networks are considered to be *tanh* functions. In the last layer, this implies that the actions will lie in $[-1, 1]^m$.

The critic network is operating on the same state information, while the m dimensional action information vector is only concatenated to the output of the third layer. The first three layers of the critic network follow the same structure as the actor network in terms of the number of neurons, then after concatenating the action into the network, the two fully connected layers following the concatenation maps the number of neurons again to $k = 32$. Finally, the last layer is a fully connected layer with one neuron to make sure that the output is a scalar representing the approximated Q value function.

2.4.2 ACRL training procedure for DRM and the role of translation invariance

We now explain the training procedure employed for the actor and critic networks in the DRM. Recall that in an SRM setting, overfitting of any DRL algorithm can be controlled by measuring the performance of the trained policy on a validation data set using an empirical estimate of the risk-averse objective as validation score. Unfortunately, this is no longer possible in the case of DRMs since the risk measure relies on conditional risk measurements of the trajectories produced by our policy. In theory, estimates of such conditional measurements could be obtained by training a new critic network using the

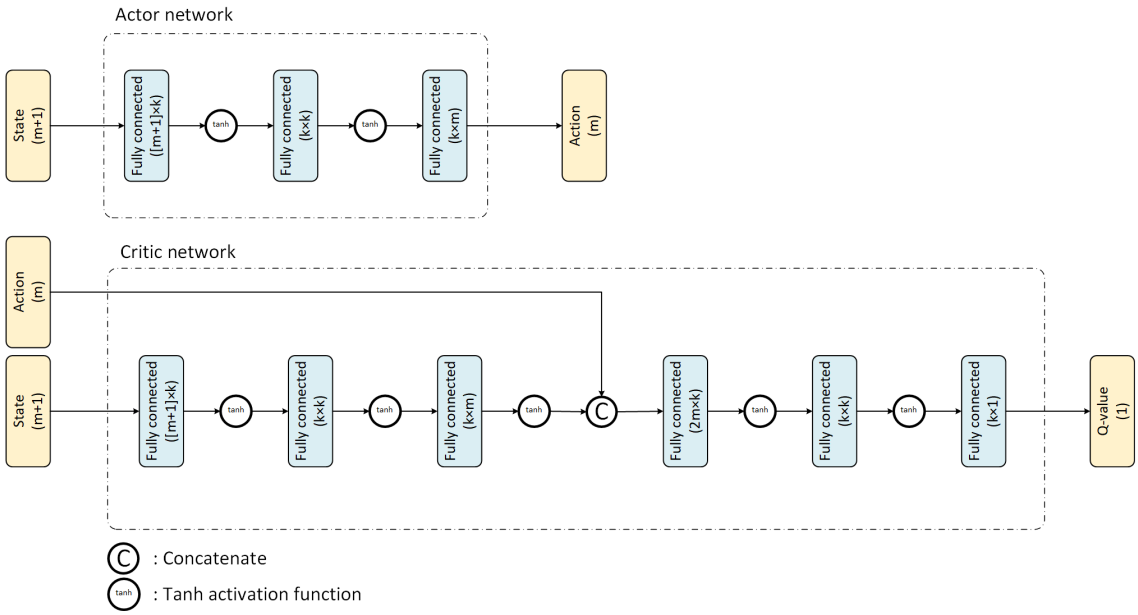
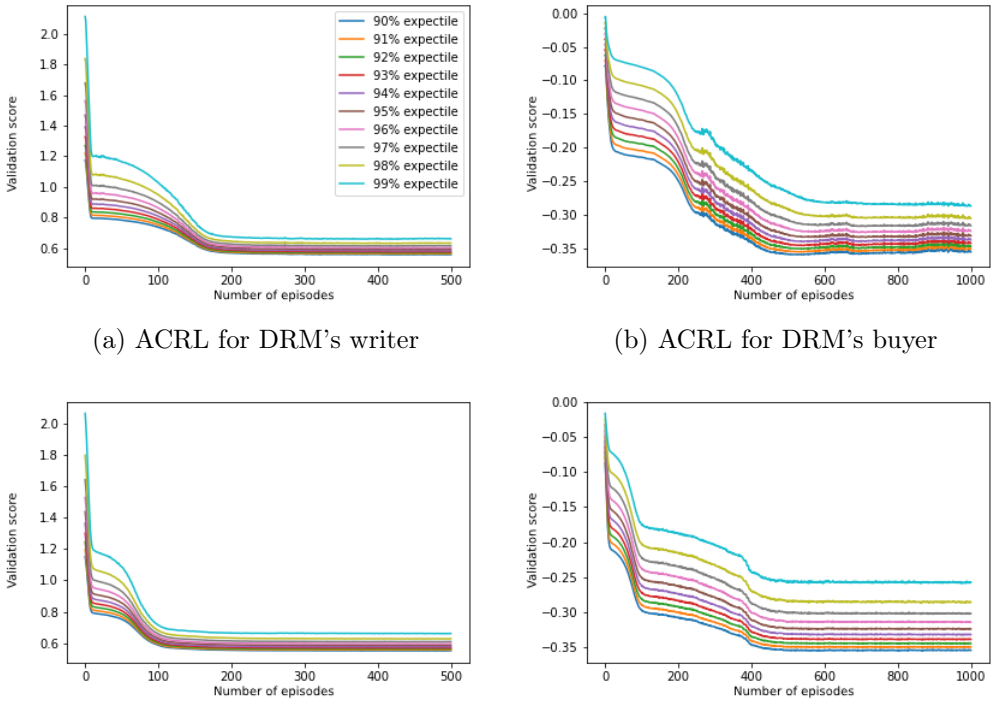


Figure 2.1 – The architecture of the actor and critic networks in ACRL algorithm.

validation set (while maintaining the policy fixed to the trained one). In practice, this is highly computationally demanding to perform in the training stage and raises a new issue of how to control overfitting of the validation score estimate. Our solution for this problem is to rely on using a static risk measure as validation score. Given that it is unclear how to best replace a dynamic expectile risk measure with a static one, we choose to compute a set of validation scores that report the performance for a set of static expectiles at risk levels that are larger or equal to the risk level of the DRM. Relying on higher risk levels is motivated by the fact that dynamic expectile measures capture a more risk-averse attitude than their static counterpart at the same risk level τ. Figure 2.2(a) and (b) show examples of learning curves for the validation performance of a DRM when trained to hedge the writer and buyer positions of a vanilla option at a risk level of $\tau = 90\%$. In this experiment, there is evidence that convergence happens for all levels of $\tau \geq 90\%$. This approach is applied in all of our experiments for choosing the optimal number of episodes. We also note that both our training and validation sets included 1000 trajectories from the underlying geometric Brownian motion process. This implies that the training procedure used in these experiments can naturally extend to settings where only historical data is available.

We close this section with a short discussion about the role of the translation invariance



(a) ACRL for DRM's writer

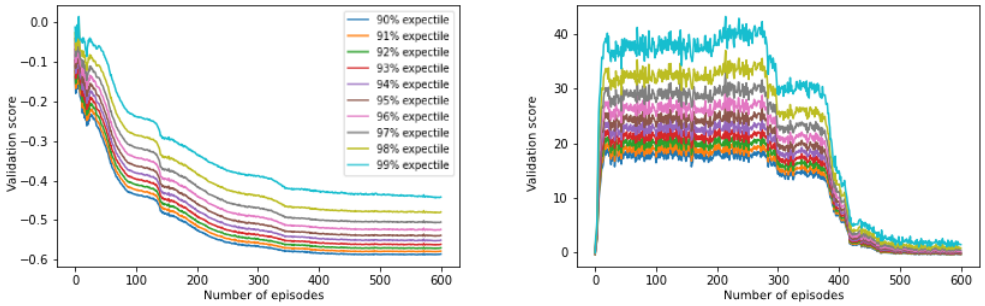
(b) ACRL for DRM's buyer

(c) AORL for SRM's writer

(d) AORL for SRM's buyer

Figure 2.2 – Learning curves of the DRM and SRM for an at-the-money vanilla call option on AAPL when a 90% expectile measure is used. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.

property of dynamic risk measures. In particular, the work of Marzban et al. (2020) explains how, without this property, the dynamic programming equations need to keep track of the wealth accumulated since $t = 0$ using an additional state variable that gets only employed at $t = T$. More importantly, without translation invariance, the MDP representation ends up only having a reward at $t = T$ thus preventing the ACRL algorithm from receiving quick feedback about the quality of the actions that it is proposing. To illustrate the effect of this property, we compared the convergence of the training process for the ACRL algorithm under both form of DP representation of the buyer's DRM. Namely, Figure 2.3 presents the learning curves of ACRL with immediate rewards as described in Section 2.3.2, while (b) presents the learning curves for an implementation in which all the rewards are delayed (using an additional state variable) until $t = T$. These figures clearly show that the MDP with immediate rewards is much easier to train than the delayed rewards MDP. In



(a) With immediate rewards (b) With delayed rewards

Figure 2.3 – Learning curves of the ACRL algorithm for the buyer's DRM when using (a) the immediate rewards versus (b) delayed rewards in the hedging of a vanilla call at-the-money option.

particular, not only does this model converge in less number of episodes, it also ends up converging to a better solution: the immediate rewards MDP converges to a risk of -0.59 for the buyer (0.91 for the writer), while with delayed rewards it converges to -0.41 (1.01).

2.4.3 Vanilla call option pricing and hedging

In our first set of experiments, we consider pricing and hedging an at-the-money vanilla call option on AAPL. We should first note that solving a hedging problem, e.g. DRM, for a vanilla option is not particularly difficult since the number of state variables in this case is small. It is possible to obtain (approximately) optimal solutions by dynamic programming (Marzban et al. (2020)). Our purpose of considering the case of vanilla option is twofold. First, it provides a useful basis for checking the accuracy of solutions obtained from our deep reinforcement-learning (DRL) methods against the "true" optimal solutions, namely by comparing against the DP solutions. Such an accuracy check would be useful for justifying our use of DRL later in this article as a general means to evaluate hedging performance and calculate the equal risk price (which becomes necessary for problems that cannot be solved by DP such as the case of basket options discussed in the next section). Second, the setting of a vanilla option also allows us to provide a more accurate comparison between DRM and SRM and demonstrate the advantage of the former, i.e. the benefit of time-consistent hedging policies, particularly when options with different time to maturity need to be considered.

To proceed, we first detail how the experiments are conducted. First, the initial price of the underlying stock AAPL is always set to be 78.81, and the hedging portfolio is rebalanced on a monthly basis. Options with different time to maturity are considered, ranging from one month to one year. We generate from a Brownian motion three sets of price trajectories with one year time window, one for training, one for validation, and one for testing, and each consists of 1000 trajectories. In the training phase, we solve both DRM and SRM for the writer and buyer's hedging problems using the longest maturity time, i.e. one year, as the hedging horizon. In solving the DRM, a policy and a critic network are trained using ACRL, whereas in solving the SRM, only a policy network is trained using AORL. See also Section 4.2 regarding how the validation is done to guide the training. Figure 2.2 presents the learning curves for the training of the hedging policies of the DRM and the SRM with a risk level of $\tau = 90\%$. SRM appears to have a faster rate of convergence than DRM, which might not be surprising given that the architecture of SRM is simpler [5]. It is however worth noting that the issue of time inconsistency for SRM implies that it can potentially produce poor quality policies and prices when the maturity of the option is modified unless it is completely retrained for each type of maturity. This is not the case for DRM and will be further discussed below.

With the trained DRM and SRM policy networks for a fixed 1 year maturity and risk aversion level $\tau \in \{75\%, 90\%, 95\%\}$, we can evaluate the writer and the buyer's (out-of-sample) risk exposure over a pre-specified time horizon so as to calculate the corresponding ERP. We consider the following three metrics for measuring the realized risk under different hedging policy and explain the methods used for calculating the metrics:

- *Out-of-sample static expectile risk*: Given a trained policy network, use the testing data to calculate the static expectile risk obtained when hedging the option using this policy.

- *RL based out-of-sample dynamic expectile risk estimation*: Use the testing data to train only the critic network in ACRL for evaluating the out-of-sample dynamic risk. In particular, by fixing the policy network in ACRL to a trained policy network, the

[5]The policy network at SRM model is exactly the actor network of DRM, while the quality of actions are directly evaluated in the absence of a trained critic network.

critic network trained based on testing data provides an estimate of the out-of-sample dynamic expectile risk. To speed up the training of the critic network, one may initialize the critic network using the network trained previously with the training data.

- *DP based out-of-sample dynamic expectile risk estimation*: Given a trained policy network, evaluate the "true" dynamic expectile risk by solving the dynamic programming equations, under the fixed policy, using a high precision discretization of the states, actions, and transitions. Note that this metric is available neither for the case of basket option nor in a data-driven environment where the stochastic process is unknown.

We note that our RL based estimate of out-of-sample dynamic risk is a novel concept, which refers to the calculation of dynamic risk based on testing data. This is possible, as explained above, by training only the critic network using ACRL on the test data. This metric is especially relevant given that classical methods for calculating dynamic risk, such as our DP based estimate, assume full knowledge of the stochastic model that captures the dynamics of an underlying system, i.e. stock price, and require the resolution of dynamic programming equations, which is known to suffer from the curse of dimensionality. Consequently, such methods can no longer be used when the DP equations require a large state space, as can be the case with basket options, or when the description of the underlying stochastic process is unknown.

In our experiments, we apply the second and third metric to the trained DRM policies and the first metric to both the trained DRM and SRM policies. In the former case, we are interested in demonstrating that the RL based out-of-sample expectile risk estimate is an accurate metric. Namely, we will compare the RL based estimate against the "true" DP based estimate. In the latter case, we will shed light on how the DRM policy performs when evaluated according to other metrics that are also of interest to practitioners. In particular, the static expectile risk measure, despite its issue of time inconsistency, can still have its intuitive appeal as a metric, and one may be interested in knowing how a DRM policy performs against this metric as compared to an SRM policy.

Figure 2.4 summarizes the evaluations of out-of-sample dynamic risk for DRM policies trained for 1 year maturity then applied to options of different maturities ranging from 12 months to 2 months. One can observe that the risk of the writer decreases monotonically for options of shorter maturities, whereas the risk of the buyer increases monotonically. This is consistent with the fact that there is less uncertainty for a shorter hedging horizon, which favors the writer's risk exposure more than the buyer's when considering an at-the-money option. This also provides the evidence that the DRM policies, albeit only trained based on the longest time to maturity, i.e. one year, can be well applied to hedge options with shorter time to maturity and be used to draw consistent conclusion. The observation that the DRM policies remain good policies for problems with shorter time to maturity testifies of the value of using a time-consistent hedging model. Another important observation one can make is that the RL based out-of-sample dynamic risk estimate is generally very close to the DP based estimate across all conditions. The difference between the two appears to be more noticeable for the case of high risk aversion, i.e. $\tau = 95\%$ and long time to maturity, but the difference remains minor overall. This observation allows us to confirm the accuracy of our RL based out-of-sample dynamic risk estimation procedure as a replacement for the DP based estimation in settings where the latter cannot be used.

