Journal of Machine Learning Research 26 (2025) 1-90 Submitted 1/25; Revised 7/25; Published 8/25

Optimizing Return Distributions with Distributional Dynamic Programming

Bernardo Ávila Pires BAVILAPIRES@GOOGLE.COM
Google DeepMind, London, UK

Mark Rowland
Google DeepMind

Diana Borsa
Google DeepMind

Zhaohan Daniel Guo
Google DeepMind

Khimya Khetarpal
Google DeepMind

André Barreto
Google DeepMind

David Abel
Google DeepMind

Rémi Munos
FAIR, Meta; work done at Google DeepMind

Will Dabney
Google DeepMind

Editor: Martha White

Abstract

We introduce distributional dynamic programming (DP) methods for optimizing statistical functionals of the return distribution, with standard reinforcement learning as a special case. Previous distributional DP methods could optimize the same class of expected utilities as classic DP. To go beyond, we combine distributional DP with *stock augmentation*, a technique previously introduced for classic DP in the context of risk-sensitive RL, where the MDP state is augmented with a statistic of the rewards obtained since the first time step. We find that a number of recently studied problems can be formulated as stock-augmented return distribution optimization, and we show that we can use distributional DP to solve them. We analyze distributional value and policy iteration, with bounds and a study of what objectives these distributional DP methods can or cannot optimize. We describe a number of applications outlining how to use distributional DP to solve different stock-augmented return distribution optimization problems, for example maximizing conditional value-at-risk, and homeostatic regulation. To highlight the practical potential of stock-augmented return distribution optimization and distributional DP, we introduce an agent that combines DQN and the core ideas of distributional DP, and empirically evaluate it for solving instances of the applications discussed.

Keywords: reinforcement learning, distributional reinforcement learning, risk-sensitive reinforcement learning, dynamic programming, stock-augmented Markov decision process

1. Introduction

Reinforcement learning (RL; Sutton and Barto, 2018; Szepesvári, 2022) is a powerful framework for building intelligent agents, and it has been successfully applied to solve many practical problems (Mnih et al., 2015; Silver et al., 2018; Bellemare et al., 2020; Degrave et al., 2022; Fawzi et al., 2022). In the standard formulation of the RL problem, the objective is to find a policy (a decision rule for selecting actions) that maximizes the expected (discounted) return in a Markov decision process (MDP; Puterman, 2014). A similar, related problem is what we refer to as *return distribution optimization*, where the objective is to optimize a functional of the return distribution (Marthe et al., 2024), which may not be the expectation. For example, we could maximize an *expected utility* (Von Neumann and Morgenstern, 2007; Bäuerle and Rieder, 2014; Marthe et al., 2024), that is, the expectation of the return "distorted" by some function.

By varying the choice of statistical functional being optimized (be it an expected utility or more general), we can model various RL-like problems as return distribution optimization, including problems in the field of risk-sensitive RL (Chung and Sobel, 1987; Chow and Ghavamzadeh, 2014; Noorani et al., 2022), homeostatic regulation (Keramati and Gutkin, 2011) and satisficing (Simon, 1956; Goodrich and Quigley, 2004).

The fact that return distribution optimization captures many problems of interest makes it appealing to develop solution methods for the general problem. At first glance, the apparent benefits of solving the general problem are offset by the fact that, for many instances, optimal stationary Markov policies do not exist (see, for example, Marthe et al., 2024). This can be problematic, because it rules out dynamic programming (DP; value iteration and policy iteration; Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018; Szepesvári, 2022) and various other RL methods that are designed to output stationary Markov policies. Defaulting to solution methods that produce history-based policies is an alternative we would like to avoid, under the premise that learning history-based policies can be intractable (Papadimitriou and Tsitsiklis, 1987; Madani et al., 1999).

We show that we can reclaim optimality of stationary Markov policies for many instances of return distribution optimization by augmenting the state of the MDP with a simple statistic we call *stock*. Stock is a backward looking quantity related to the agent's accumulated past rewards, including an initial stock (the precise definition is given in Section 2). It was introduced by Bäuerle and Ott (2011)[1] for maximizing conditional value-at-risk (Rockafellar et al., 2000). The MDP state and stock together provide enough information for stationary Markov policies (with respect to the state-stock pair) to optimize various statistical functionals of the distribution of returns offset by the agent's initial stock.

Incorporating stock into return distribution optimization gives rise to the specific formulation we consider in this paper, where the environment is assumed to be an MDP with states augmented by stock, and the return is offset by an initial stock. We refer to this formulation as *stock-augmented return distribution optimization*.

[1] Kreps (Example b, p. 269; 1977) outlined a similar statistic in the undiscounted setting.

The optimality guarantee for stationary stock-augmented Markov policies in return distribution optimization suggests that we may be able to develop DP solution methods for the instances where the guarantee applies. Classic value/policy iteration cannot cope with return distributions, but this limitation can be overcome using distributional RL (Chung and Sobel, 1987; Morimura et al., 2010; Bellemare et al., 2017, 2023). That is, we may resort to *distributional dynamic programming* to tackle return distribution optimization.

In the standard MDP setting, without stock, distributional DP methods already exist for policy evaluation (Chapter 5; Bellemare et al., 2023), for maximizing expected return (as an obvious adaptation), and for expected utilities (Marthe et al., 2024). However, these methods can only solve problems that classic DP can also solve (Marthe et al., 2024), namely, the return distribution optimization problems for which an optimal stationary Markov policy exists (with respect to the MDP states alone). Notably, by incorporating stock into distributional DP, we can optimize statistical functionals of the return distribution that we could not otherwise. Moreover, stock-augmented distributional DP is a single solution method for a variety of return distribution optimization problems (which so far have been studied and solved in isolation), and also a blueprint for practical methods to solve return distribution optimization, in much the same way the principles of classic DP and distributional policy evaluation factor into previously proposed, successful RL methods.

1.1 Paper Summary and Contributions

This paper is an in-depth study of distributional dynamic programming for solving stock-augmented return distribution optimization, and we make the following contributions:

1. We identify conditions on the statistical functional being optimized under which distributional DP can solve stock-augmented return distribution optimization, and develop a theory of distributional DP for solving this problem, including:

    - principled distributional DP methods (distributional value/policy iteration),
    - performance bounds and asymptotic optimality guarantees (for the cases that distributional DP can solve),
    - necessary and sufficient conditions for the finite-horizon case, plus mild sufficient conditions to the infinite-horizon discounted case.

2. We demonstrate multiple applications of distributional value/policy iteration for stock-augmented return distribution optimization, namely:

    - Optimizing expected utilities (Von Neumann and Morgenstern, 2007; Bäuerle and Rieder, 2014).
    - Maximizing conditional value-at-risk, a form of risk-sensitive RL, both the risk-averse conditional value-at-risk (Bäuerle and Ott, 2011), and a risk-seeking variant that we introduce.
    - Homeostatic regulation (Keramati and Gutkin, 2011), where the agent aims to maintain vector-valued returns near a target.
    - Satisfying constraints, and trading off minimizing constraint violations with maximizing expected return.

3. We show how to reduce stock-augmented return distribution optimization objective to a stock-augmented RL objective (via reward design); and that, in stock-augmented settings, classic DP cannot optimize all the return distribution optimization objectives that distributional DP can.

4. We introduce DηN (pronounced *din*), a deep RL agent that combines QR-DQN (Dabney et al., 2018) with the principles of distributional value iteration and stock augmentation to optimize expected utilities. Through experiments, we demonstrate DηN's ability to learn effectively under objectives in toy gridworld problems and the game of Pong in Atari (Bellemare et al., 2013).

1.2 Paper Outline

Section 2 introduces notation and basic definitions. In Section 3.1 we formalize the problem of stock-augmented return distribution optimization, and provide some basic example instances. Section 4 introduces distributional value/policy iteration and presents our main theoretical results. In Section 5, we discuss multiple applications of our results and show concrete examples of how to model different problems using stock augmentation and distributional DP (Sections 5.1 to 5.5 and 5.8).[2] In Section 5 we also explore implications of our results in different contexts: Generalized policy evaluation (Barreto et al., 2020; Section 5.6); reward design and the relationship between stock-augmented RL and stock-augmented return distribution optimization (Section 5.7). In Section 6, we introduce DηN and show how distributional DP can inform the design of deep RL agents. To highlight the practical implications of our contributions, in Section 7 we present an empirical study of DηN in different gridworld instances of some of the applications considered in Section 5. In Section 8 we complement our gridworld results with a demonstration of DηN controlling returns in a more complex setting: The Atari game of Pong, where we show that a single trained DηN agent can obtain various specific scores in a range, and where we use stock augmentation to specify the scores we want the agent to achieve. Section 9 concludes our work and presents directions for future work, notably practical questions revealed by our empirical study. We provide additional theoretical results in Appendix A. Appendix B contains the full analysis of distributional value/policy iteration, and Appendix C contains the full analysis of the conditions for our main theorems. Appendices D to G contain proofs for the results in Section 5. Appendix H contains implementation details for DηN and our experiments. Appendix I provides a summary guarantees for classic and distributional DP in the various different settings considered throughout this paper, and is a useful map for readers interested in understanding the kinds of problems that DP can solve.

2. Preliminaries

We write $\mathbb{N} \doteq \{1, 2, \ldots\}$ for the natural numbers excluding zero, and $\mathbb{N}_0 \doteq \{0, 1, 2, \ldots\}$. For $n \in \mathbb{N}_0$, $\Delta(n)$ denotes the $|n|$-dimensional simplex. For $m \in \mathbb{N}$, $\Delta(\mathbb{R}^m)$ denotes the set of probability distribution functions of $\mathbb{R}^m$-valued random variables.

[2] Some of these problems have been previously studied, and distributional DP is a novel solution approach in some cases (see Section 5).

We study the problems where an agent interacts with a Markov decision process (MDP; Puterman, 2014) with (possibly infinite) state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Rewards can be stochastic and the discount is $\gamma \in (0, 1]$. We adopt the convention that R_{t+1} is the reward random variable observed jointly with S_{t+1}, that is, R_{t+1}, S_{t+1} result from taking action A_t at state S_t, according to the MDP's reward and transition kernels.

The reward signal may be a vector-valued pseudo-reward (cumulant) signal (Sutton et al., 2011) in $\mathcal{C} \doteq \mathbb{R}^m$. The vector-valued case allows us to capture interesting applications that are worth the extra generality. However, to avoid unnecessary complication, our presentation is intentionally in terms of $\mathcal{C}$, so that the reader can easily appreciate the results in the scalar case ($\mathcal{C} = \mathbb{R}$) if they wish. We use the terms reward and returns to avoid an excess of *pseudo* prefixes in the text.

Some of our results only apply to finite-horizon MDPs. We say an MDP *has finite horizon* if there exists a constant $n \in \mathbb{N}$ such that S_n is terminal with probability one for any trajectory $S_0, A_0, S_1, A_1, \ldots, S_n$ generated in the MDP. We call the smallest such n the *horizon* of the MDP. A state s is terminal if $(S_{t+1}, R_{t+1}) = (s, 0)$ with probability one whenever $S_t = s$ (regardless of A_t). We refer to the case where the MDP has finite horizon as the *finite-horizon case* (complementary to the *infinite-horizon* case), and to the case where $\gamma < 1$ as the *discounted case* (complementary to the *undiscounted* case, where $\gamma = 1$).

We make the following assumption throughout the work, similar to Assumption 2.5 by Bellemare et al. (p. 19; 2023).

Assumption 1 (All rewards have uniformly bounded first moment)

$$\sup_{s,a \in \mathcal{S} \times \mathcal{A}} \mathbb{E}\left(\|R_{t+1}\|_1 \,|\, S_t = s, A_t = a \right) < \infty$$

Similar to Bäuerle and Ott (2011), we consider an augmented MDP state space $\mathcal{S} \times \mathcal{C}$. If s, a, r', s' is a transition in the original MDP, then for any $c \in \mathcal{C}$ the augmented MDP transitions as $(s, c), a, r', (s', \gamma^{-1}(c + r'))$, that is:

$$c_{t+1} = \frac{c_t + r_{t+1}}{\gamma}. \tag{1}$$

We refer to c_t as the agent's *stock*.[3] If we unroll the recursion in Equation 1 up to an *initial stock* c_0 (see Remark 2 below), we can interpret the stock, in a forward view, as a scaled sum of the initial stock c_0 and the discounted return from time step zero up to time step t:

$$c_t = \underbrace{\gamma^{-t}}_{\substack{\text{time-dependent}\\ \text{scaling}}} \Big(\underbrace{c_0}_{\substack{\text{initial}\\ \text{stock}}} + \underbrace{\sum_{i=0}^{t-1} \gamma^i r_{i+1}}_{\substack{\text{partial discounted}\\ \text{return}}} \Big).$$

In a backward view, the stock can be seen as a backward reverse-discounted return:

$$c_t = \gamma^{-1} r_t + \gamma^{-2} r_{t-1} + \cdots + \gamma^{-t} r_1 + \gamma^{-t} c_0.$$

[3] In our formulation the stock and the rewards are m-dimensional, whereas Bäuerle and Ott (2011) consider 1-dimensional stock.

Importantly, the stock allows us to keep track of the discounted return (plus the initial stock c_0) from time step 0, since, for all $t \geq 0$,

$$c_0 + \sum_{i=0}^{\infty} \gamma^i r_{i+1} = \gamma^t \left(c_t + \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \right),$$

When rewards (and stocks) are random, the above holds with probability one, written as

$$C_t + G_t = \gamma^{-t} \left(C_0 + G_0 \right), \tag{2}$$

with $G_t \doteq \sum_{i=0}^{\infty} \gamma^i R_{t+i+1}$ denoting the respective discounted return from time step t. Equation 2 will be key to optimizing return distributions: The distribution of $C_t + G_t$ will work as an "anytime proxy" for the distribution of $C_0 + G_0$, and by controlling the former distribution we can also control the latter—provided the objective is such that the γ^{-t} factor does not interfere with the optimization (we will later introduce this as an indifference of the objective to the discount γ).

Remark 2 (The Initial Stock c_0) *The expansion of stock includes an initial stock c_0 that is unspecified. Together with the initial MDP state s_0, this stock will form the initial augmented state (s_0, c_0). While the initial s_0 is often "given", c_0 can be set (even as a function of s_0). This will provide extra flexibility to policies, which may display diverse behaviors in response to changes in c_0, and it will allow us to reduce different problems to return distribution optimization by plugging in specific choices of c_0 (as a function of s_0). For example, as shown by Bäuerle and Ott (2011), we can choose c_0 in such a way that optimizing conditional value-at-risk reduces to an instance of return distribution optimization with stock augmentation (see Theorem 16 in Section 5.2).*

Remark 3 (Dynamics Influenced by Stock) *Our results do not rely on the transitions and rewards of the augmented MDP depending only on s. In a transition $(s, c), a, r', (s', c')$, c' must be updated according to Equation 1, but s', r' may depend on c. This can be useful, for example, to define termination conditions: The state s' may be terminal when $c' = 0$ or when $|c'|$ is too large.*

Stationary Markov policies with respect to stock are $\mathcal{S} \times \mathcal{C} \to \Delta(\mathcal{A})$ functions, and the space of these policies is $\Pi \doteq \Delta(\mathcal{A})^{\mathcal{S} \times \mathcal{C}}$. A Markov policy π is a sequence $\pi = \pi_0, \pi_1, \pi_2, \ldots$ of stationary policies $\pi_n : \mathcal{S} \times \mathcal{C} \to \Delta(\mathcal{A})$, and the space of these policies is $\Pi_{\mathrm{M}} \doteq \Pi^{\mathbb{N}}$. For a policy $\pi = \pi_0, \pi_1, \pi_2, \ldots$, returns are written as $G^\pi(s, c) \doteq \sum_{t=0}^{\infty} \gamma^t R_{t+1}$ where R_{t+1} are the rewards generated by starting at state $(S_0, C_0) = (s, c)$, then selecting $A_t \sim \pi_t(S_t, C_t)$ for $t \geq 0$. The return $G^\pi(s, c)$ may depend on c (even when rewards do not depend on the stock), because π may choose actions differently depending on c, so the trajectories generated depend on c as well. If π is stationary, then $A_t \sim \pi(S_t, C_t)$ for all $t \geq 0$.

A *history* is the sequence of everything observed preceding action A_t, that is,

$$H_t \doteq (S_0, C_0), A_0, R_1, (S_1, C_1), A_1, \ldots, R_t, (S_t, C_t),$$

The history at $t = 0$ is S_0, C_0. The set of possible histories of finite length is

$$\mathcal{H} \doteq \bigcup_{n \in \mathbb{N}_0} \underbrace{\mathcal{S} \times \mathcal{C}}_{(s_0, c_0)} \times (\underbrace{\mathcal{A}}_{a_t} \times \underbrace{\mathcal{C}}_{r_{t+1}} \times \underbrace{\mathcal{S} \times \mathcal{C}}_{(s_{t+1}, c_{t+1})})^n,$$

and a *history-based policy* is a function $\mathcal{H} \to \Delta(\mathcal{A})$. That is, a history-based policy makes decisions based on everything observed so far. For π history-based and $t \geq 0$ we have $A_t \sim \pi(H_t)$, and the set of all history-based policies is $\Pi_{\mathrm{H}} \doteq \Delta(\mathcal{A})^{\mathcal{H}}$.

We let $\Delta(\mathbb{R})$ be the set of distributions of $\mathbb{R}$-valued random variables. With $X \sim \nu$, we write $\mathrm{df}X = \nu$. For two $\mathcal{C}$-valued random variables X, X' we say $X \overset{\mathcal{D}}{=} X'$ if $\mathrm{df}X = \mathrm{df}X'$. For $\nu \in \Delta(\mathbb{R})$, we let QF_ν be the quantile function of ν:

$$\mathrm{QF}_\nu(\tau) \doteq \inf\{t \in \mathbb{R} : \mathbb{P}(X \leq t) \geq \tau\}. \qquad (X \sim \nu)$$

For $c \in \mathcal{C}$, we denote by δ_c the Dirac measure on c, that is, the distribution such that if $\mathbb{P}(G = c) = 1$ when $G \sim \delta_c$. The Dirac on zero is δ_0 (where in the vector-valued case it is understood that 0 refers to the all-zeros vector).

We define $\mathcal{D} \doteq \Delta(\mathcal{C})$ as the set of distributions of $\mathcal{C}$-valued random variables. The 1-*Wasserstein distance* for $\nu, \nu' \in \mathcal{D}$ is defined as (Definition 6.1, p. 105; Villani, 2009)

$$\mathrm{w}(\nu, \nu') \doteq \inf\left\{\mathbb{E}\|X - X'\|_1 : \mathrm{df}(X) = \nu, \mathrm{df}(X') = \nu'\right\},$$

where X and X' may be jointly distributed. In the scalar case ($\mathcal{C} = \mathbb{R}$), we have

$$\mathrm{w}(\nu, \nu') = \|\mathrm{QF}_\nu - \mathrm{QF}_{\nu'}\|_{\ell_1} = \mathbb{E}_{\tau \sim u_{(0,1)}}|\mathrm{QF}_\nu(\tau) - \mathrm{QF}_{\nu'}(\tau)|,$$

where $u_{(0,1)}$ denotes the uniform distribution in $(0, 1)$. Sometimes we will say the sequence $\nu_1, \nu_2, \ldots$ converges to ν_∞; when we say this, we mean convergence in 1-Wasserstein distance: $\lim_{n\to\infty} \mathrm{w}(\nu_n, \nu_\infty) = 0$. The *supremum* 1-*Wasserstein distance* is defined for $\eta, \eta' \in \mathcal{D}^{\mathcal{S}\times\mathcal{C}}$ as

$$\overline{\mathrm{w}}(\eta, \eta') \doteq \sup_{s\in\mathcal{S}, c\in\mathcal{C}} \mathrm{w}(\eta(s, c), \eta'(s, c)). \tag{3}$$

With a slight abuse of notation, we let $\mathrm{w}(\nu) \doteq \mathrm{w}(\nu, \delta_0)$ and $\overline{\mathrm{w}}(\eta) \doteq \sup_{s\in\mathcal{S}, c\in\mathcal{C}} \overline{\mathrm{w}}(\eta(s, c), \delta_0)$.

Given a policy $\pi \in \Pi_{\mathrm{H}}$, we define its *return distribution function* $\eta^\pi : \mathcal{S} \times \mathcal{C} \to \mathcal{D}$ by $\eta^\pi(s, c) \doteq \mathrm{df}(G^\pi(s, c))$ (for $(s, c) \in \mathcal{S} \times \mathcal{C}$).

We will make ample use of Banach's fixed point theorem (Theorem 1, p. 77, Szepesvári, 2022) and the following spaces:

$$(\mathcal{D}, \mathrm{w}) \doteq \{\nu \in \mathcal{D} : \mathrm{w}(\nu) < \infty\},$$
$$(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \doteq \{\eta \in \mathcal{D}^{\mathcal{S}\times\mathcal{C}} : \overline{\mathrm{w}}(\eta) < \infty\}.$$

These spaces are complete as shown in Lemma 23, Appendix A. Assumption 1 combined with $\gamma < 1$ or a finite-horizon MDP ensure that the return distributions of all policies are uniformly bounded, that is, $\sup_{\pi\in\Pi_{\mathrm{H}}} \overline{\mathrm{w}}(\eta^\pi) < \infty$.

Given a stationary policy $\pi \in \Pi$, we define the *stock-augmented distributional Bellman operator* $T_\pi : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ for $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ as follows: $(T_\pi\eta)(s, c)$ is the distribution of $R_{t+1} + \gamma G(S_{t+1}, C_{t+1})$ when $(S_t, C_t) = (s, c)$, $A_t \sim \pi(S_t, C_t)$, and $G(s, c) \sim \eta(s, c)$. We require that if s is terminal then $(T_\pi\eta)(s, c) = \delta_0$ for all $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and $c \in \mathcal{C}$.

On occasion, we will refer back to classic RL operators for comparison against the distributional case. We will denote the space of possible (state-) value functions by $(\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty) \doteq \left\{V \in \mathbb{R}^{\mathcal{S}} : \sup_{s\in\mathcal{S}} |V(s)| < \infty\right\}$. To avoid introducing further notation, we will also denote the classic Bellman operator by T_π. Whether the Bellman operator is classic or distributional will be clear from whether its argument is a return distribution function or a value function.

We let $x_+ \doteq \max\{x, 0\}$, $x_- \doteq \min\{x, 0\}$, and $\mathbb{I}(\cdot)$ be the indicator function.

3. Stock-Augmented Return Distribution Optimization

3.1 Problem Formulation

We are concerned with building intelligent agents that can do various things. When the agent can be expressed in terms of its behavior (a policy) and the outcome of the agent acting can be modeled as the stock-augmented discounted return generated by that policy, we can frame the problem of building intelligent agents as an optimization problem. A person looking to build an intelligent agent in this framework (we will call them *the designer*) is thus tasked with expressing what they want of agents as an objective to be optimized—where the better the agent, the higher the objective value of its policy.[4]

We propose to control the distribution of the quantity $c_0 + G^\pi(s_0, c_0)$,[5] which is the return generated by π from the initial augmented state $(s_0, c_0) \in \mathcal{S} \times \mathcal{C}$, offset by the initial stock c_0. We want an objective that quantifies how preferred $\mathrm{df}(c_0 + G^\pi(s_0, c_0))$ is for each policy π, so that we can phrase the problem of finding the most preferred policy. We can accomplish this with a statistical functional $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ that assigns a real number to each possible distribution of $c_0 + G^\pi(s_0, c_0)$, to phrase the optimization problem as:

$$\sup_{\pi \in \Pi_\mathrm{H}} K\mathrm{df}\left(c_0 + G^\pi(s_0, c_0)\right). \tag{4}$$

As an example, the standard RL problem can be expressed in Equation 4 by taking K to be the expectation:

$$\sup_{\pi \in \Pi_\mathrm{H}} \mathbb{E}(c_0 + G^\pi(s_0, c_0)) = c_0 + \sup_{\pi \in \Pi_\mathrm{H}} \mathbb{E}(G^\pi(s_0, c_0)).$$

The optimization, for the moment, is over the (most general) class of history-based policies Π_H. In standard RL, this problem formulation (adopted, for example, by Altman, 1999) differs from the more frequent optimization over stationary Markov policies (adopted, for example, by Sutton and Barto, 2018; Szepesvári, 2022), but the two formulations are equivalent in MDPs because of the existence of optimal stationary Markov policies (Puterman, 2014). For stock-augmented return distribution optimization, we have elected to introduce the problem in terms of history-based policies, and to address the existence of optimal stationary Markov policies on the solution side of the results (in connection to DP; see Appendix B.1).

Because the supremum in Equation 4 is over all history-based policies, it makes sense to talk about optimizing $K\mathrm{df}\left(c_0 + G^\pi(s_0, c_0)\right)$ simultaneously for all $(s_0, c_0) \in \mathcal{S} \times \mathcal{C}$. We can state this problem concisely, using an objective functional applied to the return distribution function η^π:

[4] In practice, the designer is also tasked with modeling the environment as an MDP. In standard RL, this means designing the states, actions and rewards. Stock-augmented MDPs additionally require designing the stock and the pseudo-rewards.

[5] In terms of a problem/solution separation, incorporating stock is part of the solution (distributional DP). However, because the scope of our work is DP, it is convenient for our presentation to incorporate stock augmentation and the offset by c_0 as part of the problem (return distribution optimization). The simpler formulation without stock augmentation or the offset is limiting for distributional DP: Marthe et al. (2024) studied return distribution optimization without stock augmentation in the finite-horizon undiscounted setting, and concluded that only exponential utilities could be optimized through distributional DP—the same class that classic DP can optimize. On the other hand, as our analysis will show, the distributional DP with stock can optimize a broader class of objectives than without, and, surprisingly, than classic DP with stock augmentation.

Definition 4 (Stock-Augmented Return Distribution Optimization) *Given $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$, define the* stock-augmented objective functional $F_K : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to \mathbb{R}^{\mathcal{S}\times\mathcal{C}}$ *as*

$$(F_K\eta)(s, c) \doteq K\mathrm{df}(c + G(s, c)). \qquad (G(s, c) \sim \eta(s, c))$$

The stock-augmented return distribution optimization problem *is*

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^\pi. \tag{5}$$

We will often drop the subscript and refer to a stock-augmented objective as F, in which case a corresponding K is implied. We will also drop df and write $K(G) = K\mathrm{df}(G)$.

To recap Equation 5: The stock-augmented return distribution optimization problem consists of optimizing, over all policies $\pi \in \Pi_{\mathrm{H}}$, a preference specified by a statistical functional $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$, over the distribution of the policy's discounted return offset by the stock $(c_0 + G^\pi(s_0, c_0))$. The optimization is considered simultaneously for all (s_0, c_0), as allowed by history-based policies.

3.2 Example: Expected Utilities

Equation 5 provides a flexible problem formulation for controlling $\mathrm{df}(c_0 + G^\pi(s_0, c_0))$, based on a choice of $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ provided by a designer to capture what they want an agent to achieve. We have already shown that the RL problem can be recovered by taking K to be the expectation ($K\nu = \mathbb{E}G$, $G \sim \nu$), so what else can we do? We can obtain an interesting family of objective functionals by considering the expected value of transformations of the return specified by a function $f : \mathcal{C} \to \mathbb{R}$: $K\nu = \mathbb{E}f(G)$ $(G \sim \nu)$. These are the *expected utilities*, which have been widely studied in decision-making theory (Von Neumann and Morgenstern, 2007), and also used for sequential decision-making in RL (Bäuerle and Rieder, 2014; Bowling et al., 2023).

Definition 5 *A stock-augmented objective functional F_K is an* expected utility *if there exists $f : \mathcal{C} \to \mathbb{R}$ such that*

$$K\nu = \mathbb{E}f(G). \qquad (G \sim \nu)$$

In this case, we write $F_K = U_f$, which can be written as

$$(U_f\eta)(s, c) \doteq \mathbb{E}f(c + G(s, c)). \qquad (G(s, c) \sim \eta(s, c))$$

Table 1 gives examples of return distribution optimization problems resulting from different choices of f in the scalar case[6] ($\mathcal{C} = \mathbb{R}$), with some notable risk-sensitive examples: Maximizing conditional value-at-risk (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Lim and Malik, 2022) and maximizing the probability of the discounted return being above a threshold. Recall that the choice of initial stock c_0 is "up to the user" and can be made as a function of the starting state s_0.

We will later show that the examples in the first part of the table can be optimized by distributional DP both in the finite-horizon and discounted cases, the ones in the second

[6] Expected utilities are not restricted to the scalar case, as implied by Definition 5, since the domain of f is $\mathcal{C}$. We provide some concrete examples in Section 5 of expected utilities for the vector-valued case.

Problem	$f(x)$	Formulation
Standard RL	x	$\sup_{\pi\in\Pi_{\mathrm{H}}}\mathbb{E}(c_0+G^\pi(s_0,c_0))$ $\equiv\sup_{\pi\in\Pi_{\mathrm{H}}}\mathbb{E}G^\pi(s_0,\cdot)$
Minimize the expected absolute distance to a target c_0 (Section 5.1)	$-\lvert x\rvert$	$\inf_{\pi\in\Pi_{\mathrm{H}}}\mathbb{E}\lvert G^\pi(s_0,-c_0)-c_0\rvert$
Optimizing τ-CVaR (conditional value-at-risk, Section 5.2)	x_-	$\inf_{\pi\in\Pi_{\mathrm{H}},c_0}\frac{1}{\tau}\int_0^\tau \mathrm{QF}_{\eta^\pi(s_0,c_0)}(t)\mathrm{d}t$
Maximize the probability of the return above a threshold c_0	$\mathbb{I}(x>0)$	$\sup_{\pi\in\Pi_{\mathrm{H}}}\mathbb{P}(G^\pi(s_0,-c_0)>c_0)$
Minimize the expected square distance to a target c_0	$-x^2$	$\inf_{\pi\in\Pi_{\mathrm{H}}}\mathbb{E}\left((G^\pi(s_0,-c_0)-c_0)^2\right)$
Maximize the probability of the return above a threshold plus a margin c_0+c	$\mathbb{I}(x>c)$	$\sup_{\pi\in\Pi_{\mathrm{H}}}\mathbb{P}\left(G^\pi(s_0,-c_0)>c_0+c\right)$

Table 1: Example problems that can be formulated as optimizing an expected utility, with the respective choices of f and the formulation.

part of the table can be optimized in the finite-horizon case, and the example in the third part can only be optimized in the finite-horizon undiscounted case (see Theorems 6 and 8, Section 4.3, and Appendix C.2).

We will also establish that distributional DP can, in fact, optimize any expected utility in the finite-horizon undiscounted case (see Lemma 12). Going beyond expected utilities, we will see that is an open question whether it is possible for distributional DP to optimize any non-expected utility in the infinite-horizon discounted case, but we provide examples that can be optimized in the finite-horizon case (see Section 5.8).

4. Distributional Dynamic Programming

Dynamic programming (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018) is at the heart of RL theory and many RL algorithms.[7] For this reason, we have chosen to establish the basic theory of solving stock-augmented return distribution optimization by studying how we can solve these problems using DP. We refer to the solution methods we introduce as *distributional dynamic programming.* As in the case of distributional DP for policy evaluation (Chapter 5; Bellemare et al., 2023), return distribution functions (in $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}},\overline{\mathrm{w}})$) are the main object of distributional value/policy iteration, whereas, in contrast, classic DP, namely value/policy iteration, work directly with value functions (see, for example, Szepesvári, 2022).

[7] As pointed out by Szepesvári (2022) many RL algorithms can be thought of as dynamic programming methods modified to cope with scale and complexity of practical problems.

4.1 Distributional Value Iteration

Classic value iteration computes the iterates $V_1, V_2, \dots$ satisfying, for $n \geq 0$,

$$V_{n+1} = \sup_{\pi \in \Pi} T_\pi V_n, \tag{6}$$

and the procedure enjoys the following optimality guarantees. In finite-horizon MDPs, V_n is optimal if n is at least the horizon of the MDP and in the discounted case (Section 2.4; Szepesvári, 2022):

$$V^* - V_n \leq \gamma^n \|V^* - V_0\|_\infty \tag{7}$$

pointwise for all $s \in \mathcal{S}$, where $V^* \doteq \sup_{\pi \in \Pi_\mathrm{H}} V^\pi$ and V^π denotes the value function of a policy π.

Note how the bounds are distinct for the finite-horizon case and the discounted case. This distinction recurs in results for both classic and distributional value/policy iteration, and it will merit further discussion in the case of distributional DP.

In classic value iteration, the iterates correspond to the values of the objective functional being optimized, and the iteration in Equation 6 makes a one-step decision that maximizes that objective functional. We typically use the value iterates to obtain policies via a greedy selection, and leverage a near-optimality guarantee for these greedy policies. We say $\tilde{\pi}_n$ is a greedy policy with respect to V_n if it satisfies the following:

$$T_{\tilde{\pi}_n} V_n = \sup_{\pi \in \Pi} T_\pi V_n.$$

Classic value iteration results give us the following optimality guarantees for the greedy policies: In finite-horizon MDPs, $\tilde{\pi}_n$ is optimal when n is at least the horizon of the MDP, and in the discounted case (Section 2.4, Szepesvári, 2022; Singh and Yee, 1994):

$$V^* - V^{\tilde{\pi}_n} \leq \frac{2\gamma^n}{1-\gamma} \|V^* - V_0\|_\infty. \tag{8}$$

Distributional value iteration, while similar to value iteration, maintains distributional iterates $\eta_1, \eta_2, \ldots \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, which means the iterates no longer correspond to values of the objective functional. The distributional analogue of Equation 6 makes a one-step decision that maximizes F_K, and this iteration of locally optimal one-step decisions gives guarantees similar to the classic case. Theorem 6 formalizes this claim:[8]

Theorem 6 (Distributional Value Iteration) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, then for every $\eta_0 \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, if the iterates $\eta_1, \eta_2, \ldots$ satisfy (for $n \geq 0$)*

$$F_K \eta_{n+1} = \sup_{\pi \in \Pi} F_K T_\pi \eta_n, \tag{Distributional Value Iterates}$$

and the policies $\overline{\pi}_0, \ldots, \overline{\pi}_n$ satisfy (for $n \geq 0$),

$$F_K T_{\overline{\pi}_n} \eta_n = \sup_{\pi \in \Pi} F_K T_\pi \eta_n, \tag{Greedy Policies}$$

[8] To simplify the presentation, we have chosen to present the distributional DP results upfront, and discuss the conditions on the objective functional F_K in Section 4.3.

then the following hold.

