Bachelor's Thesis

Categorical Distributional Reinforcement Learning

Finite-Time Analysis and Application with Risk-Sensitive Policies

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Technical Mathematics

by

Markus Böck
Registration Number: 01634838

under the supervision of
Assoc. Prof. Dr.techn. Dipl.-Ing. Clemens Heitzinger
Institute of Analysis and Scientific Computing, TU Wien

Vienna, August 29, 2020

Abstract

In distributional reinforcement learning the entire distribution of returns is modelled, rather than just their expected value. In this work, the particular framework of using categorical distributions as approximation method is reviewed and subject to finite-time analysis. It is shown that even though one gains significantly more information about the return, the sample complexity of categorical distributional reinforcement learning algorithms is essentially equivalent to their value-based counterparts in the tabular case. While distributional methods proved popular in large scale applications when maximising the expected return, it is argued that the real benefit of having the entire return distributions at hand lies in the optimisation of risk-sensitive objectives. Incorporating risk measures into policy iteration methods is shown to be theoretically limited; however, in a simple gridworld problem such algorithms achieved the desired goal. Lastly, the application of distributional reinforcement learning methods to the problem of finding optimal sepsis treatment strategies is considered.

Kurzfassung

Reinforcement Learning mit Verteilungen befasst sich damit, anstelle des erwarteten Gewinns, die gesamte Verteilungsfunktion des Gewinns zu modellieren. In dieser Arbeit wird der Ansatz kategorische Verteilungen als Annäherungsmethode zu verwenden behandelt. Im tabularen Fall ist diese Methode Gegenstand von Konvergenzanalyse. Es wird gezeigt, dass obwohl man signifikant mehr Information über den Gewinn erhält, ist die Stichprobenkomplexität von Algorithmen basierend auf kategorischen Verteilungen im Wesentlichen äquivalent zu deren Erwartungswert-basierenden Pendants. Während sich Reinforcement Learning mit Verteilungen vor allem bei Anwendungen im großen Maßstab ausgezeichnet hat, wird argumentiert, dass die Verfügbarkeit der Gewinnverteilung den Einsatz in der Optimierung von risikosensitiven Problemen als wahren Vorteil hat. Es stellt sich heraus, dass das Verwenden von Risikomaße in Policy-Iteration Algorithmen theoretisch eingeschränkt ist, jedoch wurde in einem einfachen Gridworld-Problem das angestrebte Ziel mit solchen Methoden erreicht. Abschließend wird der Einsatz von Reinforcement Learning mit Verteilungen in der Suche nach einer optimalen Behandlung von Sepsis in Erwägung gezogen.

Contents

Chapter 1

Introduction

Besides supervised and unsupervised learning, *reinforcement learning (RL)* is a fundamental paradigm of machine learning. RL comprises goal-oriented methods, in which an agent aims to learn optimal action strategies through interaction with his environment. Only a reward signal helps the agent to deduce which actions have favourable or unfavourable outcome. This trial and error style of learning closely resembles human learning and is suited for many real world applications.

The effectiveness of RL proved itself throughout history. In 2016, a team at Google Deepmind presented AlphaGo, a program which learns to play the popular Chinese board game Go based on RL techniques. The resulting agent remarkably defeated the 18-time world champion Lee Sedol, winning 4 out of 5 games. The challenge of developing strong Go programs is the difficulty of finding an evaluation function – a function determining which board configuration favours which player. This renders search methods based on expert knowledge, which were used for chess programs, not applicable. Learning by playing, as suggested by RL, turned out to be the better choice [23, Chapter 16].

In standard RL, the performance of the agent is modelled as the expected sum of all future rewards – the expected return. In this work however, we focus on *distributional reinforcement learning (DRL)* where the *entire* distribution of the return is modelled directly. The interest in this subfield was rekindled by [2] in 2017 and DRL is currently an active area of research. As we have more information about the consequences of actions, a wide range of new possibilities for designing algorithms is offered.

Firstly, in order to introduce the mathematical notation used throughout this work, the formalisation of the RL objective in terms of a Markov Decision Processes will be stated. The DRL framework will be presented and a motivating example showing the potential benefits of DRL will be given. Naturally, the challenge of choosing a distribution approximation method arises and in this work the focus will be on categorical distributions. This method will be subject to complexity analysis in order to compare its performance to value-based RL methods. As already mentioned, modelling the distribution allows us a detailed view on the outcome of actions and offers the possibility of learning risk-sensitive strategies. However, designing algorithms for such tasks is challenging and the limitations of including risk measures in policy iteration methods will be examined. Lastly, the discussed methods will be tested on toy problems and an application to the medical problem of sepsis treatment will be considered.

1.1 Markov Decision Processes

The RL objective is formalised by a Markov Decision Process. The three main components are a set of states which represents the environment, a set of actions through which the agent can interact with the environment and a reward signal [2, 23].

Definition 1.1. A *Markov Decision Process (MDP)* is a tuple $\langle \mathcal{X}, \mathcal{A}, R, P \rangle$, where $\mathcal{X}$ is a set of states, $\mathcal{A}$ is a set of actions and $R(x, a)$ is a random variable for each $(x, a) \in \mathcal{X} \times \mathcal{A}$ representing the immediate reward when transitioning from (x, a).
Trajectories (X_t, A_t) are obtained through the selection of actions at given states, where the probabilities of state transitions are defined by the deterministic function P such that

$$\mathbf{P}\left[X_{t+1} = x' | X_t = x, A_t = a\right] = P(x'|x, a).$$

Furthermore, a MDP has the *Markov property*. State transition probabilities only depend on the current state and not on the history of predecessors, i.e.,

$$\mathbf{P}\left[X_{t+1} = x' | X_t, A_t, X_{t-1}, A_{t-1}, \ldots, X_0, A_0\right] = \mathbf{P}\left[X_{t+1} = x' | X_t, A_t\right].$$

The dynamics of a MDP are determined by a single function $p \colon \mathbb{R} \times \mathcal{X} \times \mathcal{X} \times \mathcal{A} \to [0, 1]$ such that

$$\int_{\mathbb{R}} p(r, x'|x, a) \mathrm{d}r = P(x'|x, a).$$

In addition to the state transition probabilities, the function p also captures the randomness of the immediate reward $R(x, a)$.

Action strategies of the agent are described by the notion of a policy.

Definition 1.2. A *policy* π is a mapping from states to probabilities of selecting each possible action. For all $s \in \mathcal{X}$ and $a \in \mathcal{A}$

$$0 \leq \pi(a|s) \leq 1, \quad \sum_{a \in \mathcal{A}} \pi(a|s) = 1.$$

However, it is often convenient to think of policies as random variables from states to actions $\pi \colon \mathcal{X} \to \mathcal{A}$ with

$$\mathbf{P}\left[\pi(s) = a\right] = \pi(a|s).$$

If there is an action a for each state s such that $\mathbf{P}\left[\pi(s) = a\right] = 1$, the policy can be considered a mapping from states to actions and is called *deterministic*, otherwise *stochastic*.

1.2 Distributional Reinforcement Learning

1.2.1 The Return Distribution

Definition 1.3. For a given MDP $\langle \mathcal{X}, \mathcal{A}, R, P \rangle$ and discount factor $\gamma \in (0, 1)$, the *return* at $(x, a) \in \mathcal{X} \times \mathcal{A}$ is the sum of discounted rewards along a trajectory following the policy π after starting in state x and taking action a, i.e.,

$$Z^\pi(x, a) := \sum_{t=0}^{\infty} \gamma^t R(X_t, A_t),$$
$$\text{where} \quad X_0 = x, \;\; A_0 = a, \;\; X_{t+1} \sim P(\cdot|X_t, A_t), \;\; A_{t+1} \sim \pi(\cdot|X_{t+1}).$$

The function Z^π, mapping state-action pairs to random variables, is called *return distribution function*. The set of all return distribution functions is denoted by $\mathcal{Z}$.
In order to shorten notation, we often write $R_t := R(X_t, A_t)$ and $Z^\pi(x, a) = \sum_{t=0}^{\infty} \gamma^t R_t$.

For proofs it is often useful to make use of the induced probability measure instead of the random variable. For this matter, we define the probability measure

$$\eta_\pi^{(x,a)}((-\infty, z]) := \mathbf{P}\left[Z^\pi(x,a) \leq z\right].$$

With this construction we have

$$Z^\pi(x,a) \sim \eta_\pi^{(x,a)} \quad \text{for all } (x,a) \in \mathcal{X} \times \mathcal{A}. \tag{1.1}$$

The function $\eta_\pi \colon \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R}), (x,a) \mapsto \eta_\pi^{(x,a)}$ may also be referred as *return distribution function* [2, 19]. Here $\mathscr{P}(\mathbb{R})$ denotes the set of all probability measures supported on $\mathbb{R}$.