Figure 2.5 reports the out-of-sample static risk for both SRM policies and DRM policies. The results are interesting and perhaps surprising. First, unlike the consistent behavior observed in the case of dynamic risk, i.e. Figure 2.4, the static risk of SRM policies for the seller (resp. buyer) may increase (resp. decrease) when hedging an option with shorter maturity. The possibility that a seller's policy may actually increase risk when applied to an option with shorter maturity is clearly problematic when the underlying asset follows a geometric Brownian motion with positive drift, as it is inconsistent with the fact that there is less uncertainty (and lower expected value) regarding the payout of such options. This inconsistency occurs because the SRM policies are only trained based on the longest time to maturity, i.e. one year, and they cannot be well applied, unlike for the case of DRM policies, to problems with shorter time to maturity due to the violation of the time-consistency property. It is clear from the figures that the SRM policies can be far from the optimal policies when applied to a shorter time to maturity. On the other hand, the DRM policies can actually be found not only to outperform SRM policies in terms of static risk exposure

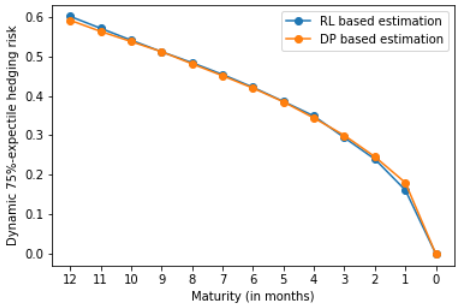
but also to generate consistent results across time, i.e. risk decreases (resp. increases) for the seller (resp. buyer) as the time to maturity decreases. This can be somewhat surprising, as the DRM policies are optimized based on dynamic risk measures rather than the static ones, but the policies can still perform well when evaluated according to static risk measures. Overall, the results presented in Figure 2.5 best showcase the strength of time-consistent policies and why such policies are important to consider in settings where risk needs to be re-evaluated across different time points or maturity dates.[6] We suspect that the possibility that SRM policies may not account properly for risk aversion at some future time point or for other range of option maturities should seriously hinder their use in practice.

In order to be more precise about results presented in figures 2.4 and 2.5, we detail in Table 2.3 all the numerical results for the case of high risk aversion, i.e. $\tau = 90\%$, along with the equal risk prices calculated based on RL based out-of-sample dynamic risk estimate and based on the discretized DP (referred as True ERP).[7] One first confirm that the RL based estimateof ERP is a high quality approximation of the true ERP in this vanilla option pricing setting, with a maximum approximation error of 0.01 over all maturity dates. Moreover, we can see that the prices for the SRM polices are generally higher than the prices for the DRM polices. The observation is that while DRM policies are less risky than SRM policies across different time to maturity, it is the writer that benefits more from the use of DRM than the buyer. This could be related to the fact that the writer's loss due to the option payout is unbounded while the option protects the buyer from losses. This in turns implies that the writer's risk exposure is larger in this transaction. Thus, the choice of a policy can be more critical to the writer than the buyer. As the risk exposure of the writer decreases more than for the buyer, this leads to lower ERP price for DRM policies.

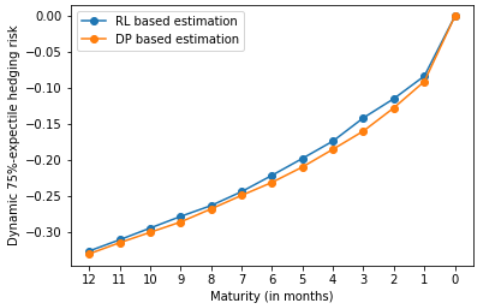
Finally, Figure 2.6 presents the optimal policies of the two models (i.e., DRM and SRM), together with the actual optimal policy of DRM, obtained using a high precision dynamic program (referred as DP-DRM). Each subfigure shows the policy as a function of current

[6]Indeed, recall that the example in Section 2.2.2 demonstrated that the fact that SRM was time inconsistent implied that its policy might not remain a reasonable risk-averse policy at future time points. This phenomenon is implicitly observed in Figure 2.5 given that the MDP is stationary so that the risk measured for a maturity t is exactly equal to the risk measured at time $T - t$ when $S_t = S_0$.

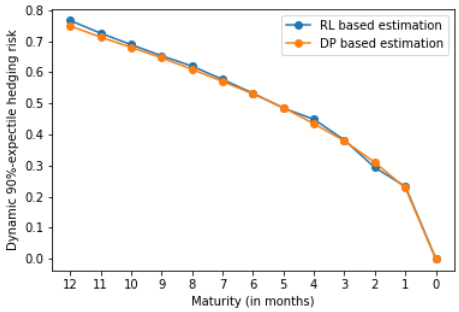
[7]Note that in a purely data-driven setting, the ERP could either be estimated using the in-sample trained critic network, or by calculating our RL based estimate using some freshly reserved data to reduce overfitting biases.



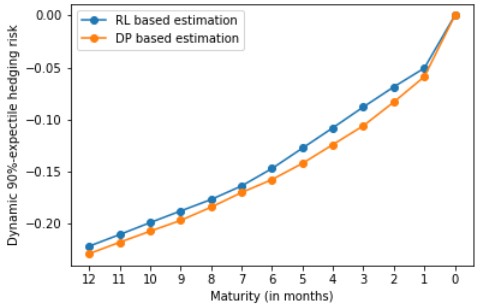
(a) Writer, $\tau = 75\%$



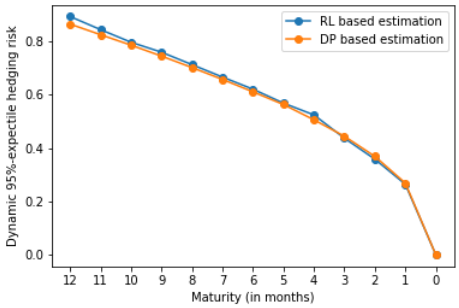
(b) Buyer, $\tau = 75\%$



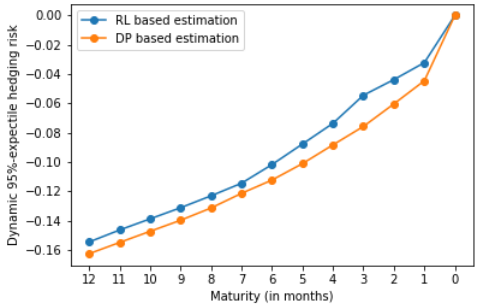
(c) Writer, $\tau = 90\%$



(d) Buyer, $\tau = 90\%$

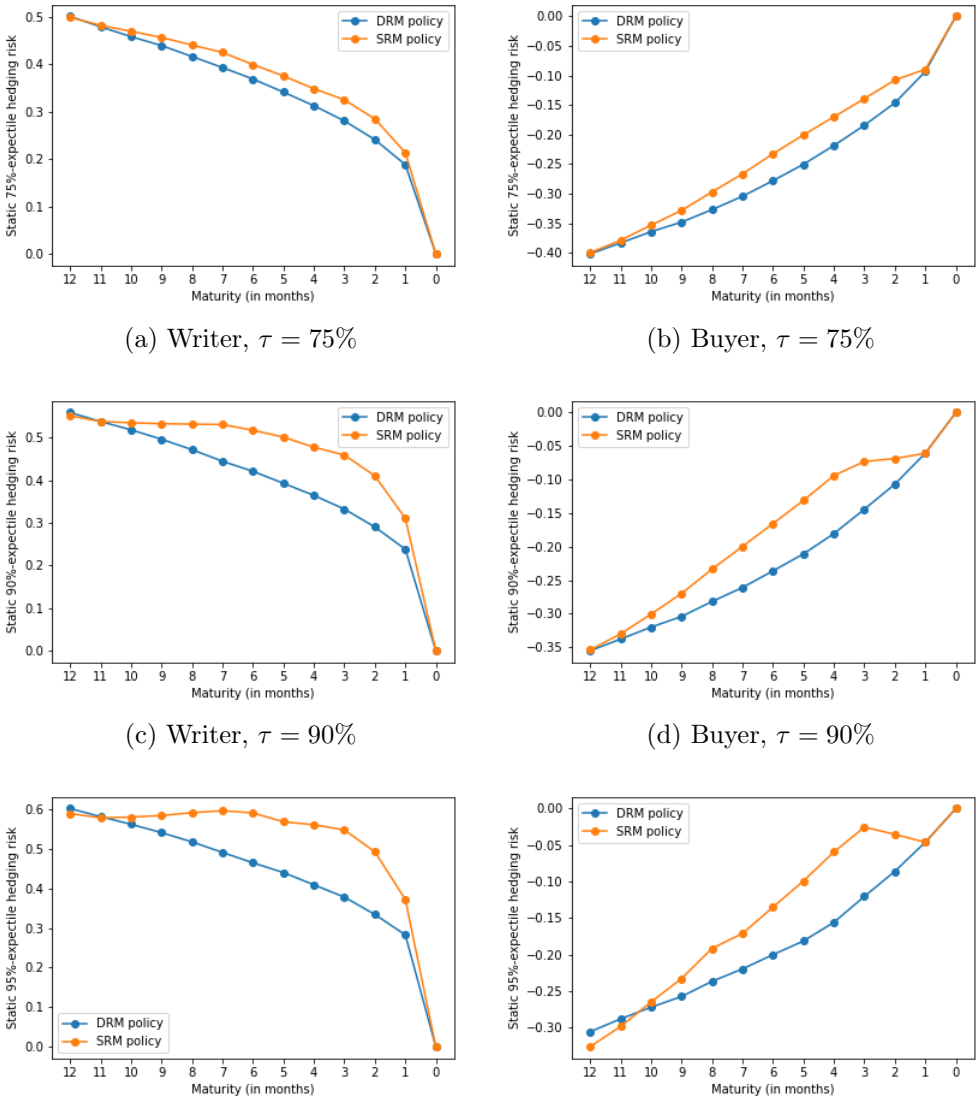


(e) Writer, $\tau = 95\%$



(f) Buyer, $\tau = 95\%$

Figure 2.4 – The out-of-sample dynamic risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 0 months) under the DRM policy trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.



(a) Writer, $\tau = 75\%$

(b) Buyer, $\tau = 75\%$

(c) Writer, $\tau = 90\%$

(d) Buyer, $\tau = 90\%$

(e) Writer, $\tau = 95\%$

(f) Buyer, $\tau = 95\%$

Figure 2.5 – The out-of-sample static risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 2 months) under the DRM and SRM policies trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.

Table 2.3 – The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of vanilla at-the-money call options over AAPL, with maturities ranging from 12 to 0 months, when hedged using the DRM and the SRM policies trained at risk level $\tau = 90\%$ and for a 12 months maturity. Associated ERPs under the DRM are also compared to the "true" ERP measured using a discretized MDP.

		Time to maturity											
Policy	Est.†	12	11	10	9	8	7	6	5	4	3	2	1
Dynamic 90%-expectile risk													
Writer's DRM	RL	0.77	0.73	0.69	0.65	0.62	0.58	0.53	0.48	0.45	0.38	0.29	0.23
	DP	0.75	0.71	0.68	0.65	0.61	0.57	0.53	0.49	0.43	0.38	0.31	0.23
Buyer's DRM	RL	-0.22	-0.21	-0.20	-0.19	-0.18	-0.16	-0.15	-0.13	-0.11	-0.09	-0.07	-0.05
	DP	-0.23	-0.22	-0.21	-0.20	-0.18	-0.17	-0.16	-0.14	-0.12	-0.11	-0.08	-0.06
Static 90%-expectile risk													
Writer's SRM	ED	0.55	0.54	0.54	0.53	0.53	0.53	0.52	0.50	0.48	0.46	0.41	0.31
Writer's DRM	ED	0.56	0.54	0.52	0.50	0.47	0.44	0.42	0.39	0.36	0.33	0.29	0.24
Buyer's SRM	ED	-0.35	-0.33	-0.30	-0.27	-0.23	-0.20	-0.17	-0.13	-0.09	-0.07	-0.07	-0.06
Buyer's SRM	ED	-0.36	-0.34	-0.32	-0.30	-0.28	-0.26	-0.24	-0.21	-0.18	-0.14	-0.11	-0.06
Equal risk prices with DRM													
True ERP		0.49	0.47	0.45	0.42	0.40	0.37	0.34	0.31	0.28	0.24	0.19	0.14
DRM	RL	0.50	0.47	0.45	0.42	0.40	0.37	0.34	0.31	0.28	0.24	0.18	0.14
SRM	RL	0.49	0.46	0.44	0.43	0.40	0.38	0.35	0.33	0.30	0.27	0.24	0.22

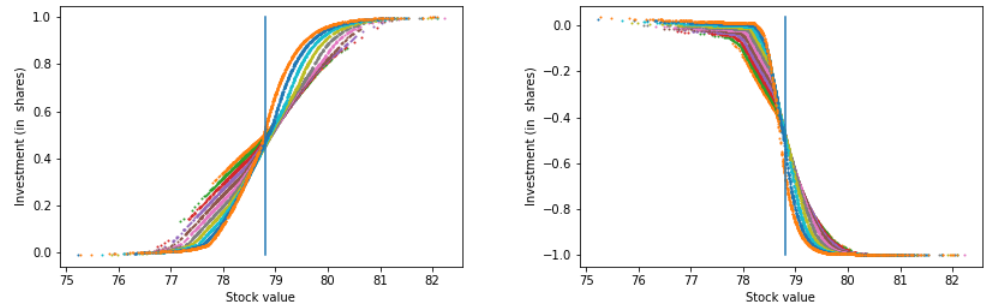
† Estimation (Est.) is either made based on reinforcement-learning (RL), discretized dynamic programming (DP), or with the empirical distribution (ED).

price (x-axis) and time period (colors). The figure further confirms that the policies of both DRM and SRM follow a similar pattern as DP-DRM, which ensures the quality of implementation of both AORL for SRM and ACRL for DRM.

2.4.4 Basket options

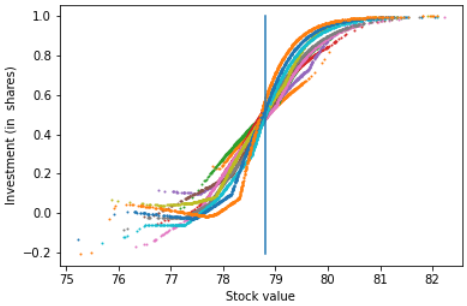
In our second set of experiments, we extend the application of ERP pricing framework to the case of basket options where traditionnal DP solution schemes are not computationally tractable. In particular, we consider an at-the-money basket option with the strike price of 753$ on five underlying assets: AAPL, AMZN, FB, JPM, and GOOGL, where the option payoff is determined by the average price of the underlyings. Similarly to the case of the vanilla option, the rebalancing of the portfolio is happening once per month, options with different maturities from one month to twelve months are considered, and three sets of price trajectories are used for training, validation, and testing the models. We train the ACRL and AORL networks for a one year basket option and then use the same policy network for hedging options with shorter time to maturity.