Finite-horizon case: *for all n greater or equal to the horizon of the MDP,*

$$F_K\eta_n = \sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi, \tag{9}$$

and

$$F_K\eta^{\overline{\pi}_n} = \sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi. \tag{10}$$

Discounted case ($\gamma < 1$): *If K is L-Lipschitz, then for all $n \geq 0$*

$$\sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi - F_K\eta_n \leq L\gamma^n \cdot \sup_{\pi\in\Pi_\mathrm{M}} \overline{\mathrm{w}}(\eta_0, \eta^\pi), \tag{11}$$

and

$$\sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi - F_K\eta^{\overline{\pi}_n} \leq L\gamma^n \cdot \left(\frac{1}{1-\gamma}\sup_{\pi\in\Pi} \overline{\mathrm{w}}(T_\pi\eta_0, \eta_0) + \sup_{\pi\in\Pi_\mathrm{M}} \overline{\mathrm{w}}(\eta_0, \eta^\pi)\right). \tag{12}$$

Next, we discuss a number of aspects of our value iteration result.

Iterates may not converge. The guarantees in Theorem 6 only apply to values of the objective functional $F_K\eta_n$, and iterate convergence cannot be guaranteed because multiple iterates may be tied at the optimum. Iterate non-convergence has been identified before in distributional RL, as multiple return distributions can be optimal (Example 7.11, p. 210, Bellemare et al., 2023).

Comparison to classic DP bounds in the finite-horizon case. The guarantees for finite-horizon MDPs are essentially the same for distributional and classic value iteration: Namely, optimality after iterating at least as many times as the MDP horizon.

Comparison to classic DP bounds in the discounted case. In the discounted case, the bounds for distributional value iteration (Equations 11 and 12) are similar to the classic value iteration bounds (Equations 7 and 8) with three notable differences:

i) The bounding terms are 1-Wasserstein distances, rather than ∞-norms. This is inherent to the fact that our iterates are distributional.

ii) The Lipschitz constant of K is present. This constant is 1 when F_K is the standard RL objective functional.

iii) The classic value iteration bounds are given in terms of V^*, but the distributional value iteration bounds are not. This is because it is still an open question whether an optimal return distribution η^* exists in the discounted case in general. However, if we assume η^* exists, we can replace the bounding term in Equation 11 with $L\gamma^n \cdot \overline{\mathrm{w}}(\eta_0, \eta^*)$, which is comparable to the classic DP bounds.

The considerations above apply similarly to the greedy policy bounds for distributional and classic DP.

When an optimal return distribution η^* exists, we can also show an optimality guarantee for policies that are greedy with respect to η^*, similar to the classic case:

Theorem 7 (Greedy Optimality) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, and if: i) the MDP has finite horizon; or ii) $\gamma < 1$ and K is Lipschitz, then the following hold.*

There exists an optimal return distribution $\eta^* \in \mathcal{D}^{\mathcal{S}\times\mathcal{C}}$ satisfying

$$F_K\eta^* = \sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi,$$

iff the supremum on the right-hand side is attained (that is, an optimal policy exists).

If such η^* exists, then any greedy policy with respect to η^* is optimal (and thus attains the supremum above).

4.2 Distributional Policy Iteration

Classic policy iteration computes the iterates $\pi_1, \pi_2, \ldots$ satisfying, for $n \geq 0$,

$$T_{\pi_{n+1}}V^{\pi_n} = \sup_{\pi\in\Pi} T_\pi V^{\pi_n},$$

that is, each iterate π_{n+1} is greedy with respect to the value of the previous iterate π_n. In finite-horizon MDPs, V^{π_n} is optimal if n is at least the horizon of the MDP. In the discounted case, we have (Proposition 2.8, p. 45; Bertsekas and Tsitsiklis, 1996):

$$V^* - V^{\pi_n} \leq \gamma^n \|V^* - V^{\pi_0}\|_\infty.$$

Distributional policy iteration is similar to its classic counterpart (he main difference being that the objective functional F_K determines the greedy policy selection) and also enjoys similar guarantees, as formalized by Theorem 8:

Theorem 8 (Distributional Policy Iteration) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, for every stationary policy $\pi_0 \in \Pi$ if the iterates $\pi_1, \pi_2, \ldots$ satisfy (for $n \geq 0$)*

$$F_K T_{\pi_{n+1}}\eta^{\pi_n} = \sup_{\pi\in\Pi} F_K T_\pi \eta^{\pi_n} \qquad \text{(Distributional Policy Iterates)}$$

then the following hold.

Finite-horizon case: *For all n greater or equal to the horizon of the MDP,*

$$F_K\eta^{\pi_n} = \sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi. \tag{13}$$

Discounted case ($\gamma < 1$): *If K is L-Lipschitz, then for all $n \geq 0$*

$$\sup_{\pi\in\Pi_\mathrm{H}} F_K\eta^\pi - F_K\eta^{\pi_n} \leq L\gamma^n \cdot \sup_{\pi\in\Pi_\mathrm{M}} \overline{\mathrm{w}}(\eta^{\pi_0}, \eta^\pi), \tag{14}$$

Comparison to classic policy iteration bounds. Essentially the same considerations apply here as in Section 4.1, for comparing the respective value iteration bounds. This is because we obtain the policy iteration bounds from the value iteration bounds, using the fact that $V^{\pi_n} \geq V_n$ for classic DP, and $F_K\eta^{\pi_n} \geq F_K\eta_n$ for distributional DP (see the proof of Theorem 8 in Appendix B.6).

4.3 Conditions Overview

Theorems 6 and 8 only apply to objective functionals that satisfy certain properties: Indifference to mixtures and indifference to γ in the finite-horizon case, plus Lipschitz continuity in the infinite-horizon discounted case. In this section we give an overview of these conditions and test them: How restrictive are these conditions? Can they be weakened? The proofs for the results in this section can be found in Appendix C. Recall that we are abusing notation and writing $K(G) = K\mathrm{df}(G)$.

Definition 9 (Indifference to Mixtures (of Initial Augmented States)) *We say* $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ *is* indifferent to mixtures *(of initial augmented states) if for every* $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ *such that*

$$K\eta(s,c) \geq K\eta'(s,c),$$

for all $(s,c) \in \mathcal{S} \times \mathcal{C}$*, then for all random variables* (S,C) *taking values in* $\mathcal{S} \times \mathcal{C}$ *we also have*

$$K(G(S,C)) \geq K(G'(S,C)). \qquad (G(s,c) \sim \eta(s,c),\ G'(s,c) \sim \eta'(s,c))$$

Definition 10 (Indifference to γ) *We say* $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ *is* indifferent to γ *if, for every* $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$

$$K\nu \geq K\nu' \Rightarrow K(\gamma G) \geq K(\gamma G'). \qquad (G \sim \nu,\ G' \sim \nu')$$

Definition 11 (Lipschitz Continuity) *We say* $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ *is* L-Lipschitz *(or* Lipschitz*, for simplicity) if there exists* $L \in \mathbb{R}$ *such that*

$$\sup_{\substack{\nu,\nu':\\ \mathrm{w}(\nu)<\infty\\ \mathrm{w}(\nu')<\infty\\ \mathrm{w}(\nu,\nu')>0}} \frac{|K\nu - K\nu'|}{\mathrm{w}(\nu,\nu')} \leq L.$$

L *is the* Lipschitz constant *of* K*.*

We believe that in general these conditions are fairly easy to verify for different choices of K. As an example, Lemma 12 does part of the verification for expected utilities.

Lemma 12 (Conditions for Expected Utilities) *Let* U_f *be an expected utility, which is an objective functional* F_K *with* $K\nu = \mathbb{E}f(G)$ $(G \sim \nu)$*. Then the following hold:*

1. K *is indifferent to mixtures.*

2. K *is indifferent to* γ *iff there exists* $\alpha \in (0,1]$ *such that* $\gamma < 1 \Rightarrow \alpha < 1$ *and, for all* $c \in \mathcal{C}$*,*

$$f(\gamma c) = \alpha f(c) + (1-\alpha) f(0). \tag{15}$$

3. K *is* L*-Lipschitz iff* f *is* L*-Lipschitz.*

The condition for indifference to γ is interesting because it means $c \mapsto f(c) - f(0)$ is positively homogeneous with degree $\log_\gamma \alpha$.

If we refer back to Table 1, we see that the choices of f in the first part of the table satisfy all three conditions, so distributional DP can optimize the corresponding U_f both in the finite-horizon and discounted cases. The choices of f in the second part of the table are not Lipschitz, so we know that DP can optimize the corresponding U_f in the finite-horizon setting. The choice of U_f in the third part of the table is neither Lipschitz nor indifferent to $\gamma < 1$, so distributional DP is only guaranteed to optimize the corresponding U_f in the finite-horizon *undiscounted* setting. A consequence of Lemma 12, since indifference to $\gamma = 1$ is trivially true, is that distributional DP can optimize any expected utility in the finite-horizon undiscounted case.

We have investigated the three conditions (Definitions 9 to 11) to determine how restrictive they are. We have found that indifference to mixtures and indifference to γ are necessary and sufficient, so they are minimal. In the absence of either, even a basic greedy optimality guarantee (Theorem 7) fails:

Proposition 13 *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is not indifferent to mixtures or not indifferent to γ, then there exists an MDP, an $\eta^* \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and a $\overline{\pi} \in \Pi$ such that $\overline{\pi}$ is greedy with respect to η^* and*

$$F_K \eta^* = \sup_{\pi \in \Pi_{\mathrm{H}}} F_K \eta^\pi,$$

however, for some $(s, c) \in \mathcal{S} \times \mathcal{C}$

$$F_K \eta^{\overline{\pi}}(s, c) < \sup_{\pi \in \Pi_{\mathrm{H}}} F_K \eta^\pi(s, c).$$

We have found that the relationship between Lipschitz continuity and the infinite-horizon discounted case is less clear, and it is still an open question whether this property is necessary. However, we can show that indifference to mixtures and indifference to γ are not sufficient for the infinite-horizon discounted case, so there is a real distinction between the finite-horizon and infinite-horizon discounted cases, in line with our results for distributional DP (Theorems 6 and 8).

In Appendix C.2, we show an instance where distributional value/policy iteration fail for the expected utility U_f with $f(x) = \mathbb{I}(x > 0)$, even though the starting iterate is optimal. The intuition for this is simple and we outline it here (the key is to exploit the fact that f is not continuous). Consider an MDP with $\mathcal{S} = \{s_0, s_1\}$, $\mathcal{A} = \{a_0, a_1\}$, $r(\cdot, a_i) = i$ and $\gamma < 1$. The initial state is s_0 and s_1 is terminal, and taking a_i in s_0 transitions to s_i. A stationary policy $\pi \in \Pi$ satisfying $\pi(a_1|s_0, \cdot)$ is optimal, so let us denote it by π^* and its return distribution function by η^*. Thanks to η^*, we have

$$U_f T_\pi \eta^* = U_f \eta^*$$

for all $\pi \in \Pi$, including a policy that always selects a_0, and, in fact, by induction, any non-stationary policy that selects a_0 finitely many times is also optimal, even though selecting a_0 *always* is suboptimal. In the case of distributional value iteration with $\eta_0 = \eta^*$, if we take $\overline{\pi}_n$ to be the policy that always selects a_0 we will have $U_f \eta_n = U_f \eta^*$ for all n, however, $U_f \eta^{\overline{\pi}_n} < U_f \eta^*$ also for all n, which means distributional value iteration has failed.

Distributional policy iteration fails too, except that when starting from π^* every other iterate may be suboptimal depending on how ties are broken.

The assumption on Lipschitz continuity of f for the infinite-horizon discounted case prevents failures like the example above (which we attributed to the fact that f is not continuous). In Appendix C.2 we also show that the lack of Lipschitz continuity affects our ability to evaluate policies, in the sense that if we take $f(x) = x^2$ (which is continuous but not Lipschitz) we can construct an MDP and a policy $\pi \in \Pi$ such that $T_\pi^n \eta$ converges to η^π as $n \to \infty$, but $U_f T_\pi^n \eta$ does not converge uniformly to $U_f \eta^\pi$ (though it converges pointwise).

It is unclear whether the lack of uniform convergence for non-Lipschitz f can be translated to a failure of distributional value/policy iteration, however we have a failure case example of a discontinuous f, so it suggests that some property related to continuity of f (and K more generally) is necessary.

4.4 Analysis Overview

The valuable insight in this work is that we can use distributional DP to optimize different objective functionals F_K of the (stock-augmented) return distribution (and a broader class than without). Once we identify the right conditions and the core components for distributional value/policy iteration to work, the remaining work is relatively straightforward: We retrace the steps of classic DP and ensure technical correctness. Most of the challenge is, in fact, ensuring technical correctness with a generic objective functional—for example, we need to be careful to make correct statements about convergence; we cannot rely on the existence of an optimal return distribution η^* or on the convergence of distributional value iterates.

In this section, we give an outline of our analysis with the most interesting points and a focus on how we can obtain asymptotic optimality guarantees. This will allow us to understand how the different conditions factor into our proofs, and how they work in essence. We defer the technical proofs to Appendix B, including details about performance bounds.

A fundamental component for DP is monotonicity. In classic RL (see Lemma 2.1, p. 21, Bertsekas and Tsitsiklis, 1996), it states that if we have $V \geq V'$, then following a policy π for one step and having a value of V afterward is always better than following the same policy but obtaining a value of V' afterward, regardless of the policy π. That is, we have

$$V \geq V' \Rightarrow T_\pi V \geq T_\pi V'$$

for all $\pi \in \Pi$. In distributional DP, it translates to the following:

Lemma 14 (Monotonicity) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, then, for every $\pi \in \Pi$, the distributional Bellman operator T_π is* monotone *(or* order-preserving*) with respect to the preference induced by F_K on $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$. That is, for every stationary policy $\pi \in \Pi$ and $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, we have*

$$F_K \eta \geq F_K \eta' \Rightarrow F_K T_\pi \eta \geq F_K T_\pi \eta'.$$

Monotonicity is a powerful result that underpins value iteration, policy iteration and also *policy improvement*.[9] Classic policy improvement (see Proposition 2.4, p. 30, Bertsekas

[9] To underscore the importance of monotonicity, we note that the result in Proposition 13 holds essentially because monotonicity is equivalent to K being indifferent to mixtures and indifferent to γ, and it is the absence of monotonicity that causes greedy optimality (Theorem 7) to fail.

and Tsitsiklis, 1996) states that if a policy $\tilde{\pi}$ is greedy with respect to V^{π}, then $\tilde{\pi}$ is better than π ($V^{\tilde{\pi}} \geq V^{\pi}$). We have a similar result for distributional DP, given as Lemma 15. This result is of particular interest here because its proof gives a good sense of how to provide asymptotic guarantees for distributional DP, and how the different conditions factor in, in particular how departing from the standard RL case in classic DP demands special attention to convergence guarantees.

Lemma 15 (Distributional Policy Improvement) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, and if: i) the MDP has finite horizon; or ii) $\gamma < 1$ and K is Lipschitz, then for $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and any stationary policy $\overline{\pi} \in \Pi$ if*

$$F_K T_{\overline{\pi}} \eta \geq F_K \eta, \tag{16}$$

then

$$F_K \eta^{\overline{\pi}} \geq F_K \eta.$$

In particular, for any stationary policy $\pi \in \Pi$, if $\overline{\pi}$ satisfies

$$F_K T_{\overline{\pi}} \eta^{\pi} = \sup_{\pi' \in \Pi} F_K T_{\pi'} \eta^{\pi}, \tag{Greedy Policy}$$

then Equation 16 is satisfied with $\eta = \eta^{\pi}$ and we have

$$F_K \eta^{\overline{\pi}} \geq F_K \eta^{\pi}. \tag{Greedy Policy Improvement}$$

Proof We write $F = F_K$ for simplicity, and fix $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ arbitrary. Indifference to mixtures and indifference to γ give us monotonicity. By induction, for all $n \geq 1$, if we assume that Equation 16 holds and that $F T_{\overline{\pi}}^{n} \eta \geq F\eta$, then

$$\begin{aligned} F T_{\overline{\pi}}^{n+1} \eta &= F T_{\overline{\pi}} T_{\overline{\pi}}^{n} \eta \\ &\geq F T_{\overline{\pi}} \eta && \text{(Monotonicity, induction assumption)} \\ &\geq F\eta. && \text{(Equation 16)} \end{aligned}$$

Thus, if Equation 16 holds, then, for all $n \geq 1$,

$$F T_{\overline{\pi}}^{n} \eta \geq F\eta. \tag{17}$$

In the finite-horizon case, we can take n to be the horizon of the MDP and the result follows, since $T_{\overline{\pi}}^{n} \eta = \eta^{\overline{\pi}}$.

In the infinite-horizon discounted case, the induction argument is not enough to show that $F\eta^{\overline{\pi}} \geq F\eta$, since we need Equation 17 to hold in the limit. In this case, we have $\gamma < 1$, $T_{\overline{\pi}}$ is a contraction (see Lemma 26 and Proposition 4.15, p. 88, Bellemare et al., 2023) and $\overline{\mathrm{w}}(\eta) < \infty$, so $T_{\overline{\pi}}^{n} \eta$ converges to $\eta^{\overline{\pi}}$. K Lipschitz implies F Lipschitz by Proposition 31, and because F is Lipschitz, the convergence of $T_{\overline{\pi}}^{n} \eta$ to $\eta^{\overline{\pi}}$ implies the convergence of $F T_{\overline{\pi}}^{n} \eta$ to $F\eta^{\overline{\pi}}$ (see Proposition 32). Thus, Equation 17 holds in the limit of $n \to \infty$, which gives the result:

$$F\eta^{\overline{\pi}} = \lim_{n\to\infty} F T_{\overline{\pi}}^{n} \eta \geq F\eta.$$

For the greedy policy improvement result for stationary π, it suffices to use the fact that $T_\pi \eta^\pi = \eta^\pi$, so the choice of greedy policy gives

$$FT_{\overline{\pi}}\eta^\pi = \sup_{\pi' \in \Pi} FT_{\pi'}\eta^\pi \geq FT_\pi \eta^\pi = F\eta^\pi.$$

which gives us Equation 16. ■

As we can see in the proof of Lemma 15, indifference to mixtures and indifference to γ are connected to monotonicity, whereas Lipschitz continuity is used to ensure that $FT_{\overline{\pi}}^n \eta^\pi$ converges to $F\eta^{\overline{\pi}}$ as $n \to \infty$. In terms of asymptotic convergence, the main additional technical challenge in the proofs of Theorems 6 and 8 comes from the fact that iterates do not necessarily converge. However, it is still possible to show that the value of the objective functional converges uniformly for all starting augmented states. Then the induction argument for chaining improvements (Equation 17), and the use of monotonicity and Lipschitz continuity are essentially the same as in the proof of Lemma 15.

The condition in Equation 16 in Lemma 15 corresponds to the assumption that $\overline{\pi}$ is a one-step improvement on η. We can always improve on return distributions of stationary policies with a greedy policy (as the second part of Lemma 15 shows), however improvement is not always possible for return distributions of non-stationary policies. To see this, consider a finite-horizon binary-tree MDP and a non-stationary policy $\pi = \pi_1, \pi_2, \ldots$ where each π_t has optimal performance on the t-th level of the tree, but poor performance in all other states. The policy $(\overline{\pi}, \pi_1, \pi_2, \ldots)$ would suffer from the poor performance of all π_t because of the time-shift introduced by first following $\overline{\pi}$ and then π. Importantly, however, when η is optimal, even over non-stationary policies, we can satisfy Equation 16. This is used in the proof of the distributional value iteration result (Theorem 6) for finite-horizon MDPs: In an MDP with horizon n, the iterates η_n and $\eta_{n+1} = T_{\overline{\pi}_n}\eta_n$ are optimal (where, recall, $\overline{\pi}_n$ is greedy with respect to η_n), so we can use Lemma 15 to show that $F\eta^{\overline{\pi}_n} \geq F\eta_n$ and therefore $\overline{\pi}_n$ is optimal.

4.5 Previous Distributional Dynamic Programming Results

From the vantage point provided by the results in this section, we can better appreciate the landscape of distributional DP in the literature: The core elements of distributional DP for stock-augmented return distribution optimization have been studied before, albeit separately, and with different analysis techniques for the standard case and the stock-augmented case. Our results expand the stock-augmented problems that can be demonstrably solved by distributional DP beyond what was previously known and beyond what can be achieved without stock augmentation, and our analysis adapts the commonly used tools for the standard case (see, for example, Bertsekas and Tsitsiklis, 1996) to the stock-augmented case. Moreover, previous work only considered the scalar case ($\mathcal{C} = \mathbb{R}$), and we are the first to provide the extension to the vector-valued case ($\mathcal{C} = \mathbb{R}^m$).

In the standard case, the theory of distributional DP for policy evaluation has been known prior to this work, as well as distributional value and policy iteration for the standard RL objective (Bellemare et al., 2023). Marthe et al. (2024) posed the return distribution optimization without stock augmentation and, having demonstrated that only expected utilities could be optimized, introduced distributional value iteration for optimizing expected utilities. As they show, only affine utilities (U_f with $f(x) = ax + b$ for $a, b \in \mathbb{R}$) and exponential utilities (U_f with $f(x) = ae^{\lambda x} + b$ for $a, b, \lambda \in \mathbb{R}$) can be optimized without stock augmentation (in the finite-horizon undiscounted setting; Marthe et al., 2024).

In stock-augmented problems, classic and distributional DP have been considered primarily in the context of optimizing risk measures. Bäuerle and Ott (2011) introduced a value iteration procedure that maintains $U_f\eta_n$ (with $f(x) = x_-$) as iterates, so it is not distributional. Bäuerle and Rieder (2014); Bäuerle and Glauner (2021) employed the methodology with an augmentation other than stock, for optimizing expected utilities U_f with continuous and increasing f in the former work, and increasing and convex f in the latter.[10] Parallel to the development of this work, Moghimi and Ku (2025) introduced a related policy iteration method that can optimize expected utilities where f has the form $f(x) = \mathbb{E}(x - Z)_-$ and Z satisfies certain conditions (see Equation 6 in Moghimi and Ku, 2025). While they built their analysis on the work introduced by Bäuerle and Ott (2011), the iterates used by their method are return distributions, so it is fair to say that their method is stock-augmented distributional policy iteration.

The distributional Q-learning method introduced by Lim and Malik (2022) for optimizing expected utilities U_f with $f(x) = x_-$ can be associated with a partially stock-augmented DP. The method tracks the stock throughout each episode and uses it during action selection, however it does not employ stock-augmented states for the return distribution functions. In other words, their method adopts a hybrid greedy selection that we can write as $\sup_{\pi \in \Pi} T_\pi \eta$, but with $\eta : \mathcal{S} \to \mathcal{D}$ rather than $\eta : \mathcal{S} \times \mathcal{C} \to \mathcal{D}$.

In terms of analysis, ours is distinct from Bäuerle and Rieder (2014). Instead, we use results and proofs from classic-DP theory (Bertsekas and Tsitsiklis, 1996; Szepesvári, 2022) as a roadmap, incorporate techniques from distributional policy evaluation (Bellemare et al., 2017) to cope with return distributions, and employ novel results required to cope, additionally, with stock augmentation and statistical functionals of the return distribution.

5. Applications

5.1 Generating Desired Returns

In many cases, we want to instruct agents to perform tasks in highly controllable environments, but not necessarily the tasks with a "do something as much as possible" nature that are a clear fit for RL. For example, we may want to specify the task of collecting a given number of objects in a room, or obtaining a score equal to two in the game of Pong in the Atari Benchmark (Bellemare et al., 2013). The standard RL framework can be unwieldy for this type of task, but this type of task can be easily modeled as a stock-augmented problem.

If we were to model this an RL problem without stock augmentation, we would likely have to use a non-Markov reward that tracks how many apples have been collected, give a reward of 1 to the agent when the third apple is collected, and zero otherwise. Moreover,

[10] Rather than the stock C_t, their augmentation is the pair $\left(\sum_{i=0}^{t-1} \gamma^i R_{i+1}, \gamma^t\right)$. While this setting is the same for the finite-horizon undiscounted case, the settings are different in the infinite-horizon discounted case. Notably, the approach of Bäuerle and Glauner (2021) can optimize increasing and convex objectives, which need not be Lipschitz. We hypothesize that the different augmentation allows removing the requirement for indifference to γ.

we would have one reward function for each number of apples to be collected, which might require training one agent per reward function (which seems wasteful).

With stock augmentation, on the other hand, this type of task can be tackled effectively. We can frame it as a stock-augmented return distribution optimization problem with an expected utility U_f and $f(x) = -|x|$, where the stock is the number of apples collected so far by the agent. Moreover, we can get *a single stock-augmented agent* to perform various instances of the same task—for example, collect one apple, or collect three apples—simply by changing the agent's initial stock: Without discounting and with a reward of 1 for each apple, a stock of -3 will cause an optimal stock-augmented agent to collect 3 apples, a stock of -2 will cause the agent to collect 2 apples, and so forth.

5.2 Maximizing the Conditional Value-at-Risk of Returns

The problem of maximizing *conditional value-at-risk* (CVaR; Rockafellar et al., 2000), also known as *average value-at-risk* or *expected shortfall*, has received attention both in the context of risk-sensitive RL (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Chow et al., 2015; Bäuerle and Glauner, 2021; Greenberg et al., 2022) and in non-sequential decision-making (Rockafellar et al., 2000).

It was for this problem that stock-augmented methods were originally developed and studied (see Section 4.5 and Bäuerle and Ott, 2011; Bäuerle and Rieder, 2014; Bäuerle and Glauner, 2021; Lim and Malik, 2022; Moghimi and Ku, 2025). Other works have also proposed methods for optimizing the CVaR and other risk measures, in approaches that can be seen as alternatives to stock augmentation (Chow and Ghavamzadeh, 2014; Chow et al., 2015; Tamar et al., 2015; Greenberg et al., 2022).

The *τ-CVaR* of returns with distribution $\nu \in (\Delta(\mathbb{R}), \mathrm{w})$ is defined as

$$\mathrm{CVaR}(\nu, \tau) \doteq \frac{1}{\tau} \int_0^\tau \mathrm{QF}_\nu(t) \mathrm{d}t.$$

We can see the τ-CVaR as an "expected return in the worst-case", since it corresponds to the expected return of $X \sim \nu$ in the lower-tail of the return distribution (where the tail has mass τ).

For any starting augmented state (s_0, c_0), a history-based policy $\pi \in \Pi_\mathrm{H}$ generates returns distributed according to $\eta^\pi(s_0, c_0)$, and we want to find a policy π and a c_0 to maximize the τ-CVaR of these returns:

$$\sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau).$$

It is easy to see that this problem does not admit an optimal stationary Markov policy on states alone, however Bäuerle and Ott (2011) showed that we can solve it as follows (see Appendix D for the proof):

Theorem 16 (Adapted from Bäuerle and Ott, 2011) *For every $\tau \in (0, 1)$ and $s_0 \in \mathcal{S}$,*

$$\sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) = -c_0^* + \frac{1}{\tau} \sup_{\pi \in \Pi_\mathrm{H}} \mathbb{E}(c_0^* + G^\pi(s_0, c_0^*))_-,$$

where c_0^* is the solution of

$$\max_{c_0}\left(-c_0+\frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}}\mathbb{E}(c_0+G^{\pi}(s_0,c_0))_-\right). \tag{18}$$

The main algorithmic difference between our work and that of Bäuerle and Ott (2011) is how to obtain π^*.[11] While we propose to use distributional DP with $F_K = U_f$ and $f(x) = x_-$, Bäuerle and Ott (2011) used a modified classic value iteration, but required the iterates to satisfy specific conditions (see $\mathbb{M}$, p. 45, Bäuerle and Ott, 2011). With distributional DP, on the other hand, it is possible to establish approximate guarantees for τ-CVaR optimization, for both distributional value/policy iteration, with minimal conditions on the starting iterates (return distribution iterates must have uniformly bounded first moment). This is what the following result shows, if we combine distributional DP with a grid search procedure to approximately solve the optimization in Equation 18:

Theorem 17 *For every $\tau\in(0,1)$, $s_0\in\mathcal{S}$ and $\varepsilon>0$, there exists a stationary policy $\overline{\pi}\in\Pi$ (obtainable through distributional DP) and a $\overline{c}_0^*$ (obtainable through grid search) such that*

$$\sup_{\pi\in\Pi_{\mathrm{H}},c_0\in\mathcal{C}}\mathrm{CVaR}(\eta^{\pi}(s_0,c_0),\tau)-\mathrm{CVaR}(\eta^{\overline{\pi}}(s_0,\overline{c}_0^*),\tau)\le 4\varepsilon.$$

In particular, $\overline{\pi}$ satisfies (for $f(x)=x_-$)

$$\sup_{\pi\in\Pi_{\mathrm{H}}}U_f\eta^{\pi}-U_f\eta^{\overline{\pi}}\le\varepsilon,$$

and

$$\overline{c}_0^*=\arg\max_{c_0\in\overline{\mathcal{C}}}\left(-c_0+\frac{1}{\tau}\mathbb{E}(c_0+G^{\overline{\pi}}(s_0,c_0))_-\right), \tag{19}$$

where $\overline{\mathcal{C}}\doteq\{c_{\min}+i\varepsilon : i\in\mathbb{N}_0, c_{\min}+i\varepsilon\le c_{\max}\}$ and $c_{\min}$ and $c_{\max}$ are chosen so that

$$\begin{aligned}&\max_{c_0}\left(-c_0+\frac{1}{\tau}\mathbb{E}(c_0+G^{\overline{\pi}}(s_0,c_0))_-\right)\\&=\max_{c_{\min}\le c_0\le c_{\max}}\left(-c_0+\frac{1}{\tau}\mathbb{E}(c_0+G^{\overline{\pi}}(s_0,c_0))_-\right).\end{aligned}$$

The key insight in Theorem 17 is that the objective functional being maximized over c_0 in Equation 18 is 1-Lipschitz, so we can approximate it through a grid search with an approximately optimal return distribution (Equation 19). A remaining limitation of the approach is how to choose $c_{\min}, c_{\max}$ in practice. We know from Theorems 16 and 17 that we can choose $c_{\min}$ small enough and $c_{\max}$ large enough to satisfy the requirement, but how large/small they need to be is left to a case-by-case basis.

[11] The differences in analysis are discussed in Section 4.5.

5.3 Maximizing the Optimistic Conditional Value-at-Risk of Returns

The τ-CVaR is the expectation of the return over the lower tail of the distribution (with tail mass τ), and maximizing it is a risk-averse approach. With $\tau = 0$, the τ-CVaR is the risk-neutral expected return, and as τ decreases the amount of risk-aversion increases.

We can also consider the problem of maximizing the upper tail of the return distribution, which we call the *optimistic τ-CVaR*, defined for returns with distribution $\nu \in (\Delta(\mathbb{R}), \mathrm{w})$ as

$$\mathrm{OCVaR}(\nu, \tau) \doteq \frac{1}{\tau} \int_{1-\tau}^{1} \mathrm{QF}_\nu(t) \mathrm{d}t.$$

This application is interesting to analyze because it is similar to the optimism used by Fawzi et al. (2022) in AlphaTensor. More generally, risk-seeking behavior can be useful for "scientific discovery" problems like discovering matrix multiplication algorithms, where it is more helpful to attain exceptional outcomes some of the time, even at the expense of performance in most cases, than to perform well on average. This is because in this type of problem the RL agent is being used to generate solutions to a search-like problem where exceptional solutions are very valuable, but low-quality solutions are harmless, as they can simply be discarded.

We can show that analogues of Theorems 16 and 17 hold for optimizing the optimistic τ-CVaR.

Theorem 18 *For every $\tau \in (0, 1)$ and $s_0 \in \mathcal{S}$,*

$$\sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^\pi(s_0, c_0), \tau) = -c_0^* + \frac{1}{\tau} \sup_{\pi \in \Pi_\mathrm{H}} \mathbb{E}(c_0^* + G^\pi(s_0, c_0^*))_+,$$

where c_0^* is the solution of

$$\min_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_\mathrm{H}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_+ \right).$$

The proof of Theorem 18 is more subtle than the proof of its risk-averse counterpart. In Theorem 16, we can exploit the equivalence

$$\sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) = \sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_+ \right).$$

The similar step in the case of the optimistic τ-CVaR gives

$$\sup_{\pi \in \Pi_\mathrm{H}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) = \sup_{\pi \in \Pi_\mathrm{H}} \inf_{c_0 \in \mathcal{C}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_+ \right).$$

Thanks to distributional DP, we can optimize U_f with $f(x) = x_+$ uniformly for all (s_0, c_0), and we use this to swap the supremum and the infimum above, which gives Theorem 18.

The approximate version of Theorem 19 then follows analogously to Theorem 17.

Theorem 19 *For every $\tau \in (0,1)$, $s_0 \in \mathcal{S}$ and $\varepsilon > 0$, there exists a stationary policy $\overline{\pi} \in \Pi$ (obtainable through distributional DP) and a $\overline{c}_0^*$ (obtainable through grid search) such that*

$$\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) - \mathrm{OCVaR}(\eta^{\overline{\pi}}(s_0, \overline{c}_0^*), \tau) \leq 4\varepsilon.$$

In particular, $\overline{\pi}$ satisfies (for $f(x) = x_+$)

$$\sup_{\pi \in \Pi_{\mathrm{H}}} U_f \eta^{\pi} - U_f \eta^{\overline{\pi}} \leq \varepsilon,$$

and

$$\overline{c}_0^* = \arg\min_{c_0 \in \overline{\mathcal{C}}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_+ \right),$$

where $\overline{\mathcal{C}} \doteq \{c_{\min} + i\varepsilon : i \in \mathbb{N}_0, c_{\min} + i\varepsilon \leq c_{\max}\}$ and $c_{\min}$ and $c_{\max}$ are chosen so that

$$\begin{aligned} &\min_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_+ \right) \\ &= \min_{c_{\min} \leq c_0 \leq c_{\max}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_+ \right). \end{aligned}$$

5.4 Homeostatic Regulation

Homeostatic regulation is a computational model for the behavior of natural agents (Keramati and Gutkin, 2011) whereby they aim to reduce *drive* (Hull, 1943), the mismatch between their current internal state and a stable state. Drive reduction aims to explain empirical observations about the behavior of natural agents (Hull, 1943)—a simplistic instance being the hypothesis that an animal feeds to reduce its hunger.

