Essays in information relaxations and scenario analysis for partially observable settings

Octavio Ruiz Lacedelli

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Essays in information relaxations and scenario analysis for partially observable settings

Octavio Ruiz Lacedelli

This dissertation consists of three main essays in which we study important problems in engineering and finance.

In the first part of this dissertation, we study the use of Information Relaxations to obtain dual bounds in the context of Partially Observable Markov Decision Processes (POMDPs). POMDPs are in general intractable problems and the best we can do is obtain suboptimal policies. To evaluate these policies, we investigate and extend the information relaxation approach developed originally for Markov Decision Processes. The use of information relaxation duality for POMDPs presents important challenges, and we show how change-of-measure arguments can be used to overcome them. As a second contribution, we show that many value function approximations for POMDPs are supersolutions. By constructing penalties from supersolutions we are able to achieve significant variance reduction when estimating the duality gap directly, and the resulting dual bounds are guaranteed to provide tighter bounds than those provided by the supersolutions themselves. Applications in robotic navigation and telecommunications are given in Chapter 2. A further application of this approach is provided in Chapter 5 in the context of personalized medicine.

In the second part of this dissertation, we discuss a number of weaknesses inherent in traditional scenario analysis. For instance, the standard approach to scenario analysis aims to compute the P&L of a portfolio resulting from joint stresses to underlying risk factors, leaving all unstressed risk factors set to zero. This approach ignores thereby the conditional distribution of the unstressed risk factors given the stressed risk factors. We address these weaknesses by embedding the scenario

analysis within a dynamic factor model for the underlying risk factors. We recur to multivariate state-space models that allow the modeling of real-world behavior of financial markets, like volatility clustering for example. Additionally, these models are sufficiently tractable to permit the computation (or simulation from) the conditional distribution of unstressed risk factors. Our approach permits the use of observable and unobservable risk factors. We provide applications to fixed income and options portfolios, where we are able to show the degree in which the two scenario analysis approaches can lead to dramatic differences.

In the third part, we propose a framework to study a Human-Machine interaction system within the context of financial Robo-advising. In this setting, based on risk-sensitive dynamic games, the robo-advisor adaptively learns the preferences of the investor as the investor makes decisions that optimize her risk-sensitive criterion. The investor and machine's objectives are aligned but the presence of asymmetric information makes this joint optimization process a game with strategic interactions. By considering an investor with mean-variance risk preferences we are able to reduce the game to a POMDP. The human-machine interaction protocol features a trade-off between allowing the robo-advisor to learn the investors preferences through costly communications and optimizing the investor's objective relying on outdated information.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to start by expressing my most sincere gratitude to my advisor, Professor Martin Haugh, whose mentorship, support and guidance were fundamental in the development of this thesis. Through his example, Martin taught me the value of high standards in both doing and communicating research work. I truly value the mentorship he offered me and his availability during my time at Columbia, which were key to my personal and professional development.

I also want to thank my dissertation committee Professors Garud Iyengar, Agostino Capponi, Ali Hirsa and Ciamac Moallemi. Their careful reading of this work and their comments are highly appreciated. I'm very grateful to Professor Agostino Capponi and to Matt Stern for their contribution to this work and for the interesting discussions we had. A special thank you to Professors Mariana Olvera-Cravioto, Yuan Zhong, Jose Blanchet and Jay Sethuraman for their guidance provided early on.

I am also grateful to the department staff for their help in innumerable occasions and always being happy to help. All those friendships made are invaluable and I thank them for all the time shared and interesting discussions about work and life, especially to those who shared an office with me for most of my time at Columbia, Francois, Camilo and Enrique.

To my parents Octavio and Elizabeth I wish to thank for always supporting me, bringing me up with the values of hard-work and teaching me the high importance of learning. Through their example and support it was possible for me to obtain a good education and always fight to improve and reach my goals. I also want to thank Omar for his company and support throughout the years.

My time at Columbia and this work would have been impossible without the support of my wife Petra, whose patience, love and companionship I value more than anything. Thank you for all your love, advice and support now and always!

To Petra

Chapter 1

Introduction

This dissertation consists of three main and independent essays in which we study important problems in engineering and finance.

In Chapter 2, we investigate how Information Relaxations can be used to obtain dual bounds in the context of Partially Observable Markov Decision Processes (POMDPs). In general, POMDPs result in intractable problems and we must be satisfied with sub-optimal policies. The question of evaluating these policies has been addressed in the Markov decision process (MDP) literature through the use of information relaxation based duality. In this chapter we study and extend this approach to POMDPs, where we highlight the challenges presented in the partially observable setting. We use recently-developed change-of-measure arguments to be able to solve the so-called inner problems and use standard filtering arguments to identify the appropriate Radon-Nikodym derivatives. As a second contribution we show that standard value function approximations for POMDPs are in fact supersolutions. This is of interest for of two important reasons: 1) if penalties are constructed from supersolutions, then absolute continuity of the change-of-measure is not required and we can achieve significant variance reduction when estimating the duality gap directly, and 2) dual bounds constructed from supersolution-based penalties are guaranteed to provide tighter bounds than those provided by the supersolutions themselves. We finally provide results for applications in robotic navigation, telecommunications, and a further application in personalized medicine is provided in Chapter 5.

In Chapter 3, we discuss a number of inherent weaknesses of scenario analysis as typically applied in practice. For instance, in an index options portfolio, a risk manager would compute the stressed P&L of her portfolio resulting from joint stresses to the underlying index and parallel movements to the implied volatility surface of the index options. The scenario analysis report would then be presented as a grid of stressed P&L numbers for each stress scenario under consideration. The implicit assumption of this approach is that all other risk factors are set to zero. However, the expected values of non-stressed factors conditional on the stresses are generally non-zero. Moreover, convexity effects of portfolios that depend non-linearly on the risk factors may result in further inaccuracy of the standard approach. In this chapter, we address these weaknesses by embedding the scenario analysis within a dynamic factor model for the underlying risk factors. In order to model the real-world behavior of financial markets, e.g volatility clustering, we use multivariate state-space models that are sufficiently tractable so that we can compute (or simulate from) the conditional distribution of unstressed risk factors. We demonstrate our approach for observable and unobservable risk factors in applications to fixed income and option markets. In these applications, we are able to show how these two approaches can lead to dramatically different results. Finally, we argue for a more accurate and scientific approach for scenario analysis, where the reported P&L numbers of a given model can be back-tested and therefore possibly rejected.

In Chapter 4, we propose a framework to study a Human-Machine interaction system within the context of financial Robo-advising. In this setting, based on risk-sensitive dynamic games, the robo-advisor adaptively learns the preferences of the investor as the investor makes decisions that optimize her risk-sensitive criterion. The investor and machine's objectives are aligned but the presence of asymmetric information makes this joint optimization process a game with strategic interactions. By considering an investor with mean-variance risk preferences we are able to reduce the game to a POMDP. The human-machine interaction protocol features a trade-off between allowing the robo-advisor to learn the investors preferences through costly communications and optimizing the investor's objective relying on outdated information.

Chapter 2

Information Relaxation Bounds for Partially Observed Markov Decision Processes

Partially observed Markov decision processes (POMDPs) are an important class of control problems that are ubiquitous in a wide range of fields. Unfortunately these problems are generally intractable and so in general we must be satisfied with sub-optimal policies. But how do we evaluate the quality of these policies? This question has been addressed in recent years in the Markov decision process (MDP) literature through the use of information relaxation based duality where the non-anticipativity constraints are relaxed but a penalty is imposed for violations of these constraints. In this chapter we extend the information relaxation approach to POMDPs. It is of course well known that the belief-state formulation of a POMDP is an MDP and so the previously developed results for MDPs also apply to POMDPs. Under the belief-state formulation, we use recently developed change-of-measure arguments to solve the so-called inner problems and we use standard filtering arguments to identify the appropriate Radon-Nikodym derivatives. We also show, however, that dual bounds can also be constructed without resorting to the belief-state formulation. In this case, change-of-measure arguments are required for the evaluation of so-called dual feasible penalties rather than for the solution of the inner problems. We compare dual bounds for both

formulations and argue that in general the belief-state formulation provides tighter bounds. The second main contribution of this chapter is to show that several value function approximations for POMDPs are in fact *supersolutions*. This is of interest because it can be particularly advantageous to construct penalties from supersolutions since absolute continuity (of the change-of-measure) is no longer required and so significant variance reduction can be achieved when estimating the duality gap directly. Dual bounds constructed from supersolution based penalties are also guaranteed to provide tighter bounds than the bounds provided by the supersolutions themselves. We use applications from robotic navigation and telecommunication to demonstrate our results.

2.1 Introduction

Partially observed Markov decision processes (POMDPs) are an important class of control problems with wide-ranging applications in fields as diverse as engineering, machine learning and economics. The resulting problems are often very difficult to solve, however, due to the so-called curse of dimensionality. In general then, these problems are intractable and so we must make do with constructing sub-optimal policies that are (hopefully) close to optimal. But how can we evaluate a given sub-optimal policy? We can of course simulate it many times and obtain a *primal* bound, i.e. a lower (upper) bound in the case of a maximization (minimization) problem, on the true optimal value function. But absent a *dual* bound, i.e. an upper (lower) bound, there is no easy way in general to conclude that the policy is close to optimal.

In the case of Markov decision processes (MDPs), we can construct such dual bounds using the information relaxation approach that was developed independently by Brown, Smith and Sun [17] (hereafter BSS) and Rogers [69]. The information relaxation approach proceeds in two steps: (i) relax the non-anticipativity constraints that any feasible policy must satisfy and (ii) include a penalty that punishes violations of these constraints. In a finite horizon setting BSS showed how to construct a general class of dual feasible penalties and proved versions of weak and strong duality. In particular, they showed that if the dual feasible penalties were constructed using the optimal value function, then the resulting dual bound would be tight, i.e. it would equal the optimal value function. In practice of course, the optimal value function is unknown but the strong duality result

suggests that a penalty constructed from a good approximate value function (AVF) should lead to a good dual bound. If a good primal bound is also available, e.g. possibly by simulating the policy that is greedy with respect to the approximate value function, then the primal and dual bounds will be close and therefore yield a "certificate" of near-optimality for the policy.

The main goal of this work is to extend the information relaxation approach to POMDPs. It is well known of course that POMDPs can be formulated as MDPs by working with the belief-state formulation of the POMDP and so the results established for MDPs therefore also apply to POMDPs. Under the belief-state formulation, we use the recently developed change-of-measure arguments of Brown and Haugh [15] (hereafter BH) to solve the so-called inner problems and we use standard filtering arguments to identify the appropriate Radon-Nikodym derivatives. We also show that information relaxation bounds can also be constructed without resorting to the belief-state formulation of the POMDP. In particular, we can still construct these bounds if we work with the *non*-belief-state formulation of the POMDP, i.e. with the explicit dynamics for the hidden state transitions and observations. If we work with the non-belief-state formulation, however, then the evaluation of so-called dual feasible penalties requires the evaluation of expectations that in general are not available explicitly and are strongly action-dependent. Indeed we need to be able to calculate these expectations efficiently for all possible action histories at each time point on each of the simulated inner problems (see (2.15)). We show that this obstacle can be overcome by again using a change-of-measure argument that limits dramatically the number of expectations that must be computed. The expectations that are required can then be computed using standard filtering techniques and so we can proceed to compute the corresponding dual bounds in the usual manner.

Regardless then of the formulation of the POMDP that we choose to work with, we can use change-of-measure arguments to ensure that dual bounds can be computed efficiently. It is perhaps worth emphasizing, however, that the motivation for using a change-of-measure depends on the POMDP formulation that we work with. With the belief-state formulation evaluating the dual penalties is easy but solving the inner problems is hard. In contrast, when we work with the explicit dynamics for the hidden state transitions and observations, then evaluating the dual penalties is hard but solving the inner problems is easy.

We compare the perfect-information (PI) relaxation bounds that arise from the belief-state and non-belief-state formulation of the POMDP. We argue that the two bounds should be identical in general but that this changes for a specific but natural choice of the change-of-measures. In particular, when calculating the belief-state bound we can use a suitably integrated version of the change-of-measure that we used for the non-belief-state formulation. In that case we argue that the resulting information relaxation bound for the belief-state formulation will be tighter than than information relaxation bound for the non-belief-state formulation.

The second main contribution of this chapter is to show that several standard value function approximations for POMDPs are in fact *supersolutions*. Supersolutions are feasible solutions for the linear programming formulation of an MDP and are therefore upper bounds (in the case of a maximization problem) on the unknown optimal value function. Desai et al. [27] showed how to obtain bound improvements in approximate linear programming with perfect information relaxations, and BH showed information relaxation bounds constructed from supersolution based penalties are guaranteed to provide tighter bounds than the bounds provided by the supersolutions themselves. A further advantage of constructing penalties from supersolutions is that absolute continuity (of the change-of-measure) is no longer required and so significant variance reduction can be achieved when estimating the duality gap directly. These advantages were identified by BH although perhaps not emphasized sufficiently. We therefore believe that the information relaxation approach is particularly valuable in the context of POMDPs. One of the standard AVFs we consider is the so-called fast informed bound update AVF [42]. We extend this approach in a natural way to construct what we call the Lag-2 AVF. We show the Lag-2 AVF is a supersolution and prove that it is a tighter upper bound than that provided by the fast informed bound update AVF.

We demonstrate our results in applications from robotic navigation and telecommunications. The robotic navigation application requires controlling the movements of a robot in a maze with the goal of reaching a desired state within a finite number of time-steps. Our telecommunications application concerns packet transmissions in a multi-access communication setting that uses the *slotted aloha* protocol. In both cases we use the aforementioned supersolutions to construct penalties for the dual bounds. We also use them to construct primal bounds by simulating the policies that

are greedy with respect to them. We demonstrate the bound improvement results of BH and also show that tight duality gaps can be achieved in these applications. In particular, the duality gap can be as much as 85% smaller than the gap given by the primal bound and the corresponding supersolution. (This reduction in duality gap under-estimates the upper bound improvement since the duality gap includes the gap from the primal lower bound to the unknown optimal value function.) In our robotic navigation application, for example, we will see that the tightest duality gap, i.e. the gap between our best lower bound and our best information relaxation-based upper bound, is obtained using the Lag-2 AVF. Moreover, the duality gap is so small that we could argue that we have essentially succeeded in solving the problem.

A further contribution of this work is the implication that the information relaxation approach can be extended to other non-Markovian settings beyond POMDPs. The basic underlying probability structure of a POMDP is a (controlled) hidden Markov model (HMM) where the filtered probability distributions that we need can be computed efficiently. It should be clear from this work that other structures, specifically controlled hidden singly-connected graphical models, would also be amenable to the information relaxation approach since filtered probability distributions for these models can also be computed very quickly. More generally, it should be possible to tackle control problems where the controlled hidden states form a multiply-connected graphical model as is often the case with influence diagrams in the decision sciences literature. In this latter case, we suspect that the non-belief-state formulation is the more natural approach to take.

2.1.1 Literature Review and Chapter Outline

The work of BSS and [69] follows earlier work by [38] and [68] on the pricing of high-dimensional American options. Other related work on American option pricing includes [21] and [2]. The pricing of swing options with multiple exercise opportunities is an important problem in energy markets and the information relaxation approach was soon extended to this problem via the work of [58], [74], [1], [12] and [20] among others. BSS were the the first to extend the information relaxation approach to general MDPs *and* demonstrate the tractability of the approach on large-scale problems. Other notable developments include work by [18] and [16] on the structure of dual feasible penalties,

extensions by BH and [87] to infinite horizon settings, bound improvements in approximate linear programming with perfect information relaxations in [27], the bound improvement guarantees of BH who also use change-of-measure arguments (building in part on Rogers [69]) to solve intractable inner problems. The approach has also been extended to continuous-time stochastic control by [86], and dynamic zero sum-games by [41] and [13]. Recently [8] and [7] have shown how information relaxations can be used to construct *analytical* bounds on the suboptimality of heuristic policies for problems including the stochastic knapsack and scheduling.

The information relaxation methodology has now become well established in the operations research and quantitative finance community with applications in revenue management, inventory control, portfolio optimization, multi-class queuing control and finance. Other interesting applications and developments include [53], [49], [37], [40], [30] and [88].

Finally, we note that POMDPs are a well-established and important class of problems and doing justice to the enormous literature on POMDPs is beyond the scope of this chapter. Instead we refer the interested reader to the recent text [51] for a detailed introduction to the topic as well as an extensive list of references.

The remainder of this chapter is organized as follows. In Section 2.2 we formulate our discrete-time, discrete-state POMDP and also discuss its belief-state formulation there. In Section 2.3 we review information relaxations and the change-of-measure approach of BH for solving the difficult inner problems that arise in the belief-state formulation of POMDPs. In Section 2.4 we consider information relaxations for the *non*-belief-state formulation and then compare information relaxation bounds from the belief-state and non-belief state formulations in Section 2.5. We construct several standard value function approximations for POMDPs in Section 2.6. We also introduce our Lag-2 AVF there and prove that all of these AVFs are in fact supersolutions. We describe our applications to robotic navigation and multiaccess communication in Sections 2.7 and 2.8, respectively. We conclude in Section 2.9. Derivations, proofs and various technical details including how to extend our approach to the infinite horizon setting are relegated to the appendices.

2.2 Discrete-Time POMDPs

We begin with the standard POMDP formulation where we explicitly model the hidden state transitions and observations. We consider a discrete-time setting with a finite horizon T and time indexed by $t \in \{0, 1, \ldots, T\}$. At each time t there is a hidden state, $h_t \in \mathcal{H}$, as well as a noisy observation, $o_t \in \mathcal{O}$, of h_t. After observing o_t at time $t > 0$, the decision maker (DM) chooses an action $a_t \in \mathcal{A}$. We also assume a known prior distribution, π_0, on the initial hidden state, h_0, and the initial action a_0 is based on π_0. For ease of exposition we assume that $\mathcal{H}$, $\mathcal{O}$ and $\mathcal{A}$ are all finite. It is standard to describe the dynamics[1] for $t = 1, \ldots, T$ via the following:

- A $|\mathcal{H}| \times |\mathcal{H}|$ matrix, $P(a)$, of transition probabilities for each action $a \in \mathcal{A}$ with

$$P_{ij}(a) := \mathbb{P}(h_t = j \mid h_{t-1} = i, a_{t-1} = a), \quad i, j \in \mathcal{H}. \tag{2.1}$$

- A $|\mathcal{H}| \times |\mathcal{O}|$ matrix, $B(a)$, of observation probabilities for each action $a \in \mathcal{A}$ with

$$B_{ij}(a) := \mathbb{P}(o_t = j \mid h_t = i, a_{t-1} = a), \quad i \in \mathcal{H}, j \in \mathcal{O}. \tag{2.2}$$

Our POMDP formulation is therefore time-homogeneous but there is no difficulty extending our results to the time-inhomogeneous setting where P and B may also depend on t. Rather than using (2.1) and (2.2), however, we will find it more convenient to use the following alternative, but equivalent, dynamics. In particular, we assume the hidden state and observation dynamics satisfy

$$h_{t+1} = f_h(h_t, a_t, w_{t+1}), \tag{2.3}$$

$$o_{t+1} = f_o(h_{t+1}, a_t, v_{t+1}) \tag{2.4}$$

for $t = 0, 1, \ldots, T-1$ and where the v_t's and w_t's are IID $\mathrm{U}(0, 1)$ random variables for $t = 1, \ldots, T$. We can interpret the v_t's and w_t's as being the IID uniform random variables that are required by the inverse transform approach to generate the state transitions and observations of (2.1) and (2.2), respectively. At each time t, we assume the DM obtains a reward, $r_t(h_t, a_t)$, which is a function of

[1] It may be the case that an initial observation, o_0, is also available and this presents no difficulty as long as its distribution conditional on h_0 is known.

the hidden state, h_t, and the action, a_t. As rewards depend directly on hidden states, but not the observations, the DM does not have perfect knowledge of the rewards obtained. We will assume, however, that the final observation satisfies $o_T = h_T$ so that $r_T(h_T) = r_T(o_T)$. This is without loss of generality since the DM cannot act at time T and so there is no benefit to receiving any information at time T.

A policy $\mu = (\mu_0, \mu_1, \ldots, \mu_T)$ is non-anticipative if it only depends on past and current observations (as well as on the initial distribution, π_0, over h_0). For such a policy we can therefore write the time t action a_t as $a_t = \mu_t(o_{1:t})$ where $o_{1:t} := (o_1, \ldots, o_t)$ and where we have omitted the implicit dependence on π_0. We define a filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ to be the filtration generated by the observations so that $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$. A non-anticipative policy is therefore $\mathbb{F}$-adapted. We also define $\mathcal{F} := \mathcal{F}_T$. We denote the class of all non-anticipative policies by $\mathcal{U}_{\mathbb{F}}$. The objective of the DM is to find an $\mathbb{F}$-adapted policy, μ^*, that maximizes the expected total reward. The POMDP problem is therefore to solve for

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0 \right\} \tag{2.5}$$

and where we acknowledge[2] a slight abuse of notation in (2.5) since there is no time T action μ_T.

2.2.1 The Belief State Formulation of the POMDP

Rather than use the hidden state and observation dynamics of (2.3) and (2.4), we can instead define the POMDP state dynamics in terms of the belief state process, π_t, which lies in the $|\mathcal{H}|$-dimensional simplex. Specifically we can equivalently write the POMDP dynamics as

$$\pi_{t+1} = f_\pi(\pi_t, a_t, u_{t+1}), \quad t = 0, 1, \ldots, T-1 \tag{2.6}$$

where the u_t's are IID U(0,1) random variables and f_π is the state transition function which is only defined implicitly via the filtering[3] algorithm. We now define the filtration $\mathbb{F}^\pi = (\mathcal{F}_0^\pi, \ldots, \mathcal{F}_T^\pi)$ where $\mathcal{F}_t^\pi$ is the σ-algebra generated by $\pi_{0:t}$. We note that the filtrations $\mathbb{F}$ and $\mathbb{F}^\pi$ are not identical

[2] This abuse is also found elsewhere in this article but we can resolve it by simply assuming the existence of a dummy action at time T which has no impact on the time T reward.

[3] The filtering algorithm takes π_t and o_{t+1} (which is a function of π_t, a_t and u_{t+1}) as inputs and outputs π_{t+1}.

and while they are of course related, they actually live on different probability spaces. We can also write the time t reward as a function of the belief state by setting[4] $r(\pi_t, a_t) := \mathbb{E}[r(h_t, a_t) \mid \mathcal{F}_t^\pi]$. The analog of (2.5) under the belief-state formulation is then

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right\} \tag{2.7}$$

where we use $\mathcal{U}_{\mathbb{F}^\pi}$ to denote the class of $\mathbb{F}^\pi$-adapted policies. The advantage of formulating the POMDP via the belief-state is that the problem becomes an MDP albeit a potentially high-dimensional one.

2.3 A Review of Information Relaxations

We now briefly describe the information relaxation approach for obtaining dual bounds. Because this theory has been developed for MDPs, we will focus on the belief-state formulation of (2.7). Solving (2.7) is generally an intractable problem so the best we can hope for is to construct a good sub-optimal policy. In order to evaluate the quality of such a policy, however, we need to know how far its value is from the (unknown) optimal value function, $V_0^*(\pi_0)$. If we could somehow bound $V_0^*(\pi_0)$ with a lower bound, V_0^{lower}, and an upper bound, V_0^{upper}, satisfying $V_0^{\text{lower}} \leq V_0^*(\pi_0) \leq V_0^{\text{upper}}$ with $V_0^{\text{lower}} \approx V_0^{\text{upper}}$ then we can answer this question by simulating the policy in question and comparing its value to V_0^{upper}. In practice, we take V_0^{lower} to be the value of our best $\mathbb{F}^\pi$-adapted policy which can typically be estimated to any required accuracy via Monte-Carlo. The goal then is to construct V_0^{upper} and if it is sufficiently close to V_0^{lower} then we have a "certificate" of near-optimality for the policy in question.

Towards this end we will use the concept of information relaxations and our development will follow that of BSS which can be consulted for additional details and proofs. An information relaxation $\mathbb{G}^\pi$ of the filtration $\mathbb{F}^\pi$ is a filtration $\mathbb{G}^\pi = (\mathcal{G}_0^\pi, \mathcal{G}_1^\pi, \ldots, \mathcal{G}_T^\pi)$, where $\mathcal{F}_t^\pi \subseteq \mathcal{G}_t^\pi$ for each t. We denote by $\mathcal{U}_{\mathbb{G}^\pi}$ the set of $\mathbb{G}^\pi$-adapted policies. Then, $\mathcal{U}_{\mathbb{F}^\pi} \subseteq \mathcal{U}_{\mathbb{G}^\pi}$. Note that a $\mathbb{G}^\pi$-adapted

[4] Indeed, when simulating a policy to compute a *primal* bound using the original POMDP formulation of Section 2.2, we can use $r_t(\pi_t, a_t)$ instead of $r_t(h_t, a_t)$ to compute the rewards. Using $r_t(\pi_t, a_t)$ instead of $r_t(h_t, a_t)$ to estimate a primal bound amounts to performing a *conditional* Monte-Carlo which is a standard variance reduction technique.

policy is generally not feasible for the original *primal* problem in (2.7) as such a policy can take advantage of information that is not available to an $\mathbb{F}^\pi$-adapted policy.

Before proceeding we also need the concept of dual penalties. Penalties, like rewards, depend on states and actions and are incurred in each period. Specifically, for each t, we define a dual penalty, c_t, according to

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi] - \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{G}_t^\pi] \tag{2.8}$$

where $\vartheta_{t+1}(\pi_{t+1})$ is[5] a bounded real-valued function of the time $t+1$ state π_{t+1}. It is straightforward to see that $\mathbb{E}[c_t \,|\, \mathcal{F}_t^\pi] = 0$ for all t and any $\mathbb{F}^\pi$-adapted policy. (In general this is not the case for a $\mathbb{G}^\pi$-adapted policy.) This in turn implies $\mathbb{E}[\sum_{t=0}^T c_t \,|\, \mathcal{F}_0^\pi] = 0$ for any $\mathbb{F}^\pi$-adapted policy. Beginning with (2.7) we now obtain

$$\begin{aligned} V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^T r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi\right] &= \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^T r_t(\pi_t, \mu_t) + c_t \,\middle|\, \mathcal{F}_0^\pi\right] \\ &\leq \max_{\mu \in \mathcal{U}_{\mathbb{G}^\pi}} \mathbb{E}\left[\sum_{t=0}^T r_t(\pi_t, \mu_t) + c_t \,\middle|\, \mathcal{F}_0^\pi\right]. \end{aligned} \tag{2.9}$$

BSS also showed that *strong duality* holds. Specifically, if we could take $\vartheta_{t+1}(\pi_{t+1}) = V_{t+1}^*(\pi_{t+1})$, i.e. use the (unknown) optimal value function as our generating function in (2.8), then we would have equality in (2.9). Indeed a simple inductive proof that works backwards from time T establishes strong duality and also shows that equality holds in (2.9) *almost surely*. That is, if we could use the optimal value function V_t^* to construct the dual penalties then the optimal value of the inner problem (inside the expectation in (2.9)) would equal $V_0^*(\pi_0)$ almost surely. This result has two implications when we have a good approximation, $\widetilde{V}_t$, to V_t^* and we take $\vartheta_{t+1}(\pi_{t+1}) = \widetilde{V}_{t+1}(\pi_{t+1})$. First it suggests that (2.9) should yield a good upper bound on V_0^* and second, the almost sure property of the preceding paragraph suggests that relatively few sample paths should be needed to estimate V_0^{upper} to any given accuracy.

[5] In practice we will take $\vartheta_{t+1}(\pi_{t+1})$ to be an approximation to the time $t+1$ optimal value function. We note that dual feasible penalties are essentially *action-dependent* control variates, a standard variance reduction technique in the simulation literature. Recall also that π_{t+1} is a function of the actions $a_{0:t}$ as well as exogenous noise as described in (2.6).

We can use (2.9) to construct upper bounds on $V_0^*(\pi_0)$ for general information relaxations $\mathbb{G}^\pi$ but it is perhaps easier to understand how to do this when we use the *perfect information* relaxation, which is the most common choice in applications. We will actually refer to this relaxation as the belief-state perfect information relaxation (BSPI) as it is the perfect information relaxation for the belief-state formulation of the problem.

2.3.1 The BSPI Relaxation

The BSPI information relaxation is given by the filtration $\mathbb{B}^\pi := (\mathcal{B}_0^\pi, \ldots, \mathcal{B}_T^\pi)$ where $\mathcal{B}_0^\pi = \mathcal{B}_1^\pi = \cdots = \mathcal{B}_T^\pi := \sigma(u_{1:T})$ where the u_t's are as in (2.6). The DM therefore gets to observe $u_{1:T}$ at time 0 under the BSPI relaxation. Moreover, knowledge of $u_{1:T}$ implies knowledge of the belief states $\pi_{0:T}$ corresponding to all possible action sequences, which implies that $\mathcal{F}_t^\pi \subseteq \mathcal{B}_t^\pi$ for all t so that $\mathbb{B}^\pi$ is indeed a relaxation of $\mathbb{F}^\pi$. The upper bound of (2.9) now yields

$$V_0^*(\pi_0) \leq \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r(\pi_t, a_t) + c_t \,\middle|\, \mathcal{F}_0^\pi\right] \tag{2.10}$$

where c_t now takes the form

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi] - \vartheta_{t+1}(\pi_{t+1}). \tag{2.11}$$

In principle we can evaluate the right-hand-side of (2.10) by simulating J sample paths, $\left(u_{1:T}^{(j)}\right)$, for $j = 1, \ldots, J$, and solving the deterministic maximization problem inside the expectation in (2.10) (the *inner* problem) for each path. If we let $V^{(j)}$ denote the optimal value of the j^{th} inner problem, then $\sum_j V^{(j)}/J$ provides an unbiased estimator of an upper bound, V_0^{upper}, on the optimal value function, $V_0^*(\pi_0)$. Moreover standard methods can be used to construct approximate confidence intervals for V_0^{upper}.

In the BSPI setting, however, the state space is the $|\mathcal{H}|$-dimensional simplex. As a result, solving the inner problem in (2.10) amounts to solving a deterministic DP with a $|\mathcal{H}| - 1$-dimensional state space. For all but the smallest problems, these deterministic DPs will in generally be intractable.

2.3.2 The Uncontrolled Formulation

BH showed how this problem could be solved using a change-of-measure approach. In particular they reformulated the primal problem of (2.7) using an equivalent probability measure under which the chosen actions do not influence the state transition dynamics. Instead, the actions are accounted for by the Radon-Nikodym (RN) derivatives which adjust for the change-of-probability measure. BH called this an *uncontrolled formulation* and showed that their weak and strong duality results continued to hold under such a formulation. In this case the analog of (2.10), i.e. weak duality under the uncontrolled BSPI relaxation, is given by

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t^{\pi}[r_t(\pi_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{2.12}$$

where

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \phi(\pi_t, \pi_{t+1}, a_t)\, \vartheta_{t+1}(\pi_{t+1}) \tag{2.13}$$

$$\Phi_t^{\pi}(\pi_{0:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(\pi_s, \pi_{s+1}, a_s) \tag{2.14}$$

and where $\widetilde{\mathbb{E}}[\cdot]$ denotes an expectation under the new probability[6] measure, $\widetilde{\mathbb{P}}$. The $\phi(\pi_t, \pi_{t+1}, a_t)$ terms in (2.13) and (2.14) are appropriately defined one-step RN derivative terms. Explicit expressions for these RN derivatives are provided and justified in Appendix A.1.1.

Using an uncontrolled formulation results in a dramatic reduction of the state space that needs to be considered in solving the inner problem in (2.12). In particular, when we solve the inner problem as a deterministic dynamic program, we do not need to solve this DP for all possible states π_t in the $|\mathcal{H}|$-dimensional simplex. This is because the sequence of states $\pi_0, \ldots, \pi_T$ is fixed inside the inner problem of (2.12) due to the uncontrolled nature of the formulation where the history of actions does not influence the state transition dynamics. As such, the deterministic DP that solves the inner problem only needs to be solved along a single path of states $\pi_0, \ldots, \pi_T$. Of course this state path will vary across inner problem instances.

[6] Throughout the chapter we will use $\mathbb{P}$ to denote the probability measure for a controlled POMDP formulation such as (2.6) or (2.3) and (2.4). We will use $\widetilde{\mathbb{P}}$ to denote the probability measure for any uncontrolled POMDP formulation. The particular controlled or uncontrolled formulation should be clear from the context.

2.4 Information Relaxations for the Non-Belief-State Formulation

Until now we have followed the approach of BSS and BH to outline how information relaxation dual bounds can be computed for POMDPs using the belief-state (and hence MDP) formulation of these problems. In this section we will show that information relaxation bounds for POMDPs can also be obtained using the non-belief-state formulation of the problem as described in the first part of Section 2.2. This leads to a very different form of inner problem which in principle is much simpler to solve. We will still need to use an uncontrolled formulation, however, in order to evaluate the dual penalties. This is in contrast to the inner problems of the BSPI relaxation where, as discussed in Section 2.3.2, an uncontrolled formulation was required to reduce the effective dimension of the inner problem.

In Section 2.5 we will argue that the information relaxation bounds provided by the non-belief state formulation of this section are weaker than the corresponding bounds provided by the belief-state formulation. Nonetheless, some subtleties (regarding how the inner paths are generated) arise in our argument. Moreover, we believe the non-belief state formulation (and the resulting PI relaxation) may potentially be useful for other non-Markovian control problems where a belief-state formulation doesn't arise as naturally as it does in the case of POMDPs. Influence diagrams, for example, is one such class of problems. See [46] or Chapter 23 of [50] for an introduction to influence diagrams.

2.4.1 The Perfect Information Relaxation

We now assume that the POMDP is formulated using the hidden state and observation dynamics of (2.3) and (2.4). We recall that the filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ is the filtration generated by the observations so that $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$ and π_0. The perfect information (PI) relaxation corresponds to the filtration $\mathbb{I} = (\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_T)$, with $\mathcal{I}_t = \sigma(h_0, w_{1:T}, v_{1:T})$ for all t. In particular, the DM gets to observe all of the w_t's, v_t's and h_0 at time 0 under $\mathbb{I}$. It is worth noting that knowledge of the w_t's, v_t's and h_0 implies knowledge of the observations $o_{1:T}$ corresponding to all possible action sequences. It therefore follows that $\mathcal{F}_t \subseteq \mathcal{I}_t$ for all t so that $\mathbb{I}$ is indeed a relaxation of $\mathbb{F}$. Under the PI relaxation, the equivalent of (2.10), i.e. weak duality for

the non-belief-state formulation, corresponds to

$$V_0^*(\pi_0) \leq \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r_t(h_t, a_t) + c_t \,\middle|\, \mathcal{F}_0\right] \tag{2.15}$$

where the c_t's now take the form

$$c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{1:t+1}). \tag{2.16}$$

We note that that the same ϑ_{t+1}'s that we use in (2.11) can also be used in (2.16). This follows because π_{t+1} is in fact a function of $o_{1:t+1}$ and so it is perfectly fine to write $\vartheta_{t+1}(o_{1:t+1})$ instead of $\vartheta_{t+1}(\pi_{t+1})$.

In principle we can again compute an unbiased estimate of the right-hand-side of (2.15) by first simulating J sample paths, $\left(h_0^{(j)}, w_{1:T}^{(j)}, v_{1:T}^{(j)}\right)$, for $j = 1, \ldots, J$. We solve the inner problem inside the expectation in (2.15) for each such path and then average the corresponding optimal objective functions. It is perhaps worth emphasizing that we still inherit strong duality from the BSPI formulation of the POMDP. In particular, this suggests that a good choice of ϑ_{t+1} should lead to good upper bounds on $V_0^*(\pi_0)$.

2.4.2 Solving the Inner Problem in (2.15)

We would therefore like to use the PI relaxation to construct an upper bound on V_0^* by solving the inner problem in (2.15) as a deterministic dynamic program. The main obstacle we will encounter under the PI relaxation, however, is computing the c_t's as defined in (2.16). We can see this most clearly if we consider the zero-penalty case where we set $\vartheta_{t+1} \equiv 0$. In that case $c_t \equiv 0$ for all t and the inner problem in (2.15) is a simple deterministic DP with just $|\mathcal{H}|$ states. In contrast, when $c_t \equiv 0$ in (2.10), we see that the inner problem in (2.10) is still a deterministic DP but now the state space lies in the $|\mathcal{H}|$-dimensional simplex. The inner problems in (2.10) for the BSPI relaxation are therefore in principle considerably more challenging than the inner problems in (2.15) and this is why the uncontrolled formulation of (2.12) was required.

Unfortunately, if we want to use a non-zero ϑ_{t+1} (as is typically the case), then evaluating the $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ term in (2.16) is challenging. With the PI relaxation of the non-belief-state formulation of (2.3) and (2.4), however, this is not possible because the probability distribution

required to compute $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ depends on the entire history of actions, $a_{0:t}$, up to time t. Moreover, this probability distribution is not available explicitly and must be calculated via a filtering algorithm. This means that in solving the inner problem in (2.15) as a deterministic dynamic program, we would need to compute $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ at each time t for *all* possible action histories, $a_{0:t}$. In fact this is also true for the second term in (2.16), $\vartheta_{t+1}(o_{1:t+1})$. Evaluating the penalties c_t for all possible action histories is therefore clearly impractical for any realistic application. Once again, however, we can use an uncontrolled formulation to resolve this problem.

Before proceeding to the uncontrolled formulation, however, it is worth emphasizing why the calculation of these penalty terms is straightforward for the BSPI relaxation. Consider the term $\mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi]$ that arises in the calculation of the penalty in (2.11) in the case of the BSPI relaxation. Because we are conditioning on $\mathcal{F}_t^\pi$ the calculation of $\mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi]$ depends on π_t (which is known given $\mathcal{F}_t^\pi$) and the time t action a_t. In particular, it does *not depend* on the action history $a_{0:t-1}$ which is in contrast to the term $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ that arises in the PI penalty of (2.16). Therefore under the BSPI relaxation the penalties are easy to calculate for any state π_t. Of course, what is really happening here is that the complexity of evaluating penalties for the inner problems of the PI relaxation is transferred to the complexity of working with a much higher dimensional state-space when solving inner problems for the BSPI relaxation. Either way then, we must use an uncontrolled formulation.

2.4.3 The Uncontrolled Formulation

In order to define an action-independent change-of-probability-measure, we simply define a hidden Markov model (HMM) on the same hidden state and observation spaces as our POMDP. Specifically, for $t = 1, \ldots, T$ we define:

- A $|\mathcal{H}| \times |\mathcal{H}|$ matrix, Q, of transition probabilities, with

$$Q_{ij} := \mathbb{P}(h_t = j \mid h_{t-1} = i), \quad i, j \in \mathcal{H}. \tag{2.17}$$

- A $|\mathcal{H}| \times |\mathcal{O}|$ matrix, E, of observation probabilities with

$$E_{ij} := \mathbb{P}(o_t = j \mid h_t = i), \quad i \in \mathcal{H}, j \in \mathcal{O}. \tag{2.18}$$

Note that both Q and E are *action independent* although in general they could depend on time in which case we would write Q^t_{ij} and E^t_{ij}. In general[7] we will also require them to satisfy the following absolute continuity conditions:

1. $Q_{ij} > 0$ for any $i, j \in \mathcal{H}$ for which there exists an action $a \in \mathcal{A}$ such that $P_{ij}(a) > 0$
2. $E_{ij} > 0$ for any $i \in \mathcal{H}$ and $j \in \mathcal{O}$ for which there exists an action $a \in \mathcal{A}$ such that $B_{ij}(a) > 0$.

A trivial way to ensure these conditions is to have $Q_{ij} > 0$ and $E_{ik} > 0$ for all $i, j \in \mathcal{H}$ and $k \in \mathcal{O}$. As mentioned earlier, we let $\tilde{\mathbb{P}}$ denote the probability measure induced by Q and E with $\widetilde{\mathbb{E}}$ denoting expectations under $\tilde{\mathbb{P}}$. We now proceed by reformulating our POMDP under $\tilde{\mathbb{P}}$ and adjusting rewards (and penalties) with appropriate Radon-Nikodym (RN) derivatives. In Appendix A.1.2 we show that these RN derivatives are of the form $d\mathbb{P}/d\tilde{\mathbb{P}} = \Phi_T(h_{0:T}, o_{1:T}, a_{0:T-1})$ with

$$\phi(i, j, k, a) := \frac{P_{ij}(a)}{Q_{ij}} \cdot \frac{B_{jk}(a)}{E_{jk}} \tag{2.19}$$

$$\Phi_t(h_{0:t}, o_{1:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(h_s, h_{s+1}, o_{s+1}, a_s). \tag{2.20}$$

It is then straightforward to see that

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=0}^{T} r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0\right] = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \widetilde{\mathbb{E}}\left[\sum_{t=0}^{T} \Phi_t r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0\right]. \tag{2.21}$$

We refer to (2.21) as an uncontrolled formulation of the non-belief-state POMDP formulation. The "uncontrolled" terminology reflects the fact that the policy, μ, does not influence the dynamics of the system which are now determined by the action independent transition and observation distributions in Q and E, respectively. The impact of the policy instead manifests itself via the Φ_t's. With this uncontrolled formulation the analog of (2.15), i.e. weak duality for the PI relaxation, is given by

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t[r_t(h_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0\right] \tag{2.22}$$

with

$$c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \phi(h_t, h_{t+1}, o_{t+1}, a_t)\, \vartheta_{t+1}(o_{1:t+1}). \tag{2.23}$$

[7] We will see later in Section 2.6.2 that we can ignore these absolutely continuity conditions when we take the ϑ_t's to be *supersolutions*.

Returning to the penalty in (2.16) we recall that we need to compute $\mathbb{E}\left[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t\right]$ but note that we no longer need to compute it for all possible action histories, $a_{0:t}$, when solving an inner problem in (2.22). This is because the action histories under $\tilde{\mathbb{P}}$ influence neither the dynamics of the hidden states nor the observations. This means we only need to compute $\mathbb{E}\left[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t\right]$ once for each time t in each inner problem. This is a straightforward calculation and the expectation can be computed as

$$\mathbb{E}\left[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t\right] = \sum_{o\in\mathcal{O};h,h'\in\mathcal{H}} \pi_t(h) P_{hh'}(a_t) B_{h'o}(a_t) \vartheta_{t+1}(o_{1:t}, o) \tag{2.24}$$

where $\pi_t(h) := \tilde{\mathbb{P}}(h_t = h \mid o_{1:t})$ can be calculated efficiently using standard HMM filtering methods. As discussed in Section 2.4.1, we can now calculate an unbiased upper bound on V_0^* by solving J instances of the inner problems in (2.22) and averaging their optimal objective values. Note that an inner problem can be solved recursively according to

$$V_t^{\mathbb{I}} = \max_a \{r_t(h_t, a) + c_t + \phi(h_t, h_{t+1}, o_{t+1}, a)\, V_{t+1}^{\mathbb{I}}\} \tag{2.25}$$

for $t = 0, \ldots, T-1$ and where $h_{0:T}$ and $o_{1:T}$ are the hidden states and observations that were generated for that specific inner problem. We also have the terminal condition $V_T^{\mathbb{I}} = r_T(h_T)$ since $c_T = 0$ as each ϑ_{T+1} can be assumed to be identically zero. Each of these J inner problem instances should be independently generated via $\tilde{\mathbb{P}}$ and they can be solved as deterministic dynamic programs. Strong duality suggests that if ϑ_t is a "good" approximation to the optimal value function, V_t^*, then we should obtain tight upper bounds on V_0^*. We will see that this is indeed the case in the robotic navigation and multi-access communication applications of Sections 2.7 and 2.8, respectively.

2.5 Comparing the BSPI and PI Dual Bounds

Consider now the primal problems in (2.5) and (2.7) corresponding to the non-belief-state and belief-state formulations, respectively. In (2.5) the rewards are $r_t(h_t, a_t)$ and the optimisation is over $\mathbb{F}$-adapted policies. In contrast, the rewards are $r_t(\pi_t, a_t)$ and the optimisation is over $\mathbb{F}^\pi$-adapted policies in (2.7). Of course the two objectives are equal since $r(\pi_t, a_t) := \mathbb{E}[r(h_t, a_t) \mid \mathcal{F}_t^\pi]$ and because $\mathcal{F}_t$ contains no relevant information beyond what is in $\mathcal{F}_t^\pi$ (even though $\mathcal{F}_t^\pi \subset \mathcal{F}_t$).

Consider now a third equivalent formulation where the rewards are $r(\pi_t, a_t)$ but the optimisation is over $\mathbb{F}$-adapted policies. In this case we have

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^{\pi} \right\} \tag{2.26}$$

where we note the only difference between (2.7) and (2.26) is that the optimisation is over $\mu \in \mathcal{U}_{\mathbb{F}^\pi}$ in the former and over $\mu \in \mathcal{U}_{\mathbb{F}}$ in the latter. Despite the presence of $r_t(\pi_t, \mu_t)$ in (2.26), this is also a non-belief-state formulation of the problem because $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ where $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$ (and π_0).

The PI relaxation bound corresponding to formulation (2.26) is given by

$$\begin{aligned}
\mathbb{E}[V_0^{\mathbb{I}}] &:= \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r_t(\pi_t, a_t) + c_t \,\middle|\, \mathcal{F}_0^{\pi} \right] \\
&= \mathbb{E}_{h_{0:T}, o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{1:t+1}) \right] \,\middle|\, \mathcal{F}_0^{\pi} \right]
\end{aligned} \tag{2.27}$$

where we have substituted for c_t using (2.16) and where we have used $\mathbb{E}_x$ to denote an expectation taken w.r.t. the random vector x. As we shall see in Section 2.6 all our AVFs $\vartheta(o_{1:t})$ can be written equivalently as $\vartheta(\pi_t)$. Together with the fact that $\mathcal{F}_t$ contains no relevant information beyond what is in $\mathcal{F}_t^{\pi}$, this implies we can write (2.27) as

$$\begin{aligned}
\mathbb{E}[V_0^{\mathbb{I}}] &= \mathbb{E}_{h_{0:T}, o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1}) \right] \,\middle|\, \mathcal{F}_0^{\pi} \right] \\
&= \mathbb{E}_{o_{1:T}}\left[\mathbb{E}_{h_{0:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1}) \right] \mid o_{1:T}, \mathcal{F}_0^{\pi} \right] \,\middle|\, \mathcal{F}_0^{\pi} \right]
\end{aligned} \tag{2.28}$$

where the second equality follows from the tower property. Note that the π_t's appearing inside the inner expectation in (2.28) are deterministic functions of π_0, $o_{1:t}$ and $a_{0:t-1}$ and as such, are independent of $h_{0:T}$, given π_0, $o_{1:T}$ and $a_{0:T}$. It therefore follows that (2.28) becomes

$$\mathbb{E}[V_0^{\mathbb{I}}] = \mathbb{E}_{o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1}) \right] \,\middle|\, \mathcal{F}_0^{\pi} \right] \tag{2.29}$$

$$= \mathbb{E}_{\pi_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1}) \right] \,\middle|\, \mathcal{F}_0^{\pi} \right] \tag{2.30}$$

$$= \mathbb{E}[V_0^{\mathbb{B}^\pi}] \tag{2.31}$$

where we recognize the right-hand-side of (2.30) as the BSPI relaxation bound in (2.10) with penalties given by (2.11) and we use $V_0^{\mathbb{B}^\pi}$ to denote the optimal value of a BSPI inner problem. We therefore have the following result.

Proposition 2.5.1. *Given penalties constructed from the same AVF, the BSPI information relaxation bound is equal to the PI information relaxation bound with rewards $r_t(\pi_t, a_t)$.*

Remark 2.5.1. *One direction of Proposition 2.5.1 is quite obvious and follows immediately from BSS. In particular we note that the BSPI relaxation is weaker than the PI relaxation, i.e. $\mathcal{B}_t^\pi \subseteq \mathcal{I}_t$ for all t. This follows because knowledge of $(v_{1:t}, w_{1:t})$ together with π_0 and the action history $a_{0:t-1}$ is sufficient to determine π_t. That the BSPI bound is at least as good as the PI bound (with rewards $r_t(\pi_t, a_t)$) now follows immediately from Prop. 2.3(i) of BSS since the rewards are identical in both formulations.*

Note that it's clear that Proposition 2.5.1 continues to hold under the *same* absolutely continuous change-of-measure. In particular, such a measure change will preserve equality in (2.29) to (2.31). That said, we never use the same change-of-measure for the PI and BSPI bounds. In general, it is difficult to compare bounds constructed via different changes-of-measure but that will not be true in our POMDP case as the change-of-measures that we propose to use for the PI and BSPI bounds will be closely related. This is most easily explained by way of a simple example where to make matters simple, we will assume the penalties are identically zero.