Our first observation in this set of experiments relates to the training time of the model for the basket option with five assets. Figure 2.7 presents the convergence of the training of the ACRL model under $\tau = 90\%$. When comparing to the case of the vanilla option,

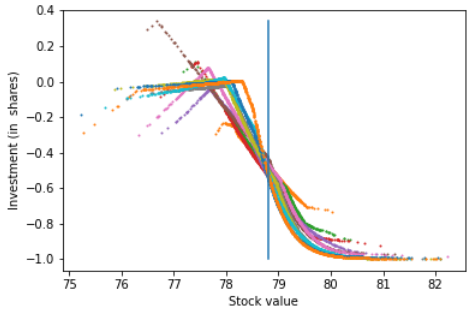


(a) DRM's writer

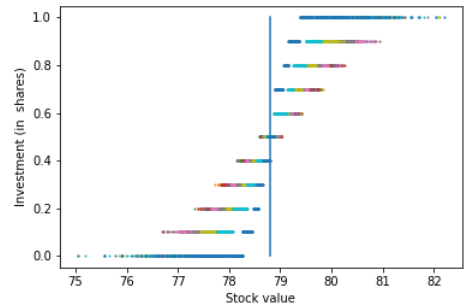
(b) DRM's buyer



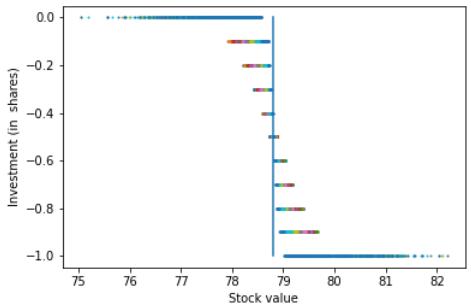
(c) SRM's writer



(d) SRM's buyer



(e) DP-DRM's writer



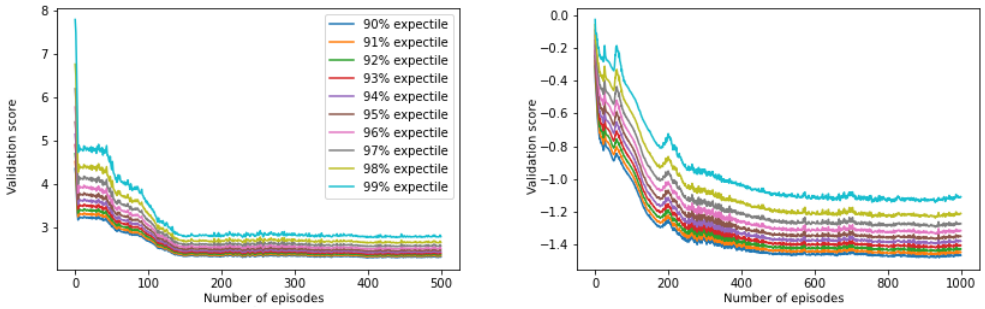
(f) DP-DRM's buyer

Figure 2.6 – Comparison of the optimal DRL policies obtained for DRM and SRM (with 90% expectile measures) to the discretized DP solution (DP-DRM) for an at-the-money vanilla call option on AAPL with a one year maturity. Each figure presents the sampled actions in our simulated trajectories as a function of the AAPL stock value. The strike price is marked at 78.81.

the convergence rate appears to have a similar behavior, i.e., the number of episodes and the time spent on each episode is similar for both the case of the writer and the buyer. This is important as it indicates that the training time might not be very sensitive to the number of assets, while traditional DP approaches are known to become intractable when the option is written on multiple assets.

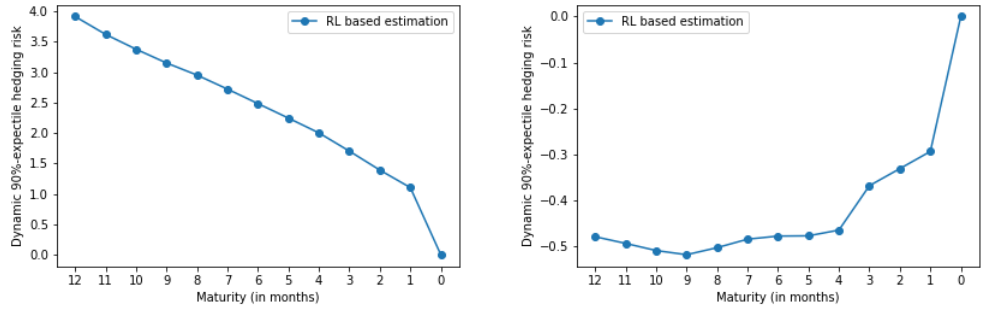
In this section, dynamic risk is estimated using the RL based estimator described in Section 2.4.3 given that the DP estimator requires too much computations and that the RL based one was shown to provide a relatively high precision estimation of the "true" dynamic risk. Following this, in Figure 2.8 (a) and (b) we present the dynamic risk obtained from applying the DRM policy on the test data when the model is trained for a one year maturity option. Hedging risk using the same trained policy is presented for 12 different options with maturity ranging from 0 to 12 months. Similar to the vanilla option case, the dynamic risk of the writer is monotonically decreasing as we get closer to the maturity of the option, which can be attributed to the reduced probability that the average price of the assets significantly diverges from the initial average (i.e., the strike price of the option). On the other side, i.e. for the buyer of the option, although overall the risk is increasing to zero as the maturity gets closer to zero, for longer time to maturities we observe some degradation of risk. We attribute this behavior to the estimation error of the RL based dynamic risk estimator.

In order to have a view of risk that is not perturbed by estimation errors, we also compare the static risk under DRM and SRM as we did for vanilla options. Figure 2.9 (a) and (b) shows the static risk under $\tau = 90\%$. One can first recognize the same monotone convergence to zero of the two sides of the options. However, contrary to the case of the vanilla option, the difference between the static risk performance of DRM and SRM policies are rather similar for all maturity times. It therefore appears that in these experiments with a basket option, both SRM and DRM produce similar polices. One possible reason could be that the range of "optimal" risk-averse investment plans, whether using DRM or SRM, is more limited. Indeed, while for the vanilla option, we observed that the optimal policies generated investments in the range $[0, 1]$ and $[-1, 0]$ for the writer and the buyer respectively, for the basket option we observed wealth allocations that are more concentrated around 0.20 (i.e. the uniform portfolio known for its risk hedging properties) and -0.20 for each of



(a) Writer (b) Buyer

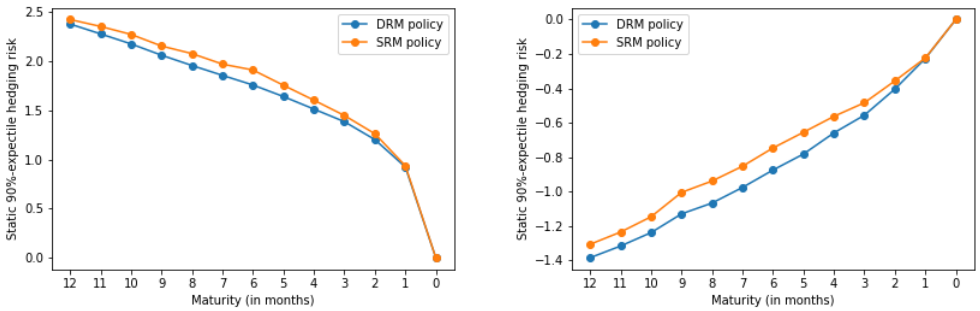
Figure 2.7 – Learning curves of the ACRL algorithm for the writer and buyer's DRM for a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.



(a) Writer (b) Buyer

Figure 2.8 – The out-of-sample dynamic risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under a DRM policy trained for a 12 months maturity.

the 5 assets asset respectively. Finally, similar to the vanilla option case, Table 2.4 presents more details on the results used to produce figures 2.8 and 2.9, along with the equal risk prices computed based on our RL based out-of-sample dynamic risk estimator. The higher ERP price for the SRM policy is an obvious observation in this table, which again can be attributed to the better performing (in terms of dynamic risk) hedging policy produced by ACRL for the DRM, compared to the policy produced by AORL for the SRM.



(a) Static risk, writer

(b) Static risk, buyer

Figure 2.9 – The out-of-sample static risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under the DRM and SRM policies trained for a 12 months maturity.

Table 2.4 – The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of basket at-the-money call options over AAPL, AMZN, FB, JPM, and GOOGL, with maturities ranging from 12 to 0 months, when hedged using the DRM and the SRM policies trained at risk level $\tau = 90\%$ and for a 12 month maturity. Associated ERPs under the DRM are also compared.

		Time to maturity											
Policy	Est.†	12	11	10	9	8	7	6	5	4	3	2	1
Dynamic 90%-expectile risk													
Writer's DRM	RL	3.92	3.62	3.38	3.15	2.95	2.72	2.48	2.25	2.00	1.70	1.39	1.10
Buyer's DRM	RL	-0.48	-0.49	-0.51	-0.52	-0.50	-0.49	-0.48	-0.48	-0.47	-0.37	-0.33	-0.29
Static 90%-expectile risk													
Writer's SRM	ED	2.43	2.36	2.28	2.16	2.08	1.97	1.91	1.76	1.61	1.45	1.26	0.94
Writer's DRM	ED	2.38	2.28	2.18	2.06	1.96	1.86	1.76	1.64	1.51	1.39	1.20	0.92
Buyer's SRM	ED	-1.31	-1.24	-1.15	-1.01	-0.94	-0.85	-0.75	-0.66	-0.56	-0.48	-0.36	-0.22
Buyer's SRM	ED	-1.39	-1.32	-1.24	-1.13	-1.07	-0.98	-0.88	-0.78	-0.66	-0.56	-0.40	-0.23
Equal risk prices with DRM													
DRM	RL	2.20	2.06	1.95	1.84	1.73	1.61	1.48	1.37	1.24	1.04	0.86	0.70
SRM	RL	2.23	2.10	2.01	1.91	1.79	1.65	1.52	1.39	1.21	1.03	0.92	0.82

† Estimation (Est.) is either made based on reinforcement-learning (RL), discretized dynamic programming (DP), or with the empirical distribution (ED).

2.5 Conclusion

In this article, we developed and implemented the first deep reinforcement-learning algorithm for calculating equal risk prices under time-consistent dynamic risk measures. This algorithm exploits the elicitability property of the expectile risk measure to extend in a natural way the famous off-policy deterministic actor-critic method presented in Silver et al. (2014) to the risk-averse setting. Our numerical experiments confirmed that it can identify risk-averse hedging strategies of good quality and be used to estimate the ERP, simultaneously for a range of maturities, using a reasonable amount of computational resources in conditions where

traditional DP methods are impracticable. We also demonstrated important issues regarding the implementability of hedging strategies that are based on static (time inconsistent) risk measures. Namely, both our illustrative example and two numerical experiments demonstrated how the time-consistent policy produced using the DRM might in fact appear preferable to the investor (from the point of view of the time inconsistent static risk measure) as the risk is measured at later points of time, i.e. with shorter maturity. We only evaluated the performance of our model in a synthetic environment using a simple neural network architecture. One may be interested in examining the performance of this model under real market conditions. In particular, in a simulation environment having access to infinite i.i.d. samples makes training much easier to machine learning models. In a real market environment, where the available data is limited to some non-stationary samples of past historical prices, training an on-policy network will face serious issues associated to lack of exploration. This might result in even higher out performance of the ACRL model compared to the AORL, and therefore superior hedging precision under the time-consistent dynamic risk measure. In addition, we only consider European style options in this article, where as demonstrated in Marzban et al. (2020), the ERP model can also be investigated in the case of American options.

References

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Bellini, F. and Bernardino, E. D. (2017). Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, 15(5):725–733.

Bertsimas, D., Kogan, L., and Lo, A. W. (2001). Hedging derivative securities and incomplete markets: an ϵ-arbitrage approach. *Operations Research*, 49(3):372–397.

Carbonneau, A. and Godin, F. (2020). Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, pages 1–16.

Carbonneau, A. and Godin, F. (2021). Deep equal risk pricing of financial derivatives with multiple hedging instruments. *arXiv preprint arXiv:2102.12694*.

Castro, D. D., Oren, J., and Mannor, S. (2019). Practical risk measures in reinforcement learning. *ArXiv*, abs/1908.08379.

Chen, J. M. (2018). On exactitude in financial regulation: Value-at-risk, expected shortfall, and expectiles. *Risks*, 6(2):61.

Degris, T., White, M., and Sutton, R. S. (2012). Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 179–186, Madison, WI, USA. Omnipress.

Detlefsen, K. and Scandolo, G. (2005). Conditional and dynamic convex risk measures. *Finance and stochastics*, 9(4):539–561.

Guo, I. and Zhu, S.-P. (2017). Equal risk pricing under convex trading constraints. *Journal of Economic Dynamics and Control*, 76:136–151.

Huang, A., Leqi, L., Lipton, Z. C., and Azizzadenesheli, K. (2021). On the convergence and optimality of policy gradient for markov coherent risk.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Marzban, S., Delage, E., and Li, J. Y. (2020). Equal risk pricing and hedging of financial derivatives with convex risk measures. *arXiv preprint arXiv:2002.02876*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Rudloff, B., Street, A., and Valladão, D. M. (2014). Time consistency and risk averse dynamic decision models: Definition, interpretation and practical consequences. *European Journal of Operational Research*, 234(3):743–750.

Shapiro, A. (2017). Interchangeability principle and dynamic equations in risk averse stochastic programming. *Operations Research Letters*, 45(4):377–381.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR.

Singh, R., Zhang, Q., and Chen, Y. (2020). Improving robustness via risk averse distributional reinforcement learning. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M., editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 958–968.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2015). Policy gradient for coherent risk measures. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Urpí, N. A., Curi, S., and Krause, A. (2021). Risk-averse offline reinforcement learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Chapter 3

WaveCorr: Correlation-savvy Deep Reinforcement-Learning for Portfolio Management

Chapter information

This article is a joint work with my supervisors, Erick Delage, and Jonathan Yu-Meng Li, and also my colleagues at Evovest, Jeremie Desgagne-Bouchard, and Carl Dussault. It is submitted to the International Conference on Learning Representations (ICLR).

Abstract

The problem of portfolio management represents an important and challenging class of dynamic decision making problems, where rebalancing decisions need to be made over time with the consideration of many factors such as investors' preferences, trading environments, and market conditions. In this article, we present a new portfolio policy network architecture for deep reinforcement-learning (DRL) that can more effectively exploit cross-asset dependency information and achieve better performance than state-of-the-art architectures. In particular, we introduce a new property, referred to as *asset-permutation invariance*, for portfolio policy networks that exploit multi-asset time series data, and design the first portfolio policy network, named WaveCorr, that preserves this

invariance property when treating asset correlation information. At the core of our design is an innovative permutation invariant correlation processing layer. An extensive set of experiments are conducted using data from both Canadian and American stock markets, and WaveCorr consistently outperforms other architectures with an impressive 3%-25% absolute improvement in terms of average annual return, and up to more than 200% relative improvement in average Sharpe ratio. We also measured an improvement of a factor of up to 5 in the stability of performance under random choices of initial asset ordering and weights. The stability of the network has been found as particularly valuable by our industrial partner.

3.1 Introduction

In recent years, there has been a growing interest in applying Deep Reinforcement-Learning (DRL) to solve dynamic decision problems that are complex in nature. One representative class of problems is portfolio management, whose formulation typically requires a large amount of continuous state/action variables and a sophisticated form of risk function for capturing the intrinsic complexity of financial markets, trading environments, and investors' preferences.