We can formalize the homeostatic regulation problem considered by Keramati and Gutkin (2011) as:

$$\sup_{\pi \in \Pi_{\mathrm{H}}} -\mathbb{E}\|c_0 + G^{\pi}(s_0, c_0)\|_p^q,$$

where $p, q \geq 1$, $\mathcal{C} = \mathbb{R}^m$, $-c_0$ is the "ideal" setpoint for the agent's internal state, and the agent's stock C_t represents its drive (the deviation from the desired state to be reduced).

"Minimizing drive in norm" above corresponds to the expected utility U_f with $f(x) = -\|x\|_p^q$. This choice of f is positively homogeneous (since $f(\gamma x) = \gamma^{\frac{q}{p}} f(x)$), but Lipschitz only when $q = 1$, so by Lemma 12 and Theorems 6 and 8 distributional DP can solve this variant of homeostatic regulation in the finite-horizon case (regardless of q) and in the infinite-horizon discounted case if $q = 1$ and if we consider the variant where the agent's drive increases over time due to the reverse-discounting, as $C_{t+1} = \gamma^{-1}(C_t + R_{t+1})$.

The formulation where f is a norm presumes that there is an ideal setpoint (namely, $-c_0$), and that the agent wants to keep its stock as close to that as possible, that is, the agent wants its drive (positive or negative) to be as close to zero as possible. This is different from minimizing positive drive—intuitively, a sated agent would not actively drive itself back to the threshold of being hungry.

To accommodate for minimizing only positive drive, we can consider a homeostatic regulation problem with an expected utility, but a different choice of f:

$$f(x) = \sum_{i=1}^{m} \alpha_i \cdot (x_i)_-,$$

where $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ are fixed weights. Once again, this choice of f is positively homogeneous (since $f(\gamma x) = \gamma f(x)$) and Lipschitz (since $f(x) \leq \max_i |\alpha_i| \cdot \|x\|_1$), so by Lemma 12 and Theorems 6 and 8 distributional DP can also solve this variant of homeostatic regulation both in the finite-horizon case and in the infinite-horizon discounted case.

These two reductions are examples of how we can use the framework of stock-augmented return distribution optimization to provide simple solution methods for a problem that has been otherwise complicated to solve with RL. Previously, solving homeostatic regulation with RL methods required the design of an appropriate reward signal (as done by Keramati and Gutkin, 2011). Considering that Keramati and Gutkin (2011) aimed to reconcile the differences between the drive reduction model and the RL-based computational model proposed by Schultz et al. (1997), perhaps the framework of stock-augmented return distribution optimization will help bring the two models closer together.

The reward signal designed by Keramati and Gutkin (2011) to reduce homeostatic regulation to RL corresponds precisely to the reward signal that we have identified as the way to reduce stock-augmented return distribution optimization to stock-augmented RL (see Theorem 20).

5.5 Constraint Satisfaction

In this application, we want an agent to generate returns that satisfy various constraints, with probability one if they are feasible. Our proposal is to model constraint satisfaction as minimizing constraint violations in expectation, which is a variation of minimizing only positive drive discussed in Section 5.4 and generating exact returns from Section 5.1. Constraint satisfaction is related to satisficing problems (Simon, 1956; Goodrich and Quigley, 2004), though satisficing proposes to use constraint satisfaction as a means to find acceptable suboptimal policies when finding optimal policies is inviable.

If we want a policy with return above a threshold g, we can implement the constraint satisfaction as a stock-augmented return distribution optimization problem with U_f, $f(x) = x_-$ and set $c_0 = -g$. This choice of f satisfies Equation 15 (the condition for U_f to be indifferent to γ), so distributional DP can optimize U_f. Maximizing the expected utility will correspond to minimizing the expected violation:

$$\mathbb{E}(c_0 + G^\pi(s_0, c_0))_- = -\mathbb{E}(g - G^\pi(s_0, -g))_+.$$

For any π, we have $G^\pi(s_0, -g) \geq g$ with probability one iff $\mathbb{E}(g - G^\pi(s_0, -g))_+ = 0$. So if the constraint can be satisfied, optimizing U_f will suffice. If we want a policy with return below a threshold g, we optimize U_f with $f(x) = -(x_+)$ and set $c_0 = g$, and for any π, we have $G^\pi(s_0, -g) \leq g$ with probability one iff $\mathbb{E}(G^\pi(s_0, -g) - g)_+$ is zero. For an equality constraint, we can use $f(x) = -|x|$ as in Section 5.1.

Distributional DP can also optimize any weighted combination of the constraints above, with a different stock and reward vector coordinate per constraint, since the weighted combination will also satisfy Equation 15. For example, to generate a return in the interval $[g_1, g_2]$, assume the return is replicated, so that $G_1 = G_2$, set $c_0 = (-g_1, -g_2)$ and optimize U_f with

$$f(x) = (x_1)_- - (x_2)_+.$$

Then for any π, we have $G^\pi(s_0, (-g_1, -g_2)) \in [g_1, g_2]$ with probability one iff

$$\mathbb{E}\,(G^\pi(s_0, (-g_1, -g_2))_1 - g_1)_- - \mathbb{E}\,(G^\pi(s_0, (-g_1, -g_2))_2 - g_2)_+ = 0.$$

Finally, we can also trade off minimizing constraint violations and minimizing or maximizing expected return. An example of this kind of problem is when we want an agent achieve a certain goal "as fast as possible" (Section 3.2, Sutton and Barto, 2018). Traditionally, this kind of goal is normally implemented in episodic settings by terminating the episode when the goal is achieved, with a constant negative reward at each step, or in discounted settings with a reward of 1 when the goal is achieved, and zero otherwise. This is manageable when the goal is achieved instantaneously,[12] but otherwise specifying a reward can be tricky. Return distribution optimization with vector-valued rewards allows for an alternative formulation of this problem with U_f and

$$f(x) = -x_1 + \sum_{i=2}^{m} \alpha_i \cdot (x_i)_-,$$

where the first coordinate of the reward vector is always -1 (representing the time penalty), and the remaining $\alpha_i \cdot (x_i)_-$ regularize the agent's behavior to achieve the multiple goals. It is easy to see that this choice of f is Lipschitz and satisfies Equation 15, so by Lemma 12 and Theorems 6 and 8 distributional DP can solve this problem both in the finite-horizon case and in the infinite-horizon discounted case. We will explore this application in an empirical setting in Section 7.4.

5.6 Generalized Policy Evaluation

One interesting aspect of stock-augmented return distribution optimization is that policy evaluation is not bound to any particular objective functional: If we know the return distribution for a policy π, we can evaluate it under various different choices of F_K, which means the setting is amenable to Generalized Policy Evaluation (GPE; Barreto et al., 2020). In the standard RL setting, GPE is "the computation of the value function of a policy π on a set of tasks" (Barreto et al., 2020). Its natural adaptation to our setting can be stated as the evaluation of a policy under multiple objective functionals $F_{K_1}, \ldots, F_{K_n}$, each corresponding to a different task. This adaptation can be used without stock, with the caveat that removing stock augmentation limits the objectives that distributional DP can optimize (cf. Section 4.5 and Appendix I).

We can also adapt Generalized Policy Improvement (GPI; Barreto et al., 2020) in a similar way: Given policies $\pi_1, \ldots, \pi_n$ and an objective functional F_K, the following is an improved policy using GPI:

$$\overline{\pi}(s, c) \doteq \operatorname*{arg\,max}_{\pi \in \{\pi_1, \ldots, \pi_{n'}\}} (F_K \eta^\pi)(s, c).$$

[12] Admittedly neither a sparse reward nor a constant reward of -1 may be easy for deep RL agents to optimize in practical settings.

The individual policies $\pi_1, \ldots, \pi_n$ may have been obtained by optimizing different objective functionals $F_{K_1}, \ldots, F_{K_n}$, and they can be combined into a policy $\overline{\pi}$ for a new objective functional F_K. Thanks to distributional policy improvement (Lemma 15), we know that $\overline{\pi}$ is, fact, at least as good for F_K as any of the individual policies $\pi_1, \ldots, \pi_n$.

5.7 Reward Design

In deploying RL algorithms on real-world sequential decision-making problems, it is often required to explicitly design a reward signal to codify the intended outcomes. As the reward hypothesis states (Section 3.2, Sutton and Barto, 2018): "All of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)." This hypothesis has been explored and disproved for some interpretations of what constitutes a "goal" (Pitis, 2019; Abel et al., 2021; Shakerinava and Ravanbakhsh, 2022; Bowling et al., 2023). However, even when the hypothesis holds, the reward signal is not necessarily the simplest tool for expressing goals and purposes.

Designing rewards is notoriously difficult. For instance, Knox et al. (2023) present a systematic examination of the perils of designing effective rewards for autonomous driving. They found that, among publicly available reward functions for autonomous driving, "the most risk-averse reward function [. . .] would approve driving by a policy that crashes 2000 times as often as our estimate of drunk 16–17 year old US drivers" (p. 7). Earlier work by Hadfield-Menell et al. (2017) reveals the difficulty of hand-designing rewards, with common failures including unintentional positive reward cycles.

We contend that, in some cases, the framework of stock-augmented return distribution optimization eliminates the need for bespoke reward design. To support this claim, we extend a reward-design result by Bowling et al. (2023) to the stock-augmented setting, showing, once the objective functional has been chose, how to define an RL reward signal so that the RL objective is equivalent to the stock-augmented return distribution optimization objective. The result also shows that this reduction between objectives is only possible if the statistical functional is an expected utility and indifferent to γ.

Theorem 20 *A stock-augmented return distribution optimization objective functional U_f can be reduced to an equivalent stock-augmented reinforcement learning objective (expected return) with discount $\alpha \in (0, 1]$ with $\gamma < 1 \Rightarrow \alpha < 1$ and reward proportional to*

$$\widetilde{R}_{t+1} \doteq \alpha f(C_{t+1}) - f(C_t) + (1 - \alpha) f(0) \tag{20}$$

if f satisfies, for all $c \in \mathcal{C}$,

$$f(\gamma c) = \alpha f(c) + (1 - \alpha) f(0), \tag{21}$$

and:

- *in the finite-horizon case,*

$$\sup_{s,c,a \in \mathcal{S} \times \mathcal{C} \times \mathcal{A}} \mathbb{E}\left(|\widetilde{R}_{t+1}| \,\middle|\, S_t = s, C_t = c, A_t = a \right) < \infty; \tag{22}$$

- *in the discounted case, f is Lipschitz.*

A stock-augmented return distribution optimization objective that is not an expected utility or not indifferent to γ cannot be reduced via reward design to a stock-augmented reinforcement learning objective.

The reward construction used in Theorem 20 may seem obvious in hindsight, but we believe that it can be much less evident if the corresponding U_f has not been identified, and that this may account for some of the challenges in designing rewards straight from imprecise "goals and purposes". However, once U_f has been identified, the construction essentially automates away one step in the design of RL agents. For example, the construction used in Theorem 20 can be seen to be the same as the one used by Keramati and Gutkin (2011) to reduce homeostatic regulation to an RL problem, and Theorem 20 provides this reduction immediately.

Theorem 20 allows us to optimize certain stock-augmented return distribution optimization objectives with classic DP. In the discounted case, these are the same objectives we have shown that can be solved with distributional DP. In the finite-horizon undiscounted case, there are two main differences. First, distributional DP can optimize (arguably pathological) objectives where Assumption 1 is satisfied, but not Equation 22.[13] Second, and more importantly, distributional DP can optimize certain objective functionals that are not expected utilities, whereas classic DP, at least via reward design, cannot.

5.8 Beyond Expected Utilities

In all the applications we have presented so far, the objective functionals being optimized by distributional DP were expected utilities. While expected utilities cover many common use cases of stock-augmented return distribution optimization, it is worth considering which non-expected utilities distributional DP can optimize. Without stock augmentation, distributional DP cannot optimize non-expected utilities, even in the finite-horizon undiscounted case (Marthe et al., 2024), which is the most permissive as far as conditions for optimizing F_K go. We also saw in Theorem 20 that, at least through reward design, classic DP cannot optimize non-expected utilities, even with stock augmentation. What about distributional DP with stock augmentation?

The answers differ depending on whether we consider the infinite-horizon discounted case, or the finite-horizon case. In the infinite-horizon discounted case, the following theorem states that only Lipschitz expected utilities satisfy indifference to mixtures and Lipschitz continuity, which are required in our distributional DP guarantees.

Theorem 21 *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and Lipschitz, then F_K is an expected utility, that is, there exists an $f : \mathcal{C} \to \mathbb{R}$ such that $K\nu = \mathbb{E}f(G)$ $(G \sim \nu)$ and f is Lipschitz.*

Theorem 21 does not necessarily rule out distributional DP optimizing non-expected utilities in the infinite-horizon discounted case, because it is still an open question whether

[13] For example, consider optimizing U_f with $f(c) = c^2$ in the finite-horizon case with $\gamma = 1$ (f need not be Lipschitz in this case), where rewards for each state-action are Pareto$(2, 1)$ random variables multiplied by ±1. These pseudo-rewards have bounded first moment, but unbounded second moment, so $\widetilde{R}_{t+1}$ may have unbounded first moment, and Bellman equations may be invalid.

Lipschitz continuity is necessary. However, it does rule out Lipschitz functionals that are not expected utilities, including, for example, the τ-CVaR:

$$K\nu = \frac{1}{\tau}\int_0^\tau \mathrm{QF}_\nu \mathrm{d}t. \tag{23}$$

This choice of K is Lipschitz, but F_K is not an expected utility.[14] This may seem to contradict the claims in Section 5.2, but it does not. Theorem 36 shows that distributional DP can optimize the τ-CVaR by transforming the problem into the optimization of an expected utility, and specifying how to select c_0. The objective that distributional DP cannot optimize is F_K with K set to be exactly the τ-CVaR functional (as in Equation 23). To emphasize the difference between the two cases, compare which K is used in the greedy policies of Theorems 6 and 8.

As another example of non-expected utilities with Lipschitz K, consider minimizing the 1-Wasserstein distance to a reference distribution $\overline{\nu}$ in the scalar case ($\mathcal{C} = \mathbb{R}$), that is, $K\nu = -\mathrm{w}(\nu, \overline{\nu})$. This K is Lipschitz (by the triangle inequality), however F_K is not an expected utility unless $\overline{\nu}$ is a Dirac. By Theorem 21, distributional DP cannot optimize this objective functional if $\overline{\nu}$ is not a Dirac. We can verify that the K is not indifferent to mixtures, for example, when $\overline{\nu}$ is the distribution of a Bernoulli-$\frac{1}{2}$ random variable (in this case, $K\delta_0 = K\delta_1$, so indifference to mixtures requires that $K\left(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_0\right)$ equal $K\left(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1\right)$, which is not the case). When $\overline{\nu} = \delta_c$ for some $c \in \mathbb{R}$, it is easy to see that $K\nu = -\mathbb{E}|G - c|$ ($G \sim \nu$), and we have already established that K is indifferent to $\gamma < 1$ iff $c = 0$.

Turning to the finite-horizon case, can we claim that distributional DP cannot optimize non-expected utilities? A positive answer here would imply that distributional and classic DP are essentially equivalent in the finite-horizon *undiscounted* case, with stock augmentation as well as without.[15]

As the next result shows, it *is* possible for distributional DP to optimize non-expected utilities in the finite-horizon case. The choice of functional in Proposition 22 can be phrased as "any negative return is (equally) unacceptable," and is known not to be an expected utility (Juan Carreño, 2020; Bowling et al., 2023).

Proposition 22 *The statistical functional $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ satisfying, for $\nu \in (\mathcal{D}, \mathrm{w})$,*

$$K\nu = \mathbb{I}(\nu([0,\infty)) = 1)$$

is indifferent to mixtures and F_K is not an expected utility.

The choice of K in Proposition 22 does not allow for a reduction to a stock-augmented RL objective via reward design (cf. Theorem 20), because it is not an expected utility. However, since it is indifferent to mixtures, distributional DP can optimize the corresponding F_K in the finite-horizon undiscounted case.

[14] K violates the von-Neumann-Morgenstern (Von Neumann and Morgenstern, 2007) axiom of independence. See Axiom 3 in Appendix F with ν uniform in $\{0\}$, ν' uniform in $\{-1, 2\}$, $\overline{\nu}$ uniform in $\{2\}$ and $\tau, p = \frac{1}{2}$.

[15] Save for the extreme case where the pseudo-rewards in the stock-augmented return distribution optimization have bounded first moment, but not the designed stock-augmented RL rewards, as discussed in the context of Theorem 20.

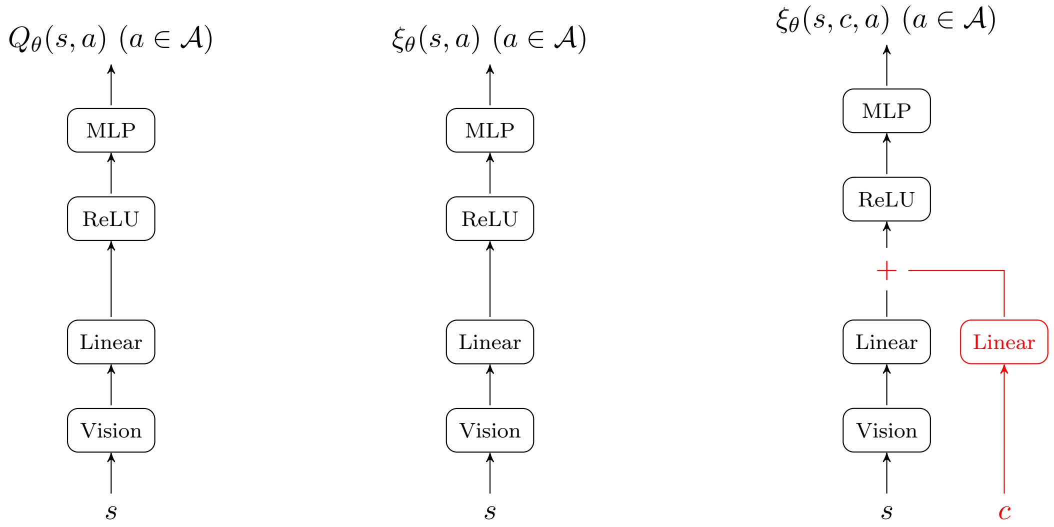

$Q_\theta(s,a)$ $(a \in \mathcal{A})$
MLP
ReLU
Linear
Vision
s
$\xi_\theta(s,a)$ $(a \in \mathcal{A})$
MLP
ReLU
Linear
Vision
s
$\xi_\theta(s,c,a)$ $(a \in \mathcal{A})$
MLP
ReLU
+
Linear
Linear
Vision
s
c

Figure 1: Architecture diagrams for DQN (left), QR-DQN (center) and DηN (right). In red, the elements introduced specifically for DηN. The QR-DQN and DηN networks output return distribution quantiles for each input (s or (s,c)) and action.

6. DηN

To highlight the practical potential of distributional DP for solving return distribution optimization problems, we adapted QR-DQN (DQN with quantile regression; Dabney et al., 2018) to optimize expected utilities U_f and evaluated it empirically. We call this new method *Deep η-Networks*, or *DηN* (pronounced *din*). In this section introduce DηN and describe how it incorporates the principles of distributional DP. We present the empirical study in Sections 7 and 8.

DηN uses a neural-network estimator for the stock-augmented return distribution, similar to QR-DQN, with one difference: The stock embedding. In DηN, we input the stock to a linear layer[16] and then add the output of this linear layer to output to the of the agent's vision network.[17] The architecture diagrams for DQN (Mnih et al., 2015), QR-DQN (Dabney et al., 2018)) and DηN are given in Figure 1. The output $\xi_\theta(s,c,a)$ of the network is a return distribution parameterized as quantiles (see Appendix H.1 for implementation details).

To explain the remaining differences between QR-DQN and DηN, it is useful to understand how QR-DQN is adapted from classic DP, and then see how distributional DP is adapted into DηN. This adaptation is necessary because DP is designed for a *planning setting* (where the transition and reward dynamics of the MDP are known), but planning methods are rarely tractable or feasible in practice (where state spaces can be very large and the dynamics

[16] While this simple design decision proved sufficient for our experiments, we believe that improved scalar embedding should be considered in the future (for example, Springenberg et al., 2024).

[17] In practice, the MDP state s is converted to an image observation before being input to the vision network, and the conversion is domain-dependent.

can only be observed through interaction with the environment). Practical settings are more closely modeled as prediction and control settings (Sutton and Barto, 2018) with a function approximator learned through deep learning, that is, the typical setting for deep reinforcement learning.

Given a (state-) value function $V \in (\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$, the corresponding *action-value function* is defined as

$$Q(s,a) = (T_{\pi_a} V)(s),$$

where π_a denotes the policy that selects action a with probability one at all states. It is convenient to denote this transformation with an operator, commonly known as the classic *Bellman lookahead* (p. 30, Szepesvári, 2022):

$$(AV)(s,a) \doteq (T_{\pi_a} V)(s).$$

We also let $M : (\mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \|\cdot\|_\infty) \to (\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$ be the *max operator on action-value functions* defined as

$$(MQ)(s) \doteq \max_a Q(s,a) = \sup_{p\in\Delta(\mathcal{A})} \mathbb{E}Q(s,A). \qquad (A \sim p)$$

Each iterate V_n in classic value iteration has a corresponding $Q_n \doteq AV_n$, and it holds that $V_{n+1} = MQ_n$. Thus, we can equivalently carry out value iteration on action-value functions, via the relation

$$Q_{n+1} = AMQ_n. \tag{24}$$

Q-learning (Watkins, 1989; Sutton and Barto, 2018) aims to approximate value iteration through multiple asynchronous stochastic updates per transition. Given a *transition* $(s_t, a_t, r_{t+1}, s_{t+1})$, the Q-learning update is:

$$Q_\theta(s_t,a_t) \leftarrow (1-\alpha)Q_\theta(s_t,a_t) + \alpha\cdot\left(r_{t+1} + \gamma(MQ_\theta)(s_{t+1})\right), \tag{25}$$

where Q_θ is the action-value function estimator being learned and α is a learning rate. Note how the term in parentheses resembles the right-hand side of Equation 24. Roughly speaking, it serves as an estimate of AMQ_θ on the given transition.[18]

DQN (Mnih et al., 2015) implements the Q-learning update with a deep neural network estimator for Q_θ, and in addition, an estimator $Q_{\overline{\theta}}$ with *target parameters* $\overline{\theta}$ on the right-hand side of Equation 25. The target parameters slowly track θ, and the DQN value update only modifies θ. The updates to θ are performed through regression, similar to fitted Q-iteration (Ernst et al., 2005) with a Huber loss, and with the *prediction targets*

$$r_{t+1} + \gamma(MQ_{\overline{\theta}})(s_{t+1}),$$

which, as before, are meant to serve as an estimate of $AMQ_{\overline{\theta}}$ on the given transition.

The implementation of DηN can be thought of as applying the adaptations above to distributional DP with an expected-utility objective U_f. This is a stock-augmented setting, so note the use of the augmented state $(s,c) \in \mathcal{S}\times\mathcal{C}$, in contrast to the use of the plain states

[18] The precise relationship between the two quantities can be understood from the analysis of Q-learning (Dayan and Watkins, 1992).

$s \in \mathcal{S}$ for classic DP, Q-learning, DQN and QR-DQN. The *stock-augmented distributional Bellman lookahead* operator is defined as

$$(A\eta)(s,c,a) \doteq (T_{\pi_a}\eta)(s,c),$$

where, as before, π_a selects a with probability one at all $(s,c) \in \mathcal{S} \times \mathcal{C}$. The distributional analogue of action-value functions are action-dependent return distribution functions. From a return distribution η, the distributional Bellman lookahead gives the corresponding action-dependent return distribution function $\xi = A\eta$.

The analogue of the max operator M for optimizing U_f must take f into account, so we denote it by M_f to highlight this dependence, and we define it so that:

$$U_f(M_f\xi)(s,c) = \sup_{p \in \Delta(\mathcal{A})} \mathbb{E} f(c + G(s,c,A)). \qquad (A \sim p,\, G(s,c,a) \sim \xi(s,c,a))$$

M_f may not be unique because U_f may allow multiple policies to realize the supremum on the right-hand side, but any valid M_f is acceptable. Because the right-hand side above is linear in π, we can write M_f via a simple maximization over actions:

$$U_f(M_f\xi)(s,c) = \max_a \mathbb{E} f(c + G(s,c,a)). \qquad (G(s,c,a) \sim \xi(s,c,a))$$

As in the classic case, we can carry out distributional value iteration on action-dependent return distribution function iterates:

$$\xi_{n+1} = AM_f\xi_n.$$

DηN adapts distributional value iteration similarly to how QR-DQN adapts classic value iteration. QR-DQN replaces DQN's action-value function estimator with a return distribution estimator (see the middle diagram in Figure 1), and employs quantile regression to fit it, rather than ordinary scalar regression with a Huber loss. The return distribution estimator used by DηN is $\xi_\theta : \mathcal{S} \times \mathcal{C} \times \mathcal{A} \to \mathcal{D}$ and the distributional prediction target can be written as

$$\mathrm{df}\left(r_{t+1} + \gamma(M_f\xi_{\overline{\theta}})(s_{t+1}, c_{t+1})\right), \tag{26}$$

and QR-DQN is analogous, but without the stock augmentation. In analogy to DQN, the distributional prediction target in Equation 26 is meant to serve as an estimate of $AM_f\xi_{\overline{\theta}}$ on the observed data.

In QR-DQN, f is the identity function and U_f is the standard RL objective, so

$$\mathbb{E}(M_f\xi_{\overline{\theta}})(s_{t+1}) = \max_a \mathbb{E}\left(G(s_{t+1},a)\right). \qquad (G(s,a) \sim \xi_{\overline{\theta}}(s,a))$$

This is an equation over action-values, and it naturally resembles the action choice used in the Q-learning update and DQN's prediction targets. Similar to how the greedy action for Q-learning and DQN is a maximizing action, DηN's greedy action at (s_t, c_t) maximizes U_f:

$$\mathbb{E} f(c_t + G(s_t,c_t,a_t)) = \max_a \mathbb{E} f(c_t + G(s_t,c_t,a)). \tag{27}$$

with $G(s,c,a) \sim \xi_{\overline{\theta}}(s,c,a)$.

In summary, DηN is similar to QR-DQN in many ways, with two notable differences: The neural network supports stock augmentation (Figure 1), and the stock and the utility factor into the action selection, both for the quantile regression targets (Equation 26) and for the agent's interaction with the environment (Equation 27).

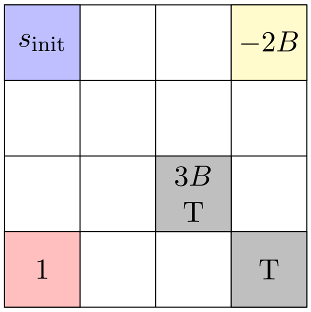


Figure 2: Example gridworld (with cells indexed as matrix entries). The starting cell s_{init} is the upper-left corner cell $(1,1)$. The bottom-left corner (red, $(4,1)$) has a deterministic reward of 1. The upper-right corner (yellow, $(1,4)$) has a stochastic reward $-2B$, where $B \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ (sampled independently each time step). The bottom-right corner (gray, $(4,4)$) is terminal. The cell $(3,3)$ (gray) is terminal and has a stochastic reward of $3B$.

7. Gridworld Experiments

In this section we present experiments to illustrate how DηN solves different toy instances of stock-augmented return distribution optimization, corresponding to some of the applications discussed in Section 5. These experiments are also interesting because they reveal practical challenges of training stock-augmented return distribution optimization agents.

The environments are 4×4 gridworlds (Sutton and Barto, 2018). The agent's actions are up, down, left, right, and no-op. If the agent takes a no-op action or attempts to go outside the grid, it stays in the same cell. The starting cell is always the top-left corner of the grid, which we denote by $s_0 = s_{\text{init}}$, and the starting stock c_0 is set per experiment. For a transition $(s,c), a, r', (s',c')$, if s is terminal, then $c' = c$, $s' = s$ and $r' = 0$. Otherwise, $c' = \gamma^{-1}(c + r')$ (as in Equation 1). Some cells are terminating; if the agent enters a terminating cell, then s' is terminal (and absorbing). Some cells are rewarding: If s is non-terminal and s' is rewarding, then the agent receives r' associated with s'. The reward may be deterministic, or it may be $r' \cdot B$ where $B \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ (independently for each transition). A cell may be both rewarding and terminal, in which case the agent receives the reward for the cell upon entering it, but not afterwards. Figure 2 gives an example gridworld with the notation we use. At an augmented state (s,c), besides the stock c, the input to DηN's vision network (see Figure 1) is a one-channel 4×4 frame with 1 in the cell corresponding to s and zero otherwise.

During training, it was essential to randomize the starting c_0, by sampling values uniformly from a range (implementation details are given in Appendix H.2). This was meant to introduce diversity in the training data and ensure that the agent could solve problems for a variety of c_0.

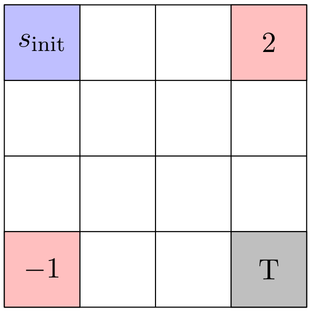


Figure 3: Gridworld for the first experiment for generating returns.

7.1 Generating Desired Returns

Our two first experiments illustrate how DηN with $\mathcal{C} = \mathbb{R}$ and $f(x) = -|x|$ can generate desired outcomes in a deterministic environment (see the application discussed in Section 5.1). In this setting the trained DηN agent displays different behaviors depending on c_0.

We first consider generating specific returns in the gridworld given in Figure 3. Because this gridworld is deterministic, we can set c_0 to different values to generate different desired discounted returns, and the agent must do so by combining the rewards of 2 on the top-right corner and the rewards of -1 on the bottom-left corner.

Because in practice DQN-like agents tend not to cope well with $\gamma = 1$, we set $\gamma = 0.997$ and assessed whether the agent can approximately generate the values of c_0 provided. Table 2 shows the agent's average return for different choices of c_0, with confidence interval bounds in parentheses. In each independent run, we trained the agent and then measured its average discounted return (over 200 episodes) for each of the values of c_0 considered. We then computed 95%-confidence intervals based on the 30 independent averages using bias-corrected and accelerated bootstrap (James et al., 2013; Virtanen et al., 2020). Each row of Table 2 shows the "desired" return ($-c_0$), the average discounted return obtained by the agent ($\mathbb{E}G(s_0, c_0)$) and the "error" $\mathbb{E}|c_0 + G(s_0, c_0)|$, the negative of the objective. We can see that, as intended, the trained DηN agent can approximately produce the desired discounted returns.

The mismatch between $-c_0$ and average discounted returns is likely due to the function approximation and discounting, which makes the exact c_0 challenging to realize for arbitrary c_0. However, the agent should generate returns equal to $-c_0$ when it corresponds to a realizable discounted return. To test this hypothesis, we carried out a follow-up evaluation where, for each trained agent, each choice of c_0, and each evaluation episode generated with discounted return $G(s_0, c_0)$, we ran that agent starting from (s_0, c'_0) with $c'_0 = -G(s_0, c_0)$, and measured the discounted return $G(s_0, c'_0)$ obtained. The observed values for $|c'_0 + G(s_0, c'_0)|$ were less than $3.02 \cdot 10^{-2}$ uniformly for *all runs* (across all independent runs, c_0 and episodes). Thus DηN can closely reproduce realizable discounted returns, and the mismatches in Table 2 are likely related to γ and function approximation.

This first experiment is an illustration of the ability of methods like DηN to control deterministic environments and generate desired outcomes, which is a desirable capability for artificial agents. Besides combining different rewards, another means to control the

Desired discounted return $-c_0$	Discounted return $\mathbb{E}G(s_0, c_0)$	Error $\mathbb{E}\lvert c_0 + G(s_0, c_0)\rvert$
7.00	$6.95\ (6.95, 6.95)$	$0.05\ (0.05, 0.05)$
5.00	$4.98\ (4.98, 4.98)$	$0.02\ (0.02, 0.02)$
3.00	$3.00\ (3.00, 3.00)$	$0.00\ (0.00, 0.00)$
1.00	$1.01\ (1.01, 1.01)$	$0.01\ (0.01, 0.01)$
-2.00	$-1.85\ (-1.99, -1.59)$	$0.15\ (0.01, 0.41)$
-4.00	$-3.96\ (-3.96, -3.96)$	$0.04\ (0.04, 0.04)$
-6.00	$-5.92\ (-5.92, -5.92)$	$0.08\ (0.08, 0.08)$
-8.00	$-7.87\ (-7.87, -7.87)$	$0.13\ (0.13, 0.13)$

Table 2: Evaluation results for DηN optimizing U_f with $f(x) = -|x|$ in the gridworld from Figure 3, and $\gamma = 0.997$. Entries are averages with bootstrap confidence intervals in the format "average (low, high)" where low and high are the interval bounds.