As alluded to in the introduction, the objective of RL is to maximise the expected return at each state-action pair. For this task it is sufficient to consider only the *state-action value function* of a policy π defined by

$$Q^\pi(x,a) = \mathbb{E}\left[Z^\pi(x,a)\right].$$

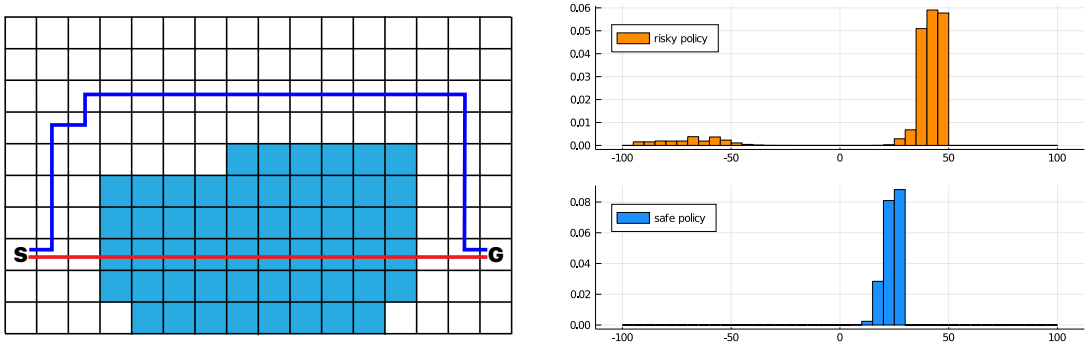
The state-value function of an optimal policy π^* satisfies

$$Q^{\pi^*}(x,a) = Q^*(x,a) := \sup_\pi Q^\pi(x,a),$$

where Q^* is called the optimal state-action value function.

But if we only care about the expected value, why bother modelling the return distribution as a whole? Following example shows that the expected value cannot be trusted in all applications and may lead to unexpected results. In Chapter 4 we address this problem by considering the inclusion of risk measures in the action selection process.

Example. Consider a gridworld environment in which the agent has the possibility of moving in four directions. In order to add randomness, the environment only accepts the action 90% of the times, otherwise it moves the agent in a random direction. The objective of the agent is to move from the start cell (marked with S) to the goal cell (marked with G). To make things difficult, there is a lake between these cells (marked with blue colour). If the agent lands on a lake cell, there is 1% chance of drowning, resulting in a reward of -100. Note that this setup is similar to the cliff walking problem in [23].
Consider two policies, whose trajectories are drawn in Figure 1.1a. The policy going straight through the lake will be regarded as the risky policy. The second policy, the safe policy, takes the path around the lake, which increases the probability of reaching the goal cell and receiving the reward of $+100$. However, more steps are required and we have a discount factor of $\gamma = 0.95$ which decreases the return for long trajectories.



(a) Gridworld with lake and two policies

(b) Histograms of return at start state

Figure 1.1: Environment with a risky and a safe policy

In Figure 1.1b you can see the histograms of the return after simulating 10^6 trajectories for both policies. The risky policy has a return of 30.84 ± 35.52, whereas the safe one has a return of 24.08 ± 3.74. In standard RL one would conclude that the risky policy is better. However, in DRL we can do further analysis. For example, the risky policy has around 10% chance of drowning (note the probability mass left of 0), while the safe one only drowns in 0.01% of the tries and only due to the randomness of the environment.
One could argue that if we care so much about drowning, we would choose a different setup. For example, punishing drowning by a greater negative reward or increasing the discount factor. However, in many applications the environment is difficult to understand and the choice of a reward function is not straightforward. In these case, the availability of the return distribution is undeniably of great benefit.

1.2.2 Distributional Bellman Equation and Operators

In the heart of RL lies the Bellman equation [3]. It relates the return of a state to the return of its successor states.

Theorem 1.4. *For the random transition $(x, a) \to (X', A')$, the distributional Bellman equation is given by*

$$Z^\pi(x, a) \overset{D}{=} R(x, a) + \gamma Z^\pi(X', A').$$

The equality indicates that the random variable on the left hand side and the one on the right hand side are identically distributed.

Like the Bellman equation, Bellman operators play an important role in reinforcement learning. In standard RL they are defined in terms of the state-action value functions $Q\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ by

$$\begin{aligned} \mathcal{T}^\pi Q(x, a) &:= \mathbb{E}\left[R(x, a) + \gamma Q(X', A')\right], \\ \mathcal{T} Q(x, a) &:= \mathbb{E}\left[R(x, a) + \gamma \max_{a' \in \mathcal{A}} Q(X', a')\right], \end{aligned}$$

where $X' \sim p(\cdot|x, a)$ and $A' \sim \pi(\cdot|X')$. These operators are called *Bellman operator* and *Bellman optimality operator*, respectively. They are clearly defined with the Bellman equation in mind and have the property that a fixed point of $\mathcal{T}^\pi$ is the state-action value function Q^π and a fixed point of $\mathcal{T}$ is the optimal state-action value function Q^*. Both are γ-contractions in the supremum norm $\|.\|_\infty$, which is used in convergence proofs for Q-learning [24].
The extension of Bellman operators to distributions is straightforward.

Definition 1.5. The *distributional Bellman operator* $\mathcal{T}^\pi\colon \mathcal{Z} \to \mathcal{Z}$ is defined by

$$\mathcal{T}^\pi Z(x, a) := R(x, a) + \gamma Z(X', A'), \quad X' \sim p(\cdot|x, a), \;\; A' \sim \pi(\cdot|X')$$

and the *distributional Bellman optimality operator* $\mathcal{T}\colon \mathcal{Z} \to \mathcal{Z}$ by

$$\mathcal{T} Z := R(x, a) + \gamma Z(X', A^*), \quad X' \sim p(\cdot|x, a), \;\; A^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[Z(X', a)\right].$$

Proposition 1.6. *$\mathcal{T}^\pi Z^\pi(x, a) \sim \mathcal{T}^\pi \eta^{(x,a)}$, where the probability measure $\mathcal{T}^\pi \eta^{(x,a)}$ is defined by its cumulative distribution function*

$$\begin{aligned} F_{\mathcal{T}^\pi \eta^{(x,a)}}(z) &= \mathbb{E}\left[F_{\eta^{(X',A')}}\left(\frac{z - R(x, a)}{\gamma}\right)\right] \\ &= \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} F_{\eta^{(x',a')}}\left(\frac{z - r}{\gamma}\right) \pi(a'|x') p(r, x'|x, a) \mathrm{d}r. \end{aligned} \tag{1.2}$$

Proof. Follows immediately from relationship (1.1) and Definition 1.5. □

Analogously to value-based RL, we would like to find a metric such that $\mathcal{T}^\pi$ and $\mathcal{T}\colon \mathcal{Z} \to \mathcal{Z}$ are contraction mappings with a unique fixed point. [2] proposed to use the Wasserstein distance.

Definition 1.7. Let $\mu, \nu \in \mathscr{P}_p(\mathbb{R})$ be probability measures with p-th finite moments. The *p-Wasserstein distance* between μ and ν is given by

$$w_p(\mu, \nu) := \left\| F_\mu^{-1} - F_\nu^{-1} \right\|_{L^p(0,1)},$$

where F^{-1} denotes the respective quantile function and $\|.\|_{L^p(0,1)}$ is the norm of the L^p-space on the interval $(0, 1)$. The Wasserstein distance between two random variables is defined equivalently,

$$w_p(U, V) := \left\| F_U^{-1} - F_V^{-1} \right\|_{L^p(0,1)}.$$

By taking the supremum over all state-action pairs this distance can be extended to return distribution functions, i.e.,

$$\overline{w}_p(Z_1, Z_2) := \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} w_p(Z_1(x, a), Z_2(x, a)).$$

Indeed, [2, Lemma 3] confirms that $\mathcal{T}^\pi$ is a contraction in the Wasserstein distance.

Proposition 1.8. *$\mathcal{T}^\pi$ is a γ-contraction in $\overline{w}_p$.*

By Banach's fixed point theorem Proposition 1.8 guarantees that $Z^\pi = \lim_{n\to\infty}(\mathcal{T}^\pi)^n Z_0$ in $\overline{w}_p$ for any initial guess $Z_0 \in \mathcal{Z}$. Essentially, this verifies that policy evaluation works in the distributional setting.
However, policy control, where we are concerned with $\mathcal{T}$ and seek to find optimal policies, causes problems. [2, Prop. 1 – 3] provides three negative results:

- $\mathcal{T}$ is not a contraction in $\overline{w}_p$.
- In general, $\mathcal{T}$ does not have a fixed point.
- Even if $\mathcal{T}$ has a fixed point, convergence of $\mathcal{T}^n Z_0$ to it is not guaranteed.

This constitutes a fundamental challenge in DRL. However, it was possible to recover convergence in the control setting in a weaker sense [2, Theorem 1], which we do not examine further.
In the following chapter we will consider the Cramér distance. For this metric the optimality operator is also not a contraction. Generally, it is very unlikely that there is a metric which yields the desired contraction properties of $\mathcal{T}\colon \mathcal{Z} \to \mathcal{Z}$.

Chapter 2

Categorical Distributional Reinforcement Learning

In categorical DRL the return distributions are approximated by a finite set of fixed atoms. [2] used this approximation method in conjunction with a deep neural network and achieved convincing experimental results. After presenting this algorithm, called C51, we take a look at categorical DRL methods in the tabular case, where convergence results have been established [19]. Finally, a few alternative approximation methods will be presented.

2.1 The C51 Algorithm

As mentioned, the approach of [2] was to approximate the return distributions with a finite set of fixed atoms. They set bounds for the return $V_{\min}$, $V_{\max}$ and then used N equally spaced atoms $\{V_{\min} + (i-1)\Delta z : i = 1, \dots, N\}$, $\Delta z = (V_{\max} - V_{\min})/(N-1)$ as support for the distributions. The algorithm is called C51 because the choice of 51 atoms resulted in especially good empirical performance. The distributions supported on a finite support are called categorical distributions.