In this article, we propose a new architecture of DRL for solving portfolio management problems that optimize a Sharpe ratio criterion (Moody and Saffell, 1999). While several works in the literature apply DRL for portfolio management problems such as Moody et al. (1998); He et al. (2016); Liang et al. (2018) among others, little has been done to investigate how to improve the design of a neural network (NN) in DRL so that it can capture more effectively the nature of dependency exhibited in financial data. In particular, it is known that extracting and exploiting cross-asset dependencies over time is crucial to the performance of portfolio management. The neural network architectures adopted in most existing works, such as Long-Short-Term-Memory (LSTM) or convolutional neutral network (CNN), however, only process input data on an asset-by-asset basis and thus lack a mechanism to capture cross-asset dependency information. The architecture presented in this paper, named as WaveCorr, offers a mechanism to extract the information of both time-series dependency and cross-asset dependency. It is built upon the WaveNet structure

(Oord et al., 2016), which uses dilated causal convolutions at its core, and a new design of correlation block that can process and extract cross-asset information.

In particular, throughout our development, we identify and define a property that can be used to guide the design of a network architecture that takes multi-asset data as input. This property, referred to as *asset-permutation invariance*, is motivated by the observation that the dependency across assets (cross-asset correlation) has a very different nature from the dependency across time (auto correlation). Namely, given a multivariate time series data, the information that can be extracted from the data for prediction purpose would not be considered the same if the time indices are permuted, while it should remain the same if the asset indices are permuted. While this property may appear more than reasonable, as discussed in section 3, a naive extension of CNN that accounts for both time and asset dependencies can easily fail to satisfy this property. To the best of our knowledge, the only other works that have also considered extracting cross-asset dependency information in DRL for portfolio management are the recent works of Zhang et al. (2020) and Xu et al. (2020). While Zhang et al.'s work is closer to ours in that it is also built upon the idea of adding a correlation layer to a CNN-like module, its overall architecture is different from ours and, most noticeably, their design does not follow the property of asset-permutation invariance and thus its performance can vary significantly when the ordering of assets changes. As further shown in the numerical section, our architecture, which has a simpler yet permutation invariant structure, outperforms in many aspects Zhang et al.'s architecture. The work of Xu et al. (2020) takes a very different direction from ours, which follows a so-called attention mechanism and an encoder-decoder structure. A more detailed discussion is beyond the scope of this paper.

Overall, the contribution of this paper is threefold. First, we introduce a new property, referred to as asset-permutation invariance, for portfolio policy networks that exploit multi-asset time series data. Second, we design the first portfolio policy network, named WaveCorr, that accounts for asset dependencies in a way that preserves this invariance. This achievement relies on the design of an innovative permutation invariant correlation processing layer. Third, and most importantly, we present evidence that WaveCorr significantly outperforms state-of-the-art policy network architectures using data from both Canadian (TSX) and American (S&P 500) stock markets. Specifically, our new architecture leads

to an impressive 5%-25% absolute improvement in terms of average annual return, up to more than 200% relative improvement in average Sharpe ratio, and reduces, during the period of 2019-2020 (i.e. the Covid-19 pandemic), by 16% the maximum daily portfolio loss compared to the best competing method. Using the same set of hyper-parameters, we also measured an improvement of up to a factor of 5 in the stability of performance under random choices of initial asset ordering and weights, and observe that WaveCorr consistently outperforms our benchmarks under a number of variations of the model: including the number of available assets, the size of transaction costs, etc. Overall, we interpret this empirical evidence as a strong support regarding the potential impact of the WaveCorr architecture on automated portfolio management practices, and, more generally, regarding the claim that asset-permutation invariance is an important NN property for this class of problems.

3.2 Problem statement

3.2.1 Portfolio management problem

The portfolio management problem consists of optimizing the reallocation of wealth among many available financial assets including stocks, commodities, equities, currencies, etc. at discrete points in time. In this paper, we assume that there are m risky assets in the market, hence the portfolio is controlled based on a set of weights $\boldsymbol{w}_t \in \mathbb{W} := \{\boldsymbol{w} \in \mathbb{R}_+^m | \sum_{i=1}^m w^i = 1\}$, which describes the proportion of wealth invested in each asset. Portfolios are rebalanced at the beginning of each period $t = 0, 1, ..., T-1$, which will incur proportional transaction costs for the investor, i.e. commission rates are of c_s and c_p, respectively. We follow Jiang et al. (2017) to model the evolution of the portfolio value and weights (see Figure 3.1). Specifically, during period t the portfolio value and weights start at p_{t-1} and $\boldsymbol{w}_{t-1}$, and the changes in stock prices, captured by a random vector of asset returns $\boldsymbol{\xi}_t \in \mathbb{R}^m$, affect the end of period portfolio value $p'_t := p_{t-1}\boldsymbol{\xi}_t^\top \boldsymbol{w}_{t-1}$, and weight vector $\boldsymbol{w}'_t := (p_{t-1}/p'_t)\boldsymbol{\xi}_t \bullet \boldsymbol{w}_{t-1}$, where $\bullet$ is a term-wise product. The investor then decides on a new distribution of his wealth $\boldsymbol{w}_t$, which triggers the following transaction cost:

$$c_s \sum_{i=1}^{m} (p'_t w'^i_t - p_t w^i_t)^+ + c_p \sum_{i=1}^{m} (p_t w^i_t - p'_t w'^i_t)^+ .$$

Denoting the net effect of transaction costs on portfolio value with $\nu_t := p_t/p'_t$, as reported in Li et al. (2018) one finds that ν_t is the solution of the following equations:

$$\nu_t = f(\nu_t, \boldsymbol{w}'_t, \boldsymbol{w}_t) := 1 - c_s \sum_{i=1}^{m} (w'^i_t - \nu_t w^i_t)^+ - c_p \sum_{i=1}^{m} (\nu_t w^i_t - w'^i_t)^+.$$

This, in turn, allows us to express the portfolio's log return during the $t+1$-th period as:

$$\zeta_{t+1} := \ln(p'_{t+1}/p'_t) = \ln(\nu_t p'_{t+1}/p_t) = \ln(\nu_t(\boldsymbol{w}'_t, \boldsymbol{w}_t)) + \ln(\boldsymbol{\xi}^\top_{t+1} \boldsymbol{w}_t) \tag{3.1}$$

where we make explicit the influence of $\boldsymbol{w}'_t$ and $\boldsymbol{w}_t$ on ν_t.

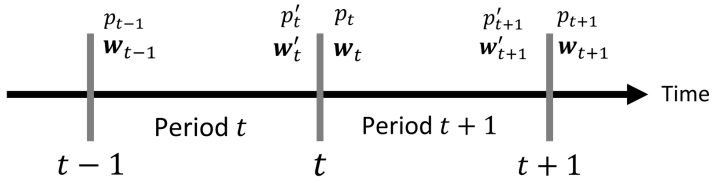


Figure 3.1 – Portfolio evolution through time

We note that in Jiang et al. (2017), the authors suggest to approximate ν_t using an iterative procedure. We actually show in Appendix 3.5.1 that ν_t can easily be identified with high precision using the bisection method.

3.2.2 Risk-averse Reinforcement-Learning Formulation

In this section, we formulate the portfolio management problem as a Markov Decision Process (MDP) denoted by $(\mathcal{S}, \mathcal{A}, r, P)$. In this regard, the agent (i.e. an investor) interacts with a stochastic environment by taking an action $a_t \equiv \boldsymbol{w}_t \in \mathbb{W}$ after observing the state $s_t \in \mathcal{S}$ composed of a window of historical market observations, which include the latest stock returns $\boldsymbol{\xi}_t$, along with the final portfolio composition of the previous period $\boldsymbol{w}'_t$. This action results in the immediate stochastic reward that takes the shape of an approximation of the realized log return, i.e. $r_t(s_t, a_t, s_{t+1}) := \ln(f(1, \boldsymbol{w}'_t, \boldsymbol{w}_t)) + \ln(\boldsymbol{\xi}^\top_{t+1} \boldsymbol{w}_t) \approx \ln(\nu(\boldsymbol{w}'_t, \boldsymbol{w}_t)) + \ln(\boldsymbol{\xi}^\top_{t+1} \boldsymbol{w}_t)$, for which a derivative is easily obtained. Finally, P captures the assumed Markovian transition dynamics of the stock market and its effect on portfolio weights: $P(s_{t+1}|s_0, a_0, s_1, a_1, ..., s_t, a_t) = P(s_{t+1}|s_t, a_t)$.

Following the works of Moody et al. (1998) and Almahdi and Yang (2017) on risk-averse DRL, our objective is to identify a deterministic trading policy μ_θ (parameterized by θ), i.e.

$\mu_\theta : \mathcal{S} \to \mathcal{A}$, that maximizes the expected value of the Sharpe ratio measured on T-periods log return trajectories generated by μ_θ. Namely:

$$\max_\theta \; J_F(\mu_\theta) := \mathbb{E}_{\substack{s_0 \sim F \\ s_{t+1} \sim P(\cdot|s_t, \mu_\theta(s_t))}} [SR(r_0(s_0, \mu_\theta(s_0), s_1), ..., r_{T-1}(s_{T-1}, \mu_\theta(s_{T-1}), s_T))] \quad (3.2)$$

where F is some fixed distribution and

$$SR(r_{0:T-1}) := \frac{(1/T)\sum_{t=0}^{T-1} r_t}{\sqrt{(1/(T-1))\sum_{t=0}^{T-1}(r_t - (1/T)\sum_{t=0}^{T-1} r_t)^2}} \,.$$

The choice of using the Sharpe ratio of log returns is motivated by modern portfolio theory (see Markowitz (1952)), which advocates a balance between expected returns and exposure to risks, and where it plays the role of a canonical way of exercising this trade-off (Sharpe, 1966). While it does not distinguish downside from upside risk, it is still considered a "gold standard of performance evaluation" by the financial community (Bailey and Lopez de Prado, 2012). In Moody et al. (1998), the trajectory-wise Sharpe ratio is used as an estimator of the instantaneous one in order to facilitate its use in RL. A side-benefit of this estimator is to offer some control on the variations in the evolution of the portfolio value which can be reassuring for the investor.

In the context of our portfolio management problem, since s_t is composed of an exogeneous component s_t^{exo} which includes $\boldsymbol{\xi}_t$ and an endogenous state $\boldsymbol{w}'_t$ that becomes deterministic when a_t and s_{t+1}^{exo} are known, we have that:

$$J_F(\mu_\theta) := \mathbb{E}_{\substack{s_0 \sim F \\ s_{t+1} \sim P(\cdot|s_t, \beta(s_t)))}} [SR(r_0(\bar{s}_0^\theta, \mu_\theta(\bar{s}_0^\theta), \bar{s}_1^\theta), \ldots, r_{T-1}(\bar{s}_{T-1}^\theta, \mu_\theta(\bar{s}_{T-1}^\theta), \bar{s}_T^\theta))]$$

where $\beta(s_t)$ is an arbitrary policy[1], and where the effect of μ_θ on the trajectory can be calculated using

$$\bar{s}_t^\theta := \left(s_t^{exo}, \frac{\boldsymbol{\xi}_t \bullet \mu_\theta(\bar{s}_{t-1}^\theta)}{\boldsymbol{\xi}_t^\top \mu_\theta(\bar{s}_{t-1}^\theta)} \right) ,$$

for $t \geq 1$, while $\bar{s}_0^\theta := s_0$. Hence,

$$\nabla_\theta J_F(\mu_\theta) := \mathbb{E}[\nabla_\theta SR(r_0(\bar{s}_0^\theta, \mu_\theta(\bar{s}_0^\theta), \bar{s}_1^\theta), \ldots, r_{T-1}(\bar{s}_{T-1}^\theta, \mu_\theta(\bar{s}_{T-1}^\theta), \bar{s}_T^\theta))] \,, \quad (3.3)$$

where $\nabla_\theta SR$ can be obtained by backpropagation using the chain rule. This leads to the following stochastic gradient step:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta SR(r_0(\bar{s}_0^\theta, \mu_\theta(\bar{s}_0^\theta), \bar{s}_1^\theta), \ldots, r_{T-1}(\bar{s}_{T-1}^\theta, \mu_\theta(\bar{s}_{T-1}^\theta), \bar{s}_T^\theta) \,,$$

[1]This model is insensitive to the choice of the β-policy given that the effect of the actions on the state is known and will be exactly recomputed for μ_θ.

with $\alpha > 0$ as the step size.

3.3 The New Permutation Invariant WaveCorr Architecture

There are several considerations that go into the design of the network for the portfolio policy network μ_θ. First, the network should have the capacity to handle long historical time series data, which allows for extracting long-term dependencies across time. Second, the network should be flexible in its design for capturing dependencies across a large number of available assets. Third, the network should be parsimoniously parameterized to achieve these objectives without being prone to overfitting. To this end, the WaveNet structure (Oord et al., 2016) offers a good basis for developing our architecture and was employed in Zhang et al. (2020). Unfortunately, a direct application of WaveNet in portfolio management struggles at processing the cross-asset correlation information. This is because the convolutions embedded in the WaveNet model are 1D and extending to 2D convolutions increases the number of parameters in the model, which makes it more prone to the issue of over-fitting, a notorious issue particuarly in RL. Most importantly, naive attempts at adapting WaveNet to account for such dependencies (as done in Zhang et al. (2020)) can make the network become sensitive to the ordering of the assets in the input data, an issue that we will revisit below.

We first present the general architecture of WaveCorr in Figure 3.2. Here, the network

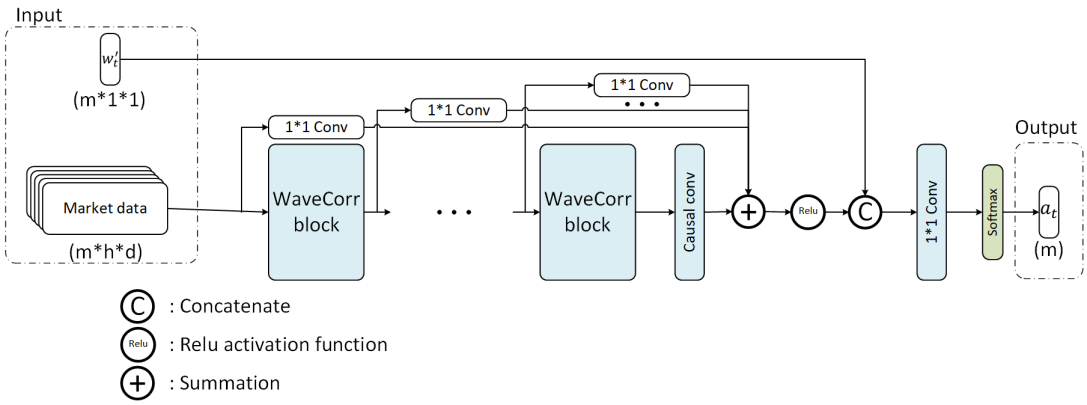


Figure 3.2 – The architecture of the WaveCorr model

takes as input a tensor of dimension $m \times h \times d$, where m : the number of assets, h : the size

of look-back time window, d : the number of channels (number of features for each asset), and generates as output an m-dimensional wealth allocation vector. The WaveCorr blocks, which play the key role for extracting cross time/asset dependencies, form the body of the architecture. In order to provide more flexibility for the choice of h, we define a causal convolution after the sequence of WaveCorr blocks to adjust the receptive field so that it includes the whole length of the input time series. Also, similar to the WaveNet structure, we use skip connections in our architecture.