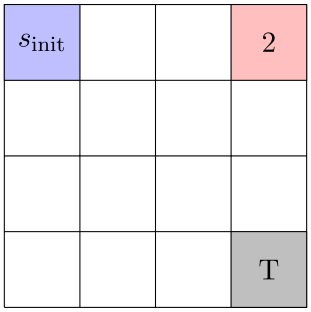


Figure 4: Gridworld for the second experiment.

returns is to use the discounting. Intuitively, in this case, instead of collecting a unit of reward as soon as possible, the agent may choose to "wait" for a few time steps until the discounted reward (from the starting state) achieves the desired value. To illustrate this point, in our second experiment we removed the negative reward from the gridworld in the first experiment, and set $\gamma = \frac{1}{2}$. The gridworld diagram is given in Figure 4.

The results are in Table 3, and the agent successfully generates the desired discounted returns. From an observer's point of view, the perceived behavior of the agent is that it "correctly times" the rewarding transitions; in reality, the agent uses the stock to decide whether or not to collect a reward at a particular augmented state.

7.2 Maximizing the τ-CVaR

We can use DηN to optimize τ-CVaR of the return, the risk-averse RL setup outlined in Section 5.2. The 1-CVaR is risk-neutral (stock-augmented RL), and as τ goes to zero optimizing the τ-CVaR requires more risk aversion. In this setting, DηN displays behaviors with different risk profiles in response to changing τ.

Desired discounted return	Discounted return	Error
$-c_0$	$\mathbb{E}G(s_0, c_0)$	$\mathbb{E}\lvert c_0 + G(s_0, c_0)\rvert$
1.00	$1.00\ (1.00, 1.00)$	$0.00\ (0.00, 0.00)$
0.50	$0.50\ (0.50, 0.50)$	$0.00\ (0.00, 0.00)$
0.25	$0.25\ (0.25, 0.25)$	$0.00\ (0.00, 0.00)$
0.12	$0.12\ (0.12, 0.12)$	$0.00\ (0.00, 0.00)$
0.06	$0.06\ (0.06, 0.06)$	$0.00\ (0.00, 0.00)$

Table 3: Evaluation results for DηN optimizing U_f with $f(x) = -|x|$ in the gridworld from Figure 4 and $\gamma = \frac{1}{2}$. Entries are averages with bootstrap confidence intervals in the format "average (low, high)" where low and high are the interval bounds.

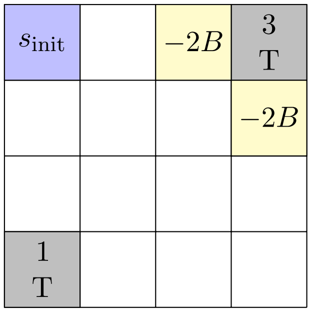


Figure 5: Gridworld for the first risk-averse RL experiment.

The objective functional is U_f with $f(x) = x_-$, but we do not specify c_0 directly. Instead, given a desired τ, we compute c_0^* according to Theorem 16 and start the agent in the augmented state (s_0, c_0^*). The gridworld for this experiment is given in Figure 5. It has a "safe" terminating cell in the bottom-left corner, and a "high-risk" terminating cell in the upper-right corner. This cell has high risk because it is surrounded by cells that give -2 reward with probability $\frac{1}{2}$ (and zero otherwise). With $\gamma = 0.997$ the high-risk cell is better in expectation, so an optimal risk-neutral agent ($\tau = 1$) would go there. However, an optimal risk-averse agent (with respect to the τ-CVaR and for small enough τ) will avoid the high-risk cell and go to the safe cell in the bottom-left corner.

DηN's performance is consistent with these behaviors, as we see in Figure 6, which shows the histograms of the returns obtained by DηN over several runs. As before, we trained the DηN agent in 30 independent training runs. After training the agent in each of the runs, we ran the agent with different values of τ for 200 episodes. It is worth emphasizing that we run the *same* trained agent with different values of τ, as discussed in Section 5.1. We binned the observed returns and computed their frequencies for each independent run, and we report the average frequencies per bin with 95% bootstrap confidence intervals. For smaller τ, the agent goes to the safe terminating cell. As τ increases, the frequency of returns corresponding to the high-risk cell also increases.

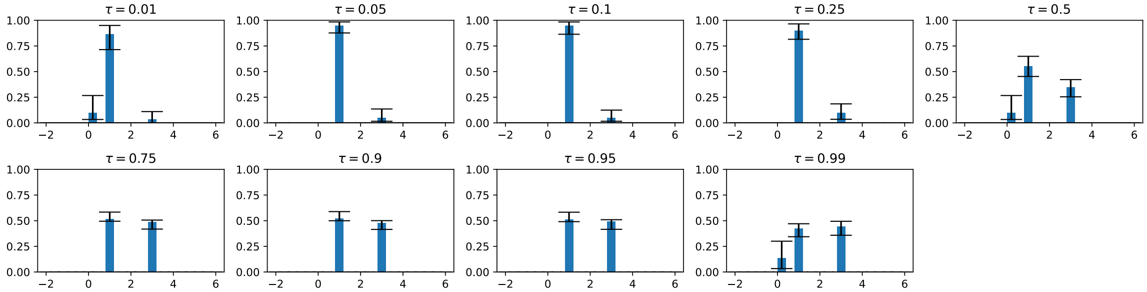


Figure 6: Discounted return histogram for different values of τ, obtained by a trained DηN agent. Error bars correspond to bootstrap confidence intervals.

DηN generated zero returns in some instances, which are suboptimal behaviors regardless of τ. The selection of c_0^* uses grid search and approximate return estimates from $\xi_{\bar{\theta}}$, and estimation errors may cause $\mathbb{E}(c_0^* + \xi_{\bar{\theta}}(s_0, c_0^*, a))_-$ to be zero for all actions, even for the down action. When this is the case, DηN selects actions uniformly at random (because all actions are greedy). The stock, which starts often at a negative value, inflates due to the γ^{-1} factor and becomes more negative. Eventually it is so large in magnitude that the future discounted return can never exceed the stock, and the result is degenerate behavior.

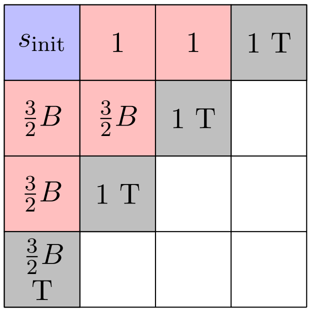
7.3 Maximizing the Optimistic τ-CVaR

Similar to how we can use DηN to produce risk-averse behavior, we can also use it to produce risk-seeking behavior, by following the outline in Section 5.3. In this case we also observe DηN display behaviors with different risk profiles: When the agent is risk-seeking, it tries to maximize its best-case expected performance, and as it becomes more risk neutral its performance resembles that of an RL agent maximizing value.

The objective functional is U_f with $f(x) = x_+$ and as before we do not specify c_0 directly. Instead, given τ, we compute c_0^* according to Theorem 18, and run the agent from (s_0, c_0^*). The optimistic 1-CVaR is risk-neutral, and as τ goes to zero the optimistic τ-CVaR demands more risk-seeking behavior. The gridworld for this experiment is given in Figure 7. The only allowed actions are down and right, and $\gamma = 0.997$. In this environment, the higher the risk, the higher the best-case return, but the lower the expected return. A risk-neutral agent will go right twice and then either right or down, terminating with a discounted return of $1 + \gamma + \gamma^2$. These are the low-risk paths. In any given cell and whatever the stock, moving to a cell with Bernoulli rewards increases the risk relative to choosing a cell with deterministic reward. Going down three times is the path with highest risk, with expected discounted return $\frac{3}{4}(1 + \gamma + \gamma^2)$, but twice that amount with probability $\frac{1}{8}$ (the best case).

DηN's performance is consistent with the risk profile given by τ, as we see in Figure 8, which shows the histograms of the returns obtained by DηN over several runs. We trained the DηN agent and computed histograms in the same way as in Figure 6.

For $\tau \leq 0.1$ we see that the agent is maximally risk-seeking, as the support of the distribution includes the maximum possible return (approximately 4.5) with probability around $\frac{1}{8}$. As τ increases, the agent becomes less risk-seeking, and eventually ($\tau = 0.25$) the



s_{init}	1	1	1 T
$\frac{3}{2}B$	$\frac{3}{2}B$	1 T	
$\frac{3}{2}B$	1 T		
$\frac{3}{2}B$ T			

Figure 7: Gridworld for the risk-seeking RL experiment. The only allowed actions are down and right.

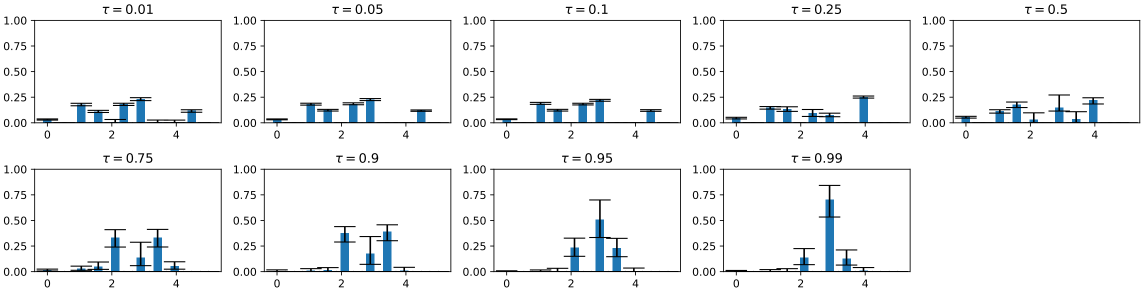


Figure 8: Discounted return histogram for different values of τ, obtained by a trained DηN agent. Error bars correspond to bootstrap confidence intervals.

agent stops going for the riskiest path and visits cells with deterministic rewards more often. At $\tau \approx 1$ the agent is nearly risk-neutral, with a mean discounted return of 2.6 ± 0.485. The optimal risk-neutral expected discounted return is approximately 2.99, and we believe the mismatch is due to approximation errors on the choice of the starting c_0^*.

To highlight the agent's ability to adapt to different stochastic outcomes, notice how the frequency of zero returns is quite low, even for the highly risk-seeking behavior ($\tau = 0.01$). This may seem counter-intuitive if we consider that the highest-risk path has the same probability of a best discounted return (4.48 with probability $\frac{1}{8}$) as of a worst discounted return (zero). Yet DηN with $\tau = 0.01$ observes a discounted return of 4.48 with probability around $\frac{1}{8}$, and worst-case returns with probability around 0.03 ± 0.02. This happens because DηN adapts its behaviors to the observed returns, through stock augmentation. If we look back at Figure 7, we can see that there is always a path such that, if the agent observes a zero reward at one of the non-terminal cells with Bernoulli rewards, it can go right and avoid a return of zero. For example, for a low enough τ, the agent's starting stock will be $c_0^* \leq -4$. If the agent goes down on its first action and observes a reward of zero, it is no longer able to generate a discounted return above 4. Because $f(x) = x_+$, all actions will have expected utility zero (modulo estimation errors), and because DηN breaks ties by uniform

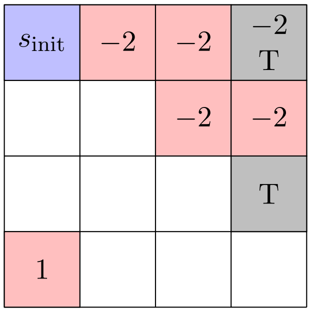


Figure 9: Gridworld for the experiment with trading off minimizing constraint violation and maximizing expected return.

sampling, the agent will follow a uniformly random policy. So the probability of observing a zero discounted return is $\mathbb{P}(R_1 = 0, A_1 = \text{down}, R_2 = 0, A_2 = \text{down}, R_3 = 0)$. Since there are only two actions, this probability is $\left(\frac{1}{2}\right)^5 = 0.03125$, which is consistent with our data.

7.4 Trading Off Minimizing Constraint Violation and Maximizing Expected Return

In this section, we consider the application outlined at the end of Section 5.5: To obtain a certain amount of reward in as few steps as possible. This application requires DηN to optimize an objective functional with vector-valued rewards.

In this setting, we have $\mathcal{C} = \mathbb{R}^2$. The first coordinate of the reward is always -1, and corresponds to the "time-to-termination" penalty to be minimized. The values observed in the second coordinate of the reward vector are given in Figure 9. The objective functional is U_f with $f(x) = -x_1 + \alpha \cdot (x_2)_-$. We set $\alpha = 50$ to encourage prioritizing the term on the second coordinate of the reward vector, so the semantics of the objective functional is to get to termination as fast as possible, keeping $G(s, c)_2 \geq -(c_0)_2$, but allowing for small violations to be traded off for faster termination. For this experiment, we estimate the marginal distributions (per coordinate) of the vector-valued returns. This simplifies the prediction in DηN, and is sufficient for the expected utility being optimized.[19]

An optimal policy with respect to U_f will display different behaviors depending on the choice of $(c_0)_2$. If $-(c_0)_2 \leq -2(\gamma^2 + \gamma^1 + \gamma^0) \approx -5.98$, the policy will go straight from s_{init} to terminate at the top-right corner. This is the shortest possible path to termination, but it is "costly" in terms of the cell rewards. With $-2(\gamma^2 + \gamma^1 + \gamma^0) < -(c_0)_2 \leq 0$, the policy goes to the "lower" terminating cell ($(3, 4)$) in 5 steps and with $G(s_0, c_0)_2 = 0$. For $-(c_0)_2 > 0$, the policy must stay at the cell in the lower-left corner for multiple steps before going to the "lower" terminating cell ($(3, 4)$). The number of steps it stays will depend on α and $-(c_0)_2$: As $\alpha \to \infty$ the policy will stay longer to make $G(s_0, c_0)_2$ closer to $-(c_0)_2$ (either larger or

[19] When $f(x)$ does not decouple as $\sum_i f_i(x_i)$ for some choice of f_i (for example, $f(x) = -\|x\|_2$), the distribution of the quantile vectors is needed. For those cases, one may consider building on results for multivariate distributional RL (Zhang et al., 2021; Wiltzer et al., 2024).

Lower-bound $-(c_0)_2$	Discounted Return $\mathbb{E}G(s_0, c_0)_2$	Penalty term $\mathbb{E}\left((c_0)_2 + G(s_0, c_0)_2\right)_-$	Episode duration
3.00	$3.62\ (3.39, 3.88)$	$-0.05\ (-0.18, -0.01)$	$10.83\ (10.23, 11.70)$
2.00	$2.47\ (2.14, 2.77)$	$-0.14\ (-0.41, -0.04)$	$11.00\ (10.00, 12.20)$
1.00	$1.41\ (1.08, 1.77)$	$-0.20\ (-0.37, -0.10)$	$11.57\ (10.20, 12.97)$
0.00	$0.20\ (0.07, 0.55)$	$0.00\ (0.00, 0.00)$	$5.87\ (5.37, 7.20)$
−1.00	$0.06\ (0.00, 0.39)$	$0.00\ (0.00, 0.00)$	$5.37\ (5.00, 6.83)$
−2.00	$0.03\ (0.00, 0.16)$	$0.00\ (0.00, 0.00)$	$5.37\ (5.00, 6.83)$
−6.00	$-0.40\ (-1.20, 0.00)$	$0.00\ (0.00, 0.00)$	$4.93\ (4.67, 5.00)$
−7.00	$-4.79\ (-5.58, -3.79)$	$0.00\ (0.00, 0.00)$	$3.40\ (3.13, 3.73)$

Table 4: Performance of DηN trading off minimizing constraint violation and maximizing expected return. The weight of the second term is $\alpha = 50$. Entries are averages with bootstrap confidence intervals in the format "average (low, high)" where low and high are the interval bounds.

slightly smaller). For example, it would take the optimal policy at most 8 steps to reach termination with $-(c_0)_2 = 1$, 9 steps with $-(c_0)_2 = 2$ and 10 steps with $-(c_0)_2 = 3$.

The results for DηN are in Table 4. DηN did not produce optimal behaviors, but aligned with them. In the first three settings (upper rows of the table), visiting the bottom-left corner was required by U_f. The agent did that (albeit overstaying) and then went to the lower terminating cell. In the second three settings (middle rows of the table), visiting the bottom-left corner was not required by U_f; the agent went to the lower terminating cell. In the last two settings (bottom rows of the table), U_f allowed the agent to suffer the -2 rewards on the path to the upper terminating cell, in exchange for a shorter time to termination. An optimal agent would go in a straight line to the right and terminate in three steps, but DηN behaved suboptimally most of the time. For $c_0 = 7$ (last row), we see that the agent often took the path to the upper terminating cell, however, for $c_0 = 6$ (second to last line) the agent rarely did so, often going for the lower terminating cell.

Why did DηN overshoot the second coordinate of the discounted return on the first three settings, and why did it rarely go for the upper terminating cell when $c_0 = 6$? We hypothesize that the cause was inaccuracy in the return distribution estimates. A small underestimation of $\mathbb{E}\left((c_t)_2 + G(s_t, c_t)_2\right)_-$ will be amplified by $\alpha = 50$ and may cause the agent to become "conservative" in optimizing for this term of the objective, relative to term on the first coordinate of the discounted return. To test this hypothesis, we ran a second version of our experiment with $\alpha = 500$. The choice of $\alpha \in \{50, 500\}$ should have little impact on an optimal agent's behavior with the values of c_0 we considered, however, larger α should make an agent with imperfect return estimates seem more conservative. The results are in Table 5. Consistent with our hypothesis, we observe that DηN with $\alpha = 500$ appears more conservative, with longer episodes than with $\alpha = 50$, especially for $c_0 = 0$ and $c_0 = 7$. For $c_0 = 0$, the agent did not take the zero-reward path to the lower terminating cell, but

Lower-bound $-(c_0)_2$	Discounted Return $\mathbb{E}G(s_0, c_0)_2$	Penalty term $\mathbb{E}\left((c_0)_2 + G(s_0, c_0)_2\right)_-$	Episode duration
3.00	$5.83\ (5.09, 7.06)$	$-0.00\ (-0.00, 0.00)$	$12.97\ (12.17, 13.83)$
2.00	$4.75\ (3.89, 5.92)$	$-0.04\ (-0.20, 0.00)$	$12.47\ (11.43, 13.53)$
1.00	$3.38\ (2.73, 4.40)$	$-0.00\ (-0.01, 0.00)$	$11.83\ (10.73, 13.03)$
0.00	$1.73\ (1.24, 2.44)$	$0.00\ (0.00, 0.00)$	$12.07\ (10.77, 13.30)$
-1.00	$0.36\ (0.13, 0.84)$	$0.00\ (0.00, 0.00)$	$6.80\ (5.80, 8.47)$
-2.00	$0.19\ (-0.07, 0.63)$	$0.00\ (0.00, 0.00)$	$6.77\ (5.67, 8.47)$
-6.00	$-0.27\ (-1.14, -0.01)$	$0.00\ (0.00, 0.00)$	$6.50\ (5.43, 8.30)$
-7.00	$-0.74\ (-1.74, -0.07)$	$0.00\ (0.00, 0.00)$	$5.97\ (5.03, 7.67)$

Table 5: Performance of DηN trading off minimizing constraint violation and maximizing expected return. The weight of the second term is $\alpha = 500$. Entries are averages with bootstrap confidence intervals in the format "average (low, high)" where low and high are the interval bounds.

first visited the rewarding cell in the bottom-left corner, and for $c_0 = 7$ the agent did not go to the upper terminating cell.

8. Atari Experiment

Atari 2600 (Bellemare et al., 2013) is a popular RL benchmark where several deep RL agents have been evaluated, including DQN (Mnih et al., 2015) and QR-DQN (Dabney et al., 2018). It provides us with a more challenging setting for deep RL agents than gridworld instances, since agents must overcome multiple learning challenges—to name a few: perception, exploration and control over longer timescales.

Atari 2600 is very much an RL benchmark, with games framed as RL problems in which the goal is to maximize the score. However, we can use the game of Pong to create an interesting setting for generating returns—an Atari analogue of the gridworld experiments in Section 7.1. In Pong, the agent plays against an opponent controlled by the environment. The goal of the game is for each player to get the ball to cross the edge of the opponent's side of the screen. Each time this happens, the player gets a point. Each player controls a paddle that can be used for hitting back the ball, preventing the opponent from scoring a point and sending the ball toward the opponent in a straight trajectory.

In a typical RL setting, we train agents to maximize the score (the difference between the player's and the opponent's scores), but in this section we are interested in using DηN to achieve different scores, which entails both scoring against the opponent, and being scored upon. We trained DηN and evaluated the trained agent with different values of c_0, corresponding to different desired discounted returns, $\gamma = 0.997$, and reduced episode duration from thirty minutes to twenty-five seconds (implementation details are given in Appendix H.3). This dramatic reduction is related to the interaction between γ and the objective functional. The goal is to control the distribution of the discounted return from the start of the episode. A reward at time step $t + 1$ offsets this discounted return by

Desired discounted return $-c_0$	Discounted return $\mathbb{E}G(s_0, c_0)$	Error $\mathbb{E}\lvert c_0 + G(s_0, c_0)\rvert$
4.00	$2.26\ (2.22, 2.28)$	$1.74\ (1.72, 1.78)$
2.00	$1.90\ (1.88, 1.92)$	$0.15\ (0.13, 0.18)$
1.00	$0.88\ (0.82, 0.95)$	$0.23\ (0.21, 0.27)$
0.00	$-0.23\ (-0.33, -0.15)$	$0.29\ (0.22, 0.37)$
-1.00	$-1.03\ (-1.09, -0.95)$	$0.19\ (0.16, 0.21)$
-2.00	$-2.06\ (-2.11, -1.96)$	$0.18\ (0.16, 0.22)$
-4.00	$-3.97\ (-4.01, -3.94)$	$0.14\ (0.11, 0.16)$

Table 6: Evaluation results generating discounted returns with DηN in Pong and $\gamma = 0.997$. Entries are averages with bootstrap confidence intervals in the format "average (low, high)" where low and high are the interval bounds.

$\gamma^t R_{t+1}$. The rewards in Pong are ± 1 and the agent acts at 15Hz, so after 25s an observed reward only offsets the discounted return by approximately ± 0.32. As the episode advances, the effect of the agent's actions on the value of the objective decreases, and at a minute this effect has reduced to ± 0.07. The agent's behavior after that is unlikely to make any meaningful difference to the return and collected data may be less useful for training. For these experiments, we have sidestepped the issue by reducing the episode duration, but the interaction between the timescale and γ for stock-augmented return distribution optimization is an important practical consideration that deserves a systematic study in future work.

Table 6 shows the performance of DηN. Similar to the setting in Table 2, we trained the agent and, for evaluation, conditioned its policy on different values of c_0 corresponding to the negative of the desired discounted return. We measured the agent's average discounted return ($\mathbb{E}G(s_0, c_0)$) and the "error" $\mathbb{E}\lvert c_0 + G(s_0, c_0)\rvert$. The confidence intervals correspond to 95%-confidence bootstrap intervals over 12 independent repetitions of training and evaluation (differently from the 30 independent runs in the gridworld setting). DηN approximately and reliably generated the desired discounted returns for various choices of c_0, with the exception of discounted returns to approximate $-c_0 = 4$ (first row). We believe that the agent's training regime explains the successes, as well as the failure for $-c_0 = 4$.

We used DηN's policy for data collection during training, which required us to select c_0 during training. At the beginning of each episode, we sampled a value for c_0 uniformly at random from $[-9, 9)$. This was the strategy used in the gridworld experiments (albeit with a different interval) and it was meant to increase data diversity. Because the episodes in Atari were much longer than in the gridworld experiment (375 versus 16 steps), this strategy likely yielded little diversity in the stocks observed later in the episode. Diversity is important because we need to train the stock-augmented agent to optimize the objective for a variety of augmented states. Similar to how certain RL problems may pose exploration challenges in the state space $\mathcal{S}$, stock-augmented problems may suffer from exploration challenges in the augmented-state space ($\mathcal{S} \times \mathcal{C}$).

Fortunately, we can reintroduce diversity across stocks after generating data, based on the following observation: When the state dynamics are independent of the stock, from a single transition $(S_t, C_t), A_t, R_{t+1}, (S_{t+1}, C_{t+1})$, it is possible to generate counterfactual transitions with the correct distribution for the whole spectrum of stocks $c \in \mathcal{C}$, that is, the following transitions:

$$\left\{(S_t, c), A_t, R_{t+1}, (S_{t+1}, \gamma^{-1}(c + R_{t+1})) : c \in \mathcal{C}\right\}.$$

We refer to this change on C_t and C_{t+1} as *stock editing.* DηN updates parameters using a minibatch of trajectories with subsequent transitions. In this setting, before performing each update, we edited the stocks in the minibatch as follows: We sampled a value of C'_0 uniformly at random from $[-9, 9)$ for the first step of each trajectory, and edited the whole trajectory to create new transitions $(S_{t+k}, C'_k), A_{t+k}, R_{t+k+1}, (S_{t+k+1}, C'_{k+1})$ with, for $k \geq 0$,

$$C'_{k+1} = \gamma^{-k}\left(C'_0 + \sum_{i=0}^{k} \gamma^i R_{t+i+1}\right).$$

Stock editing was essential for our results, and we were unable to reproduce the outcomes in Table 6 without it.

We believe that the failure for $-c_0 = 4$ happened because there was not enough data for learning to generate discounted returns of approximately 4. As $-c_0$ increases, the behaviors generated for the diverse stocks through stock editing are likely not as useful for solving the problem at c_0. In other words, we conjecture that the data was diverse but imbalanced, and we pose this issue of data balance as a question for future work.

9. Conclusion

While standard RL has been successfully employed to solve various practical problems, its formulation as maximizing expected return limits its use in the design of intelligent agents. The problem of return distribution optimization aims to address this limitation by posing the optimization of a statistical functional of the return distribution. While this is a more general problem, the additional flexibility cannot be exploited by DP, as distributional DP can only solve the instances that classic DP can solve (Marthe et al., 2024). We showed that this limitation can be addressed by augmenting the state of the MDP with *stock* (Equation 1), a statistic originally introduced by Bäuerle and Ott (2011) for optimizing the τ-CVaR with classic DP, and recurrent within the risk-sensitive RL literature (Lim and Malik, 2022; Moghimi and Ku, 2025), but not beyond. It is through the combination of distributional RL, stock augmentation and optimizing statistical functionals of the return distribution that distributional DP can tackle a broader class of return distribution optimization problems than what is possible when any of the components are missing.

We introduced distributional value iteration and distributional policy iteration as principled distributional DP methods for stock-augmented return distribution optimization, that is, optimizing various objective functionals F_K of the return distribution. These methods enjoy performance bounds that resemble the classic DP bounds, and they can be applied to various RL-like problems that have been the subject of interest in previous work, including instances of risk-sensitive RL (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Noorani et al., 2022; Moghimi and Ku, 2025), homeostatic regulation (Keramati and Gutkin, 2011) and constraint satisfaction.

Distributional DP offers a clear path for developing practical return distribution optimization methods based on existing deep RL agents, as exemplified by our empirical results. We adapted QR-DQN (Dabney et al., 2018) to incorporate the principles of distributional DP into a novel agent called DηN (Deep η-Networks, pronounced *din*), and illustrated that it works as intended in different simple scenarios for return distribution optimization in gridworld and Atari.

We believe there are a number of interesting directions for future work in stock-augmented return distribution optimization. Besides open theoretical questions, there are various practical challenges to be studied systematically on the path to developing strong practical methods for return distribution optimization. Because return distribution optimization formalizes a wide range of problems, these solution methods can have broad applicability in practice.

9.1 Open Theoretical Questions

Does an optimal return distribution exist when K is indifferent to γ, indifferent to mixtures and Lipschitz? If this is the case, the proofs of Theorems 6 and 8 can be simplified and the bounds can be tightened to depend on the optimal return distribution, similar to how the classic DP error bounds depend on the optimal value function.

What is needed for DP to optimize an objective functional in the infinite-horizon discounted case? We conjecture some form of uniform continuity may be necessary (see Appendix C.2, where we show a failure case with U_f and $f(x) = \mathbb{I}(x > 0)$). We also conjecture that Lipschitz continuity is needed for uniform bounds to be possible.

Can we develop distributional DP methods to solve constrained problems? We have come close to constrained problems in Section 5.5, and it would be interesting to develop a theory of stock-augmented constrained return distribution optimization, somewhat like constrained MDPs (Altman, 1999) are related to RL.

9.2 Addressing DηN's Limitations

DηN is a proof-of-concept stock-augmented agent that we used for illustrating how the principles underlying distributional value/policy iteration can be incorporated into a deep reinforcement learning agent. Below, we list some limitations of the method that we believe should be addressed on the path to developing full-fledged stock-augmented agents for optimizing return distributions in challenging environments.

How to embed the stock? We have employed a simple embedding strategy for the stock in DηN's network, which relies on inputting the stock to an MLP and adding out result to the output of the agent's vision network (see Figure 1). This was sufficient for our experiments, however improved scalar embedding should be considered in the future (for example, Springenberg et al., 2024), as it may improve the agent's data efficiency and performance, especially in more challenging environments.

How to go beyond expected utilities? The fact that DηN can only optimize expected utilities is also a limitation worth addressing. DηN relies on the existence of greedy actions, which holds for expected utilities, but not for other objective functionals. That is, other stock-augmented return distribution optimization problems may only admit optimal stochastic policies. Perhaps an approach based on policy gradient (Sutton and Barto, 2018; Espeholt et al., 2018) or policy optimization (Schulman et al., 2017; Abdolmaleki et al., 2018) may be therefore more suited for going beyond expected utilities.

How to estimate distributions of vector-valued returns? DηN maintains estimates of the marginal distributions (per coordinate) of the vector-valued returns (see Appendix H.1). This was enough for our experiments, but our simplification highlights an important consideration: We want practical methods that can estimate the distributions of vector-valued returns. This capability is needed, for example, to tackle the formulation of homeostatic regulation proposed by Keramati and Gutkin (2011). Zhang et al. (2021); Wiltzer et al. (2024) have studied learning distributional estimates with vector-valued returns, so their results can inform the design of distributional estimators for vector-valued returns.

9.3 Practical Challenges

Our experimental results revealed a number of interesting challenges in stock-augmented return distribution optimization that we believe should be addressed in order to develop effective agents for practical settings.

In our experiments we mitigated these issues with simple ideas, and we were helped by the simplicity of the experimental settings, but stronger solutions may be required in more challenging environments. We typically need to apply interventions to the stock during training, in order to generate diverse data (Sections 7 and 8). The interaction of objective functional, c_0 and approximate return distribution estimates may result in degenerate behavior (Sections 7.2 and 7.3) and this can be worsened when c_0 is selected through a procedure like grid-search to optimize an approximate objective (as in the case of τ-CVaR, both risk-averse and risk-seeking). Depending on the objective functional, near-optimal decision making may require substantially accurate return estimates (Section 7.4). Over long timescales, the discount factor may limit the agent's ability to influence the returns (Section 8). In more complex environments, we need to ensure the training data is not only diverse across the stock spectrum, but also balanced, lest the learned policies underperform for certain choices of c_0.

Acknowledgments

We thank Csaba Szepesvári for reviewing our draft of this work. We thank Kalesha Bullard, Noémi Éltető, András György, Lucia Cipolina Kun, Dale Schuurmans, and Yunhao Tang for helpful discussions. We thank Yang Peng for identifying issues on a previous version of this paper and proposing fixes, along with technical feedback and discussions. We also thank the anonymous JMLR Reviewers, for their technical review and the thoughtful suggestions for improvement, and Martha White for her work as Action Editor to this work. Our experimental infrastructure was built using Python 3, Flax (Heek et al., 2024), Haiku (Hennigan et al., 2020), JAX (Bradbury et al., 2018), and NumPy (Harris et al., 2020). We have used Matplotlib (Hunter, 2007), NumPy (Harris et al., 2020), pandas (Wes McKinney, 2010; pandas development team, 2020) and SciPy (Virtanen et al., 2020) for analyzing and plotting our experimental data.

Appendix A. Additional Theoretical Results

A.1 Complete Spaces

Lemma 23 *The spaces* $(\mathcal{D}, \mathrm{w})$ *and* $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ *are complete.*

Proof We know that $(\mathcal{D}, \mathrm{w})$ is complete (Theorem 6.18, p. 116; Villani, 2009), so it remains to show that $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ is complete. Let $\eta_1, \eta_2, \ldots$ be a Cauchy sequence in $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$. For each (s, c), the sequence $\eta_1(s, c), \eta_2(s, c), \ldots$ is Cauchy in $(\mathcal{D}, \mathrm{w})$ and by completeness it has a limit $\eta_\infty(s, c)$.

We claim that η_∞ is the limit of $\eta_1, \eta_2, \ldots$ in $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$. Given $\varepsilon > 0$, we can take n such that $\sup_{n'\geq n} \overline{\mathrm{w}}(\eta_{n'}, \eta_n) < \varepsilon$, which means

$$\begin{aligned}
\varepsilon &> \sup_{n'\geq n} \overline{\mathrm{w}}(\eta_{n'}, \eta_n) \\
&= \sup_{n'\geq n} \sup_{s,c} \mathrm{w}(\eta_{n'}(s, c), \eta_n(s, c)) \\
&\geq \sup_{n'\geq n} \sup_{s,c} \mathrm{w}(\eta_{n'}(s, c), \eta_\infty(s, c)) \\
&= \sup_{n'\geq n} \overline{\mathrm{w}}(\eta_{n'}, \eta_\infty),
\end{aligned}$$

and since this holds for all $\varepsilon > 0$ we have that $\limsup_{n\to\infty} \overline{\mathrm{w}}(\eta_n, \eta_\infty) = 0$. Combining the above with the fact that $\overline{\mathrm{w}}$ is a norm gives

$$0 \leq \liminf_{n\to\infty} \overline{\mathrm{w}}(\eta_n, \eta_\infty) \leq \limsup_{n\to\infty} \overline{\mathrm{w}}(\eta_n, \eta_\infty) = 0,$$

so, indeed, η_∞ is the limit of $\eta_1, \eta_2, \ldots$.