Consider a POMDP with just two periods, $t = 0$ and $t = 1$. For the PI bound, we consider the change-of-measure given by (2.17) and (2.18), so that the PI relaxation bound corresponding to formulation (2.26) is given by

$$\begin{aligned}\mathbb{E}[V_0^{\mathbb{I}}] &:= \widetilde{\mathbb{E}}\left[\max_{a_0} r_0(\pi_0, a_0) + \phi(h_{0:1}, o_1, a_0) r_1(\pi_1) \,\middle|\, \mathcal{F}_0^\pi\right] \\ &= \widetilde{\mathbb{E}}_{o_1}\left[\widetilde{\mathbb{E}}_{h_{0:1}}\left[\max_{a_0} r_0(\pi_0, a_0) + \phi(h_{0:1}, o_1, a_0) r_1(\pi_1) \mid o_1, \mathcal{F}_0^\pi\right] \,\middle|\, \mathcal{F}_0^\pi\right] \qquad (2.32)\end{aligned}$$

where (2.32) follows from the tower property. We can no longer ignore the $\widetilde{\mathbb{E}}_{h_{0:1}}$ expectation in (2.32), however, because the term $\phi(h_{0:1}, o_1, a_0)$ is not independent of $h_{0:1}$, given π_0, o_1 and a_0. However, we can use Jensen's inequality to exchange the maximization with the expectation w.r.t

$h_{0:1}$ to obtain

$$\mathbb{E}[V_0^{\mathbb{I}}] \geq \widetilde{\mathbb{E}}_{o_1}\left[\max_{a_0} \widetilde{\mathbb{E}}_{h_{0:1}}\left[r_0(\pi_0, a_0) + \phi(h_{0:1}, o_1, a_0) r_1(\pi_1) \mid o_1, \mathcal{F}_0^{\pi}\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{2.33}$$

$$= \widetilde{\mathbb{E}}_{o_1}\left[\max_{a_0} r_0(\pi_0, a_0) + r_1(\pi_1) \widetilde{\mathbb{E}}_{h_{0:1}}\left[\phi(h_{0:1}, o_1, a_0) \mid o_1, \mathcal{F}_0^{\pi}\right] \,\middle|\, \mathcal{F}_0^{\pi}\right]. \tag{2.34}$$

On the other hand, the corresponding uncontrolled BSPI bound is given by the r.h.s of (2.12) with zero penalties and satisfies

$$\mathbb{E}[V_0^{\mathbb{B}^{\pi}}] := \widetilde{\mathbb{E}}_{\pi_1}\left[\max_{a_0} r_0(\pi_0, a_0) + \phi(\pi_{0:1}, a_0) r_1(\pi_1) \,\middle|\, \mathcal{F}_0^{\pi}\right]. \tag{2.35}$$

If we now define the RN derivative

$$\phi(\pi_{0:1}, a_0) := \widetilde{\mathbb{E}}_{h_{0:1}}\left[\phi(h_{0:1}, o_1, a_0) \mid \mathcal{F}_0^{\pi}, o_1\right] \tag{2.36}$$

so that the change-of-measure (2.36) for the belief-state formulation is simply an integrated version of the change-of-measure for the non-belief-state formulation, then we recognize that the r.h.s of (2.34) is equal to the BSPI dual bound in (2.35). In particular, the BSPI bound is tighter than the PI bound when the BSPI change-of-measure is an integrated version of the PI change-of-measure for the PI bound. Such an argument provides some intuition for why we see the BSPI bounds outperforming the corresponding PI bounds in the numerical applications of Sections 2.7 and 2.8.

Which PI Dual Bound is Better?

Based on the previous discussion there are two PI dual bounds of interest, the original with rewards $r_t(h_t, a_t)$ and the new one with rewards $r_t(\pi_t, a_t)$. The latter bounds will be tighter in general than the former. To see this, consider the following POMDP again with just two periods, $t = 0$ and $t = 1$. There are two possible hidden states h_{good} and h_{bad} and the initial belief-state distribution π_0 puts equal probability on each of h_{good} and h_{bad}. The only possible actions are a_{stay} and a_{switch}. If the chosen action at time $t = 0$ is a_{stay} then at time $t = 1$ you will stay in the same hidden state that you were in at time $t = 0$. If the chosen action is a_{switch} at time $t = 0$ then at time $t = 1$ you will move to the other hidden state. So for example, if $h_0 = h_{\text{bad}}$ and you choose action a_{switch} then w.p.1 $h_1 = h_{\text{good}}$. A reward of 1 is realised at $t = 1$ if $h_1 = h_{\text{good}}$ and this in the only possible reward. The observations in this POMDP are completely uninformative.

Consider now an inner problem in the PI formulation with rewards $r_t(h_t, a_t)$ and zero penalties. In this case the DM is guaranteed to get a reward of 1 since she will see h_0. In particular, she will know which of a_{stay} and a_{switch} she should choose to guarantee she is in state $h_1 = h_{\text{good}}$ at time $t = 1$ and therefore earn the reward of 1. For the inner problem in the PI formulation with rewards $r_t(\pi_t, a_t)$ and (zero penalties), the DM can again guarantee that $h_1 = h_{\text{good}}$. This time, however, the reward is $r_1(\pi_1, a_1) = 1/2$ because the observations are non-informative and so π_1 puts equal weight on the two possible hidden states at time $t = 1$. So even though the PI decision-maker knows what the true state is at $t = 1$ she only receives a reward of $1/2$ for this.

More generally, suppose that the observations were informative although in general still noisy. With rewards $r_t(h_t, a_t)$ the DM can always guarantee a reward of 1 at time $t = 1$ in the PI relaxation. In contrast, with rewards $r_t(\pi_t, a_t)$, the DM would receive a reward of $r_t(\pi_t, a_t) \in (1/2, 1]$ at time $t = 1$ if she ensured $h_1 = h_{\text{good}}$ since π_i would then put more weight on $h_1 = h_{\text{good}}$ given that the observations are informative. We note in passing that for the PI bounds of Sections 2.7 and 2.8, we always use the $r_t(\pi_t, a_t)$ form of the rewards.

2.6 Approximate Value Functions and Supersolutions

We now discuss several standard approaches for obtaining approximations to the optimal value function in our POMDP setting. In general we can use each such approximation, $\widetilde{V}_t$, to:

1. Construct a lower bound, V_0^{lower}, on V_0^*, by simulating the policy that is greedy[8] with respect to $\widetilde{V}_t$. Towards this end, we can generate J independent sample paths $\big(h_0^{(j)}, w_{1:T}^{(j)}, v_{1:T}^{(j)}\big)$, for $j = 1, \ldots, J$, where we recall the w's and v's are used for generating the hidden and observation states in equations (2.3) and (2.4) in Section 2.2. For each sample path j we calculate at time t the corresponding belief state π_t using standard filtering techniques, and take the action a_t that obtains the maximum in the chosen AVF from each of (C.5), (C.7) or (2.44) below. If we denote by $V_{\text{lower}}^{(j)}$ the reward obtained from following one of these policies

[8] Recall that a policy is said to be greedy with respect to $\widetilde{V}_t$ if the action, a_t, chosen by the policy at time t is an action that maximizes the current time t reward plus the expected discounted value of $\widetilde{V}_{t+1}$, i.e. $a_t = \text{argmax}_a\{r_t(\pi_t, a) + \mathbb{E}[\widetilde{V}_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}]\}$.

on the j^{th} sample path, then an unbiased estimator of a lower bound on the true optimal value function is given by $\sum_j V_{\text{lower}}^{(j)}/J$.

2. Construct an upper bound, V_0^{upper}, via our BSPI and PI uncontrolled information relaxations by setting $\vartheta_t = \widetilde{V}_t$ in (2.13) and (2.23). This of course is motivated by the strong duality result of BSS which states that if we take $\vartheta_t = V_t^*$ then the dual bound will be tight and coincide almost surely with V_0^*.

If our best lower bound is close to our best upper bound then we will have a certificate of near-optimality for the policy that yielded the best lower bound. Later in Section 2.6.1 we will discuss the concept of supersolutions and state a proposition asserting that the approximate-value functions that we define below are indeed supersolutions. The significance of supersolutions will then be discussed in Sections 2.6.1 and 2.6.2.

We now describe the *MDP*, *QMDP* and *Fast Informed (Lag-1)* value function approximations together with the *Lag-2* approximation which we propose as a natural extension of the Lag-1 approximation. More generally, we could define a Lag-d approximation but the computational requirements for calculating it scale exponentially in the number of lags d. Other approximate solution approaches can be found, for example, in [51]. Before proceeding further, we note that the optimal value function $V_T^*(\pi_T)$ is known at time T and satisfies $V_T^*(\pi_T) = r_T(o_T)$ because of our earlier w.l.o.g. assumption that $o_T = h_T$. This means that each of our AVFs can also be assumed to satisfy $\widetilde{V}_T(\pi_T) = r_T(o_T)$.

The MDP Approximate Value Function

The MDP AVF is constructed from $V_t^{\text{MDP}}(h)$, the optimal value function from the corresponding fully observable MDP formulation where the hidden state, h_t, is actually observed at each time t. It is generally easy to solve for V_t^{MDP} in typical POMDP settings and we can use it to construct an AVF according to

$$\widetilde{V}_t^{\text{MDP}}(\pi_t) := \mathbb{E}[V_t^{\text{MDP}}(h_t) \mid \mathcal{F}_t^{\pi}] = \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\text{MDP}}(h) \tag{2.37}$$

where $V_T^{\text{MDP}}(h) := r_T(h)$ and for $t \in \{0, \ldots, T-1\}$ we define

$$V_t^{\text{MDP}}(h) := \max_{a_t \in \mathcal{A}} \big\{ r_t(h, a_t) + \mathbb{E}[V_{t+1}^{\text{MDP}}(h_{t+1}) \mid h_t = h] \big\}. \tag{2.38}$$

The QMDP Approximate Value Function

The QMDP AVF is constructed using the Q-values [54] which are defined as

$$V_t^{\text{Q}}(h, a) := r_t(h, a) + \sum_{h' \in \mathcal{H}} P_{hh'}(a) V_{t+1}^{\text{MDP}}(h') \tag{2.39}$$

for $t \in \{0, \ldots, T-1\}$. The QMDP AVF is then defined according to

$$\widetilde{V}_t^{\text{Q}}(\pi_t) := \max_{a_t} \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\text{Q}}(h, a_t). \tag{2.40}$$

Note that by exchanging the order of the expectation and max operators in (C.5) and then applying Jensen's inequality, we easily obtain that the QMDP value function is less than or equal to the MDP value function in (2.37).

The Lag-1 Approximate Value Function

The Lag-1 approximation was first proposed in [42] as the *fast informed bound update*. This approximation uses the optimal value function, $V_t^{\text{L1}}(h_{t-1}, a_{t-1}, o_t)$, from the corresponding lag-1 formulation of the POMDP where the hidden state, h_{t-1}, is observed before deciding on the time t action a_t for all $t < T$. We can calculate V_t^{L1} recursively via

$$V_t^{\text{L1}}(h_{t-1}, a_{t-1}, o_t) = \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L1}}(h_t, a_t, o_{t+1}) \mid h_{t-1}, o_t] \tag{2.41}$$

for $t \in \{1, \ldots, T-1\}$ and with terminal condition $V_T^{\text{L1}}(h_{T-1}, a_{T-1}, o_T) := r_T(h_T)$ (since $o_T = h_T$). The corresponding AVF is then defined according to

$$\widetilde{V}_t^{\text{L1}}(\pi_t) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L1}}(h_t, a_t, o_{t+1}) \mid \mathcal{F}_t^{\pi}] \tag{2.42}$$

where the expectation is taken with respect to o_{t+1} and h_t, given the current belief state, π_t. Further details on calculating V_t^{L1} can be found in Appendix A.2.

The Lag-2 Approximate Value Function

The Lag-2 approximation is derived by first constructing the optimal value function

$$V_t^{\mathrm{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t})$$

corresponding to the MDP where the hidden state, h_{t-2}, is observed before taking the decision a_t at time t for all $t < T$. Again the terminal value function is

$$V_T^{L_2}(h_{T-2}, a_{T-2:T-1}, o_{T-1:T}) := r_T(o_T) = r_T(h_T)$$

and the optimal value function, $V_t^{L_2}$, at earlier times is computed iteratively according to

$$V_t^{\mathrm{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t}) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid h_{t-2}, o_{t-1:t}] \tag{2.43}$$

for $t \in \{2, \ldots, T-1\}$. When $t = 0$ or 1 we must adjust (2.43) appropriately so that we only condition on o_0 and $o_{0:1}$, respectively. The calculation of $V_t^{L_2}$ is clearly more demanding than the calculation of $V_t^{\mathrm{L}_1}$ since its state space is larger and since the expectation in (2.43) over (h_{t-1}, h_t, o_{t+1}) is more demanding to compute than the expectation in (C.6) which is over (h_t, o_{t+1}). We define the corresponding Lag-2 AVF according to

$$\widetilde{V}_t^{\mathrm{L}_2}(\pi_t) := \max_{a_t} \mathbb{E}[\max_{a_{t+1}} \mathbb{E}[r_t(h_t, a_t) + r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{L_2}(h_t, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}] \mid \mathcal{F}_t^{\pi}] \tag{2.44}$$

for $t \in \{0, \ldots, T-2\}$, with the understanding that when $t = T-1$, the Lag-2 approximation is equal to the Lag-1 approximation, as there is only one time period remaining at that point. While more demanding to compute, we show in Appendix A.2.3 that the Lag-2 AVF is superior to the Lag-1 AVF in that $V_t^*(\pi_t) \leq \widetilde{V}_t^{\mathrm{L}_2}(\pi_t) \leq \widetilde{V}_t^{\mathrm{L}_1}(\pi_t)$. (The first inequality follows from the supersolution property of the AVFs as discussed in Section 2.6.1 below.) Before proceeding we mention that an alternative and perhaps more natural definition of the Lag-2 AVF is

$$\widetilde{V}_t^{\mathrm{Alt}_2}(\pi_t) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid \mathcal{F}_t^{\pi}]. \tag{2.45}$$

However, it is straightforward to show that $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t) \leq \widetilde{V}_t^{\mathrm{Alt}_2}(\pi_t)$ and so we prefer to use $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$ as our generalization of the Lag-1 AVF.

2.6.1 Supersolutions and Bound Guarantees

We begin by defining the concept of a supersolution.

Definition 2.6.1. *Let ϑ_t be any AVF that satisfies*

$$\vartheta_t(\pi_t) \geq \max_{a_t \in \mathcal{A}} \left\{ r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] \right\} \tag{2.46}$$

for all belief states π_t, and all $t \in \{0, \ldots, T\}$. Then we say that ϑ_t is a supersolution.

It is well-known[9] that a supersolution ϑ_t is an upper bound on the optimal value function $V_t^*(\pi_t)$. Indeed the condition (2.46) is simply the feasibility condition for the linear programming formulation of the belief-state MDP. The supersolution property is particularly important in the context of information relaxations and there are two reasons for this, the first of which is Proposition 2.6.1 below from BH.[10]

Proposition 2.6.1. *(Prop 4.1 in Brown & Haugh, 2017) An information relaxation upper bound based on a penalty constructed from a supersolution is guaranteed to be at least as good as the upper bound provided by the supersolution itself.*

We now state the main result of this section. A proof can be found in Appendix A.3.

Proposition 2.6.2. *The MDP, QMDP, Lag-1 and Lag-2 AVFs are all supersolutions.*

The significance of Proposition 2.6.2, however, is that a dual upper bound (as given by (2.12)) based on a penalty constructed from a supersolution is guaranteed to be no worse than the original upper bound provided by the supersolution itself. We will see this result in action in the numerical results of Sections 2.7 and 2.8 when we see that the information relaxation upper bound is typically significantly better than the bound provided by the supersolution.

---

[9] A proof can be found in standard dynamic programming texts and is based on the linear-programming formulation of the Bellman equation. Note that the "supersolution" terminology was introduced by BH.

[10] This result was first developed by Desai et al. [27] in the context of approximate linear programming with perfect information relaxations.

The Non-Belief State Formulation

While the MDP, QMDP, Lag-1 and Lag2 AVFs were all defined for the belief-state formulation of the POMDP it is clear that they can be viewed as functions of the observation history $o_{1:t}$ (and implicitly the action history $a_{0:t-1}$) rather than the belief state π_t. As such, there is no problem in using these AVFs to construct penalties for the PI relaxation upper bounds corresponding to the non-belief-state formulation of the POMDP. Moreover, Propositions 2.6.1 and 2.6.2 will still apply as long as we take $r_t(\pi_t, a_t) := \mathbb{E}[r_t(h_t, a_t) \mid \mathcal{F}_t]$ rather than $r_t(h_t, a_t)$ to be the rewards in the non-belief-state formulation. This is because the constraint (2.46) defining a supersolution requires $r_t(\pi_t, a_t)$ rather than $r_t(h_t, a_t)$. As discussed at the end of Section 2.5, however, using $r_t(\pi_t, a_t)$ rather than $r_t(h_t, a_t)$ is straightforward and indeed should lead to tighter bounds for the PI relaxation.

2.6.2 Using Supersolutions to Estimate the Duality Gap Directly

A second advantage of working with a supersolution AVF is that when the dual penalties are constructed using a supersolution then the requirement that $\mathbb{P} \ll \tilde{\mathbb{P}}$ can be ignored. This was shown by BH who then exploited[11] this fact by directly estimating the duality gap $V_0^{\text{upper}} - V_0^{\text{lower}}$. We describe their approach here and defer to Appendix A.4 an explanation for why the absolute continuity condition, i.e. $\mathbb{P} \ll \tilde{\mathbb{P}}$, can be ignored when the dual penalties are constructed using a supersolution.

Specifically, suppose we have a good candidate $\mathcal{F}^\pi$-adapted policy, μ, and let $\tilde{\mathbb{P}}$ be the probability measure induced by following this policy. If we set V_0^{lower} to be the expected value of this policy, we then have

$$\begin{aligned} V_0^{\text{lower}} &= \mathbb{E}\Big[\sum_{t=0}^{T} \big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^\pi\Big] \\ &= \widetilde{\mathbb{E}}\Big[\sum_{t=0}^{T} \Phi_t(\mu)\big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^\pi\Big] \end{aligned} \tag{2.47}$$

[11] BH discussed this in their Section 4.3.1 but perhaps under-emphasized this practically important aspect of working with supersolutions.

where the c_t's now play the role of (action-dependent) control variates and where $\Phi_t = \Phi_t(\mu) = 1$ for all t in (2.47) because $\mathbb{P}$ and $\tilde{\mathbb{P}}$ coincide when the policy μ is followed. We can use this same $\tilde{\mathbb{P}}$ to estimate an upper bound

$$V_0^{\text{upper}} = \widetilde{\mathbb{E}}\Big[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t(a_{0:t-1}; \mu)\big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^{\pi} \Big] \tag{2.48}$$

as long as ϑ_t is constructed from a supersolution and where (2.48) now explicitly recognizes the dependence of the Φ_t's on $a_{0:t-1}$ and μ. Since both lower and upper bounds (2.47) and (2.48) are simulated using the same measure, $\tilde{\mathbb{P}}$, we may as well use the same set of paths to estimate each bound. This has an obvious computational advantage since the $r_t(\pi_t, \mu_t)$'s and c_t's that were computed along each sample path for estimating (2.47) can now be re-used on the corresponding inner problem in (2.48).

There is a further benefit to this proposal, however. Because the actions of the policy, μ, are feasible for the inner problem in (2.48), it is clear the term inside the expectation in (2.47) will be less than or equal to the optimal objective of the inner problem in (2.48) along each simulated path. In fact the difference, D, between the two terms satisfies

$$0 \leq D := \max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t(a_{0:t-1}; \mu)\big(r_t(\pi_t, \mu_t) + c_t\big) - \sum_{t=0}^{T} \big(r_t(\pi_t, \mu_t) + c_t\big) \quad \tilde{\mathbb{P}} \text{ a.s.} \tag{2.49}$$

and provides an unbiased estimate of the duality gap, $V_0^{\text{upper}} - V_0^{\text{lower}}$. Finally, we expect that the variance of the random variable, D, should be very small due to a strong positive correlation between each of the terms in (2.49). As a result, we anticipate that very few sample paths should be required to estimate the duality gap to a given desired accuracy as long as μ is sufficiently close to optimal. This approach to evaluating a strategy, i.e. by estimating the duality gap, requires very little work over and beyond the work required to estimate V_0^{lower}. And because the variance of D is often extremely small, we generally only need to estimate the duality gap and solve the inner problem on a small subset of the paths that may have been used to estimate V_0^{lower} directly.

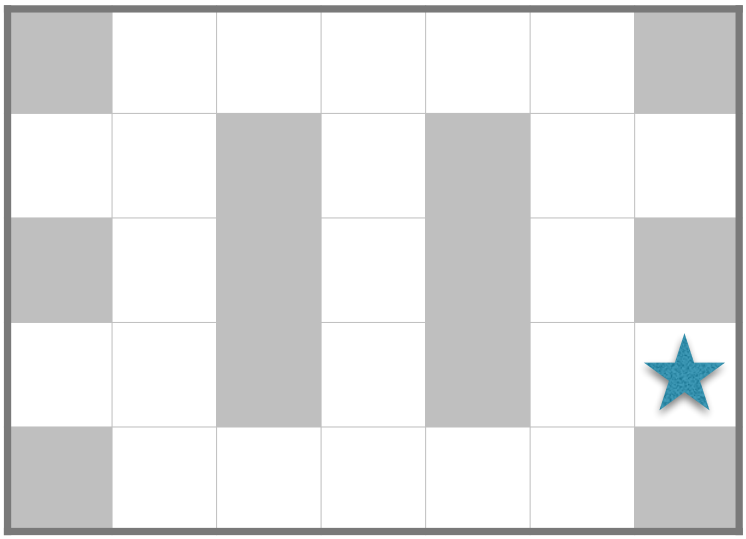


Figure 2.1: Maze representation for the robot navigation problem. The white spaces indicate the possible hidden states where the robot can be located. The star indicates the goal state.

2.7 An Application to Robotic Navigation

We now apply our results to a well-known robotic navigation application and our problem formulation follows [54][43][63]. A robot is placed randomly in one of the 22 white squares (excluding the goal state) inside the maze depicted in Figure 2.1. The robot must navigate the maze, one space at a time, with the objective of reaching the goal state in 10 movements and only traversal along white squares is possible. The exact position within the maze is not directly known to the robot. Sensors placed on the robot provide noisy information on whether or not a wall (depicted as grey squares and edges of the maze) is present on the neighboring space for each of the four compass directions. After taking these readings, the robot must choose one of five possible actions: (attempt to) move north, east, south or west, or stay in the current position.

The sensors have a noise factor of $\alpha \in [0, 1]$. This factor represents two types of errors: a wall will fail to be recognized with probability α when a wall exists, and a wall is incorrectly observed with probability $\alpha/2$ when it does not exist. A second source of uncertainty results from the imperfect movements of the robot. Specifically, after a decision to move has been made, the

robot will move in the opposite direction with probability 0.001, the +90 degree direction with probability 0.01, the -90 degree direction with probability 0.01 and it will fail to move at all with probability 0.089. The robot therefore succeeds in moving in the desired direction with probability 0.89. These movement probabilities are normalized in the event that a particular direction is not possible due to the presence of a wall. The robot may also choose to stay in its current location and such a decision is successful with probability 1.

We formulate the control problem as a POMDP with horizon $T = 10$ periods, 23 hidden states including the goal state h_{goal}, five actions and 16 possible observations. The hidden state h_t at time t is the current position of the robot and is 1 of the 23 white squares in the maze. The observation at time $t < T$ is a 4×1 binary vector of sensor readings indicating whether or not a wall was observed in each compass direction. The possible actions are the direction of desired movement or the decision to stay. Note the observation probabilities are action-independent conditional on the current hidden state. That is, B_{ij} in (2.2) (or equivalently f_o in (2.4)) does not depend on the current action a *given* the current hidden state h. At time $t = 0$ the robot is allowed to take an initial sensor reading o_0, with the distribution of o_0 as described above. Prior to this initial observation, the robot has a prior distribution over the initial hidden state h_0 that is uniform over the 22 non-goal states.

There is a reward function at time T which is defined as $r_T(h_T) = 1$ if $h_T = h_{goal}$, and zero otherwise. All intermediate rewards are zero. Finally, we define $o_T \equiv h_T$ so that we know for certain whether or not the terminal reward was earned or not at the end of the horizon.

2.7.1 The Uncontrolled Formulation

Because all of our AVFs are supersolutions we were able to ignore the absolute continuity requirement when defining the change-of-measures for the uncontrolled formulations. Specifically we used the policies that were greedy w.r.t the QMDP, Lag-1 and Lag-2 AVFs to define uncontrolled-measure changes for the PI bounds. The corresponding measure change for the BSPI bound was then obtained by filtering the actions (that were greedy w.r.t the AVF under consideration) and observations to obtain an action-independent belief-state distribution. This amounts to integrating

the RN derivatives for the uncontrolled non-belief-state formulation to obtain the RN derivatives for the uncontrolled belief-state formulation as discussed in Section 2.5. Further details and explicit calculation of these RN derivatives are described in Appendix A.1.

We can then solve the inner problems in (2.12) and (2.22) as simple deterministic dynamic programs with terminal value $V_T(o_{0:T}) := 1_{\{h_T = h_{goal}\}}$. Because the hidden states and observations on each simulated path are fixed, only one expectation needs to be computed at each time t to evaluate the penalty in (2.13) or (2.23). We can then calculate an unbiased upper bound on V_0^* by averaging the optimal values of each the J inner problem instances for the PI and BSPI relaxations, respectively. Moreover, since our penalties are constructed from supersolutions we are guaranteed to obtain dual upper bounds that improve on the upper bounds provided by the supersolutions themselves. Furthermore, we can use these penalties as control-variates for the primal problem and therefore estimate the duality gap directly as explained in Section 2.6.2.

2.7.2 Numerical Results

Figures 2.2 and 2.3 display numerical results from our experiments. Specifically, Figure 2.2 displays[12] the MDP, QMDP, Lag-1 and Lag-2 AVFs at time $t = 0$. Since these approximations are supersolutions we know they are also valid upper bounds on the true unknown optimal value function. We also display the dual upper bounds obtained from the uncontrolled PI and BSPI relaxations when the penalties were constructed from the Lag-1 and Lag-2 AVFs, respectively. All of these bounds are displayed as a function of α with the time horizon fixed at $T = 10$ periods. The best lower bound was obtained by simulating the policy that is greedy w.r.t the Lag-2 AVF.

Several observations are in order. We see that each of the dual upper bounds improves upon the respective supersolution that was used to construct the dual penalty in each case. We also see from Figure 2.2 that the duality gap decreases as α decreases and this of course is to be expected. Indeed when $\alpha = 0$ all of the bounds coincide and the duality gap is zero. This is because at that point the robot has enough accuracy and time to be able to infer its position in the maze,

[12] The figures actually report $\mathbb{E}[\widetilde{V}_0^{\mathrm{MDP}}(o_0) \mid \pi_0]$, $\mathbb{E}[\widetilde{V}_0^{\mathrm{Q}}(o_0) \mid \pi_0]$ etc. All of the numerical results in this section and the next were obtained using MATLAB release 2016b on a MacOS Sierra with a 1.3 GHz Intel Core i5 processor and 4 GB of RAM.

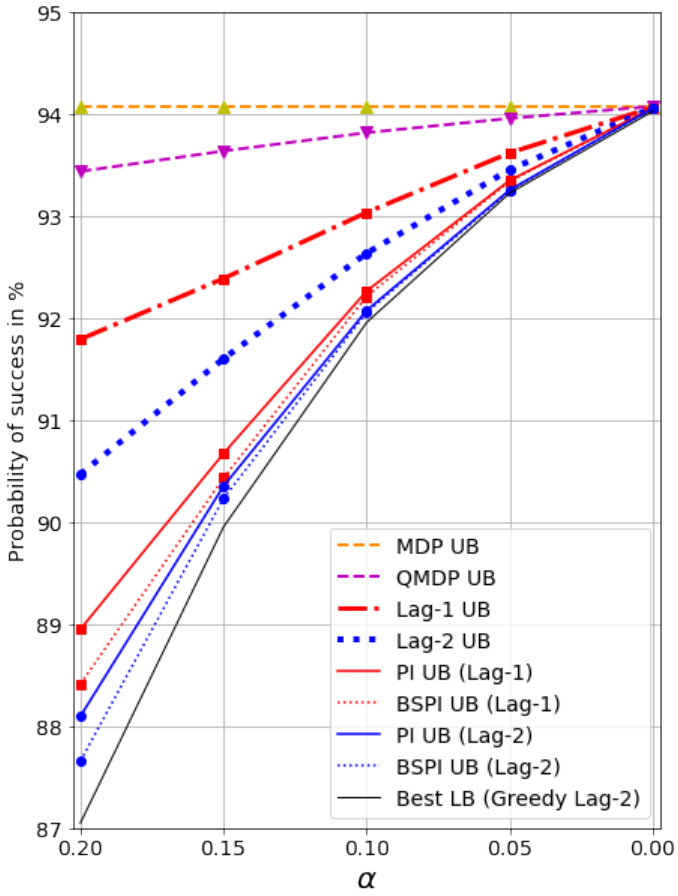
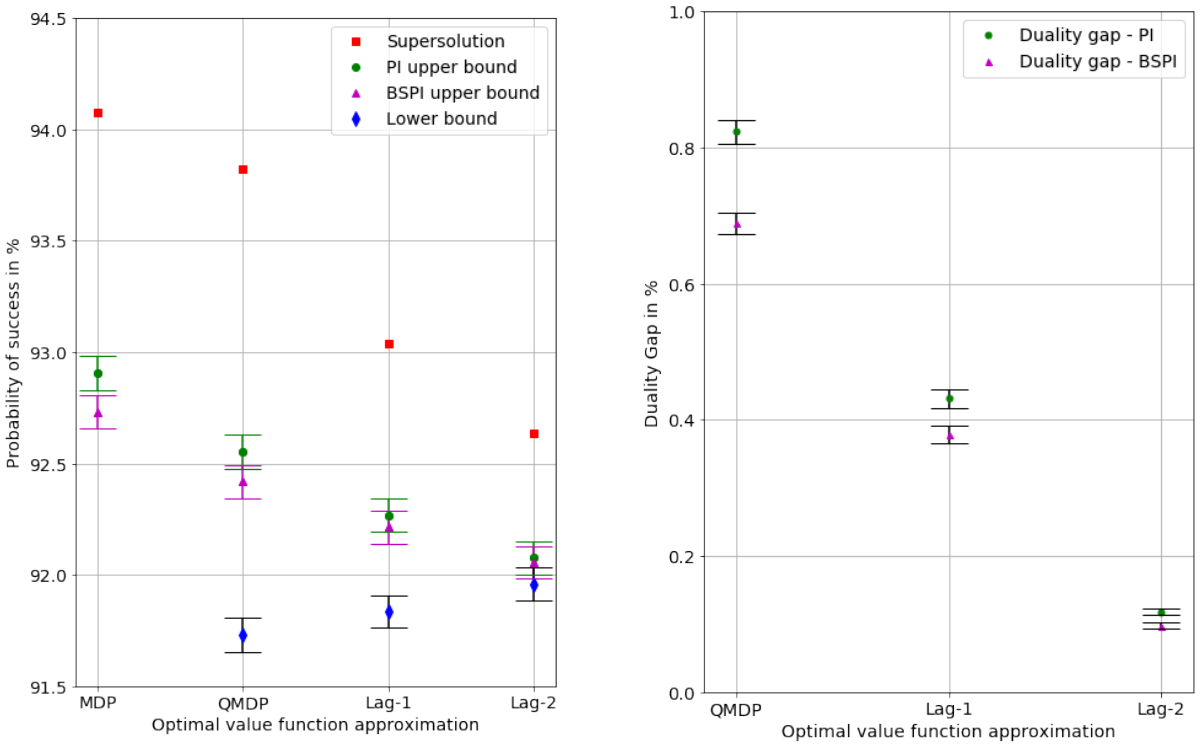


Figure 2.2: Comparison of upper bounds as a function of the noise factor α. The thick dotted lines correspond to the MDP, QMDP, Lag-1 and Lag-2 approximations. The solid (thin dotted) red and blue lines correspond to the dual PI (BSPI) relaxation upper bounds resulting from penalties constructed using the Lag-1 and Lag-2 approximations, respectively. The solid black line displays the best lower bound which in this case is obtained by simulating the policy that is greedy w.r.t. the Lag-2 AVF.

essentially collapsing the POMDP into the MDP version of the problem where the hidden state, h_t, is correctly observed at each time t.

Figure 2.3a displays lower and upper bounds corresponding to each of the four AVFs with $\alpha = 0.10$ and $T = 10$ while Figure 2.3b focuses directly on the corresponding duality gaps. Approximate 95% confidence intervals are also provided and so we see that the various bounds are computed



(a) (b)

Figure 2.3: (a) Lower and upper bounds corresponding to each of the four AVFs. The supersolution upper bound is plotted together with the corresponding dual upper bounds obtained from the perfect information (PI) and belief state perfect information (BSPI) relaxations. Approximate 95% confidence intervals are also provided via error bars. The model parameters were $\alpha = 0.10$ and $T = 10$. (b) Duality gap estimates and confidence intervals for the value function approximations from Figure 2.3a. Details on how the duality gap can be estimated directly are provided in Appendix A.4.

to a high degree of accuracy. Several observations are again in order. First, we note the lower and upper bounds improve as we go from the MDP approximation to the QMDP approximation to the Lag-1 and Lag-2 approximations. This is not surprising since each of these approximations uses successively less information regarding the true hidden state at each time t. Second, we again see that each of the dual upper bounds improves upon its corresponding supersolution. We also observe that regardless of the AVF (that we used to construct the penalties and resulting change-

of-measure), the BSPI bound is always superior to the corresponding PI bound.

We also note that the best duality gap (approximately $92.06\% - 91.96\% = 0.10\%$) is approximately an 85% relative improvement over the gap between the Lag-2 supersolution and the best lower bound (which is given by the policy that is greedy w.r.t the Lag-2 supersolution). While these numbers may not appear very significant on an absolute (rather than relative) basis, in many applications these differences can be significant at the margin. Moreover, there are undoubtedly applications where the best available supersolution will not be close to its corresponding lower bound in which case the improvement provided by the best information relaxation dual bound could be very significant.

The number of simulated paths that we used to generate the various PI and BSPI bounds and duality gaps are reported in Table 2.1 together with corresponding run-times and mean standard errors. All of the numbers are reported as percentages so for example, the BSPI Lag-2 duality gap is a mere 0.10%. The most obvious feature of the tables is how little time was required to compute the dual bounds in comparison to the lower bounds. This comparison is a somewhat misleading, however. In particular, the lower bounds were constructed using the penalties as (action-dependent) control variates, a standard variance reduction technique. Once these control variates were calculated on each simulated path, they could then be re-used as penalties when solving the inner problem along the same path. These control variates were quite expensive to compute, however, and in Table 2.1 this cost has been allocated to the run times for the lower bound. It is therefore fairer to add the run-times for the LB and DG columns and interpret that as the overall time required to compute the lower bounds and duality gap. We do note, however, that the reported standard errors are very small and so we could have used significantly fewer sample paths to still obtain sufficiently accurate estimates of the lower bounds and duality gaps.

Approx.	MDP		QMDP		Lag-1		Lag-2	
	LB*	DG	LB	DG	LB	DG	LB	DG
PI results								
Mean	-	1.15	91.73	0.82	91.84	0.43	91.96	0.12
Std. dev.	-	0.016	0.039	0.009	0.038	0.007	0.038	0.002
Run time (in minutes)	-	0.26	6.56	0.49	6.97	0.55	233	1.04
Supersolution UB	94.08		93.82		93.04		92.64	
DG reduction	51%		61%		64%		82%	
BSPI results								
Mean	-	0.97	91.73	0.69	91.84	0.38	91.96	0.10
Std. dev.	-	0.015	0.039	0.008	0.038	0.007	0.038	0.002
Run time (in minutes)	-	0.41	6.58	0.77	6.89	0.81	235	1.59
Supersolution UB	94.08		93.82		93.04		92.64	
DG reduction	59%		67%		68%		85%	

*There is no greedy policy w.r.t. the MDP AVF.

Table 2.1: Numerical results for the maze application with $\alpha = 0.10$. We used 50,000 sample paths to estimate the lower bounds and their corresponding dual upper bounds and duality gaps (DG). All numbers are expressed as percentages.

2.8 An Application to Multiaccess Communication

Our second application is a well-known[13] multiaccess communication problem in which multiple remote users share a common channel. Users with information packets wish to transmit them through the channel and this can only be done at integer times. Users only submit at most one packet per time slot. If only one user submits a packet through the channel in a given time slot then the packet will be successfully transmitted in that slot. If more than one user submits a

[13]See, for example, Chapter 4 of [14] for an overview of the problem.

packet, however, then the packets will collide, transmission fails and the packets are returned to their respective users to be sent at a later time slot. If no packet was sent during a time slot, then the system is said to be idle in that slot. Users cannot communicate with each other and therefore do not know the action histories of other users.

The total number of packets waiting to be delivered at time t is called the *backlog* and is denoted by h_t. While the backlog is not directly observed by the users, they do know the history of the channel activity via observations of collisions ($o_t = 2$), successful transmissions ($o_t = 1$) and idle time slots ($o_t = 0$). In addition, new packets arrive randomly to the backlog at the end of period t. The number of arrivals, denoted by $z_t \geq 0$, are assumed to follow some discrete probability distribution independent of prior arrivals, and they can be first scheduled for transmission beginning in period $t + 1$. The backlog therefore evolves according to

$$h_{t+1} = \begin{cases} h_t + z_t - 1, & \text{if } o_t = 1 \\ h_t + z_t, & \text{otherwise.} \end{cases} \tag{2.50}$$

The *slotted Aloha* scheduling strategy prescribes each packet in the backlog to be scheduled for transmission with probability $a_t \in \mathcal{A} := [0, 1]$. This probability is common to all waiting packets and transmission attempts are independent across packages. It is therefore easy to see that the probability of a transmission ($o_t = 1$) during slot t is $h_t a_t (1 - a_t)^{h_t - 1}$. We assume a reward of $r_t(h_t)$ is obtained at time t where $r_t(\cdot)$ is a monotonically decreasing function of the backlog. The objective is to choose a transmission probability a_t to maintain a small backlog or equivalently, to maximize the probability of a transmission. In the fully-observable case where h_t is observed by the DM, it is straightforward to see that the maximum transmission probability is attained at $a_t = 1/h_t$ when $h_t \geq 1$. However, in the POMDP setting where h_t is not directly observable computing an optimal policy is generally intractable.

In order to adapt this problem to our finite state and action framework, we restrict the maximum number of packets in the backlog to be $M_h = 30$, so that $h_t \in \mathcal{H} = \{0, 1, \ldots, M_h\}$. We assume that arrivals z_t follow a Poisson distribution with mean λ, but truncate this distribution so that, if the current backlog is h_t, then the maximum number of arrivals is limited to $M_h - h_t$. This is easily

accomplished by taking

$$P_z(k \mid h_t) := P(z_t = k \mid h_t) = \frac{f(k;\lambda)}{F(M_h - h_t;\lambda)}, \text{ for } k = 0, \ldots, M_h - h_t \tag{2.51}$$

where $f(\cdot;\lambda)$ and $F(\cdot;\lambda)$ denote the PMF and CDF, respectively, of the Poisson distribution with parameter λ. To deal with the continuous action space, we must discretize $[0,1]$. Following [19], and recalling that $a_t = 1/h_t$ maximizes the transmission probability for a given known state h_t, we set the discrete action set to be

$$\mathcal{A} := \left\{ \frac{1}{m} : m = 1, \ldots, M_h \right\} \tag{2.52}$$

As stated earlier, observations o_t of the channel history satisfy $o_t \in \mathcal{O} = \{0, 1, 2\}$. The observation probabilities depend on the current backlog h_t and decision a_t, and satisfy

$$B_{ho}(a) := \begin{cases} (1-a)^h, & \text{if } o = 0 \\ ha(1-a)^{h-1}, & \text{if } o = 1 \\ 1 - (1-a)^h - ha(1-a)^{h-1}, & \text{if } o = 2 \end{cases} \tag{2.53}$$

where $B_{ho}(a) := \mathbb{P}(o_t = o \mid h_t = h, a_t = a)$. The state transmission probabilities implied by (2.50) satisfy for $h, h' \in \{0, 1, \ldots, M_h\}$

$$P_{hh'}(o) = \begin{cases} 0, & \text{if } h' < h - 1, \\ P_z(h' - h + 1 \mid h) & \text{if } o = 1 \text{ and } h' \geq h - 1, \\ P_z(h' - h \mid h) & \text{if } o \in \{0, 2\} \text{ and } h' \geq h \end{cases} \tag{2.54}$$

where $P_{hh'}(o) := \mathbb{P}(h_{t+1} = h' \mid h_t = h, o_t = o)$ and where $P_z(k \mid h)$ corresponds to the probability mass function of the truncated Poisson arrivals given in (2.51).

A couple of observations are in order. First, we note that in contrast to our earlier description of the POMDP framework, we assume here that the observation o_t is a function of the *current action* a_t rather than the previous action a_{t-1}. This results in a slightly different but equally straightforward filtering algorithm to compute the belief-state any point in time. Second, we note that conditional on the observation o_t, the hidden-state dynamics are action-independent. This means that in defining an action-independent change-of-measure it will only be necessary to change the observation probabilities $B_{ho}(a)$.

2.8.1 Value Function Approximations

To simplify matters we only consider the MDP and QMDP AVFs in this application. They satisfy

$$\widetilde{V}_t^{\mathrm{MDP}}(\pi_t) := \sum_{h\in\mathcal{H}} \pi_t(h) \max_{a_t\in\mathcal{A}} V_t^{\mathrm{Q}}(h, a_t) \tag{2.55}$$

$$\widetilde{V}_t^{\mathrm{Q}}(\pi_t) := \max_{a_t} \sum_{h\in\mathcal{H}} \pi_t(h) V_t^{\mathrm{Q}}(h, a_t) \tag{2.56}$$

where

$$V_t^{\mathrm{Q}}(h, a) := r_t(h) + \sum_{h'\in\mathcal{H}} \sum_{o\in\mathcal{O}} P_{hh'}(o) B_{ho}(a) V_{t+1}^{\mathrm{MDP}}(h')$$

$$V_t^{\mathrm{MDP}}(h) := \max_{a_t\in\mathcal{A}} V_t^{\mathrm{Q}}(h, a_t)$$

for $t \in \{0, \ldots, T\}$ with terminal condition $V_{T+1}^{\mathrm{MDP}} := 0$. Note that because the time t observation o_t is now a function of a_t, the belief state π_t is a function of the observation and action histories $o_{0:t-1}$ and $a_{0:t-1}$, respectively, rather than $o_{1:t}$ and $a_{0:t-1}$.

2.8.2 The Uncontrolled Formulation

Since the MDP and QMDP AVFs are[14] supersolutions, we can ignore the absolute continuity requirement and define an uncontrolled emission probability matrix according to

$$E_{ij}^t \equiv B_{ij}\big(\operatorname*{argmax}_{a\in\mathcal{A}} V_t^{\mathrm{Q}}(i, a)\big), \tag{2.57}$$

That is, we use the emission probability matrix induced by following a policy that is greedy w.r.t the QMDP value function approximation. Because the hidden-state transitions are already action-independent (given the current observation) we leave those dynamics unchanged under $\tilde{\mathbb{P}}$. As previously mentioned, the POMDP dynamics here are different to the baseline case as defined in Section 2.2 because of the timing of observations and actions whereby the the observation o_t is a function of a_t rather than a_{t-1}. This results in slightly different filtering updates and RN derivative calculations and we give them explicitly in Appendix A.5.

[14] It is easy to adapt the proofs of Appendix A.3 (to handle the fact that the observation o_t is a function of a_t rather than a_{t-1}) to show that the MDP and QMDP AVFs are supersolutions.

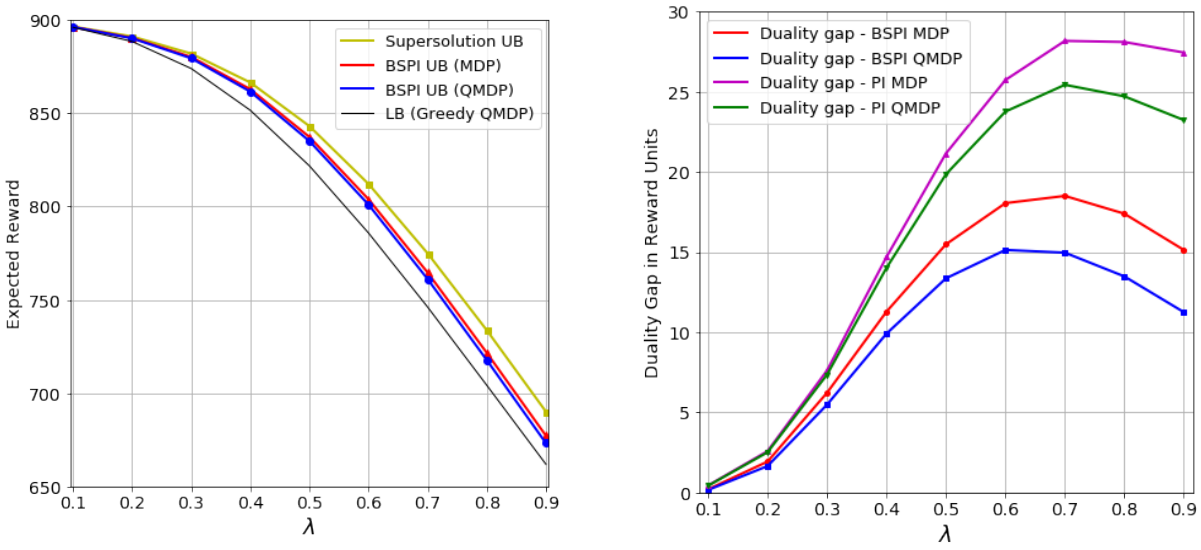
2.8.3 Numerical Results

We consider a system with $T = 30$ periods and initial belief-state $\pi_0 = [1, 0, \ldots, 0]$ so that the system is initially empty w.p. 1. We assume a linear function $r_t(h_t) := M_h - h_t$ so that the reward is maximal (and equal to M_h) when the backlog is zero and minimal (and equal to zero) when the backlog is at its maximum. We used 1,000 sample paths to estimate the dual upper bounds and duality gaps for the PI and BSPI relaxations.

Figure 2.4a displays the lower and upper bounds corresponding to each of the two AVFs used for various values of λ. We display the dual bounds in that figure for the BSPI relaxation but we remark that the PI dual bounds lie between the supersolution upper bound (the yellow curve) and the BSPI upper bound with penalties constructed using the MDP AVF (the red curve). We also note that the MDP and QMDP supersolution upper bounds are equal because by assumption the system is empty initially so that the left-hand-sides of (2.55) and (2.56) are equal at time $t = 0$. Figure 2.4b illustrates the duality gaps that we estimated directly for both value function approximations and for both relaxations.

A few additional observations are in order. First, we note the dual bounds for the QMDP approximation outperform the corresponding dual bounds for the MDP approximation. This is not surprising since we believe the QMDP AVF to be a better approximation to the unknown optimal value function than the MDP approximation. Second, we observe from Figure 2.4a that both dual bounds obtained from the MDP and QMDP approximations improve upon the supersolution upper bound. (This was also true for the PI relaxation dual bounds.) Finally, we observe that the dual gaps increase in λ up to values of $\lambda \approx 0.7$, and decrease in λ thereafter. This non-monotonicity in λ can be explained by the fact that as $\lambda \nearrow 1$ the system becomes rapidly saturated in which case the DM can infer with a higher degree of confidence (than he would be able to at intermediate values of λ) that the time t backlog is likely to be close to the system cap M_h. As a result we expect the duality gap to decrease as $\lambda \nearrow 1$. Likewise when $\lambda \searrow 0$, we expect the best duality gap to also converge to 0 since the system will generally be empty and the DM will be able to infer this with increasing confidence as fewer and fewer collisions ($o_t = 2$) occur.