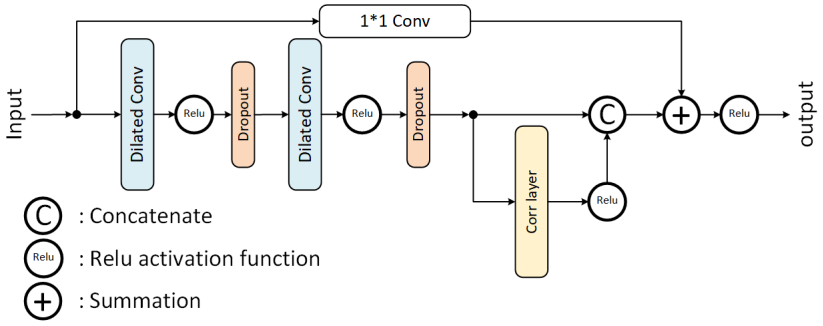


Figure 3.3 – WaveCorr residual block

The design of the WaveCorr residual block in WaveCorr extends a simplified variation (Bai et al., 2018) of the residual block in WaveNet by adding our new correlation layers (and Relu, concatenation operations following right after). As shown in Figure 3.3, the block includes two layers of dilated causal convolutions followed by Relu activation functions and dropout layers. Having an input of dimensions $m \times h \times d$, the convolutions output tensors of dimension $m \times h \times d'$ where each slice of the output tensor, i.e. an $m \times 1 \times d$ matrix, contains the dependency information of each asset over time. By applying different dilation rates in each WaveCorr block, the model is able of extracting the dependency information for a longer time horizon. A dropout layer with a rate of 50% is considered to prevent over-fitting, whereas for the gradient explosion/vanishing prevention mechanism of residual connection we use a 1×1 convolution (presented on the top of Figure 3.3), which inherently ensures that the summation operation is over tensors of the same shape. The *Corr* layer generates an output tensor of dimensions $m \times h \times 1$ from an $m \times h \times d$ input, where each slice of the output tensor, i.e. an $m \times 1 \times 1$ matrix, is meant to contain cross-asset dependency information. The concatenation operator combines the cross-asset

dependency information obtained from the *Corr* layer with the cross-time dependency information obtained from the causal convolutions.

Before defining the *Corr* layer , we pause to introduce a property that will be used to further guide its design, namely the property of asset-permutation invariance. This property is motivated by the idea that the set of possible investment policies that can be modeled by the portfolio policy network should not be affected by the way the assets are indexed in the problem. On a block per block level, we will therefore impose that, when the asset indexing of the input tensor is reordered, the set of possible mappings obtained should also only differ in its asset indexing. More specifically, we let $\sigma : \mathbb{R}^{m\times h\times d} \to \mathbb{R}^{m\times h\times d}$ denote a permutation operator over a tensor $\mathcal{T}$ such that $\sigma(\mathcal{T})[i,:,:] = \mathcal{T}[\pi(i),:,:]$, where $\pi : \{1,...,m\} \to \{1,...,m\}$ is a bijective function. Furthermore, we consider $\sigma^{-1} : \mathbb{R}^{m\times h\times d'} \to \mathbb{R}^{m\times h\times d'}$ denote its "inverse" such that $\sigma^{-1}(\mathcal{O})[i,:,:] := \mathcal{O}[\pi^{-1}(i),:,:]$, with $\mathcal{O} \in \mathbb{R}^{m\times h\times d'}$.

Definition 14. (Asset-Permutation Invariance) A block capturing a set of functions $\mathcal{B} \subseteq \{B : \mathbb{R}^{m\times h\times d} \to \mathbb{R}^{m\times h'\times d'}\}$ is **asset-permutation invariant** if given any permutation operator σ, we have that $\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} = \mathcal{B}$, where $\circ$ stands for function composition.

One can verify, for instances, that all the blocks described so far in WaveCorr are permutation invariant and that asset-permutation invariance is preserved under composition (see Appendix 3.5.2).

With this property in mind, we can now detail the design of a permutation invariant *Corr* layer via Algorithm2, where we denote as $CC : \mathbb{R}^{(m+1)\times h\times d} \to \mathbb{R}^{1\times h\times 1}$ the operator that applies an $(m+1)\times 1$ convolution, and as $Concat_1$ the operator that concatenates two tensors along the first dimension. In Algorithm 2, the kernel is applied to a tensor $\mathcal{O}_{mdl} \in \mathbb{R}^{(m+1)\times h\times d}$ constructed from adding the i-th row of the input tensor on the top of the input tensor. Concatenating the output tensors from each run gives the final output tensor. Figure 3.4 gives an example for the case with $m = 5$, and $h = d = 1$. Effectively, one can show that *Corr* layer satisfies asset-permutation invariance (proof in Appendix).

Proposition 3.3.1. *The Corr layer block satisfies asset-permutation invariance.*

Table 3.1 summarizes the details of each layer involved in the WaveCorr architecture: including kernel sizes, internal numbers of channels, dilation rates, and types of activation

Algorithm 2 : *Corr* layer

Result : Tensor that contains correlation information, $\mathcal{O}_{out}$ of dimension $m \times h \times 1$
Inputs: Tensor $\mathcal{O}_{in}$ of dimension $m \times h \times d$;
Define an empty tensor $\mathcal{O}_{out}$ of dimension $0 \times h \times 1$;
for $i = 1 : m$ **do**
  Set $\mathcal{O}_{mdl} = Concat_1(\mathcal{O}_{in}[i, :, :], \mathcal{O}_{in})$;
  Set $\mathcal{O}_{out} = Concat_1(\mathcal{O}_{out}, CC(\mathcal{O}_{mdl}))$;
end

functions. Overall, the following proposition confirms that this WaveCorr portfolio policy network satisfies asset-permutation invariance (see Appendix for proof).

Proposition 3.3.2. *The WaveCorr portfolio policy network architecture satisfies asset-permutation invariance.*

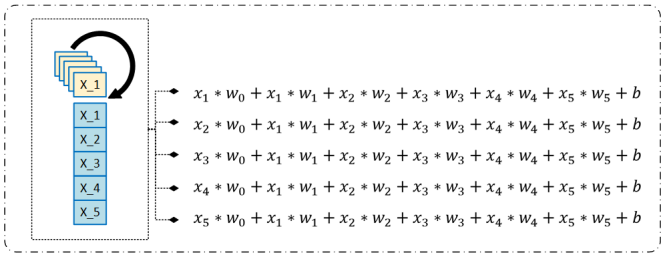


Figure 3.4 – An example of the *Corr* layer over 5 assets

Table 3.1 – The structure of the network

Layer	Input shape	Output shape	Kernel	Activation	Dilation rate
Dilated conv	$(m \times h \times d)$	$(m \times h \times 8)$	(1×3)	Relu	1
Dilated conv	$(m \times h \times 8)$	$(m \times h \times 8)$	(1×3)	Relu	1
Corr layer	$(m \times h \times 8)$	$(m \times h \times 1)$	$([m+1] \times 1)$	Relu	-
Dilated conv	$(m \times h \times 9)$	$(m \times h \times 16)$	(1×3)	Relu	2
Dilated conv	$(m \times h \times 16)$	$(m \times h \times 16)$	(1×3)	Relu	2
Corr layer	$(m \times h \times 16)$	$(m \times h \times 1)$	$([m+1] \times 1)$	Relu	-
Dilated conv	$(m \times h \times 17)$	$(m \times h \times 16)$	(1×3)	Relu	4
Dilated conv	$(m \times h \times 16)$	$(m \times h \times 16)$	(1×3)	Relu	4
Corr layer	$(m \times h \times 16)$	$(m \times h \times 1)$	$([m+1] \times 1)$	Relu	-
Causal conv	$(m \times h \times 17)$	$(m \times h \times 16)$	$(1 \times [h-28])$	Relu	-
1×1 conv	$(m \times h \times 16)$	$(m \times h \times 1)$	(1×1)	Softmax	-

Finally, it is necessary to discuss some connections with the recent work of Zhang et al. (2020), where the authors propose an architecture that also takes both sequential and cross-asset dependency into consideration. Their proposed architecture, from a high level

perspective, is more complex than ours in that theirs involves two sub-networks, one LSTM and one CNN, whereas ours is built solely on CNN. Our architecture is thus simpler to implement, less susceptible to overfitting, and allows for more efficient computation. The most noticeable difference between their design and ours is at the level of the *Corr* layer block, where they use a convolution with a $m \times 1$ kernel to extract dependency across assets and apply a standard padding trick to keep the output tensor invariant in size. Their approach suffers from two issues (see Appendix 3.5.3 for details): first, the kernel in their design may capture only partial dependency information, and second, most problematically, their design is not asset-permutation invariant and thus the performance of their network can be highly sensitive to the ordering of assets. This second issue is further confirmed empirically in section 3.4.3.

3.4 Experimental results

In this section, we present the results of a series of experiments evaluating the empirical performance of our WaveCorr DRL approach. We start by presenting the experimental set-up. We follow with our main study that evaluates WaveCorr against a number of popular benchmarks. We finally shed light on the superior performance of WaveCorr with comparative studies that evaluate the sensitivity of its performance to permutation of the assets, number of assets, size of transaction costs, and (in Appendix 3.5.6) maximum holding constraints. All code is available at **(hyperlink hidden)**.

3.4.1 Experimental set-up

Data sets: We employ three data sets. **Can-data** includes the daily closing prices of 50 Canadian assets from 01/01/2003 until 01/11/2019 randomly chosen among the 70 companies that were continuously part of the Canadian S&P/TSX Composite Index during this period. **US-data** contains 50 randomly picked US assets among the 250 that were part of S&P500 index during the same period.[2] Finally, **Covid-data** considered 50 randomly resampled assets from S&P/TSX Composite Index for period 01/11/2011-01/01/2021 and

[2]Note that by choosing stocks that are continuously part of the index for a long period of time introduces a survivorship bias in our results, however, this bias is affecting both our model and our benchmarks the same way, therefore, relative results are still meaningful.

included open, highest, lowest, and closing daily prices. The Can-data and US-data sets were partitioned into training, validation, and test sets according to the periods 2003-2009/2010-2013/2014-2019 and 2003-2009/2010-2012/2013-2019 respectively, while the Covid-data was only divided in a training (2012-2018) and testing (2019-2020) periods given that hyper-parameters were reused from the previous two studies. We assume with all datasets a constant comission rate of $c_s = c_p = 0.05\%$ in the comparative study, while the sensitivity analysis considers no transaction costs unless specified otherwise.

Benchmarks: In our main study, we compare the performance of WaveCorr to CS-PPN (Zhang et al., 2020), EIIE (Jiang et al., 2017), and the equal weighted portfolio (EW). Note that both CS-PPN and EIIE were adapted to optimize the Sharpe-ratio objective described in section 3.2.2 that exactly accounts for transaction costs.

Hyper-parameter selection: Based on a preliminary unreported investigation, where we explored the influence of different optimizers (namely ADAM, SGD, RMSProp, and SGD with momentum), we concluded that ADAM had the fastest convergence. We also narrowed down a list of reasonable values (see Table 3.6) for the following common hyper-parameters: initial learning rate, decay rate, minimum rate, look-back window size h, planning horizon T. For each method, the final choice of hyper-parameter settings was done based on the average annual return achieved on both a 4-fold cross-validation study using Can-data and a 3-fold study with the US-data. The final selection (see Table 3.6) favored, for each method, a candidate that appeared in the top 5 best performing settings of both data-sets in order to encourage generalization power among similarly performing candidates. Note that in order to decide on the number of epochs, an early stopping criteria was systematically employed.

Metrics: We evaluate all approaches using out-of-sample data ("test data"). "Annual return" denotes the annualized rate of return for the accumulated portfolio value. "Annual vol" denotes the prorated standard deviation of daily returns. Trajectory-wise Sharpe ratio (SR) of the log returns, Maximum drawdown (MDD), i.e. biggest loss from a peak, and average Turnover, i.e. average of the trading volume, are also reported (see (Zhang

et al., 2020) for formal definitions). Finaly, we report on the average "daily hit rate" which captures the proportion of days during which the log returns out-performed EW.

Important implementation details: Exploiting the fact that our SGD step involves exercising the portfolio policy network for T consecutive steps (see equation (3.3)), a clever implementation was able to reduce WavCorr's training time per episode by a factor of 4. This was done by replacing the T copies of the portfolio policy network producing a_0, a_2, ..., a_{T-1}, with an equivalent single augmented multi-period portfolio policy network producing all of these actions simultaneously, while making sure that all intermediate calculations are reused as much as possible (see Appendix 3.5.4 for details). We also implement our stochastic gradient descent approach by updating, after each episode k, the initial state distribution F to reflect the latest policy μ_{θ_k}. This is done in order for the final policy to be better adapted to the conditions encountered when the portfolio policy network is applied on a longer horizon than T.

3.4.2 Comparative Evaluation of WaveCorr

In this set of experiments the performances of WaveCorr, CS-PPN, EIIE, and EW are compared for a set of 10 experiments (with random reinitialization of NN parameters) on the three datasets. The average and standard deviations of each performance metric are presented in Table 3.2 while Figure 3.8 (in the Appendix) presents the average out-of-sample portfolio value trajectories. The main takeaway from the table is that WaveCorr significantly outperforms the three benchmarks on all data sets, achieving an absolute improvement in average yearly returns of 3% to 25% compared to the best alternative. It also dominates CS-PPN and EIIE in terms of Sharpe ratio, maximum drawdown, daily hit rate, and turnover. EW does appear to be causing less volatility in the US-data, which leads to a slightly improved SR. Another important observation consists in the variance of these metrics over the 10 experiments. Once again WaveCorr comes out as being generally more reliable than the two other DRL benchmarks in the Can-data, while EIIE appears to be more reliable in the US-data sacrificing average performance. Overall, the impressive performance of WaveCorr seems to support our claim that our new architecture allows for a better identification of the cross-asset dependencies. In conditions of market crisis (i.e.

the Covid-data), we finally observe that WaveCorr exposes the investors to much lower short term losses, with an MDD of only 31% compared to more than twice as much for CS-PPN and EIIE, which reflects of a more effective hedging strategy.

Table 3.2 – The average (and standard deviation) performances using three data sets.