It remains to show that $\eta_\infty \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, that is, that $\overline{\mathrm{w}}(\eta_\infty) < \infty$. Fix $\varepsilon > 0$ and n such that $\overline{\mathrm{w}}(\eta_n, \eta_\infty) < \varepsilon$. We have $\overline{\mathrm{w}}(\eta_n) < \infty$ since $\eta_n \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, and, by the triangle inequality, $\overline{\mathrm{w}}(\eta_n, \eta_\infty) \geq \overline{\mathrm{w}}(\eta_\infty) - \overline{\mathrm{w}}(\eta_n)$, so $\overline{\mathrm{w}}(\eta_\infty) \leq \overline{\mathrm{w}}(\eta_n) + \varepsilon < \infty$. ■

Appendix B. Analysis of Distributional Dynamic Programming

B.1 History-based policies

We start by reducing the stock-augmented return distribution optimization problem to an optimization over *Markov policies.*

Proposition 24 *If Assumption 1 holds and* $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ *is indifferent to mixtures and indifferent to* γ*, and if: i) the MDP has finite horizon; or ii)* $\gamma < 1$ *and* K *is Lipschitz, then*

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F_K \eta^\pi = \sup_{\pi\in\Pi_{\mathrm{M}}} F_K \eta^\pi = \sup_{\pi\in\Pi} F_K \eta^\pi.$$

Proof We write $F = F_K$. First note that

$$\sup_{\pi\in\Pi} F\eta^\pi \leq \sup_{\pi\in\Pi_{\mathrm{M}}} F\eta^\pi \leq \sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi,$$

so it suffices to show that

$$\sup_{\pi \in \Pi_{\mathrm{H}}} F\eta^{\pi} \leq \sup_{\pi \in \Pi} F\eta^{\pi}.$$

We will first consider history-based policies that are eventually stationary Markov. Recall the definition of a history from Section 2:

$$h_t \doteq (s_0, c_0), a_0, r_1, (s_1, c_1), \ldots, r_t, (s_t, c_t)$$

with $h_0 \doteq (s_0, c_0)$. Let $\Pi_{\mathrm{H},n}$ be the set of all history-based policies $\rho \in \Pi_{\mathrm{H}}$ for which there exists a stationary $\pi \in \Pi$ such that, for all $n' \geq n$ and every history $h_{n'}$, we have $\rho(h_{n'}) = \pi(s_{n'}, c_{n'})$. In particular, $\Pi_{\mathrm{H},0} = \Pi$.

Assume, by means of induction, that for some $n \in \mathbb{N}_0$ we have

$$\sup_{\rho \in \Pi_{\mathrm{H},n}} F\eta^{\rho} \leq \sup_{\pi \in \Pi} F\eta^{\pi}.$$

Given a $\rho \in \Pi_{\mathrm{H},n+1}$ and its corresponding stationary policy $\pi \in \Pi$, let $\overline{\pi}$ satisfy

$$FT_{\overline{\pi}}\eta^{\pi} = \sup_{\pi' \in \Pi} FT_{\pi'}\eta^{\pi}.$$

By Lemma 15, we have $F\eta^{\overline{\pi}} \geq F\eta^{\pi}$. Now, define the policy ρ' by

$$\overline{\rho}(h_t) \doteq \begin{cases} \rho(h_t) & t < n, \\ \overline{\pi}(s_t, c_t) & t \geq n. \end{cases}$$

We have that $\overline{\rho} \in \Pi_{\mathrm{H},n}$, and we now show that $F\eta^{\overline{\rho}} \geq F\eta^{\rho}$.

Define, for all $(s, c) \in \mathcal{S} \times \mathcal{C}$, $G^{\pi}(s, c) \sim \eta^{\pi}(s, c)$ (and independent from all other random variables) and $G^{\overline{\pi}}(s, c) \sim \eta^{\overline{\pi}}(s, c)$ (and independent from all other random variables). Fix $(S_0, C_0) = (s_0, c_0)$ (with probability one) and let H_n be the (random) history and G_0^{ρ} the return generated by following ρ from (S_0, C_0). Similarly, define the respective $\overline{H}_n$ and $G_0^{\overline{\rho}}$ corresponding to $\overline{\rho}$.

Equation 2 and the definitions above give

$$C_0 + G_0^{\rho} \overset{\mathcal{D}}{=} \gamma^{-n}(C_n + R_{n+1} + \gamma G^{\pi}(S_{n+1}, C_{n+1}))$$

and

$$C_0 + G_0^{\overline{\rho}} \overset{\mathcal{D}}{=} \gamma^{-n}(C_n + G^{\overline{\pi}}(S_n, C_n)).$$

The choice of $\overline{\pi}$ and the fact that K is indifferent to mixtures means that

$$K(C_n + G^{\overline{\pi}}(S_n, C_n)) \geq K(C_n + R_{n+1} + \gamma G^{\pi}(S_{n+1}, C_{n+1}))$$

with probability one. K is also indifferent to γ, so

$$K(\gamma^{-k}(C_n + G^{\overline{\pi}}(S_n, C_n))) \geq K(\gamma^{-k}(C_n + R_{n+1} + \gamma G^{\pi}(S_{n+1}, C_{n+1}))),$$

which implies that $K(C_0 + G_0^{\overline{\rho}}) \geq K(C_0 + G_0^{\rho})$ and this holds for every choice of (s_0, c_0), so $F\eta^{\overline{\rho}} \geq F\eta^{\rho}$. Thus, by induction, we have that for all $n \in \mathbb{N}_0$

$$\sup_{\rho \in \Pi_{\mathrm{H},n}} F\eta^{\rho} \leq \sup_{\pi \in \Pi} F\eta^{\pi}. \tag{28}$$

Equation 28 is sufficient for the finite-horizon case, since we can take n large enough so that

$$\sup_{\rho \in \Pi_{\mathrm{H},n}} F\eta^\rho = \sup_{\pi \in \Pi_{\mathrm{H}}} F\eta^\pi.$$

For the infinite-horizon discounted case, we proceed as follows. Fix $n \in \mathbb{N}_0$, and fix $\pi \in \Pi_{\mathrm{H}}$ and $\rho \in \Pi_{\mathrm{H},n}$ such that π and ρ are identical for all histories of size strictly less than n. For $t \in \mathbb{N}_0$, let $G_t^\pi(s, c)$ denote the return from time step t onward generated by following π from starting augmented state (s, c). Note that the arguments (s, c) are the initial state of the history, not the augmented state at time step t. Similarly, define the corresponding $G_t^\rho(s, c)$ for ρ. Because F is Lipschitz, we have, for all $(s, c) \in \mathcal{S} \times \mathcal{C}$

$$|F\eta^\pi(s,c) - F\eta^\rho(s,c)| \le \gamma^n \mathrm{w}\left(\mathrm{df}(G_n^\pi(s,c)), \mathrm{df}(G_n^\rho(s,c))\right).$$

By Assumption 1, there exists a constant κ such that

$$\sup_{s \in \mathcal{S}, c \in \mathcal{C}} \mathrm{w}\left(\mathrm{df}(G_n^\pi(s,c)), \mathrm{df}(G_n^\rho(s,c))\right) \le \kappa$$

uniformly for all π, ρ and n. Thus, for all $n \in \mathbb{N}_0$,

$$\sup_{\pi \in \Pi_{\mathrm{H}}} \inf_{\rho \in \Pi_{\mathrm{H},n}} \sup_{s \in \mathcal{S}, c \in \mathcal{C}} |F\eta^\pi(s,c) - F\eta^\rho(s,c)| \le \gamma^n \kappa, \tag{29}$$

and

$$\begin{aligned} \sup_{\pi \in \Pi_{\mathrm{H}}} F\eta^\pi &\le \sup_{\rho \in \Pi_{\mathrm{H},n+1}} F\eta^\rho + \gamma^n \kappa && \text{(Equation 29)} \\ &= \sup_{\pi \in \Pi} F\eta^\pi + \gamma^n \kappa && \text{(Equation 28)} \end{aligned}$$

Taking the limit of $n \to \infty$ gives the result. ■

Proposition 24 implies that under the conditions on F_K, for every history-based policy $\pi \in \Pi_{\mathrm{H}}$ we can find a Markov policy $\overline{\pi} \in \Pi_{\mathrm{M}}$ that is no worse than π simultaneously for all (s, c). In this sense, the quantity $\sup_{\pi \in \Pi_{\mathrm{M}}} F_K \eta^\pi$ is well-defined, even though it is a supremum of a vector-valued quantity.

B.2 Distributional Policy Evaluation

For our analysis, we also employ existing distributional RL theory for policy evaluation:

Theorem 25 (from Proposition 4.15, p. 88, Bellemare et al., 2023) *For every stationary policy $\pi \in \Pi$, the distributional Bellman operator T_π is a non-expansion in the supremum 1-Wasserstein distance. If $\gamma < 1$, then T_π is a γ-contraction in the supremum 1-Wasserstein distance.*

Proof The proof is as presented by Bellemare et al. (2023), with the caveat that to obtain the result for $\mathcal{C} = \mathbb{R}^m$ with $m > 1$ we apply Proposition 4.15 to each coordinate of the vector-valued rewards individually. ■

The following lemma uses Theorem 25 to give us a policy evaluation result for the infinite-horizon case.

Lemma 26 (Distributional Policy Evaluation) *If $\gamma < 1$ or the MDP has finite horizon, for any $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and $\pi \in \Pi_{\mathrm{M}}$ we have*

$$\lim_{n\to\infty} \overline{\mathrm{w}}(T_{\pi_1} \cdots T_{\pi_n}\eta, T_{\pi_1} \cdots T_{\pi_n}\eta') = 0.$$

Proof *Discounted Case.* In this case, $\gamma < 1$ and T_π is a γ-contraction by Theorem 25. Letting $\eta_n \doteq T_{\pi_1} \cdots T_{\pi_n}\eta$ and $\eta'_n \doteq T_{\pi_1} \cdots T_{\pi_n}\eta'$ for $n \geq 1$, for every $n \geq 1$, we have

$$\overline{\mathrm{w}}(\eta_n, \eta'_n) \leq \gamma^n \overline{\mathrm{w}}(\eta, \eta'),$$

and

$$\overline{\mathrm{w}}(\eta, \eta') \leq \overline{\mathrm{w}}(\eta) + \overline{\mathrm{w}}(\eta') < \infty.$$

so $\limsup_{n\to\infty} \overline{\mathrm{w}}(\eta_n, \eta'_n) = 0$, which implies the result.

Finite-horizon Case. In finite-horizon MDPs, if n is greater or equal to the horizon, then

$$T_{\pi_1} \cdots T_{\pi_n}\eta = T_{\pi_1} \cdots T_{\pi_n}\eta',$$

for all $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, so

$$\overline{\mathrm{w}}(T_{\pi_1} \cdots T_{\pi_n}\eta) = \overline{\mathrm{w}}(T_{\pi_1} \cdots T_{\pi_n}\eta'),$$

and we must show is that $T_{\pi_1} \cdots T_{\pi_n}\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$. When the MDP has finite horizon, T_π is a non-expansion (by Theorem 25), which implies that $\sup_{\pi\in\Pi} \overline{\mathrm{w}}(T_\pi\eta) < \infty$ and $\overline{\mathrm{w}}(T_{\pi_1} \cdots T_{\pi_n}\eta) \leq \overline{\mathrm{w}}(\eta) < \infty$ for all $n \geq 1$. ■

We refer to Lemma 26 as the distributional policy evaluation result because it implies that for a stationary policy $\pi \in \Pi$ the sequence discounted return functions given by $\eta_n \doteq T^n_\pi \eta$ converges in 1-Wasserstein distance to η^π, the distribution of discounted returns obtained by π. Moreover, the sequence of returns $G_n \sim \eta_n$ (which are distributed independently from each other) converges almost surely to a $G^\pi \overset{\mathcal{D}}{=} \sum_{t=0}^{\infty} \gamma_t R_{t+1}$ (Skorokhod's Theorem, p. 114; Shorack, 2017)

B.3 Local Policy Improvement

Informally, DP builds a globally optimal policy by "chaining" locally optimal decisions at each time step. A "distributional max operator" gives a return distribution where the first decision is locally optimal:

Definition 27 (Distributional Max Operator) *Given $F : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to \mathbb{R}^{\mathcal{S}\times\mathcal{C}}$, an operator $T_* : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ is a* distributional max operator *if it satisfies, for all $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$,*

$$FT_*\eta = \sup_{\pi\in\Pi} FT_\pi\eta.$$

The mechanism for locally optimal decision-making is the greedy policy, which is a policy that realizes a distributional max operator:

Definition 28 (Greedy Policy) *Given $F : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to \mathbb{R}^{\mathcal{S}\times\mathcal{C}}$, a policy $\pi \in \Pi$ is* greedy *with respect to $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ if*

$$FT_\pi\eta = FT_*\eta.$$

Given F_K, it is possible that K is such that for some $\nu \in (\mathcal{D}, \mathrm{w})$ we have $K\nu$ degenerate and "infinite" (for example, the expected utility U_f with $f(x) = x^{-1}$). In this case, we interpret K as encoding a preference where if $\nu_1, \nu_2, \ldots (\mathcal{D}, \mathrm{w})$ converges to ν_∞ and $K\nu_n < \infty$ for all n, but $\liminf_{n\to\infty} K\nu_n = \infty$, so there is no $\nu \in (\mathcal{D}, \mathrm{w})$ that is strictly preferred over ν_∞. In this sense, we write $K\nu_\infty \geq \sup_{\nu\in(\mathcal{D},\mathrm{w})} K\nu$. Similarly, for F_K and $\overline{\pi}$ greedy with respect to η, we write

$$F_K T_*\eta = F_K T_{\overline{\pi}}\eta \geq \sup_{\pi\in\Pi} F_K T_\pi \eta$$

even if the right-hand side is infinite for some $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and $(s, c) \in \mathcal{S} \times \mathcal{C}$.

B.4 Monotonicity

The following intermediate result will be useful for proving monotonicity, and it highlights a phenomenon in stock-augmented problems where the rewards are absorbed into the augmented state:

Lemma 29 (Reward absorption) *For every stationary policy $\pi \in \Pi$, $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and $(s, c) \in \mathcal{S} \times \mathcal{C}$, if $(S_t, C_t) = (s, c)$, $A_t \sim \pi(S_t, C_t)$, $G_{\text{lookahead}}(s, c) \sim (T_\pi\eta)(s, c)$ and $G(s, c) \sim \eta(s, c)$, then*

$$C_t + G_{\text{lookahead}}(S_t, C_t) \overset{\mathcal{D}}{=} \gamma\left(C_{t+1} + G(S_{t+1}, C_{t+1})\right).$$

Proof We have that

$$\begin{aligned} C_t + G_{\text{lookahead}}(S_t, C_t) &\overset{\mathcal{D}}{=} C_t + R_{t+1} + \gamma G(S_{t+1}, C_{t+1}) && \text{(Definition of } T_\pi\text{)} \\ &\overset{\mathcal{D}}{=} \gamma C_{t+1} + \gamma G(S_{t+1}, C_{t+1}) && \text{(Equation 1)} \\ &\overset{\mathcal{D}}{=} \gamma\left(C_{t+1} + G(S_{t+1}, C_{t+1})\right). \end{aligned}$$

■

Lemma 14 (Monotonicity) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, then, for every $\pi \in \Pi$, the distributional Bellman operator T_π is* monotone *(or* order-preserving*) with respect to the preference induced by F_K on $(\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$. That is, for every stationary policy $\pi \in \Pi$ and $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, we have*

$$F_K\eta \geq F_K\eta' \Rightarrow F_K T_\pi\eta \geq F_K T_\pi \eta'.$$

Proof Fix a stationary policy $\pi \in \Pi$ and $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ satisfying $F_K\eta \geq F_K\eta'$. Fix also $(s, c) \in \mathcal{S} \times \mathcal{C}$, and let $(S_t, C_t) = (s, c)$, $A_t \sim \pi(S_t, C_t)$, $G(s, c) \sim \eta(s, c)$, $G'(s, c) \sim \eta'(s, c)$, $G_{\text{lookahead}}(s, c) \sim (T_\pi\eta)(s, c)$ and $G'_{\text{lookahead}}(s, c) \sim (T_\pi\eta')(s, c)$

By assumption, we have $K(c + G(s, c)) \geq K(c + G'(s, c))$ for all (s, c). Combining the above with indifference to mixtures, we get

$$K(C_{t+1} + G(S_{t+1}, C_{t+1})) \geq K(C_{t+1} + G'(S_{t+1}, C_{t+1})),$$

and, thanks to indifference to γ,

$$K(\gamma\left(C_{t+1}+G(S_{t+1},C_{t+1})\right)) \geq K(\gamma\left(C_{t+1}+G'(S_{t+1},C_{t+1})\right)).$$

From Lemma 29 we have that

$$\begin{aligned}C_t + G_{\text{lookahead}}(S_t,C_t) &\overset{\mathcal{D}}{=} \gamma\left(C_{t+1}+G(S_{t+1},C_{t+1})\right),\\ C_t + G'_{\text{lookahead}}(S_t,C_t) &\overset{\mathcal{D}}{=} \gamma\left(C_{t+1}+G'(S_{t+1},C_{t+1})\right),\end{aligned}$$

so it follows that

$$K(C_t + G_{\text{lookahead}}(S_t,C_t)) \geq K(C_t + G'_{\text{lookahead}}(S_t,C_t)). \qquad \blacksquare$$

B.5 Convergence

Definition 30 (Lipschitz Continuity for Objective Functionals) *The objective functional $F : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to \mathbb{R}^{\mathcal{S}\times\mathcal{C}}$ is L-*Lipschitz *(or* Lipschitz*, for simplicity) if there exists $L \in \mathbb{R}$ such that*

$$\sup_{\substack{\eta,\eta':\\ \overline{\mathrm{w}}(\eta)<\infty\\ \overline{\mathrm{w}}(\eta')<\infty\\ \overline{\mathrm{w}}(\eta,\eta')>0}} \frac{\|F\eta - F\eta'\|_\infty}{\overline{\mathrm{w}}(\eta,\eta')} \leq L.$$

L is the Lipschitz constant *of F.*

Proposition 31 *Given $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$, F_K is L-Lipschitz iff K is L-Lipschitz.*

Proof If F_K is L-Lipschitz, then

$$\begin{aligned}L &\geq \sup_{\substack{\eta,\eta':\\ \overline{\mathrm{w}}(\eta)<\infty\\ \overline{\mathrm{w}}(\eta')<\infty\\ \overline{\mathrm{w}}(\eta,\eta')>0}} \frac{\|F_K\eta - F_K\eta'\|_\infty}{\overline{\mathrm{w}}(\eta,\eta')}\\ &\geq \sup_{c\in\mathcal{C}} \sup_{\substack{\nu,\nu':\\ \mathrm{w}(\nu)<\infty\\ \mathrm{w}(\nu')<\infty\\ \mathrm{w}(\nu,\nu')>0}} \frac{|K(c+G)-K(c+G')|}{\mathrm{w}(\mathrm{df}(c+G),\mathrm{df}(c+G'))} && (G\sim\nu,\ G'\sim\nu')\\ &\geq \sup_{\substack{\nu,\nu':\\ \mathrm{w}(\nu)<\infty\\ \mathrm{w}(\nu')<\infty\\ \mathrm{w}(\nu,\nu')>0}} \frac{|K\nu - K\nu'|}{\mathrm{w}(\nu,\nu')}, && (c=0)\end{aligned}$$

so K is L-Lipschitz. If, on the other hand, K is L-Lipschitz, then, for all $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$,

$$\begin{aligned}
\|F_K\eta - F_K\eta'\|_\infty &= \sup_{(s,c)\in\mathcal{S}\times\mathcal{C}} |K(c+G(s,c)) - K(c+G'(s,c))| \\
&\qquad\qquad\qquad\qquad (G(s,c)\sim\eta(s,c),\, G'(s,c)\sim\eta'(s,c)) \\
&\le \sup_{(s,c)\in\mathcal{S}\times\mathcal{C}} L\cdot \mathrm{w}(\eta(s,c),\eta'(s,c)) \\
&= L\cdot\overline{\mathrm{w}}(\eta,\eta'),
\end{aligned}$$

so F_K is L-Lipschitz. ■

Proposition 32 *If $F : (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}}) \to \mathbb{R}^{\mathcal{S}\times\mathcal{C}}$ is Lipschitz and the sequence $\eta_1, \eta_2, \ldots \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ converges in $\overline{\mathrm{w}}$ to some η_∞, then $F\eta_1, F\eta_2, \ldots \in \mathbb{R}^{\mathcal{S}\times\mathbb{R}}$ converges in supremum norm to $F\eta_\infty$.*

Proof If $\eta_1, \eta_2, \ldots \in (\mathcal{D}^{\mathcal{S}\times\mathbb{R}}, \overline{\mathrm{w}})$ converges in $\overline{\mathrm{w}}$ to some η_∞ and F is L-Lipschitz, then

$$\limsup_{n\to\infty} \|F\eta_n - F\eta_\infty\|_\infty \le L\cdot\limsup_{n\to\infty}\overline{\mathrm{w}}(\eta_n,\eta_\infty) = 0,$$

which gives the result. ■

The convergence highlighted in Proposition 32 is somewhat surprising: If we consider $K\nu = \mathbb{E}(G)$ $(G\sim\nu)$, we have

$$\|F_K\delta_0\|_\infty = \sup_{c\in\mathcal{C}} |K\mathrm{df}(c+0)| = \sup_{c\in\mathcal{C}} |c| = \infty,$$

so these objective functionals may have unbounded supremum norm. However, the difference of the objective functionals for $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathbb{R}}, \overline{\mathrm{w}})$ (namely, $F\eta - F\eta'$) does have bounded supremum norm when F is Lipschitz, and we can show convergence of $F\eta_n$ to $F\eta_\infty$.

Lemma 33 *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, and if: i) the MDP has finite horizon; or ii) $\gamma < 1$ and K is Lipschitz, then for all $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$*

$$\sup_{\pi\in\Pi_{\mathrm{M}}} F_K\eta^\pi = \lim_{n\to\infty} \sup_{\pi_1,\ldots,\pi_n\in\Pi} F_K T_{\pi_n}\cdots T_{\pi_1}\eta. \tag{30}$$

If $\gamma < 1$ and K is L-Lipschitz, then for all $n \ge 0$,

$$\sup_{\pi\in\Pi_{\mathrm{M}}} F_K\eta^\pi \le \sup_{\pi_1,\ldots,\pi_n\in\Pi} F_K T_{\pi_n}\cdots T_{\pi_1}\eta + L\gamma^n \cdot \sup_{\pi'\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta, \eta^{\pi'}). \tag{31}$$

Proof We write $F = F_K$ for the rest of the proof.

If the MDP has finite horizon, then for all $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$

$$\sup_{\pi\in\Pi_{\mathrm{M}}} F\eta^\pi = \sup_{\pi_1,\ldots,\pi_n\in\Pi} FT_{\pi_n}\cdots T_{\pi_1}\eta,$$

where n is the horizon of the MDP.

Otherwise, assume that $\gamma < 1$ and assume that K is L-Lipschitz. Then F is also L-Lipschitz, by Proposition 31. By the triangle inequality, the fact that $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and Assumption 1 we have

$$\sup_{\pi'\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta, \eta^{\pi'}) \leq \overline{\mathrm{w}}(\eta) + \sup_{\pi'\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta^{\pi'}) < \infty,$$

so Equation 31 implies Equation 30 in limit $n \to \infty$.

It remains to prove Equation 31. Let

$$g_{s,c}(n) \doteq \sup_{\pi_1,\ldots,\pi_n\in\Pi} (FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c) - \sup_{\pi\in\Pi_{\mathrm{M}}} (F\eta^{\pi})(s,c)$$

and

$$h(n) \doteq \sup_{\pi_1,\ldots,\pi_n\in\Pi} \sup_{\pi'\in\Pi_{\mathrm{M}}} \|FT_{\pi_1}\cdots T_{\pi_n}\eta - FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'}\|_\infty.$$

We will show that, for all $n \geq 0$ and $(s,c) \in \mathcal{S}\times\mathcal{C}$, we have

$$|g_{s,c}(n)| \leq h(n) \leq L\gamma^n \cdot \sup_{\pi'\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta, \eta^{\pi'}).$$

For all $n \geq 0$ and $(s,c) \in \mathcal{S}\times\mathcal{C}$, we have

$$\begin{aligned}
g_{s,c}(n) &= \sup_{\pi'\in\Pi_{\mathrm{M}}} (F\eta^{\pi'})(s,c) - \sup_{\pi_1,\ldots,\pi_n\in\Pi} (FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c) \\
&= \sup_{\pi'\in\Pi_{\mathrm{M}}} \inf_{\pi_1,\ldots,\pi_n\in\Pi} \Big((F\eta^{\pi'})(s,c) - (FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c)\Big) \\
&= \sup_{\pi'_1,\ldots,\pi'_n\in\Pi} \sup_{\pi'\in\Pi_{\mathrm{M}}} \inf_{\pi_1,\ldots,\pi_n\in\Pi} \Big((FT_{\pi'_1}\cdots T_{\pi'_n}\eta^{\pi'})(s,c) - (FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c)\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\pi' \text{ is non-stationary}) \\
&\leq \sup_{\pi_1,\ldots,\pi_n\in\Pi} \sup_{\pi'\in\Pi_{\mathrm{M}}} \Big((FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'})(s,c) - (FT_{\pi_1}\cdots T_{\pi_n}\eta)(s,c)\Big) \\
&\leq \sup_{\pi_1,\ldots,\pi_n\in\Pi} \sup_{\pi'\in\Pi_{\mathrm{M}}} \Big|(FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'})(s,c) - (FT_{\pi_1}\cdots T_{\pi_n}\eta)(s,c)\Big| \\
&= \sup_{\pi_1,\ldots,\pi_n\in\Pi} \sup_{\pi'\in\Pi_{\mathrm{M}}} \|FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'} - FT_{\pi_1}\cdots T_{\pi_n}\eta\|_\infty \\
&= h(n).
\end{aligned}$$

and

$$\begin{aligned}
-g_{s,c}(n) &= \sup_{\pi_1,\dots,\pi_n\in\Pi}(FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c) - \sup_{\pi'\in\Pi_{\mathrm{M}}}(F\eta^{\pi'})(s,c) \\
&= \sup_{\pi_1,\dots,\pi_n\in\Pi}\inf_{\pi'\in\Pi_{\mathrm{M}}}\Big((FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c) - (F\eta^{\pi'})(s,c)\Big) \\
&= \sup_{\pi_1,\dots,\pi_n\in\Pi}\inf_{\pi'_1,\dots,\pi'_n\in\Pi}\inf_{\pi'\in\Pi_{\mathrm{M}}}\Big((FT_{\pi_n}\cdots T_{\pi_1}\eta)(s,c) - (FT_{\pi'_1}\cdots T_{\pi'_n}\eta^{\pi'})(s,c)\Big) \\
&\qquad\qquad (\pi' \text{ is non-stationary}) \\
&\le \sup_{\pi_1,\dots,\pi_n\in\Pi}\sup_{\pi'\in\Pi_{\mathrm{M}}}\Big((FT_{\pi_1}\cdots T_{\pi_n}\eta)(s,c) - (FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'})(s,c)\Big) \\
&\le \sup_{\pi_1,\dots,\pi_n\in\Pi}\sup_{\pi'\in\Pi_{\mathrm{M}}}\Big|(FT_{\pi_1}\cdots T_{\pi_n}\eta)(s,c) - FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'})(s,c)\Big| \\
&\le \sup_{\pi_1,\dots,\pi_n\in\Pi}\sup_{\pi'\in\Pi_{\mathrm{M}}}\|FT_{\pi_1}\cdots T_{\pi_n}\eta - FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi'}\|_\infty \\
&= h(n)
\end{aligned}$$

Thus, $-h(n) \le g_{s,c}(n) \le h(n)$, which implies $|g_{s,c}(n)| \le h(n)$.

Finally, for all $n \ge 0$, we have

$$\begin{aligned}
h(n) &= \sup_{\pi_1,\dots,\pi_n\in\Pi}\sup_{\pi'\in\Pi_{\mathrm{M}}}\|FT_{\pi_1}\cdots T_{\pi_n}\eta - FT_{\pi_1}\cdots T_{\pi_n}\eta^{\pi}\|_\infty \\
&\le L\cdot\sup_{\pi_1,\dots,\pi_n\in\Pi}\sup_{\pi'\in\Pi_{\mathrm{M}}}\overline{\mathrm{w}}(T_{\pi_1}\cdots T_{\pi_n}\eta, T_{\pi_1}\cdots T_{\pi_n}\eta^{\pi}) && (F \text{ is } L\text{-Lipschitz}) \\
&\le L\gamma^n\cdot\sup_{\pi'\in\Pi_{\mathrm{M}}}\overline{\mathrm{w}}(\eta,\eta^{\pi}). && (\gamma\text{-contraction}) \quad \blacksquare
\end{aligned}$$

B.6 Distributional Dynamic Programming

Theorem 6 (Distributional Value Iteration) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, then for every $\eta_0 \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, if the iterates $\eta_1, \eta_2, \dots$ satisfy (for $n \ge 0$)*

$$F_K\eta_{n+1} = \sup_{\pi\in\Pi} F_K T_\pi \eta_n, \qquad \text{(Distributional Value Iterates)}$$

and the policies $\overline{\pi}_0, \dots, \overline{\pi}_n$ satisfy (for $n \ge 0$),

$$F_K T_{\overline{\pi}_n}\eta_n = \sup_{\pi\in\Pi} F_K T_\pi \eta_n, \qquad \text{(Greedy Policies)}$$

then the following hold.

Finite-horizon case: *for all n greater or equal to the horizon of the MDP,*

$$F_K\eta_n = \sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^{\pi}, \tag{9}$$

and

$$F_K\eta^{\overline{\pi}_n} = \sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^{\pi}. \tag{10}$$

Discounted case ($\gamma < 1$): *If K is L-Lipschitz, then for all $n \geq 0$*

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^\pi - F_K\eta_n \leq L\gamma^n \cdot \sup_{\pi\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta_0, \eta^\pi), \tag{11}$$

and

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^\pi - F_K\eta^{\overline{\pi}_n} \leq L\gamma^n \cdot \left(\frac{1}{1-\gamma}\sup_{\pi\in\Pi} \overline{\mathrm{w}}(T_\pi\eta_0, \eta_0) + \sup_{\pi\in\Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta_0, \eta^\pi)\right). \tag{12}$$

Proof We use $F = F_K$ and note that if K L-Lipschitz then F is also L-Lipschitz (by Proposition 31. Fix $\eta_0 \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and let $\eta_n \doteq T_*^n\eta_0$ for $n \geq 1$.

The sequence $\overline{\pi}_0, \overline{\pi}_1, \overline{\pi}_2, \ldots$ satisfies $F\eta_{n+1} = FT_{\overline{\pi}_n}\eta_n = FT_*\eta_n$ for all $n \geq 0$. The definition of a distributional max operator (Definition 27) gives us

$$FT_*\eta = \sup_{\pi\in\Pi} FT_\pi\eta,$$

and, by monotonicity (Lemma 14) and induction, we have for every $n \geq 1$

$$FT_*^{n+1}\eta_0 = FT_{\overline{\pi}_n}\cdots T_{\overline{\pi}_0}\eta_0 = \sup_{\pi_0,\ldots,\pi_n\in\Pi} FT_{\pi_n}\cdots T_{\pi_0}\eta_0. \tag{32}$$

Then Equations 9 and 11 follow from Lemma 33 combined with Proposition 24, which ensures that

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi = \sup_{\pi\in\Pi_{\mathrm{M}}} F\eta^\pi$$

(the conditions of Lemma 33 and Proposition 24 are satisfied).