When we used the MDP AVF to construct the penalties, the total running time (to calculate



(a) (b)

Figure 2.4: (a) Upper bounds for the slotted Aloha system as a function of the arrival parameter λ. The lower bound is obtained by simulating the policy that is greedy w.r.t. the QMDP AVF. The dual bounds are generated using the BSPI relaxation. (b) Duality gap estimates for the BSPI and PI relaxations as a function of the arrival parameter λ. The widths of the (non-displayed) 95% confidence intervals varied between approximately 0.2 for lower values of λ, to 1 for higher values of λ.

the lower bound and duality gap for each value of λ) was 45.9 seconds and 52.3 seconds for the PI and BSPI relaxations, respectively. Using the QMDP approximation, the corresponding times were 53.6 and 58.9 seconds, respectively.

2.9 Conclusions and Further Research

We have shown how change of measure arguments and an uncontrolled problem formulation can be used to extend the information relaxation approach to POMDP settings where the calculation of dual penalties would otherwise be impossible except in the smallest of problem instances. We have exploited the structure of POMDPs to construct various value function approximations and

show that they are supersolutions. Numerical applications to robotic control and multiaccess communications have demonstrated that significant bound improvements can be obtained using information relaxations when the penalties are constructed from supersolutions. We also used the supersolution property to estimate the duality gap directly and take advantage of the significant variance reduction that follows from this approach.

There are several possible directions for future research. One direction would be to extend the approach to other non-Markovian control problems where the difficulty associated with calculating dual feasible penalties would also be problematic. A particularly interesting application would be to dynamic zero-sum games (ZSG's) where the players have asymmetric information. Following [41], dual bounds on the optimal value of the game can be computed by fixing one player's strategy and bounding the other player's best response. In the case of asymmetric information (which was not considered by [41]), bounding the other player's best response amounts to finding a dual bound on a POMDP and so the techniques developed in this chapter also apply in that setting. Moreover, due to Shapley's seminal results strong duality continues to hold in the ZSG framework so the dual bounds can be used to construct a certificate of near-optimality when each player has close-to-optimal strategies. Another interesting non-Markovian setting is the *influence diagram* framework which is popular in the decision science literature.

A second direction would be to explore the relationship between the quality of the dual bound and the action-independent transition and observation distributions. While the primal, i.e. lower bound, does not depend on the action-independent distributions of the uncontrolled problem formulation, this is not true for the dual bound. Indeed as pointed out in BH, the specific value of the dual bound will depend on the quality of the penalties *and* the action-independent distributions. It would therefore be of interest to explore this dependence further. Moreover, because of the abundance of supersolutions in the POMDP setting, absolute continuity of the action-independent distributions is not a requirement and so, as discussed in Appendix A.4, we would be free to explore dual bounds when the action-independent distributions are defined by good feasible policies.

Acknowledgements

This chapter was joint work with Prof. Martin Haugh. I am very grateful for his guidance and support. The authors are also very grateful to Yuan Zhong for helpful comments and conversations. All errors are our own.

Chapter 3

Embedding Scenario Analysis in Dynamic Factor Models

A classic approach to financial risk management is the use of scenario analysis to stress test portfolios. In the case of an S&P 500 options portfolio, for example, a risk manager might compute the P&L resulting from joint stresses of the underlying security, i.e. the S&P 500, and parallel movements in the S&P 500's implied volatility surface. For example this exercise might report a P&L of $1m in the event that the S&P 500 falls 5% and its implied volatility surface increases by 3 percentage points. But how accurate is this reported value of $1m? Typically such a number is computed under the (implicit) assumption that all other risk factors are set to zero. But this assumption is generally not justified as it ignores the often substantial statistical dependence among the risk factors. In particular, the expected values of the non-stressed factors conditional on the values of the stressed factors are generally non-zero. Moreover, even if the non-stressed factors were set to their conditional expected values rather than zero, the reported P&L might still be inaccurate due to convexity effects in the case of derivatives portfolios whose values typically depend in a non-linear manner on the risk factors. A further weakness of this standard approach to scenario analysis is that the reported P&L numbers are not back-tested so that their accuracy is not subjected to any statistical tests. There are many reasons for this but the main one is that standard scenario analysis is typically conducted without having a probabilistic model for the un-

derlying dynamics of the risk factors. In this chapter we address these weaknesses by embedding the scenario analysis within a dynamic factor model for the underlying risk factors. Such an approach requires multivariate state-space models that can model the real-world behavior of financial markets, e.g volatility clustering, and that are sufficiently tractable so that we can compute (or simulate from) the conditional distribution of unstressed risk factors. We demonstrate how this can be done for observable as well as latent risk factors in examples drawn from fixed income and options markets. We show how the two forms of scenario analysis can lead to dramatically different results particularly in the case of portfolios that have been designed to be neutral to a subset of the risk factors. The contributions of this chapter are: (i) to highlight just how inaccurate the standard approach to scenario analysis can be and (ii) to argue for a more accurate and scientific approach whereby the reported P&L numbers of a given model can be back-tested and therefore possibly rejected.

3.1 Introduction

It goes without saying that financial risk management is a key function throughout the finance and insurance industries. At the aggregate level banks, investments firms and insurance companies all need to understand their exposure to adverse movements in the financial markets. This is also true within these firms at the level of a portfolio manager (p.m.) or trading desk where it is important to operate within certain risk constraints. One of the main approaches to financial risk management is the use of scalar risk measures such as Value-at-Risk (VaR) or Conditional Value-at-Risk (CVaR) to measure the riskiness of a given portfolio over a given time horizon such as one day or one week. While VaR (and to a lesser extent CVaR) are very popular and often mandated by regulators it does have serious weaknesses. First and foremost it can be extremely difficult to estimate the VaR of a portfolio and this is particularly true for portfolios containing complex derivative securities, structured products, asset-backed securities etc. Even when the VaR can be estimated accurately, it is impossible to adequately characterize the risk of a portfolio via a single scalar risk measure such as its VaR. In addition, a VaR does not identify the risk factors within the portfolio nor the exposure of the portfolio to those factors. One way to mitigate this for a derivatives portfolio is

via the so-called Greeks such as the delta, vega and theta of an options portfolio. But the Greeks are only local risk measures and can be extremely inaccurate for large moves in the corresponding risk factors. Such moves, of course, are the principal concern in risk management.

It is no surprise then that scenario analysis is one of the most popular approaches to risk management. While there are many forms of scenario analysis, the basic idea behind it is to compute the P&L of the portfolio under various combinations of stresses to one or more of the risk factors (or securities) driving the portfolio's valuation. Given these P&L numbers, the risk management team can assess whether or not the portfolio is too exposed to any of the risk factors and if so, what actions to take in order to reduce the exposure. In the case of an S&P 500 options portfolio, for example, a risk manager might compute the P&L resulting from joint stresses of the underlying security, i.e. the S&P 500, and parallel movements in the S&P 500's implied volatility surface. For example, this exercise might report a P&L of -$1m in the event that the S&P 500 falls 5% and its implied volatility surface increases by 3 points.

One supposed advantage of scenario analysis is that a probabilistic model for the risk factor dynamics is not required. In the example above, for example, a model is not required to assess how likely is the scenario that the S&P 500 falls approx. 5% and its implied volatility surface increases by approx. 3 points. Instead the portfolio manager or risk management team can use their experience or intuition to assess which scenarios are more likely. For example, it is very unlikely indeed that a large drop in the S&P 500 would be accompanied by a drop in implied volatilities and so the experienced risk manager will know that such a scenario can be discounted. Nonetheless, this approach is not scientific and we are led to wonder as to just how accurate is the reported value of -$1m in the original scenario above?

In fact we argue in this chapter that a scenario P&L number can be very inaccurate. First, such a number is typically computed under the (implicit) assumption that all other risk factors, i.e. all risk factors besides the underling and parallel shifts in the volatility surface in our example above, are set to zero. But this assumption is generally not justified as it ignores the often substantial statistical dependence among the risk factors. In particular, the expected values of the non-stressed factors conditional on the values of the stressed factors, are generally non-zero. Second, even if the

non-stressed factors were set to their conditional expected values rather than zero, the reported P&L might still be inaccurate due to convexity effects in the case of derivatives portfolios whose values typically depend in a non-linear manner on the risk factors. A further weakness of this *standard approach* to scenario analysis is that the reported P&L numbers are not back-tested so that their accuracy is not subjected to any statistical tests. There are many reasons for this but the main one is that standard scenario analysis, as mentioned above, is typically conducted without having a probabilistic model for the underlying dynamics of the risk factors. A second reason is that none of the considered scenarios ever actually occurs since they're zero probability events. After all, the probability of the S&P 500 falling exactly 5% and its entire implied volatility surface increasing by exactly 3 volatility points is zero so one can't immediately reject the number of -$1m.

This is in contrast to the use of VaR where it is quite standard to count the so-called VaR *exceptions* and subject them to various statistical tests that are used to determine the accuracy of the VaR estimation procedure. But the back-testing of VaR is inherently easier at it only requires the use of univariate time-series models for the portfolio P&L. In contrast, back-testing scenario analysis would require multivariate time-series models for the various risk-factors and they are considerably more complicated to estimate and work with than their univariate counterparts. Moreover risk-factor returns are often latent and therefore necessitate the use of *state-space* models. This adds a further complication to back-testing since after the fact one can only estimate (rather than know with certainty) what the realized latent risk factor returns were.

In this chapter we attempt to address these weaknesses with standard scenario analysis by embedding it within a dynamic factor model for the underlying risk factors. Such an approach requires multivariate time series or state-space models that can model the real-world behavior of financial markets, e.g volatility clustering, and that are sufficiently tractable so that we can compute and simulate[1] from the distribution of unstressed risk factors conditional on the given scenario. We demonstrate how this can be done for observable as well as latent risk factors in examples drawn from fixed income and options markets. We also show how the two forms of scenario analysis can

[1] One of the advantages of using simulation is that we can easily estimate other risk measures besides the expected P&L in a given scenario. For example we could estimate the P&L's standard deviation or VaR conditional on the scenario.

lead to dramatically different results particularly in the case of portfolios that have been designed to be neutral to a subset of the risk factors. The twin goals of this chapter then are: (i) to highlight just how inaccurate the standard approach to scenario analysis can be and (ii) to argue for a more accurate and scientific approach whereby the reported P&L numbers of a given model can be back-tested and therefore possibly rejected. The particular models that we use in our numerical applications are intended to simply demonstrate that it is possible and important to embed scenario analysis in a dynamic factor model framework. As such they are merely a vehicle for demonstrating our approach and we don't claim they are the "best" such models or that they would be difficult to improve upon.

The remainder of this chapter is organized as follows. In Section 3.2 we introduce standard scenario analysis and discuss in further detail its many weaknesses. We show how scenario analysis can be embedded in a dynamic factor model framework in Section 3.3 and in Section 3.4 we discuss how this framework can be used to evaluate the performance of standard scenario analysis. We then consider an application to a portfolio of U.S. Treasury securities in Section 3.5 and a portfolio of options on the S&P 500 in Section 3.6. In Section 3.7 we discuss statistical approaches for validating a dynamic factor model in the context of scenario analysis. We conclude in Section 3.8 where we also outline some directions for future research. Certain technical details are relegated to the various appendices.

3.2 Preliminaries and Standard Scenario Analysis

We assume we have a fixed portfolio of securities which in principle could include any combination of securities – derivatives or otherwise – from any combination of asset classes. In practice, however, we are limited to reasonably liquid securities for which historical price data is available. Moreover, because of the many difficulties associated with modelling across asset classes, we mainly have in mind portfolios that contain only securities from just one or two closely related asset classes. Examples include portfolios of options and futures on the S&P 500 or portfolios of US Treasury securities. We consider such portfolios in the numerical experiments of this chapter but it should be possible to handle more complex portfolios albeit at the cost of requiring more sophisticated

models. These more complex examples might include portfolios consisting of options and equity positions on US stocks, portfolios of spot and option position on the major FX currency pairs, or even more ambitiously, portfolios consisting of CDS and CDO positions on US credits.

We assume then we are given a fixed portfolio and the goal is to perform some form of scenario analysis on this portfolio. We let V_t denote the time t value so that the portfolio P&L at time $t+1$ is $\Delta V_t := V_{t+1} - V_t$. In the financial context, we have in mind that time is measured in days so that ΔV_t would then be a daily P&L. We assume V_t is known at time t but ΔV_t is random. A fundamental goal of risk managers then is to understand the distribution of ΔV_t. This is required, for example, to estimate the VaR or CVaR of the portfolio.

As is standard in the risk management literature, we will assume the portfolio value V_t is a function of n risk factors whose time t value we denote by $\mathbf{x}_t \in \mathbb{R}^n$. It therefore follows that $V_t = v(\mathbf{x}_t)$ for some function $v : \mathbb{R}^n \to \mathbb{R}$. The components of $\mathbf{x}_t$ might include stock prices in the case of equity portfolios, yields for fixed income portfolios or implied volatility levels for a number of strike-maturity combinations in the case of an equity options portfolios. While $\mathbf{x}_t$ is random, we assume it is $\mathcal{F}_t$-adapted where $\mathbb{F} := \{\mathcal{F}_t\}$ denotes the filtration generated by all relevant and observable security prices and risk factors in the market. We define the change in risk factor vector $\mathbf{\Delta x}_t := \mathbf{x}_{t+1} - \mathbf{x}_t$ so that

$$\Delta V_t(\mathbf{\Delta x}_t) = v(\mathbf{x}_t + \mathbf{\Delta x}_t) - v(\mathbf{x}_t) \tag{3.1}$$

where we have omitted the dependence of ΔV_t on $\mathbf{x}_t$ in (3.1) since $\mathbf{x}_t$ is known at time t and so the uncertainty in ΔV_t is driven entirely by $\mathbf{\Delta x}_t$.

3.2.1 Standard Scenario Analysis

In a standard scenario analysis (SSA hereafter), the risk manager would identify various stresses to apply to $\mathbf{\Delta x}_t$ in (3.1). For example, such stresses might include parallel shifts or curvature changes in the yield curve for a fixed income portfolio. In the case of a portfolio of futures and options on the S&P 500, these stresses might include shifts to the value of the underlying, i.e. the S&P 500, as well some combination of parallel shifts to the implied volatility surface and a steepening / flattening of the skew or term structure of implied volatilities.

When critiquing SSA it is convenient to work with a factor model for the risk factors $\mathbf{\Delta x}_t$. Such a factor model might take the form

$$\mathbf{\Delta x}_t = \mathbf{B f}_{t+1} + \boldsymbol{\epsilon}_{t+1}, \qquad t = 0, 1, \ldots \tag{3.2}$$

where:

- $\mathbf{f}_{t+1} \in \mathbb{R}^m$ is the **common risk factor (c.r.f.)** random return vector. Some of these factor returns may be latent.
- $\mathbf{B} = [\mathbf{b}_1 \ldots \mathbf{b}_m] \in \mathbb{R}^{n \times m}$ is the matrix of factor loadings and $\mathbf{b}_i \in \mathbb{R}^n$ is the i^{th} column of $\mathbf{B}$.
- The $\boldsymbol{\epsilon}_{t+1}$'s $\in \mathbb{R}^n$ are an i.i.d. sequence of zero-mean random vectors representing idiosyncratic error terms that are assumed to be independent of the common factors returns.

Consider for example a portfolio of US Treasury securities. Then $\mathbf{\Delta x}_t$ would naturally represent changes in yields with each component of $\mathbf{\Delta x}_t$ corresponding to a different maturity. If the first common risk factor represented parallel shifts of the yield curve, we would fix $\mathbf{b}_1$ to be a vector of ones. If we then wished to consider a scenario where all yields increase by 20 basis points, we would set $f_{1,t}$, the first component of $\mathbf{f}_{t+1}$, equal to +20 bps and set the other $m-1$ components of $\mathbf{f}_{t+1}$ (as well as $\boldsymbol{\epsilon}_{t+1}$) to zero. The portfolio P&L would then be computed via (3.1) with $\mathbf{\Delta x}_t$ determined by the stress and (3.2).

More generally, we can define a scenario by jointly stressing any number $k \leq m$ of the c.r.f.'s. Consider again our example of an options and futures portfolio on the S&P 500. In this case suppose the first component of $\mathbf{\Delta x}_t$ refers to the log-return on the S&P 500 between days t and $t+1$ with the other components of $\mathbf{\Delta x}_t$ then representing[2] changes in the implied volatilities (between days t and $t+1$) for the various strike-maturity option combinations that appear in the portfolio. If $f_{1,t+1}$ represents changes in the S&P 500 spot price then[3] $\mathbf{b}_1 = [1\, 0 \cdots 0]^\top$. Similarly, if $f_{2,t+1}$

[2] We are assuming that the main risks in the portfolio are underlying and volatility risks. If for example, the portfolio was exposed to substantial dividend or interest rate risk, which is quite possible in an S&P options portfolio, then additional risk factors for these risks should be included.

[3] We would also have $\mathrm{Var}\,(\epsilon_{1,t}) = 0$ since this would be an instance where a component of $\mathbf{\Delta x}_t$ coincides with one of the c.r.f.'s.

represents parallel shifts to the implied volatility surface then the second column of $\mathbf{B}$ would be $\mathbf{b}_2 = [0\,1\,\cdots\,1]^\top$. We can now consider a scenario where $f_{1,t+1}$ and $f_{2,t+1}$ are simultaneously stressed. For example, a scenario of interest might be one where $(f_{1,t+1},\, f_{2,t+1}) = (-5\%,\, +10)$ corresponding to a 5% fall in the S&P 500 and a 10 volatility point increase across its entire volatility surface. Once again, under the SSA approach the portfolio P&L can be computed via (3.1) with $\mathbf{\Delta x}_t$ determined by (3.2) where $(f_{1,t+1},\, f_{2,t+1}) = (-5\%,\, +10)$ and all other components of $\mathbf{f}_{t+1}$ and $\boldsymbol{\epsilon}_{t+1}$ set to zero.

In practice, a matrix of scenario P&L's might be computed as above and in fact multiple two- or even three-dimensional matrices can be computed corresponding to the simultaneous stressing of $k = 2$ or $k = 3$ different common factors. It is important to emphasize that the typical risk / portfolio manager employing SSA does not have an explicit model like (3.2) at hand nor does he / she need one. The main point of this article then is to highlight the many weaknesses of SSA and to argue for a more systematic and scientific approach to it. We can do this by *explicitly* embedding SSA in a factor model such as (3.2) and computing the scenario P&L by also accounting for the dependence structure in (3.2) and not blindly setting $\boldsymbol{\epsilon}_{t+1}$ and the unstressed components of $\mathbf{f}_{t+1}$ to zero.

3.2.2 Problems with Standard Scenario Analysis

Before proceeding, we first expand on the many weaknesses of SSA. They include:

1. A factor model of the form (3.2) is rarely explicitly stated. In fact, it may be the case that only a subset of the factors, say the first $l \leq m$, are ever considered for stressing. In that case standard scenario analysis works with a "model" of the form

$$\mathbf{\Delta x}_t = \mathbf{B}_{1:l,t}\mathbf{f}_{1:l,t+1} \tag{3.3}$$

   where $\mathbf{B}_{1:l,t}$ refers to the matrix containing the first l columns of $\mathbf{B}$ and $\mathbf{f}_{1:l,t+1}$ the vector containing the first l elements of $\mathbf{f}_{t+1}$. The important feature of (3.3) is that probability distributions are not specified and in fact play no role in it. It is therefore not a probabilistic model for the risk factor returns $\mathbf{\Delta x}_t$.

2. Let $\mathbf{f}_{\mathbf{s},t+1} := (f_{s_1,t+1}, \ldots, f_{s_k,t+1})$ denote the subset of c.r.f.'s that are stressed under a given scenario. We assume $k \leq l$ and each $s_i \leq l$. Then SSA implicitly assumes

$$\mathbb{E}_t[\mathbf{f}_{\mathbf{s}_c,t+1} \mid \mathbf{f}_{\mathbf{s},t+1}] = \mathbf{0} \tag{3.4}$$

where $\mathbf{f}_{\mathbf{s}_c,t+1}$ denotes the non-stressed risk factors in the scenario and we use (here and elsewhere) $\mathbb{E}_t[\cdot]$ to denote expectations that are conditional on $\mathcal{F}_t$. But (3.4) is typically not justified and can lead to a very inaccurate estimated P&L for the scenario.

3. Following on from the previous point, an obvious solution would be to set the unstressed factors $\mathbf{f}_{\mathbf{s}_c,t+1}$ equal to their conditional expectation $\mathbb{E}_t[\mathbf{f}_{\mathbf{s}_c,t+1} \mid \mathbf{f}_{\mathbf{s},t+1}]$ when estimating the scenario's P&L. While this should be an improvement over SSA, it ignores the uncertainty in $\boldsymbol{\epsilon}_{t+1}$ and $\mathbf{f}_{\mathbf{s}_c,t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1})$. This uncertainty may be significant, particularly for portfolios containing securities whose values depend non-linearly on $\boldsymbol{\Delta}\mathbf{x}_t$. But even setting $\mathbf{f}_{\mathbf{s}_c,t+1} = \mathbb{E}_t[\mathbf{f}_{\mathbf{s}_c,t+1} \mid \mathbf{f}_{\mathbf{s},t+1}]$ is not a straightforward task, however, as it requires a model for the common risk factor return dynamics.

4. Finally, SSA does not lend itself to rigorous back-testing and so SSA is not open to statistical rejection. There are several reasons for this. First, each of the scenarios considered by an SSA are zero probability events and none of the considered scenarios will have actually occurred on day $t+1$. If this were the only problem, then it would be easy to overcome. Specifically, on day $t+1$ we could "see" exactly what the return in the S&P 500 was over the period $[t, t+1]$. Similarly we could see what parallel change in the implied volatility surface took place over the period $[t, t+1]$.

   We could then rerun the scenario analysis for exactly this scenario, i.e. the scenario that transpired, and then compare the estimated and realised P&L's. The problem with this, however, is that we cannot directly observe the actual parallel change in the implied volatility surface that transpired. This is because this factor is a latent factor and so could only be estimated / inferred. But to do this a probabilistic model would be required and as we have noted, SSA often proceeds without a probabilistic model. Following on from this point, any probabilistic factor model as in (3.2) would surely be rejected statistically if it did not also

include a multivariate time series component that can capture the fact that the common risk factor return dynamics are not IID but in fact are dependent across time.

We now proceed to explain how SSA can be embedded in a dynamic risk factor model and therefore how the weaknesses mentioned above can be overcome. We note that the dynamic risk factor model is not intended to replace the non-probabilistic model of (3.3). Indeed it is quite possible the portfolio manager likes to think in terms of the risk factors $\mathbf{f}_{1:l,t+1}$ and would be reluctant to see these replaced by alternative risk factors. The goal here then is to embed (3.3) in a dynamic risk factor model as in (3.2).

3.3 A Dynamic Factor Model-Based Approach to Scenario Analysis

In order to embed the SSA approach within a dynamic factor model we need to be able to perform the following steps:

1. Select and estimate a multivariate times series or state-space model for the common factor returns $\mathbf{f}_{t+1}$. We need to be able to handle both observable and latent factors.
2. Specify a factor model (3.2) for the risk factor changes $\mathbf{\Delta x}_t$.
3. Simulate samples of $\boldsymbol{\epsilon}_{t+1}$ and $\mathbf{f}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1})$.
4. Compute the portfolio P&L (3.1) for each simulated sample from Step 3. Given these sample P&L's we can estimate the expected P&L for that scenario as well as any other quantities of interest, e.g. a VaR or CVar for that scenario.

Together Steps 1 and 2 enable us to estimate the joint distribution of the common factor returns conditional on time t information. Specifically, they enable us to estimate π_{t+1} where

$$\mathbf{f}_{t+1} \mid \mathcal{F}_t \sim \pi_{t+1}. \tag{3.5}$$

We assume $\mathcal{F}_t$ includes the time series of risk factor changes $\mathbf{\Delta x}_0, \ldots, \mathbf{\Delta x}_{t-1}$, as well as the time series of *observable* common factor returns. Step 3 then enables us to generate samples from the

distribution of the risk factors $\mathbf{\Delta x}_t$ conditional on $\mathcal{F}_t$ and the scenario $\mathbf{f}_{\mathbf{s},t+1}$. Given these samples, Step 4 is a matter of computing the portfolio P&L for each sample and we assume this step is a straightforward task so that any pricing models required to compute $\Delta V_t(\mathbf{\Delta x}_t)$ given $\mathbf{\Delta x}_t$ are available and easy to implement.

c.r.f.'s can be either observable or latent. Observable common factors might include market indices such as the S&P 500 or Eurostoxx index, foreign exchange rates, index CDS rates, commodity prices etc. The returns of c.r.f.'s that are latent, however, can only be inferred or estimated from other observable data such as the $\mathbf{\Delta x}_t$'s. Examples of latent common factors might include c.r.f.'s that drive the implied volatility surface of the S&P 500, for example. A popular specification would include three c.r.f.'s that drive parallel shifts, term-structure shifts and skew shifts in the implied volatility surface, respectively. Note that such shifts are never observable and can only be inferred from the changes (the $\mathbf{\Delta x}_t$'s) in the implied volatilities of S&P 500 options of various strike-maturity combinations. Another example of latent c.r.f.'s would be the factors that are motivated by a principal components analysis (PCA) of the returns on US Treasuries of various maturities. While there may be twenty or more maturities available, a PCA analysis suggests that changes in the yield curve are driven by just three factors representing, in order of importance, a parallel shift in the yield curve, a steepening / flattening of the yield curve, and a change in curvature of the yield curve, respectively.

Because most settings have one or more latent c.r.f.'s our main focus will be on the use of state-space models to tackle steps 1 to 3. We begin with the case where all c.r.f.'s are latent.

3.3.1 State-Space Modeling of the Common Factor Returns

Suppose then that all common factor returns are latent. One way to proceed is to simply construct point estimates of the latent factors by solving for $k = 1, \ldots, t$ an MLE problem[4] of the form

$$\min_{\mathbf{f}_k \in \mathbb{R}^m} -\log \mathbb{P}_\epsilon(\mathbf{\Delta x}_{k-1} - \mathbf{B}\mathbf{f}_k) \tag{3.6}$$

where $\mathbb{P}_\epsilon(\cdot)$ is the PDF of $\boldsymbol{\epsilon}_k$ from (3.2). Let $\hat{\mathbf{f}}_k$ denote the optimal solution to (3.6). We could then take the $\hat{\mathbf{f}}_k$'s to be observable risk factors and use them, for example in a multivariate GARCH

[4] As an alternative to (3.6) we could obtain the point estimate $\hat{\mathbf{f}}_t$ by solving a cross-sectional regression problem.

setting, to estimate the distribution π_{t+1} of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$. This is clearly sub-optimal, however, as the estimation of the $\mathbf{f}_k$'s ignores the temporal dependence in their dynamics. Moreover, by treating $\hat{\mathbf{f}}_t$ as the true value of $\mathbf{f}_t$ (rather than just a noisy point estimate), we are underestimating the conditional uncertainty in $\mathbf{f}_{t+1}$ when we use these point estimates to estimate π_{t+1}.

Our second and preferred approach overcomes these issues by defining a state-space model for the unobservable common factors then and treating $\mathbf{\Delta x}_0, \ldots, \mathbf{\Delta x}_{t-1}$ as noisy observations of the underlying states $\mathbf{f}_1, \ldots, \mathbf{f}_t$. For example, we could model the unobservable common factor returns via an auto-regressive stochastic process of the form

$$\mathbf{f}_{t+1} = \mathbf{G}\mathbf{f}_t + \boldsymbol{\eta}_{t+1} \tag{3.7}$$

for some matrix $\mathbf{G} \in \mathbb{R}^{m \times m}$ and where the process innovation terms $\boldsymbol{\eta}_t \in \mathbb{R}^m$ are assumed to have zero mean and constant covariance matrix $\mathbf{\Sigma}_\eta$. The initial state $\mathbf{f}_0$ is assumed to follow some probability distribution π_0. The hidden-state process (3.7) together with the observable risk factor changes $\mathbf{\Delta x}_t$ from the factor model in (3.2) now form a state-space model.

As before, our goal is to estimate π_{t+1}, the distribution of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$, where $\mathcal{F}_t$ now only includes the history of observations $\mathbf{\Delta x}_{0:t-1} := \{\mathbf{\Delta x}_0, \ldots, \mathbf{\Delta x}_{t-1}\}$. Note that if we are able to obtain the filtered probability distribution $\mathbb{P}(\mathbf{f}_t \mid \mathbf{\Delta x}_{0:t-1})$, then (3.7) implies we can obtain π_{t+1} as the convolution of the two random variables $\mathbf{G}\mathbf{f}_t \mid \mathcal{F}_t$ and $\boldsymbol{\eta}_{t+1}$. Suppose for example, that π_0 and both process innovations $\boldsymbol{\eta}_{t+1}$ in (3.7) and $\boldsymbol{\epsilon}_{t+1}$ in (3.2) are all Gaussian. Then the filtered distribution $\mathbf{f}_{t+1} \mid \mathcal{F}_t$ is also Gaussian and its mean vector and covariance matrix can be calculated explicitly via the Kalman Filter [48]. In this case π_{t+1} would then also be Gaussian.

For non-Gaussian state-space models, however, obtaining the posterior probability exactly is generally an intractable problem although there are many tractable approaches that can be used to approximate the distribution of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$. The Extended Kalman Filter and the Unscented Kalman Filter [85] can be used for non-linear Gaussian state space models, for example. More generally particle filters [34] or MCMC [82] could also be used to approximate the filtered distribution for non-gaussian state space models. Particle filters suffer from the curse of dimensionality, however, while MCMC is computationally expensive. Nonetheless implementing an MCMC or particle filter (in the lower dimensional setting) for non-linear / non-Gaussian state-space models should not be

too demanding given modern computing power.

As an alternative to computing or approximating the filtered distribution $\mathbb{P}(\mathbf{f}_t \mid \mathbf{\Delta x}_{0:t-1})$, we could simply compute its posterior mean $\mathbb{E}[\mathbf{f}_t \mid \mathbf{\Delta x}_{0:t-1}]$ or its maximum[5] a posteriori (MAP) estimate. Then, using $\hat{\mathbf{f}}_t$ as an approximation to the actual realization of $\mathbf{f}_t$, we can approximate π_{t+1} as the distribution of $\mathbf{G}\hat{\mathbf{f}}_t + \boldsymbol{\eta}_{t+1}$, i.e., the right-hand-side of (3.7), which would simply be the distribution of $\boldsymbol{\eta}_{t+1}$ shifted to have mean $\mathbf{G}\hat{\mathbf{f}}_t$. While this neglects the uncertainty in our estimation of $\mathbf{f}_t$ this is often a second-order issue relative to obtaining the correct mean of π_{t+1}. We consider both the Kalman filtering and MAP approaches in Section 3.5 when we consider scenario analysis for fixed income portfolios consisting of US Treasury securities.

3.3.2 Modeling Both Observable and Unobservable Common Factor Returns

Situations in which there are a combination of observable and latent common factor returns are not uncommon. For example, in an S&P 500 options portfolio a scenario would typically include stresses to some combination of the S&P 500 (observable) and parallel, skew or term structure shifts (latent) in the S&P 500's implied volatility surface. In this case, the challenge is to construct a multivariate state-space / time series model that can simultaneously accommodate observable and latent c.r.f. returns. While there may be many ways to tackle this modeling problem, one obvious approach is to assume all of the c.r.f.'s are latent but that the noisy signals for a subset of them (the observable ones) are essentially noiseless.

To make this more precise, we assume we have m^o observable and m^u latent common factors so that the factor model (3.2) can then be written as

$$\mathbf{\Delta x}_t = \mathbf{B}^o \mathbf{f}^o_{t+1} + \mathbf{B}^u \mathbf{f}^u_{t+1} + \boldsymbol{\epsilon}_{t+1} \tag{3.8}$$

where $\mathbf{B}^o \in \mathbb{R}^{n \times m^o}$ and $\mathbf{B}^u \in \mathbb{R}^{n \times m^u}$ are the factor loadings matrices for the observable and latent common factors $\mathbf{f}^o_{t+1}$ and $\mathbf{f}^u_{t+1}$, respectively. Our objective is to estimate π_{t+1}, the probability distribution of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$, where $\mathcal{F}_t$ now corresponds to the σ-algebra generated by the history

[5] The MAP estimator is given by $\hat{\mathbf{f}}_t := \operatorname{argmax}_{\mathbf{f}_t} \mathbb{P}(\mathbf{f}_t \mid \mathbf{\Delta x}_{0:t-1})$ which is the mode of the filtered distribution. Alternatively, we could instead compute $\operatorname{argmax}_{\mathbf{f}_{0:t}} \mathbb{P}(\mathbf{f}_{0:t} \mid \mathbf{\Delta x}_{0:t-1})$ and then take $\hat{\mathbf{f}}_t$ to be the $(t+1)^{st}$ component of the argmax. Both optimization problems can be solved efficiently using modern optimization techniques. One such technique is discussed in Appendix B.4.

of risk factor changes $\mathbf{\Delta x}_{0:t-1}$ and of the observable common factor returns $\mathbf{f}^o_{1:t}$. We define the $n^o := n + m^o$ dimensional vector

$$\mathbf{y}_t := \begin{bmatrix} \mathbf{\Delta x}_t \\ \mathbf{f}^o_{t+1} \end{bmatrix}$$

which we treat as the time $t+1$ observations vector. The model's latent state variables at time $t+1$ are given by the $m := m^o + m^u$ dimensional vector

$$\mathbf{f}_{t+1} := \begin{bmatrix} \mathbf{f}^o_{t+1} \\ \mathbf{f}^u_{t+1} \end{bmatrix}.$$

We now assume the observation dynamics satisfy

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{\Delta x}_t \\ \mathbf{f}^o_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{B}^o & \mathbf{B}^u \\ \mathbf{I}_{m^o} & \mathbf{0}_{m^{ou}} \end{bmatrix} \begin{bmatrix} \mathbf{f}^o_{t+1} \\ \mathbf{f}^u_{t+1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{t+1} \\ \mathbf{0}_{m^o} \end{bmatrix} \qquad t = 0, 1, \ldots \tag{3.9}$$

where $\mathbf{I}_{m^o}$ is the $m^o \times m^o$ identity matrix, $\mathbf{0}_{m^{ou}}$ is the $m^o \times m^o$ matrix of zeros and $\mathbf{0}_{m^o}$ is an $m^o \times 1$ vector of zeros. We can again assume latent state dynamics of the form given in (3.7). Since (3.7) and (3.9) form a state-space model, we can fit the model and estimate π_{t+1} using the various approaches described above. For instance, assuming $\boldsymbol{\epsilon}_{t+1}$ and $\boldsymbol{\eta}_{t+1}$ to be normally distributed, we could use the EM algorithm to estimate the parameters of the state-space model (3.7) and (3.9) using historical data. The Kalman Filter can then be employed to obtain the filtered probability distribution $\mathbb{P}(\mathbf{f}_t \mid \mathbf{y}_{0:t-1})$ for any sequence of observations $\mathbf{y}_{0:t-1}$. Finally, we can then obtain π_{t+1} exactly as the sum of two normal random vectors $\mathbf{G}\mathbf{f}_t \mid \mathcal{F}_t$ and $\boldsymbol{\eta}_{t+1}$, which of course is also normal. We follow this approach in Section 3.6 where we consider portfolios containing options and futures on the S&P 500 index.

3.4 Evaluating the Performance of SSA

The objective of the dynamic factor model-based scenario analysis (hereafter DFMSA) is to compute

$$\Delta V_t^{\text{dfm}}(\mathbf{c}) := \mathbb{E}_t[\Delta V_t(\mathbf{f}_{t+1}, \boldsymbol{\epsilon}_{t+1}) \mid \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}] \tag{3.10}$$

where $\mathbf{c}$ denotes the levels of the stressed factors in the given scenario, and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ denotes an expectation taken with respect to the distribution π_{t+1}. It's clear we have to be able to compute

or simulate from the distribution of $\mathbf{f}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c})$ in order[6] to calculate the conditional expectation in (3.10) .

Since π_{t+1} is the true conditional distribution of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$, we know that $\Delta V_t^{\text{dfm}}(\mathbf{c})$ is the correct way to estimate the scenario P&L. We can therefore calculate the error obtained from following the SSA approach for a given scenario $\mathbf{c}$ as

$$\mathrm{E}_t^{\text{abs}}(\mathbf{c}) := |\Delta V_t^{\text{dfm}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})| \tag{3.11}$$

where $\Delta V_t^{\text{ss}}(\mathbf{c})$ denotes the estimated scenario P&L at time t according to the SSA approach. We must of course acknowledge that the error in (3.11) is somewhat misleading in that it assumes our dynamic factor model is indeed the correct model that governs the real-world security price dynamics and that we know this. Nonetheless, it seems reasonable to assume that there is some dynamic factor model that governs the real-world security price dynamics and that if our model is not a reasonably good approximation to it, then it would be rejected by one or more of the statistical tests that are briefly discussed in Section 3.7. As such, we feel it is reasonable to take (3.11) as a ballpark estimate of the error than can arise from adopting the SSA approach.

We can also provide a partial decomposition of the error in (3.11) by calculating an alternative scenario P&L that is given by

$$\Delta V_t^{\text{alt}}(\mathbf{c}) := \Delta V_t(\mathbf{B}\boldsymbol{\mu}_t^c) \tag{3.12}$$

where $\mathbf{B}$ is the factor loadings matrix of the factor model (3.2) and

$$\boldsymbol{\mu}_t^c := \mathbb{E}_t[\mathbf{f}_{t+1} \mid \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}]. \tag{3.13}$$

This alternative scenario P&L estimator (suggested in point #3 from Section 3.2.2) goes beyond the SSA approach by using the expected value of the common factor returns conditional on the scenario to estimate the risk factor changes $\boldsymbol{\Delta}\mathbf{x}_t$ via the factor model (3.2), i.e. by setting $\boldsymbol{\Delta}\mathbf{x}_t = \mathbf{B}\boldsymbol{\mu}_t^c$. This leads to the alternative estimated scenario P&L in (3.12). Note that the alternative scenario

[6] We also note that the conditional distribution of $\boldsymbol{\epsilon}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c})$ (where $\boldsymbol{\epsilon}_{t+1}$ is given in (3.2)) is equal to its unconditional distribution since it is assumed to be independent of $\mathcal{F}_t$ and $\mathbf{f}_{t+1}$ by assumption.

P&L $\Delta V_t^{\text{alt}}(\mathbf{c})$ will in general differ from[7] and be less accurate than $\Delta V_t^{\text{dfm}}(\mathbf{c})$ as defined in (3.10). We can then decompose the error in (3.11) by

$$\begin{aligned}
\mathrm{E}_t^{\text{abs}}(\mathbf{c}) &= |\Delta V_t^{\text{dfm}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})| \\
&= |\Delta V_t^{\text{dfm}}(\mathbf{c}) - \Delta V_t^{\text{alt}}(\mathbf{c}) + \Delta V_t^{\text{alt}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})| \\
&\leq |\Delta V_t^{\text{dfm}}(\mathbf{c}) - \Delta V_t^{\text{alt}}(\mathbf{c})| + |\Delta V_t^{\text{alt}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})|. \qquad (3.14)
\end{aligned}$$

We note that $|\Delta V_t^{\text{dfm}}(\mathbf{c}) - \Delta V_t^{\text{alt}}(\mathbf{c})|$ gives a measure of the error that results from ignoring the variance in the conditional distribution of the common factor returns and the idiosyncratic error terms. In contrast, $|\Delta V_t^{\text{alt}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})|$ provides a measure of the error that results from setting the unstressed common factor returns to zero rather than their conditional expected values. While the sum of these two errors does not equal the true error we see from (3.14) that their sum does provide an upper bound on this error. In our numerical applications we found that the second term on the r.h.s. of (3.14), i.e. $|\Delta V_t^{\text{alt}}(\mathbf{c}) - \Delta V_t^{\text{ss}}(\mathbf{c})|$, is considerably more significant than the first term on the r.h.s. and is a much better approximation to the true error on the l.h.s. of (3.14). Of course this may not be the case in general, particularly with portfolios whose P&L is very non-linear in the risk factors $\mathbf{\Delta x}_t$ and where the conditional variance of the non-stressed factors is substantial.

3.4.1 Backtesting Procedure for Evaluating SSA

In our numerical experiments we will simulate a ground truth model for T periods and for each period compute the SSA error as defined in (3.11). We can then average these errors across time to get some idea of how poorly (or well) SSA performs in relation to DFMSA. Since the ground truth model will coincide with the dynamic factor-model that we use to perform the scenario analysis, this approach assumes the estimated P&Ls from the DFMSA are "correct". While of course this is optimistic, it does serve to highlight just how inaccurate the P&Ls reported by SSA can be. It is also worth emphasizing that while we assume we know the *structure* of the ground truth model in

[7] Suppose for example that $\Delta V_t(\cdot)$ is a convex function. Then Jensen's inequality implies $\Delta V_t^{\text{alt}}(\mathbf{c}) = \Delta V_t(\mathbf{B}\boldsymbol{\mu}_t^c) = \Delta V_t(\mathbb{E}_t[\mathbf{B}\mathbf{f}_{t+1} + \boldsymbol{\epsilon}_{t+1} \mid \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}]) \leq \mathbb{E}_t[\Delta V_t(\mathbf{B}\mathbf{f}_{t+1} + \boldsymbol{\epsilon}_{t+1}) \mid \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}] = \Delta V_t^{\text{dfm}}(\mathbf{c})$. In this case $\Delta V_t^{\text{alt}}(\mathbf{c})$ would underestimate the estimated scenario P&L when $\Delta V_t(\cdot)$ is convex. Similarly $\Delta V_t^{\text{alt}}(\mathbf{c})$ would overestimate the estimated scenario P&L when $\Delta V_t(\cdot)$ is concave.

these backtests, we still do not get to observe the latent c.r.f.'s. These latent factor returns must be inferred in our backtests from the risk factor returns, i.e. the $\mathbf{\Delta x}_t$'s, as well as the observable c.r.f. returns. In general, we will also be required to re-estimate the parameters of the model each day within the backtests rather than simply assuming these parameters are given and known to us.

More specifically, in each of our backtests we assume we have T days of simulated data. We choose s where $0 < s < T$ to be the size of the rolling window that we will use to re-estimate the model at each time $t \geq s$. Having estimated the dynamic-model's parameters, we then estimate π_{t+1} and use it to estimate the DFMSA P&L ΔV_t^{dfm}. The SSA P&L ΔV_t^{ss} is also computed at this time. At the end of the backtest we can calculate the average of the backtest P&L for each approach according to

$$\overline{\Delta V}_{\text{dfm}} := \frac{1}{T-s}\sum_{t=s}^{T-1} \Delta V_t^{\text{dfm}} \qquad\qquad \overline{\Delta V}_{\text{ss}} := \frac{1}{T-s}\sum_{t=s}^{T-1} \Delta V_t^{\text{ss}} \tag{3.15}$$

Comparing $\overline{\Delta V}_{\text{ss}}$ with $\overline{\Delta V}_{\text{dfm}}$ gives a measure of the bias of the SSA approach over the course of the backtest. We can also calculate the mean absolute difference between the estimated SSA P&L and the estimated DFMSA P&L. That is we define

$$\text{E}^{\text{abs}} := \frac{1}{T-s}\sum_{t=s}^{T-1} \left|\Delta V_t^{\text{dfm}} - \Delta V_t^{\text{ss}}\right| \tag{3.16}$$

as the average error in the P&L estimated by the SSA approach. Of course this error depends on the ground truth model and its parameters as well as the portfolio and scenario under consideration. Our general back-testing procedure is outlined in Algorithm 1 below.

3.4.2 What Portfolios to Backtest?

Before proceeding to our numerical experiments, it is worth discussing what kinds of portfolios we have in mind when comparing the SSA approach with the DFMSA approach. For all of the reasons outlined earlier we would argue that, regardless of the portfolio, any scenario analysis ought to be embedded in a dynamic factor model setting. Nonetheless, it stands to reason that certain types of portfolios might show little difference between the scenario P&Ls reported by the SSA and DFMSA approaches, respectively. On the other hand, it is not difficult to imagine settings where

Algorithm 1 Backtesting to Estimate Average SSA Error for a Given Scenario and Ground-Truth Model

Input: s, T, K, **gmodel**, $\mathbf{C}, \mathbf{c}$ ▷ s = # periods in rolling window for model training

▷ T = # periods in backtest horizon

▷ K = # of samples used to estimate factor model-based scenario P&L

▷ **gmodel** is the ground-truth model

▷ $\mathbf{c}$, $\mathbf{s}$ define the scenario.

Generate $\mathbf{f}_0$ from **gmodel**
for $t \leftarrow 0$ to $T-1$ **do**
  Generate $(\mathbf{f}_{t+1}, \boldsymbol{\Delta}\mathbf{x}_t) \mid \mathbf{f}_t$ from **gmodel**
  if $t \geq s$ **then**
    Estimate DFM parameters
    Estimate π_{t+1} from $(\mathbf{f}^o_{(t-s):t}, \boldsymbol{\Delta}\mathbf{x}_{(t-s):(t-1)})$ ▷ $\mathbf{f}^o_{t-s:t}$ are observable
    for $k \leftarrow 1$ to K **do**
      Generate $\mathbf{f}^{(k)}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c})$ and $\boldsymbol{\epsilon}^{(k)}_{t+1}$ to obtain $\boldsymbol{\Delta}\mathbf{x}^{(k)}_t$
      Compute scenario P&L $\Delta V_t(\boldsymbol{\Delta}\mathbf{x}^{(k)}_t)$
    end for
    Compute $\Delta V^{\text{dfm}}_t := \sum_{k=1}^{K} \Delta V_t(\boldsymbol{\Delta}\mathbf{x}^{(k)}_t)/K$ ▷ Estimated scenario P&L
    Compute ΔV^{ss}_t ▷ SSA P&L obtained by setting $\boldsymbol{\epsilon}_{t+1}$, non-stressed common factors to $\mathbf{0}$
    Compute $\text{E}^{\text{abs}}_t := \left|\Delta V^{\text{dfm}}_t - \Delta V^{\text{ss}}_t\right|$
  end if
end for
Compute $\overline{\Delta V}_{\text{dfm}}, \overline{\Delta V}_{\text{ss}}$ and E^{abs} as defined in (3.15) and (3.16)

Output: $\overline{\Delta V}_{\text{dfm}}, \overline{\Delta V}_{\text{ss}}$ and E^{abs}

the two scenario P&Ls might be very different. For example, consider a setting with securities whose daily P&L's are non-linear functions of their risk factor changes and where some of the c.r.f. returns are at least moderately[8] dependent. Consider now a portfolio that was designed to be: (i) neutral to the subset of c.r.f.'s that are stressed in scenarios and (ii) highly exposed to the c.r.f.

[8] The assumption that some of the c.r.f. returns might display moderate dependence is not a strong assumption since even uncorrelated c.r.f. returns can display moderate dependence. Suppose for example that the c.r.f. returns have a joint multivariate t distribution with ν degrees-of-freedom. These factor returns can be uncorrelated and yet still have extreme tail dependence [57]. As a result the distribution of these factors conditional on an extreme scenario can display strong dependence.

returns that are never stressed in any of the scenarios. If some of the non-stressed c.r.f. returns are conditionally dependent with some of the stressed c.r.f. returns then such a portfolio should result in very different scenario P&Ls for the SSA and DFMSA approaches.

For an adversarial example, let $\mathbf{f_e}$ where $\mathbf{e} \subset \{1, \ldots, m\}$ denote the subset of the c.r.f.'s to which the p.m. wants to be exposed. It's possible for example that the p.m. has a strong view regarding the direction of $\mathbf{f_e}$ over the short term and wishes to trade on that view. Similarly, let $\mathbf{f_n}$ where $\mathbf{n} \subset \{1, \ldots, m\}$ denote the set of c.r.f.'s to which the trader is required to be neutral according to the risk-management team. We assume the p.m. computes scenario P&Ls using the DFMSA approach uses the SSA approach. The p.m. can then easily construct a risky portfolio that gives her the desired exposure to $\mathbf{f_e}$ but that appears to have little risk according to the risk management team's perspective. If some of the c.r.f. returns in $\mathbf{f_n}$ are dependent (conditional on the scenario) with some of the c.r.f. returns in $\mathbf{f_e}$ then this portfolio should result in very different scenario P&Ls for the SSA and DFMSA approaches. In Appendix B.1 we outline a simple linear programming approach for constructing these portfolio and we will consider them portfolios in our numerical experiments of Sections 3.5 and 3.6.