Definition 2.1. Let δ_{z_i} be the Dirac measure at z_i. The set of *categorical distributions* is defined as

$$\mathcal{P}_z := \left\{ \sum_{i=1}^{N} p_i \delta_{z_i} : \ p_i \geq 0, \ \sum_{i=1}^{N} p_i = 1 \right\} \subset \mathscr{P}(\mathbb{R}).$$

Since the Bellman operator scales the return distribution by γ and translates it by the reward, the categorical distributions are not closed under this operation and it was necessary to find a way to project a distribution back on the fixed support.

Definition 2.2. The *categorical projection* operator $\Pi_{\mathcal{C}} \colon \mathcal{P}_y \to \mathcal{P}_z$ is given by

$$\Pi_{\mathcal{C}}(\delta_{y_j}) := \begin{cases} \delta_{z_1}, & y_j \leq z_1, \\ \frac{z_{i+1}-y_j}{z_{i+1}-z_i}\delta_{z_i} + \frac{y_j - z_i}{z_{i+1}-z_i}\delta_{z_{i+1}}, & z_i < y_j \leq z_{i+1}, \\ \delta_{z_N}, & y_j > z_N, \end{cases} \quad \Pi_{\mathcal{C}}\left(\sum_{i=1}^{N} p_i \delta_{y_i}\right) = \sum_{i=1}^{N} p_i \Pi_{\mathcal{C}}(\delta_{y_i}) \tag{2.1}$$

In the case of categorical distributions, the operator $\Pi_{\mathcal{C}}$ basically distributes the probability of a point among the two neighbouring fixed atoms.
This operator can also be defined on the set of all distributions $\Pi_{\mathcal{C}} \colon \mathscr{P}(\mathbb{R}) \to \mathcal{P}_z$ by specifying the cumulative distribution function

$$F_{\Pi_{\mathcal{C}}\nu}(z_i) = \frac{1}{z_{i+1} - z_i} \int_{z_i}^{z_{i+1}} F_\nu(x) \mathrm{d}x, \quad F_{\Pi_{\mathcal{C}}\nu}(z_N) = 1. \tag{2.2}$$

Definition (2.2) simplifies to (2.1) for $\nu \in \mathcal{P}_y$. In the following $\Pi_{\mathcal{C}}$ will also be used on $\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, where it is meant elementwise.

The C51 algorithm was implemented as an improvement to the DQN algorithm [16], where the objective is to play Atari 2600 games. The backbone of DQN is a deep neural network consisting of convolutional and fully connected layers. The neural network takes in pixel data and estimates the state-action value function Q^π.
Reference [2] used the same network architecture as DQN, which is formally described by a parametric model $\theta\colon \mathcal{X}\times\mathcal{A}\to\mathbb{R}^N$. In order to obtain probabilities, the softmax function is applied to the final output of the network, i.e.,

$$Z_\theta(x,a) = z_i \quad \text{w.p.} \quad p_i := p_i^\theta(x,a) = \frac{\exp(\theta_i(x,a))}{\sum_{j=1}^N \exp(\theta_j(x,a))}.$$

For a sample transition (x,a,r,x') the Bellman optimality operator $\mathcal{T}$ is estimated by

$$y_i = \hat{\mathcal{T}} Z_\theta(x,a)_i = r + \gamma z_i \quad \text{w.p.} \quad q_i := p_i^\theta(x', a^*),$$

where a^* is the current greedy action in x' (with respect to the expected return).

After calculating the categorical projection, learning is done by minimising the cross entropy loss via a gradient descent method with respect to θ, i.e.,

$$\text{minimise}_\theta \quad -\sum_{i=1}^N q_i \log(p_i) \quad \text{for} \quad \Pi_{\mathcal{C}}\hat{\mathcal{T}} Z_\theta(x,a) \sim q,\ Z_\theta(x,a)\sim p.$$

The C51 Algorithm as described is shown in Algorithm 1. Here one can see how the categorical projection can be implemented programmatically.

Algorithm 1 C51 [2, Algorithm 1]

Require: *Parametric model* $\theta(x,a)$, $0<\gamma<1$, bounds for reward $V_{\min}, V_{\max}$, number of fixed atoms N
Input: *State transition* x_t, a_t, r_t, x_{t+1}
$Q(x_{t+1},a) := \sum_{i=1}^N z_i p_i^\theta(x_{t+1},a)$
$a^* \leftarrow \arg\max_a Q(x_{t+1},a)$ # find greedy action
$q_i = 0$, $i \in 1,\dots,N$
# Compute $\Pi_{\mathcal{C}}\hat{\mathcal{T}} Z_\theta(x_t,a_t)$
for $j \in 1,\dots,N$ **do**
    $\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma z_j]_{V_{\min}}^{V_{\max}}$ # clipped to $[V_{\min}, V_{\max}]$
    $b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\min})/\Delta z + 1$ # $\in [1,N]$
    $l \leftarrow \lfloor b_j \rfloor$, $u \leftarrow \lceil b_j \rceil$ # neighbouring atoms $z_l \le \hat{\mathcal{T}} z_j \le z_u$
    # Distribute probabilities
    if l = u **then**
        $q_l \leftarrow q_l + p_j^\theta(x_{t+1},a^*)$
    else
        $q_u \leftarrow q_u + p_j^\theta(x_{t+1},a^*)(u - y_j)$
        $q_l \leftarrow q_l + p_j^\theta(x_{t+1},a^*)(y_j - l)$
    end if
end for
$\text{loss}_\theta \leftarrow -\sum_{i=1}^N q_i \log p_i^\theta(x_t,a_t)$ # cross entropy loss
Perform gradient descent step with respect to θ.

Even though rendering convincingly effective in experiments, the success could not be explained by theoretical results. In the case of tabular or linear function models of the state space, [14] proved that the distributional methods perform equivalently to their value based counterparts in the sense that the same policies are obtained when maximising the expected return. However, this equivalence ends at non-linear function approximations. They were not able to explain the performance improvements of the categorical approach in conjunction with deep neural networks, but it is conjectured that distributions may have a regularizing effect in optimisation for neural networks.

2.2 Tabular Categorical DRL

In the tabular case, instead of using function approximation in form of a parametric model, the return at each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ is kept track of individually.
For the sample (x_t, a_t, r_t, x_{t+1}) and a_{t+1}, chosen by a policy or greedily in case of policy control, updates are performed in a stepsize controlled manner, i.e.,

$$\begin{aligned} \eta_{t+1}^{(x_t,a_t)} &\leftarrow (1-\alpha_t)\eta_t^{(x_t,a_t)} + \alpha_t \Pi_{\mathcal{C}} \left(\hat{\mathcal{T}} \eta_t^{(x_t,a_t)}\right), \\ \eta_{t+1}^{(x,a)} &\leftarrow \eta_t^{(x,a)} \quad \forall (x,a) \neq (x_t, a_t), \end{aligned} \tag{2.3}$$

$$\text{where} \quad \eta_t^{(x,a)} = \sum_{i=1}^{N} p_{t,i}^{(x,a)} \delta_{z_i} \in \mathcal{P}_z \quad \text{and} \quad \hat{\mathcal{T}} \eta_t^{(x_t,a_t)} = \sum_{i=1}^{N} p_{t,i}^{(x_{t+1},a_{t+1})} \delta_{r_t + \gamma z_i}.$$

This update rule is the (categorical) distributional counterpart of the one-step temporal difference learning / Q-learning update rule [23, 25]. Therefore, there is also a close theoretical connection between these methods. The following proposition essentially shows that the same policies are obtained in both the distributional and the value-based version.

Proposition 2.3. [14, Prop. 5] *Tabular Categorical DRL is equivalent to one-step temporal difference learning / Q-learning in expectation.*
Let $z_1 \leq -\frac{R_{max}}{1-\gamma}$ *and* $z_N \geq \frac{R_{max}}{1-\gamma}$ *and* η_t *be obtained by the update rule 2.3. Furthermore, let* $\eta_0^{(x,a)} \in \mathcal{P}_z$, $Q_0(x,a) := \mathbb{E}_{Z \sim \eta_0^{(x,a)}}[Z]$. *The state-action value functions are updated by*

$$Q_{t+1}(x,a) \leftarrow \begin{cases} (1-\alpha_t) Q_t(x_t, a_t) + \alpha_t (r_t + \gamma Q_t(x_{t+1}, a_{t+1})), & (x,a) = (x_t, a_t), \\ Q_t(x,a), & \textit{otherwise.} \end{cases}$$

Then

$$Q_t(x,a) = \mathbb{E}_{Z \sim \eta_t^{(x,a)}}[Z]$$

for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ *and* $t \geq 0$.

It was discovered that there is fundamental geometric connection between the categorical projection operator $\Pi_{\mathcal{C}}$ and the Cramér distance, defined as follows.

Definition 2.4. For two probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R})$ the *Cramér distance* is given by

$$\ell_2(\mu, \nu) = \left(\int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|^2 \mathrm{d}x \right)^{1/2}. \tag{2.4}$$

Once again by taking the supremum over all state-action pairs, this metric can be extended to return distribution functions by defining

$$\bar{\ell}_2(\eta, \xi) := \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \ell_2(\eta^{(x,a)}, \xi^{(x,a)}).$$

On the set of signed measures ν with $\|F_\nu\|_{L^2} < \infty$ the Cramér distance is induced by the inner product

$$\langle \mu, \nu \rangle_{\ell_2} = \int_{\mathbb{R}} F_\mu(x) F_\nu(x) \mathrm{d}x.$$

In general, the combination of the categorical projection and the Bellman operator $\Pi_{\mathcal{C}} \mathcal{T}^\pi$ is not a contraction in the Wasserstein distance; however, it is in the Cramér distance. This follows from the aforementioned underlying geometric connection. That is, $\Pi_{\mathcal{C}}$ is an *orthogonal projection* on a certain set of measures. For more details, see the proof of the following proposition.