Equation 10 follows from Equation 9 combined with distributional policy improvement (Lemma 15). To see that Lemma 15 applies, note that, since the MDP has horizon n,

$$FT_{\overline{\pi}_n}T_{\overline{\pi}_{n-1}}\cdots T_{\overline{\pi}_0}\eta_0 = FT_{\overline{\pi}_{n-1}}\cdots T_{\overline{\pi}_0}\eta_0,$$

which satisfies Equation 16 (with $\eta = T_{\overline{\pi}_{n-1}}\cdots T_{\overline{\pi}_0}\eta_0$). Then Lemma 15 gives

$$F\eta^{\overline{\pi}_n} \geq FT_{\overline{\pi}_{n-1}}\cdots T_{\overline{\pi}_0}\eta_0 = \sup_{\pi_0,\ldots,\pi_{n-1}\in\Pi} FT_{\pi_{n-1}}\cdots T_{\pi_0}\eta_0.$$

It remains to prove Equation 12. We start by bounding the following quantity, for $n, k \geq 0$:

$$\|FT_{\overline{\pi}_n}^k\eta_{n+1} - FT_{\overline{\pi}_n}^k\eta_n\|_\infty.$$

For all $n, k \geq 0$ and $(s, c) \in \mathcal{S}\times\mathcal{C}$, we have

$$\begin{aligned}
&(FT_{\overline{\pi}_n}^k\eta_{n+1})(s,c) - (FT_{\overline{\pi}_n}^k\eta_n)(s,c)\\
&= (FT_{\overline{\pi}_n}^k T_*^n\eta_1)(s,c) - (FT_{\overline{\pi}_n}^k T_*^n\eta_0)(s,c)\\
&= \sup_{\pi_1,\ldots,\pi_n} (FT_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_1)(s,c) - \sup_{\pi_1',\ldots,\pi_n'} (FT_{\overline{\pi}_n}^k T_{\pi_1'}\cdots T_{\pi_n'}\eta_0)(s,c)\\
&\leq \sup_{\pi_1,\ldots,\pi_n} \Big((FT_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_1)(s,c) - (FT_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_0)(s,c)\Big)\\
&\leq \sup_{\pi_1,\ldots,\pi_n} \Big\|FT_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_1 - FT_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_0\Big\|_\infty\\
&\leq L\cdot \sup_{\pi_1,\ldots,\pi_n} \overline{\mathrm{w}}(T_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_1, T_{\overline{\pi}_n}^k T_{\pi_1}\cdots T_{\pi_n}\eta_0) && (F\ L\text{-Lipschitz})\\
&\leq L\gamma^{n+k}\overline{\mathrm{w}}(\eta_1, \eta_0) && (\gamma\text{-contraction})
\end{aligned}$$

and by a symmetric argument it also holds that for all $n, k \geq 0$ and $(s, c) \in \mathcal{S} \times \mathcal{C}$

$$(FT^k_{\overline{\pi}_n}\eta_{n+1})(s, c) - (FT^k_{\overline{\pi}_n}\eta_n)(s, c) \geq -L\gamma^{n+k}\overline{\mathrm{w}}(\eta_1, \eta_0).$$

so

$$\begin{aligned}\|FT^k_{\overline{\pi}_n}\eta_{n+1} - FT^k_{\overline{\pi}_n}\eta_n\|_\infty &\leq L\gamma^{n+k}\overline{\mathrm{w}}(\eta_1, \eta_0)\\ &\leq L\gamma^{n+k}\sup_{\pi\in\Pi}\overline{\mathrm{w}}(T_\pi\eta_0, \eta_0) \qquad (33)\end{aligned}$$

Recall that $\overline{\pi}_n$ realizes $T_*\eta_n$, so $T_{\overline{\pi}_n}\eta_n = \eta_{n+1}$. Then, for all $n \geq 0$, we have

$$\begin{aligned}\|F\eta^{\overline{\pi}_n} - F\eta_n\|_\infty &\leq \sum_{k=0}^{\infty}\|FT^{k+1}_{\overline{\pi}_n}\eta_n - FT^k_{\overline{\pi}_n}\eta_n\|_\infty && \text{(Telescoping and triangle inequality)}\\ &= \sum_{k=0}^{\infty}\|FT^k_{\overline{\pi}_n}\eta_{n+1} - FT^k_{\overline{\pi}_n}\eta_n\|_\infty && (T_{\overline{\pi}_n}\eta_n = \eta_{n+1})\\ &\leq \sum_{k=0}^{\infty}L\gamma^{n+k}\sup_{\pi\in\Pi}\overline{\mathrm{w}}(T_\pi\eta_0, \eta_0) && \text{(Equation 33)}\\ &= \frac{L\gamma^n}{1-\gamma}\sup_{\pi\in\Pi}\overline{\mathrm{w}}(T_\pi\eta_0, \eta_0).\end{aligned}$$

We have already established (in Equation 11) that

$$\sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi - F\eta_n \leq L\gamma^n\sup_{\pi\in\Pi_{\mathrm{M}}}\overline{\mathrm{w}}(\eta_0, \eta^\pi),$$

so

$$\begin{aligned}\sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi - F\eta^{\overline{\pi}_n} &= \sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi - F\eta_n + F\eta_n - F\eta^{\overline{\pi}_n}\\ &\leq L\gamma^n\sup_{\pi\in\Pi_{\mathrm{M}}}\overline{\mathrm{w}}(\eta, \eta^\pi) + \frac{L\gamma^n}{1-\gamma}\sup_{\pi\in\Pi}\overline{\mathrm{w}}(T_\pi\eta_0, \eta_0), \qquad \blacksquare\end{aligned}$$

A surprising technical detail about Theorem 6 is that distributional value iteration "works" (and $F\eta_n$ converges) under the given conditions, even though:

- T_* may not be a γ-contraction when $\gamma < 1$,
- T_* may not have a unique fixed point (for example, when multiple policies realize T_*),
- η_n may not converge (depending how ties are broken when realizing T_*),
- an optimal return distribution may not exist, that is, η^* such that $F\eta^* = \sup_{\pi\in\Pi_{\mathrm{H}}} F\eta^\pi$.

We can use the basic ideas from Theorem 6 so that distributional policy iteration also works under the same conditions as distributional value iteration. While distributional value iteration can start from any return distribution iterate $\eta \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$, for policy iteration we require the initial iterate to be a stationary policy $\pi_0 \in \Pi$, so that distributional policy improvement is guaranteed to work (see the discussion of Lemma 15).

Theorem 8 (Distributional Policy Iteration) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, for every stationary policy $\pi_0 \in \Pi$ if the iterates $\pi_1, \pi_2, \ldots$ satisfy (for $n \geq 0$)*

$$F_K T_{\pi_{n+1}} \eta^{\pi_n} = \sup_{\pi \in \Pi} F_K T_\pi \eta^{\pi_n} \qquad \text{(Distributional Policy Iterates)}$$

then the following hold.

Finite-horizon case: *For all n greater or equal to the horizon of the MDP,*

$$F_K \eta^{\pi_n} = \sup_{\pi \in \Pi_{\mathrm{H}}} F_K \eta^\pi. \tag{13}$$

Discounted case ($\gamma < 1$): *If K is L-Lipschitz, then for all $n \geq 0$*

$$\sup_{\pi \in \Pi_{\mathrm{H}}} F_K \eta^\pi - F_K \eta^{\pi_n} \leq L\gamma^n \cdot \sup_{\pi \in \Pi_{\mathrm{M}}} \overline{\mathrm{w}}(\eta^{\pi_0}, \eta^\pi), \tag{14}$$

Proof We use $F = F_K$. For any $n \geq 0$, we have that

$$\begin{aligned}
F\eta^{\pi_{n+1}} &= FT_{\pi_{n+1}} \eta^{\pi_{n+1}} && \text{(Distributional Bellman equation)} \\
&\geq FT_{\pi_{n+1}} \eta^{\pi_n} && \text{(Lemmas 14 and 15)} \\
&= FT_* \eta^{\pi_n} && \text{(Definition of } \pi_{n+1} \text{ and Definition 27)} \\
&\geq FT_*^{n+1} \eta^{\pi_0} && \text{(Induction)} \\
&= \sup_{\pi_1, \ldots, \pi_{n+1} \in \Pi} FT_{\pi_1} \cdots T_{\pi_{n+1}} \eta^{\pi_0}. && \text{(Definition 27)}
\end{aligned}$$

Then both Equations 13 and 14 follow by combining the above with Lemma 33 and Proposition 24, which ensures that

$$\sup_{\pi \in \Pi_{\mathrm{H}}} F\eta^\pi = \sup_{\pi \in \Pi_{\mathrm{M}}} F\eta^\pi,$$

(the conditions of Lemma 33 and Proposition 24 are satisfied). ■

Theorem 7 (Greedy Optimality) *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and indifferent to γ, and if: i) the MDP has finite horizon; or ii) $\gamma < 1$ and K is Lipschitz, then the following hold.*

There exists an optimal return distribution $\eta^* \in \mathcal{D}^{\mathcal{S} \times \mathcal{C}}$ satisfying

$$F_K \eta^* = \sup_{\pi \in \Pi_{\mathrm{H}}} F_K \eta^\pi,$$

iff the supremum on the right-hand side is attained (that is, an optimal policy exists).

If such η^* exists, then any greedy policy with respect to η^* is optimal (and thus attains the supremum above).

Proof We write $F = F_K$. We first prove the second statement, for which we assume that η^* as described exists. For every policy $\pi \in \Pi_{\mathrm{M}}$, we have $F\eta^* \geq F\eta^\pi$, so by monotonicity (Lemma 14), we also have, for all $\overline{\pi} \in \Pi$ and $\pi \in \Pi_{\mathrm{M}}$,

$$FT_{\overline{\pi}} \eta^* \geq FT_{\overline{\pi}} \eta^\pi,$$

so, for all $\overline{\pi} \in \Pi$,

$$FT_{\overline{\pi}}\eta^* \geq \sup_{\pi \in \Pi_{\mathrm{M}}} FT_{\overline{\pi}}\eta^{\pi}.$$

and thus

$$FT_*\eta^* = \sup_{\overline{\pi} \in \Pi} FT_{\overline{\pi}}\eta^* \geq \sup_{\overline{\pi} \in \Pi} \sup_{\pi \in \Pi_{\mathrm{M}}} FT_{\overline{\pi}}\eta^{\pi} = \sup_{\pi \in \Pi_{\mathrm{M}}} F\eta^{\pi}.$$

Now, let π^* be greedy with respect to η^*. Then

$$FT_{\pi^*}\eta^* = FT_*\eta^* = F\eta^*,$$

so Lemma 15 implies that $F\eta^{\pi^*} \geq F\eta^*$. The result then follows by using Proposition 24, for which the conditions are satisfied, and which states that

$$\sup_{\pi \in \Pi_{\mathrm{M}}} F\eta^{\pi} = \sup_{\pi \in \Pi_{\mathrm{H}}} F\eta^{\pi}.$$

For the first statement, under the assumption that the supremum is attained by a (possibly history-based) policy π^*, we can take $\eta^* = \eta^{\pi^*}$. If, on the other hand, we assume that η^* exists, then we have already shown that an optimal stationary policy exists, which implies that the supremum is attained. ■

Appendix C. Analysis of the Conditions for Distributional Dynamic Programming

C.1 Proofs

We start with some supporting results for the proof of Lemma 12, Item 2.

Proposition 34 *If $\nu \mapsto \mathbb{E}(f(G))$ (with $G \sim \nu$) is indifferent to γ, then for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, if $G \sim \nu$, $G' \sim \nu'$ and $\mathbb{E}(f(G)) = \mathbb{E}(f(G'))$: then $\mathbb{E}(f(\gamma G)) = \mathbb{E}(f(\gamma G'))$.*

Proof The result follows by applying indifference to γ in both directions: $\mathbb{E}(f(G)) \geq \mathbb{E}(f(G'))$ implies $\mathbb{E}(f(\gamma G)) \geq \mathbb{E}(f(\gamma G'))$, and $\mathbb{E}(f(G')) \geq \mathbb{E}(f(G))$ implies $\mathbb{E}(f(\gamma G')) \geq \mathbb{E}(f(\gamma G))$. ■

Lemma 35 *If $\nu \mapsto \mathbb{E}(f(G))$ (with $G \sim \nu$) is indifferent to γ, then there exists $\alpha > 0$ such that for all $c \in \mathcal{C}$*

$$f(\gamma c) = \alpha f(c) + (1 - \alpha) f(0). \tag{34}$$

Proof Assume $\nu \mapsto \mathbb{E}(f(G))$ (with $G \sim \nu$) is indifferent to γ.

Case 1: $f(0) = 0$. If $f(c) = 0$ for all c, then the result holds trivially (for example, we can take $\alpha = \gamma$. Otherwise, find $\overline{c} \in \mathcal{C}$ such that $f(\overline{c}) \neq 0$. We will first show that we can satisfy Equation 34 with $\alpha \doteq \frac{f(\gamma\overline{c})}{f(\overline{c})}$, and later show that $\alpha > 0$.

Fix $c \in \mathcal{C}$ arbitrary. If $f(c) = 0$, then, by Proposition 34, we have $f(\gamma c) = 0$ and Equation 34 holds for the chosen α. Let us consider the case where $f(c) \neq 0$.

If $\frac{f(c)}{f(\overline{c})} \leq 1$, we proceed as follows: Define ν, ν' such that $\nu(\overline{c}) \doteq \frac{f(c)}{f(\overline{c})}$, $\nu(0) \doteq 1 - \nu(\overline{c})$, $\nu'(c) \doteq 1$. Let $G \sim \nu$ and $G' \sim \nu'$. Then

$$\mathbb{E}(f(G)) = \frac{f(c)}{f(\overline{c})} f(\overline{c}) = f(c) = \mathbb{E}(f(G')).$$

By indifference to γ and Proposition 34, we have $\mathbb{E}(f(\gamma G)) = \mathbb{E}(f(\gamma G'))$, thus:

$$\frac{f(c)}{f(\overline{c})} f(\gamma\overline{c}) = f(\gamma c).$$

Rearranging, we get that

$$f(\gamma c) = \frac{f(\gamma\overline{c})}{f(\overline{c})} f(c),$$

which means we can satisfy Equation 34 with $\alpha = \frac{f(\gamma\overline{c})}{f(\overline{c})}$.

If $\frac{f(c)}{f(\overline{c})} > 1$, we proceed as follows: Define ν, ν' such that $\nu(c) \doteq \frac{f(\overline{c})}{f(c)}$, $\nu(0) \doteq 1 - \nu(c)$, $\nu'(\overline{c}) \doteq 1$. Let $G \sim \nu$ and $G' \sim \nu'$. Then

$$\mathbb{E}(f(G)) = \frac{f(\overline{c})}{f(c)} f(c) = f(\overline{c}) = \mathbb{E}(f(G')).$$

By indifference to γ and Proposition 34, we have $\mathbb{E}(f(\gamma G)) = \mathbb{E}(f(\gamma G'))$, thus:

$$\frac{f(\overline{c})}{f(c)} f(\gamma c) = f(\gamma\overline{c}).$$

Rearranging, we get that

$$f(\gamma c) = \frac{f(\gamma\overline{c})}{f(\overline{c})} f(c),$$

which means we can satisfy Equation 34 with $\alpha = \frac{f(\gamma\overline{c})}{f(\overline{c})}$.

We have established that Equation 34 holds for all $c \in \mathcal{C}$ with $\alpha = \frac{f(\gamma\overline{c})}{f(\overline{c})}$, provided that $f(0) = 0$. It only remains to show that $\alpha > 0$. If $f(c) > 0$, by indifference to γ we have $f(\gamma c) \geq f(0)$ (since $f(0) = 0$). Likewise, if $f(c) < 0$, then by indifference to γ we have $f(0) \geq f(\gamma c)$. In either case, $\alpha \geq 0$. Equation 34 with $c = \gamma^{-1} \cdot \overline{c}$ gives $f(\overline{c}) = \alpha f(\gamma^{-1} \cdot \overline{c})$, so $\alpha \neq 0$ (since we picked $\overline{c}$ such that $f(\overline{c}) \neq 0$). Thus, $\alpha > 0$.

Case 2: $f(0) \neq 0$. We can reduce this to the previous case with $f'(c) \doteq f(c) - f(0)$, so there exists $\alpha > 0$ such $f'(\gamma c) = \alpha f'(c)$ for all $c \in \mathcal{C}$, which means $f(\gamma c) - f(0) = \alpha f(c) - \alpha f(0)$, and rearranging gives Equation 34. ■

Lemma 12 (Conditions for Expected Utilities) *Let U_f be an expected utility, which is an objective functional F_K with $K\nu = \mathbb{E}f(G)$ ($G \sim \nu$). Then the following hold:*

1. *K is indifferent to mixtures.*

2. *K is indifferent to γ iff there exists $\alpha \in (0,1]$ such that $\gamma < 1 \Rightarrow \alpha < 1$ and, for all $c \in \mathcal{C}$,*
$$f(\gamma c) = \alpha f(c) + (1-\alpha) f(0). \tag{15}$$

3. *K is L-Lipschitz iff f is L-Lipschitz.*

Proof Item 1 follows essentially from the tower rule. Letting $G(s,c) \sim \eta(s,c)$ and $G'(s,c) \sim \eta'(s,c)$, we have $K(G(S,C)) = \mathbb{E}\left(\mathbb{E}\left(K(G(S,C))|S,C\right)\right)$. If $K\eta \geq K\eta'$, then

$$\begin{aligned}
K(G(S,C)) &= \mathbb{E} f(G(S,C)) \\
&= \mathbb{E}\left(\mathbb{E}\left(f(G(S,C))|S,C\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(K(G(S,C))|S,C\right)\right) \\
&\geq \mathbb{E}\left(\mathbb{E}\left(K(G'(S,C))|S,C\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(f(G'(S,C))|S,C\right)\right) \\
&= \mathbb{E} f(G'(S,C)) \\
&= K(G'(S,C)).
\end{aligned}$$

For Item 2, we first establish it for $\alpha > 0$, then we show that Equation 15 holds for some $\alpha \in (0,1]$ with $\gamma < 1 \Rightarrow \alpha < 1$.

Item 2 ($\Rightarrow$) follows from Lemma 35. To see the converse ($\Leftarrow$), we proceed as follows. Assume there exists $\alpha > 0$ such that for all $c \in \mathcal{C}$ Equation 15 holds and that $K(G(s,c)) \geq K(G'(s,c))$. Then

$$\begin{aligned}
K(\gamma G(s,c)) &= \mathbb{E} f(\gamma G(s,c)) \\
&= \alpha \mathbb{E} f(G(s,c)) + (1-\alpha) f(0) \\
&= \alpha K(G(s,c)) + (1-\alpha) f(0) \\
&\geq \alpha K(G'(s,c)) + (1-\alpha) f(0) \\
&= \alpha \mathbb{E} f(G'(s,c)) + (1-\alpha) f(0) \\
&= \mathbb{E} f(\gamma G'(s,c)) \\
&= K(\gamma G'(s,c)).
\end{aligned}$$

Now, define $g(c) \doteq f(c) - f(0)$, and assume Equation 15 holds for some $\alpha > 0$. If $\gamma = 1$ or f is constant, then Equation 15 holds trivially for $\alpha = \gamma$. Let us assume that $\gamma < 1$ and f is not constant. Then, by induction, we have, for all $n \in \mathbb{N}_0$, that $g(\gamma^n) = \alpha^n g(1)$, and

$$0 = \liminf_{n\to\infty} g(\gamma^n) = \liminf_{n\to\infty} \alpha^n g(1) = g(1) \cdot \lim_{n\to\infty} \alpha^n,$$

so we must have $\alpha < 1$.

For Item 3, we proceed as follows: If K is L-Lipschitz, then

$$L = \sup_{\substack{\nu,\nu': \\ \mathrm{w}(\nu)<\infty \\ \mathrm{w}(\nu')<\infty \\ \mathrm{w}(\nu,\nu')>0}} \frac{|K\nu - K\nu'|}{\mathrm{w}(\nu,\nu')} \geq \sup_{x \neq x'} \frac{|f(x) - f(x')|}{\mathrm{w}(\delta_x, \delta_{x'})} = \sup_{x \neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|_1},$$

which means f is L-Lipschitz. If f is L-Lipschitz, then, for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$,

$$\begin{aligned} |K\nu - K\nu'| &= |\mathbb{E}f(G) - \mathbb{E}f(G')| && (G \sim \nu,\, G' \sim \nu') \\ &= \inf\left\{|\mathbb{E}f(X) - \mathbb{E}f(X')| : \mathrm{df}(X) = \nu, \mathrm{df}(X') = \nu'\right\} \\ &\leq \inf\left\{\mathbb{E}|f(X) - f(X')| : \mathrm{df}(X) = \nu, \mathrm{df}(X') = \nu'\right\} \\ &\leq L \cdot \inf\left\{\|X - X'\|_1 : \mathrm{df}(X) = \nu, \mathrm{df}(X') = \nu'\right\} \\ &= L \cdot \mathrm{w}(\nu, \nu'), \end{aligned}$$

which means K is L-Lipschitz. ■

To get a better understanding of the limits of distributional DP, it is useful to inspect the necessary conditions for it to work. In the absence of indifference to mixtures or indifference to γ we can construct MDPs where greedy optimality (Theorem 7) fails due to a lack of monotonicity:

Proposition 13 *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is not indifferent to mixtures or not indifferent to γ, then there exists an MDP, an $\eta^* \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and a $\overline{\pi} \in \Pi$ such that $\overline{\pi}$ is greedy with respect to η^* and*

$$F_K\eta^* = \sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^\pi,$$

however, for some $(s, c) \in \mathcal{S} \times \mathcal{C}$

$$F_K\eta^{\overline{\pi}}(s, c) < \sup_{\pi\in\Pi_{\mathrm{H}}} F_K\eta^\pi(s, c).$$

Proof *Case 1: K is not indifferent to mixtures.* Consider $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ and a mixture distribution λ over $\mathcal{S} \times \mathcal{C}$ such that $K\eta \geq K\eta'$ but $K(G(S, C)) < K(G'(S, C))$ with $(S, C) \sim \lambda$, $G(s, c) \sim \eta(s, c)$, $G'(s, c) \sim \eta'(s, c)$.

Let $\gamma = 1$ and consider an MDP with state space $\{s_{\mathrm{init}}, s_{\mathrm{term}}\} \cup \mathcal{S}$ and action space $\{a, a'\}$ as follows: State s_{term} is terminal; either action in $(s_{\mathrm{init}}, 0)$ leads to (S, C) where $(S, C) \sim \lambda$ (in this case, the reward is C); action a on (s, c) leads to s_{term} with reward sampled according to $\eta(s, c)$; action a' on (s, c) leads to s_{term} with reward sampled according to $\eta'(s, c)$.

In this instance, there exists an optimal non-stationary policy $\pi_1^*\pi_2^*$ such that $\pi_2^*(a'|s, c) = 1$ for all $(s, c) \in \mathcal{S} \times \mathcal{C}$. Let η^* be the return distribution function for this policy. There exists a stationary policy $\overline{\pi} \in \Pi$ that is greedy respect to η^* and such that $\overline{\pi}(a|s, c) = 1$ for all $(s, c) \in \mathcal{S} \times \mathcal{C}$. Thus, letting $G(s, c) \sim \eta(s, c)$, $G'(s, c) \sim \eta'(s, c)$

$$K\eta^{\overline{\pi}}(s_{\mathrm{init}}, 0) = K(G(S, C)) < K(G'(S, C)) = K\eta^*(s_{\mathrm{init}}, 0),$$

which proves the result.

Case 2: K is not indifferent to γ. Consider ν, ν' for which $K\nu \geq K\nu'$ but $K(\gamma G) < K(\gamma G')$, with $G \sim \nu$, $G' \sim \nu'$.

Consider an MDP with $\mathcal{S} = \{s_{\mathrm{init}}, s_{\mathrm{mid}}, s_{\mathrm{term}}\}$ and $\mathcal{A} = \{a, a'\}$ as follows: State s_{init} is initial, state s_{term} terminal; state s_{init} transitions to state s_{mid} with either action and zero rewards; state s_{mid} transitions to state s_{term} with either action, but with reward distributed according to ν for a and ν' for a'.

There is an optimal non-stationary policy, corresponding to $\pi_1^*\pi_2^*$, where, for all $c \in \mathcal{C}$, $\pi_2^*(a'|s_{\text{mid}}, c) = 1$. Let η^* be the return distribution function for this policy. The stationary policy $\overline{\pi}$ that selects a always is greedy with respect to η^*, however

$$K\eta^{\overline{\pi}}(s_{\text{init}}, 0) = K(\gamma G) < K(\gamma G') = K\eta^*(s_{\text{init}}, 0),$$

which proves the result. ■

C.2 Exploring Lipschitz Continuity

We can use the examples in the second part of Table 1 to motivate why we may need Lipschitz continuity in the infinite-horizon setting. Neither $f(x) = \mathbb{I}(x > 0)$ nor $f(x) = -x^2$ are Lipschitz. $f(x) = \mathbb{I}(x > 0)$ is also not continuous, and it is informative to first consider how the lack of continuity can break distributional value/policy iteration.

Consider, by means of a counter-example, a single-state MDP with two actions $\{a_0, a_1\}$, $\gamma < 1$, and $r(a_i) = i$. The objective functional is U_f with $f(x) = \mathbb{I}(x > 0)$. Let π_i be the policy that always selects a_i. The return of π_i is deterministic and equal to $(1-\gamma)^{-1}i$. The policy π_1 and its return distribution η^{π_1} are optimal. The following is a valid greedy policy with respect to η^{π_1}:

$$\overline{\pi}(c) = \begin{cases} a_0 & c + (1-\gamma)^{-1}\gamma > 0 \\ a_1 & \text{otherwise.} \end{cases}$$

When starting from the stock $c = 0$, taking $\overline{\pi}$ for k steps followed by π_1 yields a return of $(1-\gamma)^{-1}\gamma^k > 0$ (since the first k actions are a_0). We know that the sequence $T_{\overline{\pi}}^1\eta^{\pi_1}, T_{\overline{\pi}}^2\eta^{\pi_1}, \ldots$ converges in supremum 1-Wasserstein distance to $T_{\overline{\pi}}^\infty\eta^{\pi_1} = \eta^{\overline{\pi}}$ (see Lemma 26). We also have that, for every $k \in \mathbb{N}$, $(U_f T_{\overline{\pi}}^k\eta^{\pi_1})(0) = 1$ and $(U_f\eta^{\pi_1})(0) = 1$, so $(U_f T_{\overline{\pi}}^k\eta^{\pi_1})(0) \geq (U_f\eta^{\pi_1})(0)$. However, the inequality fails in the limit: $(U_f T_{\overline{\pi}}^\infty\eta^{\pi_1})(0) = (U_f\eta^{\overline{\pi}})(0) = 0$, whereas $(U_f\eta^{\pi_1})(0) = 1$. For this reason, if π_0 is the chosen greedy policy with respect to η_1^π, then policy improvement (Lemma 15) fails, greedy optimality (Theorem 7) fails, distributional value iteration starting from $\eta^* = \eta^{\pi_1}$ fails, and distributional policy iteration starting from $\pi^* = \pi_1$ fails.

It is less clear how to design a counter-example when f is continuous but not Lipschitz, however we can show a case where where basic "evaluation" fails. Considering $f(x) = -x^2$, which is continuous but not Lipschitz, and the trivial MDP where $\mathcal{C} = \mathbb{R}$ and all rewards are zero. Consider the function $\eta_0 \doteq (s, c) \mapsto \delta_1$. This is not a value function in the MDP (no policy satisfies $\eta^\pi = \eta_0$), but we may want to use it for bootstrapping in distributional value iteration. In this particular MDP, T_* with $\gamma < 1$ is a contraction, since $\overline{\mathrm{w}}(T_*\eta, T_*\eta') \leq \gamma\overline{\mathrm{w}}(\eta, \eta')$, and the sequence $\eta_1, \eta_2, \ldots$ where $\eta_{n+1} = T_*\eta_n$ for $n \geq 0$ is Cauchy with respect to $\overline{\mathrm{w}}$, since $\overline{\mathrm{w}}(\eta_n, \eta_{n+k}) = \gamma^n(1-\gamma^k)$ for all $n, k \geq 0$. Therefore η_n converges to

$\eta_\infty = (s,c) \mapsto \delta_0$. However, letting $G_n(s,c) \sim \eta_n(s,c)$,

$$\begin{aligned}
\|U_f\eta_n - U_f\eta_{n+k}\|_\infty &= \sup_{s\in\mathcal{S},c\in\mathcal{C}} |\mathbb{E}f(c+G_n(s,c)) - \mathbb{E}f(c+G_{n+k}(s,c))| \\
&= \sup_{c\in\mathcal{C}} |(c+\gamma^n)^2 - (c+\gamma^{n+k})^2| \\
&= \sup_{c\in\mathcal{C}} |(2c+\gamma^n+\gamma^{n+k})(\gamma^n-\gamma^{n+k})| \\
&= \sup_{c\in\mathcal{C}} |(2c+\gamma^n+\gamma^{n+k})| \cdot \gamma^n \cdot (1-\gamma^k) \\
&= \infty,
\end{aligned}$$

which means the sequence $U_f\eta_n$ does not converge uniformly to $U_f\eta_\infty$ as $n \to \infty$. We have not been able to translate this failure of convergence to a failure of distributional DP, so it is unclear exactly what kind of convergence-related property of F_K is necessary for distributional DP to work in the infinite-horizon discounted case.

Appendix D. Proofs for Section 5.2

To prove Theorem 16, we follow the strategy used by Bäuerle and Ott (2011), where we reduce τ-CVaR optimization to solving the stock-augmented return distribution optimization problem with the expected utility U_f and $f(x) = x_-$, but where the starting stock c_0 must be chosen in a specific way as a function of s_0.

We start with a reduction of the τ-CVaR to an optimization problem, as shown in previous work, and some intermediate results.

Theorem 36 (Rockafellar et al., 2000) *For all $\nu \in (\Delta(\mathbb{R}), \mathrm{w})$ and $\tau \in (0,1)$,*

$$\mathrm{CVaR}(\nu,\tau) = \max_c \left(c + \frac{1}{\tau}\mathbb{E}(G-c)_- \right), \qquad (G \sim \nu)$$

and the maximum is attained at $\mathrm{QF}_\nu(\tau)$.

Proposition 37 *For all $s \in \mathcal{S}$, the function $c \mapsto -c + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s,c))_-$ is 1-Lipschitz.*

Proof Fix $s \in \mathcal{S}$ and let

$$g(c) \doteq -c + \frac{1}{\tau} \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c+G^\pi(s,c))_-.$$

For $\varepsilon \geq 0$, we have that

$$\begin{aligned}
\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c+G^\pi(s,c))_- &\leq \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c+\varepsilon+G^\pi(s,c))_- && ((x+\varepsilon)_- \geq x_-) \\
&= \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c+\varepsilon+G^\pi(s,c+\varepsilon))_-,
\end{aligned}$$

where the last line follows by noticing that the value in the stock augmentation does not change the supremum over history-based policies.

We can apply the same reasoning to see that

$$\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c-\varepsilon+G^\pi(s,c-\varepsilon))_- \leq \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c-\varepsilon+\varepsilon+G^\pi(s,c-\varepsilon+\varepsilon))_- = \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c+G^\pi(s,c))_-.$$

Thus for every $\varepsilon \geq 0$

$$\begin{aligned} g(c-\varepsilon) &= -(c-\varepsilon) + \frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c-\varepsilon+G^\pi(s,c-\varepsilon))_- \\ &\leq -c+\varepsilon+\frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c+G^\pi(s,c))_- \\ &= g(c)+\varepsilon, \end{aligned}$$

and

$$\begin{aligned} g(c+\varepsilon) &= -(c+\varepsilon) + \frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c+\varepsilon+G^\pi(s,c+\varepsilon))_- \\ &\geq -c-\varepsilon+\frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c+G^\pi(s,c))_- \\ &= g(c)-\varepsilon, \end{aligned}$$

That is:

$$g(c-\varepsilon)-\varepsilon \leq g(c) \leq g(c+\varepsilon)+\varepsilon \tag{35}$$

Thus, for $c, c' \in \mathbb{R}$, letting $c_{\max} = \max\{c, c'\}$ and $c_{\min} = \min\{c, c'\}$, we have

$$\begin{aligned} -(c_{\max}-c_{\min}) &\leq g(c_{\max}) - g(c_{\max}-(c_{\max}-c_{\min})) && \text{(Equation 35 with } \varepsilon = c_{\max}-c_{\min}) \\ &= g(c_{\max})-g(c_{\min}) \\ &= g(c_{\max})-g(c_{\min}+(c_{\max}-c_{\min})) \\ &\leq c_{\max}-c_{\min}, && \text{(Equation 35 with } \varepsilon = c_{\max}-c_{\min}) \end{aligned}$$

so

$$|g(c)-g(c')| = |g(c_{\max})-g(c_{\min})| \leq |c_{\max}-c_{\min}| = |c-c'|,$$

which means g is 1-Lipschitz. ■

Theorem 16 (Adapted from Bäuerle and Ott, 2011) *For every $\tau \in (0,1)$ and $s_0 \in \mathcal{S}$,*

$$\sup_{\pi\in\Pi_{\mathrm{H}}, c_0\in\mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0,c_0),\tau) = -c_0^* + \frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0^*+G^\pi(s_0,c_0^*))_-,$$

where c_0^* is the solution of

$$\max_{c_0}\left(-c_0+\frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0+G^\pi(s_0,c_0))_-\right). \tag{18}$$

Proof By Theorem 36, for all $s_0 \in \mathcal{S}$,

$$\begin{aligned}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathrm{CVaR}(\eta^\pi(s_0),\tau) &= \sup_{\pi\in\Pi_{\mathrm{H}}} \max_{c_0} \left(c_0 + \frac{1}{\tau}\mathbb{E}(G^\pi(s_0) - c_0)_- \right) \\ &= \sup_{c_0} \left(c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(G^\pi(s_0) - c_0)_- \right) \\ &= \sup_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0))_- \right) \\ &= \sup_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_- \right) \qquad \text{(Proposition 24)}\end{aligned}$$

It only remains to show that for all $s_0 \in \mathcal{S}$ there exists c_0^* that realizes the supremum over c_0. Note that by Assumption 1, we have

$$\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(G^\pi(s_0, c)) < \infty. \tag{36}$$

For all $s_0 \in \mathcal{S}$, we have

$$\begin{aligned}&\lim_{c_0\to\infty} -c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_- \\ &\le \lim_{c_0\to\infty} -c_0 \qquad \text{(Equation 36)} \\ &= -\infty.\end{aligned}$$

and

$$\begin{aligned}&\lim_{c_0\to-\infty} -c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_- \\ &= \lim_{c_0\to-\infty} \frac{1-\tau}{\tau} c_0 + \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(G^\pi(s_0, c_0)) - \mathbb{E}(c_0 + G^\pi(s_0, c_0))_+ \\ &\le \lim_{c_0\to-\infty} \frac{1-\tau}{\tau} c_0 + \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(G^\pi(s_0, c)) \\ &= -\infty. \qquad \text{(Equation 36)}\end{aligned}$$

Therefore there exist $c_{\min}, c_{\max} \in \mathbb{R}$ such that

$$\begin{aligned}&\sup_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_- \right) \\ &= \sup_{c_{\min}\le c_0\le c_{\max}} \left(-c_0 + \frac{1}{\tau} \sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_- \right).\end{aligned}$$

Moreover, Proposition 37 implies $c \mapsto -c + \frac{1}{\tau}\sup_{\pi\in\Pi_{\mathrm{H}}} \mathbb{E}(c + G^\pi(s_0, c))_-$ is continuous. Therefore the supremum over c_0 is attained at a maximizer $c_0^* \in \mathbb{R}$. ■

Theorem 17 *For every $\tau \in (0,1)$, $s_0 \in \mathcal{S}$ and $\varepsilon > 0$, there exists a stationary policy $\overline{\pi} \in \Pi$ (obtainable through distributional DP) and a $\overline{c}_0^*$ (obtainable through grid search) such that*

$$\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) - \mathrm{CVaR}(\eta^{\overline{\pi}}(s_0, \overline{c}_0^*), \tau) \leq 4\varepsilon.$$

In particular, $\overline{\pi}$ satisfies (for $f(x) = x_-$)

$$\sup_{\pi \in \Pi_{\mathrm{H}}} U_f \eta^\pi - U_f \eta^{\overline{\pi}} \leq \varepsilon,$$

and

$$\overline{c}_0^* = \arg\max_{c_0 \in \overline{\mathcal{C}}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_- \right), \tag{19}$$

where $\overline{\mathcal{C}} \doteq \{c_{\min} + i\varepsilon : i \in \mathbb{N}_0, c_{\min} + i\varepsilon \leq c_{\max}\}$ and $c_{\min}$ and $c_{\max}$ are chosen so that

$$\begin{aligned}&\max_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_- \right) \\ &= \max_{c_{\min} \leq c_0 \leq c_{\max}} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_- \right).\end{aligned}$$

Proof Let us fix τ, $s_0 \in \mathcal{S}$, $\varepsilon > 0$, $f(x) = x_-$, and define

$$g(c_0) \doteq -c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^\pi(s_0, c_0))_-.$$

Bäuerle and Ott (2011) (Theorem 16) established that

$$\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) = \sup_{c_0} g(c_0).$$

By using distributional DP (Theorems 6 and 8), we can find a near-optimal policy for optimizing U_f, that is, a $\overline{\pi}$ satisfying

$$\sup_{\pi \in \Pi_{\mathrm{H}}} U_f \eta^\pi - U_f \eta^{\overline{\pi}} \leq \varepsilon.$$

Let

$$\overline{g}(c_0) \doteq -c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_-.$$

Then $|g(c_0) - \overline{g}(c_0)| \leq \varepsilon$ for all $c_0 \in \mathcal{C}$. Moreover, by Proposition 37, g is 1-Lipschitz, so for all $c_0, c_0' \in \mathcal{C}$

$$|g(c_0) - g(c_0')| \leq |c_0 - c_0'|,$$

and

$$|\overline{g}(c_0) - \overline{g}(c_0')| \leq |\overline{g}(c_0) - g(c_0')| + |g(c_0) - g(c_0')| + |g(c_0') - \overline{g}(c_0')| \leq |c_0 - c_0'| + 2\varepsilon.$$

This means we can choose $c_{\min} \leq c_{\max}$ such that

$$\max_{c_0} \overline{g}(c_0) = \max_{c_{\min} \leq c_0 \leq c_{\max}} \overline{g}(c_0).$$

Define the grid $\overline{\mathcal{C}} \doteq \{c_{\min} + i\varepsilon : i \in \mathbb{N}_0, c_{\min} + i\varepsilon \leq c_{\max}\}$, Then

$$\begin{aligned}
\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{CVaR}(\eta^\pi(s_0, c_0), \tau) &= \sup_{c_0} g(c_0) && \text{(Theorem 16)} \\
&\leq \max_{c_0} \overline{g}(c_0) + \varepsilon \\
&= \max_{c_{\min} \leq c_0 \leq c_{\max}} \overline{g}(c_0) + \varepsilon \\
&\leq \sup_{c_{\min} \leq c_0 \leq c_{\max}} g(c_0) + 2\varepsilon \\
&\leq \max_{c_0 \in \overline{\mathcal{C}}} g(c_0) + 3\varepsilon \\
&\leq \max_{c_0 \in \overline{\mathcal{C}}} \overline{g}(c_0) + 4\varepsilon \\
&\leq \mathrm{CVaR}(\eta^{\overline{\pi}}(s_0, \overline{c}_0^*), \tau) + 4\varepsilon && \text{(Theorem 36)} \quad \blacksquare
\end{aligned}$$

Appendix E. Proofs for Section 5.3

Lemma 38 *For all $\nu \in (\Delta(\mathbb{R}), \mathrm{w})$ and $\tau \in (0,1)$,*

$$\mathrm{OCVaR}(\nu, \tau) = \min_c \left(c + \frac{1}{\tau}\mathbb{E}(G - c)_+ \right), \qquad (G \sim \nu)$$

and the minimum is attained at $\mathrm{QF}_\nu(\tau)$.