We also note that this setting is not at all contrived since it is quite possible for a p.m. to have a strong view on a less important risk factor which may not be a risk-factor considered by the risk-management team. Less generously, it may be the case that the p.m. is incentivized to take on a lot of risk regardless of whether or not he / she has a view justifying this risk-taking. Regardless of the p.m.'s motivation, the use of SSA instead of DFMSA can lead to very misleading scenario P&L's.

3.5 An Application to a Portfolio of U.S. Treasury Securities

We now consider a fixed income setting where the p.m. can invest in U.S. treasury securities of n distinct maturities $\tau_1, \ldots, \tau_n$. The risk factor changes for any such portfolio chosen by the p.m. will then be the vector $\mathbf{\Delta x}_t \in \mathbb{R}^n$ whose i^{th} component denotes the change in yield from dates t to $t+1$ of the zero-coupon-bond maturing at time τ_i. Our first step towards specifying a dynamic factor model is to specify the c.r.f's as in (3.2). A principal components analysis (PCA) of yield

curve data suggests there are $m = 3$ c.r.f.'s for the U.S. yield curve and that these factors can explain anywhere from 85% to 95% of the total noise in the yield curve changes. In decreasing order of importance these c.r.f.'s drive parallel, slope and curvature changes, respectively, in the yield curve. To specify a parametric model of these c.r.f.'s we will use the model of Diebold and Li [28] and modify it to include an idiosyncratic noise term $\boldsymbol{\epsilon}_{t+1}$ as in (3.2). The resulting yield curve model can then be written as

$$\Delta x_t(\tau) = f_{1,t+1} + \left(\frac{1 - e^{-\lambda\tau_i}}{\lambda\tau_i}\right) f_{2,t+1} + \left(\frac{1 - e^{-\lambda\tau_i}}{\lambda\tau_i} - e^{-\lambda\tau_i}\right) f_{3,t+1} + \epsilon_{t+1}(\tau) \tag{3.17}$$

where $\Delta x_t(\tau)$ corresponds to the change in yield curve value for maturity τ, $f_{1,t+1}, f_{2,t+1}$ and $f_{3,t+1}$ are the c.r.f. returns, and $\epsilon_{t+1}(\tau)$ is the component of $\boldsymbol{\epsilon}_{t+1}$ corresponding to maturity τ. The parameter λ is a positive scalar that can be chosen[9] to optimize the fit to the yield curve across some time window. The model (3.17) can be written in matrix form $\boldsymbol{\Delta}\mathbf{x}_t = \mathbf{B}\mathbf{f}_{t+1} + \boldsymbol{\epsilon}_{t+1}$ (as in (3.2)) with $b_{i,1} := 1$, $b_{i,2} := (1 - e^{-\lambda\tau_i})/\lambda\tau_i$ and $b_{i,3} := b_{i,2} - e^{-\lambda\tau_i}$ where $b_{i,j}$ denotes the $(i,j)^{th}$ element of $\mathbf{B}$.

It's clear that $\mathbf{b}_1$, the first column of $\mathbf{B}$, can capture parallel changes to the yield curve. For example, a value of $f_{1,t+1} = 1\%$ will result in the entire yield curve increasing by 1%. The second column $\mathbf{b}_2$ captures changes in the slope of the yield curve which are driven by $f_{2,t+1}$. We can see this in the left-hand plot of Figure 3.1 below where we see that the loadings are monotonically decreasing in τ. This means, for example, that if $f_{2,t+1} = 1\%$, for example, then short-term yields will increase considerably more than long-term yields thereby reducing the slope of the yield curve. If the current yield curve happened to be upward-sloping then this would result in a flattening of the yield curve. The third column $\mathbf{b}_3$ captures changes in the curvature of the yield curve which are driven by $f_{3,t+1}$. We can see this in the right-hand plot of Figure 3.1 where we see that the loadings are monotonically increasing in τ for the first few years after which they are monotonically decreasing. Shocks to $f_{3,t+1}$ will therefore change the curvature of the current yield curve.

Of course the c.r.f. returns are latent and so we will use the state-space model of (3.7) together with the observation process (3.17) to complete the specification of our model. Specifically, we will

[9] Diebold and Li [28] chose a value of $\lambda = 0.7308$ for the US Treasury yield curve.

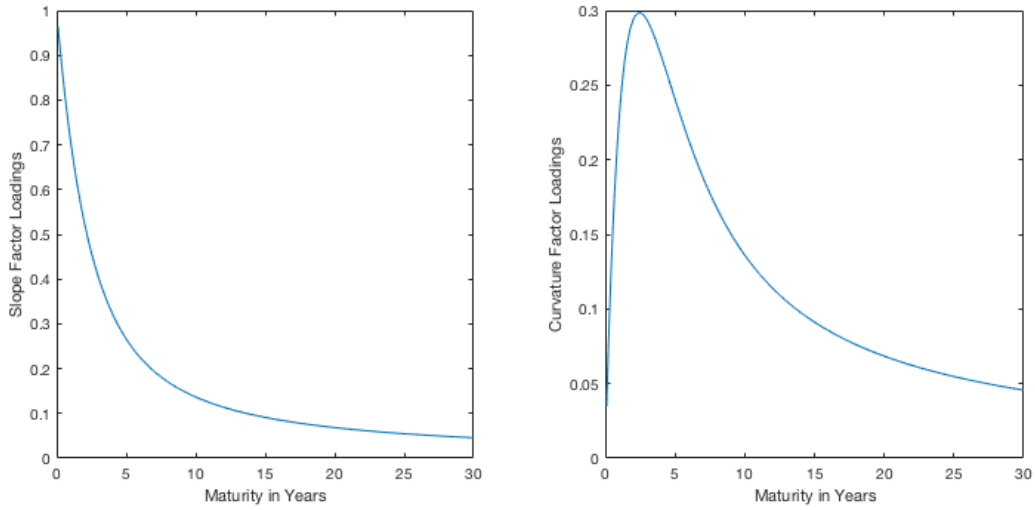


Figure 3.1: Factor Loadings for the Diebold factor model.

use a linear Gaussian state-space model as in (3.7) and assume that $\boldsymbol{\eta}_t$, $\boldsymbol{\epsilon}_t$ and $\mathbf{f}_0$ are normally distributed.

3.5.1 Model Calibration and Backtesting

In order to backtest our model we obtained US Treasury yield data from January 2008 through December 2017 for $n = 11$ maturities: 1, 3 and 6 months, and 1, 2, 3, 5, 7, 10, 20 and 30 years. We take our ground-truth model to be the model we obtain by using the EM algorithm to fit the linear Gaussian state-space model of (3.7) and (3.17) to the aforementioned yield curve data. The estimated parameters of the ground-truth model are provided in Appendix B.2.

For each day of our backtest we construct a portfolio using the linear programming approach described in Section 3.4.2 and Appendix B.1. The securities used to build the portfolio are zero-coupon risk-free bonds for the $n = 11$ maturities listed above as well as a risk-free cash account – the $(n+1)^{st}$ security – that each day returns 0% w.p. 1. We include a cash-account because it is realistic – p.m.s always have the option to take on zero risk by keeping their funds in cash – and it also provides a simple guarantee that there is a feasible portfolio, i.e. 100% in the cash-account, that satisfies all of the risk-constraints.

We consider a p.m. that on each day t believes $f_{1,t+1}$ will equal -12 basis points (1 b.p. = .01%) and that $f_{2,t+1}$ will equal -16 b.p.s. The p.m. is therefore always anticipating a parallel fall in the yield curve combined with an increase in it's slope. Note that the magnitude of these movements in the c.r.f. is given by the corresponding columns of $\mathbf{B}$ in the factor model (3.17). For example, if i corresponds to the 30 year maturity then $b_{i,1} = 1$ and $b_{i,2} = 0.05$ (see Appendix B.1) so that the resulting move in the 30-year yield is $-1 \times 0.12 - 0.05 \times 0.16 = -0.128$, i.e. a fall of 12.8 b.p.s. (This assumes the third c.r.f. return $f_{3,t+1}$ and $\boldsymbol{\epsilon}_{t+1}$ are both zero.) These anticipated movements correspond to -2 standard deviation moves in each of the first two c.r.f.s. and the p.m. wishes to construct[10] her portfolio to maximize her P&L with this view in mind.

We assume: (i) the p.m. can take on short positions so that w_i can be negative for each i and (ii) a leverage limit of 10 on each risky security so that $-10 \leq w_i \leq 10$ for each i. In addition to these constraints we assume the risk-management desk requires the p.m.'s portfolio to be "neutral" with respect to several scenarios involving joint stresses to pairwise combinations of the three c.r.f.s. They define "neutral" in such a way that the SSA P&L for the specified scenarios must be within $\pm\alpha = 3\%$ of the value of the portfolio at time t. More specifically, each scenarios is given by an element of the cross-product of $\Omega_{\text{ParallelShift}} \times \Omega_{\text{Slope}}$ or $\Omega_{\text{ParallelShift}} \times \Omega_{\text{Curvature}}$ where

$$\begin{aligned} \Omega_{\text{ParallelShift}} &:= \{-24, -12, 0, 12, 24\} \\ \Omega_{\text{Slope}} &:= \{-32, -16, 0, 16, 32\} \\ \Omega_{\text{Curvature}} &:= \{-64, -32, 0, 32, 64\}. \end{aligned} \tag{3.18}$$

The values in $\Omega_{\text{ParallelShift}}$, Ω_{Slope} and $\Omega_{\text{Curvature}}$ were calibrated to be approximately $0, \pm 2$ and ± 4 standard deviations of the three c.r.f. returns, respectively and their units are b.p.s. Once the portfolio has been constructed we then apply SSA and DFMSA on it using the following scenarios:

1. Simultaneous stresses to the parallel shift and slope c.r.f. returns, with shocks in the cross-product of $\Omega_{\text{ParallelShift}}$ and Ω_{Slope},

[10] It's worth emphasizing that our back-tests are not at all concerned with why the p.m. has this particular view or whether or not it is ever justified. The view is simply used to construct a portfolio to which we then apply SSA and DFMSA.

2. Simultaneous stresses to the parallel shift and curvature c.r.f. returns, with shocks in the cross-product of $\Omega_{\text{ParallelShift}}$ and $\Omega_{\text{Curvature}}$,

Note that the same set of scenarios are used to both construct the portfolio (via constraints on the LP) and analyze the risk of the portfolio. This of course makes sense since the constraints in the LP are driven by the scenario analysis that the risk-management desk routinely performs.

We back-tested the model using Algorithm 1 from Section 3.4.1 and where we used the ground-truth model to simulate data for a backtest horizon of $T = 1,000$ days. We set the training window to be of size $s = 500$ days. For each time $t \in \{s, \dots, T-1\}$, we use the EM algorithm on the observable simulated data $\mathbf{\Delta x}_{t-s:t-1}$ to re-estimate the model parameters $\mathbf{G}$, $\mathbf{\Sigma}_\eta$, $\mathbf{\Sigma}_\epsilon$ as well as the parameters of the normal distribution π_{t-s} governing the initial state $\mathbf{f}_{t-s}$. Once the model has been (re-)trained at time t we can use the Kalman filter to calculate the mean vector and covariance matrix of the distribution of $\mathbf{f}_t \mid \mathbf{\Delta x}_{t-s:t-1}$. Given the c.r.f. return dynamics in (3.7), it then follows that π_{t+1} is the convolution of the distribution of the Gaussian random variables $\mathbf{G}\mathbf{f}_t \mid \mathbf{\Delta x}_{t-s:t-1}$ and $\boldsymbol{\eta}_{t+1}$ and is therefore also Gaussian. Note that, even though we simulate the c.r.f. returns from the ground truth model in step 3 of Algorithm 1, these are assumed unobservable and are therefore not used by the EM algorithm to re-estimate the model parameters in step 5 of the algorithm. The SSA and DFMSA approaches are then implemented in the remaining steps of the algorithm.

3.5.2 Numerical Results

Tables 3.1 to 3.3 display the results of our backtest. Table 3.1 shows the average backtested P&L $\overline{\Delta V}_{\text{ss}}$ as reported by the SSA approach. On each day of the backtest the portfolio was constructed in such a way that the SSA loss conditional on the given scenario would be within $\pm\alpha = 3\%$. It is therefore no surprise to see that the average-backtested P&L numbers are also within $\pm 3\%$ and so this portfolio strategy appears to have relatively little risk. In contrast Table 3.2 displays the true average backtested expected P&L $\overline{\Delta V}_{\text{dfm}}$ conditional on the given scenario. These P&L numbers were computed using the DFMSA approach and we can see from them that the portfolio is not "neutral" w.r.t. the specified scenarios. For example, when the slope c.r.f. is shocked by 32 b.p.s and the parallel c.r.f. return remains flat, the SSA approach yields a 2.6% loss whereas the

	$\overline{\Delta V}_{\text{ss}}$									
	(a) Slope (bps)					(b) Curvature (bps)				
Parallel Shift (bps)	-32	-16	0	16	32	-64	-32	0	32	64
-24	3.0	1.7	0.4	-0.9	-2.2	-2.3	-1.0	0.4	1.7	3.0
-12	2.8	1.5	0.2	-1.2	-2.5	-2.5	-1.2	0.2	1.5	2.8
0	2.6	1.3	0.0	-1.3	-2.6	-2.6	-1.3	0.0	1.3	2.6
12	2.5	1.2	-0.1	-1.4	-2.7	-2.7	-1.4	-0.1	1.2	2.5
24	2.5	1.2	-0.1	-1.4	-2.7	-2.7	-1.4	-0.1	1.2	2.4

Table 3.1: Average of backtest SSA P&L $\overline{\Delta V}_{\text{ss}}$ (defined in (3.15)) for a portfolio that is constructed to have: (i) exposure to negative changes to the parallel and slope c.r.f. returns and (ii) to be approximately neutral (max. loss within $\pm\alpha := 3\%$ according to SSA) with respect to the pre-specified scenarios in the table. Subtable (a) displays the average SSA P&L when simultaneously stressing the parallel and slope c.r.f. returns. Subtable (b) displays the average SSA P&L when simultaneously stressing the parallel and the curvature c.r.f. returns. All P&L numbers are in dollars per \$100 of face value of the portfolio. The portfolio is constructed anew on each day of the back-test period.

DFMSA approach yields a 4.8% loss. We see that the differences between the two approaches can differ by up to a factor of 3. Moreover, it's possible for SSA to report a scenario loss while DFMSA reports an expected scenario profit and vice versa. We also note that it's possible to obtain more extreme discrepancies between the two approaches. For example, we could have the p.m. take a more extreme view on the parallel and slope c.r.f. returns or have her take a view on the slope and curvature c.r.f. returns. Joint movements of these two c.r.f. returns are not considered in any of the scenarios and so a view on these two c.r.f. returns might allow the p.m. to better game the risk-management constraints.

Table 3.3 displays the mean absolute error E^{abs} (as defined in (3.16)) of the SSA approach. Once again we observe the large errors produced by SSA. The largest error shown is 4.1% for the scenario in which the slope and parallel c.r.f. returns are stressed to -32 bps and -24bps, respectively.

	$\overline{\Delta V}_{\text{dfm}}$									
	(a) Slope (bps)					(b) Curvature (bps)				
Parallel Shift (bps)	-32	-16	0	16	32	-64	-32	0	32	64
-24	7.1	4.7	2.3	-0.1	-2.4	-5.3	-3.4	-1.7	0.0	1.7
-12	6.0	4.2	1.2	-1.3	-3.6	-4.2	-2.6	-0.9	0.9	2.6
0	5.0	2.4	0.1	-2.4	-4.8	-3.4	-1.8	0.0	1.7	3.4
12	3.7	1.5	-1.1	-3.3	-5.8	-2.5	-0.7	1.0	2.8	4.4
24	2.9	0.6	-1.9	-4.3	-6.6	-1.3	0.3	2.0	3.7	5.5

Table 3.2: Average of backtest DFMSA P&L $\overline{\Delta V}_{\text{dfm}}$ for the same portfolio and scenarios as reported in Table 3.1. All P&L numbers are in dollars per $100 of face value of the portfolio.

	E^{abs}									
	(a) Slope (bps)					(b) Curvature (bps)				
Parallel Shift (bps)	-32	-16	0	16	32	-64	-32	0	32	64
-24	4.1	3.0	2.0	0.9	0.5	3.0	2.4	2.1	1.7	1.3
-12	3.2	2.8	1.1	0.4	1.1	1.9	1.5	1.1	0.6	0.4
0	2.3	1.1	0.4	1.1	2.2	0.8	0.6	0.3	0.6	0.9
12	1.2	0.4	1.0	2.0	3.1	0.5	0.7	1.1	1.6	2.0
24	0.5	0.7	1.8	2.9	3.9	1.4	1.7	2.1	2.5	3.0

Table 3.3: Average backtest error E^{abs} of the SSA P&L for the same portfolio and scenarios as in Tables 3.1 and 3.2. E^{abs} is defined in (3.16).

3.6 An Application to an Equity Options Portfolio

In this application we consider a p.m. that can invest in European call and put options on the S&P 500 index as well as in the index itself. As is standard market practice, we will use the Black-Scholes formula to price these options. We will therefore[11] assume the vector of risk factor changes $\mathbf{\Delta x}_t$ to consist of the daily log-return of the S&P 500 together with daily changes in the

[11] We assume the risk-free rate of interest and dividend yield remain constant throughout and therefore do not model risk factors associated with them. This is typical for equity options setting unless the p.m. wishes to trade with a specific view on dividends. We also acknowledge that in practice one trades futures on the S&P 500 index rather than the index itself. Given the assumption of a constant risk-free rate and dividend yield, there is essentially no difference in assuming we can trade the index itself, however, and so we will make that assumption here.

implied volatilities of specific strike-maturity combinations. More precisely, we let $I_t(\xi, \tau)$ denote the implied volatility at time t of a European option with time-to-maturity τ and option moneyness $\xi := K/S_t$ where K denotes the option strike and S_t is the time t price of the S&P 500. We assume that on each day we can observe the implied volatility surface at a finite set of moneyness-maturity pairs $\{(\xi_1, \tau_1), \ldots, (\xi_{n-1}, \tau_{n-1})\}$. For a fixed pair (ξ, τ), we denote the change in implied volatility from t to $t+1$ by

$$\Delta I_t(\xi, \tau) := I_{t+1}(\xi, \tau) - I_t(\xi, \tau). \tag{3.19}$$

The risk factors changes is then given by the n-dimensional vector

$$\mathbf{\Delta x}_t := (\log(S_{t+1}/S_t), \Delta I_t(\xi_1, \tau_1), \ldots, \Delta I_t(\xi_{n-1}, \tau_{n-1}))^\top \tag{3.20}$$

where the moneyness-maturity pairs in (3.20) cover the distinct moneyness-maturity combinations of the options in the market.

Our dynamic factor model will consist of four c.r.f.'s. Naturally we will take the daily log-return of the S&P 500 to be the first c.r.f. and of course this is observable. The other $m = 3$ c.r.f.'s will be latent factors that drive changes in the implied volatility surface, specifically parallel[12] changes in the surface, a steepening / flattening of the volatility skew, and a steepening / flattening of the term structure. As our model will contain both observable and latent c.r.f.s we will proceed as discussed in Section 3.3.2 and use a linear Gaussian state-space model. In particular, we will use a slightly modified version of (3.9) and define

$$\mathbf{\Delta x}_t = \begin{bmatrix} 1 \\ \mathbf{b}^o \end{bmatrix} f^o_{t+1} + \begin{bmatrix} \mathbf{0}_3^\top \\ \mathbf{B}^u \end{bmatrix} \mathbf{f}^u_{t+1} + \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_{t+1} \end{bmatrix} \tag{3.21}$$

where $f^o_{t+1} := \log(S_{t+1}/S_t)$ is the observable c.r.f. (and coincides with the first component of $\mathbf{\Delta x}_t$), $\mathbf{f}^u_{t+1} \in \mathbb{R}^3$ denotes the vector of latent c.r.f.s and $\mathbf{0}_3 \in \mathbb{R}^3$ denotes the zero vector. The

[12] We acknowledge that the absence of arbitrage imposes restrictions on the magnitude of permissable c.r.f. stresses. For example, Rogers [70] has shown that the implied volatility surface cannot move in parallel without introducing arbitrage opportunities. Indeed it is well known that moves in the implied volatilities are more likely to follow a "square-root-of-time" rule and we will model this below with our first latent c.r.f. For another example, it is also well-known that that volatility cannot become too steep without introducing arbitrage. We don't explicitly rule out scenarios that allow for arbitrage but note that such scenarios would have to be very extreme indeed. Moreover, it is easy to check a given scenario for arbitrage and so ruling out such scenarios would be very straightforward.

factor loadings for the observable and latent c.r.f.s are denoted by $\mathbf{b}^o \in \mathbb{R}^{n-1}$ and $\mathbf{B}^u \in \mathbb{R}^{(n-1)\times 3}$, respectively. The i^{th} element of $\mathbf{b}^o$ indicates how a shock to the S&P 500 affects the implied volatility $\Delta I_t(\xi_i, \tau_i)$. The matrix $\mathbf{B}^u$ is constructed to model the aforementioned $m = 3$ types of stresses on the implied volatility surface. Specifically (and recalling that the $(i+1)^{st}$ component of $\mathbf{\Delta x}_t$ is $\Delta I_t(\xi_i, \tau_i)$), we assume

$$\Delta I_t(\xi_i, \tau_i) = b_i^o f_{t+1}^o + \left(\frac{1}{\sqrt{\tau_i}}\right) f_{1,t+1}^u + (1-\xi_i)\, f_{2,t+1}^u + \ln(2\tau_i) f_{3,t+1}^u + \epsilon_{i,t+1}, \tag{3.22}$$

for $i = 1, \ldots, n-1$, where b_i^o, $f_{i,t+1}^u$ and $\epsilon_{i,t+1}$ denote the i^{th} elements of $\mathbf{b}^o$, $\mathbf{f}_{t+1}^u$ and $\boldsymbol{\epsilon}_{t+1}$, respectively, in (3.21). Comparing (3.21) and (3.22), we see that $b_{i,1} := 1/\sqrt{\tau_i}$, $b_{i,2} := 1 - \xi_i$ and $b_{i,3} := \ln(2\tau_i)$ where $b_{i,j}$ is the $(i,j)^{th}$ element of $\mathbf{B}^u$.

A few comments on (3.22) are now in order. It is well known (see for example Natenberg [60]) that when volatility rises (falls), the implied volatility of long-term options rises (falls) less than the implied volatility of short-term options. This empirical observation has led to the commonly used "square-root-of-time" rule whereby the relative difference in implied volatility changes for options with the same moneyness but different maturities is in proportion to the square-roots of their relative maturities. We model this rule via the factor loadings for the parallel-shift c.r.f. Suppose, for example, there is a $f_{1,t+1}^u = 1$ volatility point shock to the parallel-shift c.r.f. Then the implied volatility of 1-year options would increase by 1 point exactly, whereas the implied volatility of a 1-month option would increase by $1/\sqrt{1/12} \approx 3.46$ points.

The second latent c.r.f. is used to drive changes in the implied volatility skews[13] in the surface. We use the so-called "sticky-moneyness" rule which assumes that, for a given maturity, the implied volatility is a univariate function of the moneyness $\xi = K/S$. The "sticky-moneyness" rule that we adopt in (3.22) can be motivated by first assuming

$$I_t(\xi, \tau) = I_0(1, \tau) - \beta_t\,(\xi - 1) \tag{3.23}$$

where $I_0(1, \tau)$ is the implied volatility of an at-the-money option, i.e. with $\xi = 1$, with maturity

[13] An implied volatility skew is the cross-section of the implied volatility surface that we obtain when we hold the time-to-maturity fixed. There is therefore a different skew for each time-to-maturity. There are various skew models in the literature and we refer the interested reader to the work of Derman and Miller [26] who describe some of these models.

τ at some initial time $t = 0$, and where β_t determines the slope of the skew at time t. The model (3.23) implies the implied volatility for at-the-money options remains constant for a given maturity τ, and that changes in implied volatility are given by

$$\Delta I_t(\xi, \tau) = -\Delta\beta_t \, (\xi - 1) \tag{3.24}$$

where $\Delta\beta_t := \beta_{t+1} - \beta_t$ defines the change in skew (or slope) of the implied volatility. We account for this skew behavior in our factor model (3.21) by taking $\Delta\beta_t$ to be the c.r.f. $f^u_{2,t+1}$ and setting the corresponding factor loadings to $(1-\xi)$. Then if $f^u_{2,t+1} > 0$, for example, the implied volatilities of options with moneyness < 1 (> 1) would increase (decrease) thereby resulting in the steepening of the skew for any given maturity τ. Similarly a shock $f^u_{2,t+1} < 0$ would result in a flattening of the skew.

The third c.r.f. $f^u_{3,t+1}$ models changes to the term-structure of implied volatility for any given level of moneyness. The loading term $\ln(2\tau_i)$ means that a positive shock to $f^u_{3,t+1}$ would leave 6-month volatilities unchanged, but would increase (decrease) the volatilities of options with longer (shorter) maturities thereby resulting in the flattening of an inverted term structure or steepening of an already upward sloping term structure. We note that the parallel shift c.r.f. also affects the term structure due to the square-root-of-time rule. However, including the term structure c.r.f. enriches the dynamics of the volatility surface model as it allows for a broader variety of systematic moves, i.e. moves driven only by the c.r.f.s.

Finally, we note that in this section we are neither arguing for or against the specific model of (3.7) and (3.22). We are merely using this model as an example for demonstrating the DFMSA approach where we assume the ground truth model coincides with (3.7) and (3.22). Whether or not the model would work well in practice (where we wouldn't know the ground truth model) would depend on its ability to pass the various statistical tests outlined in Section 3.7.

3.6.1 Model Calibration

We obtained implied volatility data on the S&P 500 for the period January 2006 through August 2013 from the OptionMetrics IVY database. In particular, we used the daily implied volatility data

that OptionMetrics provide for various delta-maturity[14] combinations. We transformed the data to moneyness-maturity coordinates using a non-parametric Nadaraya-Watson estimator based on an independent bivariate Gaussian kernel [32]. We can therefore obtain the implied volatilities on any given day for the fixed set of moneyness-maturity pairs given by the cross-product of

$$\xi \in \Omega_\xi := \{0.8,\, 0.9,\, 0.95,\, 1.0,\, 1.05,\, 1.1,\, 1.2\} \quad \text{and} \quad \tau \in \Omega_\tau := \{1/12,\, 2/12,\, 3/12,\, 6/12,\, 1\}\,, \tag{3.25}$$

where the time-to-maturity τ is measured in years. We then used this data to fit the linear Gaussian state-space model of (3.7) and (3.21) via the EM algorithm. We take our ground-truth model to be the resulting fitted model. The parameters of this ground-truth model are given in Appendix B.3.

We assume our portfolio can contain the S&P 500 index, at-the-money and out-of-the-money call options with moneyness ($\xi = K/S$) in the set $\{1.00, 1.025, 1.05, 1.075, 1.10, 1.15\}$ and out-of-the-money put options with moneyness in the set $\{0.85, 0.90, 0.925, 0.95, 0.975\}$. The options are assumed to have maturities in the set $\Omega'_\tau := \{i/12 : i = 1, \ldots, 12\}$ so there are a total of $N = 133$ securities in the universe. Each option is priced using the Black-Scholes formula and so we interpolate the implied volatility surface as necessary to obtain the implied volatility for certain moneyness-maturity pairs that are not explicitly modeled. As was the case in Section 3.5.1, we again assume that a risk-free cash account is also available. The cash account is the $(N+1)^{st}$ security and each day it returns 0% w.p. 1.

On each day of our backtest, we construct a portfolio using the LP approach as described in Section 3.4.2 and Appendix B.1. We consider a p.m. who on each day t believes (i) the S&P 500 will fall by 3% and (ii) the parallel shift c.r.f. will increase by 1 volatility point. We note from (3.21) that a 1 volatility point increase in the parallel shift c.r.f. would translate to a $1/\sqrt{\tau}$ volatility points increase for options with maturity τ, assuming the idiosyncratic noise and other c.r.f. returns were zero. For example, a 1-month option would then see a $1/\sqrt{1/12} = 3.46$ volatility points increase. These anticipated movements correspond to -2 and +2 standard deviation moves

[14] Roughly speaking, they build an implied volatility surface based on each day's closing prices (of the S&P 500 and its traded options) and then use this surface to read off volatilities for the various delta-maturity combinations.

in the two c.r.f.s, respectively.

We assume the p.m. can take short positions on any of the securities except for the cash account so that $0 \leq w_{N+1} \leq 1$. We also assume that $-0.3 \leq w_i \leq 0.3$ for $i = 1, \ldots, N$ so that the risk in any one security is limited. We also have the budget constraint $\sum_{i=1}^{N+1} w_i = 1$. In addition to these constraints, we assume that the risk-management desk requires portfolios to be kept "neutral" with respect to a given set of scenarios involving joint stresses to pairwise combinations of the first three c.r.f. returns, i.e., the S&P 500, the parallel shift and skew c.r.f. returns. Neutrality to a given scenario is defined as having the portfolio SSA P&L under that scenario to be within $\pm\alpha = 2\%$ of the initial portfolio value. The given scenarios are the elements of the cross-product of $\Omega_{\text{Mkt}} \times \Omega_{\text{ParallelShift}}$ or $\Omega_{\text{Mkt}} \times \Omega_{\text{Skew}}$, where

$$\begin{aligned} \Omega_{\text{Mkt}} &:= \{-4.5\%, -3.0\%, -1.5\%, 0.0\%, 1.5\%, 3.0\%, 4.5\%\} \\ \Omega_{\text{ParallelShift}} &:= \{-1.5,\ -1.0,\ -0.5,\ 0.0,\ 0.5,\ 1.0,\ 1.5\} \\ \Omega_{\text{Skew}} &:= \{-21,\ -14,\ -7,\ 0,\ 7,\ 14,\ 21\}. \end{aligned} \tag{3.26}$$

The values in $\Omega_{\text{Mkt}}, \Omega_{\text{ParallelShift}}$ and Ω_{Skew} were calibrated to be approximately $0, \pm 1, \pm 2$ and ± 3 standard deviations of the S&P 500, parallel shift and skew c.r.f. daily returns, respectively. The shocks in Ω_{Mkt} are in log-return percentage changes while the units of $\Omega_{\text{ParallelShift}}$ and Ω_{Skew} are volatility points. Recall that the magnitude of these movements in the c.r.f.s is given by the corresponding columns of $\mathbf{b}^o$ and $\mathbf{B}^u$ in the factor model (3.21). For example, a -21 unit move in the skew c.r.f. return translates to a $(1 - \xi) \times -21$ move in the implied volatility for options with moneyness ξ assuming the idiosyncratic noise and other c.r.f. returns were zero. This translates to a decrease in implied volatility of 2.1 volatility points for options with $\xi = 0.9$ and an increase of 3.15 volatility points for options with $\xi = 1.15$. Finally, we also impose the (linear equality) constraints that require the portfolio to be delta, gamma and vega neutral. We impose the latter constraints to allow for the fact that SSA is typically not done in isolation and so it would be typical for any risk manager / p.m. to also know the delta, gamma and vega of the portfolio. By insisting that the portfolio be neutral to the Greeks we are simply making it more difficult for the p.m. to "game" the fact that the scenario constraints are based on SSA rather than the correct

DFMSA.

On each day of the back-test we constructed the portfolio and then apply SSA and DFMSA to it using the same scenarios that we used to define the scenario constraints on the portfolio, i.e. scenarios corresponding to shocks in the cross-product of Ω_{Mkt} and $\Omega_{\text{ParallelShift}}$ and Ω_{Mkt} and Ω_{Skew}. We back-tested the model using Algorithm 1 and using the ground-truth model to simulate data for a backtesting horizon of $T = 1,000$ periods. We used an initial training window of size $s = 500$ and then for each $t \in \{s, \ldots, T-1\}$ we used the EM algorithm on the observable simulated data $\mathbf{\Delta x}_{(t-s):(t-1)}$ to re-estimate the model parameters $\mathbf{b}^o, \mathbf{G}, \mathbf{\Sigma}_{\boldsymbol{\eta}}$ and $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$, as well as the parameters of the normal distribution governing the initial state $\mathbf{f}_{t-s}$. Having re-trained the model at time t, we use the Kalman Filter to obtain the distribution of $\mathbf{f}_t \mid \mathbf{\Delta x}_{t-s:t-1}$. We finally obtain π_{t+1} as the distribution of the convolution of $\mathbf{G}\mathbf{f}_t \mid \mathbf{\Delta x}_{(t-s):(t-1)}$ and $\boldsymbol{\eta}_{t+1}$ and simulate samples from π_{t+1} to estimate the scenario P&L's under both SSA and DFMSA approaches.

3.6.2 Numerical Results

The results of the backtest are displayed in Tables 3.4 to 3.6 below. Table 3.4 contains the average P&L $\overline{\Delta V}_{\text{ss}}$ as estimated by the SSA approach for the same set of scenarios that were used to construct the portfolio. As a result, it is not surprising to see the reported P&L's are all less than 2% in absolute value. The average estimated P&L $\overline{\Delta V}_{\text{dfm}}$ obtained using the DFMSA approach is then reported in Table 3.5 for the same set of scenarios. It is very clear that the portfolio is (on average) not at all neutral to the various scenarios. For example, when the market c.r.f. return and the parallel shift c.r.f. are shocked by +3% and -1, respectively, the DFMSA approach estimates a loss of 8.2% whereas the SSA approach estimates a loss of just 1.3%. Similarly, a shock to the skew c.r.f. of -21 yields an estimated 4.3% loss under the DFMSA approach whereas the SSA approach only yields a loss of 0.1%.

It is also clear from Table 3.5 that, as designed, the portfolio has the correct directional exposure to positive moves in the parallel shift c.r.f. and negative moves in the market c.r.f. Furthermore, the portfolio also reacts positively to positive moves in the skew c.r.f. This can be explained by observing the correlations between the skew and the parallel shift c.r.f. returns as reported in

	$\overline{\Delta V}_{\text{ss}}$													
	(a) Parallel Shift							(b) Skew						
Mkt	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	-21	-14	-7	0	7	14	21
-4.5%	1.1	1.5	1.7	1.8	1.9	1.9	2.0	1.0	1.5	1.6	1.8	1.9	1.7	1.6
-3.0%	0.5	0.8	1.1	1.4	1.7	1.9	1.9	0.7	1.2	1.5	1.5	1.4	1.2	1.2
-1.5%	-0.8	-0.5	0.1	0.5	0.9	1.1	1.2	0.2	0.3	0.3	0.4	0.4	0.3	0.2
0.0%	-1.2	-0.8	-0.3	0.2	0.6	0.8	1.0	-0.1	-0.1	0.0	0.0	0.1	0.2	0.2
1.5%	-1.7	-1.2	-0.5	-0.1	0.1	0.3	0.6	-0.9	-0.6	-0.4	-0.1	0.0	0.1	0.1
3.0%	-1.8	-1.3	-0.5	0.0	0.3	0.4	0.7	-1.2	-0.6	-0.2	0.0	0.3	0.8	1.2
4.5%	-1.7	-1.1	-0.3	0.2	0.5	0.7	1.0	-1.3	-0.5	0.0	0.4	0.9	1.2	1.5

Table 3.4: Average of backtest SSA P&L $\overline{\Delta V}_{\text{ss}}$ (defined in (3.15)) for a portfolio that is constructed to have: (i) exposure to negative changes to the market (S&P index) c.r.f. returns and exposure to positive changes to the parallel shift c.r.f. returns and (ii) to be approximately neutral (max. loss within $\pm\alpha := 2\%$ according to SSA) with respect to the pre-specified scenarios in the table. Subtable (a) displays the average SSA P&L when simultaneously stressing the market and parallel shift c.r.f. returns. Subtable (b) displays the average SSA P&L when simultaneously stressing the market and skew c.r.f. returns. All P&L numbers are in dollars per \$100 of face value of the portfolio. The portfolio is constructed anew on each day of the back-test period.

Appendix B.3. Specifically, the skew c.r.f. return is positively correlated with the parallel shift c.r.f. (0.3002 correlation) and has close to zero correlation to the market c.r.f. return (0.0236 correlation). Since the portfolio is positively exposed to shocks in the parallel shift c.r.f. return it is therefore no surprise to see the portfolio is also positively exposed to positive shocks to the skew c.r.f. too. This of course is not captured by the SSA results in Table 3.4.

Motivated by the (partial) error decomposition in (3.14) from Section 3.4, we define

$$\mathrm{E}^{\text{abs}}_{\text{cond}} := \frac{1}{T-s}\sum_{t=s}^{T-1}\left|\Delta V_t^{\text{alt}} - \Delta V_t^{\text{ss}}\right| \qquad \mathrm{E}^{\text{abs}}_{\text{vol}} := \frac{1}{T-s}\sum_{t=s}^{T-1}\left|\Delta V_t^{\text{dfm}} - \Delta V_t^{\text{alt}}\right| \tag{3.27}$$

where ΔV_t^{alt} is defined in (3.12) and is our alternative estimated scenario P&L. We obtain ΔV_t^{alt} by setting $\mathbf{\Delta x}_t = \mathbf{B}\boldsymbol{\mu}_t^c$ where $\boldsymbol{\mu}_t^c$ (defined in (3.13)) is the expected value of the c.r.f. returns

	$\overline{\Delta V}_{\text{dfm}}$													
	(a) Parallel Shift							(b) Skew						
Mkt	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	-21	-14	-7	0	7	14	21
-4.5%	-9.1	-5.6	-1.5	2.4	5.9	9.8	13.1	-2.5	-0.7	0.5	1.7	2.9	4.7	6.1
-3.0%	-9.6	-5.9	-1.9	1.5	5.5	9.6	12.7	-3.4	-1.5	0.1	1.5	2.7	4.0	5.5
-1.5%	-10.8	-6.7	-3.5	0.4	4.1	8.9	11.2	-4.1	-2.4	-0.8	0.3	1.8	3.2	4.5
0.0%	-11.2	-7.4	-3.8	-0.1	3.6	8.2	11.1	-4.3	-2.9	-1.3	-0.1	1.7	2.5	4.2
1.5%	-11.7	-7.9	-4.2	-0.4	3.4	6.9	10.5	-4.7	-3.5	-2.2	-0.4	1.2	2.4	4.5
3.0%	-12.5	-8.2	-4.5	-0.5	3.5	7.9	11.9	-5.3	-3.8	-2.0	0.1	1.5	3.3	5.1
4.5%	-12.7	-8.9	-4.7	-0.4	3.5	8.4	12.4	-5.8	-3.9	-1.9	0.5	2.4	3.7	5.9

Table 3.5: Average of backtest DFMSA P&L $\overline{\Delta V}_{\text{dfm}}$ for the same portfolio and scenarios as reported in Table 3.4. All P&L numbers are in dollars per $100 of face value of the portfolio.

conditional on the scenario. It follows from the triangle inequality in (3.14) for each $t = s, \ldots, T-1$ that $\mathrm{E}^{\text{abs}} \leq \mathrm{E}^{\text{abs}}_{\text{cond}} + \mathrm{E}^{\text{abs}}_{\text{vol}}$.

Tables 3.6 and 3.7 display the average values of $\mathrm{E}^{\text{abs}}_{\text{vol}}$ and $\mathrm{E}^{\text{abs}}_{\text{cond}}$, respectively, in our backtest. It is clear from Table 3.6 that the error in reported P&L's that results from using the alternative ΔV^{alt}_t is relatively small and is less than 1% in all of considered scenarios. In contrast, the errors in Table 3.7 are significantly larger. These observations suggest (at least in this example), that the main source of error in the SSA approach is in setting the non-stressed factors to zero rather than their expectations conditional on the given scenario.

We can also observe from Table 3.7 (a) that the largest absolute errors occur when the parallel shift c.r.f. return is subjected to the most extreme shocks. This indicates that setting the skew and term-structure c.r.f. returns to zero (which is how SSA would proceed) results in higher errors when the parallel shift c.r.f. return is more severely stressed. Referring to the c.r.f. return correlations that are reported in Appendix B.3, we see this observation can be explained by noting that the parallel shift c.r.f. return is strongly correlated with the term-structure c.r.f. (0.9283 correlation) and is moderately correlated with the skew c.r.f. return (0.3002 correlation). Clearly setting the term-structure and skew c.r.f. returns to zero would be highly inaccurate in this setting.

	E^{abs}_{vol}													
	(a) Parallel Shift							(b) Skew						
Mkt	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	-21	-14	-7	0	7	14	21
-4.5%	0.6	0.6	0.5	0.4	0.4	0.7	0.7	0.8	0.7	0.6	0.7	0.7	0.5	0.7
-3.0%	0.5	0.5	0.5	0.6	0.6	0.5	0.6	0.8	0.7	0.6	0.6	0.7	0.5	0.6
-1.5%	0.5	0.5	0.4	0.5	0.6	0.5	0.4	0.7	0.6	0.6	0.6	0.6	0.6	0.7
0.0%	0.5	0.4	0.5	0.5	0.4	0.6	0.5	0.5	0.6	0.7	0.7	0.6	0.6	0.6
1.5%	0.4	0.5	0.4	0.5	0.5	0.5	0.4	0.6	0.7	0.5	0.6	0.6	0.7	0.6
3.0%	0.5	0.5	0.5	0.4	0.6	0.5	0.5	0.6	0.5	0.6	0.7	0.7	0.5	0.5
4.5%	0.5	0.6	0.5	0.5	0.5	0.6	0.4	0.7	0.7	0.6	0.5	0.6	0.7	0.7

Table 3.6: Average backtest error E^{abs}_{vol} of the SSA P&L for the same portfolio and scenarios as in Tables 3.4 and 3.5.

	E^{abs}_{cond}													
	(a) Parallel Shift							(b) Skew						
Mkt	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	-21	-14	-7	0	7	14	21
-4.5%	10.0	6.8	3.2	0.4	3.8	7.1	10.9	3.5	2.3	1.1	03	1.1	2.9	4.5
-3.0%	10.1	6.8	3.3	0.2	3.9	7.1	10.9	4.0	2.7	1.4	0.3	1.3	2.8	4.3
-1.5%	10.1	6.9	3.5	0.2	3.5	7.2	11.0	4.1	2.8	1.4	0.2	1.5	2.8	4.3
0.0%	10.2	7.1	3.6	0.2	3.2	7.3	11.1	4.1	2.9	1.5	0.1	1.5	2.6	4.3
1.5%	10.3	7.1	3.7	0.3	3.3	7.2	11.1	4.2	3.0	1.7	0.2	1.4	2.5	4.4
3.0%	10.8	7.3	3.9	0.4	3.3	7.4	11.3	4.4	3.2	1.8	0.2	1.4	2.6	4.4
4.5%	11.2	7.8	4.4	0.5	3.2	7.7	11.5	4.6	3.4	1.9	0.3	1.5	2.6	4.4

Table 3.7: Average backtest error E^{abs}_{cond} of the SSA P&L for the same portfolio and scenarios as in Tables 3.4, 3.5 and 3.6.

3.6.3 Historical backtesting

While DFMSA and SSA performed on simulated paths of the ground-truth model provides good insight into their relative performance, a comparison of both approaches on actual historical scenarios would provide more concrete support. To accomplish this, we perform both SSA and DFMSA for a selection of derivative securities during days of extreme market volatility in the 2008 financial

crisis. As a benchmark, we compute the true realized P&L for each of the securities during these dates. We then compare the true realized P&L to the stressed P&Ls obtained via SSA on one hand, and to the stressed P&Ls obtained via DFMSA on the other. The objective of this comparison is of course to analyze whether DFMSA provides a better picture of the risks of a security or portfolio than SSA.

To perform the historical backtest on a specific day t, we first estimate the parameters of the d.f.m. using a window of the $s > 0$ periods up to and excluding day t. The historical stress scenario is selected by choosing a subset of c.r.f.s and setting them to their *realized* values on day t. While this is straightforward for observable c.r.f. returns, it presents a difficulty if we choose any latent c.r.f. returns to stress. A good estimate of the realized c.r.f. returns can be obtained via the smoothed distribution by using the observable information in the window of periods $t - s$ through $t + s'$, with $s' > 0$. In other words, we compute $\hat{\mathbf{f}}_{t+1} := \mathbb{E}[\mathbf{f}_{t+1} \mid \mathbf{f}^o_{(t-s):(t+s')}, \mathbf{\Delta x}_{(t-s):(t+s'-1)}]$, where we recall that $\mathbf{f}^o_t$ corresponds to the observable c.r.f. returns, and set $\mathbf{c} = \hat{\mathbf{f}}_{\mathbf{s},t+1}$ as the stress scenario.[15] We then proceed to calculate the stressed P&Ls via DFMSA and SSA. The historical backtesting procedure is outlined in Algorithm 2 below.

We select 3 dates to perform this historical analysis, namely September 29, 2008, when the log-returns of the S&P500 index was -9.22%, October 13, 2008 when the S&P index rallied by 10.96%, and October 15, 2008 when the S&P sold off by 9.47%. For any given day t, we use a window of the $s = 500$ previous trading days to fit the state-space model (3.21) and (3.7), as described in Section 3.6.1. We then compute the smoothed estimate $\hat{\mathbf{f}}_{t+1}$ of the realization of the c.r.f. returns using the observations of periods $t - s$ to $t + s'$, where $s' = 250$, as described previously. Note that we know the actual realizations of the observable c.r.f., and this realization of course coincides with

[15] We acknowledge the fact that setting the scenario to the smoothed estimates of the c.r.f. returns introduces a degree of bias in our results. Indeed, by using the scenario that is most consistent with our model and with the observed risk factor returns we are giving implicit advantage to DFMSA. To see this, suppose that the true realized $\mathbf{f}_{t+1} = \mathbf{0}$ and that $\boldsymbol{\epsilon}_{t+1}$ resulted in extreme values, so that we estimate $\hat{\mathbf{f}}_{t+1} = \mathbf{c}_1$ very distinct from $\mathbf{0}$. DFMSA will then be based on the stress scenario that is a subset of $\mathbf{c}_1$ and so the resulting stressed P&L would likely be close to the true realized P&L, showing a much better performance than SSA, where in reality both DFMSA and SSA should have given similar results under the true scenario $\mathbf{c} = \mathbf{0}$. However, it should be noted that, by using the smoothed estimates instead of filtered estimates, we reduce this bias as the impact of any large $\boldsymbol{\epsilon}_t$ would be smoothed over a few periods.

Algorithm 2 Historical Backtesting to Compare SSA and DFMSA

Input: $s, t, K, \mathbf{s}$ ▷ $s = \#$ periods in window for model training

▷ $t =$ period to perform SA

▷ $K = \#$ of samples used to estimate factor model-based scenario P&L

▷ $\mathbf{s} =$ indices of c.r.f. returns to stress.