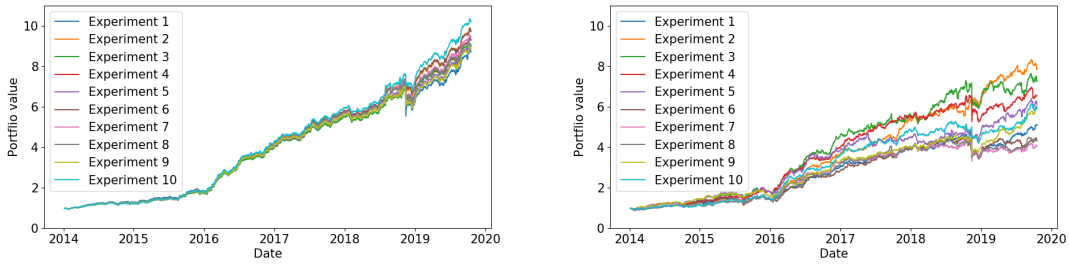
Method	Annual return	Annual vol	SR	MDD	Daily hit rate	Turnover
			Can-data			
WaveCorr	27% (3%)	16% (1%)	1.73 (0.25)	16% (2%)	52% (1%)	0.32 (0.01)
CS-PPN	21% (4%)	19% (2%)	1.14 (0.34)	17% (4%)	51% (1%)	0.38 (0.05)
EIIE	-1% (8%)	29% (4%)	-0.01 (0.28)	55% (9%)	47% (1%)	0.64 (0.08)
EW	4% (0%)	14% (0%)	0.31 (0.00)	36% (0%)	-	0.00 (0.00)
			US-data			
WaveCorr	19% (2%)	16% (2%)	1.17 (0.20)	20% (4%)	50% (1%)	0.11 (0.02)
CS-PPN	14% (2%)	15% (2%)	0.94 (0.17)	22% (6%)	49% (1%)	0.15 (0.08)
EIIE	16% (1%)	15% (0%)	1.09 (0.06)	20% (1%)	50% (0%)	0.17 (0.02)
EW	15% (0%)	13% (0%)	1.18 (0.00)	18% (0%)	-	0.00 (0.00)
			Covid-data			
WaveCorr	56% (13%)	26% (5%)	2.16 (0.50)	31% (9%)	51% (2%)	0.19 (0.05)
CS-PPN	31% (27%)	51% (6%)	0.60 (0.48)	67% (7%)	50% (2%)	0.3 (0.09)
EIIE	11% (30%)	76% (17%)	0.20 (0.43)	77% (13%)	46% (2%)	0.76 (0.27)
EW	27% (0%)	29% (0%)	0.93(0.00)	47% (0%)	-	0.01 (0.00)

3.4.3 Sensitivity Analysis

Sensitivity to permutation of the assets: In this set of experiment, we are interested in measuring the effect of asset-permutation on the performance of WaveCorr and CS-PPN. Specifically, each experiment now consists in resampling a permutation of the 50 stocks instead of the initial parameters of the neural networks. The results are summarized in Table 3.3 and illustrated in Figure 3.5. We observe that the learning curves and performance of CS-PPN are significantly affected by asset-permutation compared to WaveCorr. In particular, one sees that the standard deviation of annual return is reduced by more than a factor of about 5 with WaveCorr. We believe this is entirely attributable to the new structure of the *Corr* layer in the portfolio policy network.

Table 3.3 – The average (and standard dev.) performances over random asset-permutation in Can-data.

	Annual return	Annual vol	SR	MDD	Daily hit rate	Turnover
WaveCorr	48% (1%)	15% (1%)	3.15 (0.19)	14% (3%)	56% (0%)	0.48 (0.01)
CS-PPN	35% (5%)	18% (1%)	2.00 (0.37)	22% (4%)	54% (1%)	0.54 (0.03)



(a) WaveCorr (b) CS-PPN

Figure 3.5 – Comparison of the wealth accumulated by WaveCorr and CS-PPN under random initial permutation of assets on Can-data's test set.

Sensitivity to number of assets: In this set of experiments, we measure the effect of varying the number of assets on the performance of WaveCorr and CS-PPN. We therefore run 10 experiments (randomly resampling initial NN parameters) with growing subsets of 30, 40, and 50 assets from Can-data. Results are summarized in Table 3.4 and illustrated in Figure 3.9 (in Appendix). While having access to more assets should in theory be beneficial for the portfolio performance, we observe that it is not necessarily the case for CS-PPN. On the other hand, as the number of assets increase, a significant improvement, with respect to all metrics, is achieved by WaveCorr. This evidence points to a better use of the correlation information in the data by WaveCorr.

Table 3.4 – The average (and std. dev.) performances as a function of the number of assets in Can-data.

# of stocks	Annual return	Annual vol	SR	MDD	Daily hit rate	Turnover
			WaveCorr			
30	37.7% (4%)	19% (1%)	2.02 (0.27)	22% (3%)	55% (1%)	0.39 (0.02)
40	38.5% (4%)	21% (1%)	1.81 (0.17)	23% (2%)	55% (1%)	0.44 (0.04)
50	43.0% (5%)	17% (2%)	2.57 (0.52)	20% (6%)	55% (1%)	0.43 (0.02)
			CS-PPN			
30	30.3% (3%)	17% (1%)	1.80 (0.20)	21% (4%)	53% (1%)	0.42 (0.04)
40	29.8% (7%)	17% (2%)	1.70 (0.34)	22% (3%)	53% (1%)	0.41 (0.09)
50	32.2% (4%)	16% (1%)	2.07 (0.28)	18% (3%)	52% (1%)	0.43 (0.05)

Sensitivity to commission rate: Table 3.5 presents how the performances of WaveCorr and CS-PPN are affected by the magnitude of the commission rate, ranging among 0%, 0.05%, and 0.1%. One can first recognize that the two methods appear to have good control on turnover as the commission rate is increased. Nevertheless, one can confirm from

this table the significantly superior performance of WaveCorr prevails under all level of commission rate.

Table 3.5 – The average (and std. dev.) performances as a function of commission rate (CR) in Can-data.

Method	Annual return	Annual vol	SR	MDD	Daily hit rate	Turnover
			CR = 0			
WaveCorr	42% (3%)	15% (0%)	2.77 (0.20)	13% (1%)	55% (1%)	0.44 (0.02)
CS-PPN	35% (4%)	17% (1%)	2.04 (0.27)	14% (3%)	53% (1%)	0.47 (0.05)
			CR = 0.05%			
WaveCorr	27% (3%)	16% (1%)	1.73 (0.25)	15% (2%)	52% (1%)	0.32 (0.01)
CS-PPN	21% (4%)	19% (2%)	1.14 (0.34)	17% (4%)	51% (1%)	0.38 (0.05)
			CR = 0.1%			
WaveCorr	19% (2%)	15% (1%)	1.34 (0.16)	16% (2%)	50% (1%)	0.22 (0.01)
CS-PPN	14% (7%)	17% (3%)	0.92 (0.50)	19% (8%)	50% (1%)	0.22 (0.09)

3.5 Appendix

3.5.1 Solving $\nu = f(\nu)$

In order to apply the bisection method to solve $\nu = f(\nu)$, we will make use of the following proposition.

Proposition 3.5.1. *For any $0 < c_s < 1$ and $0 < c_p < 1$, the function $g(\nu) := \nu - f(\nu)$ is strictly increasing on $[0, 1]$ with $g(0) < 0$ and $g(1) > 0$.*

Proof. Recalling that $f(\nu, \boldsymbol{w}', \boldsymbol{w}) := 1 - c_s \sum_{i=1}^m (w'^i - \nu w^i)^+ - c_p \sum_{i=1}^m (\nu w^i - w'^i)^+$, we first obtain the two bounds at $g(0)$ and $g(1)$ as follows:

$$g(0) = 0 - \left(1 - c_s \sum_{i=1}^m (w'^i)^+ - c_p \sum_{i=1}^m (-w'^i)^+\right) = 0 - 1 + c_s < 0\,,$$

since $c_s < 1$, and

$$g(1) = 1 - \left(1 - c_s \sum_{i=1}^m (w'^i - w^i)^+ - c_p \sum_{i=1}^m (w^i - w'^i)^+\right) \geq \min(c_s, c_p)\|\boldsymbol{w}'^i - \boldsymbol{w}^i\|_1 > 0\,,$$

since $\min(c_s, c_p) > 0$. We can further establish the convexity of $g(\nu)$, given that it is the sum of convex functions. A careful analysis reveals that $g(\nu)$ is supported at 0 by the plane

$$g(\nu) \geq g(0) + \nu \left(1 - c_s + c_p \sum_{i=1}^m \mathbf{1}\{w'^i = 0\} w^i\right)\,,$$

where $\mathbf{1}\{A\}$ is the indicator function that returns 1 if A is true, and 0 otherwise. Hence, by convexity of $g(\nu)$, the fact that this supporting plane is strictly increasing implies that $g(\nu)$ is strictly increasing for all $\nu \geq 0$. □

Given Proposition 3.5.1, we can conclude that a bisection method can be used to find the root of $g(\nu)$, which effectively solves $\nu = f(\nu)$.

3.5.2 Proofs of Section 3.3

We start this section with a lemma that will simplify some of our later derivations.

Lemma 3.5.2. *A block capturing a set of functions $\mathcal{B} \subseteq \{B : \mathbb{R}^{m\times h\times d} \to \mathbb{R}^{m\times h'\times d'}\}$ is asset-permutation invariant if and only if given any permutation operator σ, we have that $\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} \supseteq \mathcal{B}$.*

Proof. The "only if" follows straightforwardly from the fact that equality between two sets implies that each set is a subset of the other.

Regarding the "if" part, we start with the assumption that

$$\forall\sigma, \{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} \supseteq \mathcal{B}\,.$$

Next, we follow with the fact that for all permutation operator σ:

$$\begin{aligned}\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} &\subseteq \{\sigma^{-1}\circ B\circ\sigma : B \in \{\sigma\circ B'\circ\sigma^{-1} : B' \in \mathcal{B}\}\}\\ &= \{\sigma^{-1}\circ\sigma\circ B'\circ\sigma^{-1}\circ\sigma : B' \in \mathcal{B}\} = \mathcal{B}\,,\end{aligned}$$

where we assumed for simplicity of exposition that $h = h'$ and $d = d'$, and exploited the fact that σ^{-1} is also a permutation operator. □

Proof of Proposition 3.3.1

We first clarify that the correlation layer is associated with the following set of functions (see Algorithm 1):

$$\mathcal{B} := \{B_{w,b} : w \in \mathbb{R}^{(m+1)\times d}, b \in \mathbb{R}\}$$

where

$$B_{w,b}(\mathcal{T})[i,:,1] := \left(\mathcal{T}[i,:,:]\bullet(\vec{1}w_0^\top) + \sum_{j=1}^{m}\mathcal{T}[j,:,:]\bullet(\vec{1}w_j^\top)\right)\vec{1} + b,\ \ \forall\, i = 1,...,m,$$

with $\bullet$ denoting the Hadamard (element-wise) product.

Let σ (associated with the bijection π) be an asset-permutation operator. For any correlation layer function $B_{w,b} \in \mathcal{B}$, one can construct a new set of parameters $w'_0 := w_0$, $w'_j := w_{\pi(j)}$, for all $j = 1, \ldots, m$, and $b' := b$ such that for all input tensor $\mathcal{T}$, we have that for all i:

$$B_{w',b'}(\sigma(\mathcal{T}))[i,:,1] = \left(\sigma(\mathcal{T})[i,:,:] \bullet (\vec{1} w_0^\top) + \sum_{j=1}^{m} \sigma(\mathcal{T})[j,:,:] \bullet (\vec{1} w_{\pi(j)}^\top) \right) \vec{1} + b \tag{3.4}$$

$$= \left(\mathcal{T}[\pi(i),:,:] \bullet (\vec{1} w_0^\top) + \sum_{j=1}^{m} \mathcal{T}[\pi(j),:,:] \bullet (\vec{1} w_{\pi(j)}^\top) \right) \vec{1} + b \tag{3.5}$$

$$= \left(\mathcal{T}[\pi(i),:,:] \bullet (\vec{1} w_0^\top) + \sum_{j'=1}^{m} \mathcal{T}[j',:,:] \bullet (\vec{1} w_{j'}^\top) \right) \vec{1} + b. \tag{3.6}$$

Hence,

$$\sigma^{-1}(B_{w',b'}(\sigma(\mathcal{T})))[i,:,1] = \left(\mathcal{T}[i,:,:] \bullet (\vec{1} w_0^\top) + \sum_{j=1}^{m} \mathcal{T}[j,:,:] \bullet (\vec{1} w_j^\top) \right) \vec{1} + b = B_{w,b}(\mathcal{T})[i,:,1]\,.$$

We can therefore conclude that $\{\sigma^{-1} \circ B \circ \sigma : B \in \mathcal{B}\} \supseteq \mathcal{B}$. Based on Lemma 3.5.2, we conclude that $\mathcal{B}$ is asset-permutation invariant.

Proof of Proposition 3.3.2

To prove Proposition 3.3.2, we demonstrate that all blocks used in the WaveCorr architecture are asset-permutation invariant (Steps 1 to 3). We then show that asset-permutation invariance is preserved under composition (Step 4). Finally, we can conclude in Step 5 that WaveCorr is asset-permutation invariant.

Step 1 - Dilated convolution, Causal convolution, Sum, and 1×1 convolution are asset-permutation invariant: The functional class of a dilated convolution, a causal convolution, a sum, and a 1×1 convolution block all have the form:

$$\mathcal{B} := \{B_g : g \in \mathcal{G}\},$$

where

$$B_g(\mathcal{T})[i,:,:] := g(\mathcal{T}(i,:,:)), \;\; \forall\, i = 1, ..., m,$$

for some set of functions $\mathcal{G} \subseteq \{G : \mathbb{R}^{1\times h\times d} \to \mathbb{R}^{1\times h\times d'}\}$. In particular, in the case of dilated, causal, and 1×1 convolutions, this property follows from the use of 1×3, $1 \times [h - 28]$, and 1×1 kernels respectively. Hence, for any $g \in \mathcal{G}$, we have that:

$$\sigma^{-1}(B_g(\sigma(\mathcal{T}))) = \sigma^{-1}(\sigma(B_g(\mathcal{T}))) = B_g(\mathcal{T})\,,$$

which implies that $\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} = \mathcal{B}$.

Step 2 - Relu and dropout are asset-permutation invariant: We first clarify that Relu and dropout on a tensor in $\mathbb{R}^{m\times h\times d}$ are singleton sets of functions:

$$\mathcal{B} := \{B_g\}$$

where $g : \mathbb{R} \to \mathbb{R}$ and $B_g(\mathcal{T})[i, j, k] := g(\mathcal{T}[i, j, k])$. In particular, in the case of Relu, we have:

$$B_g(\mathcal{T})[i, j, k] := \max(0, \mathcal{T}[i, j, k])\,,$$

while, for dropout we have:

$$B_g(\mathcal{T})[i, j, k] := \mathcal{T}[i, j, k]\,,$$

since a dropout block acts as a feed through operator. Hence, we naturally have that:

$$\sigma^{-1}(B_g(\sigma(\mathcal{T}))) = \sigma^{-1}(\sigma(B_g(\mathcal{T}))) = B_g(\mathcal{T})\,,$$

which again implies that $\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} = \mathcal{B}$.