Proof The proof of this result is derived from the proof of Theorem 36 by Rockafellar et al. (2000). Fix $\nu \in (\Delta(\mathbb{R}), \mathrm{w})$ and $\tau \in (0,1)$ and let $G \sim \nu$ and $g(c) \doteq c + \frac{1}{\tau}\mathbb{E}(G - c)_+$. The function $x \mapsto x_+$ is convex, so for $c, c' \in \mathcal{C}$ and $\alpha \in [0,1]$,

$$\mathbb{E}(G - \alpha c - (1-\alpha)c')_+ \leq \alpha\mathbb{E}(G - c)_+ + (1-\alpha)\mathbb{E}(G - c')_+,$$

which means $g(\alpha c + (1-\alpha)c') \leq \alpha g(c) + (1-\alpha)g(c')$, that is, g is convex. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}c} g = 1 - \frac{1}{\tau}\mathbb{P}(G \geq c),$$

which means $\mathrm{QF}_\nu(1-\tau)$ is a minimizer of g. Finally, with $c^* = \mathrm{QF}_\nu(1-\tau)$, we have that

$$\begin{aligned}
\min_c g(c) &= g(c^*) \\
&= c^* + \frac{1}{\tau}\mathbb{E}(G - c^*)_+ \\
&= c^* + \frac{1}{\tau}\mathbb{E}\max\{G - c^*, 0\} \\
&= \frac{1}{\tau}c^* - \frac{1-\tau}{\tau}c^* + \frac{1}{\tau}\mathbb{E}\max\{G - c^*, 0\} \\
&= -\frac{1-\tau}{\tau}c^* + \frac{1}{\tau}\mathbb{E}\max\{G, c^*\} \\
&= -\frac{1-\tau}{\tau}c^* + \frac{1}{\tau}\int_0^1 \max\{\mathrm{QF}_\nu(t), c^*\}\mathrm{d}t \\
&= -\frac{1-\tau}{\tau}c^* + \frac{1-\tau}{\tau}c^* + \frac{1}{\tau}\int_{1-\tau}^1 \mathrm{QF}_\nu(t)\mathrm{d}t \\
&= \mathrm{OCVaR}(\nu, \tau). \qquad \blacksquare
\end{aligned}$$

Proposition 39 *For all $s \in \mathcal{S}$, the function $c \mapsto -c + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s,c))_+$ is 1-Lipschitz.*

Proof This proof is essentially the proof of Proposition 37 with x_+ instead of x_-. Fix $s \in \mathcal{S}$ and let

$$g(c) \doteq -c + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s,c))_+.$$

For $\varepsilon \geq 0$, we have that

$$\begin{aligned}
\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s,c))_+ &\leq \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + \varepsilon + G^\pi(s,c))_+ && ((x+\varepsilon)_+ \geq x_+) \\
&= \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + \varepsilon + G^\pi(s,c+\varepsilon))_+,
\end{aligned}$$

where the last line follows by noticing that the stock augmentation does not change the supremum over history-based policies. We can apply the same reasoning to see that

$$\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c-\varepsilon+G^\pi(s,c-\varepsilon))_+ \leq \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c-\varepsilon+\varepsilon+G^\pi(s,c-\varepsilon+\varepsilon))_+ = \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c+G^\pi(s,c))_+.$$

Thus for every $\varepsilon \geq 0$

$$\begin{aligned}
g(c-\varepsilon) &= -(c-\varepsilon) + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c - \varepsilon + G^\pi(s, c-\varepsilon))_+ \\
&\leq -c + \varepsilon + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s,c))_+ \\
&= g(c) + \varepsilon,
\end{aligned}$$

and

$$\begin{aligned} g(c+\varepsilon) &= -(c+\varepsilon) + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c+\varepsilon+G^{\pi}(s,c+\varepsilon))_{+} \\ &\geq -c-\varepsilon + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c+G^{\pi}(s,c))_{-} \\ &= g(c) - \varepsilon, \end{aligned}$$

That is:

$$g(c-\varepsilon) - \varepsilon \leq g(c) \leq g(c+\varepsilon) + \varepsilon \tag{37}$$

Thus, for $c, c' \in \mathbb{R}$, letting $c_{\max} = \max\{c, c'\}$ and $c_{\min} = \min\{c, c'\}$, we have

$$\begin{aligned} -(c_{\max} - c_{\min}) &\leq g(c_{\max}) - g(c_{\max} - (c_{\max} - c_{\min})) && \text{(Equation 37 with } \varepsilon = c_{\max} - c_{\min}) \\ &= g(c_{\max}) - g(c_{\min}) \\ &= g(c_{\max}) - g(c_{\min} + (c_{\max} - c_{\min})) \\ &\leq c_{\max} - c_{\min}, && \text{(Equation 37 with } \varepsilon = c_{\max} - c_{\min}) \end{aligned}$$

so

$$|g(c) - g(c')| = |g(c_{\max}) - g(c_{\min})| \leq |c_{\max} - c_{\min}| = |c - c'|,$$

which means g is 1-Lipschitz. ■

Theorem 18 *For every $\tau \in (0,1)$ and $s_0 \in \mathcal{S}$,*

$$\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) = -c_0^* + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0^* + G^{\pi}(s_0, c_0^*))_{+},$$

where c_0^* is the solution of

$$\min_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_{+} \right).$$

Proof Fix τ, $s_0 \in \mathcal{S}$, and $f(x) = x_+$. By Lemma 38, we have

$$\begin{aligned} &\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) \\ &= \sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \min_{c} \left(-c + \frac{1}{\tau} \mathbb{E}(c + G^{\pi}(s_0, c_0))_{+} \right) \\ &= \sup_{\pi \in \Pi_{\mathrm{H}}} \min_{c} \left(-c + \frac{1}{\tau} \mathbb{E}(c + G^{\pi}(s_0, c))_{+} \right). \end{aligned}$$

where in the last line we use the fact that the choice of c_0 is irrelevant since the supremum is over history-based policies.

For every $\varepsilon > 0$, by using distributional DP (Theorems 6 and 8), we can find a near-optimal policy for optimizing U_f, that is, a $\overline{\pi}$ satisfying

$$\sup_{\pi \in \Pi_{\mathrm{H}}} U_f \eta^{\pi} - U_f \eta^{\overline{\pi}} < \varepsilon.$$

$$\begin{aligned}
&\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) \\
&= \sup_{\pi \in \Pi_{\mathrm{H}}} \min_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \right) \\
&\geq \min_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\overline{\pi}}(s_0, c_0))_+ \right) \\
&> \inf_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \right) - \varepsilon.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) \\
&= \sup_{\pi \in \Pi_{\mathrm{H}}} \min_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \right) \\
&< \min_{c_0} \left(-c_0 + \frac{1}{\tau} \mathbb{E}(c_0 + G^{\pi'}(s_0, c_0))_+ \right) + \varepsilon \\
&\leq \inf_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \right) + \varepsilon
\end{aligned}$$

Since the above holds for all $\varepsilon > 0$, it means that

$$\sup_{\pi \in \Pi_{\mathrm{H}}, c_0 \in \mathcal{C}} \mathrm{OCVaR}(\eta^{\pi}(s_0, c_0), \tau) = \inf_{c_0} \left(-c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \right).$$

It only remains to show that for all $s_0 \in \mathcal{S}$ there exists c_0^* that realizes the infimum over c_0. Note that by Assumption 1, we have

$$\sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(G^{\pi}(s_0, c)) < \infty. \tag{38}$$

For all $s_0 \in \mathcal{S}$, we have

$$\begin{aligned}
&\lim_{c_0 \to -\infty} -c_0 + \frac{1}{\tau} \sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}(c_0 + G^{\pi}(s_0, c_0))_+ \\
&\leq \lim_{c_0 \to -\infty} -c_0 && \text{(Equation 38)} \\
&= \infty.
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{c_0\to\infty} -c_0 + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c_0 + G^\pi(s_0,c_0))_+ \\
&= \lim_{c_0\to\infty} \frac{1-\tau}{\tau}c_0 + \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(G^\pi(s_0,c_0)) - \mathbb{E}(c_0 + G^\pi(s_0,c_0))_- \\
&\geq \lim_{c_0\to\infty} \frac{1-\tau}{\tau}c_0 + \sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(G^\pi(s_0,c)) \\
&= \infty. && \text{(Equation 38)}
\end{aligned}$$

Therefore there exist $c_\mathrm{min}, c_\mathrm{max} \in \mathbb{R}$ such that

$$\begin{aligned}
&\inf_{c_0}\left(-c_0 + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c_0 + G^\pi(s_0,c_0))_+\right) \\
&= \inf_{c_\mathrm{min}\leq c_0\leq c_\mathrm{max}}\left(-c_0 + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c_0 + G^\pi(s_0,c_0))_+\right).
\end{aligned}$$

Moreover, Proposition 39 implies $c \mapsto -c + \frac{1}{\tau}\sup_{\pi\in\Pi_\mathrm{H}} \mathbb{E}(c + G^\pi(s_0,c))_+$ is continuous. Therefore the infimum over c_0 is attained at a minimizer $c_0^* \in \mathbb{R}$. ■

Theorem 19 *For every $\tau \in (0,1)$, $s_0 \in \mathcal{S}$ and $\varepsilon > 0$, there exists a stationary policy $\overline{\pi} \in \Pi$ (obtainable through distributional DP) and a $\overline{c}_0^*$ (obtainable through grid search) such that*

$$\sup_{\pi\in\Pi_\mathrm{H}, c_0\in\mathcal{C}} \mathrm{OCVaR}(\eta^\pi(s_0,c_0),\tau) - \mathrm{OCVaR}(\eta^{\overline{\pi}}(s_0,\overline{c}_0^*),\tau) \leq 4\varepsilon.$$

In particular, $\overline{\pi}$ satisfies (for $f(x) = x_+$)

$$\sup_{\pi\in\Pi_\mathrm{H}} U_f\eta^\pi - U_f\eta^{\overline{\pi}} \leq \varepsilon,$$

and

$$\overline{c}_0^* = \arg\min_{c_0\in\overline{\mathcal{C}}}\left(-c_0 + \frac{1}{\tau}\mathbb{E}(c_0 + G^{\overline{\pi}}(s_0,c_0))_+\right),$$

where $\overline{\mathcal{C}} \doteq \{c_\mathrm{min} + i\varepsilon : i \in \mathbb{N}_0, c_\mathrm{min} + i\varepsilon \leq c_\mathrm{max}\}$ and c_min and c_max are chosen so that

$$\begin{aligned}
&\min_{c_0}\left(-c_0 + \frac{1}{\tau}\mathbb{E}(c_0 + G^{\overline{\pi}}(s_0,c_0))_+\right) \\
&= \min_{c_\mathrm{min}\leq c_0\leq c_\mathrm{max}}\left(-c_0 + \frac{1}{\tau}\mathbb{E}(c_0 + G^{\overline{\pi}}(s_0,c_0))_+\right).
\end{aligned}$$

Proof The proof of this result is essentially the same as Theorem 17, except that we use Lemma 38, Proposition 39, and Theorem 18 instead of Theorems 16 and 36 and Proposition 37. ■

Appendix F. Proofs for Section 5.7

The proof of the first statement in Theorem 20 is relatively direct and self-contained; we show that from the designed reward we can construct a valid stock-augmented RL objective, where the designed rewards $\widetilde{R}_{t+1}$ satisfy a bounded first moment condition similar to Assumption 1, and with a designed discount $\alpha < 1$ in the infinite-horizon discounted ($\gamma < 1$) case.

However, the second statement—that only expected utilities that are indifferent to γ admit a reduction to a stock-augmented RL objective—requires multiple supporting results from the theory of optimizing expected utilities. This is because, when reducing a stock-augmented RL objective to a stock-augmented return distribution optimization objective, we need to make a statement about all the objectives F_K whose optimization is equivalent to a stock-augmented RL objective. In our case, this is possible, thanks to the von-Neumann-Morgenstern theorem (Von Neumann and Morgenstern, 2007) and the results from Bowling et al. (2023).

Without stock augmentation, for each state $s \in \mathcal{S}$, the preference over policies induced by value can be mapped to a relation $\succeq$ on $(\mathcal{D}, \mathrm{w})$. The von-Neumann-Morgenstern theorem (see Theorem 40 below) states that $\succeq$ is equivalent to an expected utility function iff $\succeq$ satisfies the von-Neumann-Morgenstern axioms (Axioms 1 to 4 below). Furthermore, any such expected utility function will be unique up to affine transformations. This uniqueness is powerful, because it implies that an objective cannot be simultaneously equivalent to an expected utility and a non-expected utility.

Axiom 1 (Completeness, adapted from Bowling et al., 2023) *For all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, $\nu \succeq \nu'$ or $\nu' \succeq \nu$ (or both, if $\nu \simeq \nu'$).*

Axiom 2 (Transitivity, adapted from Bowling et al., 2023) *For all $\nu, \nu', \nu'' \in (\mathcal{D}, \mathrm{w})$, if $\nu \succeq \nu'$ and $\nu' \succeq \nu''$, then $\nu \succeq \nu''$.*

Axiom 3 (Independence, adapted from Bowling et al., 2023) *For all $\nu, \nu', \overline{\nu} \in (\mathcal{D}, \mathrm{w})$, $\nu \succeq \nu'$ iff for all $p \in (0,1)$ $p\nu + (1-p)\overline{\nu} \succeq p\nu' + (1-p)\overline{\nu}$.*

Axiom 4 (Continuity, adapted from Bowling et al., 2023) *For all $\nu, \nu', \overline{\nu} \in (\mathcal{D}, \mathrm{w})$, if $\nu \succeq \overline{\nu} \succeq \nu'$ then there exists $p \in [0,1]$ such that $p\nu + (1-p)\nu' \simeq \overline{\nu}$.*

Theorem 40 (von Neumann-Morgenstern Expected Utility Theorem) *A preference relation $\succeq$ on $(\mathcal{D}, \mathrm{w})$ satisfies Axioms 1 to 4 if and only if there exists an expected utility function $u : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ such that*

1. *for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, $\nu \succeq \nu' \iff u(\nu) \geq u(\nu')$,*
2. *for all $\nu \in (\mathcal{D}, \mathrm{w})$, $u(\nu) = \mathbb{E}\left(u(\delta_G)\right)$ $(G \sim \nu)$.*

Such u is unique up to positive affine transformations.

The main result introduced by Bowling et al. (2023) establishes that every Markovian reward function induces a value function that is equivalent to a preference $\succeq$ satisfying Axioms 1 to 4 plus a fifth axiom called *Temporal Discount Indifference.* Their temporal discount indifference axiom allows the discount to be transition-dependent, but we are interested in making statements about RL objectives with a fixed discount, so we introduce an adaptation to this special case, which we refer to as *Fixed Discount Indifference.*

Axiom 5 (Fixed Discount Indifference) *There exists $\alpha \in (0, 1]$ such that for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, with $G \sim \nu$ and $G' \sim \nu'$,*

$$\frac{1}{1+\alpha}\mathrm{df}(\gamma G) + \frac{\alpha}{1+\alpha}\nu' \simeq \frac{1}{1+\alpha}\mathrm{df}(\gamma G') + \frac{\alpha}{1+\alpha}\nu.$$

Surprisingly, for relations $\succeq$ that satisfy Axioms 1 to 4 (and thus admit an equivalent expected utility u) we can show that $\succeq$ satisfies Axiom 5 iff u is indifferent to γ (cf. Definition 10). We can prove this correspondence between the two properties (Axiom 5 and Definition 10) by combining Lemma 12 Item 2 and the following novel result.[20]

Proposition 41 *Let $\succeq$ be a relation over $(\mathcal{D}, \mathrm{w})$, and let $u : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ be an expected utility function satisfying Theorem 40 Items 1 and 2. Axiom 5 holds iff for all $c \in \mathcal{C}$*

$$\alpha \cdot (u(\delta_c) - u(\delta_0)) = u(\delta_{\gamma c}) - u(\delta_0). \tag{39}$$

Proof Since u is linear, for $c \in \mathcal{C}$ we write $u(c) = u(\delta_c)$. We first prove the result under the assumption that $u(\delta_0) = 0$, in which case we want to show that $\alpha \cdot u(\delta_c) = u(\delta_{\gamma c})$. Axiom 5 states that there exists $\alpha \in (0, 1]$ such that for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, with $G \sim \nu$ and $G' \sim \nu'$,

$$\frac{1}{1+\alpha}\mathrm{df}(\gamma G) + \frac{\alpha}{1+\alpha}\nu' \simeq \frac{1}{1+\alpha}\mathrm{df}(\gamma G') + \frac{\alpha}{1+\alpha}\nu.$$

Since u is equivalent to the preference and linear, the above is equivalent to

$$\frac{1}{1+\alpha}\mathbb{E}u(\gamma G) + \frac{\alpha}{1+\alpha}u(\nu') = \frac{1}{1+\alpha}\mathbb{E}u(\gamma G') + \frac{\alpha}{1+\alpha}u(\nu).$$

Thus, by rearranging the above, Axiom 5 holds iff there exists $\alpha \in (0, 1]$ such that, for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$,

$$\mathbb{E}u(\gamma G) - \alpha \cdot u(\nu) = \mathbb{E}u(\gamma G') - \alpha \cdot u(\nu'). \tag{40}$$

Axiom 5 implies Equation 39. Using Equation 40 with $\nu = \delta_c$ and $\nu' = \delta_0$ gives

$$u(\gamma c) - \alpha \cdot u(c) = u(\delta_0) - \alpha \cdot u(\delta_0) = 0,$$

which gives the result.

Equation 39 implies Axiom 5. We have that for all $c, c' \in \mathcal{C}$

$$u(\gamma c) - \alpha \cdot u(c) = 0 = u(\gamma c') - \alpha \cdot u(c'),$$

and since this holds "pointwise", it also holds in expectation (with random C, C'), so Equation 40 follows.

Let us now prove the general case, $u(\delta_0) \in \mathbb{R}$. Let $u'(\nu) \doteq u(\nu) - u(\delta_0)$. We have already established that Axiom 5 holds iff $\alpha \cdot u'(\delta_c) = u'(\delta_{\gamma c})$ for all $c \in \mathcal{C}$, and expanding u' in terms of u gives Equation 39. ■

[20] Note how Equation 39 in Proposition 41 is the same condition as Equation 15 in Item 2 of Lemma 12.

We can now combine Axioms 1 to 5 and Theorem 40 into the core result for characterizing what objectives stock-augmented RL can optimize—an analogue of the main result of Bowling et al. (2023) (their Theorem 4.1).

As discussed earlier, in the standard case we use $\succeq$ to compare return distributions directly, so we can connect optimizing $\succeq$ to the RL problem by comparing return distributions of policies $\pi, \pi' \in \Pi_{\mathrm{H}}$ at states $s \in \mathcal{S}$ as $\eta^{\pi}(s) \succeq \eta^{\pi'}(s)$. Therefore, expected utilities that are equivalent to $\succeq$ are naturally $(\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ functions.

With stock augmentation, whether a return distribution ν is preferable to another ν' depends on the stock c, and we compare distributions of policies $\pi, \pi' \in \Pi_{\mathrm{H}}$ at stock-augmented states $(s, c) \in \mathcal{S} \times \mathcal{C}$ as $\mathrm{df}(c + G^{\pi}(s, c)) \succeq \mathrm{df}(c + G^{\pi'}(s, c))$. We can see this as there being a different preference relation for each c, so we define one expected utility $(\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ for each c, as is the case in Theorem 42 below.

The main contribution of Theorem 42 is that we can look at properties of a stock-indexed expected utility, and make statements about the corresponding relation per stock c, as we will discuss after presenting and proving Theorem 42.

Theorem 42 *A preference relation $\succeq$ on $(\mathcal{D}, \mathrm{w})$ satisfies Axioms 1 to 5 iff there exist stock-indexed functions $\widetilde{u}_c : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ (for all $c \in \mathcal{C}$), a stock-augmented reward function $\widetilde{r} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ and $\alpha \in (0, 1]$ such that:*

1. *for all $c, r', g \in \mathcal{C}$: $\widetilde{u}_c(\delta_{r'+\gamma g}) = \widetilde{r}(c, r') + \alpha \cdot \widetilde{u}_{\gamma^{-1}(c+r')}(\delta_g)$,*
2. *for all $c \in \mathcal{C}$ and $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$: $\mathrm{df}(c + G) \succeq \mathrm{df}(c + G')$ ($G \sim \nu$, $G' \sim \nu'$) iff $\widetilde{u}_c(\nu) \geq \widetilde{u}_c(\nu')$,*
3. *for all $c \in \mathcal{C}$ and $\nu \in (\mathcal{D}, \mathrm{w})$: $\widetilde{u}_c(\nu) = \mathbb{E}\left(\widetilde{u}_c(\delta_G)\right)$ ($G \sim \nu$).*

Proof This proof retraces the steps of the proof of Theorem 4.1 by Bowling et al. (2023).

Axioms 1 to 5 imply Items 1 to 3.

From the von Neumann-Morgenstern theorem (Theorem 40), we know that Axioms 1 to 4 imply the existence of a utility function $u : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ that is equivalent to the preference (Theorem 40 Item 1), linear (Theorem 40 Item 2) and unique up to positive affine transformations (Theorem 40).

We define, for $c \in \mathcal{C}$ and $\nu \in (\mathcal{D}, \mathrm{w})$, with $G \sim \nu$,

$$\widetilde{u}_c(\nu) \doteq u(\mathrm{df}(c + G)) - u(\delta_c) \tag{41}$$

and we will show that Items 1 to 3 hold. We also define the shorthand $f(c) \doteq u(\delta_c)$, and note that, for all $c, g \in \mathcal{C}$ we have $\widetilde{u}_c(\delta_g) = f(c + g) - f(c)$.

For Item 1, we define the reward function:

$$\widetilde{r}(c, r') \doteq \alpha f\left(\gamma^{-1}(c + r')\right) - f(c) + (1 - \alpha) f(0). \tag{42}$$

From Proposition 41, we get that for all $c, r', g \in \mathcal{C}$

$$\alpha\left(f\left(\gamma^{-1}(c + r' + \gamma g)\right) - f(0)\right) = f(c + r + \gamma g) - f(0),$$

which we can rearrange as

$$f(c + r + \gamma g) = \alpha f\left(\gamma^{-1}(c + r' + \gamma g)\right) + (1 - \alpha) f(0). \tag{43}$$

Thus, for all $c, r', g \in \mathcal{C}$,

$$\begin{aligned}
\widetilde{u}_c(r' + \gamma g) &= f(c + r' + \gamma g) - f(c) \\
&= \alpha f\left(\gamma^{-1}(c + r' + \gamma g)\right) + (1-\alpha) f(0) - f(c) && \text{(Equation 43)} \\
&= \alpha \cdot \widetilde{u}_{\gamma^{-1}(c+r')}(g) + \alpha f\left(\gamma^{-1}(c + r')\right) + (1-\alpha) f(0) - f(c) \\
&= \alpha f\left(\gamma^{-1}(c + r')\right) - f(c) + (1-\alpha) f(0) + \alpha \cdot \widetilde{u}_{\gamma^{-1}(c+r')}(g) && \text{(Rearranging)} \\
&= \widetilde{r}(c, r') + \alpha \cdot \widetilde{u}_{\gamma^{-1}(c+r')}(g), && \text{(Equation 42)}
\end{aligned}$$

which proves Item 1.

Item 2 follows from the fact that the preference induced by u is equivalent to $\succeq$, and $\widetilde{u}_c(\nu) = u(\mathrm{df}(c + G)) - u(\delta_c)$, so for all $c \in \mathcal{C}$ and $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, we have

$$\widetilde{u}_c(\nu) \geq \widetilde{u}_c(\nu') \iff u(\mathrm{df}(c+G)) \geq u(\mathrm{df}(c+G')) \iff \mathrm{df}(c+G) \succeq \mathrm{df}(c+G').$$

For Item 3, we proceed as follows. For all $c \in \mathcal{C}$ and $\nu \in (\mathcal{D}, \mathrm{w})$ (with $G \sim \nu$)

$$\begin{aligned}
\widetilde{u}_c(\nu) &= u(\mathrm{df}(c + G)) - u(\delta_c) && \text{(Equation 41)} \\
&= \mathbb{E}\left(u(\delta_{c+G})\right) - u(\delta_c) && \text{(Theorem 40 Item 2)} \\
&= \mathbb{E}\left(\widetilde{u}_c(\delta_G)\right), && \text{(Equation 41)}
\end{aligned}$$

which proves Item 3.

Axioms 1 to 5 follow from Items 1 to 3. Items 1 and 3 with $c = 0$ imply Items 1 and 2 of Theorem 40 with $u = \widetilde{u}_0$, which means $\succeq$ satisfies Axioms 1 to 4.

It remains only to show that $\succeq$ satisfies Axiom 5. By rearranging Item 1, we get that, for all $g \in \mathcal{C}$,

$$\widetilde{r}(0, 0) = \widetilde{u}_0(\delta_{\gamma g}) - \alpha \cdot \widetilde{u}_0(\delta_g).$$

In particular, by taking $g = 0$, we get that $\widetilde{r}(0,0) = (1-\alpha)\widetilde{u}_0(\delta_0)$. Thus, for any $g \in \mathcal{C}$, we have

$$\widetilde{u}_0(\delta_{\gamma c}) - \alpha \cdot \widetilde{u}_0(\delta_c) = (1-\alpha) \cdot \widetilde{u}_0(\delta_0),$$

and, by rearranging,

$$\alpha \cdot (\widetilde{u}_0(\delta_c) - \widetilde{u}_0(\delta_0)) = \widetilde{u}_0(\delta_{\gamma c}) - \widetilde{u}_0(\delta_0),$$

so we can satisfy Equation 39 with $u = \widetilde{u}_0$, and, by Proposition 41, $\succeq$ satisfies Axiom 5. ■

This is how we will use Theorem 42 to prove the second statement in Theorem 20: We will show that value in stock-augmented RL is, in effect, a stock-indexed expected utility, so the stock-indexed corresponding relations satisfy Axioms 1 to 5. If this stock-augmented RL objective is equivalent to a stock-augmented return distribution optimization objective F_K, then (we show) K must be equivalent to the stock-indexed utility corresponding to value. Then we combine Theorem 42, Proposition 41, and Lemma 12 to show that K must be both an expected utility and indifferent to γ.

We are now ready to present the proof of Theorem 20.

Theorem 20 *A stock-augmented return distribution optimization objective functional U_f can be reduced to an equivalent stock-augmented reinforcement learning objective (expected return) with discount $\alpha \in (0,1]$ with $\gamma < 1 \Rightarrow \alpha < 1$ and reward proportional to*

$$\widetilde{R}_{t+1} \doteq \alpha f(C_{t+1}) - f(C_t) + (1-\alpha) f(0) \tag{20}$$

if f satisfies, for all $c \in \mathcal{C}$,

$$f(\gamma c) = \alpha f(c) + (1-\alpha) f(0), \tag{21}$$

and:

- *in the finite-horizon case,*

$$\sup_{s,c,a \in \mathcal{S}\times\mathcal{C}\times\mathcal{A}} \mathbb{E}\left(|\widetilde{R}_{t+1}| \,\middle|\, S_t = s, C_t = c, A_t = a\right) < \infty; \tag{22}$$

- *in the discounted case, f is Lipschitz.*

A stock-augmented return distribution optimization objective that is not an expected utility or not indifferent to γ cannot be reduced via reward design to a stock-augmented reinforcement learning objective.

Proof *Reduction from a stock-augmented return distribution optimization objective to a stock-augmented RL objective.* The stock-augmented RL objective we want to reduce to is an expected return where the (designed) rewards have bounded first moment ($\widetilde{R}_{t+1}$ satisfying Equation 22), the discount is $\alpha \in (0,1]$ (where $\gamma < 1 \Rightarrow \alpha < 1$), and policies $\pi \in \Pi_{\mathrm{H}}$ have value function

$$\widetilde{V}^\pi(s,c) \doteq \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \widetilde{R}_{t+1}\right).$$

We will show that, under the given conditions, for all $\pi \in \Pi_{\mathrm{H}}$ and $(s,c) \in \mathcal{S}\times\mathcal{C}$,

$$\widetilde{V}^\pi(s,c) = (U_f \eta^\pi)(s,c) - f(c) = \mathbb{E} f(c + G^\pi(s,c)) - f(c), \tag{44}$$

with $G(s,c) \sim \eta^\pi(s,c)$. If this is the case, then both stock-augmented objectives induce the same preference over policies.

Let us first establish that, under the given conditions, the designed rewards have bounded first moment. In the finite-horizon case we have imposed Equation 22 as a condition directly. In the discounted case, f is assumed to be L-Lipschitz for some L, so:

$$\begin{aligned}
|\widetilde{R}_{t+1}| &= |\alpha f(C_{t+1}) - f(C_t) + (1-\alpha) f(0)| \\
&= |f(\gamma C_{t+1}) - f(C_t)| && \text{(Equation 21)} \\
&= |f(C_t + R_{t+1}) - f(C_t)| && (C_{t+1} = \gamma^{-1}(C_t + R_{t+1})) \\
&= L \cdot \|R_{t+1}\|_1, && (f\ L\text{-Lipschitz})
\end{aligned}$$

and, by Assumption 1,

$$\sup_{s,c,a \in \mathcal{S}\times\mathcal{C}\times\mathcal{A}} \mathbb{E}\left(|\widetilde{R}_{t+1}| \,\middle|\, S_t = s, C_t = c, A_t = a\right) \leq \sup_{s,a \in \mathcal{S}\times\mathcal{A}} \mathbb{E}\left(\|R_{t+1}\|_1 \,|\, S_t = s, A_t = a\right) < \infty.$$

Next, we establish that $\gamma < 1 \Rightarrow \alpha < 1$ (that is, the α-discounting is valid for the infinite-horizon discounted case). By induction on Equation 21, we get that, for all $n \in \mathbb{N}_0$ and $c \in \mathcal{C}$, that $f(\gamma^n c) - f(0) = \alpha^n(f(c) - f(0))$, which we can rearrange as

$$f(\gamma^n c) = \alpha^n f(c) + (1 - \alpha^n) f(0). \tag{45}$$

In particular, for all $c \in \mathcal{C}$,

$$\liminf_{n\to\infty} f(\gamma^n c) = \liminf_{n\to\infty} \alpha^n f(c) + (1 - \alpha^n) f(0).$$

If $\gamma < 1$, the left-hand side is zero, so the right-hand side must be zero, thus $\alpha < 1$.