1: Estimate DFM parameters

2: Estimate π_{t+1} from $(\mathbf{f}^{o}_{(t-s):t},\ \mathbf{\Delta x}_{(t-s):(t-1)})$ ▷ $\mathbf{f}^{o}_{(t-s):t}$ are observable

3: Compute smoothed estimate $\hat{\mathbf{f}}_{t+1} := \mathbb{E}[\mathbf{f}_{t+1} \mid \mathbf{f}^{o}_{(t-s):(t+s)},\ \mathbf{\Delta x}_{(t-s):(t+s-1)}]$

4: Set $\mathbf{c} = \hat{\mathbf{f}}_{\mathbf{s},t+1}$

5: **for** $k \leftarrow 1$ to K **do**

6: Generate $\mathbf{f}^{(k)}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c})$ and $\boldsymbol{\epsilon}^{(k)}_{t+1}$ to obtain $\mathbf{\Delta x}^{(k)}_t$

7: Compute scenario P&L $\Delta V_t(\mathbf{\Delta x}^{(k)}_t)$

8: **end for**

9: Compute $\Delta V^{\text{dfm}}_t := \sum_{k=1}^{K} \Delta V_t(\mathbf{\Delta x}^{(k)}_t)/K$ ▷ Estimated scenario P&L

10: Compute ΔV^{ss}_t ▷ SSA P&L obtained by setting $\boldsymbol{\epsilon}_{t+1}$, non-stressed common factors to $\mathbf{0}$

11: Compute ΔV^{act}_t using the realized value of $\mathbf{\Delta x}_t$ ▷ Actual realized P&L

12: Compute errors $\text{E}^{\text{dfm}}_t := \left|\Delta V^{\text{dfm}}_t - \Delta V^{\text{act}}_t\right|$ and $\text{E}^{\text{ss}}_t := \left|\Delta V^{\text{ss}}_t - \Delta V^{\text{act}}_t\right|$

13: Compute the ratio $\text{E}^{\text{ratio}}_t := \text{E}^{\text{dfm}}_t / \text{E}^{\text{ss}}_t$

Output: $\Delta V^{\text{dfm}}_t, \Delta V^{\text{ss}}_t, \Delta V^{\text{act}}_t, \text{E}^{\text{dfm}}_t, \text{E}^{\text{ss}}_t$ and $\text{E}^{\text{ratio}}_t$

the smoothed estimate. We then set the scenario to be the realized return of the S&P500 index and the estimated realized return of the parallel shifts c.r.f., and proceed with SSA and DFMSA to obtain the stressed P&Ls ΔV^{dfm}_t and ΔV^{ss}_t. Denoting the actual time t realized P&L by ΔV^{act}_t, we calculate the absolute errors of each SA approach as

$$\text{E}^{\text{dfm}}_t := \left|\Delta V^{\text{dfm}}_t - \Delta V^{\text{act}}_t\right| \qquad \text{E}^{\text{ss}}_t := \left|\Delta V^{\text{ss}}_t - \Delta V^{\text{act}}_t\right| \tag{3.28}$$

Finally, we display the ratio $\text{E}^{\text{dfm}}_t / \text{E}^{\text{ss}}_t$. This ratio provides a measure of the performance of DFMSA compared to SSA. For example, if the ratio is equal to 1 then both approaches provide similar errors, whereas a ratio that is smaller (greater) than 1 indicates that DFMSA gave a more (less) accurate P&L than SSA. Evidently, the lower the ratio of absolute errors, the better the performance of DFMSA compared to SSA.

Table 3.8 shows the results of the historical backtest for out-of-the-money call and put options with 10 months to maturity, as well as for a hedged portfolio constructed via the LP procedure as described in Sections 3.4.2 and 3.6.1. For each date, the table indicates the realized S&P500 log-return and the smoothed estimate of the parallel shifts c.r.f. return, which are used as the stress scenario for that date to perform SSA and DFMSA. For example, on September 29, 2008, the S&P500 index dropped by 9.22% (in log-returns), or about 6 standard deviations, and the smoothed estimate of the parallel shifts c.r.f. return was 3.94 volatility points, or about 6.5 standard deviations, which corresponds to an equivalent increase in the volatility surface for 1 year options, as we recall that the factor loadings for the parallel shifts c.r.f. is $1/\sqrt{\tau}$. By comparing the P&L numbers obtained by each approach to the actual P&L, we can observe that the DFMSA results are consistently closer to the actual P&L than the SSA results. The absolute error ratios are below 80% for the considered securities and portfolios during these times of high market volatility, which illustrates that DFMSA is able to track better the stressed P&L.

Table 3.9 shows the results of the historical backtest for the same securities and dates as those in Table 3.8, but where the scenarios were set to the *filtered* estimates of the c.r.f. returns, i.e., using $\hat{\mathbf{f}}_{t+1} := \mathbb{E}[\mathbf{f}_{t+1} \mid \mathbf{f}^{o}_{(t-s):(t+1)}, \boldsymbol{\Delta}\mathbf{x}_{(t-s):t}]$, rather than the smoothed estimates. Here we observe that the stressed P&L numbers obtained from DFMSA are closer to the realized P&L, compared to those obtained in Table 3.8, where the scenarios were estimated using the smoothed distribution of the c.r.f. returns. This of course is to be expected, as discussed previously, since the implicit bias resulting from setting the scenarios to the estimated c.r.f. returns is smaller when smoothing large $\boldsymbol{\epsilon}_{t+1}$ over many periods.

Finally, Table 3.10 shows the results when using the S&P500 returns as the only c.r.f. to be stressed. This set of results eliminate the implicit bias as we no longer need to estimate the unobserved c.r.f. returns for setting the scenarios. We note that both DFMSA and SSA provide worse results than in the previous two tables. However, it is important to highlight that the absolute error ratios are considerably smaller than in Tables 3.9 and 3.8. This can be explained by the fact that in DFMSA we use the conditional distribution for the unstressed c.r.f. returns, instead of setting them to zero as in SSA. By using the conditional distribution we capture the correlations

Date	S&P Ret.	Parallel Shift	Security	ΔV_t^{ss}	ΔV_t^{dfm}	ΔV_t^{act}	$\text{E}_t^{\text{dfm}}/\text{E}_t^{\text{ss}}$
9/29/08	-9.22%	3.94 vol. pts	0.90 mness, 10m. Put	65.0	71.0	77.2	50.8%
9/29/08	-9.22%	3.94 vol. pts	1.05 mness, 10m. Call	-39.1	-37.3	-33.7	67.5%
9/29/08	-9.22%	3.94 vol. pts	LP Portfolio	20.5	35.3	63.9	65.9%
10/13/08	+10.96%	-6.39 vol. pts	0.90 mness, 10m. Put	-36.8	-44.6	-42.6	35.1%
10/13/08	+10.96%	-6.39 vol. pts	1.05 mness, 10m. Call	36.5	32.8	28.1	55.9%
10/13/08	+10.96%	-6.39 vol. pts	LP Portfolio	-5.5	-13.3	-15.1	19.0%
10/15/08	-9.47%	3.17 vol. pts	0.90 mness, 10m. Put	35.2	37.3	40.0	56.8%
10/15/08	-9.47%	3.17 vol. pts	1.05 mness, 10m. Call	-32.8	-32.2	-29.7	79.6%
10/15/08	-9.47%	3.17 vol. pts	LP Portfolio	-5.5	2.4	10.0	49.0%

Table 3.8: Historical SA backtest on three dates during the financial crisis for two out-of-the-money options with 10 month maturity and for the portfolio described in Section 3.6.1. For each date, we use the realized S&P500 log-return and the estimated parallel shift c.r.f. return as scenarios. We display the P&L resulting from SSA and DFMSA, as well as the actual P&L realized for each security / portfolio. We also display the ratio of the DFMSA absolute error to the SSA absolute error, serving as a measure of the relative performance between the two approaches, as mentioned in Section 3.6.3. All P&L numbers are in dollars per $100 of face value.

of the unstressed c.r.f. allowing us to estimate the stressed P&L better.

3.7 Statistical Evaluation of the Model in DFMSA

While not the focus of this chapter, a key aspect to implementing DFMSA in practice is the statistical evaluation of the dynamic factor model (d.f.m.) in question. We have argued that the SSA approach does not require or use a probabilistic model (see (3.3)) and therefore does not lend itself to any form of statistical testing. This is not true of DFMSA and in this section we briefly outline some potential approaches to the statistical validation of the underlying d.f.m.s. At a high level a data-set will consist of observations $(\mathbf{\Delta x}_t, \mathbf{f}_t^o)$ for $t = 1, \ldots, T$ of the risk factor returns and observable c.r.f. returns. While most of the state-space model literature, e.g. [82; 76], tends to focus on the estimation and implementation of these models there appears to be

Date	S&P Ret.	Parallel Shift	Security	ΔV_t^{ss}	ΔV_t^{dfm}	ΔV_t^{act}	$\text{E}_t^{\text{dfm}}/\text{E}_t^{\text{ss}}$
9/29/08	-9.22%	4.01 vol. pts	0.90 mness, 10m. Put	65.8	72.3	77.2	43.0%
9/29/08	-9.22%	4.01 vol. pts	1.05 mness, 10m. Call	-39.0	-36.6	-33.7	54.5%
9/29/08	-9.22%	4.01 vol. pts	LP Portfolio	21.6	36.4	63.9	65.1%
10/13/08	+10.96%	-6.52 vol. pts	0.90 mness, 10m. Put	-36.8	-40.8	-42.6	24.0%
10/13/08	+10.96%	-6.52 vol. pts	1.05 mness, 10m. Call	32.9	30.6	28.1	69.6%
10/13/08	+10.96%	-6.52 vol. pts	LP Portfolio	-8.7	-14.5	-15.1	9.2%
10/15/08	-9.47%	3.35 vol. pts	0.90 mness, 10m. Put	37.0	41.1	40.0	37.6%
10/15/08	-9.47%	3.35 vol. pts	1.05 mness, 10m. Call	-31.3	-30.8	-29.7	71.2%
10/15/08	-9.47%	3.35 vol. pts	LP Portfolio	-4.7	3.1	10.0	46.9%

Table 3.9: Historical SA backtest on three dates during the financial crisis for the same securities as in Table 3.8, but where the scenarios where set to the filtered estimates of the c.r.f.s, instead of the smoothed estimates. All P&L numbers are in dollars per $100 of face value.

Date	S&P Log-Return	Security	ΔV_t^{ss}	ΔV_t^{dfm}	ΔV_t^{act}	$\text{E}_t^{\text{dfm}}/\text{E}_t^{\text{ss}}$
9/29/2008	-9.22%	0.90 mness, 10 m. Put	39.8	69.0	77.2	21.9%
9/29/2008	-9.22%	1.05 mness, 10 m. Call	-56.8	-37.8	-33.7	17.9%
9/29/2008	-9.22%	LP Portfolio	-10.4	29.0	63.9	47.0%
10/13/2008	+10.96%	0.90 mness, 10 m. Put	-14.7	-30.5	-42.6	43.5%
10/13/2008	+10.96%	1.05 mness, 10 m. Call	61.1	44.7	28.1	50.3%
10/13/2008	+10.96%	LP Portfolio	12.3	-4.9	-15.1	37.4%
10/15/2008	-9.47%	0.90 mness, 10 m. Put	19.9	38.1	40.0	9.3%
10/15/2008	-9.47%	1.05 mness, 10 m. Call	-46.5	-31.6	-29.7	11.4%
10/15/2008	-9.47%	LP Portfolio	-10.4	2.6	10.0	36.2%

Table 3.10: Historical SA backtest on three dates during the financial crisis for the same securities as in Table 3.8, but where the scenarios are set to be the realized (observed) S&P c.r.f. return, to avoid the bias introduced when using smoothed or filtered estimates of the latent c.r.f. returns as scenarios. All P&L numbers are in dollars per $100 of face value.

relatively little work on the statistical testing of these models. Some notable exceptions include [67; 81]. Because the ultimate goal of these models in our context is the accurate estimation of the daily

P&L for a given portfolio (in a given scenario) we will focus here on some tests that can be applied to the one-dimensional time-series of portfolio returns.

3.7.1 VaR Exceptions for a Given Portfolio

Given any portfolio, by assumption we can use the $\mathbf{\Delta x}_t$'s to construct the univariate time series of the portfolio's realized P&L's, i.e. the $\Delta V_t(\mathbf{\Delta x}_t)$'s. As a first test of the d.f.m. it seems reasonable to require that, at the very least, the *realized* $\Delta V_t(\mathbf{\Delta x}_t)$'s should be consistent with the *estimated* $\Delta V_t(\mathbf{\Delta x}_t)$'s predicted by the d.f.m. A straightforward and commonly used approach for doing this is through the use of so-called Value-at-Risk (VaR) exceptions. Towards this end we recall that the time t α-VaR (for a given portfolio) is the $\mathcal{F}_t$-measurable random variable $\text{VaR}_{t+1}(\alpha)$ that satisfies

$$\mathbb{P}\left(\Delta V_t(\mathbf{\Delta x}_t) < \text{VaR}_{t+1}(\alpha) \mid \mathcal{F}_t\right) = 1 - \alpha$$

for any fixed $\alpha \in (0, 1)$. The time t α-VaR is therefore the $(1-\alpha)$-quantile of the distribution of the portfolio P&L conditional on $\mathcal{F}_t$. We define a VaR exception as the event that the realized ΔV_t is lower than $\text{VaR}_{t+1}(\alpha)$ and use $\mathbb{I}_{t+1}(\alpha)$ to denote the indicator function for such an event. Specifically, we define

$$\mathbb{I}_{t+1}(\alpha) := \begin{cases} 1, & \text{if } \Delta V_t(\mathbf{\Delta x}_t) < \text{VaR}_{t+1}(\alpha) \\ 0, & \text{otherwise.} \end{cases} \tag{3.29}$$

It follows that $\mathbb{I}_{t+1}(\alpha)$ is a Bernoulli random variable with success probability $1 - \alpha$. Since the $\{\mathbb{I}_t(\alpha)\}_t$'s are adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 1}$, it can in fact be easily shown[16] that they form an i.i.d. sequence of Bernoulli random variables. This result forms the basis of several simple tests for the d.f.m. under consideration.

We begin by letting $\widehat{\text{VaR}}_{t+1}(\alpha)$ be our d.f.m. estimate of $\text{VaR}_{t+1}(\alpha)$ conditional on $\mathcal{F}_t$ for $t = 1, \ldots, T$. For example, in the linear-Gaussian state-space models of Sections 3.5 and 3.6, we can use the Kalman Filter to obtain the mean vector and covariance matrix of the distribution of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$. We can then use (3.7) and (3.2), respectively, to simulate K samples from the distributions

[16] For a proof of this statement see Lemma 9.5 in [57], for example.

of $\mathbf{f}_{t+1} \mid \mathcal{F}_t$ and $\boldsymbol{\epsilon}_{t+1}$ and from there use (3.1) to obtain K samples $\Delta V_t^{(1)}, \ldots, \Delta V_t^{(K)}$ of the P&L $\Delta V_t(\boldsymbol{\Delta}\mathbf{x}_t)$. We then take the $(1-\alpha)$-quantile $\widehat{\mathrm{VaR}}_{t+1}(\alpha)$ of the empirical distribution obtained from these K samples as our d.f.m's estimate of $\mathrm{VaR}_{t+1}(\alpha)$.

We can construct the sequence of empirical VaR exception indicators $\hat{\mathbb{I}}_{t+1}(\alpha)$ by substituting $\widehat{\mathrm{VaR}}_{t+1}(\alpha)$ for $\mathrm{VaR}_{t+1}(\alpha)$ in (3.29). Under the null hypothesis that our state-space model is correct, it follows that $\sum_{t=0}^{T-1} \hat{\mathbb{I}}_{t+1}(\alpha)$ has a Binomial$(T, 1-\alpha)$ distribution. We can therefore use standard tests for the binomial to test the null hypothesis. For example, Kupiec [52] describes a two-sided binomial test with test statistic

$$Z_T = \frac{\sum_{t=1}^{T} \hat{\mathbb{I}}_t(\alpha) - T(1-\alpha)}{\sqrt{T\alpha(1-\alpha)}}. \tag{3.30}$$

In particular, we then reject the null hypothesis at the κ level if $|Z_T| \geq \Phi^{-1}(1-\kappa/2)$, where $\Phi(\cdot)$ denotes the standard normal CDF.

Various other tests can also be employed. For example, under the same null hypothesis that our state-space model is correct, it follows that the time between consecutive $\mathrm{VaR}_{t+1}(\alpha)$ exceptions are independent and geometrically distributed with success probability α. This property can be tested by approximating the geometric distribution with an exponential distribution and using a Q-Q plot or a likelihood ratio test as proposed by [22]. See also [57] for further details and additional discussion of these and other tests.

Table 3.11 shows the results of the VaR exceptions' binomial test for the dynamic factor model, as described in Section 3.6, where for each day t we fit the model using observable data for the previous $s = 500$ trading days. We note that in 2008 and 2011 the model results in a statistically significant high number of 95% and 99% VaR exceptions for most of the assets analyzed. Additionally, the model gives a statistically significant low number of 95% VaR exceptions in 2009. The reason for the poor performance in periods of changing volatilities is the fact that the state-space model assumes a static covariance matrix for the error terms $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$. A dynamic factor model with a stochastic volatility component, as in GARCH models, would be able to capture the changes in volatility levels and, as such, we would expect such a model to perform better in the binomial test.

Year	2008	2009	2010	2011	2012	2013*	All years
				95% VaR Exceptions			
Expected	13	13	13	13	13	8.5	71
S&P500 index	**34**	**4**	13	**22**	**4**	10	87
0.90 mness, 6 m. Put	**26**	**4**	14	**22**	**5**	**16**	87
1.05 mness, 6 m. Call	**39**	**2**	**4**	19	7	3	74
0.90 mness, 9 m. Put	**24**	**4**	12	**22**	**4**	13	79
1.05 mness, 9 m. Call	**39**	**2**	**5**	17	6	4	73
0.90 mness, 12 m. Put	18	**4**	10	18	**3**	4	57
1.05 mness, 12 m. Call	**42**	**4**	8	**22**	**5**	6	87
				99% VaR Exceptions			
Expected	2.6	2.6	2.6	2.6	2.6	1.7	14.3
S&P500 index	**20**	0	**7**	**14**	1	3	**45**
0.90 mness, 6 m. Put	**14**	3	5	**12**	1	**7**	**42**
1.05 mness, 6 m. Call	**18**	2	2	**10**	1	0	**33**
0.90 mness, 9 m. Put	**14**	2	4	**12**	1	2	**35**
1.05 mness, 9 m. Call	**18**	1	2	**12**	1	1	**35**
0.90 mness, 12 m. Put	**10**	1	3	**9**	1	1	**25**
1.05 mness, 12 m. Call	**17**	1	3	**10**	1	2	**34**

*Options data was available through August 2013.

Table 3.11: Number of 95% and 99% VaR exceptions of the d.f.m. for the S&P500 index and for a selection of out-of-the-money options. We highlight significant differences between the expected and realized number of exceptions, according to the binomial test at the 5% confidence level.

3.7.2 Scenario VaR Exceptions

We can use the same VaR exception framework to evaluate the state-space model within the context of scenario analysis. In particular, instead of calculating the VaR from the distribution of $\Delta V_t(\mathbf{\Delta x}_t) \mid \mathcal{F}_t$, we use the distribution of $\Delta V_t(\mathbf{\Delta x}_t) \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}_{t+1})$ for some subset $\mathbf{s}$ of the c.r.f. vector $\mathbf{f}_{t+1}$ and for some time $t+1$ scenario $\mathbf{c}_{t+1}$. In order to count the VaR exceptions, however, we must be able to obtain the realization of the P&L conditional on $\mathcal{F}_t$ and $\mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}_{t+1}$.

We therefore must set $\mathbf{c}_{t+1}$ to be equal to the realized value of $\mathbf{f}_{\mathbf{s},t+1}$. If the subset includes latent c.r.f.s, however, we need to be able to obtain good estimates of the realized c.r.f. returns, which can be obtained via the smoothing distribution $\mathbb{P}(\mathbf{f}_{0:T} \mid \mathbf{f}^{o}_{0:T}, \mathbf{\Delta x}_{0:(T-1)})$ where we recall that $\mathbf{f}^{o}_{0:T}$ corresponds to the observable c.r.f. returns.

For example,, within the linear-Gaussian state-space model framework, we can use the Kalman smoothing algorithm to obtain the smoothed estimates of the c.r.f. returns, i.e., $\hat{\mathbf{f}}_{0:T} := \mathbb{E}[\mathbf{f}_{0:T} \mid \mathbf{f}^{o}_{0:T}, \mathbf{\Delta x}_{0:(T-1)}]$. We then set $\mathbf{c}_t = \hat{\mathbf{f}}_{\mathbf{c},t}$ for each $t = 1, \ldots, T$ so that the scenario we consider at each time t is our best estimate of the scenario that actually transpired at time $t+1$. For each time t we estimate $\widehat{\text{VaR}}_{t+1}(\alpha) \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}_{t+1})$ again using Monte Carlo as described in Section 3.7.1 but where we now sample from $\mathbf{f}_{t+1} \mid (\mathcal{F}_t, \mathbf{f}_{\mathbf{s},t+1} = \mathbf{c}_{t+1})$. Having estimated each scenario-conditional $\widehat{\text{VaR}}_{t+1}(\alpha)$, we can compute the empirical VaR exception indicator $\hat{\mathbb{I}}_{t+1}(\alpha)$ and conduct the same tests as described in Section 3.7.1.

Note, however, that $\hat{\mathbb{I}}_{t+1}(\alpha)$ is no longer $\mathcal{F}_{t+1}$-adapted. Indeed, in the calculation of $\widehat{\text{VaR}}_{t+1}(\alpha)$, we use information up to the horizon T to obtain smoothed estimates of the c.r.f. returns. This introduces a bias into the results and so the resulting tests would only be approximate at best. Indeed unless we can estimate the c.r.f. returns with a high-degree of certainty the bias may be quite severe and serve to make the VaR exceptions occur less frequently than $\alpha\%$ of the time even if the null hypothesis is true. Because of this bias issue we suggest only conditioning on scenarios that only stress observable c.r.f. returns. In the options example of Section 3.6, for example, we could consider scenarios where we stress the return on the underlying security, i.e. the S&P 500, as these returns are observable.

Table 3.12 shows the results of the VaR exceptions' binomial test for the dynamic factor model, conditional on the scenario where we set the S&P500 index to its realized value. The results are qualitatively similar to the ones illustrated in Table 3.11 of Section 3.7.1, meaning that evidently the model fails to capture changes in volatility levels and therefore results in a statistically significant high number of VaR exceptions for the most part. Again, the use of a stochastic volatility component in the model would be expected to improve the performance in the binomial test.

Year	2008	2009	2010	2011	2012	2013*	All years
				95% VaR Exceptions			
Expected	13	13	13	13	13	8.5	71
0.90 mness, 6 m. Put	**27**	**5**	18	**30**	16	**16**	**112**
1.05 mness, 6 m. Call	**32**	**5**	16	**29**	14	9	**105**
0.90 mness, 9 m. Put	**30**	**5**	17	**26**	17	13	**108**
1.05 mness, 9 m. Call	**31**	**4**	14	**22**	15	5	**91**
0.90 mness, 12 m. Put	**25**	**3**	9	17	6	**2**	62
1.05 mness, 12 m. Call	**23**	**2**	8	14	10	3	60
				99% VaR Exceptions			
Expected	2.6	2.6	2.6	2.6	2.6	1.7	14.3
0.90 mness, 6 m. Put	**15**	2	**11**	**14**	7	**10**	**59**
1.05 mness, 6 m. Call	**21**	2	**8**	**14**	5	4	**54**
0.90 mness, 9 m. Put	**20**	2	**10**	**11**	3	**7**	**53**
1.05 mness, 9 m. Call	**15**	1	**7**	**11**	5	3	**42**
0.90 mness, 12 m. Put	**13**	2	**6**	**7**	2	2	**32**
1.05 mness, 12 m. Call	**15**	0	4	**9**	2	2	**32**

*Options data was available through August 2013.

Table 3.12: Number of 95% and 99% VaR exceptions of the d.f.m. conditional on the scenario where we stress the S&P500 index. We use the same selection of out-of-the-money options as in Table 3.11. We highlight significant differences between the expected and realized number of exceptions, according to the binomial test at the 5% confidence level.

3.8 Conclusions and Further Research

We have argued in this chapter for the embedding of scenario analysis inside a dynamic factor model framework so that more accurate estimates of scenario P&L's can be computed and so that these estimates can be subjected to a rigorous backtesting framework.

There are many interesting directions for future research. It would be particularly interesting to extend and develop the state-space modeling framework to more complex asset classes than

considered in Sections 3.5 and 3.6. For example, we would like to be perform DFMSA for portfolios consisting of options and equity positions on US stocks or portfolios of spot and option position on the major FX currency pairs. It would also be of interest to extend these models to allow for stochastic correlation which would by necessity move us beyond the linear-Gaussian framework. More recently Rebonato [65; 66] has proposed the use of graphical models for scenario analysis in a context where macro-economic and systemic risk factors might be stressed. It might be interesting to try and combine our DFMSA approach within such a graphical model framework.

Acknowledgements

This chapter was joint work with Prof. Martin Haugh, to whom I'm very grateful for his continued guidance and his helpful comments.

Chapter 4

Robo-Advising as a Human-Machine Interaction System

Robo-advising enhances a humans efficiency in investment decisions. We propose a framework based on risk-sensitive dynamic games, where the investor optimizes her risk-sensitive criterion while the machine adaptively learns the investors preferences. Even though the investors and machines objectives are aligned, asymmetric information makes their joint optimization process a game with strategic interactions. We consider an investor with mean-variance preferences and reduce the game to a partially observed Markov decision process. The human-machine interaction protocol features a trade-off between allowing the robo-advisor to learn the investors preferences through costly communications and optimizing the investors objective relying on outdated information.

4.1 Introduction

Robo-advising can substantially enhance human efficiency in investment decisions by handling time-intensive operations. It is crucial, however, that the investor is able to efficiently communicate her preferences to the machine to optimize her risk criterion. A machine can only provide a useful or reliable service if its valuation of the costs and risks associated with each action are aligned with the investor that it serves.

In this chapter, we propose a framework that views robo-advising as a human-machine interaction system. The objectives of the human and the machine are aligned, but there are informational asymmetries. The machine is unable to directly observe the human's preferences, and must infer them via a dynamic learning process by analyzing the human's actions. The machine is designed to serve a broad class of humans, rather than tailored to a specific category. It is thus important for the machine to personalize itself to the human, and self-calibrate as the human reveals information regarding her risk preferences and objectives.

The distinguishing feature of our human-machine interacting framework is the simultaneous handling of *human-driven* and *context-driven* risks. The uncertainty over the human's characteristics, such as her risk preferences, goals, and objectives, presents a human-driven risk to the machine. Depending on the machine's attitude toward risk, it could, for example, operate to provide a good performance to the average human. Alternatively, it could target humans whose characteristics belong to a specific quartile. On the other hand, the unpredictable nature of market conditions in which the decisions need to be executed presents context-driven risks to the human.

Both human and machine share the cooperative goal of optimizing the human's value. However, informational asymmetries make the joint minimization process of human's costs a strategic game. As such, we introduce the new equilibrium concept of *risk-sensitive Bayesian equilibrium*. In the absence of informational asymmetries, the objectives of the human and the machine are perfectly aligned, so that the game becomes cooperative. We show that, under mild assumptions on the monotonicity of the risk functions being optimized, the game theoretical problem can be reduced a related single-agent, risk-sensitive, optimization problem.

We take the perspective of an investment firm wishing to develop a robo-advising tool that constantly takes feedback from its clients, and uses it to best manage their investment portfolios. In each period, the robo-advisor must place the clients' capital into one of several pre-constructed portfolios, each having a specific risk-return profile that dynamically changes based on updated market information. Each portfolio decision reflects the robo-advisor's belief on that specific investor's risk preferences. The investor may elect to make the portfolio decisions herself over the recommendation of the robo-advisor, through the firm's communication channels and in doing so it reveals

information about her type to the machine. Overriding the portfolio choice of the robo-advisor, however, presents an opportunity cost to the investor. Through our framework, the robo-advisor can estimate the preferences of the client by observing her overriding investment decisions, or lack thereof. Additionally, the firm faces risk as aggressive portfolio choices by the robo-advisor will damage its reputation with the investor, if its estimates of the client's preferences are incorrect and the client is thus burdened with frequent override investment decisions. The tolerance that the firm has towards the uncertainty over the investor's preferences presents a human-driven risk to the machine, which is defined explicitly in our framework.

We consider an investor wishing to optimize the sum of each period's short-term risk-adjusted returns. Examples include casual investors focusing on short-term gains and other investors whose compensation package is dependent on their short-term performance. Since the robo-advisor does not know the specific risk-aversion parameter of the investor, it averages the investor's optimal value over the probability distribution on the investor's risk preferences learned on the basis of past investor's communication. We illustrate the fundamental benefit/cost trade-off faced by the investor in communicating her risk preferences to the robo-advisor to obtain more tailored investment decisions. The investor is only willing to override the machines' decision if the performance loss, defined as the difference between the risk-adjusted return of the optimal investor's portfolio and that achieved by the robo-advisor, is higher than the overriding costs. If the performance gain from human's intervention does not overcompensate for the overriding costs, then the investor would tolerate investment decisions that are suboptimal given her true risk-aversion parameter. Through numerical examples, we find that the robo-advising system achieves a value of the risk function that is lower than that of an investor-only model, in which the investor chooses the portfolio herself, but incurs opportunity costs due to market research and frequent portfolio rebalancing. The avoidance of these opportunity costs is one of the major advantages of robo-advising, because it allows the investor to delegate time-consuming activities to the machine and considerably reduce these costs.

The chapter proceeds as follows. Section 4.2 puts our chapter in perspective with existing literature. Section 4.3 develops the human-machine interaction framework. Section 4.4 specializes the framework to robo-advising. Section 4.5 concludes the chapter and discusses avenues of further

research.

4.2 Contributions and Related Work

The proposed framework describes the cooperative decision making problem of a human and a machine, that are both sensitive towards risk. In a recent work, [35] develop a framework for human-machine interactions, based on the theory of inverse reinforcement learning (IRL). Both the machine and the human are risk-neutral agents and, as such, their framework does not capture human-driven or context-driven risk. They reduce the two agent-model to a joint optimization problem building on an earlier study of [61], and compare their solution concepts to existing IRL methods. In our study, we introduce a notion of risk-sensitive equilibrium to deal with risk aversion of both human and machine, and both agents minimize a risk function.

One of the defining features of our framework is that both human and machine share the common goal of optimizing the human's objective. [62] introduce a model of decentralized stochastic control, where a team of agents work together to minimize a common objective. They show that this problem can be reduced to a POMDP by constructing a coordinator that determines strategies for the agents, based on the common information available in each period. Similar approaches have been employed by [83] [84] to solve incomplete information games between agents with conflicting objectives. The coordinator technique is appealing because it reduces a game of multiple agents to a single-agent optimization problem. In our framework, we show that, under the assumption that the risk functions are monotone with respect to risk, the solution of the coordinator problem corresponds to an equilibrium of the human-machine interaction system.

Our chapter is related to existing literature on risk-sensitive Markov decision processes (MDP). Risk aversion in MDPs has been extensively studied. Earlier contributions focused on exponential utility as in [44], mean-variance criteria an in [78], and percentile risk criteria an in [31]. [71] consider the class of risk measures, and show that these lead to tractable dynamic programming formulations. Recent contributions by [11] and [10] solve the utility maximization process and the conditional value at risk criterion for a MDP. [36] generalize these studies to a wider class of risk measures using a convex analytic approach. All these studies deal with the optimization of a single

agent. In contrast, our framework features strategic interactions between agents, and employs risk-sensitive optimization to solve for a new class of equilibria corresponding to the optimal pair of human-machine actions.

The literature on robo-advising is still at its infancy. A popular approach in the wealth management industry is goals-based investing. Investors specify quantifiable objectives such as guaranteeing the expected wealth to be above a certain threshold, given that the expected loss from return outcomes falling below the threshold is smaller than a certain value. The goals-based investment strategy is followed by Betterment, a leading robo-advisor firm, and has been investigated in academic literature by [24; 25]. [25] define a goals-based wealth management approach which restricts the efficient frontier to the subset of portfolios that achieve, with a specified probability, the investors' chosen target wealth levels. In contrast, our approach elicits information about the investor's risk preference, by offering a discrete catalogue of portfolios to the investor that may be viewed as lying on the Markowitz's efficient frontier.

Another popular robo-advising firm, Wealthfront, estimates investors' subjective risk tolerance by asking clients whether they are focused on maximizing gains, minimizing losses, or both equally. They construct a risk metric that is a weighted combination of subjective and objective risk measures, with a higher weight assigned to the component indicating higher risk aversion. The robo-advisor adopts a mean-variance optimization framework a-la [56] or variations of it. In this framework, the utility function of the investor trades off the expected return with the risk of the portfolio, weighted by the risk tolerance level of the investor. Thus, less risk-averse investors select portfolios with a higher risk and higher expected return as compared with risk-averse investors. Our approach to obtaining optimal portfolios is related to that used by Wealthfront in the short-term: in each period, the robo-advisor chooses from a catalogue of portfolios on the efficient frontier. However, in our model the investor and the machine interact throughout the whole investment horizon, and the strategy reflects the machine's learning process of the investor's risk preferences based on the investor's decisions. Our optimization criterion accounts for a long-term objective, given by the sum of single period mean-variance Markowitz utilities over the investment horizon.

Most recently, [23] develop a dynamic mean-variance framework in the context of robo-advising.

In their model, the input to the machine is the expected return of the investor, that uniquely identifies the mean-variance parameter. They argue that a quantitative asset allocation model should be based mainly on risk profile and investment horizon, while other factors such as age, labor, and income can be captured in ad-hoc ways by the financial advisor after running the asset allocation model.

4.3 The Framework

We model both the human and the machine as risk-averse agents, in order to capture context-driven and human-driven risk. We use risk functions to quantify the risk preferences of human and machine. We refer to [75] for a treatment of single agent optimization based on risk-functions. Consider a probability space $(\Omega, \mathcal{F}, P)$, and let L^∞ be the space of essentially bounded random variables.[1] A risk function is a mapping ρ: $L^\infty \to \Re$ from an uncertain outcome Z onto the set of real numbers; see also [72]. Risk functions can thus account for the entire probability distribution of an uncertain outcome, whereas expected utility functions can only depend on the realization of that outcome. We require the risk function to be monotone, i.e., that higher risk is associated with larger loss.[2]

Definition 4.3.1. *A human-machine interaction game is a T period dynamic game with asymmetric information played between two risk sensitive agents: a human, $\mathbf{H}$, and a machine, $\mathbf{M}$. The game is described by a tuple $\left\langle \mathcal{S}, \left\{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{M}}\right\}, \Theta, \left\{\rho^{\mathbf{H}}, \rho^{\mathbf{M}}\right\}, P, c, \pi_1 \right\rangle$, whose elements are defined as:*

$\mathcal{S}$ is a discrete or continuous set of system states: $s \in \mathcal{S}$;

$\mathcal{A}^{\mathbf{H}}$ is a discrete or continuous set of actions for $\mathbf{H}$: $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$;

$\mathcal{A}^{\mathbf{M}}$ is a discrete or continuous set of actions for $\mathbf{M}$: $a^{\mathbf{M}} \in \mathcal{A}^{\mathbf{M}}$;

Θ is a discrete or continuous set of possible risk parameters, only observed by $\mathbf{H}$: $\theta \in \Theta$;

[1] A random variable Z is essentially bounded if there exists $M \geq 0$ such that $P(|Z| > M) = 0$.

[2] Risk Functions which satisfy the axioms of monotonicity, translation invariance and convexity are referred to as risk measures. See [5]. In our framework, we only require the monotonicity assumption to study the risk-sensitive Bayesian equilibrium.

$\rho_\theta^{\mathbf{H}}(\cdot)$ *is* $\mathbf{H}$*'s risk function, parameterized by* θ*;*

$\rho^{\mathbf{M}}(\cdot)$ *is* $\mathbf{M}$*'s risk function;*

$P(\cdot|\cdot,\cdot,\cdot)$ *is the probability transition function on the future state, given the current state and joint action:* $P(s'|s, a^{\mathbf{H}}, a^{\mathbf{M}})$*;*

$c(\cdot,\cdot,\cdot)$ *is an instantaneous cost function that maps the system state and joint actions to a vector of real numbers:* $c : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \to \Re$*;*

$\pi_1(\cdot)$ *is a common prior distribution over the risk parameters:* $\pi_1(\theta) \in \mathcal{P}(\Theta)$*.*

Remark 4.3.1. *We assume that the set of actions* $\mathcal{A}^{\mathbf{H}}$ *does not include the action of the human directly communicating her risk-aversion parameter to the machine. In general, risk-preferences can be indirectly estimated by posing subjective questions to the human that reflect her behavioral attitudes towards risk. However, it is well known from the behavioral economics literature that humans do not provide consistent answers, for instance, research shows that individuals consistently overstate their true risk-tolerance ([9]). It is therefore the case that direct communication of the risk-aversion parameter by the human is unrealistic, and hence we exclude it from the action space.*

After each period t, the human and the machine incur a common cost, $c(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}) \in \Re$, depending on the current state of the system, and their joint action. Their incentives are partially aligned as both the human and the machine prefer to keep the total system costs low over the T period horizon. The human's objective is to minimize the costs using her risk function $\rho_\theta^{\mathbf{H}}$ as the optimization criterion, where θ is the human's risk parameter. For example, the mean-variance risk function $\rho_\theta^{\mathbf{H}} = \theta Var\left[\sum_{t=1}^{T} c(s_t, a_t^H, a_t^M)\right] - \mathbb{E}\left[\sum_{t=1}^{T} c(s_t, a_t^H, a_t^M)\right]$ maps the random outcome for the total costs to a quantity through the parameter $\theta \in \Re$. The machine does not know the value of θ at the initial stage of the game, but begins with a prior distribution $\pi_1(\cdot) \in \mathcal{P}(\Theta)$, where we have used $\mathcal{P}(\Theta)$ to denote the set of probability distributions on Θ. The machine's objective is to minimize the risk function criterion $\rho^{\mathbf{M}}$.

Denote the set of public histories as

$$H_t := \left(\mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{M}}\right)^{t-1} \times \mathcal{S}^t, \tag{4.1}$$

where $h_t = \left(s_1, a_1^{\mathbf{H}}, a_1^{\mathbf{M}}, \ldots, a_{t-1}^{\mathbf{H}}, a_{t-1}^{\mathbf{M}}, s_t\right) \in H_t$ for $t > 1$ and $h_1 = s_1$. A public history contains information that is observed by both the human and the machine, which includes the realization of the system's states and the actions executed by both agents. The machine maintains the posterior distribution over the human's type, $\pi_t(x) := P(\theta = x|h_t)$, which we refer to as the machine's belief in period t.

A Markov strategy for the human $\sigma^{\mathbf{H}} = (\sigma_1^{\mathbf{H}}, \ldots, \sigma_T^{\mathbf{H}})$ is a sequence of measurable maps $\sigma_t^{\mathbf{H}} : \mathcal{S} \times \mathcal{P}(\Theta) \times \Theta \to \mathcal{P}\left(\mathcal{A}^{\mathbf{H}}\right)$ so that

$$\sigma_t^{\mathbf{H}}(a|s_t, \pi_t, \theta) = P(a_t^{\mathbf{H}} = a|s_t, \pi_t, \theta), \ \ \forall t \in \{1, \ldots, T\}, a \in \mathcal{A}^{\mathbf{H}}.$$

A Markov strategy for the machine $\sigma^{\mathbf{M}} = (\sigma_1^{\mathbf{M}}, \ldots, \sigma_T^{\mathbf{M}})$ is a sequence of measurable maps $\sigma_t^{\mathbf{M}} : \mathcal{S} \times \mathcal{P}(\Theta) \to \mathcal{P}\left(\mathcal{A}^{\mathbf{M}}\right)$ so that

$$\sigma_t^{\mathbf{M}}(b|s_t, \pi_t) = P(a_t^{\mathbf{M}} = b|s_t, \pi_t), \ \ \forall t \in \{1, \ldots, T\}, b \in \mathcal{A}^{\mathbf{M}}$$

Notice that the human's Markov strategy depends on the machine's current beliefs because the action of the human is influenced by the action of the machine, which in turn depends on its belief over the human's type. The total (cumulative) cost is given by the random variable

$$C_T := \sum_{t=1}^{T} c\left(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}\right).$$

We define the *risk-sensitive Bayesian equilibrium* as a pair of strategies $\left(\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}\right)$ and a belief profile $\pi^* := (\pi_1^*, \ldots, \pi_T^*)$ such that

$$\begin{aligned} \rho_\theta^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1^*, h_1\right) &\leq \rho_\theta^{\mathbf{H}}\left(C_T|\tilde{\sigma}^{\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1^*, h_1\right), \\ \rho^{\mathbf{M}}\left(\rho_\theta^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1^*, h_1\right)|\pi_1^*\right) &\leq \rho^{\mathbf{M}}\left(\rho_\theta^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \tilde{\sigma}^{\mathbf{M}}, \pi_1^*, h_1\right)|\pi_1^*\right), \end{aligned} \tag{4.2}$$

for all strategies $\tilde{\sigma}^{\mathbf{H}}$, $\tilde{\sigma}^{\mathbf{M}}$. Furthermore, the machine's belief profile π^* must be consistent with the strategies $\left(\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}\right)$ in that Bayes' rule is used to update the beliefs. Specifically, the machine's belief on the true value of the human's risk parameter θ satisfies the standard nonlinear filter equation ([33])

$$\pi_{t+1}^*(\theta) := \frac{\pi_t^*(\theta)\sigma^{*\mathbf{H}}(a_t^{\mathbf{H}}|s_t, \pi_t^*, \theta)}{\sum_{\tilde{\theta}} \pi_t^*(\tilde{\theta})\sigma^{*\mathbf{H}}(a_t^{\mathbf{H}}|s_t, \pi_t^*, \tilde{\theta})}, \tag{4.3}$$

provided there exists a value of $\tilde{\theta}$ such that $\pi_t^*(\tilde{\theta}) > 0$ and $\sigma^{*\mathbf{H}}(a_t^{\mathbf{H}}|s_t, \pi_t^*, \theta) > 0$. In period 1, the belief profile π_1^* is equal to the prior π_1.

The first of the two inequalities in equation (4.2) indicates that the human has no incentive to unilaterally deviate from her action $\sigma^{*\mathbf{H}}$ to any other action $\tilde{\sigma}^{\mathbf{H}}$ because her risk-adjusted total cost would increase. Similarly, the second inequality stipulates that the machine's action yields the smallest risk-adjusted total cost, according to both the human's risk parameter and the machine's beliefs over the human's risk parameter.

The canonical solution concept for dynamic games of incomplete information is the Bayesian equilibrium (BE). However, standard equilibrium concepts rely on maximizing the expectation of utility functions assigned to each player. A Bayesian equilibrium in our setup would require that both agents minimize the expected disutility of total system costs, rather than the general risk functions we present.

Context-driven risk in our model is captured by applying the risk function $\rho^{\mathbf{H}}$ to the total system cost. This allows us to capture a wide variety of cost criteria that depend on the statistical properties of the cumulative costs, including value at risk, conditional value at risk, and worst case measures. A special case of context-driven risk is when the human minimizes the expectation of a convex utility function on costs. Human-driven risk is quantified by the risk function $\rho^{\mathbf{M}}$ over the distribution of human's risk parameters. For example, if $\rho^{\mathbf{M}}$ is the expectation operator, then the machine aims for the best service to the average human type. On the other hand, if $\rho^{\mathbf{M}}$ represents the value at risk for some level of service α, then the machine aims to provide a good service for $1-\alpha$ percentage of the human's types. Lastly, if the human's type is revealed to the machine before T, then there is no human-driven risk. In this case, $\rho^{\mathbf{M}}\left(\rho_\theta^{\mathbf{H}}(C_T)\right) = \rho_\theta^{\mathbf{H}}(C_T)$, so that the two inequalities in Eq. 4.2 coincide, and the game becomes *cooperative*.

The solution methodology that we propose to address the human-machine framework is to transform the strategic game to a single-agent problem by introducing a coordinator agent $\mathbf{C}$. The coordinator assigns a policy $\sigma^{\mathbf{C}} = (g^{\mathbf{M}}, g_\theta^{\mathbf{H}})$ such that $g^{\mathbf{M}}$ is a strategy for the machine and $g_\theta^{\mathbf{H}}$ is a decision function, which prescribes the human's strategy for each possible realization of θ. Hence, the coordinator is unaware of the human's risk parameter, but instead chooses a strategy

for every possible type of human. The coordinator's objective is to use these controls to minimize the machine's risk function

$$\min_{g^{\mathbf{M}}, g_\theta^{\mathbf{H}}} \rho^{\mathbf{M}} \left(\rho_\theta^{\mathbf{H}} \left(C_T | g_\theta^{\mathbf{H}}, g^{\mathbf{M}}, \pi_1, h_1 \right) | \pi_1 \right). \tag{4.4}$$

The resulting problem is a partially-observable, risk-sensitive, Markov decision process (risk-POMDP). The following theorem connects the solution to the coordinator problem with the equilibrium concept for the human-machine interaction game.

Theorem 4.3.1. *A solution to the coordinator problem is a risk-sensitive Bayesian equilibrium to the two-agent human-machine interaction game.*

The proof of Theorem 4.3.1 is included in appendix C.1.

4.4 Robo-Advising with Myopic Mean-Variance Preferences

We specialize the general framework presented in Section 4.3 to capture decision making in robo-advising settings. We consider a T period investment framework in which an investor hires a robo-advisor to select an investment portfolio at each period t. The robo-advisor learns the investor's risk preference over time, and selects the risk-return profile of the portfolio that best reflects the learned preferences. For instance, if the investor's tolerance for risk was known to be high, then the robo-advisor would choose a portfolio with a higher expected return, irrespective of its variance. Conversely, if the robo-advisor knew that the investor were very sensitive to risk, then it would avoid portfolios with a high variance even if they had higher expected return.

The human $\mathbf{H}$ corresponds to the investor, and the machine $\mathbf{M}$ corresponds to the robo-advisor. The system states model the market environment, assumed to be represented by the expected return and standard deviation of m available investment portfolios at each time t. Formally, $\mathcal{S} = \{s^{(1)}, \ldots, s^{(n)}\}$ represents the set of economic scenarios. Portfolio i in state $s \in \mathcal{S}$ is assumed to have a known expected return $\mu(s, i)$ and standard deviation $\sigma(s, i)$. For example, $s = s^{(1)}$ may correspond to a low return-low volatility market scenario, while $s = s^{(n)}$ may represent a high

return-high volatility scenario.[3] Note that the expected return and standard deviation parameters of each portfolio are time invariant, i.e., they depend on the actual state s, but not on the time t. The probability of a transition from state s to state s' is assumed to be independent of the human's action, and is denoted by $P(s' \mid s)$ for all $s, s' \in \mathcal{S}$. This means that the investor's decisions cannot influence the market environment.

The set of actions available for $\mathbf{M}$ corresponds to the m available portfolios, i.e., $\mathcal{A}^{\mathbf{M}} = \{a^{(1)}, \ldots, a^{(m)}\}$. An action $a_t^{\mathbf{M}} = a^{(i)} \in \mathcal{A}^{\mathbf{M}}$ corresponds to $\mathbf{M}$ choosing portfolio $a^{(i)}$ at time t. In addition, the investor is allowed to override the decision of the robo-advisor, and therefore has a set of actions $\mathcal{A}^{\mathbf{H}} = \{a^{(0)}, a^{(1)}, \ldots, a^{(m)}\}$, where $a_t^{\mathbf{H}} = a^{(0)}$ corresponds to no-override at t (so that the investor keeps the portfolio selected by $\mathbf{M}$), and $a_t^{\mathbf{H}} = a^{(i)} > 0$ corresponds to the investor overriding $\mathbf{M}$'s decision with portfolio $a^{(i)}$. We denote the actual portfolio selected at time t by

$$a_t := \begin{cases} a_t^{\mathbf{M}}, & \text{if } a_t^{\mathbf{H}} = a^{(0)} \\ a_t^{\mathbf{H}}, & \text{if } a_t^{\mathbf{H}} \neq a^{(0)} \end{cases} \tag{4.5}$$

Active intervention by the investor is costly, and we denote this cost by $\kappa(a_t^{\mathbf{H}})$. We can interpret the investor's override decision as a two-stage process: First, the investor decides whether or not the portfolio chosen by $\mathbf{M}$ is adequate (the investor's policy is discussed in Section 4.4.3). Given that the first decision (whether to override or not) is made at every period, we can assume that $\kappa(a^{(0)}) = 0$ without loss of generality. If $a_t^{\mathbf{M}}$ is inadequate, then the investor must choose an alternative portfolio by performing costly operations, including market research, etc.... Hence, in periods when an override has been decided, additional opportunity costs are incurred. Therefore, we assume that override decisions are costly, i.e., $\kappa(a_t^{\mathbf{H}}) = \kappa_c > 0$ if $a_t^{\mathbf{H}} \neq a^{(0)}$.

[3] In an empirical setting, we would define $s \in \mathcal{S}$ to be a specific economic regime, and then estimate the expected return and standard deviation of a pre-defined set of portfolios from historical data. Note, however, that the focus of the present work is to analyze how the machine learns from the decisions of the investor and to quantify the value added by a robo-advisor. Hence, we abstract away from the inference and selection of those scenarios, and assume that they have been computed beforehand.