Step 3 - Softmax is asset-permutation invariant: We first clarify that softmax on a vector in $\mathbb{R}^{m\times h\times 1}$ is a singleton set of functions:

$$\mathcal{B} := \{B\}$$

where

$$B(\mathcal{T})[i, j, 1] := \frac{\exp(\mathcal{T}[i, j, 1])}{\sum_{i'=1}^{m} \exp(\mathcal{T}[i', j, 1])}\,.$$

Hence, we have that:

$$B(\sigma(\mathcal{T}))[i, j, 1] := \frac{\exp(\mathcal{T}[\pi(i), j, 1])}{\sum_{i'=1}^{m} \exp(\mathcal{T}[\pi(i'), j, 1])} = \frac{\exp(\mathcal{T}[\pi(i), j, 1])}{\sum_{i'=1}^{m} \exp(\mathcal{T}[i', j, 1])}\,.$$

This allows us to conclude that:

$$\sigma^{-1}(B(\sigma(\mathcal{T})))[i, j, 1] = B(\mathcal{T})[i, j, 1]\,.$$

Hence, we conclude that $\{\sigma^{-1}\circ B\circ\sigma : B \in \mathcal{B}\} = \mathcal{B}$.

Step 4 - Asset-permutation invariance is preserved under composition: Given two asset-permutation invariant blocks representing the set of functions $\mathcal{B}_1$ and $\mathcal{B}_2$, one can define the composition block as:

$$\mathcal{B} := \{B_1 \circ B_2 : B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}\,.$$

We have that for all $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$:

$$\begin{aligned} B &= B_1 \circ B_2 \\ &= (\sigma^{-1} \circ B_1' \circ \sigma) \circ (\sigma^{-1} \circ B_2' \circ \sigma) \\ &= \sigma^{-1} \circ B_1' \circ B_2' \circ \sigma \\ &= \sigma^{-1} \circ B' \circ \sigma\,, \end{aligned}$$

where $B_1' \in \mathcal{B}_1$ and $B_2' \in \mathcal{B}_2$ come from the definition of asset-permutation invariance, and where $B' := B_1' \circ B_2' \in \mathcal{B}$. We therefore have that $\{\sigma^{-1} \circ B \circ \sigma : B \in \mathcal{B}\} \supseteq \mathcal{B}$. Finally, Lemma 3.5.2 allows us to conclude that $\mathcal{B}$ is asset-permutation invariant.

Step 5 - WaveCorr is asset-permutation invariant: Combing Step 1 to 4 with Proposition 3.3.1, we arrive at the conclusion that the architecture presented in Figure 3.2 is asset-permutation invariant since it is composed of a sequence of asset-permutation invariant blocks. □

3.5.3 Correlation Layer in Zhang et al. (2020) Violates Asset-Permutation Invariance

Assuming for simplicity that m is odd, the "correlational convolution layer" proposed in Zhang et al. (2020) takes the form of the following set of functions:

$$\mathcal{B} := \{B_{w,b} : \mathcal{W} \in \mathbb{R}^{m \times d \times d}, b \in \mathbb{R}\}$$

where

$$B_{w,b}(\mathcal{T})[i,j,k] := \sum_{\ell=1}^{m} \sum_{k'=1}^{d} \mathcal{T}[i - (m+1)/2 + \ell, j, k'] \mathcal{W}[\ell, k, k'] + b, \quad \begin{matrix} \forall\, i = 1, \ldots, m \\ \forall j = 1, \ldots, h \\ \forall k = 1, \ldots, d \end{matrix},$$

where $\mathcal{T}[i', :, :] := 0$ for all $i' \notin \{1, \ldots, m\}$ to represent a zero padding. Figure 3.6 presents an example of this layer when $m = 5$, $h = 1$, and $d = 1$. One can already observe in this figure that correlation information is only partially extracted for some of the assets, e.g. the convolution associated to asset one (cf. first row in the figure) disregards the influence of the fifth asset. While this could perhaps be addressed by using a larger kernel, a more important issue arises with this architecture, namely that the block does not satisfy asset-permutation invariance.

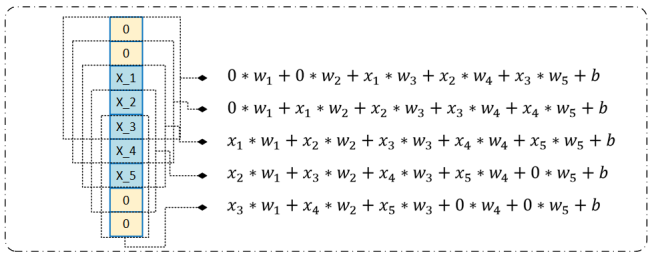


Figure 3.6 – An example of the correlation layer in Zhang et al. (2020)'s work over 5 assets

Proposition 3.5.3. *The correlational convolution layer block used in Zhang et al. (2020) violates asset-permutation invariance already when $m = 5$, $h = 1$, and $d = 1$.*

Proof. When $m = 5$, $h = 1$, and $d = 1$, we first clarify that the correlational convolution layer from Zhang et al. (2020) is associated with the following set of functions:

$$\mathcal{B} := \{B_{w,b} : w \in \mathbb{R}^5, b \in \mathbb{R}\}$$

where

$$B_{w,b}(\mathcal{T})[i] := \begin{cases} w_3\mathcal{T}[1] + w_4\mathcal{T}[2] + w_5\mathcal{T}[3] + b & \text{if } i = 1 \\ w_2\mathcal{T}[1] + w_3\mathcal{T}[2] + w_4\mathcal{T}[3] + w_5\mathcal{T}[4] + b & \text{if } i = 2 \\ w_1\mathcal{T}[1] + w_2\mathcal{T}[2] + w_3\mathcal{T}[3] + w_4\mathcal{T}[4] + w_5\mathcal{T}[5] + b & \text{if } i = 3 \\ w_1\mathcal{T}[2] + w_2\mathcal{T}[3] + w_3\mathcal{T}[4] + w_4\mathcal{T}[5] + b & \text{if } i = 4 \\ w_1\mathcal{T}[3] + w_2\mathcal{T}[4] + w_3\mathcal{T}[5] + b & \text{if } i = 5 \end{cases},$$

where we shortened the notation $\mathcal{T}[i, 1, 1]$ to $\mathcal{T}[i]$. Let's consider the asset-permutation operator that inverts the order of the first two assets: $\pi(1) = 2$, $\pi(2) = 1$, and $\pi(i) = i$ for all $i \geq 3$. We will prove our claim by contradiction. Assuming that $\mathcal{B}$ is asset-permutation invariant, it must be that for any fixed values $\bar{w}$ such that $\bar{w}_4 \neq \bar{w}_1$, there exists an associated

pair of values (w', b') that makes $B_{w',b'} \equiv \sigma^{-1} \circ B_{\bar{w},0} \circ \sigma$. In particular, the two functions should return the same values for the following three "tensors": $\mathcal{T}_0[i] := 0$, $\mathcal{T}_1[i] := \mathbf{1}\{i = 1\}$, and at $\mathcal{T}_2[i] := \mathbf{1}\{i = 2\}$. The first implies that $b' = 0$ since

$$b' = B_{w',b'}(\mathcal{T}_0)[1] = \sigma^{-1}(B_{\bar{w},0}(\sigma(\mathcal{T}_0)))[1] = 0\,.$$

However, it also implies that:

$$w'_2 = B_{w',0}(\mathcal{T}_1)[2] = \sigma^{-1}(B_{\bar{w},0}(\sigma(\mathcal{T}_1)))[2] = B_{\bar{w},0}(\mathcal{T}_2)[1] = \bar{w}_4$$

and that

$$w'_2 = B_{w',0}(\mathcal{T}_2)[3] = \sigma^{-1}(B_{\bar{w},0}(\sigma(\mathcal{T}_2)))[3] = B_{\bar{w},0}(\mathcal{T}_1)[3] = \bar{w}_1\,.$$

We therefore have a contradiction since $\bar{w}_4 = w'_2 = \bar{w}_1 \neq \bar{w}_4$ is impossible. We must therefore conclude that $\mathcal{B}$ was not asset-permutation invariant. □

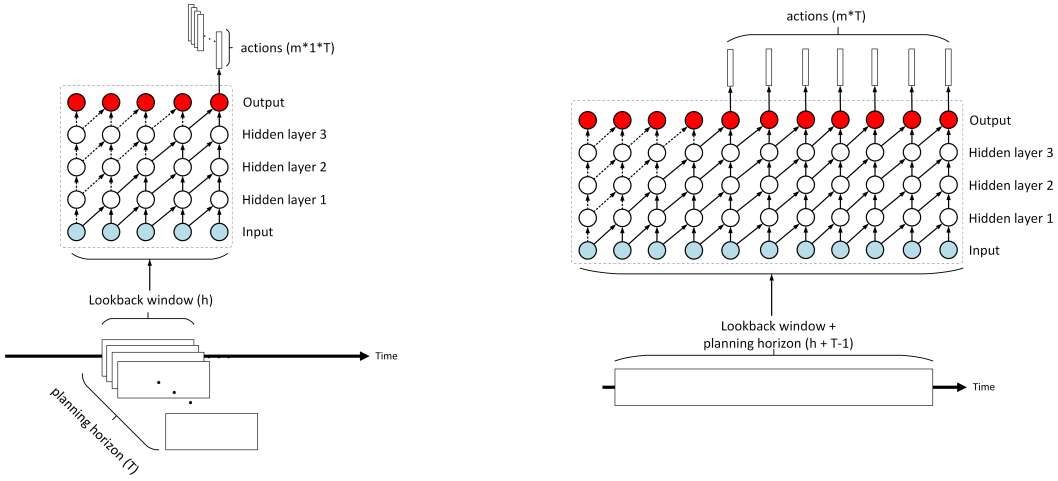
We close this section by noting that this important issue cannot simply be fixed by using a different type of padding, or a larger kernel in the convolution. Regarding the former, our demonstration made no use of how padding is done. For the latter, our proof would still hold given that the fixed parameterization $(\bar{w}, 0)$ that we used would still identify a member of the set of functions obtained with a larger kernel.

3.5.4 Augmented policy network to accelerate training

We detail in this section how the structure of the portfolio management problem (3.2) can be exploited for a more efficient implementation of a policy network, both in terms of computation time and hardware memory. This applies not only to the implementation of WaveCorr policy network but also policy networks in Jiang et al. (2017) and Zhang et al. (2020). In particular, given a a multiperiod objective as in (3.2), calculating the gradient $\nabla_\theta SR$ involves the step of generating a sequence of actions $a_0, a_1, ..., a_{T-1}$ from a sample trajectory of states $s_0, s_1, ..., s_{T-1} \in \mathbb{R}^{m \times h \times d}$ over a planning horizon T, where m : the number of assets, h : the size of a lookback window, d : the number of features. The common way of implementing this is to create a tensor $\mathcal{T}$ of dimension $m \times h \times d \times T$ from $s_0, ..., s_{T-1}$ and apply a policy network $\mu_\theta(s)$ to each state s_t in the tensor $\mathcal{T}$ so as to generate each action a_t. Assuming for simplicity of exposition that the state is entirely exogenous, this

procedure is demonstrated in Figure 3.7(a), where a standard causal convolution with $d = 1$ and kernel size of 2 is applied. In this procedure, the memory used to store the tensor $\mathcal{T}$ and the computation time taken to generate all actions $a_0, ..., a_{T-1}$ grow linearly in T, which become significant for large T. It is possible to apply the policy network $\mu_\theta(s)$ to generate all the actions $a_0, ..., a_{T-1}$ more efficiently than the procedure described in Figure 3.7(a). Namely, in our implementation, we exploit the sequential and overlapping nature of sample states $s_0, ..., s_{T-1}$ used to generate the actions $a_0, ..., a_{T-1}$, which naturally arises in the consideration of a multiperiod objective. Recall firstly that each sample state $s_t \in \mathbb{R}^{m \times h \times d}$, $t \in \{0, ..., T-1\}$, is obtained from a sample trajectory, denoted by $\mathcal{S} \in \mathbb{R}^{m \times (h+T-1) \times d}$, where $s_t = \mathcal{S}[:, t+1 : t+h, :]$, $t = 0, ..., T-1$. Thus, between any s_t and s_{t+1}, the last $h-1$ columns in s_t overlap with the first $h-1$ columns in s_{t+1}. The fact that there is a significant overlap between any two consecutive states s_t, s_{t+1} hints already that processing each state s_{t+1} separately from s_t, as shown in Figure 3.7(a), would invoke a large number of identical calculations in the network as those that were already done in processing s_t, which is wasteful and inefficient. To avoid such an issue, we take an augmented approach to apply the policy network. The idea is to use a sample trajectory $\mathcal{S}$ directly as input to an augmented policy network $\vec{\mu}_\theta : \mathbb{R}^{m \times (h+T-1) \times d} \to \mathbb{R}^{m \times T}$, which reduces to exactly the same architecture as the policy network $\mu_\theta(s_t)$ when generating only the t-th action. Figure 3.7(b) presents this augmented policy network $\vec{\mu}_\theta(S)$ for our example, and how it can be applied to a trajectory $\mathcal{S}$ to generate all actions $a_0, ..., a_{T-1}$ at once. One can observe that the use of an augmented policy network allows the intermediate calculations done for each state s_t (for generating an action a_t) to be reused by the calculations needed for the other states (and generating other actions). With the exact same architecture as the policy network $\mu_\theta(s)$, the augmented policy network $\vec{\mu}_\theta(S)$, which takes a trajectory with width $h + T - 1$ (thus including T many states), would by design generate T output, each corresponds to an action a_t. This not only speeds up the generation of actions $a_0, ..., a_{T-1}$ significantly but also requires far less memory to store the input data, i.e. the use of a tensor with dimension $(m \times (h + T) \times d)$ instead of $m \times h \times d \times T$. The only sacrifice that is made with this approach is regarding the type of features that can be integrated. For instance, we cannot include features that are normalized with respect to the most recent history (as done in Jiang et al. (2017)) given that this breaks the data redundancy between

two consecutive time period. Our numerical results however seemed to indicate that such restrictions did not come at a price in terms of performance.



(a) $\mu_\theta(s)$ applied to each state separately

(b) $\vec{\mu}_\theta(S)$ applied to the full trajectory

Figure 3.7 – Comparison between the use of policy network $\mu_\theta(s)$ and of the augmented policy network $\vec{\mu}_\theta(S)$

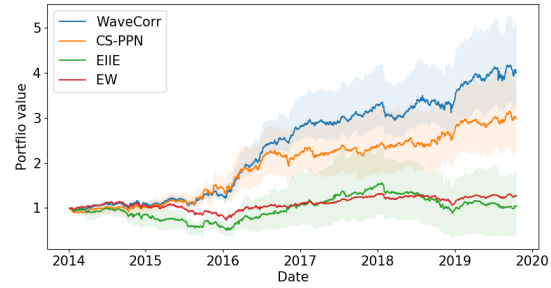
3.5.5 Hyper-parameters Selection

Table 3.6 – List of Selected Hyper-parameters.