Finally, we prove Equation 44. For all $\pi \in \Pi_{\mathrm{H}}$ and $(s, c) \in \mathcal{S} \times \mathcal{C}$, with $(S_0, C_0) = (s, c)$ with probability one, we have

$$\begin{aligned}
\widetilde{V}^\pi(s, c) &= \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \widetilde{R}_{t+1} \middle| C_0 = c\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(\sum_{t=0}^{n-1} \alpha^t \widetilde{R}_{t+1} \middle| C_0 = c\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(\sum_{t=0}^{n-1} \alpha^{t+1} f(C_{t+1}) - \alpha^t f(C_t) + \alpha^t (1-\alpha) f(0) \middle| C_0 = c\right) && \text{(Equation 20)} \\
&= \lim_{n\to\infty} \mathbb{E}\left(\sum_{t=0}^{n-1} \alpha^{t+1} f(C_{t+1}) - \alpha^t f(C_t) \middle| C_0 = c\right) + (1 - \alpha^n) f(0) \\
&= \lim_{n\to\infty} \mathbb{E}\left(\alpha^n f(C_n) \middle| C_0 = c\right) - f(c) + (1 - \alpha^n) f(0) && \text{(Telescoping, } C_0 = c\text{)} \\
&= \lim_{n\to\infty} \mathbb{E}\left(\alpha^n f(C_n) + (1 - \alpha^n) f(0) \middle| C_0 = c\right) - f(c) \\
&= \lim_{n\to\infty} \mathbb{E}\left(f\left(\gamma^n C_n\right) \middle| C_0 = c\right) - f(c) && \text{(Equation 45)} \\
&= \lim_{n\to\infty} \mathbb{E}\left(f\left(C_0 + \sum_{t=0}^{n-1} \gamma^t R_{t+1}\right) \middle| C_0 = c\right) - f(c) \\
&= \mathbb{E}\left(f\left(C_0 + \sum_{t=0}^{\infty} \gamma^t R_{t+1}\right) \middle| C_0 = c\right) - f(c) && (f \text{ Lipschitz or finite horizon}) \\
&= (U_f \eta^\pi)(s, c) - f(c),
\end{aligned}$$

which proves Equation 44 and concludes the proof of the first statement.

Impossible reduction via reward design when the objective F_K is not an expected utility or not indifferent to γ. We will show the contrapositive: If the reduction is possible, then F_K is an expected utility and indifferent to γ.

Assume we can reduce it to an equivalent stock-augmented RL objective with a suitably designed reward function. It is important to stress that the reduction must be valid regardless of the underlying MDP transition or reward kernels, as long as Assumption 1 is satisfied.

Let us define the "stock-indexed value functional" $\widetilde{v}_c : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ as follows: For a Markov chain $C_0 \to R_1 \to R_2 \to \ldots$ all taking values in $\mathcal{C}$ (and satisfying Assumption 1) with $G_0 \doteq \sum_{t=0}^{\infty} \gamma^t R_{t+1}$ and $C_0 = c$, we let

$$\widetilde{v}_c(\mathrm{df}(G_0)) \doteq \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \cdot \widetilde{r}(C_t, R_{t+1})\right),$$

where $\widetilde{r} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ is the designed (Markov) reward.

The value functional and the designed reward function do not directly depend on states and actions. This is natural, as trajectories with the same $C_0 \to R_1 \to R_2 \to \ldots$, regardless of the underlying $S_0, A_0, S_1, \ldots$, must be equivalent in terms of the objective (either return distribution optimization or RL).

The reduction requires the same augmented state space $\mathcal{S} \times \mathcal{C}$ to be used for both return distribution optimization and RL objectives, so, for all $(c, \nu), (c', \nu') \in \mathcal{C} \times (\mathcal{D}, \mathrm{w})$ we have $K\mathrm{df}(c + G) \geq K\mathrm{df}(c' + G') \iff \widetilde{v}_c(\nu) \geq \widetilde{v}_{c'}(\nu')$, with $G \sim \nu$ and $G' \sim \nu'$.

We can now apply Theorem 42 to conclude that the relation induced by K on $(\mathcal{D}, \mathrm{w})$ must satisfy Axioms 1 to 5. To do so, we must prove that Items 1 to 3 hold for $\widetilde{v}_c$ and $\widetilde{r}$.

For Item 1, consider $C_0 = c$, $R_1 = r_1$ and so forth, with probability one, such that $g_1 \doteq \sum_{t=1}^{\infty} \gamma^t r_{t+2} < \infty$. Then

$$\begin{aligned}
\widetilde{v}_c(\delta_{r_1 + \gamma g_1}) &= \widetilde{v}_c(\mathrm{df}(G_0)) \\
&= \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \cdot \widetilde{r}(C_t, R_{t+1})\right) \\
&= \mathbb{E}\left(\widetilde{r}(C_0, R_1) + \alpha \cdot \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \cdot \widetilde{r}(C_{t+1}, R_{t+2}) \middle| C_1\right)\right) \\
&= \mathbb{E}\left(\widetilde{r}(C_0, R_1) + \alpha \cdot \widetilde{v}_{C_1}(\mathrm{df}(G_1))\right) \\
&= \widetilde{r}(c, r_1) + \alpha \cdot \widetilde{v}_{\gamma^{-1}(c + r_1)}(\delta_{g_1}),
\end{aligned}$$

which gives us Item 1.

Item 2 follows by assumption that the reduction is possible.

Item 3 can be proved as follows: For all $c \in \mathcal{C}$, with $C_0 = c$ and $C_0 \to R_1 \to R_2 \to \ldots$ satisfying Assumption 1:

$$\begin{aligned}
\widetilde{v}_c(\mathrm{df}(G_0)) &= \mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \cdot \widetilde{r}(C_t, R_{t+1})\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{t=0}^{\infty} \alpha^t \cdot \widetilde{r}(C_t, R_{t+1}) \middle| C_0, R_1, R_2, \ldots\right)\right) \\
&= \mathbb{E}\left(\widetilde{v}_c(\delta_{G_0})\right).
\end{aligned}$$

Hence, by Theorem 42, the relation induced by K on $(\mathcal{D}, \mathrm{w})$ satisfies Axioms 1 to 5, which implies that K is an expected utility.

We know from Theorem 40 that there exist $a > 0$ and $b \in \mathbb{R}$ such that, for all $c, g \in \mathcal{C}$, we have $aK\delta_{c+g} + b = \widetilde{v}_c(\delta_g)$. So define $f(c) \doteq aK\delta_c + b$. Then, for all $c \in \mathcal{C}$,

$$\begin{aligned} a \cdot K\delta_{\gamma c} + b &= f(\gamma c) \\ &= \widetilde{v}_0(\delta_{\gamma c}) \\ &= \widetilde{r}(0,0) + \alpha \cdot \widetilde{v}_0(\delta_c) \\ &= \widetilde{r}(0,0) + \alpha f(c). \end{aligned}$$

In particular, for $g = 0$, the above implies that $\widetilde{r}(0,0) = (1-\alpha)f(0)$, so, for all $c \in \mathcal{C}$,

$$f(\gamma c) = \alpha f(c) + (1-\alpha)f(0).$$

The assumption that the reduction is possible ensures that $\alpha \in (0,1]$ and $\gamma < 1 \Rightarrow \alpha < 1$, so, by Lemma 12 Item 2, K is indifferent to γ. ■

Appendix G. Proofs for Section 5.8

Our characterization builds on and extends the results by Marthe et al. (2024), which characterized objective functionals that distributional DP can optimize in the finite-horizon undiscounted setting, *without* stock augmentation. Our proof strategy is to connect indifference to mixtures, indifference to γ and Lipschitz continuity to the von Neumann-Morgenstern axioms (from Appendix F), so that we can apply the powerful von Neumann-Morgenstern theorem (or show that it cannot apply, in the case of the non-expected-utility objective functional that distributional DP can optimize).

The following results connect Lipschitz continuity and indifference to mixtures to the von Neumann-Morgenstern independence axiom (Axiom 3).

Proposition 43 (If K Lipschitz then Axiom 3's $\Leftarrow$ is satisfied.) *If K is Lipschitz, the following holds: For every $\nu, \nu', \overline{\nu} \in (\mathcal{D}, \mathrm{w})$ if for all $p \in (0,1)$ we have*

$$K((1-p)\nu + p\overline{\nu}) \geq K((1-p)\nu' + p\overline{\nu}),$$

then

$$K\nu \geq K\nu'.$$

Proof Fix $\nu, \nu', \overline{\nu} \in (\mathcal{D}, \mathrm{w})$ and assume that for all $p \in (0,1)$ we have

$$K((1-p)\nu + p\overline{\nu}) \geq K((1-p)\nu' + p\overline{\nu}).$$

Define the sequences of distributions

$$\begin{aligned} \nu_n &\doteq \frac{1}{n}\overline{\nu} + \left(1 - \frac{1}{n}\right)\nu \\ \nu_n' &\doteq \frac{1}{n}\overline{\nu} + \left(1 - \frac{1}{n}\right)\nu'. \end{aligned}$$

We have that ν_n converges to ν in w as $n \to \infty$ (and ν_n' to ν'). Because K is Lipschitz, and by assumption $K\nu_n - K\nu_n' \geq 0$ for all $n \in \mathbb{N}$, we get

$$K\nu - K\nu' = \lim_{n\to\infty} K\nu_n - K\nu_n' \geq 0. \qquad \blacksquare$$

Proposition 44 (If K is indifferent to mixtures, then Axiom 3's $\Rightarrow$ is satisfied.) *If K is indifferent to mixtures, then the following holds: For every $\nu, \nu', \overline{\nu} \in (\mathcal{D}, \mathrm{w})$ if*

$$K\nu \geq K\nu',$$

then for all $p \in (0,1)$ we have

$$K((1-p)\nu + p\overline{\nu}) \geq K((1-p)\nu' + p\overline{\nu}).$$

Proof Definition 9 with $\nu_1, \nu_2, \nu_1', \nu_2'$ such that $K\nu_1 \geq K\nu_1'$ and $\nu_2' = \nu_2$, gives us that for all $p \in (0,1)$

$$K\nu \geq K\nu' \Rightarrow K((1-p)\nu + p\overline{\nu}) \geq K((1-p)\nu' + p\overline{\nu}). \qquad \blacksquare$$

Next, we apply the von Neumann-Morgenstern theorem to characterize objective functionals that distributional DP can optimize in the infinite-horizon discounted case.

Theorem 21 *If $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ is indifferent to mixtures and Lipschitz, then F_K is an expected utility, that is, there exists an $f : \mathcal{C} \to \mathbb{R}$ such that $K\nu = \mathbb{E}f(G)$ ($G \sim \nu$) and f is Lipschitz.*

Proof Consider the relation $\succeq$ over $(\mathcal{D}, \mathrm{w})$ defined by $\nu \succeq \nu' \iff K\nu \geq K\nu'$. It is easy to see that $\succeq$ satisfies completeness and transitivity (Axioms 1 and 2 in Appendix F). K Lipschitz implies that $\succeq$ also satisfies continuity (Axiom 4). K Lipschitz and K indifferent to mixtures implies that K satisfies Axiom 3 (Propositions 43 and 44).

Then by the von Neumann-Morgenstern theorem (Theorem 40) there exists an expected utility function $u : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ satisfying Items 1 and 2, and it is unique up to affine transformations. By Item 1, for all $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$, with $G \sim \nu$ and $G' \sim \nu'$, we have $\nu \succeq \nu' \iff u(\nu) \geq u(\nu)$, and thus $K\nu \geq K\nu' \iff u(\nu) \geq u(\nu)$. Moreover, by Theorem 40, we know u is unique up to positive affine transformations, so there exist $a > 0$ and $b \in \mathbb{R}$ such that $K\nu = a \cdot u(\nu) + b$ for all $\nu \in (\mathcal{C}, \mathrm{w})$. Without loss of generality we can consider u in the rest of this proof such that $a = 1$ and $b = 0$. Since u is linear, we know there exists $f : \mathcal{C} \to \mathbb{R}$ such that $u(\nu) = \mathbb{E}f(G)$ ($G \sim \nu$) for all $\nu \in (\mathcal{C}, \mathrm{w})$. The statement that f is Lipschitz follows from Lemma 12. $\blacksquare$

Proposition 22 *The statistical functional $K : (\mathcal{D}, \mathrm{w}) \to \mathbb{R}$ satisfying, for $\nu \in (\mathcal{D}, \mathrm{w})$,*

$$K\nu = \mathbb{I}(\nu([0,\infty)) = 1)$$

is indifferent to mixtures and F_K is not an expected utility.

Proof *K is indifferent to mixtures.* Consider $\eta, \eta' \in (\mathcal{D}^{\mathcal{S}\times\mathcal{C}}, \overline{\mathrm{w}})$ such that, for all $(s,c) \in \mathcal{S}\times\mathcal{C}$,

$$K\eta(s,c) \geq K\eta'(s,c), \tag{46}$$

and let (S,C) be a random variable taking values in $\mathcal{S}\times\mathcal{C}$, $\nu \doteq \mathrm{df}(G(S,C))$ and $\nu' \doteq \mathrm{df}(G'(S,C))$.

Equation 46 implies that $\{\nu'(S,C)([0,\infty)) = 1\} \subseteq \{\nu(S,C)([0,\infty)) = 1\}$, which in turn implies that

$$\mathbb{I}(\nu'(S,C)([0,\infty)) = 1) \leq \mathbb{I}(\nu(S,C)([0,\infty)) = 1),$$

which proves the result.

K is indifferent to γ. Given $\nu, \nu' \in (\mathcal{D}, \mathrm{w})$ and letting $G \sim \nu$ and $G' \sim \nu'$, note that $\nu([0,\infty)) = 1 \iff \mathrm{df}(\gamma G)([0,\infty)) = 1$ (and similarly for ν' and G'), so $K(\gamma G) = K\nu$ and $K(\gamma G') = K\nu'$, which means $K\nu \geq K\nu'$ implies $K(\gamma G) \geq K(\gamma G')$.

F_K is not an expected utility. It suffices to show that K violates at least one of the von Neumann-Morgenstern axioms, otherwise Theorem 40 applies and F_K is an expected utility. K invariably satisfies completeness and transitivity (Axioms 1 and 2), however it violates independence and continuity (Axioms 3 and 4; cf. Juan Carreño, 2020, p. 15). ■

Appendix H. Implementation details

H.1 DηN

The architecture diagram for DηN's stock-augmented return distribution estimator is given in Figure 1. The training and network parameters were set per domain (see Appendices H.2 and H.3). The target parameters $\overline{\theta}$ were updated via exponential moving average updates, as done by Schwarzer et al. (2023), and differently from the periodic updates used by Mnih et al. (2015). Our intent was to have smoother quantile regression targets, rather than sudden changes introduced by the periodic update. The target network is updated as an exponential moving average with step size α as $\overline{\theta} \leftarrow (1-\alpha)\overline{\theta} + \alpha\theta$. D$\eta$N uses the target network parameters $\overline{\theta}$ for both training and evaluation (similar to Abdolmaleki et al., 2018). Our intent was to slower-changing behavior and quantile regression targets.

As in DQN (Mnih et al., 2015) and QR-DQN (Dabney et al., 2018), the action selection used by DηN during data collection is ε-greedy. For greedy policy selection during both data generation (Equation 27) and learning (Equation 26), given a return distribution function $\xi : \mathcal{S}\times\mathcal{C}\times\mathcal{A} \to \mathcal{D}$, D$\eta$N selects the greedy policy $\overline{\pi} \in \Pi$ satisfying

$$U_f(M_f\xi)(s,c) = \mathbb{E}f(c + G(s,c,A)) \qquad\qquad (A \sim \overline{\pi}(s,c),\, G(s,c,a) \sim \xi(s,c,a))$$

and, for all $(s,c) \in \mathcal{S}\times\mathcal{C}$ and $a \in \mathcal{A}$,

$$\overline{\pi}(a|s,c) > 0 \Rightarrow \overline{\pi}(a|s,c) = \max_{a'} \overline{\pi}(a'|s,c).$$

We chose this because ties may happen often in return distribution optimization. This is not the case in standard deep RL with DQN, and we rarely need to resort to tie-breaking, because action-value estimates are often noisy. However, the choice of U_f may introduce ties in practice. For example, when maximizing the risk-averse τ-CVaR, we have $f(x) = x_-$, which can introduce ties among maximizing actions.

With vector-valued returns, DηN maintains estimates of the quantiles each individual return coordinate, rather than an estimate of the joint distribution of the vector-valued return. This means we cannot optimize all expected utilities over vector-valued returns, but only the ones with the form:

$$f(x) = \sum_i f_i(x_i).$$

We believe this is acceptable for a proof-of-concept algorithm, and that future work will address this limitation based on results for multivariate distributional RL (Zhang et al., 2021; Wiltzer et al., 2024).

For the quantile regression loss, the greedy policy $\overline{\pi}$ breaks ties via uniform random action selection, but to avoid having to sample multiple actions from $\overline{\pi}$ we use the policy directly. For a transition $(s, c), a, r', (s', c')$, the loss estimate is:

$$\frac{1}{n^2} \sum_{i,j \in \{1,\ldots,n\}} \sum_{a' \in \mathcal{A}} \overline{\pi}(a'|s, c) \ell(r' + \gamma \xi_{\overline{\theta}}(s', a', c')_j - \xi_\theta(s, a, c)_i, \tau_i),$$

where ℓ is the quantile regression loss (Dabney et al., 2018)

$$\ell(x, \tau) \doteq |\mathbb{I}(x > 0) - \tau| \cdot |x|,$$

and the quantiles are the bin centers of an n-bin discretization of $[0, 1]$, that is, for $i \in \{1, \ldots, n\}$ we have $\tau_i \doteq \frac{2i-1}{2n}$. As in DQN (Mnih et al., 2015) and QR-DQN (Dabney et al., 2018), we explicitly use δ_0 as the return distribution of the terminal state.

H.2 Gridworld

In these experiments we trained DηN on an Nvidia V100 GPU. For simplicity, DηN did not use a replay in these experiments. Instead, it alternated generating a minibatch of transitions by having the agent interact with the environment, and then updating the network with the generated minibatch (the "learner update"). The transitions were generated in episodic fashion, with the agent starting at s_{init} and acting in the environment until the end of the episode. The episode ended when the agent reached a terminating cell, or when it was interrupted on the 16-th step. Upon interruption, s' was not treated as terminal. Each minibatch consisted of 64 trajectories of length 16, and each transition had the form $(s_k, c_k), a_k, r'_k, (s'_k, c'_k)$. If a termination or interruption happened at the k-th step in a trajectory, the next transition would start from the initial state, in which case $s'_k \neq s_{k+1}$ ($s'_k = s_{k+1}$ held otherwise).

Tables 7 and 8 contain additional implementation details. For training, we have used the Adam optimizer (Kingma, 2014) with defaults from the Optax library (DeepMind et al., 2020) unless otherwise stated.

During evaluation, DηN followed greedy policies ($\varepsilon = 0$ for the ε-greedy exploration). For the τ-CVaR experiments (Sections 7.2 and 7.3), we selected c_0^* based on Theorems 17 and 19, with a grid search of 256 equally spaced points on the interval $[-10, 10]$ (with points on the interval limits).

Parameter	Value
Batch size	64
Trajectory length	16
Training duration (environment steps)	$\approx 2M$
Training duration (learner updates)	$2K$
Adam optimizer learning rate	10^{-4}
Target network exponential moving average step size (α)	10^{-2}
Discount (γ)	0.997
ε-greedy parameter	0.1
Interval for sampling c_0	$[-10, 10)$

Table 7: Training parameters for DηN in the gridworld experiments.

Component	Parameter	Value
Vision (ConvNet)		
	Output channels (per layer)	$(32, 64, 64)$
	Kernel sizes (per layer)	$((8, 8), (4, 4), (3, 3))$
	Strides (all layers)	$(1, 1)$
	Padding	`SAME`
Linear		
	Output size	512
MLP		
	Number of quantiles (per action)	128
	Hidden layer size	512

Table 8: Neural network parameters for DηN's return distribution estimator ξ_θ in the gridworld experiments. See Figure 1 for reference.

The vision network in the gridworld experiments is a ConvNet (LeCun et al., 2015) following the implementation used by Mnih et al. (2015). Convolutional layers used ReLU activations (Nair and Hinton, 2010), as did the MLP hidden layer. The "Linear" components in Figure 1 did not use an activation function on the outputs (with the exception of the explicit ReLU activation shown in the diagrams). The outputs of the ConvNet were flattened before being input to the "Linear" component.

H.3 Atari

In these experiments we trained DηN in a distributed actor-learner setup (Horgan et al., 2018) using TPUv3 actors and learners. The data was generated in episodic fashion (with multiple asynchronous actors). The episode duration was set to 25s, at 15Hz and 4 frames per environment step due to action repeats (Mnih et al., 2015). The Atari benchmark typically

Parameter	Value
Batch size (global, across 6 learners)	144
Trajectory length	19
Training duration (environment steps)	$75M$
Training duration (learner updates)	$\approx 3.44K$
Adam optimizer learning rate	10^{-4}
Weight decay	10^{-2}
Gradient norm clipping	10
Target network exponential moving average step size α	10^{-2}
Discount (γ)	0.997
Interval for sampling c_0	$[-9, 9)$

Table 9: Training parameters for DηN in the Atari experiments.

has sticky actions (Machado et al., 2018), but we disabled them for these experiments, to have deterministic returns. DηN, similar to DQN (Mnih et al., 2015) and QR-DQN (Dabney et al., 2018), observes 84×84 grayscale Atari frames with frame stacking of 4.

DηN was trained with a $3:7$ mixture of online and replay data in each learner update. Each minibatch consisted of 144 sampled trajectories (sequences of subsequent transitions) of length 19 (the minibatch was distributed across multiple learners, and updates were combined before being applied). The data generated in the actors was added simultaneously to a queue (for the online data stream) and to the replay (for the replay data stream). The replay was not prioritized, and we edited the stocks in each minibatch as explained in Section 8.

Tables 9 and 10 contain additional implementation details. For training, we have used the Adam optimizer (Kingma, 2014) with defaults from the Optax library (DeepMind et al., 2020) unless otherwise stated, as well as gradient norm clipping and weight decay.

Similar to DQN, we annealed the ε-greedy parameter linearly from 1.0 at the start to 0.1 at the end of training, and used 10^{-2}-greedy policies for evaluation.

The convolutional network in the Atari experiments is a ResNet (He et al., 2016) as used by Espeholt et al. (2018). Convolutional layers and residual blocks used ReLU activations (Nair and Hinton, 2010), as did the MLP hidden layer. The "Linear" components in Figure 1 did not use an activation function on the outputs (note that the explicit ReLU activation in the diagrams is used). The outputs of the ResNet were flattened before being input to the "Linear" component.

Appendix I. Summary of Guarantees

Table 11 provides a summary of the necessary and sufficient conditions for the objective F_K to be optimizable by DP in the different scenarios considered in this work. The table includes references to specific results in this work and in previous work, as applicable. For a more detailed discussion on DP guarantees from previous work, see Section 4.5. For a comparison between classic and distributional DP bounds (value iteration and policy iteration) refer to Sections 4.1 and 4.2.

Component	Parameter	Value
Vision (ResNet)		
	Output channels (per for Conv2D and residual layers per section)	$(64, 128, 128)$
	Kernel sizes (all Conv2D and residual layers)	$(3, 3)$
	Strides (all Conv2D and residual layers)	$(1, 1)$
	Padding	`SAME`
	Pool sizes (all sections)	$(3, 3)$
	Pool strides (all sections)	$(3, 3)$
	Residual blocks (per section)	$(2, 2, 2)$
Linear		
	Output size	512
Quantile MLP		
	Number of quantiles (per action)	100
	Hidden layer size	512

Table 10: Neural network parameters for DηN's return distribution estimator ξ_θ in the Atari experiments. See Figure 1 for reference.

Counter-examples. In the standard setting, without stock augmentation, classic and distributional DP can solve the same set of problems (see Table 11). With stock augmentation in the finite-horizon undiscounted case, see Proposition 22 for a functional that distributional DP can optimize, but classic DP cannot. In the stock-augmented infinite-horizon setting, we are not aware of any functionals that can only be optimized by either classic or distributional DP (cf. Theorems 6, 8 and 20). In the finite-horizon undiscounted setting with stock augmentation, there exist functionals that distributional DP can optimize but classic DP cannot (see Proposition 22). In the stock-augmented infinite-horizon setting, we are not aware of any functionals that can only be optimized by either classic or distributional DP (cf. Theorems 6, 8 and 20). If a counter-example exists, it must fall in one of the following two cases: i) an expected utility with f non-Lipschitz but $c \mapsto f(c) - f(0)$ positively homogeneous; ii) a non-Lipschitz non-expected-utility that is indifferent to mixtures and indifferent to γ (classic DP cannot optimize this; see Proposition 13 and Theorem 21).

References

A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller. Maximum a Posteriori Policy Optimisation. In *Proceedings of the 35th International Conference on Learning Representations*, 2018.

D. Abel, W. Dabney, A. Harutyunyan, M. K. Ho, M. L. Littman, D. Precup, and S. Singh. On the Expressivity of Markov Reward. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Setting	DP	Case	Conditions on the Objective (and references)
Standard	Classic or distributional	Finite horizon ($\gamma = 1$)	*Necessary and sufficient*: Expected utility U_f with (up to affine transformations) $f(c) = e^{\lambda c}$ for $\lambda \in \mathbb{R}$ or f identity (Marthe et al., 2024).
		Infinite horizon ($\gamma < 1$)	*Necessary and sufficient*: Expected utility U_f with f (up to affine transformations) positively homogeneous (see Proposition 41 and Theorems 6 and 8 and Bowling et al., 2023).
Stock-augmented	Classic	Finite horizon ($\gamma = 1$)	*Necessary and sufficient*: Expected utility, RL rewards with bounded first moment (Theorem 20).
		Infinite horizon ($\gamma < 1$)	*Necessary*: Expected utility U_f with f (up to affine transformations) positively homogeneous (Theorem 20 and Lemma 12). *Sufficient*: Expected utility U_f with f Lipschitz and f (up to affine transformations) positively homogeneous (Theorem 20).
Stock-augmented	Distributional	Finite horizon ($\gamma = 1$)	*Necessary and sufficient*: Indifferent to mixtures (Theorems 6 and 8 and Proposition 13).
		Infinite horizon ($\gamma < 1$)	*Necessary*: Indifferent to mixtures and indifferent to γ (Theorems 6 and 8 and Proposition 13). *Sufficient*: Lipschitz, indifferent to mixtures and indifferent to γ (Theorems 6 and 8 and Proposition 13).

Table 11: Summary of necessary and sufficient conditions on F_K for classic and distributional DP in various scenarios, including references. Previous work only considered the scalar case ($\mathcal{C} = \mathbb{R}$); our results also apply to the vector-valued case ($\mathcal{C} = \mathbb{R}^m$). All instances of positive homogeneity mentioned on this table have the following condition: $(1 - \alpha)(f(c) - f(0)) = f(\gamma c) - f(0)$ with $\alpha \in (0, 1]$ and $\gamma < 1 \Rightarrow \alpha < 1$ (see Equation 34).

E. Altman. *Constrained Markov Decision Processes.* Routledge, 1999.

A. Barreto, S. Hou, D. Borsa, D. Silver, and D. Precup. Fast Reinforcement Learning with Generalized Policy Updates. *Proceedings of the National Academy of Sciences*, 117(48): 30079–30087, 2020.

N. Bäuerle and A. Glauner. Minimizing Spectral Risk Measures Applied to Markov Decision Processes. *Mathematical Methods of Operations Research*, 94(1):35–69, 2021.

N. Bäuerle and J. Ott. Markov Decision Processes with Average-Value-at-Risk Criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

N. Bäuerle and U. Rieder. More Risk-Sensitive Markov Decision Pprocesses. *Mathematics of Operations Research*, 39(1):105–120, 2014.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

M. G. Bellemare, W. Dabney, and R. Munos. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang. Autonomous Navigation of Stratospheric Balloons Using Reinforcement Learning. *Nature*, 588(7836):77–82, 2020.

M. G. Bellemare, W. Dabney, and M. Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

D. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the Reward Hypothesis. In *Proceedings of the 40th International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

K.-J. Chung and M. J. Sobel. Discounted MDP's: Distribution Functions and Exponential Utility Maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

W. Dabney, M. Rowland, M. Bellemare, and R. Munos. Distributional Reinforcement Learning with Quantile Regression. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 32, 2018.

P. Dayan and C. Watkins. Q-Learning. *Machine Learning*, 8(3):279–292, 1992.

DeepMind, I. Babuschkin, K. Baumli, A. Bell, S. Bhupatiraju, J. Bruce, P. Buchlovsky, D. Budden, T. Cai, A. Clark, I. Danihelka, A. Dedieu, C. Fantacci, J. Godwin, C. Jones, R. Hemsley, T. Hennigan, M. Hessel, S. Hou, S. Kapturowski, T. Keck, I. Kemaev, M. King, M. Kunesch, L. Martens, H. Merzic, V. Mikulik, T. Norman, G. Papamakarios, J. Quan, R. Ring, F. Ruiz, A. Sanchez, L. Sartran, R. Schneider, E. Sezener, S. Spencer, S. Srinivasan, M. Stanojević, W. Stokowiec, L. Wang, G. Zhou, and F. Viola. The DeepMind JAX Ecosystem, 2020.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, et al. Magnetic Control of Tokamak Plasmas Through Deep Reinforcement Learning. *Nature*, 602(7897):414–419, 2022.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, 6, 2005.

L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. Impala: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz, et al. Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. *Nature*, 610(7930):47–53, 2022.

M. A. Goodrich and M. Quigley. Satisficing Q-Learning: Efficient Learning in Problems with Dichotomous Attributes. In *Proceedings of the International Conference on Machine Learning and Applications*, 2004.

I. Greenberg, Y. Chow, M. Ghavamzadeh, and S. Mannor. Efficient Risk-Averse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A. Dragan. Inverse Reward Design. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array Programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee. Flax: A Neural Network Library and Ecosystem for JAX, 2024.

T. Hennigan, T. Cai, T. Norman, L. Martens, and I. Babuschkin. Haiku: Sonnet for JAX, 2020.

D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. Van Hasselt, and D. Silver. Distributed Prioritized Experience Replay. *International Conference on Learning Representations*, 2018.

C. L. Hull. *Principles of Behavior: An Introduction to Behavior Theory.* Appleton-Century, 1943.

J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

D. Juan Carreño. The Von Neumann-Morgenstern Theory and Rational Choice. *Treballs Finals de Grau (TFG) – Matemàtiques, Universitat de Barcelona*, 2020.

M. Keramati and B. Gutkin. A Reinforcement Learning Theory for Homeostatic Regulation. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

D. P. Kingma. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2014.

W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone. Reward (Mis) Design for Autonomous Driving. *Artificial Intelligence*, 316:103829, 2023.

D. M. Kreps. Decision Problems with Expected Utility Criteria, ii: Stationarity. *Mathematics of Operations Research*, 2(3):266–274, 1977. ISSN 0364765X, 15265471.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

S. H. Lim and I. Malik. Distributional Reinforcement Learning for Risk-Sensitive Policies. In *Advances in Neural Information Processing Systems*, volume 35, pages 30977–30989, 2022.

M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

O. Madani, S. Hanks, and A. Condon. On the Undecidability of Probabilistic Planning and Infinite-Horizon Partially Observable Markov Decision Problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 10, 1999.

A. Marthe, A. Garivier, and C. Vernade. Beyond Average Return in Markov Decision Processes. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, 2015.

M. Moghimi and H. Ku. Beyond CVaR: Leveraging Static Spectral Risk Measures for Enhanced Decision-Making in Distributional Reinforcement Learning. *arXiv preprint arXiv:2501.02087*, 2025.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric Return Distribution Approximation for Reinforcement Learning. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 799–806, Haifa, Israel, June 2010. Omnipress.

V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

E. Noorani, C. Mavridis, and J. Baras. Risk-Sensitive Reinforcement Learning with Exponential Criteria. *arXiv preprint arXiv:2212.09010*, 2022.

T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020.

C. H. Papadimitriou and J. N. Tsitsiklis. The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

S. Pitis. Rethinking the Discount Factor in Reinforcement Learning: A Decision Theoretic Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

R. T. Rockafellar, S. Uryasev, et al. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–42, 2000.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

W. Schultz, P. Dayan, and P. R. Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, 1997.

M. Schwarzer, J. S. O. Ceron, A. Courville, M. G. Bellemare, R. Agarwal, and P. S. Castro. Bigger, Better, Faster: Human-Level Atari with Human-Level Efficiency. In *Proceedings of the 40th International Conference on Machine Learning*, pages 30365–30380. PMLR, 2023.

M. Shakerinava and S. Ravanbakhsh. Utility Theory for Sequential Decision Making. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19616–19625, 2022.

G. R. Shorack. *Probability for Statisticians*. Springer, 2017.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play. *Science*, 362(6419):1140–1144, 2018.

H. A. Simon. Rational Choice and the Structure of the Environment. *Psychological Review*, 63(2):129, 1956.

S. P. Singh and R. C. Yee. An Upper Bound on the Loss from Approximate Optimal-Value Functions. *Machine Learning*, 16:227–233, 1994.

J. T. Springenberg, A. Abdolmaleki, J. Zhang, O. Groth, M. Bloesch, T. Lampe, P. Brakel, S. Bechtle, S. Kapturowski, R. Hafner, et al. Offline Actor-Critic Reinforcement Learning Scales to Large Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46323–46350. PMLR, 2024.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A Scalable Real-Time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems–Volume 2*, pages 761–768, 2011.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Springer nature, 2022.

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via Sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

J. Von Neumann and O. Morgenstern. Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition. In *Theory of Games and Economic Behavior*. Princeton University Press, 2007.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. King's College, Cambridge United Kingdom, 1989.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

H. Wiltzer, J. Farebrother, A. Gretton, and M. Rowland. Foundations of Multivariate Distributional Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

P. Zhang, X. Chen, L. Zhao, W. Xiong, T. Qin, and T.-Y. Liu. Distributional Reinforcement Learning for Multi-Dimensional Reward Functions. In *Advances in Neural Information Processing Systems*, volume 34, 2021.