Proposition 2.5. [19, Prop. 2] *$\Pi_{\mathcal{C}} \mathcal{T}^\pi \colon \mathcal{P}_z \to \mathcal{P}_z$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$.*

Proof. First, $\mathcal{T}^\pi$ will be shown to be a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$. Secondly, by considering a certain Hilbert space of signed measures, $\Pi_{\mathcal{C}}$ will be shown to be a orthogonal projection. As concatenation of these operators, $\Pi_{\mathcal{C}} \mathcal{T}^\pi$ is then also a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$.
For each (x, a) in $\mathcal{X} \times \mathcal{A}$ we have

$$\begin{aligned}
&\ell_2^2 \left((\mathcal{T}^\pi \eta)^{(x,a)}, (\mathcal{T}^\pi \xi)^{(x,a)} \right) \\
&= \int_{\mathbb{R}} |F_{(\mathcal{T}^\pi \eta)^{(x,a)}}(z) - F_{(\mathcal{T}^\pi \xi)^{(x,a)}}(z)|^2 \mathrm{d}z \\
&= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \left(F_{\eta^{(x',a')}}(\frac{z-r}{\gamma}) - F_{\xi^{(x',a')}}(\frac{z-r}{\gamma}) \right) \pi(a'|x') p(r, x'|x, a) \mathrm{d}r \right|^2 \mathrm{d}z \\
&\leq \gamma \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} \left| F_{\eta^{(x',a')}}(y) - F_{\xi^{(x',a')}}(y) \right|^2 \mathrm{d}y \, \pi(a'|x') p(r, x'|x, a) \mathrm{d}r \\
&\leq \gamma \sup_{(x',a') \in \mathcal{X} \times \mathcal{A}} \int_{\mathbb{R}} \left| F_{\eta^{(x',a')}}(y) - F_{\xi^{(x',a')}}(y) \right|^2 \mathrm{d}y = \gamma \bar{\ell}_2^2(\eta, \xi).
\end{aligned}$$

The first inequality follows by substituting $y = \frac{z-r}{\gamma}$ in the inner integral, applying Jensen's inequality for $|.|^2$, and finally rearranging the integrals. By taking the supremum over all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and taking the square root it follows that $\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$.
Let $\mathcal{M}_0$ be the Hilbert space of finite signed measures with $\mu(\mathbb{R}) = 0$ and $\|F_\mu\|_{L^2} < \infty$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\ell_2}$. Then $\delta_0 + \mathcal{M}_0$ contains the set of measures ν with $\int_{-\infty}^0 F_\nu(t)^2 \mathrm{d}t < \infty$ and $\int_0^\infty (1 - F_\nu(t))^2 \mathrm{d}t < \infty$ and is also a Hilbert space with $\langle \delta_0 + \mu, \delta_0 + \nu \rangle = \langle \mu, \nu \rangle_{\ell_2}$.
It follows from 2.2 that the projection $\Pi_{\mathcal{C}}$ minimises $\langle \Pi\nu - \nu, \Pi\nu - \nu \rangle_{\ell_2} = \int_{\mathbb{R}} |F_{\Pi\nu}(x) - F_\nu(x)|^2 \mathrm{d}x$. As $\mathrm{span}(\mathcal{P}_z)$ is a finite dimensional (and thus closed) subspace, $\Pi_{\mathcal{C}}$ is an orthogonal projection from $\delta_0 + \mathcal{M}_0$ to $\mathrm{span}(\mathcal{P}_z)$ and therefore a non-expansion. □

The analysis of [19] was concluded by showing that we get improved accuracy as we increase the number of atoms and that we have convergence both in policy evaluation and policy control.

Proposition 2.6. *Let $\eta_{\mathcal{C}} = \lim_{n \to \infty} (\Pi_{\mathcal{C}} \mathcal{T}^\pi)^n \eta_0$ proven to exist by Proposition 2.5. If the true return distribution $\eta_\pi^{(x,a)}$ is supported on $[z_1, z_N]$ for all states and actions (x, a), then*

$$\bar{\ell}_2^2(\eta_{\mathcal{C}}, \eta_\pi) \leq \frac{1}{1-\gamma} \max_{1 \leq i < N} (z_{i+1} - z_i).$$

Theorem 2.7 (Policy Evaluation)**.** *Let π be a policy. Suppose that*

i) the usual stepsize conditions hold, i.e. $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$;

ii) the distributions of the initial estimate $\eta_0^{(x,a)}$ have support contained in $[z_1, z_N]$.

Then

$$\bar{\ell}_2(\eta_t, \eta_{\mathcal{C}}) \overset{t\to\infty}{\longrightarrow} 0 \quad \text{almost surely,}$$

where $\eta_{\mathcal{C}} = \lim_{n\to\infty}(\Pi_{\mathcal{C}}\mathcal{T}^{\pi})^n\eta_0$ is the limiting return distribution function proven to exist by Proposition 2.5.

Theorem 2.8 (Policy Control)**.** *Suppose that the assumptions of Theorem 2.7 hold. Furthermore, let $z_1 \leq -\frac{R_{max}}{1-\gamma}$ and $z_N \geq \frac{R_{max}}{1-\gamma}$ and assume that there exists a unique optimal policy π^*.*
Then there exists a return distribution function $\eta_{\mathcal{C}}^*$ with

$$\bar{\ell}_2(\eta_t, \eta_{\mathcal{C}}^*) \overset{t\to\infty}{\longrightarrow} 0 \quad \text{almost surely.}$$

The greedy policy with respect to $\eta_{\mathcal{C}}^*$ is π^*.

2.3 Alternative Approximation Methods

2.3.1 Model-based Approximations

As discussed in the previous section, in categorical DRL we consider distributions over a fixed finite support, i.e., we consider

$$\mathcal{P}_z = \left\{ \sum_{i=1}^{N} p_i \delta_{z_i} : \;\; p_i \geq 0, \;\; \sum_{i=1}^{N} p_i = 1 \right\}.$$

This approximation method is highly expressive as one obtains increased accuracy compared to the true return distribution by increasing the number of atoms, see Proposition 2.6. However, it comes with two drawbacks: First, we need to be able to set reasonable bounds for the return ($V_{\min}$ and $V_{\max}$). Secondly, we have to store N parameters per state-action pair, which can lead to huge tables.
Reference [18] considered the following distribution models with only a few parameters:

i) Gaussian distributions

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \text{where } \mu \in \mathbb{R}, \sigma > 0,$$

ii) (skewed) Laplace distributions

$$f(x; m, b, c) = \frac{c(1-c)}{b} \begin{cases} \exp(\frac{1-c}{b}(x-m)), & \text{if } x < m, \\ \exp(-\frac{c}{b}(x-m)), & \text{otherwise,} \end{cases}$$

where $m \in \mathbb{R}, b > 0, c \in (0, 1)$.

For each state-action pair only two or three parameters are needed. In the work of [18], the cross entropy loss between return distribution η_t and estimated Bellman update $\hat{\mathcal{T}}\eta_t$ was minimised via gradient descent, similar to the C51 algorithm. The selection of the above distribution families allowed them to derive explicit update rules for the parameters.

2.3.2 Model-free Approximations

Quantile Regression

In DRL with Quantile Regression [6] one makes no assumption on the return distribution and aims to learn the quantile function F^{-1}. Consider N fixed, equally spaced quantiles $\tau_i := \frac{i}{N}$, $i = 1, \dots, N$. Then the expected value of a random variable Z can be approximated by

$$\mathbb{E}[Z] = \mathbb{E}\left[F_Z^{-1}(U)\right] \approx \frac{1}{N}\sum_{i=1}^{N} F_Z^{-1}(\tau_i),$$

where $U \sim \mathcal{U}([0,1])$ is uniformly distributed on the interval $[0,1]$.
In a way, this strategy transposes the parametrisation of the categorical approach. Instead of fixing the locations and learning their probabilities, one fixes the probabilities and learns their locations (quantiles). An immediate advantage is that one no longer needs to determine bounds for the return.
Reference [6] developed an algorithm based on the method of quantile regression. Here one finds a τ-quantile of a distribution η by minimising the loss

$$\mathcal{L}_{\mathrm{QR}}^{\tau}(\theta) = \mathbb{E}_{Z\sim\eta}\left[(\tau \mathbb{1}_{Z>\theta} + (1-\tau)\mathbb{1}_{Z\leq\theta})|Z-\theta|\right],$$

where $\mathbb{1}$ denotes the indicator function. They were also able to backup their algorithm by establishing a contraction result in the 1-Wasserstein distance.

Particles

Reference [17] proposed to use a finite set of particles $v_{x,n}$, $n = 1, \dots, N$, for each state $x \in \mathcal{X}$. The return distribution in this state is then estimated by the empirical cumulative distribution function,

$$F_{\eta^x}(z) \approx \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{v_{x,n}\leq z}.$$

For a transition $x \to x'$ with reward r updates are performed by randomly (uniformly) selecting indices $p, q \in 1, \dots, N$ and setting

$$v_{x,p} \leftarrow r + \gamma v_{x',q}.$$

The number of updated particles of state x is controlled by a learning rate. Increasing the total number of particles achieves higher approximation accuracy.
Storing a large amount of particles for each state is very expensive and only feasible for small state-actions spaces. However, we have the benefit that we need not have knowledge about bounds for the return as in categorical DRL.