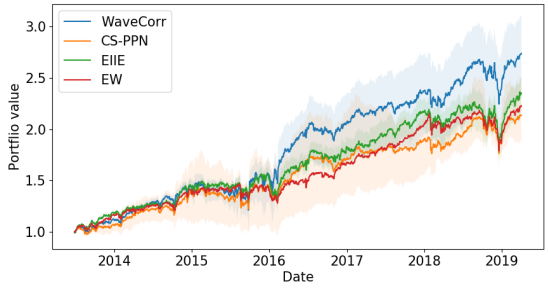
Hyper-parameter	Search range	WaveCorr	CS-PPN	EIIE
Learning rate	$\{5 \times 10^{-5},\ 10^{-4},\ 10^{-3},\ 5 \times 10^{-3}\}$	5×10^{-5}	5×10^{-5}	10^{-4}
Decay rate	{0.9999, 0.99999, 1}	0.99999	0.99999	1
Minimum rate	$\{\ 10^{-6},\ 10^{-5}\}$	10^{-5}	10^{-5}	10^{-5}
Planning horizon T	{32, 64}	32	32	32
Look back window size h	{32, 64}	32	32	32
Number of epochs	$[0,\ \infty)$	5000	5000	5000

3.5.6 Additional results

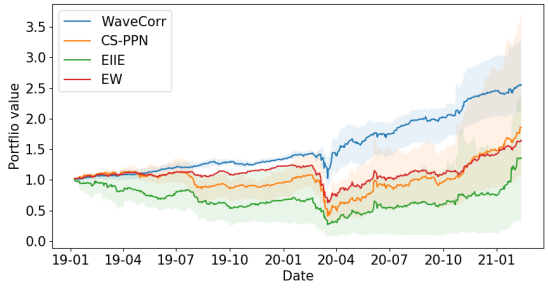
Comparative Study



(a) Can-data



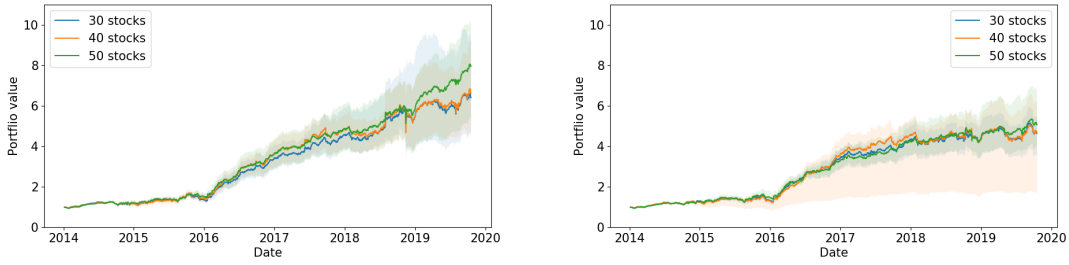
(b) US-data



(c) Covid-data

Figure 3.8 – Average (solid curve) and range (shaded region) of out-of-sample wealth accumulated by WaveCorr, CS-PPN, EIIE, and EW over 10 experiments using Can-data, US-data, and Covid-data.

Sensitivity to number of assets



(a) WaveCorr (b) CS-PPN

Figure 3.9 – Average (solid curve) and range (shaded region) of the out-of-sample wealth accumulated, on 10 experiments using Can-data, by WaveCorr and CS-PPN when increasing the number of assets.

Performance comparison under maximum holding constraint

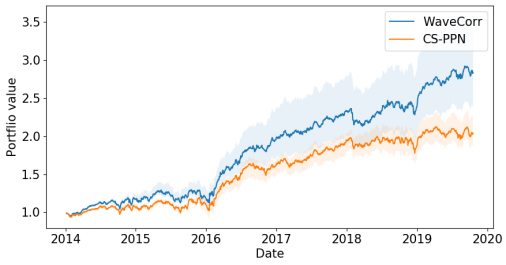
In practice, it is often required that the portfolio limits the amount of wealth invested in a single asset. This can be integrated to the risk-averse DRL formulation:

$$\bar{J}_F(\mu_\theta) := \mathbb{E}_{\substack{s_0 \sim F \\ s_{t+1} \sim P(\cdot|s_t, \mu_\theta(s_t))}} [SR(r_0(s_0, \mu_\theta(s_0), s_1), ...)] - \frac{M}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{m} \max(0,\ w_t^i - w_{max})$$

where w_{max} is the maximum weight allowed in any asset, and M is a large constant. This new objective function penalizes any allocation that goes beyond w_{max}, which will encourage μ_θ to respects the maximum weight allocation condition. The commission rates are considered to be $c_s = c_p = 0.5\%$, and the experiments here are done over Can-data using the full set of 70 stocks, with a maximum holding of 20%. The results are summarized in Table 3.7 and illustrated in Figure 3.10. As noted before, we observe that WaveCorr outperforms CS-PPN with respect to all performance metrics.

Table 3.7 – The average (and standard dev.) performances when imposing a maximum holding constraints over 10 random initial NN weights in Can-data.

	Annual return	Annual vol	SR	MDD	Daily hit rate	Turnover
WaveCorr	20% (2%)	13% (0%)	1.55 (0.18)	14% (1%)	53% (1%)	0.17 (0.01)
CS-PPN	13% (1%)	13% (1%)	1.00 (0.15)	15% (2%)	50% (1%)	0.22 (0.03)



(a) Out-of-sample cumulative returns

Figure 3.10 – Average (solid curve) and range (shaded region) of the out-of-sample wealth accumulated, on 10 experiments using Can-data, by WaveCorr and CS-PPN under maximum holding constraint.

References

Almahdi, S. and Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, 87:267–279.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bailey, D. H. and Lopez de Prado, M. (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, 15(2):13.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jiang, Z., Xu, D., and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

Li, B., Wang, J., Huang, D., and Hoi, S. C. H. (2018). Transaction cost optimization for online portfolio selection. *Quantitative Finance*, 18(8):1411–1424.

Liang, Z., Chen, H., Zhu, J., Jiang, K., and Li, Y. (2018). Adversarial deep reinforcement learning in portfolio management. *arXiv preprint arXiv:1808.09940*.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Moody, J., Wu, L., Liao, Y., and Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5-6):441–470.

Moody, J. E. and Saffell, M. (1999). Reinforcement learning for trading. *Advances in Neural Information Processing Systems*, pages 917–923.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, 39(1):119–138.

Xu, K., Zhang, Y., Ye, D., Zhao, P., and Tan, M. (2020). Relation-aware transformer for portfolio policy learning. IJCAI.

Zhang, Y., Zhao, P., Li, B., Wu, Q., Huang, J., and Tan, M. (2020). Cost-sensitive portfolio selection via deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*.

General Conclusion

In this thesis, we investigated two problems whose solutions are the main component for many other problems in finance, namely option pricing and portfolio management. Chapter 1 extends the theoretical ground for a new option pricing scheme called equal risk pricing (ERP), where the determining factor in pricing an option is the risk exposure of the two sides in the contract. We studied this framework in an incomplete market where a complete hedging of all the risk is impossible. We derived the risk-averse dynamic programming equations required to solve two hedging problems for each side of the contract, where the risk measures chosen by each side are convex. The equations were derived for both American and European options, where by exploiting the translation invariance property of the risk measure, we were able to reduce the dimensionality of the state space in the dynamic programming model to be able to solve it by a grid based method in tractable time. We further used worst-case risk measures to test the performance of this new pricing and hedging scheme using synthetic data compared to ϵ-arbitrage and Black–Scholes models. The numerical results confirmed that under the ERP both the writer and the buyer have risks that are more similar and on average smaller than the risks that they would experience by the two other methods.

In the second chapter we extended the ERP framework by exploiting Deep Reinforcement-Learning (DRL). We differentiate our work as being the first to study ERP under a time-consistent dynamic risk measure. In particular, we exploited the elicitability property of the expectile risk measure to extend the famous off-policy deterministic actor-critic method to the risk-averse setting. In our numerical results, we illustrated how our model can benefit from time-consistency of the optimal solution in order to price a range of option contracts with different maturities by training the model only on the option with longest

time to maturity. This is in contrast to time inconsistent models where the optimal policy is potentially lost if risk is evaluated at any later time, hence we need to retrain a different model to price options with different maturities.

As the potential limitation of our works on ERP, we only evaluated the performance of our model in a synthetic environment using a simple neural network architecture. One may be interested in examining the performance of this model under real market conditions, as training a model in a synthetic environment is significantly simpler than a real market that often exhibits non-stationary behaviors. In addition, we only considered European style options in our implementation of DRL. Extending our DRL approach to solve the case of American options is an interesting direction that has not been well explored in our works. We leave these as the future work that can reveal more interesting results regarding our framework.

In our last chapter we developed a DRL model for time series analysis in a portfolio management problem using real market data. In particular, we built on the literature on DRL models suggested for the task of portfolio management and introduced a new architecture that can effectively take the cross asset correlations into account and come up with high performance portfolio policies. We called this new architecture WaveCorr. As an important contribution, in the last chapter we introduced an essential property that is called asset-permutation invariance. We numerically demonstrated the importance of satisfying this property by running experiments on real market data that provided highly stable portfolio policies compared to our benchmark model that does not satisfy this property. Also, by testing the model on a diverse set of data sets, including one that spans through the COVID-19 pandemic, we showed that our framework is able to outperform other state-of-the-art DRL based portfolio selection models.

Although the numerical experiments in the last chapter is performed for a portfolio management problem, the model can be easily embedded in our ACRL framework that is introduced in the second chapter. In fact, we believe the results of this thesis can be complemented by further studying WaveCorr under ACRL setting where we can benefit from the exploration ability of ACRL for both option pricing and portfolio management applications.

References

Almahdi, S. and Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, 87:267–279.

Amin, K. I. (1993). Jump diffusion option valuation in discrete time. *The Journal of Finance*, 48(5):1833–1863.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bailey, D. H. and Lopez de Prado, M. (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, 15(2):13.

Bandi, C. and Bertsimas, D. (2014). Robust option pricing. *European Journal of Operational Research*, 239(3):842–853.

Bellini, F. and Bernardino, E. D. (2017). Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, 15(5):725–733.

Bernhard, P. (2003). A robust control approach to option pricing. *Applications of Robust Decision Theory and Ambiguity in Finance. City University Press, London.*

Bernhard, P., Engwerda, J., Roorda, B., M. Schumacher, J., Kolokoltsov, V., Saint-Pierre, P., and Aubin, J.-P. (2013). *The Interval Market Model in Mathematical Finance*, pages 293–317.

Bertsekas, D. P. (2015). *Convex optimization algorithms*. Athena Scientific Belmont.

Bertsimas, D., Kogan, L., and Lo, A. W. (2001). Hedging derivative securities and incomplete markets: an ϵ-arbitrage approach. *Operations Research*, 49(3):372–397.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Brennan, M. J. (1979). The pricing of contingent claims in discrete time models. *The Journal of Finance*, 34(1):53–68.

Carbonneau, A. and Godin, F. (2020). Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, pages 1–16.

Carbonneau, A. and Godin, F. (2021). Deep equal risk pricing of financial derivatives with multiple hedging instruments. *arXiv preprint arXiv:2102.12694*.

Carr, P., Geman, H., and Madan, D. B. (2001). Pricing and hedging in incomplete markets. *Journal of Financial Economics*, 62(1):131 – 167.

Castro, D. D., Oren, J., and Mannor, S. (2019). Practical risk measures in reinforcement learning. *ArXiv*, abs/1908.08379.

Chen, J. M. (2018). On exactitude in financial regulation: Value-at-risk, expected shortfall, and expectiles. *Risks*, 6(2):61.

Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263.

Degris, T., White, M., and Sutton, R. S. (2012). Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 179–186, Madison, WI, USA. Omnipress.

Delbaen, F. and Schachermayer, W. (1995). The variance-optimal martingale measure for continuous processes. *Bernoulli*, 2:81–105.

Detlefsen, K. and Scandolo, G. (2005). Conditional and dynamic convex risk measures. *Finance and stochastics*, 9(4):539–561.

Föllmer, H. and Schied, A. (2011). *Stochastic finance: an introduction in discrete time*. Walter de Gruyter.

Föllmer, H., Sondermann, H., and Sondermann, D. (1985). *Hedging of non-redundant contingent claims*, pages 205 – 223.

François, P., Gauthier, G., and Godin, F. (2014). Optimal hedging when the underlying asset follows a regime-switching markov process. *European Journal of Operational Research*, 237(1):312–322.

Gourieroux, C., Laurent, J. P., and Pham, H. (1998). Mean-variance hedging and numéraire. *Mathematical Finance*, 8(3):179–200.

Guo, I. and Zhu, S.-P. (2017). Equal risk pricing under convex trading constraints. *Journal of Economic Dynamics and Control*, 76:136–151.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Huang, A., Leqi, L., Lipton, Z. C., and Azizzadenesheli, K. (2021). On the convergence and optimality of policy gradient for markov coherent risk.

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300.

Jaschke, S. and Küchler, U. (2001). Coherent risk measures and good-deal bounds. *Finance and Stochastics*, 5(2):181–200.

Jiang, Z., Xu, D., and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

King, A. J. (2002). Duality and martingales: a stochastic programming perspective on contingent claims. *Mathematical Programming*, 91(3):543–562.

Li, B., Wang, J., Huang, D., and Hoi, S. C. H. (2018). Transaction cost optimization for online portfolio selection. *Quantitative Finance*, 18(8):1411–1424.

Liang, Z., Chen, H., Zhu, J., Jiang, K., and Li, Y. (2018). Adversarial deep reinforcement learning in portfolio management. *arXiv preprint arXiv:1808.09940*.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Marzban, S., Delage, E., and Li, J. Y. (2020). Equal risk pricing and hedging of financial derivatives with convex risk measures. *arXiv preprint arXiv:2002.02876*.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, pages 141–183.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moody, J., Wu, L., Liao, Y., and Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5-6):441–470.

Moody, J. E. and Saffell, M. (1999). Reinforcement learning for trading. *Advances in Neural Information Processing Systems*, pages 917–923.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pichler, A. and Shapiro, A. (2018). Risk averse stochastic programming: time consistency and optimal stopping. *arXiv preprint arXiv:1808.10807*.

Rudloff, B., Street, A., and Valladão, D. M. (2014). Time consistency and risk averse dynamic decision models: Definition, interpretation and practical consequences. *European Journal of Operational Research*, 234(3):743–750.

Ruszczyński, A. (2010). Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261.

Ruszczynski, A. and Shapiro, A. (2006). Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561.

Schweizer, M. (1996). Approximation pricing and the variance-optimal martingale measure. *The Annals of Probability*, 24(1):206–236.

Schweizer, M. (1999). A guided tour through quadratic hedging approaches. Technical report, SFB 373 Discussion Paper.

Shapiro, A. (2012). Time consistency of dynamic risk measures. *Operations Research Letters*, 40(6):436 – 439.

Shapiro, A. (2017). Interchangeability principle and dynamic equations in risk averse stochastic programming. *Operations Research Letters*, 45(4):377–381.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, 39(1):119–138.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR.

Singh, R., Zhang, Q., and Chen, Y. (2020). Improving robustness via risk averse distributional reinforcement learning. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M., editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 958–968.

Staum, J. (2007). *Chapter 12 Incomplete Markets*, volume 15 of *Handbooks in Operations Research and Management Science*, pages 511 – 563. Elsevier.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2015). Policy gradient for coherent risk measures. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Urpí, N. A., Curi, S., and Krause, A. (2021). Risk-averse offline reinforcement learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Xu, K., Zhang, Y., Ye, D., Zhao, P., and Tan, M. (2020). Relation-aware transformer for portfolio policy learning. IJCAI.

Xu, M. (2006). Risk measure pricing and hedging in incomplete markets. *Annals of Finance*, 2(1):51–71.

Zhang, Y., Zhao, P., Li, B., Wu, Q., Huang, J., and Tan, M. (2020). Cost-sensitive portfolio selection via deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*.