4.4.1 The Risk-Aversion Parameter

The investor's risk-aversion levels are assumed to belong to a finite set Θ, such that $|\Theta| = p$. The robo-advisor's initial belief over the risk-aversion levels is given by $\pi_1 \in \mathbb{R}^p$. This initial belief could be obtained and estimated by the robo-advisor from a series of questionnaires given to the investor during sign-up.

We model the investor's decision making process as if she were aware of her own risk-aversion parameter, while **M** does not know it and must estimate it using available information. We highlight the difference between the prior distribution π_1 on the initial risk-aversion parameter from the implied risk-aversion parameter that the investor indirectly communicates through her trading decisions. In practice, an investor may not be aware of what her risk-aversion parameter is at any given point in time. For example, it is well known that investors consistently overestimate their risk-tolerance, hence relying on an investor's self-reported risk-tolerance may lead to a suboptimal choice of portfolios. Instead, she makes decisions in accordance with an internal system of beliefs which implicitly, rather than explicitly, quantifies risk.

Our modeling framework provides a mechanism to infer the implied risk-aversion parameter of the investor. The decisions of the investor allow the robo-advisor to learn the risk-aversion parameter via a standard Bayesian update, as described in equation (4.3).

4.4.2 Costs and Objective Functions

As discussed in Section 4.3, the objective of the human-machine interaction system is to minimize the risk-adjusted expected cost for each period of the investment horizon. In particular, the cost in period t is given by[4]

$$c_\theta(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}) = \theta\sigma^2(s_t, a_t) - \mu(s_t, a_t) + \kappa(a_t^{\mathbf{H}}),$$

where a_t represents the chosen portfolio and is given by (4.5). The cost function above weights the risk associated with the investment decision against the expected portfolio return, and accounts for the costs of overriding the robo-advisor's decision. This cost function penalizes the amount of risk

[4] Transaction costs, although an important factor in any investment strategy, are not considered in this framework.

undertaken (captured by the variance of the selected portfolio) according to the risk-aversion level of the investor. The total cumulative cost is then given by

$$C_T(s_{1:T}, a^{\mathbf{H}}_{1:T}, a^{\mathbf{M}}_{1:T}) := \sum_{t=1}^{T} c_\theta(s_t, a^{\mathbf{H}}_t, a^{\mathbf{M}}_t),$$

where $s_{1:T}$ denotes the state path $s_1, \ldots, s_T$, while $a^{\mathbf{H}}_{1:T} := \{a^{\mathbf{H}}_1, \ldots, a^{\mathbf{H}}_T\}$ and $a^{\mathbf{M}}_{1:T} := \{a^{\mathbf{M}}_1, \ldots, a^{\mathbf{M}}_T\}$ denote the set of investor and robo-advisor actions, respectively. We then define the coordinator policy as $\sigma^{\mathbf{C}} = (g^{\mathbf{M}}, g^{\mathbf{H}}_\theta)$, where, as discussed in Section 4.3, $g^{\mathbf{M}} := \{g^{\mathbf{M}}_1, \ldots, g^{\mathbf{M}}_T\}$ denotes a strategy for the machine and $g^{\mathbf{H}}_\theta := \{g^{\mathbf{H}}_{\theta,1}, \ldots, g^{\mathbf{H}}_{\theta,T}\}$ prescribes the human's strategy for each possible realization of θ. Note that $g^{\mathbf{M}}$ and $g^{\mathbf{H}}_\theta$ are policies that are adapted to the set of public histories given by (4.1). The risk function of the human is then given by $\rho^{\mathbf{H}}_\theta(C_T \mid g^{\mathbf{M}}, g^{\mathbf{H}}_\theta, \pi_1, h_1) := \mathbb{E}^s[C_T(s_{1:T}, g^{\mathbf{M}}, g^{\mathbf{H}}_\theta) \mid \pi_1, h_1]$, where the expectation is taken with respect to the probability distribution of the state path $s_{1:T}$. We assume that the robo-advisor captures the human-driven risk using $\rho^{\mathbf{M}}(\rho^{\mathbf{H}}_\theta \mid \pi_1) := \mathbb{E}^\pi[\rho^{\mathbf{H}}_\theta \mid \pi_1]$, where in this case the expectation is taken with respect to the robo-advisor's belief states $\pi_1, \ldots, \pi_T$ on the risk-aversion parameter θ. In other words, we assume that the robo-advisor is neutral with respect to the investor's type.

The objective function of the human-machine interaction system corresponds to the minimization criterion in the coordinator problem given by (4.4), and takes the explicit form

$$\min_{g^{\mathbf{M}}, g^{\mathbf{H}}_\theta} \mathbb{E}^{\pi,s}\Big[\sum_{t=1}^{T} \theta\sigma^2(s_t, g_t) - \mu(s_t, g_t) + \kappa(g^{\mathbf{H}}_{\theta,t}))\Big], \tag{4.6}$$

where the expectation is taken with respect to the joint distribution of the belief states $\pi_1, \ldots, \pi_T$ and the state path $s_{1:T}$, and where g_t is defined similar to a_t in (4.5), but replacing actions $a^{\mathbf{M}}_t$ and $a^{\mathbf{H}}_t$ with strategies $g^{\mathbf{M}}_t$ and $g^{\mathbf{H}}_{\theta,t}$, respectively. This choice of objective function reflects that the robo-advisor averages an investor's optimal risk criterion, including the cost of communication, over the filtering probability distribution of the investor's risk preference conditioned on the set of public histories. The investor wishes to optimize the sum of each period's risk function.

4.4.3 Investor's Policy

From the form of the objective function in (4.6) and the learning capabilities of the robo-advisor, it is evident that the investor faces a trade-off. On the one hand, the investor would like to frequently

communicate her risk preferences (through overriding actions) so that the robo-advisor is better informed to make investment decisions. On the other hand, the investor would like to keep the costs low and not override, unless communication leads to significant improvements in the robo-advisor's portfolio selection strategy. In other words, if the override costs $\kappa(a_t^{\mathbf{H}})$, for $a_t^{\mathbf{H}} > 0$, is large enough, the investor would not have any incentive to override the robo-advisor's decisions, even if they appear suboptimal for a given risk-aversion parameter. Under these circumstances, the robo-advisor will not be able to learn the risk-aversion of the investor. On the other hand, if the override cost is sufficiently low, the investor may find it optimal to communicate her preferences very frequently, and the robo-advisor will be able to learn the investor's risk preferences fast.

Assuming that the investor behaves myopically, as described in Section 4.4.2, we can explicitly write the investor's myopic policy $a_t^{\mathbf{H}}$. First, we denote by a^* the myopic optimal portfolio at time t, i.e.,

$$a^* := \underset{a \in \mathcal{A}^{\mathbf{H}} \setminus \{a^{(0)}\}}{\operatorname{argmin}} \; \theta\sigma^2(s_t, a) - \mu(s_t, a) \tag{4.7}$$

Then the myopic investor's policy, after observing the machine decision $a_t^{\mathbf{M}}$, is given by

$$a_t^{\mathbf{H}} = \begin{cases} a^{(0)}, & \text{if } \theta\sigma^2(s_t, a_t^{\mathbf{M}}) - \mu(s_t, a_t^{\mathbf{M}}) \leq \theta\sigma^2(s_t, a^*) - \mu(s_t, a^*) + \kappa(a^*) \\ a^*, & \text{otherwise.} \end{cases}$$

Hence, the investor will only override if the risk-adjusted cost of portfolio $a_t^{\mathbf{M}}$ is lower than the risk-adjusted cost of the myopic optimal portfolio a^* plus the override cost.

The above policy assumes that the investor always acts optimally, so that any override decision only happens if the portfolio chosen by the robo-advisor significantly differs from the optimal myopic portfolio of the investor. However, there can be situations in which an investor does not have the time or flexibility to override a suboptimal decision made by the robo-advisor. The frequency of these errors would be higher for short time-scales, because the investor has a smaller amount of time at her disposal to make decisions. Therefore, we consider a situation in which an investor behaves as an imperfect agent. More specifically, we allow the investor to commit a *missed override* error, in which she fails to override a suboptimal decision taken by the robo-advisor. More specifically,

the error is captured by the imperfect human policy given by

$$a_t^{\mathbf{H}} = a^{(0)} \text{ if } D_t > 0.$$

where D_t is a measure of the sub-optimality of the robo-advisor's decision, and is given by

$$D_t := \theta_t \sigma^2(s_t, a_t^{\mathbf{M}}) - \mu(s_t, a_t^{\mathbf{M}}) - \Big[\theta_t \sigma^2(s_t, a^*) - \mu(s_t, a^*) + \kappa(a^*)\Big]$$

We assume the missed override error occurs randomly with probability $P_m(D_t)$, conditional on $D_t > 0$. We expect that larger discrepancies would be easier to perceive for an investor, while lower discrepancies would be harder to detect or less important to correct. To capture this behavior we choose $P_m(D_t)$ to be a non-increasing function of D_t, so that the probability of a missed override is smaller if the differences between the robo-advisor chosen portfolio and the optimal myopic investor portfolio is larger. Note that these errors would slow down the learning process of the robo-advisor, who will take longer to learn the risk-aversion parameter of the investor.

4.4.4 Robo-advisor's Policy

As discussed in Section 4.3, the optimization criterion of the robo-advising system may be formulated as a POMDP. It is well known that finding the optimal solution of a POMDP is, in general, computationally intractable. Many approximation algorithms have been proposed in the literature, and we refer to [51] for a comprehensive review of POMDPs.[5]

We consider a simple heuristic that is based on the greedy policy with respect to the so-called Q-function, and defined by

$$Q_t(\theta, s_t, a_t^{\mathbf{M}}, a_t^{\mathbf{H}}) := \theta \sigma^2(s_t, a_t) - \mu(s_t, a_t) + \kappa(a_t^{\mathbf{H}}) + \mathbb{E}\Big[V_{t+1}(\theta, s_{t+1}) \mid s_t\Big], \quad t = 0, \ldots, T \tag{4.8}$$

where we recall that a_t is defined in (4.5), and we define

$$V_t(\theta, s_{t+1}) := \min_{a_t^{\mathbf{M}}, a_t^{\mathbf{H}}} Q_t(\theta, s_t, a_t^{\mathbf{M}}, a_t^{\mathbf{H}}), \quad t = 0, \ldots, T \tag{4.9}$$

[5] See also [55], [54] and [47], who review several methods that yield near-optimal policies, along with the efficiencies and weaknesses of these procedures. Recent work by [39] uses the so-called supersolutions to construct efficient approximate value functions.

with boundary condition $V_{T+1} = 0$. The greedy policy with respect to the Q-function is then given by

$$(g_t^{\mathbf{M}}(\pi_t), g_t^{\mathbf{H}}(\pi_t)) := \underset{a_t}{\operatorname{argmin}} \sum_{\theta \in \Theta} \pi_t(\theta) Q_t(\theta, s_t, a_t), \tag{4.10}$$

where π_t represents the filtering distribution (or belief state) on the risk-aversion parameter at time t.[6] We refer the reader to Appendix C.2 for a description of alternative heuristic procedures to approximate solutions of POMDPs.

Having set a heuristic policy for the robo-advisor, we construct an upper bound V_0^{upper} on (4.6) via Monte Carlo simulation. We sample J paths of the state process and calculate the greedy policy (4.10) at each time t. Using this policy, we compute the cumulative cost on each sample path, and then take the average to obtain V_0^{upper}. To measure the performance of the heuristic, however, we need to calculate a so-called *dual bound*, which is a lower bound for the minimization problem in (4.6). A dual bound is obtained directly from the Q-function at time $t = 0$, by setting

$$V_0^{lower} = \min_{a_0} \sum_{\theta \in \Theta} \pi_0(\theta) Q_0(\theta, s_0, a_0). \tag{4.11}$$

As our numerical results in Section 4.4.5 show, the above described heuristic above is close to optimal because the duality gap is relatively small. It is worth highlighting that the quality of the approximation depends on the particular POMDP that is being solved. For completeness, we provide a brief discussion on more sophisticated approaches to calculate dual bounds in Appendix C.2.

4.4.5 Numerical Results

This section develops a numerical study to analyze the rate at which the machine learns the human's risk preferences, and to measure the value added by the robo-advisor over the stand-alone investor.

[6] We remark that the Q-function heuristic does not consider exploration-exploitation tradeoffs. Despite this being an important concept in the reinforcement learning literature, exploration is not desirable in a system where the client pays the robo-advisor to be given optimal investment decisions. Proposing portfolio choices that explore the risk-aversion space of the investor, for example by choosing a random portfolio with probability $\epsilon > 0$, would reflect badly on the robo-advisor system and may burden the client with unnecessary costly override decisions.

We use Monte Carlo simulation of the greedy policy given in 4.4.4 to estimate an upper bound for the solution of the robo-advising system (4.6). We fix the number of investment periods to $T = 10$, and set the number of portfolios available to the investor at each time t to $n = 4$. We assume that there are $m = 20$ admissible values for the investor's risk-aversion parameter. Additional details on the numerical study are reported in Appendix C.3.

We start analyzing how the machine learns, over time, the investor's risk aversion parameter θ. Figure 4.1 illustrates the learning process on two distinct simulated paths of the system, for an error-prone investor with $P_m = 0.4$ for $D_t \leq 3\%$, and $P_m = 0$ otherwise. Based on the investor's decisions to override, the machine revises its belief on the investor's risk-aversion parameter via Bayesian updating (4.3). At time $t = 1$, the robo-advisor places a uniform prior distribution on the set Θ of possible risk-aversion parameters for the investor (see Appendix C.3 for details). With time, the mass of the posterior distribution concentrates on the set of plausible values, i.e., those that are consistent with the investor's decisions so far.

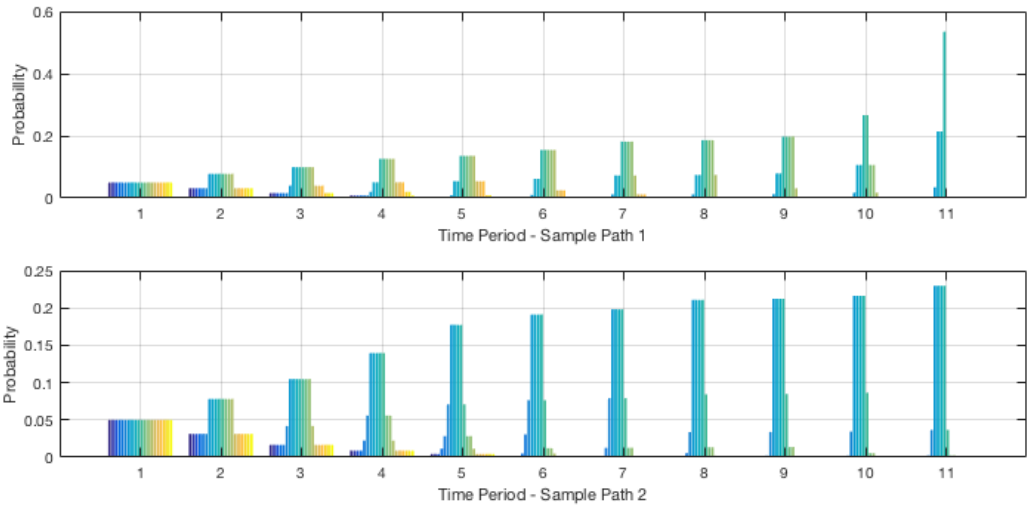


Figure 4.1: Updating of beliefs on the risk-aversion parameter for an error-prone investor with $P_m = 0.4$ for $D_t \leq 3\%$, and $P_m = 0$ otherwise. We illustrate the result on two sample paths of the robo-advising system.

Next, we analyze the value added by the robo-advisor in making decisions, as compared to an

investor-only model who makes decisions without any machine support. To perform this comparison, we first calculate the approximate expected optimal value produced by the chosen heuristic in the risk-POMDP, for a given override cost $\kappa(a_t^{\mathbf{H}}) = k$ if $a_t^{\mathbf{H}} > 0$. Then, we consider an investor-only system by reducing the action space of the investor to be $\mathcal{A}_0^{\mathbf{H}} := \mathcal{A}^{\mathbf{H}} \setminus \{a^{(0)}\}$. This means that the investor needs to choose her own portfolio at every period t (or equivalently, she must always override the choice of the machine). For comparison purposes, in the investor-only setting we assume that any action $a_t^{\mathbf{H}} \in \mathcal{A}_0^{\mathbf{H}}$ has the same cost $\kappa(a_t^{\mathbf{H}}) = k$. This cost may be interpreted as the effort incurred by the investor for choosing a portfolio. She needs to closely monitor financial markets, solving her own optimization problem, and communicating her choice to an asset manager. Moreover, the attention span required to make decisions on short time-scales is subtracted to other activities, and thus represents an opportunity cost for the investor. Clearly, the investor-only system corresponds to a fully-observed problem, because the investor is aware of her own risk-aversion parameter, and so the criterion (4.6) becomes a fully-observed Markov Decision Process (MDP) which can be solved to optimality. Figure 4.2 shows the approximate optimal value of the risk-POMDP corresponding with the human-machine interaction system, and compares it to the optimal value of a human-only MDP, for different choices of the cost parameter k.

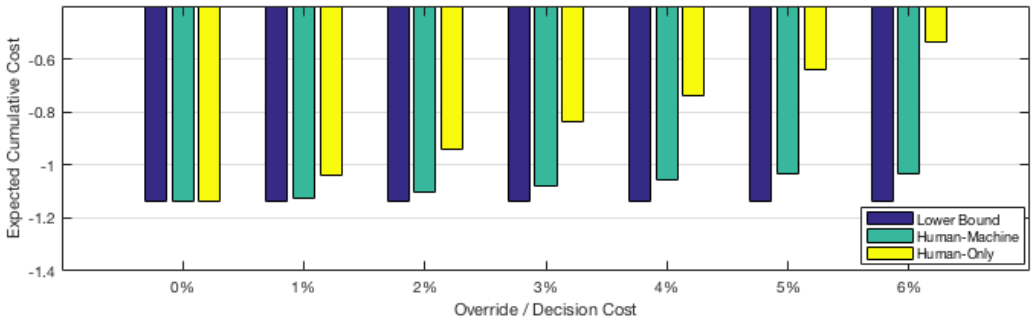


Figure 4.2: Approximate value of the minimum expected cost of the robo-advisor system (green) and expected minimum cost of the investor-only system (yellow), as a function of the cost parameter k. We assume that the cost of an override decision in the robo-advisor system is equal to the cost of an investment decision in the investor-only system. The blue bars represent the lower bound on the true-optimal value of the robo-advisor system computed using Eq. (4.11).

Figure 4.2 illustrates the value added by the robo-advisor. Assuming the override cost in the robo-advisor system equals the decision cost in the investor-only system, we observe that the robo-advisor system yields a lower expected cumulative cost over the investment horizon, compared to the investor-only system. This difference can be explained by two main observations. First, the investor will not incur costs if the robo-advisor selects a portfolio that is close to the myopically optimal one, given the true (unknown) risk-aversion parameter (i.e., no override is needed). By contrast, these decision costs are incurred every period in the investor-only system. From an operational perspective, this is one of the primary advantages of robo-advising, in that it allows the investor to delegate research on investment instruments, times for portfolio re-balancing, and other time-consuming activities to the robo-advisor. Such a delegating process may considerably reduce the investor's costs. It also appears from Figure 4.2 that, as the override / decision cost increases, the overall expected cumulative cost increases (i.e., it becomes less negative). However, this increase in cost is not reflected in a similar fashion by the human-only and the robo-advisor system. In the human-only system, we observe a linear increase in expected cumulative cost, while in the robo-advisor system, the expected cumulative cost increases at a slower rate for override costs greater than 4%. This effect can be explained by the previously mentioned trade-off faced by the investor when deciding on overriding: if the override cost is too high, the investor never chooses to override and the robo-advisor does not efficiently learn the risk-aversion parameter of the investor. As a result, it will make decisions that satisfy the *average* investor, where the average is taken with respect to the initial belief on the investor's risk aversion.

4.4.6 Model Extensions - Dynamic Risk-Aversion

We present an extension of the modeling framework, that can accommodate risk-preferences which are not necessarily static, but rather dynamically change overtime as the market moves and investment decisions are made.

We consider a dynamic risk-aversion parameter $\theta_t \in \Theta$, whose transitions are determined by the following function

$$\theta_{t+1} = f(\theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t, s_{t+1}) \text{ for } t = 1, \ldots, T-1. \tag{4.12}$$

The transition function f may be designed by the modeler to reflect typical behavior of an investor. Note that f is not only a function of the decisions $a_t^{\mathbf{H}}$ and $a_t^{\mathbf{M}}$, but also of the current state s_t and next state s_{t+1}. Hence, risk-aversion parameter transitions are both impacted by investment decisions and changes in the market environment. For example, the modeler may believe that riskier choices, i.e. portfolios with a higher standard deviation, should have a higher impact on the risk-aversion parameter. More specifically, an investor could end up with a higher appetite for risk if the market moves in a favorable direction and the portfolio chosen was high-risk high-return, because the resulting capital and the investor's optimism would then have increased. Similarly, if the high-risk high-return portfolio was chosen but the market moved in an adverse direction, the appetite for risk could be lower because both the resulting capital and her optimism would have taken a hit. Additionally, the magnitude of the change in risk preferences may also depend on the riskiness of the chosen portfolio, so that a high-risk high-return portfolio may have a higher impact on the capital and optimism than a low-risk low-return portfolio.

We can combine the transition function of the risk-aversion parameter (4.12) with the state transitions $P(s_{t+1} \mid s_t)$ to obtain a risk-aversion transition probability function, given by

$$P(\theta_{t+1} \mid \theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t) := \sum_{s_{t+1} \in \mathcal{S}} \mathbb{I}_{\{\theta_{t+1} = f(\theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t, s_{t+1})\}} P(s_{t+1} \mid s_t), \tag{4.13}$$

where $\mathbb{I}$ denotes the indicator function. The above expression is useful to perform Bayesian updating of the risk-aversion parameter. For a given investor strategy $\sigma^{\mathbf{H}}$, the robo-advisor learns and tracks the risk-aversion parameter using via Bayesian updating. The resulting formula is an extension of that given in Eq. (4.3) for the static case, and takes the following form in the case of dynamically changing risk-aversion

$$\pi_{t+1}^*(\theta_{t+1}) := \frac{\sum_{\theta_t} \pi_t^*(\theta_t) \sigma^{*\mathbf{H}}(a_t^{\mathbf{H}} | s_t, \pi_t^*, \theta_t) P(\theta_{t+1} \mid \theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t)}{\sum_{\tilde{\theta}} \sum_{\theta_t} \pi_t^*(\theta_t) \sigma^{*\mathbf{H}}(a_t^{\mathbf{H}} | s_t, \pi_t^*, \theta_t) P(\tilde{\theta} \mid \theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t)}, \tag{4.14}$$

Figure 4.3 illustrates the estimation process of an error-prone investor on one simulated path of the state process, assuming that risk preferences change dynamically as prescribed by Eq. (4.12). Figure 4.3 shows how the robo-advisor tracks the risk-aversion parameter, as it changes according to market movements and past decisions. Noticeably, the mode of the filtering probability mass adapts to reflect the actual dynamics of the risk-aversion parameter.

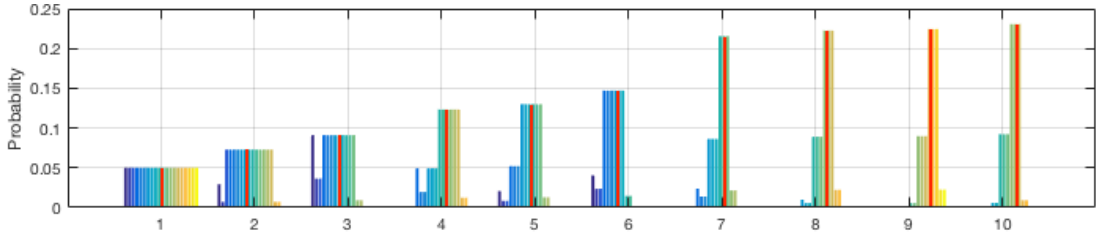


Figure 4.3: Updating of beliefs that track the dynamic risk-aversion parameter, using the Bayesian filtering distribution in Eq. (4.14). We consider an error-prone investor with $P_m = 0.4$ for $D_t \leq 3\%$, and $P_m = 0$ otherwise. The red vertical lines correspond to the value of the true (unknown) risk aversion parameter in that period.

4.5 Conclusion and Future Work

In this chapter, we presented a framework for human-machine decision making, accounting for both human-driven and context-driven risk. Due to the different sensitivities to risk by the human and the machine, respectively, to the context in which the task is being executed and to the category of humans served, the optimal decision making problem may be formulated as a game with strategic interactions. We have introduced the concept of risk-sensitive equilibria to deal with the corresponding game, and shown that it can be computed by solving a risk-POMDP through a coordinator problem.

We have specialized our framework to capture the interactions between an investor and a robo-advising firm. Our numerical study highlights the trade-off between frequent communication of preferences by the investor and the costs of such a communication. If the investor intervenes frequently, the machine can learn the risk-aversion parameter of the investor faster, and therefore make more tailored portfolio decisions. On the other hand, each override decision of the investor is costly, and these total costs may exceed the performance gain stemming from more informed investment decisions by the machine. The robo-advising firm provides a service to the investor that may be superior to a stand-alone investor making all investment decisions on her own. Assuming that override costs occurring in the human-machine system and market research costs occurring in the human-only system are equal, our numerical analysis suggests that the objective risk function is

lower in the human-machine interacting system. More importantly, since human costs are incurred in all periods for the human-only system, an increase in these costs translates to a linear increase in the human-only expected cumulative costs. On the other hand, the cost increase in the human-machine system is bounded. This is because if the investor does not communicate her preferences, the robo-advisor will make portfolio decisions using its initial belief on the human's risk aversion, without updating it. These decisions, however, will not be tailored to the specific risk-profile of the investor.

Future directions for this research include the development of new solution methods to integrate risk optimization techniques with concepts from game theory. A key refinement to equilibrium in dynamic games is the notion of subgame perfection. This enforces incentive compatibility for both agents in each subgame initiated at the start of each period. However, many commonly used risk functions are not time-separable, i.e. the risk over the entire horizon cannot be decomposed into a set of risks, each allocated to a different time period. Without time separability, the risk-POMDP no longer satisfies the Markov property. For example, when an investor chooses an action at a specific time, she may account for the implications of such an investment decision on her future risk preferences. Changes in the investor's risk attitudes depend both on machine observable information, such as the current wealth level of the investor, and on investor-specific information, such as updates on her educational or family status, that is unobserved by the machine. The establishment of an effective communication protocol, accounting for the fact that the investor will optimize a different objective functional at later points in time, is left for future research.

Acknowledgements

This chapter was joint work with Prof. Agostino Capponi and Matt Stern. I am very grateful for their useful ideas, guidance and commitment to this project.

Chapter 5

Information Relaxation Bounds for POMDPs: An Application to Personalized Medicine in Mammography Screening

To date there have been relatively few[1] successful medical applications of POMDPs. The reasons for this include the difficulty of determining a suitable objective to optimize, the difficulty of estimating the POMDP parameters and the general difficulty of solving POMDP problems. Recently Ayer et al. [6] proposed a POMDP formulation with the goal of determining an optimal screening policy for breast cancer, the most common cancer among U.S. women according to the American Cancer Society (ACS). The recommendation guidelines provided by the ACS in 2015 [64] is for women with an average risk of breast cancer to take mammograms beginning at age 45, and to continue annually until age 54. Beginning at age 55, they are then recommended to undergo biannual screenings (but they have the opportunity to continue annually if desired) and to continue taking mammograms as long as their life expectancy is at least 10 years. In addition, the ACS indicates

[1] A review of applications of MDPs and POMDPs to medical decision problems can be found in [73].

that women aged 40 to 44 may choose to begin mammogram screening if desired. In contrast, in 2016 the U.S. Preventive Services Task Force (USPSTF) [77] recommended that women aged 50 to 74 screen biannually using mammography, and they left open the decision for women aged 40-49. In addition, they did not find enough evidence to recommend taking mammograms beyond the age of 75.

5.1 Modeling Screenings as a POMDP

In this chapter we apply the information relaxation approach to the POMDP formulation of Ayer et al. We will use the term decision-maker (DM) to refer to the woman or patient in question but the decision-maker could also refer to a doctor or some other medical professional. We assume the DM has the objective of maximizing her total expected quality-adjusted life years (QALYs). We assume a finite-horizon discrete-time model where the time intervals correspond to six-month periods beginning at age 40 and ending at age 100 so that $t \in \{0, \dots, 120\}$. The hidden state space represents the true health state of the patient with $\mathcal{H} = \{0, 1, 2, 3, 4, 5\}$. Specifically:

- State 0 represents a cancer-free patient.
- States 1 and 2 indicate the presence of *in situ* and *invasive* cancer, respectively.
- States 3 and 4 represent fully observed absorbing states in which the patient has been diagnosed with in situ and invasive cancer, respectively, and has begun treatment.
- State 5 is a fully observed absorbing state representing the death of the patient.

Clearly states 3, 4 and 5 can be explicitly observed and are therefore not actually *hidden*. We include them among the set of hidden states, however, to account for the possible transition dynamics of the other hidden states into these absorbing states. The knowledge of being in these hidden absorbing states can then be modeled correctly through noiseless observations of them. We will refer to the subset of hidden states $\{0, 1, 2\}$ as *pre-cancer* states and the absorbing states $\{3, 4, 5\}$ as *post-cancer* states.

At each time t, the DM can choose to either have a mammography screening (M) or wait (W). If the decision to wait is made, the patient may perform a self-detection screening which will have either a positive or negative result. That is, if through self-detection the patient has reason to be concerned about the presence of cancer, we say the self-test is positive. The possible results of a mammogram are also positive or negative. In the former case, an accurate procedure, e.g. a biopsy, is then prescribed to precisely determine the true cancer status of the patient. If the biopsy result is positive and cancer is found with certainty, the patient will then exit the screening process and move into one of the absorbing states, 3 or 4, to indicate that cancer treatment has commenced. To code this behavior, Ayer uses hidden state transitions that are functions of the observations. To model this behavior as a conventional POMDP (where hidden state transitions do not depend on observations), we introduce an exit action (E) as the only available action after a positive biopsy has been observed. The transition into absorbing state 3 or 4 will now only depend on the current hidden state and the exit action which must be taken if the biopsy result is positive and cancer is found with certainty. The set of possible observations is therefore $\mathcal{O} = \{R-, R+, B_1, B_2, D\}$ where:

- $R-$ is a negative test result (either from a mammography or self-detection).
- $R+$ is a positive test result (including a negative biopsy if the test was a positive mammogram).
- B_1 and B_2 represent in situ cancer and invasive cancer, respectively, and they can be observed via a biopsy following a positive mammogram. If B_1 or B_2 are observed, the action space is then restricted to the exit action E which transfers the patient to the corresponding absorbing state.
- D represents the death of the patient.

We assume a prior probability distribution, π_0, on the true health-state of the woman at age 40. The transition probabilities of the latent pre-cancer health states are assumed to be age-specific and therefore a function of time t. We assume that a screening decision does not influence the development of cancer and therefore have $P_{ij}^t(M) = P_{ij}^t(W)$ for all t and for all $i, j \in \mathcal{H}$. The time

t transition matrices for the screening and wait actions, M and W, are then given by

$$P^t(M) = P^t(W) = \begin{bmatrix} p_{00}^t & p_{01}^t & p_{02}^t & 0 & 0 & m_0^t \\ 0 & p_{11}^t & p_{12}^t & 0 & 0 & m_1^t \\ 0 & 0 & p_{22}^t & 0 & 0 & m_2^t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

where m_i^t represents the mortality rates for each health-state, i, p_{01}^t and p_{02}^t represent the in situ and invasive cancer incidence rates, respectively, and p_{12}^t is the probability that in situ cancer develops into invasive cancer. Recalling that time steps in the POMDP correspond to half-year periods, all rates correspond to effective semiannual rates. Estimates for some of the parameters in (5.1) were obtained from various sources (see Table 5.1 below), and we used reasonable assumptions to estimate the parameters for which we could not find external estimates. We note that we have not conducted a full study on the appropriateness[2] of these parameters, but rather we treat them as ballpark estimates in order to illustrate the information relaxation POMDP methodology. Finally, the exit action, E, will take pre-cancer states to post-cancer treatments with probability 1, i.e. $P_{1,3}^t(E) = 1$ and $P_{2,4}^t(E) = 1$. Since this action is only available to true health-states 1 and 2, we need not define the transitions for other health-states.

The observation probabilities are determined by the accuracy of the examinations, which are commonly referred to as *specificity* and *sensitivity*. The specificity of a test corresponds to the true negative rate, i.e. the probability that a cancer-free woman obtains a negative test result, while the sensitivity of a test is the true positive rate, i.e. the probability of a positive test result given that the woman has cancer. For each test we employ the age-specific sensitivity and specificity factors

[2] Experts in the field of breast cancer could almost certainly provide superior estimates for those parameters where we could not find external estimates.

Parameter	Source
Mortality m_0	SSA Period Life Table, 2013, Female mortality [79]
Mortality m_1, m_2	SEER [45] Table 4.13, all stages and all ages[3]
Incidence p_{01}	SEER Table 4.12, all races
Incidence p_{02}	SEER Table 4.11, all races
Incidence p_{12}	Assumed equal to p_{02}
Initial risk π_0	SEER Table 4.24, female 40-49 [4]

Table 5.1: Sources of the demographic rates for the transition probabilities.

that were computed and reported by Ayer et al. They are:

$$\text{spec}_t(W) = 0.92, \quad \forall t \qquad\qquad \text{sens}_t(W) = 0.44, \quad \forall t$$

$$\text{spec}_t(M) = \begin{cases} 0.889, & \text{if } t \in \{0, \ldots, 19\} \\ 0.893, & \text{if } t \in \{20, \ldots, 39\} \\ 0.897, & \text{if } t \geq 40 \end{cases} \qquad \text{sens}_t(M) = \begin{cases} 0.722, & \text{if } t \in \{0, \ldots, 29\} \\ 0.81, & \text{if } t \in \{30, \ldots, 59\} \\ 0.862, & \text{if } t \geq 60. \end{cases}$$

Using these rates, we define the age-specific observation matrices according to

$$B^t(W) = \begin{bmatrix} \text{spec}_t(W) & 1 - \text{spec}_t(W) & 0 & 0 & 0 \\ 1 - \text{sens}_t(W) & \text{sens}_t(W) & 0 & 0 & 0 \\ 1 - \text{sens}_t(W) & \text{sens}_t(W) & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

[3] We approximated the invasive cancer mortality rate by inferring the 6-month mortality rate from the 5 year survival rate (0.897) and used the maximum of this 6-month rate and the average female 6-month mortality for a woman of that age. We assumed that in situ mortality is equal to the female mortality times 1.02 for women of the same age.

[4] The initial risk for an average woman was taken from the breast cancer prevalence rate (0.9462%) and split 80% for invasive cancer and 20% for in situ cancer, as discussed in [80].

and

$$B^t(M) = \begin{bmatrix} \text{spec}_t(M) & 1-\text{spec}_t(M) & 0 & 0 & 0 \\ 1-\text{sens}_t(M) & 0 & \text{sens}_t(M) & 0 & 0 \\ 1-\text{sens}_t(M) & 0 & 0 & \text{sens}_t(M) & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $B^t_{ij}(a)$ is the probability of observation $j \in \mathcal{O}$ when action a is taken and the hidden state is $i \in \mathcal{H}$. Note that the observability of the hidden absorbing states 3, 4 and 5 is made evident through these matrices. It is worth pointing out that once action E has been chosen, the DM immediately transitions to an absorbing fully-observable state, and therefore there is no need to define $B^t(E)$.

A characteristic of medical decision problems, as pointed out in Ayer et al., is that the observation at time t is a function of the current action, $B_{ij}(a) := \mathbb{P}(o_t = j \mid h_t = i, a_t = a)$, as opposed to a conventional POMDP where the observation is a function of the prior action; see (2.2). This means that events take place in the following order: given a belief state the DM first takes an action, then immediately observes the result of the action and updates the belief, then a transition takes place and the belief is "carried forward". This technicality results in a different version of the standard filtering update in which the transition occurs prior to the observation. Nonetheless, filtering in this non-standard form of the POMDP is stilla straightforward task. And for the same reason, the natural filtration for the medical decision problem is one where $\mathcal{F}_t$ is defined to be the σ-algebra generated by $o_{0:t-1}$, for $t \geq 1$, and with $\mathcal{F}_0$ defined to be the σ-algebra generated by π_0, the prior distribution on the initial hidden state.

We define the reward obtained at time t, $r_t(h_t, a_t, o_t)$, as the expected QALYs between times t and $t+1$ that a person in true health-state h_t would accrue after making decision a_t and obtaining observation o_t. Note that although the reward is a function of the as yet unseen observation (see

previous paragraph), o_t, we can instead use[5] its expected value

$$r_t(h_t, a_t) := \mathbb{E}[r_t(h_t, a_t, o_t) \mid h_t] \tag{5.2}$$

which is easy to calculate and is now in the standard form for a POMDP.

We follow the same calculations as Ayer et al. to define the reward functions. If the patient is in a pre-cancer state $i \in \{0, 1, 2\}$, the wait action reward is given by $r_t(i, W, o_t) = 0.25m_i^t + 0.5(1 - m_i^t)$ where the m_i^t's are the (semi-annual) mortality rates given in Table 5.1. This is the reward for a woman in period t and true pre-cancer health state, i, and in fact does not depend on the observation o_t. Specifically, if death occurs in the next six months (which occurs w.p. m_i^t), it is assumed to happen exactly at the three month mark and so the woman will therefore obtain 0.25 years of lifetime. In contrast, if she survives (which occurs w.p. $1 - m_i^t$) she obtains the 0.5 half-years of lifetime in that period.

For the mammography screening action, we subtract a disutility function, $du(h_t, o_t)$, from the reward so that $r_t(h_t, M, o_t) := r_t(h_t, W, o_t) - du(h_t, o_t)$. The disutility is given a value of 0.5 days for a negative mammogram, two weeks for a true positive mammogram and four weeks for a false positive mammogram. True positive mammograms will in addition force the DM to exit the system in the next period, and provide a lump-sum reward of $R_t(i) := r_t(i, E)$ for $i = 1, 2$. Recall that a true positive mammogram followed by an exit action refers to a woman being accurately diagnosed with cancer and then going into treatment immediately. We expect that a patient under treatment would have a lower remaining expected lifetime than the remaining expected lifetime, $e_t(0)$ say, of a healthy woman of the same age, but higher than the remaining expected lifetime, $e_t(i)$ say, of a woman with cancer $i \in \{1, 2\}$ who is undiagnosed and of the same age. (Note that the expected remaining lifetimes can be calculated using the corresponding mortality rates from times t to T.) We therefore assume $e_t(0) < R_t(i) < e_t(i)$ and in our numerical example, we set $R_t(i) = 0.5e_t(0) + 0.5e_t(i)$ for $i = 1, 2$. We also assume that the absorbing states provide no rewards.

It is perhaps worth noting how the benefit of mammography screening is modelled in our POMDP setting. Specifically, it arises from the possibility of identifying a cancer early and therefore

[5] We acknowledge a slight abuse of notation here in that we are using the same r_t to denote time t rewards $r_t(h_t, a_t)$, $r_t(h_t, a_t, o_t)$ and $r_t(\pi_t, a_t)$. It should be clear from the context what version of the reward we have in mind.

entering treatment and having an expected remaining lifetime that is greater than if the cancer went undiagnosed. The reduced expected lifetime of a woman with an undiagnosed cancer will be reflected via the specific values of the transition and mortality rates of the second and third rows (corresponding to undiagnosed cancer states 1 and 2) in (5.1). There is a cost to mammography screening, however, which is reflected via the disutility function and so the ultimate goal is to find a policy that trades the benefits of mammography screening off against its disutility.

5.1.1 Value Function Approximations

Two methods were used to obtain value function approximations: a QMDP approximation, adapted from the robot navigation problem to include intermediate rewards, and a grid-based approximation. The QMDP approximation is given by

$$\widetilde{V}_t(o_{0:t-1}) := \max_{a_t} \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\mathrm{Q}}(h, a_t) \tag{5.3}$$

with the understanding that at $t = 0$, $\widetilde{V}_0 := \widetilde{V}_0(\pi_0)$, and where V_t^Q is the Q-function of the corresponding fully observable MDP formulation, i.e.

$$V_t^{\mathrm{Q}}(h, a) := r_t(h, a) + \sum_{h' \in \mathcal{H}} P_{hh'}(a) V_{t+1}^{\mathrm{MDP}}(h') \tag{5.4}$$

$$V_t^{\mathrm{MDP}}(h) := \max_{a_t \in \mathcal{A}} V_t^{\mathrm{Q}}(h, a_t) \tag{5.5}$$

for $t \in \{0, \ldots, T\}$ with terminal condition $V_{T+1}^{MDP} := 0$. Note that the only difference between these definitions and those given for the robot navigation application is the inclusion here of intermediate rewards.

The *grid approximation* corresponds to a point-based value iteration method using a fixed and finite grid approximation of the belief space, Π (see [55][42]). A standard approximation tool in dynamic programming is to represent an infinite state space as a finite grid of points, $P \subset \Pi$, and obtain an AVF by linear interpolation for points not in P. Specifically, the AVF is obtained by solving a dynamic program with terminal condition $\widetilde{V}_{T+1} = 0$ and Bellman equation

$$\widetilde{V}_t(\pi) = \max_{a_t} \left[r_t(\pi, a_t) + \sum_o \mathbb{P}_{a_t}(o \mid \pi) \widetilde{V}_{t+1}(f(\pi, a_t, o)) \right] \tag{5.6}$$

for $t \in \{0, \ldots, T\}$, $\pi \in P$ and where $r_t(\pi, a) := \sum_h r_t(h, a)\pi(h)$, $f(\pi, a, o)$ is the belief update function, and $\mathbb{P}_a(o \mid \pi) := \sum_h \mathbb{P}_a(o \mid h)\pi(h)$. Note that in general $f(\pi, a_t, o)$ will not be an element in P and so we use linear interpolation to evaluate the AVF at those points. To tie in the grid approximation with our application, we take the 3-dimensional subspace corresponding to the pre-cancer states

$$\tilde{\Pi} := \{\pi \in \Pi \mid \pi = (\pi_0, \pi_1, \pi_2, 0, 0, 0),\ \pi_0 + \pi_1 + \pi_2 = 1\}$$

of the 6-dimensional simplex Π. We call $\tilde{\Pi}$ the pre-cancer belief space simplex[6] and form a finite grid $P \subset \tilde{\Pi}$. We then solve the dynamic program (5.6) for all elements of P union the elements $(0, 0, 0, 1, 0, 0)$, $(0, 0, 0, 0, 1, 0)$ and $(0, 0, 0, 0, 0, 1)$. For our application, we use a grid P with elements $0.05 \times (i_1, i_2, i_3)$ with i_1, i_2, i_3 integer valued and such that they lie on $\tilde{\Pi}$, i.e. $0.05 \times (i_1 + i_2 + i_3) = 1$.

We can now generate lower bounds on the optimal value function, $V_0^*(\pi_0)$, by simulating the policies that are greedy w.r.t. each of the value function approximations. We will compare the performance of these greedy policies to the official policies recommended by ACS and USPSTF.

5.1.2 The Uncontrolled Formulation

The action-independent transition and emission matrices are built using different approaches for each AVF. First, using the fact that the QMDP approximation is a supersolution, we can drop the absolute continuity requirement and set the transition matrices, Q^t, using (A.7) and, similarly, we set the uncontrolled emission matrices according to

$$E_{ij}^t \equiv B_{ij}\big(\operatorname*{argmax}_{a \in \mathcal{A}} V_t^Q(i, a)\big). \tag{5.7}$$

In contrast, there is no guarantee that the AVF based on the grid approximation is a supersolution and so we must satisfy the absolute continuity conditions. To achieve this, we add a small positive quantity $\epsilon = 0.001$ to each Q_{ij}^t if j can be reached from i under some action, and then normalize the probabilities. Similarly, we add ϵ to B_{ik}^t only if k can be observed from state i under some action and again we then normalize the probabilities. This approach allows our transition and emission

[6] Although the dimension of the hidden state space is 6, in reality the uncertainty in the process is entirely restricted to the 3 pre-cancer states. We can therefore reduce our analysis to the 3-dimensional pre-cancer belief space simplex.

probabilities to satisfy absolute continuity for the PI relaxation. For the BSPI relaxation we would need to make an additional adjustment (as described in Appendix A.1.1) but the BSPI results were slightly inferior to the PI results (as was the case with the maze application) and so we don't report them in our numerical results.

5.1.3 Numerical Results

We consider two different test cases: case 1 represents a woman at age 40 with an average risk of having cancer and therefore an initial distribution over hidden states given by

$$\pi_0 = [0.9905, 0.0019, 0.0076, 0, 0, 0].$$

Case 2 represents a woman at age 40 with a high-risk of having cancer; she has an initial distribution of $\pi_0 = [0.96, 0.02, 0.02, 0, 0, 0]$. In Figure 5.1a we display the lower bounds obtained by simulating each of the four policies, namely the policies recommended USPSTF and ACS, as well as the policies that are greedy w.r.t. the QMDP and grid-based AVFs. We note that the latter two policies outperform the official recommendations of USPSTF and ACS, with the best lower bound coming from the grid approximation.

Figure 5.1b displays the upper bounds obtained with the PI relaxation using penalties constructed from each of the two AVFs. Since the QMDP AVF is a supersolution and therefore also an upper bound we also plot its value in the figure. As a reference, we also display the value of the best lower bound to obtain a visual representation of the duality gap. The duality gap reduction of the best dual bound with respect to the supersolution is 57% in case 1, and 51% in case 2, or equivalently, 19.7 and 29.6 days respectively.

In Ayer et al., the authors were able to solve the POMDP to optimality using Monahan's algorithm [59] with Eagle's reduction [29]. The authors used an Intel Xeon 2.33 GHz processor with 16 GB RAM for their computations, and were able to solve the problem in 55.95 hours. As with our robot navigation application, we used MATLAB Release 2016b, and a MacOS Sierra with 1.3 GHz Intel Core i5 processor with 4 GB RAM. The numerical results in Table 5.2 display the running times and other statistics for the various Case 1 bounds as well as the best bounds in bold font. As we have noted, our bound approximations result in a very tight duality gap (19.7 days or

Bound	Expected value	Standard dev.	Number of paths	Running time
USPSTF (Lower)	41.15	0.0188	400,000	6.75 mins.
ACS (Lower)	41.14	0.0185	400,000	6.79 mins.
Greedy QMDP (Lower)	41.56	0.0160	500	11.1 secs.*
Greedy Grid (Lower)	**41.72**	0.0128	800,000	35.05 mins
Grid PI (Upper)	**41.77**	0.0001	100	9.9 secs.
QMDP PI (Upper)	41.83	0.00004	500	8.1 secs.
QMDP Supersol. (Upper)	41.84	-	1	0.02 secs.

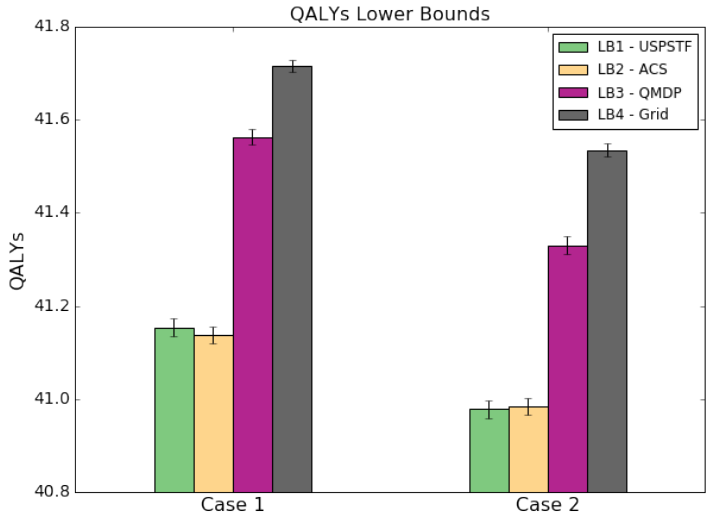
*Lower bound for QMDP greedy strategy was estimated using the penalties as control variates - see Appendix A.4

Table 5.2: Summary statistics for the lower and upper bounds for the Case 1 scenario.