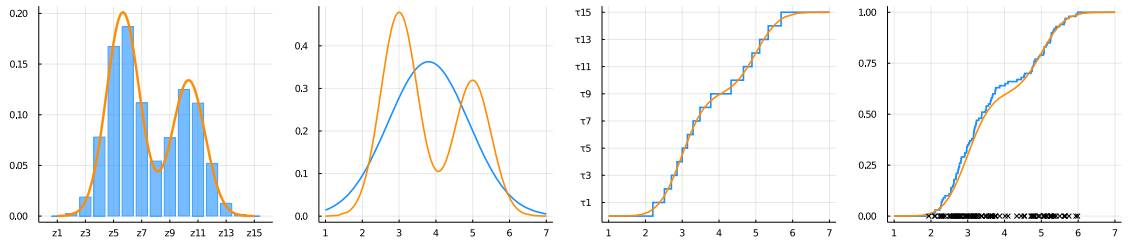


Figure 2.1: Comparison of approximation methods (blue) and true distribution (orange). Left to right: Categorical distribution with 15 atoms, Gaussian model, 15 quantiles, 100 particles (samples drawn).

Chapter 3

Finite-Time Analysis

As the goal of DRL is to estimate return distributions, rather than only the expected return, one might assume that with this more challenging task comes also the need for more transition samples to accurately approximate the distributions. By performing a finite-time analysis so called probably approximately correct (PAC) bounds for tabular categorical DRL algorithms are established and we come to the conclusion that the sample complexity is essentially the same as for the value based algorithmic counterparts.

3.1 Complexity of Q-Learning and Speedy Q-Learning

The following notation is used to compare the complexity of algorithms.

Definition 3.1. For two functions f and $g\colon D \subseteq \mathbb{R}^d \to [0,\infty)$ we define

$$\begin{aligned} f = \mathcal{O}(g) &:\Longleftrightarrow \exists C > 0 : f(x) \leq Cg(x) \quad \forall x \in D, \\ f = \Omega(g) &:\Longleftrightarrow g = \mathcal{O}(f), \\ f = \Theta(g) &:\Longleftrightarrow f = \mathcal{O}(g) \text{ and } f = \Omega(g). \end{aligned}$$

Lastly, since logarithmic factors are often insignificant we write

$$f = \tilde{\mathcal{O}}(g) :\Longleftrightarrow \exists C_1, C_2 > 0 : f(x) \leq C_1 g(x) \log^{C_2}(g(x)) \quad \forall x \in D.$$

Note that all inequalities are required to hold on the entire domain D. This is uncommon but sufficient for our purposes.

Now we are ready to examine the complexity of standard Q-learning [25]. Recall that for a sample (x, a, r, x') the update rule reads as

$$Q_{k+1}(x,a) = (1 - \alpha_k(x,a))Q_k(x,a) + \alpha_k(x,a)(r + \gamma \max_{a \in \mathcal{A}} Q_k(x',a)). \tag{3.1}$$

If we assume the usual stepsize conditions $\sum_{k=0}^{\infty} \alpha_k(x,a) = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2(x,a) < \infty$, convergence to the optimal state-action-value function Q^* is guaranteed [24].
It is convenient to neglect the exploration process at first and assume that updates are performed *synchronously* – at each timestep k all $(x,a) \in \mathcal{X} \times \mathcal{A}$ are updated. The result can then be extended to the asynchronous case, which we will not consider, since the synchronous case will be sufficient for the comparison to categorical DRL.
In [7], the authors assume a finite state-action space $n = |\mathcal{X} \times \mathcal{A}|$, a discount factor $\gamma < 1$, that the reward is bounded by $R_{\max}$ and polynomial learning rates $\alpha_k = \frac{1}{(k+1)^\omega}$, where

$0.5 < \omega < 1$. It follows that the return is also bounded by $V_{\max} := \beta R_{\max}$, where $\beta := \frac{1}{1-\gamma}$. Under these assumptions it was proved that

$$\mathbf{P}\left[\|Q^* - Q_T\|_\infty \leq \epsilon\right] \geq 1-\delta \quad \text{for } T = \Omega\left(\left[\frac{\beta^2 V_{\max}^2 \log \frac{nV_{\max}}{\delta\epsilon}}{\epsilon^2}\right]^{\frac{1}{\omega}} + \left[\beta \log \frac{V_{\max}}{\epsilon}\right]^{\frac{1}{1-\omega}}\right). \tag{3.2}$$

Inequalities of this form are called probably approximately correct (PAC) bounds, as they provide a lower bound for the time T such that Q_T is close to the solution with high probability.
The bound (3.2) seems quite complicated at first glance, but if γ is close to 1, one can argue that β becomes the dominant term and the bound is optimised for $\omega = 4/5$. Recall that increasing T by 1 means looping over the entire state-action space and thus $2n$ samples (reward and next state) are taken. This yields a sample complexity of $\tilde{\mathcal{O}}(n\beta^5/\epsilon^{2.5})$, omitting logarithmic factors.

In more recent developments, [10] introduced a faster variant of Q-learning and gave it the name *speedy Q-learning (SQL)*. They defined an update rule based on two previous timesteps instead of just one, i.e.,

$$Q_{k+1}(x,a) = Q_k(x,a) + \alpha_k(\mathcal{T}_k Q_{k-1}(x,a) - Q_k(x,a)) + (1-\alpha_k)(\mathcal{T}_k Q_k(x,a) - \mathcal{T}_k Q_{k-1}(x,a)), \tag{3.3}$$

where $\mathcal{T}_k Q(x,a) = r + \gamma \max_{a\in\mathcal{A}} Q(x',a)$ and the learning rate is linear, $\alpha_k := \frac{1}{k+1}$. The key difference to Q-learning is that SQL uses a more aggressive learning rate for the third term. Changing the third summand to $\alpha_k(\mathcal{T}_k Q_k(x,a) - \mathcal{T}_k Q_{k-1}(x,a))$ would simplify to standard Q-learning (3.1). The difference seems small; however, it yields faster convergence, which is very noticeable in experimental results, see Section 5.1.
In fact, for the sequence obtained by (3.3) we have

$$\mathbf{P}\left[\|Q^* - Q_T\|_\infty \leq \epsilon\right] \geq 1-\delta \quad \text{for } T = \Omega\left(\frac{\beta^2 V_{\max}^2 \log \frac{2n}{\delta}}{\epsilon^2}\right). \tag{3.4}$$

Again, viewing β as the dominant term we have a sampling complexity of $\tilde{\mathcal{O}}(n\beta^4/\epsilon^2)$.

Because of this performance improvement we will extend the analysis of SQL to categorical distributions, rather than standard Q-learning. However, it is worth mentioning that the main idea of the proof is also applicable if one uses (3.1).

3.2 Speedy Categorical Policy Evaluation

In order to translate SQL to categorical distributions, we combine (3.3) and (2.3) for the evaluation of a fixed policy π, i.e.,

$$\eta_{k+1}^{(x,a)} = \eta_k^{(x,a)} + \alpha_k(\Pi_{\mathcal{C}} \mathcal{T}_k^\pi \eta_{k-1}^{(x,a)} - \eta_k^{(x,a)}) + (1-\alpha_k)(\Pi_{\mathcal{C}} \mathcal{T}_k^\pi \eta_k^{(x,a)} - \Pi_{\mathcal{C}} \mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}). \tag{3.5}$$

The two initial return distribution functions are assumed to be the same, i.e., $\eta_0 = \eta_{-1} \in \mathcal{P}_z$. Like in the value-based algorithm, we also use a fixed linear stepsizes $\alpha_k := \frac{1}{k+1}$.
$\mathcal{T}_k^\pi$ is the stochastic Bellman operator at time k, which depends on samples $x'_k \sim p(\cdot|x,a)$, $a'_k \sim \pi(\cdot|x'_k)$, and the reward sample $r_k(x,a) \sim R(x,a)$ for each $(x,a) \in \mathcal{X}\times\mathcal{A}$. In terms of the cumulative distribution function the operator can be written as

$$F_{\mathcal{T}_k^\pi \eta^{(x,a)}}(z) = F_{\eta^{(x'_k,a'_k)}}\left(\frac{z - r_k(x,a)}{\gamma}\right), \tag{3.6}$$

which is a random variable for all $z \in \mathbb{R}$ and we have $F_{\mathcal{T}^\pi \eta^{(x,a)}}(z) = \mathbb{E}\left[F_{\mathcal{T}_k^\pi \eta^{(x,a)}}(z)\right]$, see equation (1.2). It is easy to see that (3.5) can be rewritten as a convex combination

$$\eta_{k+1}^{(x,a)} = \frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)},$$

where we define the *sample update* as

$$\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} := k\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_k^{(x,a)} - (k-1)\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}.$$

Note that it is ad-hoc not clear whether the $\eta_k^{(x,a)}$ obtained by (3.5) are in fact probability measures. Writing $\eta_{k+1}^{(x,a)}$ as a convex combination of the current distribution and the sample update distribution reduces the problem to showing that $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ is a probability measure for all k. In general, both $\eta_k^{(x,a)}$ and $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ are finite signed measures. The consideration of this problem makes up a substantial part of the analysis in Section 3.2.2. In Lemma 3.7 it is proved that we indeed have probability measures.
We call the described method *speedy categorical policy evaluation (SCPE)*, which is summarised in Algorithm 2.

Algorithm 2 Synchronous Speedy Categorical Policy Evaluation (SCPE)

Require: $\eta_k^{(x,a)} = \sum_{i=1}^N p_{k,i}^{(x,a)}\delta_{z_i}$ for fixed atoms $z_1, \dots z_N$
Input: discount factor γ, policy π, number of iterations T, initial guess η_0
$\eta_{-1} \leftarrow \eta_0$
for $k \in 0, \dots, T-1$ **do**
    $\alpha_k \leftarrow \frac{1}{k+1}$
    for $(x,a) \in \mathcal{X} \times \mathcal{A}$ **do**
        Sample $x_k' \sim p(\cdot|x,a)$, $a_k' \sim \pi(\cdot|x_k')$, $r_k \sim R(x,a)$
        $\mathcal{T}_k^\pi \eta_k^{(x,a)} \leftarrow \sum_{i=1}^N p_{k,i}^{(x_k', a_k')}\delta_{r_k + \gamma z_i}$ # Bellman update
        $\mathcal{T}_k^\pi \eta_{k-1}^{(x,a)} \leftarrow \sum_{i=1}^N p_{k-1,i}^{(x_k', a_k')}\delta_{r_k + \gamma z_i}$ # Bellman update
        # Project onto support $z_1, \dots z_N$ and calculate difference
        $\mathcal{D}_k^{(x,a)} \leftarrow k\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_k^{(x,a)} - (k-1)\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}$
        # Update η
        $\eta_{k+1}^{(x,a)} \leftarrow (1-\alpha_k)\eta_k^{(x,a)} + \alpha_k \mathcal{D}_k^{(x,a)}$
    end for
end for

3.2.1 Complexity of SCPE

For the complexity results we make almost exactly the same assumptions as [10] but, of course, in terms of categorical distributions.

Assumption 1. The state-action space is finite with $n = |\mathcal{X} \times \mathcal{A}|$ elements. The rewards are bounded by $R_{\max}$. The discount factor γ is smaller than 1 and let $\bar{\beta} := \frac{1}{1-\sqrt{\gamma}}$. Let $V_{\max} := \frac{1}{1-\gamma}R_{\max}$ be the maximal attainable return. The N fixed atoms cover all returns, $z_1 = -V_{\max}, z_N = V_{\max}$. The categorical distribution $\eta_{\mathcal{C}}$ is the unique fixed point of $\Pi_{\mathcal{C}}\mathcal{T}^\pi$. Lastly, the two initial return distribution functions are equal $\eta_{-1} = \eta_0$ and η_k are obtained by update rule (3.5).

Theorem 3.2. *Under Assumption 1, with probability at least* $1-\delta$*, the inequality*

$$\bar{\ell}_2(\eta_{\mathcal{C}}, \eta_T) \leq \sqrt{2V_{\max}}\bar{\beta}\left[\frac{\sqrt{\gamma}}{T} + \sqrt{\frac{2\log\frac{2nN}{\delta}}{T}}\right]$$

holds.

Corollary 3.3. *Under Assumption 1, for any* $0 < \epsilon \leq \sqrt{V_{\max}}$*, after*

$$T = \frac{6.53\bar{\beta}^2 V_{\max}\log\frac{2nN}{\delta}}{\epsilon^2}$$

steps of SCPE, $\bar{\ell}_2(\eta_{\mathcal{C}}, \eta_T) \leq \epsilon$ *holds with probability at least* $1-\delta$*.*

Corollary 3.4. *Under Assumption 1,* η_T *converges to* $\eta_{\mathcal{C}}$ *almost surely in* $\bar{\ell}_2$*.*

Before proving Theorem 3.2 in the next section, let's compare the complexity to the value-based method. For each timestep k the algorithm sweeps over the entire state-action space, so after T iterations, a total of $3nT$ samples (reward, next state, next action) are taken. For γ close to 1, we have $\bar{\beta} \approx 2\beta$. Recall that $V_{\max} = \beta R_{\max}$. Therefore, the sample complexity of SCPE is $\tilde{\mathcal{O}}(n\beta^3/\epsilon^2)$ (omitting the logarithmic factor). The number of atoms N only contributed to the logarithmic factor. Thus, increasing the accuracy of the distribution approximation causes only a small performance penalty.
Further, SCPE has *essentially the same* sample complexity as value-based SQL, which is $\tilde{\mathcal{O}}(n\beta^4/\epsilon^2)$. The difference in the exponent of β stems from the fact that a different metric was used. This is quite an interesting result. We *do not* need more samples when modelling the entire distribution. However, the computational complexity is higher with $\tilde{\mathcal{O}}(nN\beta^3/\epsilon^2)$ as one has to loop over all atoms when updating the return distributions. The space complexity is also higher with $\Theta(nN)$ because N atoms have to be stored for all $(x,a) \in \mathcal{X}\times\mathcal{A}$.

3.2.2 Analysis

The analysis will follow the outline of [10]. Since in DRL the return distributions depend on state, action and reward samples, it is imperative to extend the notion of random variables to random distributions. We define signed random measures according to [11].

Definition 3.5. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and define

$$M := \{\nu \text{ signed measure on } (\mathbb{R}, \mathcal{B}) : \ |\nu(B)| < \infty \text{ for all bounded } B \in \mathcal{B}\},$$

where $\mathcal{B}$ is the Borel-σ-field on $\mathbb{R}$. M is equipped with the σ-field $\mathcal{M}$, which is the smallest σ-field such that $\nu \mapsto \nu(B)$ is measurable for all $B \in \mathcal{B}$.
Measurable functions $X\colon (\Omega, \mathcal{A}, \mathbf{P}) \to (M, \mathcal{M})$, $\omega \mapsto X_\omega$ are called *signed random measures.*
The *expected measure* $\mathbb{E}[X] \in M$ is given by

$$\mathbb{E}[X](A) := \mathbb{E}[X(A)], \quad \text{where } X(A)\colon \Omega \to \mathbb{R}, \ \omega \mapsto X_\omega(A).$$

Further, $F_X(z) := \omega \mapsto F_{X_\omega}(z)$ is a random variable for all $z \in \mathbb{R}$ and we have

$$F_{\mathbb{E}[X]}(z) = \mathbb{E}[X]((-\infty, z]) = \mathbb{E}[X(-\infty, z]] = \mathbb{E}[F_X(z)]. \tag{3.7}$$

The set of all signed random measures on $E \subseteq M$ is denoted by

$$\mathscr{P}(E) := \{f\colon (\Omega, \mathcal{A}, \mathbf{P}) \to (E, \mathcal{M}\,|_E) \text{ measurable}\}.$$

Step 1. Stability

As mentioned, we do not know whether $\eta_k^{(x,a)}$ are indeed probability measures. For that reason, we first define a vector space of finite signed measures, which allows us to freely perform addition and scalar multiplication.

Definition 3.6. Let $\mathcal{L}$ be the set of finite signed Lebesgue-Stieltjes measures,

$$\begin{aligned}\mathcal{L} = \{\nu \text{ signed measure}: \ \exists F_\nu \colon \mathbb{R} \to \mathbb{R} \text{ right continuous,} \\ \nu((a,b]) = F_\nu(b) - F_\nu(a), \lim_{z\to-\infty} F_\nu(z) = 0, |\lim_{z\to\infty} F_\nu(z)| < \infty\}\end{aligned}$$

$\mathcal{L}$ is a real vector space by defining

$$(a\mu + b\nu)(A) := a\mu(A) + b\nu(A), \quad \mu, \nu \in \mathcal{L},\ a, b \in \mathbb{R},\ A \text{ a measureable set.} \tag{3.8}$$

It follows immediately from (3.8) that

$$F_{a\mu+b\nu} = aF_\mu + bF_\nu. \tag{3.9}$$

The categorical distributions are also extended to a subspace of signed measures,

$$\mathcal{P}_z \subseteq \mathcal{L}_z := \left\{\sum_{i=1}^N c_i \delta_{z_i} : \ c_i \in \mathbb{R}\right\} \subseteq \mathcal{L}.$$

The categorical projection operator $\Pi_\mathcal{C}$ can be easily applied to $\mathcal{L}$,

$$\Pi_\mathcal{C} \colon \mathcal{L} \to \mathcal{L}_z, \quad F_{\Pi_\mathcal{C}\nu}(z_i) = \frac{1}{z_{i+1} - z_i} \int_{z_i}^{z_{i+1}} F_\nu(z)\mathrm{d}z, \quad F_{\Pi_\mathcal{C}\nu}(z_N) = \lim_{z\to\infty} F_\nu(z). \tag{3.10}$$

From (3.10) and (3.9), it is not difficult to see that $\Pi_\mathcal{C} \colon \mathcal{L} \to \mathcal{L}_z$ still is a linear projection. Furthermore, from characterisation (3.6) and (3.9) it follows that $\mathcal{T}_k^\pi \colon \mathcal{L}^{\mathcal{X}\times\mathcal{A}} \to \mathcal{L}^{\mathcal{X}\times\mathcal{A}}$ is a linear mapping.

Recall that $\mathscr{P}(\mathcal{P}_z)$ is the set of random measures with values in $\mathcal{P}_z$.

Lemma 3.7. *For all $k \geq 0$ it holds that $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ and $\eta_k^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$.*

Proof. This result is proved by induction. Since we only extended $\Pi_\mathcal{C}\mathcal{T}_k^\pi$ to signed measures, it is still true that when passed a (random) probability measure $\Pi_\mathcal{C}\mathcal{T}_k^\pi$ outputs a random probability measure.