0.054 QALYs for an average woman) and we were able to obtain the best lower and upper bounds in 35.05 minutes and 9.9 seconds, respectively, with narrow confidence intervals. So while Ayer et al. were able to solve the problem to optimality, we were able to get provably close to optimality using[7] a slower processor and less RAM with a total runtime that was approximately 2 orders of magnitude smaller. We also note that even tighter bounds information relaxation bounds should be attainable here if so desired using a *partially controlled* formulation as introduced in BH.

[7] We do not know what software Ayer et al. used

(a)



(b)

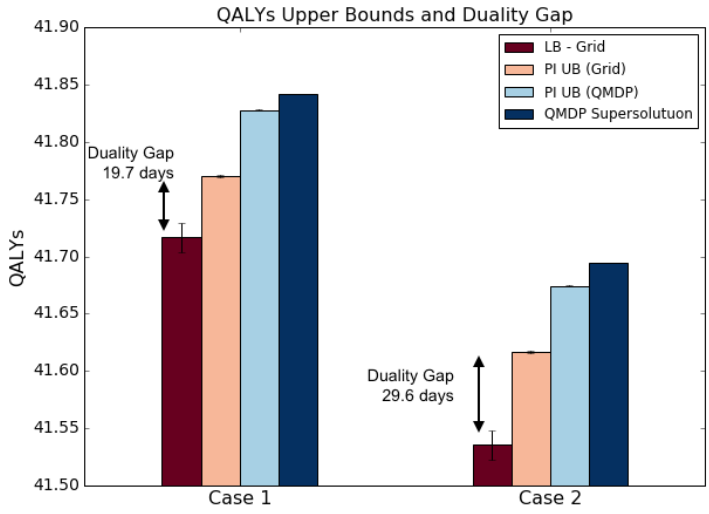


Figure 5.1: (a) Lower bounds on the optimal value function obtained from simulating the USPSTF and ACS recommended policies as well as policies that are greedy w.r.t. the QMDP and grid-based AVFs. Case 1 corresponds to an average risk 40-year old woman while case 2 corresponds to a high risk 40-year old woman. The vertical lines on each bar represent 95% confidence intervals. (b) Upper bounds on the optimal value function compared to the best lower bound which was obtained by simulating the policy that is greedy w.r.t. the grid-based AVF. The best upper bound was also obtained by constructing penalties for the PI relaxation from the grid-based AVF. The optimal duality gap is displayed in each case.

Bibliography

[1] N. Aleksandrov and B.M. Hambly. A dual approach to multiple exercise option problems under constraints. *Mathematical Methods of Operations Research*, 71(3):503–533, 2010.

[2] L. Andersen and M. Broadie. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science*, 50(9):1222–1234, 2004.

[3] A.Y. Aravkin, J.V. Burke, and G. Pillonetto. Optimization viewpoint on kalman smoothing, with applications to robust and sparse estimation. In A.Y. Carmi, L.S. Mihaylova, and S.J. Godsill, editors, *Compressed Sensing and Sparse Filtering*, pages 237–280. Springer, Berlin, Germany, 2014.

[4] A.Y. Aravkin, J.V. Burke, and G. Pillonetto. Robust and trend following student's t kalman smoothers. *SIAM J. Control Optim.*, 52(5):2891–2916, 2014.

[5] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 2015.

[6] Turgay Ayer, Oguzhan Alagoz, and Natasha K. Stout. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012.

[7] S. R. Balseiro, D. B. Brown, and C. Chen. Static routing in stochastic scheduling: performance guarantees and asymptotic optimality. Working paper, Duke University, 2016.

[8] S.R. Balseiro and D.B. Brown. Approximations to stochastic dynamic programs via information relaxation duality. Working paper, Duke University, 2016.

[9] B. Barber and T. Odean. Risk-sensitive Markov decision processes. *Quarterly Journal of Economics*, 116(1):261–292, 2001.

[10] Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

[11] Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[12] C. Bender. Primal and dual pricing of multiple exercise options in continuous time. *SIAM Journal of Financial Mathematics*, 2:562–586, 2011.

[13] C. Bender, C. Gartner, and N. Schweizer. Pathwise dynamic programming. Preprint, 2016.

[14] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 4th edition, 2017.

[15] D. B. Brown and M. B. Haugh. Information relaxation bounds for infinite horizon markov decision processes. *Operations Research*, 65(5):1355–1379, 2017.

[16] D. B. Brown and J. E. Smith. Information relaxations, duality, and convex stochastic dynamic programs. *Operations Research*, 62(6):1394–1415, 2014.

[17] D. B. Brown, J. E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4-part-1):785–801, 2010.

[18] D.B. Brown and J.E. Smith. Dynamic portfolio optimization with transaction costs: heuristics and dual bounds. *Management Science*, 57(10):1752–1770, 2011.

[19] Anthony R. Cassandra. *Exact and Approximate Algorithms for Partially Observed Markov Decision Processes*. PhD thesis, Brown University, 1998.

[20] S. Chandramouli and M.B. Haugh. A unified approach to multiple stopping and duality. *Operations Research Letters*, 2012.

[21] N. Chen and P. Glasserman. Additive and multiplicative duals for american option pricing. *Finance and Stochastics*, 11:153–179, 2007.

[22] P.F. Christoffersen and D. Pelletier. Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics*, 2(1):84–108, 2004.

[23] Min Dai, Jin Hanqing, Steven Kou, and Yuhong Xu. A dynamic mean-variance analysis with application to robo-advising. *Working Paper, National University of Singapore*, pages 33–94, 2018.

[24] S. Das, Harry Markowitz, Jonathan Scheid, and M. Statman. Portfolio optimization with mental accounts. *Journal of Financial and Quantitative Analysis*, 45(2):331–334, 2000.

[25] S.R. Das, D. Ostrov, A. Radhakrishnan, and D. Srivastav. A new approach to goals-based wealth management. *Journal of Investment Management*, 16(3):1—27, 2018.

[26] E. Derman and M.B. Miller. *The Volatility Smile*. John Wiley & Sons, 2016.

[27] V.V. Desai, V.F. Farias, and C.C. Moallemi. Bounds for markov decision processes. In *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, (F. L. Lewis, D. Liu, eds.)*, pages 452–473. IEEE Press, December 2012.

[28] F. X. Diebold and C. Li. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130:337–364, 2006.

[29] J N Eagle. The optimal search for a moving target when the search path is constrained. *Operations Research*, 32(5):1107–1115, 1984.

[30] A. Federgruen, D. Guetta, and G. Iyengar. Information relaxation-based lower bounds for the stochastic lot sizing problem with advanced demand information. Working paper, Columbia University, 2015.

[31] J. Filar. The variance of discounted Markov decision processes. *IEEE Transaction of Automatic Control*, 40(1):2–10, 1985.

[32] J. Da Fonseca and K. Gottschalk. A joint analysis of the term structure of credit default swap spreads and the implied volatiltiy surface. *Journal of Futures Markets*, 33(6):494–517, 2013.

[33] Drew Fudenberg and Jean Tirole. Game theory, 1991. *Cambridge, Massachusetts*, 1991.

[34] N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.

[35] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 25, 2017.

[36] William B Haskell and Rahul Jain. A convex analytic approach to risk-aware Markov decision processes. *SIAM Journal on Control and Optimization*, 53(3):1569–1598, 2015.

[37] M. B. Haugh, G. Iyengar, and C. Wang. Tax-aware dynamic asset allocation. *Operations Research*, 64(4):849–866, 2016.

[38] M. B. Haugh and L. Kogan. Pricing american options: A duality approach. *Operations Research*, 52(2):258–270, 2004.

[39] M. B. Haugh and O. Ruiz-Lacedelli. Information relaxation bounds for partially observed Markov decision processes. *Working Paper, Columbia University*, 2018.

[40] M. B. Haugh and C. Wang. Dynamic portfolio execution and information relaxations. *SIAM J. Financial Math.*, 5:316–359, 2014.

[41] M. B. Haugh and C. Wang. Information relaxations and dynamic zero-sum games. Working paper, Columbia University, 2014.

[42] M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

[43] M. C. Horsch and D. Poole. An anytime algorithm for decision making under uncertainty. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 246–55, 1998.

[44] R. Howard and J. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

[45] N. Howlader, A.M. Noone, M. Krapcho, and et al. (editors). SEER Fast Stats, 2009 - 2013. National Cancer Institute. Bethesday, MD., 2016.

[46] F.V. Jensen and T.D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2007.

[47] L.P. Kaelbling, M.L. Littman, and A.R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[48] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions of the ASME*, 82:35–46, 1960.

[49] L. Kogan and I. Mitra. Accuracy verification for numerical solutions of equilibrium models. Working paper, Massachusetts Insitute of Technology, 2013.

[50] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[51] V. Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambride University Press, 2016.

[52] P.H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2):73–84, 1995.

[53] G. Lai, F. Margot, and N. Secomandi. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Operations Research*, 58(3):564–582, 2010.

[54] M.L. Littman, A.R. Cassandra, and L.P. Kaelbling. Learning policies for partially observable environments. In *Proceedings of the 12th International Conference on Machine Learning*, pages 362–70, 1995.

[55] W.S. Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

[56] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[57] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: concepts, techniques, and tools*. Princeton university press, 2015.

[58] N. Meinshausen and B.M. Hambly. Monte carlo methods for the valuation of multiple-exercise options. *Mathematical Finance*, 14(4):557–583, 2004.

[59] G E Monahan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[60] S. Natenberg. *Option Pricing and Volatility: Advandec Trading Strategies and Techniques*. McGraw-Hill, 2nd edition, 1994.

[61] A Nayyar, A Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644—1658, 2013.

[62] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.

[63] D. Nilsson and M. Hohle. Computing bounds on expected utilities for optimal policies based on limited information. Technical report, Dina Research, 2001.

[64] K.C. Oeffinger, E.H. Fontham, R. Etzioni, and et al. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *JAMA*, 314(15):1599–1614, 2015.

[65] R. Rebonato. A bayesian approach to stress testing and scenario analysis. *Journal of Investment Management*, 8(3):1–13, 2010.

[66] R. Rebonato. *Coherent Stress Testing: A Bayesian Approach to the Analysis of Financial Stress*. John Wiley & Sons, 2010.

[67] Alejandro Rodriguez and Esther Ruiz. Bootstrap prediction intervals in state-space models. *Journal of Time Series Analysis*, 30(2):167–178, 2009.

[68] L.C.G. Rogers. Monte carlo valuation of american options. *Mathematical Finance*, 12:271–286, 2002.

[69] L.C.G. Rogers. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization*, 46:1116–1132, 2007.

[70] L.C.G. Rogers and M.R. Tehranchi. Can the implied volatility surface move by parallel shifts? *Finance and Stochastics*, 14(2):235–248, 2010.

[71] A. Ruszczynski. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125:235–261, 2010.

[72] A. Ruszczynski and A. Shapiro. Optimization of risk measures. In *Probabilistic and Randomized Methods for Design under Uncertainty*, pages 117—158. Springer-Verlag, 2005.

[73] Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. Modeling medical treatment using markov decision processes. In *Operations research and health care*, pages 593–612. Springer, 2005.

[74] J. Schoenmakers. A pure martingale dual for multiple stopping. *Finance and Stochastics*, pages 1–16, 2010.

[75] A. Shapiro, D. Dentcheva, and A.P. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial Mathematics, 2009.

[76] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer International Publishing, 4 edition, 2017.

[77] A.L. Siu and U.S. preventive services task force. Screening for breast cancer: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine*, 164:279–96, 2016.

[78] M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

[79] Social Security Administration. Period life table, 2013.

[80] B. L. Sprague and A. Trentham-Dietz. Prevalence of breast carcinoma in situ in the united states. *JAMA: The Journal of the American Medical Association*, 302(8):846–848, 2009.

[81] David S. Stoffer and Kent D. Wall. Bootstrapping state-space models: Gaussian maximum likelihood estimation and the kalman filter. *Journal of the American Statistical Association*, 86(416):1024–1033, 1991.

[82] R. S. Tsay. *Analysis of Financial Time Series*. John Wiley & Sons, 2010.

[83] Deepanshu Vasal and Achilleas Anastasopoulos. Signaling equilibria for dynamic lqg games with asymmetric information. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 6901–6908. IEEE, 2016.

[84] Deepanshu Vasal and Achilleas Anastasopoulos. A systematic process for evaluating structured perfect bayesian equilibria in dynamic games with asymmetric information. In *American Control Conference (ACC), 2016*, pages 3378–3385. IEEE, 2016.

[85] E. Wan and R. Van der Merwe. The unscented kalman filter for nonlinear estimation. *Proc. Adaptive Systems for Signal Process., Commun., and Control Symp.. AS-SPCC.*, pages 153–158, 2000.

[86] F. Ye and E. Zhou. Information relaxation and dual formulation of controlled markov diffusions. *IEEE Transactions on Automatic Control*, 2014.

[87] F. Ye, H. Zhu, and E. Zhou. Weakly coupled dynamic program: Information and lagrangian relaxations. Working paper, Georgia Institute of Technology, 2014.

[88] H. Zhu, F. Ye, and E. Zhou. Solving the dual problems of dynamic programs via regression. arXiv preprint arXiv:1610.07726, 2016.

Appendix A

Chapter 2 - Supplemental Content

A.1 RN Derivative Calculations

A.1.1 The Uncontrolled Belief State POMDP Formulation

In order to compute the RN derivatives for the uncontrolled belief state POMDP formulation we must first define the uncontrolled belief-state dynamics for π_t which lies in the $|\mathcal{H}|$-dimensional simplex. Note that while there are infinitely many points in the simplex only a finite number of these points will have a strictly positive probability under $\mathbb{P}$ conditional on π_0 which is assumed known. These points with strictly positive $\mathbb{P}$-probability arise from the various possible combinations of action / observation sequences which are finite in number by assumption.

An obvious approach to defining uncontrolled belief-state dynamics for π_t would be to use (2.17) and (2.18) to generate uncontrolled hidden state / observation sequences and then simply use the generated observations to update the belief state appropriately, beginning with π_0. The only problem with this is that $\mathbb{P}$ will not be absolutely continuous w.r.t $\tilde{\mathbb{P}}$ even if Q and E as defined in (2.17) and (2.18) do satisfy their absolute continuity conditions. To see this note that the belief state updates under $\mathbb{P}$ are computed according to

$$\pi_{t+1}(h';a,o) = \frac{\sum_h \pi_t(h) P_{hh'}(a) B_{h'o}(a)}{\sum_{h,h'} \pi_t(h) P_{hh'}(a) B_{h'o}(a)} \tag{A.1}$$

where we explicitly recognize the $\mathbb{P}$-dependence of π_{t+1} on $a_t = a$ and $o_{t+1} = o$. In contrast, the

belief state updates under $\tilde{\mathbb{P}}$ are computed according to

$$\tilde{\pi}_{t+1}(h') = \frac{\sum_h \pi_t(h) Q_{hh'} E_{h'o}}{\sum_{h,h'} \pi_t(h) Q_{hh'} E_{h'o}}. \tag{A.2}$$

Even if Q and E satisfy their absolute continuity conditions, there will in general be $\pi_{t+1}(\cdot\,;a,o)$'s that satisfy $\mathbb{P}\left(\pi_{t+1}(\cdot\,;a,o) \mid \pi_t\right) > 0$ and $\tilde{\mathbb{P}}\left(\pi_{t+1}(\cdot\,;a,o) \mid \pi_t\right) = 0$. As such, $\mathbb{P}$ will not be absolutely continuous w.r.t. $\tilde{\mathbb{P}}$. There are many ways to resolve this issue and we mention just two of them:

1. We can instead assume that under $\tilde{\mathbb{P}}$ the current belief state π transitions with strictly positive probability to any belief state π' which is feasible for some available action $a \in \mathcal{A}$ given π. Specifically, we define the belief-state transition probability

$$\tilde{\mathbb{P}}(\pi' \mid \pi) := \frac{1}{|\mathcal{A}| \times |\mathcal{O}|} \sum_{(a,o) \in \mathcal{A} \times \mathcal{O}} \mathbf{1}_{\{\pi' = f(\pi;a,o)\}} \tag{A.3}$$

   where each component of $f(\pi;a,o)$ in the $|\mathcal{H}|$-dimensional simplex is defined according to (A.1) with $\pi_t = \pi$. While it is of course possible to define other $\tilde{\mathbb{P}}$'s, (A.3) seems like a particularly easy way (in our finite action and observation setting) to guarantee that $\mathbb{P}$ is absolutely continuous w.r.t. $\tilde{\mathbb{P}}$.

2. As before use (2.17) and (2.18) to generate action-independent hidden state and observation sequences. Given these sequences, we generate an action $a \in \mathcal{A}$ randomly (with each a having strictly positive probability) and then generate π_{t+1} using (A.1) (rather than (A.2)). It is also straightforward to write down $\tilde{\mathbb{P}}(\pi' \mid \pi)$ for this absolutely continuous change-of-measure.

Regardless of the specific form of $\tilde{\mathbb{P}}$, the RN derivatives take the form

$$\begin{aligned} \frac{d\mathbb{P}}{d\tilde{\mathbb{P}}} =: \Phi_T^\pi(\pi_{0:T}, a_{0:T-1}) &:= \prod_{s=0}^{T-1} \phi(\pi_s, \pi_{s+1}, a_s) \\ \phi(\pi, \pi', a) &:= \frac{\sum_{i,j,k} \pi(i) P_{ij}(a) B_{jk}(a) \mathbf{1}_{\{\pi' = f(\pi,a,k)\}}}{\tilde{\mathbb{P}}(\pi' \mid \pi)}. \end{aligned} \tag{A.4}$$

In order to justify (A.4) we first express the time t reward as a function of the belief state according to $r_t(\pi_t, a_t) := \mathbb{E}[r_t(h_t, a_t) \mid \mathcal{F}_t^\pi] = \sum_{h_t} r_t(h_t, a_t)\pi_t(h_t)$. The RN derivatives must then satisfy (by a standard conditioning argument to obtain the second equality)

$$\mathbb{E}\left[r_t(\pi_t, a_t) \,\middle|\, \mathcal{F}_0^\pi\right] = \widetilde{\mathbb{E}}\left[\Phi_T^\pi r_t(\pi_t, a_t) \,\middle|\, \mathcal{F}_0^\pi\right] = \widetilde{\mathbb{E}}\left[\Phi_t^\pi r_t(\pi_t, a_t) \,\middle|\, \mathcal{F}_0^\pi\right].$$

Writing the expectations explicitly, we must have

$$\sum_{\pi_{1:t}} r_t(\pi_t, a_t)\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) = \sum_{\pi_{1:t}} \Phi_t^\pi r_t(\pi_t, a_t)\tilde{\mathbb{P}}(\pi_{1:t})$$

where $\mathbb{P}_{a_{0:t-1}}$ explicitly recognizes the dependence of the given probabilities on $a_{0:t-1}$ and $\pi_{1:t} := \{\pi_1, \ldots, \pi_t\}$. It is clear then that the RN derivatives must satisfy

$$\Phi_t^\pi := \frac{\mathbb{P}_{a_{0:t-1}}(\pi_{1:t})}{\tilde{\mathbb{P}}(\pi_{1:t})}. \tag{A.5}$$

We can compute the numerator of (A.5) as

$$\begin{aligned}
\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) &= \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(\pi_{s+1} \mid \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{o_{s+1}} \mathbb{P}_{a_s}(o_{s+1} \mid \pi_s)\mathbb{P}_{a_s}(\pi_{s+1} \mid o_{s+1}, \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o_{s+1}} \pi_s(h)\mathbb{P}_{a_s}(h' \mid h)\mathbb{P}_{a_s}(o_{s+1} \mid h')\mathbf{1}_{\{\pi_{s+1}=f(\pi_s,a_s,o_{s+1})\}} \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o} \pi_s(h)P_{hh'}(a_s)B_{h'o}(a_s)\mathbf{1}_{\{\pi_{s+1}=f(\pi_s,a_s,o)\}}.
\end{aligned} \tag{A.6}$$

Substituting $\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)$ and (A.6) into (A.5) then establishes that (A.4) is correct.

The Robotic Navigation Application

It is worth emphasizing that in the numerical results of Sections 2.7 and 2.8, our penalties were constructed using supersolutions. As explained in Appendix A.4, the absolute continuity of $\mathbb{P}$ w.r.t. $\tilde{\mathbb{P}}$ is no longer required in this case and so in fact we did not need to define $\tilde{\mathbb{P}}$ using either of the two options described above. Instead we defined $\tilde{\mathbb{P}}$ to be the measure induced by following the policy that was greedy with respect to the AVF under consideration, i.e. the QMDP, Lag-1 or Lag-2 AVF. In the case of robotic navigation application of Section 2.7, the action-independent transition probabilities[1] induced by following the policy that is greedy with respect to the QMDP AVF were

[1] Recall that the emission matrix B was already action-independent, and so we continued to use B in the uncontrolled formulation

defined according to

$$Q_{ij}^t \equiv P_{ij}\big(\operatorname*{argmax}_{a\in\mathcal{A}} V_t^Q(i,a)\big) \tag{A.7}$$

for $t \in \{0,\ldots,T-1\}$. Similarly, the action-independent transition probabilities induced by following the policy that is greedy with respect to the Lag-1 AVF (C.7) were defined according to

$$Q_{ij}^t \equiv P_{ij}\left(\operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[r_t(h_t,a) + V_{t+1}^{\mathrm{L}_1}(h_t,a,o_{t+1}) \mid h_t = i\right]\right) \tag{A.8}$$

and for the Lag-2 AVF (2.44) we defined

$$Q_{ij}^t \equiv P_{ij}\left(\operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[\max_{a_{t+1}} \mathbb{E}\left[r_t(h_t,a) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{L_2}(h_t,a_{t:t+1},o_{t+1:t+2}) \mid h_t = i, o_{t+1}\right] \mid h_t = i\right]\right). \tag{A.9}$$

For each AVF under consideration, the denominator of (A.5) is computed as

$$\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s) = \prod_{s=0}^{t-1} \sum_{h,h',o} \pi_s(h) Q_{hh'}^s B_{h'o} \mathbf{1}_{\{\pi_{s+1} = \tilde{f}_s(\pi_s,o)\}} \tag{A.10}$$

where $Q_{hh'}^s$ is given by either (A.7), (A.8) or (A.9), and where each component of $\tilde{f}_s(\pi_s;o)$ in the $|\mathcal{H}|$-dimensional simplex is defined according to

$$\tilde{f}_s(\pi_s;o)(h') = \frac{\sum_h \pi_s(h) Q_{hh'}^s B_{h'o}}{\sum_{h,h'} \pi_s(h) Q_{hh'}^s B_{h'o}}. \tag{A.11}$$

A.1.2 The Uncontrolled Non-Belief-State POMDP Formulation

To show that the general RN derivatives in (2.19) and (2.20) are correct under the PI relaxation framework, it suffices to prove that $\mathbb{E}\left[\sum_{t=0}^T r_t(h_t,a_t) \,\middle|\, \mathcal{F}_0\right] = \widetilde{\mathbb{E}}\left[\sum_{t=0}^T \Phi_t r_t(h_t,a_t) \,\middle|\, \mathcal{F}_0\right]$ or equivalently that

$$\mathbb{E}\left[r_t(h_t,a_t) \,\middle|\, \mathcal{F}_0\right] = \widetilde{\mathbb{E}}\left[\Phi_t r_t(h_t,a_t) \,\middle|\, \mathcal{F}_0\right] \tag{A.12}$$

for all $t \in \{0,\ldots,T\}$, where we recall that $\mathbb{E}$ and $\widetilde{\mathbb{E}}$ correspond to expectations under $\mathbb{P}$ and $\tilde{\mathbb{P}}$, respectively. We first write the expectation on the r.h.s. of (A.12) explicitly to obtain

$$\sum_{o_{1:t},h_{0:t}} \Phi_t(h_{0:t},o_{1:t},a_{0:t-1}) r_t(h_t,a_t) \tilde{\mathbb{P}}(o_{1:t},h_{0:t} \mid \pi_0) \tag{A.13}$$

where π_0 is the initial hidden state distribution. Recalling (2.19) and (2.20) we have

$$\begin{aligned}\Phi_t(\cdot) = \prod_{s=0}^{t-1} \frac{P_{h_s h_{s+1}}(a_s) B_{h_{s+1} o_{s+1}}(a_s)}{Q_{h_s h_{s+1}} E_{h_{s+1} o_{s+1}}} &\equiv \prod_{s=0}^{t-1} \frac{\mathbb{P}_{a_s}(h_{s+1} \mid h_s)\mathbb{P}_{a_s}(o_{s+1} \mid h_{s+1})}{\tilde{\mathbb{P}}(h_{s+1} \mid h_s)\tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1})} \\ &= \frac{\mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{0:t} \mid h_0)\pi_0(h_0)}{\tilde{\mathbb{P}}(o_{1:t}, h_{0:t} \mid h_0)\pi_0(h_0)} \\ &= \frac{\mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{0:t} \mid \pi_0)}{\tilde{\mathbb{P}}(o_{1:t}, h_{0:t} \mid \pi_0)} \end{aligned} \tag{A.14}$$

where $\mathbb{P}_{a_{0:t-1}}$ and $\mathbb{P}_{a_s}$ explicitly recognize the dependence of the given probabilities on $a_{0:t-1}$ and a_s, respectively. If we substitute (A.14) into (A.13) we obtain

$$\sum_{o_{1:t}, h_{0:t}} r_t(h_t, a_t)\mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{0:t} \mid \pi_0) = \mathbb{E}\left[r_t(h_t, a_t) \,\middle|\, \mathcal{F}_0\right] \tag{A.15}$$

which establishes the correctness of the RN derivatives in (2.19) and (2.20). Once again for the robotic navigation application, we did not need to impose absolute continuity of the measure change due to our use of supersolution AVFs and so we used the Q's of (A.7), (A.8) or (A.9) in the denominator of (A.14).

A.2 The Lag-1 and Lag-2 Approximate Value Functions

A.2.1 Computing the Optimal Value Function for the Lag-1 MDP

The Lag-1 formulation corresponds to the relaxed problem in which the time t DM knows the true state h_{t-1} that prevailed at time $t-1$, the observation history $o_{0:t}$ and the action history $a_{0:t-1}$. Given the dependence structure of the hidden states and observations in the POMDP, it follows that the Lag-1 optimal value function $V_t^{\mathrm{L}_1}$ only depends on (h_{t-1}, a_{t-1}, o_t). The terminal value function satisfies $V_T^{L_1}(h_{T-1}, a_{T-1}, o_T) := r_T(o_T) = r_T(h_T)$ with

$$\begin{aligned} V_t^{\mathrm{L}_1}(h_{t-1}, a_{t-1}, o_t) &:= \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_1}(h_t, a_t, o_{t+1}) \mid h_{t-1}, o_t] \\ &= \max_{a_t} \sum_{h_t, o_{t+1}} \mathbb{P}_{a_{t-1:t}}(h_t, o_{t+1} \mid h_{t-1}, o_t)[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_1}(h_t, a_t, o_{t+1})] \end{aligned}$$

for $t \in \{1, \ldots, T-1\}$ and where $\mathbb{P}_{a_{t-1:t}}$ recognizes the dependence of the conditional PMF on the actions $a_{t-1:t}$. These probabilities can be calculated explicitly using standard manipulations. In

particular, we have

$$\begin{aligned}\mathbb{P}_{a_{t-1:t}}(h_t, o_{t+1} \mid h_{t-1}, o_t) &= \frac{\mathbb{P}_{a_{t-1:t}}(h_t, o_t, o_{t+1} \mid h_{t-1})}{\mathbb{P}_{a_{t-1:t}}(o_t \mid h_{t-1})} \\ &= \frac{\sum_{h_{t+1}} \mathbb{P}_{a_{t-1:t}}(h_t, o_t, h_{t+1}, o_{t+1} \mid h_{t-1})}{\sum_{h_t} \mathbb{P}_{a_{t-1:t}}(h_t, o_t \mid h_{t-1})} \\ &= \frac{P_{h_{t-1}h_t} B_{h_t o_t} \sum_{h_{t+1}} P_{h_t h_{t+1}} B_{h_{t+1} o_{t+1}}}{\sum_{h_t} P_{h_{t-1}h_t} B_{h_t o_t}} \end{aligned} \tag{A.16}$$

where for ease of exposition we have suppressed[2] the dependence of the various quantities in (A.16) on the various actions. We can calculate $V_0^{L_1}$ in a similar fashion by noting that

$$\begin{aligned} V_0^{\mathrm{L}_1}(o_0) &:= \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + V_1^{L_1}(h_0, a_0, o_1) \mid o_0] \\ &= \max_{a_0} \sum_{h_0, o_1} \mathbb{P}_{a_0}(h_0, o_1 \mid o_0)[r_0(h_0, a_0) + V_1^{\mathrm{L}_1}(h_0, a_0, o_1)] \end{aligned}$$

where $\mathbb{P}_{a_0}(h_0, o_1 \mid o_0)$ can be calculated as in (A.16) but with $P_{h_{t-1}h_t}(a_{t-1})$ replaced by $P(h_0)$.

A.2.2 The Lag-2 Approximate Value Function

We must first show how the optimal value function for the Lag-2 MDP can be calculated.

Computing the Optimal Value Function for the Lag-2 MDP

The Lag-2 formulation corresponds to the relaxed problem in which the time t DM knows the true state h_{t-2} that prevailed at time $t-2$, the observation history $o_{0:t}$ and the action history $a_{0:t-1}$. The terminal value function satisfies $V_T^{\mathrm{L}_2}(h_{T-2}, a_{T-2:T-1}, o_{T-1:T}) := r_T(o_T) = r_T(h_T)$ with

$$\begin{aligned} &V_t^{\mathrm{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t}) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid h_{t-2}, o_{t-1:t}] \\ &= \max_{a_t} \sum_{h_{t-1:t}, o_{t+1}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t+1} \mid h_{t-2}, o_{t-1:t})[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1})] \end{aligned} \tag{A.17}$$

[2] In these appendices we will often suppress the dependence of the various transition and observation probabilities on the chosen actions. For example, it should be clear in (A.16) that $P_{h_{t-1}h_t}$ depends on a_{t-1} while $B_{h_{t+1}o_{t+1}}$ depend on a_t.

for $t \in \{2, \ldots, T-1\}$ and where we use $\mathbb{P}_{a_{t-2:t}}$ to denote a probability that depends on $a_{t-2:t}$. We note it is straightforward to calculate $\mathbb{P}_{a_{t-2:t}}(\cdot \mid \cdot)$ using standard arguments. Specifically, we have

$$\begin{aligned}\mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t+1} \mid h_{t-2}, o_{t-1:t}) &= \frac{\mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t-1:t+1} \mid h_{t-2})}{\mathbb{P}_{a_{t-2:t}}(o_{t-1:t} \mid h_{t-2})} \\ &= \frac{\sum_{h_{t+1}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t+1}, o_{t-1:t+1} \mid h_{t-2})}{\sum_{h_{t-1:t}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t-1:t} \mid h_{t-2})} \\ &= \frac{PB_{t-2} PB_{t-1} \sum_{h_{t+1}} PB_t}{\sum_{h_{t-1}, h_t} PB_{t-2} PB_{t-1}} \end{aligned} \tag{A.18}$$

where we use PB_t to denote $P_{h_t h_{t+1}} B_{h_{t+1} o_{t+1}}$ and again we have suppressed the action dependence of the various terms. A slightly different calculation is required for each of $V_0^{\mathrm{L}2}$ and $V_1^{\mathrm{L}2}$ as there is no hidden state information available at times $t = 0$ and $t = 1$. For $t = 1$ we have

$$\begin{aligned} V_1^{\mathrm{L}2}(o_{0:1}, a_0) &:= \max_{a_1} \mathbb{E}[r_1(h_1, a_1) + V_2^{\mathrm{L}2}(h_0, a_{0:1}, o_{1:2}) \mid o_{0:1}] \\ &= \max_{a_1} \sum_{h_0, o_2} \mathbb{P}_{a_{0:1}}(h_{0:1}, o_1 \mid o_{0:1})[r_1(h_1, a_1) + V_2^{\mathrm{L}2}(h_0, a_{0:1}, o_{1:2})] \end{aligned}$$

where $\mathbb{P}_{a_{0:1}}(h_{0:1}, o_1 \mid o_{0:1})$ is calculated as in (A.18) but where we replace $P_{h_{t-2} h_{t-1}}(a_{t-2})$ in PB_{t-2} with the initial distribution $P(h_0)$. Similarly, at $t = 0$ we have

$$\begin{aligned} V_0^{\mathrm{L}2}(o_0) &:= \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + V_1^{\mathrm{L}2}(o_{0:1}, a_0) \mid o_0] \\ &= \max_{a_0} \sum_{o_1} \mathbb{P}_{a_0}(h_0, o_1 \mid o_0)[r_0(h_0, a_0) + V_1^{\mathrm{L}2}(o_{0:1}, a_0)] \end{aligned}$$

and where

$$\mathbb{P}_{a_0}(h_0, o_1 | o_0) = \frac{P(h_0) B_{h_0 o_0} \sum_{h_1} PB_0}{\sum_{h_0} P(h_0) B_{h_0 o_0}}.$$

Computing the Lag-2 Approximate Value Function for the POMDP

Following (2.44) we can write the Lag-2 AVF as

$$\begin{aligned} &\widetilde{V}_t^{\mathrm{L}2}(\pi_t) = \\ &\max_{a_t} \mathbb{E}[r_t(h_t, a_t) + \max_{a_{t+1}} \mathbb{E}[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}2}(h_t, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}] \mid \mathcal{F}_t^{\pi}]. \end{aligned} \tag{A.19}$$

The inner expectation in (A.19) can be calculated according to

$$\sum_{h_{t:t+1}, o_{t+2}} \mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+2} \mid \pi_t, o_{t+1})[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}2}(h_t, a_{t:t+1}, o_{t+1:t+2})]. \tag{A.20}$$

The probability in (A.20) can then be computed using standard arguments. In particular, we have

$$\begin{aligned}\mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+2} \mid \pi_t, o_{t+1}) &= \frac{\mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+1:t+2} \mid \pi_t)}{\mathbb{P}_{a_t}(o_{t+1} \mid \pi_t)} \\ &= \frac{\sum_{h_{t+2}} \mathbb{P}_{a_{t:t+1}}(h_{t:t+2}, o_{t+1:t+2} \mid \pi_t)}{\sum_{h_{t:t+1}} \mathbb{P}_{a_t}(h_{t:t+1}, o_{t+1} \mid \pi_t)} \\ &= \frac{\pi_t(h_t) PB_t \sum_{h_{t+2}} PB_{t+1}}{\sum_{h_t, h_{t+1}} \pi_t(h_t) PB_t} \end{aligned} \tag{A.21}$$

where we once again denote by $PB_t \equiv P_{h_t h_{t+1}}(a_t) B_{h_{t+1} o_{t+1}}(a_t)$.

Remark A.2.1. *We note that if $T = 2$, then we recover the optimal value $V_0^*(\pi_0)$ of the POMDP. In particular,*

$$\begin{aligned}\widetilde{V}_0^{L_2}(\pi_0) &= \max_{a_0} \mathbb{E}[\max_{a_1} \mathbb{E}[r_0(h_0, a_0) + r_1(h_1, a_1) + V_2^{L_2}(h_0, a_{0:1}, o_{1:2}) \mid \mathcal{F}_0^\pi, o_1] \mid \mathcal{F}_0^\pi] \\ &= \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + \max_{a_1} \mathbb{E}[r_1(h_1, a_1) + r_2(h_2) \mid \mathcal{F}_0^\pi, o_1] \mid \mathcal{F}_0^\pi] \;=\; V_0^*(\pi_0)\end{aligned}$$

where the second equality follows from the tower property of conditional expectations.

A.2.3 Comparing the Lag-1 and Lag-2 Approximate Value Functions

We begin by proving the unsurprising result that the Lag-2 AVF is tighter than the Lag-1 AVF.

Proposition A.2.1. *For all t we have $V_t^*(\pi_t) \leq \widetilde{V}_t^{L_2}(\pi_t) \leq \widetilde{V}_t^{L_1}(\pi_t)$.*

Proof. We show in Appendix A.3 that $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$ is a supersolution and so it follows that $V_t^*(\pi_t) \leq \widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$. To prove the second inequality we begin with the definition of $\widetilde{V}_t^{\mathrm{L}_1}(\pi_t)$ in (C.7) for $t = 0, \ldots, T-2$. (We recall that at $t = T-1$ and $t = T$ we have that $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t) = \widetilde{V}_t^{\mathrm{L}_1}(\pi_t)$ for all

π_t.) We obtain

$$
\begin{aligned}
\widetilde{V}_t^{\mathrm{L}_1}(\pi_t) &:= \max_{a_t} \mathbb{E}\left[r_t(h_t,a_t) + V_{t+1}^{\mathrm{L}_1}(h_t,a_t,o_{t+1}) \mid \mathcal{F}_t^{\pi}\right] \\
&\overset{(a)}{=} \max_{a_t} \mathbb{E}_{h_t,o_{t+1}}\left[r_t(h_t,a_t) + \max_{a_{t+1}} \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}\right] \\
&\overset{(b)}{=} \max_{a_t} \mathbb{E}_{h_t,o_{t+1}}\left[\max_{a_{t+1}} r_t(h_t,a_t) + \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}\right] \\
&\overset{(c)}{\geq} \max_{a_t} \\
&\mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_t}\left[r_t(h_t,a_t) + \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}\right] \\
&\overset{(d)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\left[r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1},a_{t+1},o_{t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}\right] \\
&\overset{(e)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\left[r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_2}(h_{t+1},a_{t:t+1},o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}\right] \mid \mathcal{F}_t^{\pi}\right] \\
&= \widetilde{V}_t^{\mathrm{L}_2}(\pi_t)
\end{aligned}
$$

where (a) results from using the definition of $V_{t+1}^{\mathrm{L}_1}$ and (b) follows by simply moving $r_t(h_t,a_t)$ inside the maximization of a_{t+1}. Inequality (c) results from applying Jensen's inequality when exchanging the order of the expectation w.r.t. h_t and the maximization of a_{t+1}. Equality (d) results from the tower property and noting that the argument inside the inner expectation, conditional on h_t, is independent of $\mathcal{F}_t^{\pi}$. Inequality (e) follows by replacing $V_{t+2}^{\mathrm{L}_1}$ with $V_{t+2}^{\mathrm{L}_2}$ and then using Lemma A.2.1 below. Finally, the last equality follows from the definition of $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$ in (2.44). □

Lemma A.2.1. $\mathbb{E}[V_{t+2}^{L_1}(h_{t+1},a_{t+1},o_{t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}] \geq \mathbb{E}[V_{t+2}^{L_2}(h_t,a_{t:t+1},o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}]$ *for all* $t = 0,\ldots,T-2$.

Proof. To begin we note that it suffices to prove that

$$
\mathbb{E}[V_{t+2}^{\mathrm{L}_1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1:t+2}] \geq V_{t+2}^{\mathrm{L}_2}(h_t,a_{t:t+1},o_{t+1:t+2}) \tag{A.22}
$$

since taking expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t^{\pi}, o_{t+1}]$ on both sides of (A.22) and applying the tower property yields[3] the desired result. We now prove (A.22) by induction for $t = 0,\ldots,T-2$. The base case follows immediately since $V_T^{\mathrm{L}_1} = V_T^{\mathrm{L}_2} = r_T(h_T)$ and so $\mathbb{E}[V_T^{\mathrm{L}_1} \mid h_{T-2}, o_{T-1:T}] = r_T(h_T) = V_T^{\mathrm{L}_2}$ where we recall that $o_T \equiv h_T$. We now assume the result is true for time $t+3$ so that $\mathbb{E}[V_{t+3}^{\mathrm{L}_1} \mid$

[3] Note that the term inside the expectation on the left-hand-side of (A.22), conditional on h_t, is independent of $\mathcal{F}_t^{\pi}$, and so the tower property indeed yields the result.

$h_{t+1}, o_{t+2:t+3}] \geq V_{t+3}^{\mathrm{L}_2}$. An application of the tower property then implies

$$\mathbb{E}[V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}] \geq \mathbb{E}[V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t+1}, o_{t+2}]. \tag{A.23}$$

It then follows that

$$\begin{aligned}
&\mathbb{E}\left[V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\max_{a_{t+2}} \mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(b)}{\geq} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(c)}{\geq} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(d)}{=} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t:t+1}, o_{t+1:t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(e)}{=} \max_{a_{t+2}} \mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\right] \\
&= V_{t+2}^{\mathrm{L}_2}(h_t, a_{t:t+1}, o_{t+1:t+2})
\end{aligned}$$

where we use the definition of $V_{t+2}^{\mathrm{L}_1}$ in (a). Inequality (b) follows from Jensen's inequality after exchanging the outer expectation with $\max_{a_{t+2}}$. We obtain (c) from the induction hypothesis and inequality (A.23). Equality (d) follows by noting that the argument inside the inner expectation, conditional on h_{t+1}, is independent of h_t and o_{t+1}. Equality (e) then follows from the tower property and the final equality results from the definition of $V_{t+2}^{\mathrm{L}_2}$. We have therefore shown the desired result for time $t+2$ and so the proof is complete. □

A.3 Proving that the Approximate Value Functions Are Supersolutions

We now prove Proposition 2.6.2 which states that all of our AVFs are supersolutions. Recall that a supersolution is an AVF ϑ that for all possible time t belief states π_t satisfies

$$\vartheta_t(\pi_t) \geq \max_{a_t \in \mathcal{A}} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\vartheta_{t+1}(\pi_t) \mid \mathcal{F}_t^{\pi}\right] \right\}. \tag{A.24}$$

Before proceeding we note that given the current belief state π_t and the next observation o_{t+1}, the belief state π_{t+1} can be computed according to

$$\pi_{t+1}(h') = \frac{\sum_h \pi_t(h) P_{hh'} B_{h'o}}{\sigma(o, \pi_t)} \tag{A.25}$$

where[4] $\sigma(o, \pi_t) := P(o_{t+1} \mid \pi_t) = \sum_{h,h'} \pi_t(h) P_{hh'} B_{h'o}$ for $t \in \{0, \ldots, T-1\}$.

Proof that the MDP Approximation is a Supersolution

Following (2.37) and (2.38) we have

$$\begin{aligned}
\widetilde{V}_t^{\mathrm{MDP}}(\pi_t) &= \sum_h \pi_t(h) \max_{a_t} \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(a)}{\geq} \max_{a_t} \sum_h \pi_t(h) \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(b)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} \left[\sum_h \pi_t(h) P_{hh'}(a_t) B_{h' o_{t+1}}(a_t) \right] V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(c)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(d)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1}) \right\} \\
&\equiv \max_{a_t} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi} \right] \right\}
\end{aligned}$$

where (a) results from Jensen's inequality and (b) follows from including the factor

$$\sum_{o_{t+1}} B_{h' o_{t+1}} = 1$$

and then a simple re-ordering of the terms. Equality (c) follows from (A.25) while we have used the definition of $\widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1})$ to obtain (d).

[4] As before, we will often suppress the dependence of the various transmission and emission probabilities on the actions.

Proof that the QMDP Approximation is a Supersolution

The proof for the QMDP approximation follows a similar argument. From (C.4) and (C.5) we have

$$\begin{aligned}
\widetilde{V}_t^{\mathrm{Q}}(\pi_t) &= \max_{a_t} \sum_h \pi_t(h) \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(a)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(b)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') \max_{a'} V_{t+1}^{Q}(h', a') \right\} \\
&\overset{(c)}{\geq} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \max_{a'} \sum_{h'} \pi_{t+1}(h') V_{t+1}^{\mathrm{Q}}(h', a') \right\} \\
&\overset{(d)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1}) \right\} \\
&\equiv \max_{a_t} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi} \right] \right\}
\end{aligned}$$

where (a) follows from following steps (b) to (d) of the MDP proof above and (b) then follows from the definition of both V_{t+1}^{MDP} and V_{t+1}^{Q}. Inequality (c) follows from Jensen's inequality after changing the order of $\max_{a'}$ and the marginalization of h'. Finally (d) follows from the definition of $\widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1})$.

Proof that the Lag-1 Approximation is a Supersolution

Because of the many terms involved, throughout the proof we will write the relevant quantities as expectations and we will use $\mathbb{E}_X$ to denote an expectation taken over the random variable X.

Following its definition in (C.7), the Lag-1 AVF satisfies

$$
\begin{aligned}
\widetilde{V}_t^{\mathrm{L}_1}(\pi_t) &\overset{(a)}{=} \max_{a_t} \mathbb{E}_{h_t, o_{t+1}} \Big[r_t(h_t, a_t) \\
&\qquad + \max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1} \right] \mid \mathcal{F}_t^{\pi} \Big] \\
&\overset{(b)}{=} \max_{a_t} \Big\{ r_t(\pi_t, a_t) + \\
&\qquad \mathbb{E}_{h_t, o_{t+1}} \left[\max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1} \right] \mid \mathcal{F}_t^{\pi} \right] \Big\} \\
&\overset{(c)}{=} \max_{a_t} \Big\{ r_t(\pi_t, a_t) + \mathbb{E}_{o_{t+1}} \Big[\mathbb{E}_{h_t} \Big[\\
&\qquad \max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1} \right] \mid o_{t+1}, \mathcal{F}_t^{\pi} \Big] \mid \mathcal{F}_t^{\pi} \Big] \Big\} \\
&\overset{(d)}{\geq} \max_{a_t} \Big\{ r_t(\pi_t, a_t) + \mathbb{E}_{o_{t+1}} \Big[\max_{a_{t+1}} \mathbb{E}_{h_t} \Big[\\
&\qquad \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1} \right] \mid o_{t+1}, \mathcal{F}_t^{\pi} \Big] \mid \mathcal{F}_t^{\pi} \Big] \Big\} \\
&\overset{(e)}{=} \max_{a_t} \Big\{ r_t(\pi_t, a_t) + \mathbb{E}_{o_{t+1}} \Big[\max_{a_{t+1}} \mathbb{E}_{h_t} \Big[\\
&\qquad \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1}, \mathcal{F}_t^{\pi} \right] \mid o_{t+1}, \mathcal{F}_t^{\pi} \Big] \mid \mathcal{F}_t^{\pi} \Big\} \\
&\overset{(f)}{=} \max_{a_t} \Big\{ r_t(\pi_t, a_t) + \\
&\qquad \mathbb{E}_{o_{t+1}} \Big[\max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} \left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid o_{t+1}, \mathcal{F}_t^{\pi} \right] \mid \mathcal{F}_t^{\pi} \Big] \Big\} \\
&\overset{(g)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \mathbb{E} \left[\widetilde{V}_{t+1}^{\mathrm{L}_1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi} \right] \right\}
\end{aligned}
$$

where (a) follows from the definition of $V_{t+1}^{\mathrm{L}_1}$ in (C.6) and (b) follows from noting that the expectation of $r_t(h_t, a_t)$ conditional on $\mathcal{F}_t^{\pi}$ is $r_t(\pi_t, a_t)$. Equality (c) follows from the tower property while (d) follows from Jensen's inequality after changing the order of $\max_{a_{t+1}}$ and the expectation over h_t. Equality (e) follows since the function inside the expectation $\mathbb{E}[\cdot \mid h_t, o_{t+1}]$ is independent of $\mathcal{F}_t^{\pi}$ after conditioning on h_t. Equality (f) follows from applying the tower property to the nested expectations. Finally (g) follows from the definition of $\widetilde{V}_{t+1}^{\mathrm{L}_1}(\pi_{t+1})$ and where we note that π_{t+1} is completely determined given π_t, o_{t+1} and a_t.