Recall that $\mathcal{D}_k[\eta_k, \eta_{k-1}] = k\Pi_\mathcal{C}\mathcal{T}_k^\pi\eta_k - (k-1)\Pi_\mathcal{C}\mathcal{T}_k^\pi\eta_{k-1}$. As the initial return distributions are identical, we have

$$\mathcal{D}_0[\eta_0, \eta_{-1}]^{(x,a)} = \Pi_\mathcal{C}\mathcal{T}_0^\pi\eta_{-1}^{(x,a)} = \Pi_\mathcal{C}\mathcal{T}_0^\pi\eta_0^{(x,a)}.$$

$\mathcal{D}_0[\eta_k, \eta_{k-1}]^{(x,a)}$ is a random probability measure and in $\mathscr{P}(\mathcal{P}_z)$, since $\eta_0^{(x,a)} \in \mathcal{P}_z$. Of course, $\eta_0^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ also.

Assume that $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ and $\eta_k^{(x,a)}$ are random probability measures. For the induction step we can relate $\mathcal{D}_{k+1}[\eta_{k+1}, \eta_k]$ to $\mathcal{D}_k[\eta_k, \eta_{k-1}]$ by observing that

$$\begin{aligned}\mathcal{D}_{k+1}[\eta_{k+1}, \eta_k]^{(x,a)} &= (k+1)\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\eta_{k+1}^{(x,a)} - k\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\eta_k^{(x,a)} \\ &= (k+1)\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi(\frac{k}{k+1}\eta_k + \frac{1}{k+1}\mathcal{D}_k[\eta_k, \eta_{k-1}])^{(x,a)} - k\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\eta_k^{(x,a)} \\ &= k\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\eta_k^{(x,a)} + \Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} - k\Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\eta_k^{(x,a)} \\ &= \Pi_\mathcal{C}\mathcal{T}_{k+1}^\pi\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)},\end{aligned}$$

where we used the fact that $\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi}$ is linear. Thus, $\mathcal{D}_{k+1}[\eta_{k+1},\eta_k]^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ also.

Since $\eta_{k+1} = \frac{k}{k+1}\eta_k + \frac{1}{k+1}\mathcal{D}_k[\eta_k,\eta_{k-1}]$ and $\mathcal{P}_z$ is a convex set, we have $\eta_{k+1}^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$. □

Step 2. Error Martingal
The history of the algorithm at time time k can be captured in form of the filtration

$$\mathcal{F}_k := \sigma\text{-field generated by } r_1(x,a), x_1', a_1', \ldots, r_k(x,a), x_k', a_k',\ (x,a) \in \mathcal{X}\times\mathcal{A}.$$

The expected update is given by

$$\mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} := \mathbb{E}\left[\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right] \overset{(3.6)}{=} k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k^{(x,a)} - (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1}^{(x,a)}.$$

The error $\epsilon_k^{(x,a)}$ and the cumulative error to the sample update $E_k^{(x,a)}$ are given by

$$\epsilon_k^{(x,a)} := \mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} - \mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)}, \quad E_k^{(x,a)} := \sum_{j=0}^{k}\epsilon_j^{(x,a)}.$$

Again, we can rewrite the update rule in terms of the expected update and the error as

$$\eta_{k+1}^{(x,a)} = \frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}(\mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} - \epsilon_k^{(x,a)}). \tag{3.11}$$

It is not immediately clear how one can turn the errors into a martingal. The following Lemma shows that we have to look at the cumulative distribution function at each atom. Lemma 3.7 and Lemma 3.8 are the core results that allow us to extend the analysis of [10] to categorical distributions. One can extend the result (3.2) from [7] in a similar fashion.

Lemma 3.8. *$\epsilon_k^{(x,a)} \in \mathscr{P}(\mathcal{L}_z)$ and $E_k^{(x,a)} \in \mathscr{P}(\mathcal{L}_z)$ for all $k \geq 0$. For each atom z_i it holds that the cumulative distribution functions of the error ϵ_k evaluated at z_i form a uniformly bounded martingal difference sequence, i.e.,*

$$\forall k \geq 0: \quad \mathbb{E}\left[F_{\epsilon_k^{(x,a)}}(z_i) \,\middle|\, \mathcal{F}_{k-1}\right] = 0, \quad \left|F_{\epsilon_k^{(x,a)}}(z_i)\right| \leq 1. \tag{3.12}$$

Proof. By Lemma 3.7 $\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ holds. It follows from (3.7) that the expected measure $\mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} \in \mathcal{P}_z$. This makes $\epsilon_k^{(x,a)}$ the difference of a random probability measure in $\mathscr{P}(\mathcal{P}_z)$ and a probability measure in $\mathcal{P}_z$. So it is an element of $\mathscr{P}(\mathcal{L}_z)$. Further, $E_k^{(x,a)}$ is the sum of elements of $\mathscr{P}(\mathcal{L}_z)$ and thus also in $\mathscr{P}(\mathcal{L}_z)$.

By definition,

$$\begin{aligned}\mathbb{E}\left[\epsilon_k^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} - \mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right] \\ &= \mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)} - \mathbb{E}\left[\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right] = 0 \in \mathcal{L}_z,\end{aligned}$$

and therefore

$$\mathbb{E}\left[F_{\epsilon_k^{(x,a)}}(z_i) \,\middle|\, \mathcal{F}_{k-1}\right] = F_{\mathbb{E}\left[\epsilon_k^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right]}(z_i) = 0 \in \mathbb{R}.$$

Furthermore, we have that

$$F_{\epsilon_k^{(x,a)}}(z_i) = F_{\mathcal{D}[\eta_k,\eta_{k-1}]^{(x,a)}}(z_i) - F_{\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)}}(z_i)$$

is the difference of a real value in $[0,1]$ and a random variable with values in $[0,1]$. This makes it a random variable which is bounded by 1.

□

Step 3. Upper bound
The following lemma shows that $\eta_k \approx \Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1}$.

Lemma 3.9. *For all $k \geq 1$ it holds that*

$$\eta_k = \frac{1}{k}(\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_0 + (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1} - E_{k-1}).$$

Proof. The equation is proved by induction. The result holds for $k = 1$ since

$$\eta_1 = \mathcal{D}[\eta_0, \eta_{-1}] - \epsilon_0 = \Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{-1} - \epsilon_0 = \Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_0 - E_0.$$

Assume that the equation holds for $k \geq 1$. From the definition of $\mathcal{D}[\eta_k, \eta_{k-1}]$ and E_k it follows that

$$\begin{aligned}
\eta_{k+1} &= \frac{k}{k+1}\eta_k + \frac{1}{k+1}(\mathcal{D}[\eta_k,\eta_{k-1}] - \epsilon_k) \\
&= \frac{k}{k+1}\eta_k + \frac{1}{k+1}(k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k - (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1} - \epsilon_k) \\
&= \frac{k}{k+1}\left(\frac{1}{k}(\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_0 + (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1} - E_{k-1})\right) \\
&\qquad + \frac{1}{k+1}(k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k - (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1} - \epsilon_k) \\
&= \frac{1}{k+1}(\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_0 + k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k - E_{k-1} - \epsilon_k) = \frac{1}{k+1}(\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_0 + k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k - E_k).
\end{aligned}$$

□

As $\mathcal{L}_z$ is a vector space, it is more convenient to work with norms instead of metrics. For that matter, we define

$$\|\nu\|_{\ell_2} := \left(\sum_{i=1}^{N-1}(z_{i+1} - z_i)F_\nu(z_i)^2 + F_\nu(z_N)^2\right)^{1/2} \tag{3.13}$$

for $\nu \in \mathcal{L}_z$. It is not difficult to see that $\|.\|_{\ell_2}$ is a norm on $\mathcal{L}_z$ and induces the metric ℓ_2 on $\mathcal{P}_z$. By taking the supremum over all state-action pairs this property extends to $\overline{\ell}_2$. Further we define

$$\|\nu\|_{\overline{\ell}_\infty} := \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\nu\|_{\ell_\infty} := \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \max_{1\leq i\leq N} |F_\nu(z_i)|$$

for $\nu \in \mathcal{L}_z$. The inequalities

$$\ell_2(\mu,\nu) = \|\mu - \nu\|_{\ell_2} \leq \sqrt{2V_{\max}}\,\|\mu - \nu\|_{\ell_\infty} \leq \sqrt{2V_{\max}} \tag{3.14}$$

hold for $\mu, \nu \in \mathcal{P}_z$. Lastly, since $\epsilon_k^{(x,a)}$ is the difference of a random probability measure and a probability measure in $\mathcal{P}_z$ (see proof of Lemma 3.8), $F_{\epsilon_k^{(x,a)}}(z_N) = 0$ and thus $F_{E_k^{(x,a)}}(z_N) = 0$ also. The inequality

$$\|E_k\|_{\bar{\ell}_2} \leq \sqrt{2V_{\max}} \|E_k\|_{\bar{\ell}_\infty} \tag{3.15}$$

follows from (3.13).