Proof that the Lag-2 Approximation is a Supersolution

Proving that the Lag-2 AVF is a supersolution is similar to proving that the Lag-1 AVF is a supersolution but the details are a little more involved. From (2.44) we have

$$
\begin{aligned}
\widetilde{V}_t^{\mathrm{L}_2}(\pi_t) &:= \max_{a_t} \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[\\
&\qquad r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}_2}(h_t,a_{t:t+1},o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}\Big] \mid \mathcal{F}_t^{\pi}\Bigg] \\
&\overset{(a)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + \max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^{\pi}, o_{t+1}\Big] \mid \mathcal{F}_t^{\pi}\Bigg] \\
&\overset{(b)}{=} \max_{a_t} \Bigg\{ r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[r_{t+1}(h_{t+1},a_{t+1}) + \max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^{\pi}, o_{t+1}\Big] \mid \mathcal{F}_t^{\pi}\Bigg]\Bigg\} \\
&\overset{(c)}{=} \max_{a_t} \Bigg\{ r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[\max_{a_{t+2}} \Big\{ r_{t+1}(h_{t+1},a_{t+1}) + \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^{\pi}, o_{t+1}\Big] \mid \mathcal{F}_t^{\pi}\Bigg]\Bigg\}
\end{aligned}
\tag{A.26}
$$

where (a) follows from the definition of $V_{t+2}^{\mathrm{L}_2}$ in (A.17). We obtain (b) by taking the expectation of $r_t(h_t,a_t)$ outside the maximization of a_{t+1} (which is fine since a_{t+1} has no bearing on $r_t(h_t,a_t)$) and then using the tower property with the outer expectation to obtain $r_t(\pi_t,a_t)$. Equality (c) follows from taking $r_{t+1}(h_{t+1},a_{t+1})$ inside the maximization of a_{t+2} which is again fine since a_{t+2} has no bearing on $r_{t+1}(h_{t+1},a_{t+1})$. We focus now on the term inside the outermost expectation

$\mathbb{E}_{o_{t+1}}[\cdot \mid \mathcal{F}_t^\pi]$ of (A.26). It satisfies

$$
\begin{aligned}
&\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Bigg[\max_{a_{t+2}}\Big\{r_{t+1}(h_{t+1},a_{t+1})+ \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(d)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Bigg[\mathbb{E}_{h_{t:t+1}}\Big[\max_{a_{t+2}}\Big\{r_{t+1}(h_{t+1},a_{t+1})+\mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big] \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(e)}{\geq} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Bigg[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t:t+1}}\big[r_{t+1}(h_{t+1},a_{t+1})+\mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(f)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Bigg[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t+1}}\big[r_{t+1}(h_{t+1},a_{t+1}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big]+\mathbb{E}_{h_t}\Big[\mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[\\
&\qquad r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(g)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Bigg[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t+1}}\Big[r_{t+1}(h_{t+1},a_{t+1}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big]+ \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\Big[r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(h)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Bigg[\max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\Big[r_{t+1}(h_{t+1},a_{t+1})+ \\
&\qquad r_{t+2}(h_{t+2},a_{t+2})+V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big] \mid \mathcal{F}_t^\pi, o_{t+1}\Bigg] \\
&\overset{(i)}{=} \widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})
\end{aligned}
$$

where (d) follows from the tower property so that

$$
\mathbb{E}_{h_{t:t+1},o_{t+2}}\left[\cdot \mid \mathcal{F}_t^\pi, o_{t+1}\right] = \mathbb{E}_{o_{t+2}}\Big[\mathbb{E}_{h_{t:t+1}}\left[\cdot \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\right] \mid \mathcal{F}_t^\pi, o_{t+1}\Big]
$$

and (e) follows from Jensen's inequality after changing the order of the $\max_{a_{t+2}}$ operator and the marginalization of h_t and h_{t+1}. We obtain (f) by simply writing the conditional expectation of a sum as the sum of conditional expectations. Equality (g) follows from applying the tower property to the nested expectations while (h) follows from grouping together the two conditional expectations $\mathbb{E}[\cdot \mid \mathcal{F}_t^\pi, o_{t+1:t+2}]$. Finally, (i) follows from the definition of the $\widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})$ and where we note again that π_{t+1} is completely determined given π_t, o_{t+1} and a_t.

The overall result now follows by substituting $\widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})$ in for the conditional expectation $\mathbb{E}_{o_{t+1}}[\cdot \mid \mathcal{F}_t^\pi]$ in (A.26) with the equality there replaced by a greater-than-or-equal to inequality.

A.4 Dropping the Requirement that $\mathbb{P} \ll \tilde{\mathbb{P}}$

We explain here why we do not require $\mathbb{P}$, the probability measure for the controlled formulation, to be absolutely continuous w.r.t $\tilde{\mathbb{P}}$ (the probability measure for the original uncontrolled formulation), when the penalties in (2.16) are constructed from supersolutions. This result was originally shown by BH in [15] but we outline the details here in the finite horizon case for the sake of completeness. We will work with the PI relaxation of belief-state POMDP formulation, i.e. the BSPI relaxation, but it should be clear that the result is general and holds for general information relaxations.

We therefore assume the penalty function, $c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1})$, is such that ϑ_t is a supersolution satisfying[5] $\vartheta_{T+1} \equiv 0$. From Definition 2.6.1, it follows that for each $t \in \{0, \ldots, T\}$ and π_t we have

$$\vartheta_t(\pi_t) \geq r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] \quad \forall a_t \in \mathcal{A}. \tag{A.27}$$

Subtracting $\vartheta_t(\pi_t)$ from both sides of (A.27), summing over t and recalling that $\vartheta_{T+1} \equiv 0$, we obtain

$$\begin{aligned} 0 &\geq \sum_{t=0}^{T} \Big\{ r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{0:t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_t(\pi_t) \Big\} \\ &= \sum_{t=0}^{T} \Big\{ r_t(\pi_t, a_t) + c_t \Big\} - \vartheta_0(\pi_0). \end{aligned} \tag{A.28}$$

We now obtain

$$\begin{aligned} V_0^* - \vartheta_0 &= \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} V_0^{\mu} - \vartheta_0 \\ &\overset{(a)}{=} \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\Big[\sum_{t=0}^{T} (r_t + c_t) - \vartheta_0 \mid \mathcal{F}_0^{\pi} \Big] \qquad (A.29) \\ &\overset{(b)}{\leq} \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \widetilde{\mathbb{E}}\Big[\sum_{t=0}^{T} \Phi_t\big(r_t + c_t\big) - \vartheta_0 \mid \mathcal{F}_0^{\pi} \Big] \\ &\overset{(c)}{\leq} \widetilde{\mathbb{E}}\Big[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t\big(r_t + c_t\big) \mid \mathcal{F}_0^{\pi} \Big] - \vartheta_0. \qquad (A.30) \end{aligned}$$

[5] There is no difficulty in assuming $\vartheta_{T+1} \equiv 0$ since $\vartheta_t(\pi_t)$ represents an AVF and all of our AVFs naturally satisfy this assumption.

where we have omitted the arguments of r_t and ϑ_0 for the sake of clarity. Equality (a) follows since $\mathbb{E}[\sum_{t=0}^{T} c_t \mid \mathcal{F}_0^\pi] = 0$ for any $\mathbb{F}^\pi$-adapted policy and since $\vartheta_0(\pi_0)$ is $\mathcal{F}_0^\pi$-adapted. In order to establish inequality (b), we first note that (A.28) implies the random quantity inside the expectation in (A.29) is non-positive w.p. 1. The inequality then follows[6] for any probability measure, $\tilde{\mathbb{P}}$, regardless of whether or not $\mathbb{P}$ is absolutely continuous w.r.t $\tilde{\mathbb{P}}$. Inequality (c) follows from the usual weak duality argument. We also note that $\Phi_0 \equiv 1$ which explains why there is no RN term multiplying $\vartheta_0(\pi_0)$.

We can now add $\vartheta_0(\pi_0)$ across both sides of (A.30) to establish the result, i.e. weak duality continues to hold even if the probability measure, $\mathbb{P}$, is not absolutely continuous w.r.t $\tilde{\mathbb{P}}$ as long as the penalty is constructed from a supersolution. It is also interesting to note that inequality (b) will in fact be an equality if $\tilde{\mathbb{P}}$ is the measure induced by following an optimal policy for the primal problem since in that case $\mathbb{P}$ and $\tilde{\mathbb{P}}$ will coincide. Strong duality will then also continue to hold. In particular, (c) will then also be an equality if ϑ_t coincides with the optimal value function, V_t^*, which is itself a supersolution.

A.5 Further Details for the Multiaccess Communication Application

The main difference between the multiaccess communication application and the POMDP framework as defined in Section 2.2 is the timing of observations. Specifically, in the multiaccess communication application an observation occurs immediately after an action is taken and is therefore a function of the current hidden state and the *current* action. In contrast, in the usual POMDP setting, an observation is a function of the current hidden state and the action from the previous period. Therefore the filtering algorithm for the belief-state update is different than the standard update as given in (A.25) (where the action dependence was suppressed). The belief update for

[6] This result was stated as Lemma A.1 in [15] and we state it here for the sake of completeness. Consider a measurable space (Ω, Σ) and two probability measures P and Q. Let ϕ represent the Radon-Nikodym derivative of the absolutely continuous component of P with respect to Q. If $Y = Y(\omega)$ is a bounded random variable such that $Y(\omega) \leq 0$ for all $\omega \notin \Omega_Q := \{\omega \in \Omega : Q(\omega) > 0\}$, then $\mathbb{E}^P[Y] \leq \mathbb{E}^Q[\phi Y]$.

the slotted Aloha dynamics satisfies

$$\pi_{t+1}(h') = \frac{\sum_h \pi_t(h) B_{ho_t}(a_t) P_{hh'}(o_t)}{\sum_h \pi_t(h) B_{ho_t}(a_t)} \tag{A.31}$$

for $t \in \{0, \ldots, T-1\}$ and where we recognize the denominator in (A.31) as $\mathbb{P}_{a_t}(o_t \mid \pi_t)$. It is worth emphasizing that the belief state for time $t+1$ is a function of the time t action and observation. Moreover, the hidden-state transition probabilities under $\mathbb{P}$ are action-independent given the current observation. As a result we assume the hidden state transitions probabilities are unchanged when we go from $\mathbb{P}$ to $\tilde{\mathbb{P}}$.

These alternative dynamics also impact the calculations of the RN derivatives. In the case of the belief-state formulation, the arguments in Appendix A.1.1 that led to (A.5) still apply. However, in the multiaccess communication application the numerator of (A.5) now satisfies

$$\begin{aligned}
\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) &= \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(\pi_{s+1} \mid \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{o_s} \mathbb{P}_{a_s}(o_s \mid \pi_s) \mathbb{P}_{a_s}(\pi_{s+1} \mid o_s, \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o_s} \pi_s(h) \mathbb{P}_{a_s}(o_s \mid h) \mathbb{P}(h' \mid h, o_s) \mathbf{1}_{\{\pi_{s+1} = f(\pi_s, a_s, o_s)\}} \\
&= \prod_{s=0}^{t-1} \sum_{h,o} \pi_s(h) B_{ho}(a_s) \mathbf{1}_{\{\pi_{s+1} = f(\pi_s, a_s, o)\}}
\end{aligned} \tag{A.32}$$

where $f(\pi_s, a_s, o_s)$ lies in the $|\mathcal{H}|$-dimensional simplex with each of its components defined according to

$$f(\pi_s, a_s, o_s)(h') := \frac{\sum_h \pi_s(h) B_{ho}(a_s) P_{hh'}(o)}{\sum_h \pi_s(h) B_{ho}(a_s)}.$$

Using similar arguments, we see that the denominator of (A.5) satisfies

$$\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \sum_{h,o} \pi_s(h) E^s_{ho} \mathbf{1}_{\{\pi_{s+1} = \tilde{f}_s(\pi_s; o)\}}$$

where E^s_{ho} is the uncontrolled emission matrix defined in (5.7) and where $\tilde{f}_s(\pi_s; o)$ lies in the $|\mathcal{H}|$-dimensional simplex with each of its components defined according to

$$\tilde{f}_s(\pi_s; o)(h') := \frac{\sum_h \pi_s(h) E^s_{ho} P_{hh'}(o)}{\sum_h \pi_s(h) E^s_{ho}}.$$

In the case of the non-belief-state formulation of the problem, the RN derivatives satisfy

$$\Phi_t := \frac{\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t})}{\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t})} \tag{A.33}$$

where

$$\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(o_s \mid h_s) \mathbb{P}(h_{s+1} \mid h_s, o_s) \tag{A.34}$$

$$\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(o_s \mid h_s) \mathbb{P}(h_{s+1} \mid h_s, o_s). \tag{A.35}$$

It immediately follows from (A.34) and (A.35) that the RN derivatives for the uncontrolled non-belief-state formulation satisfy

$$\phi_t(i, k, a) := \frac{B_{ik}(a)}{E^t_{ik}}$$

$$\Phi_t(h_{0:t}, o_{0:t-1}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi_s(h_s, o_s, a_s).$$

A.6 Extension to Infinite Horizon Problems

We can extend these techniques to the infinite horizon class of POMDPs with discounted rewards following the approach of BH and [87]. Let the discount factor be denoted by $\delta \in [0, 1)$, indicating that rewards received at a later time contribute less than rewards received earlier. The corresponding infinite-horizon POMDP can be stated as solving the following optimization problem

$$V_0^* := \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t r(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right] \tag{A.36}$$

In order to solve the dual problem using a BSPI relaxation, we would have to simulate an infinite sequence of random variables $\{u_t\}_{t \geq 0}$, which is not possible in practice. An equivalent formulation, however, is to replace the discounting by a costless, absorbing state π^a which can be reached from every belief-state and feasible action with probability $1 - \delta$, at each t. The state transition distribution remains as in (2.6), conditional on not reaching the absorbing state. The equivalent absorbing state formulation is then given by

$$V_0^* := \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^{\tau} r(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right] \tag{A.37}$$

where $\tau = \inf\{t : \pi_t = \pi^a\}$ is the absorption time, distributed as a geometric random variable with parameter $1 - \delta$. In (A.37) the expected value is calculated over the modified state transition function that accounts for the presence of the absorbing state. In the dual problem formulation, knowledge of the absorption time should be included in the relevant information relaxation. For example, under the BSPI relaxation, the dual upper bound can be expressed as

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:\tau-1}} \sum_{t=0}^{\tau} \Phi_t[r_t(\pi_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0^{\pi}\right]. \tag{A.38}$$

An inner problem inside the expectation on the r.h.s of (A.38) can be generated by first simulating the absorption time $\tau \sim \text{Geom}(1-\delta)$, and then generating the belief states π_t using some action-independent change of measure. A lower bound can be obtained of course by simply simulating many paths of some feasible policy.

One concern with the bound of (A.38) is that the optimal objective of the inner problem in (A.38) might have an infinite variance. This was not a concern in the finite horizon setting with finite state and action spaces. It is a concern, however, in the infinite horizon setting where τ is now random and the presence of the RN derivative terms Φ_t might now cause the variance to explode. BH resolved this issue through the use of supersolutions to construct dual penalties. In that case their bound improvement result[7] and other considerations allowed them to conclude that the variance of the upper bound estimator in (A.38) would remain bounded.

Of course an alternative approach to guarantee finite variance estimators is to truncate the infinite horizon to some large fixed value, T, and then add $\delta^T \bar{r}/(1-\delta)$ as a terminal reward where $\bar{r} := \max_{\pi,a} r(\pi, a)$. Because the terminal reward is an upper bound on the total discounted remaining reward after time T in the infinite horizon problem, we are guaranteed that a dual upper bound for the truncated problem will also be a valid upper bound on the infinite horizon problem. By choosing T suitably large we can minimize the effect of truncation on the quality of the dual bound for the infinite horizon problem.

[7] See also the discussion immediately following our Proposition 2.6.2.

Appendix B

Chapter 3 - Supplemental Content

B.1 Portfolio Construction Via Linear Programming

We develop a simple linear programming (LP) approach to construct a portfolio according to the setting and notation introduced in Section 3.4.2. We assume the p.m can trade in N securities and that their daily P&L, $\Delta v_i(\mathbf{\Delta x})$, for $i = 1, \ldots, N$, depends on the vector of risk factor changes $\mathbf{\Delta x} \in \mathbb{R}^n$. The p.m. wishes to determine the portfolio weights $w_1, \ldots, w_N$ where $\sum_{i=1}^N w_i = 1$ and where w_i is the percentage of the portfolio value allocated to security i. The p.m. believes $\mathbf{f_e} = \mathbf{c}$ at the end of the next period and wishes to construct per portfolio to take advantage of this belief. The p.m. also believes and uses the DFM approach and has therefore estimated π_{t+1} as well as the parameters of the model (3.2) and the corresponding dynamic factor model for $\mathbf{f}_t$. She can therefore easily simulate K samples of the risk factor changes, $\mathbf{\Delta x}^{(1)}, \ldots, \mathbf{\Delta x}^{(K)}$ and the use these samples to estimate the expected P&L for each of the N securities conditional on the view, i.e. scenario, that $\mathbf{f_e} = \mathbf{c}$. We let $\Delta v_i^{\text{fm}} := \frac{1}{K} \sum_{k=1}^K \Delta v_i(\mathbf{\Delta x}^{(k)})$ denote these expected conditional P&Ls. Letting $\mathbf{w} := (w_1, \ldots, w_N)$, the p.m.'s objective function will therefore be given by

$$F(\mathbf{w}) := \sum_{i=1}^N w_i \Delta v_i^{\text{fm}} \tag{B.1}$$

which is her expected portfolio P&L conditional on the view $\mathbf{f_e} = \mathbf{c}$.

The p.m. must also satisfy certain constraints imposed by the risk management team. In particular, the risk management team require that the estimated scenario P&L's for L different

scenarios must lie between $-\alpha\%$ and $\alpha\%$. These estimated scenario P&L's are computed using the SSA approach and involve stresses to combinations of the c.r.f.'s in $\mathbf{f_n}$. For each security $i = 1, \ldots, N$ and each scenario $l = 1, \ldots, L$, we can use the SSA approach to estimate the P&L of the i^{th} security in that scenario. If we denote this estimated P&L by $\Delta v_i^{(l)}$ then these risk constraints will result in the following linear constraints for the LP:

$$\begin{aligned} A_{l+}(\mathbf{w}) &:= \sum_{i=1}^{N} w_i \Delta v_i^{(l)} \leq \alpha \quad \text{for } l = 1, \ldots, L \\ A_{l-}(\mathbf{w}) &:= \sum_{i=1}^{N} w_i \Delta v_i^{(l)} \geq -\alpha \quad \text{for } l = 1, \ldots, L \end{aligned} \tag{B.2}$$

We can then combine (B.1) and (B.2) together with the constraint $\mathbf{1}^\top \mathbf{w} = 1$ to obtain the full LP that the p.m. must solve to obtain her optimal portfolio.

We note that it's easy to formulate more realistic LPs. For example, it would make sense to allow α to be scenario dependent and only limit the downside risk in the L scenarios. Similarly, we could assume the risk-management team is more sophisticated and therefore use DFMSA when estimating the scenario P&Ls. Likewise, it is easy to include constraints imposed by the p.m. rather than the risk-management team. Additional constraints on the so-called Greeks, e.g. delta, gamma, vega etc, of the overall portfolio as well as position constraints could also be imposed in the LP. Nonetheless the LP formulated above seems like a very straightforward way to highlight the problems that can arise when using SSA rather than DFMSA.

B.2 Ground Truth Parameters for the Yield Curve Model of Section 3.5

Our yield curve model from (3.17) in matrix form is $\mathbf{\Delta x}_t = \mathbf{B f}_{t+1} + \boldsymbol{\epsilon}_{t+1}$ where we recall $\mathbf{\Delta x}_t$ denotes the yield changes in b.p.s for the n maturities and

$$\mathbf{f}_{t+1} := \begin{bmatrix} \text{ParallelShift}_{t+1} & \text{Slope}_{t+1} & \text{Curvature}_{t+1} \end{bmatrix}^\top$$

denotes the 3×1 vector of c.r.f. returns between dates t and $t+1$. Following Diebold and Li [28] we take $\lambda = 0.7308$ which results in the loadings matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.97 & 0.91 & 0.84 & 0.71 & 0.53 & 0.41 & 0.27 & 0.19 & 0.14 & 0.07 & 0.05 \\ 0.03 & 0.08 & 0.14 & 0.23 & 0.29 & 0.29 & 0.24 & 0.19 & 0.14 & 0.07 & 0.05 \end{bmatrix}^\top.$$

The parameter estimates for $\Sigma_{\boldsymbol{\epsilon}}, \mathbf{G}$ and $\Sigma_{\boldsymbol{\eta}}$ in (3.7) were obtained from the EM algorithm and $\mathbf{G}$ was constrained to be diagonal so that each c.r.f. return follows a univariate AR(1) process with no exogenous covariates, i.e., $f_{i,t+1} = g_{i,i} f_{i,t} + \eta_{i,t+1}$ for $i = 1, 2, 3$ where $g_{i,i}$ denotes the i^{th} diagonal element of $\mathbf{G}$. As a result, the cross-sectional dependence between c.r.f's in $\mathbf{f}_{t+1}$ are induced exclusively via the covariance of the innovation process $\boldsymbol{\eta}_{t+1}$. The ground-truth model parameters were estimated to be

$$\mathbf{diag}(\Sigma_{\boldsymbol{\epsilon}}^{1/2}) = \begin{bmatrix} 0.0600 & 0.0312 & 0.0146 & 0.0165 & 0.0158 & 0.0109 & 0.0112 & 0.0135 & 0.0107 & 0.0056 & 0.0097 \end{bmatrix}^\top$$

$$\mathbf{G} = \begin{bmatrix} 0.0383 & 0.0000 & 0.0000 \\ 0.0000 & 0.0727 & 0.0000 \\ 0.0000 & 0.0000 & 0.0399 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \begin{bmatrix} 0.0036 & -0.0038 & -0.0002 \\ -0.0038 & 0.0066 & -0.0039 \\ -0.0002 & -0.0039 & 0.0266 \end{bmatrix}.$$

The initial distribution π_0 of the c.r.f. returns was assumed to be Gaussian with mean zero and diagonal covariance matrix with diagonal elements equal to 0.01.

B.3 Ground Truth Parameters for the Options Portfolio Model of Section 3.6

Our state-space model (3.21) from Section 3.6 assumes the observation model

$$\boldsymbol{\Delta}\mathbf{x}_t = \begin{bmatrix} 1 \\ \mathbf{b}^o \end{bmatrix} f^o_{t+1} + \begin{bmatrix} \mathbf{0}_3^\top \\ \mathbf{B}^u \end{bmatrix} \mathbf{f}^u_{t+1} + \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_{t+1} \end{bmatrix}$$

where the first component of $\boldsymbol{\Delta}\mathbf{x}_t$ represents the daily log-return (in percentage points) of the S&P 500 and the remaining components represent the daily changes (in volatility points) in implied

volatility for $n-1$ moneyness-maturity pairs. The factor loadings matrix $\mathbf{B}^u$ corresponding to the latent c.r.f. returns is given explicitly for each moneyness-maturity pair by. The parameter estimates for $\mathbf{b}^o, \mathbf{G}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ in (3.7) and (3.21) were obtained via the EM algorithm where we also imposed the constraint that $\mathbf{G}$ is diagonal. While not strictly necessary, this assumption was made to help the convergence of the EM algorithm and it implies that (i) each c.r.f. return follows a univariate AR(1) process with no exogenous covariates and (ii) the dependence in $\mathbf{f}_{t+1}$ conditional on $\mathcal{F}_t$ is induced via the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

The ground-truth model parameters for the observation model (3.21) were obtained as

$$\mathbf{b}^o = \begin{bmatrix} -1.22 \\ -1.19 \\ -1.12 \\ -1.04 \\ -0.94 \\ -0.83 \\ -0.60 \\ \vdots \end{bmatrix} \qquad \mathbf{B}^u = \begin{bmatrix} 3.46 & 0.20 & -1.79 \\ 3.46 & 0.10 & -1.79 \\ 3.46 & 0.05 & -1.79 \\ 3.46 & 0.00 & -1.79 \\ 3.46 & -0.05 & -1.79 \\ 3.46 & -0.10 & -1.79 \\ 3.46 & -0.20 & -1.79 \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad \mathbf{diag}(\Sigma_\epsilon^{1/2}) = \begin{bmatrix} 0.0196 \\ 0.0091 \\ 0.0058 \\ 0.0031 \\ 0.0002 \\ 0.0040 \\ 0.0142 \\ \vdots \end{bmatrix}$$

where we show only[1] the rows of $\mathbf{b}^o, \mathbf{B}^u$ and $\mathbf{diag}(\Sigma_\epsilon^{1/2})$ that correspond to the first seven moneyness-maturity pairs (ξ, τ): $(0.80, 30d)$, $(0.90, 30d)$, $(0.95, 30d)$, $(1.00, 30d)$, $(1.05, 30d)$, $(1.10, 30d)$ and $(1.20, 30d)$.

The estimated parameters of the c.r.f. returns model (3.7) are

$$\mathbf{G} = \begin{bmatrix} -0.1161 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.1176 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & -0.4127 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -0.0466 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \begin{bmatrix} 0.00018 & 0.00001 & 0.00002 & 0.00000 \\ 0.00001 & 0.00003 & 0.00013 & 0.00002 \\ 0.00002 & 0.00013 & 0.00523 & 0.00010 \\ 0.00000 & 0.00002 & 0.00010 & 0.00002 \end{bmatrix}$$

where the first, second, third and fourth rows (and columns) of $\mathbf{G}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ represent the S&P 500, parallel shift, skew and term structure c.r.f. returns, respectively. For reference, the standard

[1] The complete model parameters are available upon request.

deviations and correlation matrix of the innovations $\boldsymbol{\eta}_t$'s are given by

$$\mathbf{diag}(\Sigma_{\eta}^{1/2}) = \begin{bmatrix} 0.0138 \\ 0.0058 \\ 0.0723 \\ 0.0039 \end{bmatrix} \qquad \boldsymbol{\rho}_{\boldsymbol{\eta}} = \begin{bmatrix} 1.0000 & 0.1525 & 0.0236 & 0.0921 \\ 0.1525 & 1.0000 & 0.3002 & 0.9283 \\ 0.0236 & 0.3002 & 1.0000 & 0.3543 \\ 0.0921 & 0.9283 & 0.3543 & 1.0000 \end{bmatrix}$$

The initial distribution π_0 was assumed to be normal with a zero mean vector and a diagonal covariance matrix with all diagonal elements set to 0.005.

B.4 Obtaining MAP Estimates of the Latent C.R.F. Returns

Here we provide a brief outline of the optimization approach to obtaining smoothed MAP estimates of the latent state variables, developed by [4] and [3]. In the following discussion, we refer to $\mathbf{f}_{t+1}$ as the state-vector and $\boldsymbol{\Delta}\mathbf{x}_t$ as the observation vector, in the sense than $\boldsymbol{\Delta}\mathbf{x}_t$, through the factor model (3.2), provides a noisy observation of the underlying *latent* state variables $\mathbf{f}_{t+1}$, which follow the dynamics (3.7).

In this appendix, and following [4] and [3], we use a general state space model given by

$$\begin{aligned} \mathbf{f}_t &= g_t(\mathbf{f}_{t-1}) + \boldsymbol{\eta}_t, \qquad t = 1, \ldots, T \\ \mathbf{y}_t &= h_t(\mathbf{f}_t) + \boldsymbol{\epsilon}_t, \qquad t = 1, \ldots, T \end{aligned} \tag{B.3}$$

with initial condition $\mathbf{f}_0$ is a known constant vector. In (B.3), $g_t : \mathbb{R}^m \to \mathbb{R}^m$ and $h_t : \mathbb{R}^m \to \mathbb{R}^n$ are known smooth functions, $\mathbf{y}_t \equiv \boldsymbol{\Delta}\mathbf{x}_{t-1}$, and $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are mutually independent zero-mean random vectors with $\mathbb{P}_{\boldsymbol{\eta}_t}(\cdot)$ and $\mathbb{P}_{\boldsymbol{\epsilon}_t}(\cdot)$ probability density functions, respectively. Note that by setting $h_t(\mathbf{f}_t) = \mathbf{B}\mathbf{f}_t$ and $g_t(\mathbf{f}_{t-1}) = \mathbf{G}\mathbf{f}_{t-1}$ for each t, we obtain equations (3.2) and (3.7) as special cases. Using the notation

$$\mathbf{F} := vec(\{\mathbf{f}_t\}_{t=1}^T) := \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_T \end{bmatrix} \in \mathbb{R}^{mT}$$

and similarly, denoting $\mathbf{Y} := vec(\{\mathbf{y}_t\}_{t=1}^T) \in \mathbb{R}^{nT}$, we can write the likelihood of the latent common factor returns given the observations as

$$\mathbb{P}(\mathbf{F} \mid \mathbf{Y}) \propto \mathbb{P}(\mathbf{Y} \mid \mathbf{F})\mathbb{P}(\mathbf{F}) = \prod_{t=1}^{T} \mathbb{P}(\mathbf{y}_t \mid \mathbf{f}_t)\mathbb{P}(\mathbf{f}_t \mid \mathbf{f}_{t-1}) = \prod_{t=1}^{T} \mathbb{P}_{\boldsymbol{\epsilon}_t}\left(\mathbf{y}_t - h_t(\mathbf{f}_t)\right)\mathbb{P}_{\boldsymbol{\eta}_t}\left(\mathbf{f}_t - g_t(\mathbf{f}_{t-1})\right) \quad \text{(B.4)}$$

We obtain the MAP estimates of the common factors by solving the optimization problem

$$\max_{\mathbf{f}_{1:T}} \prod_{t=1}^{T} \mathbb{P}_{\boldsymbol{\epsilon}_t}\left(\mathbf{y}_t - h_t(\mathbf{f}_t)\right)\mathbb{P}_{\boldsymbol{\eta}_t}\left(\mathbf{f}_t - g_t(\mathbf{f}_{t-1})\right) \quad \text{(B.5)}$$

i.e., by maximizing the objective function (B.4) for a given set of observations $\mathbf{y}_{1:T}$ and initial condition $\mathbf{f}_0$. In the case where $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are normally distributed, and $g_t(\cdot)$ and $h_t(\cdot)$ are linear, the MAP estimates can be obtained explicitly via the Kalman Filter and Kalman Smoother algorithms [48]. If either one or both of $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are not normally distributed, solving the optimization problem (B.5) to obtain the MAP estimates of the common factors results in an intractable problem in general. The recent work of Aravkin [4] proposes an optimization technique to solve (B.5) for the case in which $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are mutually independent Student-t distributed random variables.

Maximizing the likelihood (B.4) is equivalent to minimizing the negative log-posterior. If we let $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ be mutually independent Student-t distributed random variables, with covariance matrices $\mathbf{S}_t$ and $\mathbf{R}_t$, respectively, and degrees of freedom parameters s and r, respectively. Then the negative log-posterior can be written as proportional to

$$L(\mathbf{F}) := \sum_{t=1}^{T} r \ln\left(1 + \frac{\|\mathbf{y}_t - h_t(\mathbf{f}_t)\|^2_{\mathbf{R}_t^{-1}}}{r}\right) + s \ln\left(1 + \frac{\|\mathbf{f}_t - g_t(\mathbf{f}_{t-1})\|^2_{\mathbf{S}_t^{-1}}}{s}\right) \quad \text{(B.6)}$$

where $\|\mathbf{u}\|^2_{\mathbf{A}} := \mathbf{u}^\top \mathbf{A}\mathbf{u}$, for any vector $\mathbf{u}$ and matrix $\mathbf{A}$ of suitable sizes.

Note that the objective function (B.6) is a non-convex function of the common factors. A solution method is proposed in [4], in which the objective function is iteratively approximated locally using a convex function. This method follows a modified Gauss-Newton procedure in which information about the curvature of the log-likelihood is included in the Hessian approximation. More specifically, the modified Gauss-Newton procedure is an iterative method of the form

$$\mathbf{F}^{k+1} = \mathbf{F}^k + \gamma^k \mathbf{d}^k \quad \text{(B.7)}$$

where $\mathbf{F}^k$ is the k-th iterate approximation to the optimal $\mathbf{F}^*$ in (B.6), starting from some approximation $\mathbf{F}^0$, and where γ^k is a scalar that guarantees that $L(\mathbf{F}^{k+1}) < L(\mathbf{F}^k)$ and is obtained by a standard backtracking line-search procedure.Finally, $\mathbf{d}^k$ is the *modified* Gauss-Newton search direction, obtained by solving the subproblem

$$\min_{\mathbf{d}\in\mathbb{R}^{mT}} L(\mathbf{F}^k) + L^{(1)}(\mathbf{F}^k)^\top \mathbf{d} + \tfrac{1}{2}\mathbf{d}^\top \mathbf{U}(\mathbf{F}^k)\mathbf{d} \tag{B.8}$$

where $L^{(1)}(\mathbf{F}^k)$ denotes the gradient of the objective function (B.6) at current estimate $\mathbf{F}^k$, and $\mathbf{U}(\mathbf{F}^k)$ is a matrix that approximates the curvature of the log-likelihood around $\mathbf{F}^k$. The form of the matrix $\mathbf{U}$ is given in [4] as a symmetric positive definite block tridiagonal matrix. The block tridiagonal and positive definite structure of $\mathbf{U}$ allows an efficient calculation of the optimal $\mathbf{d}$ in (B.8), with the solution given by $\mathbf{d}^* = -\mathbf{U}^{-1}L^{(1)}$.

Note that the solution method can also handle the inclusion of a regularization term ρ to the objective function (B.6), as long as $\rho(\cdot)$ is a smooth convex function.

Appendix C

Chapter 4 - Supplemental Content

C.1 Proof of Theorem 1

Given a solution to the coordinator problem, $\sigma^{*\mathbf{C}} = (g^{*\mathbf{M}}, g_\theta^{*\mathbf{H}})$, assume that the human's true type is an arbitrary value $\bar{\theta} \in \Theta$. Then, it is sufficient to confirm that the strategies $\sigma^{*\mathbf{H}} = g_{\bar{\theta}}^{*\mathbf{H}}$ and $\sigma^{*\mathbf{M}} = g^{*\mathbf{M}}$ satisfy the three properties for the risk-sensitive Bayesian equilibrium:

(I) Machine's incentive compatibility

$$\rho^{\mathbf{M}}\left(\rho_\theta^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1, h_1\right)|\pi_1\right) \leq \rho^{\mathbf{M}}\left(\rho_\theta^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \tilde{\sigma}^{\mathbf{M}}, \pi_1, h_1\right)|\pi_1\right).$$

(II) Human's incentive compatibility

$$\rho_{\bar{\theta}}^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1, h_1\right) \leq \rho_{\bar{\theta}}^{\mathbf{H}}\left(C_T|\tilde{\sigma}^{\mathbf{H}}, \sigma^{*\mathbf{M}}, \pi_1, h_1\right).$$

(III) The consistent belief profile

$$\pi_{t+1}^*(\theta) := \frac{\pi_t^*(\theta)\sigma^{*\mathbf{H}}(a_t^{\mathbf{H}}|s_t, \pi_t^*, \theta)}{\sum_{\tilde{\theta}} \pi_t^*(\tilde{\theta})\sigma^{*\mathbf{H}}(a_t^{\mathbf{H}}|s_t, \pi_t^*, \tilde{\theta})}. \tag{C.1}$$

The machine's incentive compatibility (I) is satisfied since the objective function of the coordinator is equal to the objective function of the machine. The consistent belief profile (III) follows directly from the formulation of the coordinator problem as a POMDP. The human's incentive

compatibility condition (II) follows from the monotonicity property of the risk measures $\rho_{\theta}^{\mathbf{H}}$ and $\rho^{\mathbf{M}}$ by the following logic. Assume that there exists a strategy $\tilde{\sigma}_{\bar{\theta}}^{\mathbf{H}}$ such that

$$\rho_{\bar{\theta}}^{\mathbf{H}}\left(C_T|\sigma^{*\mathbf{H}},\sigma^{*\mathbf{M}},\pi_1,h_1\right) > \rho_{\bar{\theta}}^{\mathbf{H}}\left(C_T|\tilde{\sigma}^{\mathbf{H}},\sigma^{*\mathbf{M}},\pi_1,h_1\right).$$

Then the monotonicity property implies that the coordinator's objective function can be decreased using the strategy $g_{\bar{\theta}}^{\mathbf{H}} = \tilde{\sigma}^{\mathbf{H}}$. Thus we arrive at the desired contradiction that the solution to the coordinator problem is optimal.

C.2 Approximate Solutions and Bounds for POMDPs

In this appendix, we discuss various approaches for finding approximate solutions to POMDPs. We also discuss approaches for obtaining *dual bounds*, that provide a lower (upper) bound for the optimal value of a minimization (maximization) problem, and therefore serve to evaluate how close the approximate solutions are to optimality. Before going further, it is important to highlight that POMDPs can be formulated as MDPs, if we work with the belief-state (instead of hidden states and observations).

If we have a general POMDP with hidden states $x_t \in \mathcal{X}$, observations $o_t \in \mathcal{O}$, actions $a_t \in \mathcal{A}$ and cost functions $c_t(x_t, a_t)$, so that the POMDP problem can be written as

$$V_1^* := \min_{g\in\mathcal{U}} \mathbb{E}\left\{\sum_{t=1}^{T} c_t(x_t, g_t) \mid \pi_1\right\} \tag{C.2}$$

where π_1 denotes the initial distribution over x_1, and where $g = (g_1, g_1, \ldots, g_T)$ is taken from the set of *non-anticipative* policies $\mathcal{U}$, i.e., g_t depends on π_1 and the history of observations $o_{1:t}$ so that the action at time t is given by $a_t = g_t(\pi_1, o_{1:t})$. Equivalently, we can instead define the POMDP in terms of the belief-state π_t[1]. In this case, we can write the time t cost as a function of the belief-state by setting $\tilde{c}_t(\pi_t, a_t) := \mathbb{E}^{\pi}[c_t(x_t, a_t) \mid \pi_t]$ so that the POMDP problem is reformulated

[1] The belief-state is defined on the $|\mathcal{X}|$-dimensional simplex and can be calculated by a standard filtering algorithm, which takes π_t, a_t and o_{t+1} as inputs, and outputs the belief-state π_{t+1}. In our specific robo-advising framework, the filtering updates are given by (4.3).

as

$$V_1^* := \min_{g \in \mathcal{U}^\pi} \mathbb{E}\left\{ \sum_{t=1}^{T} \tilde{c}_t(\pi_t, g_t) \mid \pi_1 \right\} \tag{C.3}$$

where in this case the set of non-anticipative policies $\mathcal{U}^\pi$ corresponds to those g_t that depend on the belief-state history up to time t, $\pi_{1:t}$.

In our robo-advising framework, the hidden states x_t correspond to the pair (θ, s_t). In the model extension with a dynamic risk-aversion parameter, the pair would be (θ_t, s_t), i.e., both components change over time.) The observations o_t correspond to the market state s_t and the investor's actions $a_t^{\mathbf{H}}$. Such an action provides information about the unknown parameter θ_t to the robo-advisor. Therefore, the natural filtration of the POMDP, given by the σ-algebra generated by the observations $o_{1:t}$, corresponds to the set of public histories H_t, defined by (4.1).

Even though we can reformulate a POMDP as an MDP, the state space of the resulting MDPs is the belief-state simplex, typically high dimensional, which makes the MDP formulation intractable as well. This formulation nevertheless helps in obtaining approximate value functions and dual bounds, as discussed in the next section.

C.2.1 Approximate Value Functions and Primal Bounds

Approximate value functions (AVF) can be used to obtain sub-optimal policies for an MDP or POMDP. If we simulate such a policy, we obtain an unbiased estimator of a *primal* bound, which represents an upper (lower) bound for the optimal value of a minimization (maximization) problem. Section 4.4.4 describes a direct approach for constructing an AVF via the Q-function. Here, we show how that approach can be extended to obtain improved approximations that may yield better policies.

The QMDP AVF, introduced in Section 4.4.4, formulates the POMDP as a fully observed problem, i.e., the hidden states are fully observed at each time t, and defines the Q-function as

$$V_t^Q(x, a) := c_t(x, a) + \sum_{x' \in \mathcal{X}} P_{xx'}(a) \min_{a'} V_t^Q(x', a') \tag{C.4}$$

for $t \in \{0, \ldots, T-1\}$. The QMDP AVF is then defined by

$$\widetilde{V}_t^Q(\pi_t) := \min_{a_t} \sum_{x \in \mathcal{X}} \pi_t(x) V_t^Q(x, a_t). \tag{C.5}$$

[42] proposed the *fast informed* bound which improves on the QMDP AVF. This AVF formulates the POMDP as a problem where the hidden state x_{t-1} is known when action a_t is selected, for all $t < T$, so that the *Lag-1* value function, $V_t^{L_1}$, is calculated recursively via

$$V_t^{L_1}(x_{t-1}, a_{t-1}, o_t) := \min_{a_t} \mathbb{E}^{x_t, o_{t+1}}[r_t(x_t, a_t) + V_{t+1}^{L_1}(x_t, a_t, o_{t+1}) \mid x_{t-1}, o_t], \tag{C.6}$$

for $t \in \{1, \ldots, T-1\}$, with terminal condition $V_{T+1}^{L_1} \equiv 0$. We then define the Lag-1 AVF as

$$\widetilde{V}_t^{L_1}(\pi_t) := \min_{a_t} \mathbb{E}^{x_t, o_{t+1}}[r_t(x_t, a_t) + V_{t+1}^{L_1}(x_t, a_t, o_{t+1}) \mid \pi_t] \tag{C.7}$$

where the expectation is calculated with respect to the joint distribution of o_{t+1} and x_t, conditional on the current belief state, π_t.

More recently, [39] formulate a natural extension to the fast-informed bound by exchanging the order of the minimization and expectation operation, and obtain the so-called Lag-2 AVF. The complexity increases considerably for the calculation of such bound, but it provably provides a bound that is tighter than the fast informed bound.

The greedy policies corresponding to the AVF methods discussed above can then be accordingly defined as

$$g_t^Q(\pi_t) := \operatorname*{argmin}_{a \in \mathcal{A}} \sum_{x \in \mathcal{H}} \pi_t(x) V_t^Q(x, a_t) \tag{C.8}$$

$$g_t^{L_1}(\pi_t) := \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}\left[r_t(x_t, a) + V_{t+1}^{L_1}(x_t, a, o_{t+1}) \mid \pi_t\right] \tag{C.9}$$

for $t \in \{0, \ldots, T-1\}$.

We can then simulate each policy by first simulating the hidden states and observations to calculate the belief-state in each time t, and then taking the action prescribed by the greedy policy (C.8) or (C.9).

C.2.2 Dual Bounds

As discussed in Section 4.4.4, a dual bound is useful to be able to conclude that a policy is close enough to optimal. [39] have shown that the aforementioned AVFs are subsolutions[2], where a

[2] We highlight that [15] introduced the *subsolution* terminology.

subsolution ϑ is defined as any AVF that satisfies

$$\vartheta_t(\pi_t) \leq \min_{a_t \in \mathcal{A}} \{r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \pi_t]\} \tag{C.10}$$

for all belief states π_t, and all $t \in \{0, \ldots, T\}$. From (C.10), it follows immediately that any subsolution ϑ is a feasible solution of the linear programming formulation of the Bellman equation. Therefore, the subsolutions presented above are all dual (lower) bounds on the optimal value of the original minimization problem.

It is often the case that in realistic applications, it is impractical to calculate the tighter AVFs, and this may result in unsatisfactory duality gaps. [39] recently generalized information relaxation approaches to obtaining dual bounds for POMDPs, which are guaranteed to improve on the dual bounds given by the subsolutions themselves. Their methodology extends the information relaxation approach for MDPs, developed independently by [17] and [69]. Such an approach first relaxes the non-anticipativity constraints of feasible policies, and then penalizes violations of these constraints through the so-called *dual penalties* that act as action-dependent control variates in the optimization problem. In the context of MDPs, [15] showed that the use of subsolutions in the construction of the dual penalties guarantees a tighter bound than that obtained by the subsolution itself. In POMDPs, this becomes particularly useful since many AVFs are also subsolutions, as discussed above.

In our context, the duality gap resulting from the QMDP subsolution (dual bound) and greedy policy (primal bound) is small enough (see figure 4.2) to highlight the main qualitative properties of the solution.

C.3 Details on the Numerical Study

We provide further details on the numerical study conducted in Section 5.1. Recall that we are assuming the time horizon $T = 10$, and setting the number of portfolios to $n = 4$.

We take the state space to be the set of indices $\mathcal{S} := \{1, 2, \ldots, s^{(n)} = 21\}$, with state transitions

given by

$$P(s'|s) = \begin{cases} 0.5, & \text{if } s < s^{(n)} \text{ and } s' = s+1, \\ 0.5, & \text{if } s = s^{(n)} \text{ and } s' = s^{(n)}, \\ 0.5, & \text{if } s > 1 \text{ and } s' = s-1, \\ 0.5, & \text{if } s = 1 \text{ and } s' = 1, \\ 0, & \text{otherwise} \end{cases} \tag{C.11}$$

We take the portfolio space to be the set of indices $\mathcal{A}^{\mathbf{M}} := \{a^{(1)}, a^{(2)}, a^{(3)}, a^{(4)}\} \equiv \{1, 2, 3, 4\}$. As mentioned in Section 4.4, for each state $s \in \mathcal{S}$ and portfolio $i \in \{1, 2, 3, 4\}$ we have expected return $\mu(s, i)$ and standard deviation $\sigma(s, i)$. We define these state and portfolio dependent parameters using a parametric specification. For each portfolio i, we specify the expected return and standard deviation for the state in the middle of the index set $\mathcal{S}$ and then define the expected returns and standard deviations of the other states relative to $\tilde{s}$, i.e., for $\tilde{s} = 11$ we set $\mu(\tilde{s}, 1) = 0.05, \mu(\tilde{s}, 2) = 0.10, \mu(\tilde{s}, 3) = 0.15, \mu(\tilde{s}, 4) = 0.20$, and $\sigma(\tilde{s}, 1) = 0.05, \sigma(\tilde{s}, 2) = 0.15, \sigma(\tilde{s}, 3) = 0.30, \sigma(\tilde{s}, 4) = 0.50$. Hence, $\mu(\tilde{s}, i)$ and $\sigma(\tilde{s}, i)$ are increasing with respect to the index i, with portfolio 4 being the riskiest and portfolio 1 being the safer. The remaining parameters are given by

$$\begin{aligned} \mu(s, i) &= \mu(\tilde{s}, i) + 0.02(s - \tilde{s}), && \text{for } s \in \mathcal{S}, i = 1, 2, 3, 4, \\ \sigma(s, 1) &= \sigma(\tilde{s}, 1) - 0.005(s - \tilde{s}), && \text{for } s \in \mathcal{S}, \\ \sigma(s, 2) &= \sigma(\tilde{s}, 2) - 0.01(s - \tilde{s}), && \text{for } s \in \mathcal{S}, \\ \sigma(s, 3) &= \sigma(\tilde{s}, 3) - 0.02(s - \tilde{s}), && \text{for } s \in \mathcal{S}, \\ \sigma(s, 4) &= \sigma(\tilde{s}, 4) - 0.04(s - \tilde{s}), && \text{for } s \in \mathcal{S}. \end{aligned}$$

Hence, portfolios feature the risk-return tradeoff. Riskier portfolios have a higher standard deviation, i.e., $\sigma(s, 1) < \sigma(s, 2) < \sigma(s, 3) < \sigma(s_0, 4)$ for all $s \in \mathcal{S}$, and higher expected returns, i.e., $\mu(s, 1) < \mu(s, 2) < \mu(s, 3) < \mu(s, 4)$. Note also that, when the market move causes an increase in expected return, the standard deviation corresponding decreases. This characteristic is also observed in equity markets, and known as the leverage effect.

The space Θ of risk-aversion parameters consists of equally spaced points on a grid of size $m = 20$ on the interval $[0, 1]$, i.e., $\Theta = \{0.05, 0.10, \ldots, 0.95, 1.0\}$. The initial probability is chosen to be uniform over Θ, i.e., $\pi_1(\theta) = 0.05$ for all $\theta \in \Theta$, and zero otherwise. For the numerical study on the dynamic risk-aversion parameter, we define the transition function f as

$$f(\theta_t, a_t^{\mathbf{H}}, a_t^{\mathbf{M}}, s_t, s_{t+1}) := \begin{cases} \min(\theta_t + 0.05 \times a_t, 1.0), & \text{if } s_t < s_{t+1} \\ \theta_t, & \text{if } s_t = s_{t+1} \\ \max(\theta_t - 0.05 \times a_t, 0.05), & \text{if } s_t > s_{t+1} \end{cases}$$

where a_t is given by (4.5). Note that this definition of f captures the properties discussed in Section 4.4, i.e., changes in the risk-parameter are greater for riskier portfolios (as given by a_t), while the direction of the change in risk-aversion is determined by the sign of the state change, so that $\theta_{t+1} > \theta_t$ if $s_t < s_{t+1}$ and $\theta_{t+1} < \theta_t$ if $s_t > s_{t+1}$, as discussed in Section 4.4.