Lemma 3.10. *For all $k \geq 1$ it holds that*

$$\|\eta_{\mathcal{C}} - \eta_k\|_{\bar{\ell}_2} \leq \frac{\sqrt{\gamma}\bar{\beta}}{k}\sqrt{2V_{\max}} + \frac{1}{k}\sum_{j=1}^{k} \sqrt{\gamma}^{k-j} \|E_{j-1}\|_{\bar{\ell}_2}.$$

Proof. Again, this is proved by induction. We use the fact that $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$ (Proposition 2.5), plug in the equality from Lemma 3.9 and apply the norm inequality (3.14).
For $k = 1$ the inequality holds as

$$\begin{aligned}
\|\eta_{\mathcal{C}} - \eta_1\|_{\bar{\ell}_2} &= \|\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_{\mathcal{C}} - \Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_0 + E_0\|_{\bar{\ell}_2} \\
&\leq \sqrt{\gamma}\,\|\eta_{\mathcal{C}} - \eta_0\|_{\bar{\ell}_2} + \|E_0\|_{\bar{\ell}_2} \\
&\leq \sqrt{\gamma}\sqrt{2V_{\max}} + \|E_0\|_{\bar{\ell}_2} \leq \sqrt{\gamma}\bar{\beta}\sqrt{2V_{\max}} + \|E_0\|_{\bar{\ell}_2}.
\end{aligned}$$

Assume that the equation holds for $k \geq 1$. It also holds for $k+1$ as

$$\begin{aligned}
&\|\eta_{\mathcal{C}} - \eta_{k+1}\|_{\bar{\ell}_2} \\
&= \left\|\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_{\mathcal{C}} - \frac{1}{k+1}(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_0 + k\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_k - E_k)\right\|_{\bar{\ell}_2} \\
&= \left\|\frac{1}{k+1}(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_{\mathcal{C}} - \Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_0) + \frac{k}{k+1}(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_{\mathcal{C}} - \Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_k) + \frac{1}{k+1}E_k\right\|_{\bar{\ell}_2} \\
&\leq \frac{\sqrt{\gamma}}{k+1}\|\eta_{\mathcal{C}} - \eta_0\|_{\bar{\ell}_2} + \frac{k\sqrt{\gamma}}{k+1}\|\eta_{\mathcal{C}} - \eta_k\|_{\bar{\ell}_2} + \frac{1}{k+1}\|E_k\|_{\bar{\ell}_2} \\
&\leq \frac{\sqrt{\gamma}}{k+1}\sqrt{2V_{\max}} + \frac{k\sqrt{\gamma}}{k+1}\left[\frac{\sqrt{\gamma}\bar{\beta}}{k}\sqrt{2V_{\max}} + \frac{1}{k}\sum_{j=1}^{k}\sqrt{\gamma}^{k-j}\|E_{j-1}\|_{\bar{\ell}_2}\right] + \frac{1}{k+1}\|E_k\|_{\bar{\ell}_2} \\
&= \frac{\frac{\sqrt{\gamma}-\sqrt{\gamma}^2}{1-\sqrt{\gamma}}}{k+1}\sqrt{2V_{\max}} + \frac{\sqrt{\gamma}^2\bar{\beta}}{k+1}\sqrt{2V_{\max}} + \frac{1}{k+1}\sum_{j=1}^{k+1}\sqrt{\gamma}^{k+1-j}\|E_{j-1}\|_{\bar{\ell}_2} \\
&= \frac{\sqrt{\gamma}\bar{\beta}}{k+1}\sqrt{2V_{\max}} + \frac{1}{k+1}\sum_{j=1}^{k+1}\sqrt{\gamma}^{k+1-j}\|E_{j-1}\|_{\bar{\ell}_2}.
\end{aligned}$$

□

Step 4. Bounding the error in probability

Lemma 3.11 (Maximal Hoeffding-Azuma Inequality [10])**.** *Let $\mathcal{V} = \{V_1, V_2, \ldots, V_T\}$ be a martingal difference w.r.t. to the filtration $\mathcal{F}_k$ ($\mathbb{E}[V_k|\mathcal{F}_{k-1}] = 0$) such that $\mathcal{V}$ is uniformly bounded by $L > 0$, then for any $\epsilon > 0$ the inequality*

$$\mathbf{P}\left[\max_{1\leq k\leq T}\left|\sum_{i=1}^{k} V_i\right| > \epsilon\right] \leq 2\exp\left(\frac{-\epsilon^2}{2TL^2}\right)$$

holds.

Lemma 3.12. *For all $\epsilon > 0$ and all timesteps T, the inequality*

$$\mathbf{P}\left[\max_{1\le k\le T} \|E_{k-1}\|_{\bar{\ell}_\infty} > \epsilon\right] \le 2nN \exp\left(\frac{-\epsilon^2}{2T}\right)$$

holds.

Proof. Fix $(x,a) \in \mathcal{X} \times \mathcal{A}$ and define

$$E_k^i := F_{E_k^{(x,a)}}(z_i) = \sum_{j=0}^{k} F_{\epsilon_j^{(x,a)}}(z_i).$$

By Lemma 3.8, $V_j = F_{\epsilon_j^{(x,a)}}(z_i)$, $j = 0, \dots, T$, is a martingal difference sequence w.r.t $\mathcal{F}_j$ and uniformly bounded by 1. Therefore, we can apply the maximal Hoeffding-Azuma inequality, which takes the form

$$\mathbf{P}\left[\max_{1\le k\le T} |E_{k-1}^i| > \epsilon\right] \le 2\exp\left(\frac{-\epsilon^2}{2T}\right).$$

By taking the union over all atoms we have

$$\begin{aligned}
\mathbf{P}\left[\max_{1\le k\le T} \left\|E_{k-1}^{(x,a)}\right\|_{\ell_\infty} > \epsilon\right] &= \mathbf{P}\left[\max_{1\le k\le T}\max_{1\le i\le N} |E_{k-1}^i| > \epsilon\right] \\
&= \mathbf{P}\left[\bigcup_{i=1}^{N}\left\{\max_{1\le k\le T} |E_{k-1}^i| > \epsilon\right\}\right] \\
&\le 2N\exp\left(\frac{-\epsilon^2}{2T}\right).
\end{aligned}$$

Similarly, taking the union over all $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have

$$\mathbf{P}\left[\max_{1\le k\le T} \|E_{k-1}\|_{\bar{\ell}_\infty} > \epsilon\right] \le 2nN \exp\left(\frac{-\epsilon^2}{2T}\right).$$

□

Step 5. Concluding the proof

Proof of Theorem 3.2. By Lemma 3.10 and inequality (3.15), we find

$$\begin{aligned}
\|\eta_{\mathcal{C}} - \eta_T\|_{\bar{\ell}_2} &\le \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{\max}} + \frac{1}{T}\sum_{k=1}^{T}\sqrt{\gamma}^{T-k}\|E_{k-1}\|_{\bar{\ell}_2} \\
&\le \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{\max}} + \frac{\bar{\beta}}{T}\sqrt{2V_{\max}}\max_{1\le k\le T}\|E_{k-1}\|_{\bar{\ell}_\infty}.
\end{aligned}$$

By Lemma 3.12 the inequality

$$\mathbf{P}\left[\max_{1\le k\le T} \|E_{k-1}\|_{\bar{\ell}_\infty} > \epsilon\right] \le 2nN \exp\left(\frac{-\epsilon^2}{2T}\right) =: \delta.$$

holds. Setting δ as above and solving for ϵ yields

$$\mathbf{P}\left[\max_{1\le k\le T} \|E_{k-1}\|_{\bar{\ell}_\infty} \le \sqrt{2T\log\frac{2nN}{\delta}}\right] \ge 1-\delta.$$

Therefore, with probability at least $1-\delta$ we have

$$\overline{\ell}_2(\eta_{\mathcal{C}}, \eta_T) = \|\eta_{\mathcal{C}} - \eta_T\|_{\overline{\ell}_2} \leq \sqrt{2V_{\max}}\overline{\beta}\left[\frac{\sqrt{\gamma}}{T} + \sqrt{\frac{2\log\frac{2nN}{\delta}}{T}}\right].$$

□

Proof of Corollary 3.3. Write $T = \frac{C\bar{\beta}^2 V_{\max}\log\frac{2nN}{\delta}}{\epsilon^2}$, assume $t := \frac{\bar{\beta}^2 V_{\max}\log\frac{2nN}{\delta}}{\epsilon^2} \geq 1$ and so $\frac{1}{t} \leq \frac{1}{\sqrt{t}}$.

For $C = 2 + \sqrt{2} + 2\sqrt{1+\sqrt{2}} \leq 6.53$ it follows that

$$\overline{\ell}_2(\eta_{\mathcal{C}}, \eta_T) \leq \epsilon\sqrt{2}\left(\frac{\sqrt{\gamma}}{C\sqrt{\log\frac{2nN}{\delta}}} + \sqrt{\frac{2}{C}}\right) \leq \epsilon\sqrt{2}\left(\frac{1}{C} + \sqrt{\frac{2}{C}}\right) \leq \epsilon.$$

□

Proof of Corollary 3.4. After rearranging we have

$$\mathbf{P}\left[\overline{\ell}_2(\eta_{\mathcal{C}}, \eta_T) > \epsilon\right] \leq 2nN\exp\left(\frac{\sqrt{\gamma}\epsilon}{\sqrt{2V_{\max}}\overline{\beta}} - \frac{\gamma}{2T} - \frac{T\epsilon^2}{4V_{\max}\bar{\beta}^2}\right).$$

As $\frac{\gamma}{2T} \geq 0$, we can omit this term. Since $\exp\left(-\frac{\epsilon^2}{4V_{\max}\bar{\beta}^2}\right) < 1$, we have an inequality of form

$$\mathbf{P}\left[\overline{\ell}_2(\eta_{\mathcal{C}}, \eta_T) > \epsilon\right] \leq Cq^T, \quad C > 0, \quad 0 < q < 1.$$

Therefore $\sum_{T=0}^{\infty}\mathbf{P}\left[\overline{\ell}_2(\eta_{\mathcal{C}}, \eta_T) > \epsilon\right] < \infty$ and by the Lemma of Borel-Cantelli we have almost sure convergence. □

3.3 Policy Control

In the previous section we only considered a fixed policy π and analysed the convergence to the fixed point of $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$. If we try to extend the result to the control case, where in each timestep actions are chosen such that the expected return is maximised, we run into following problems:

i) First and foremost, $\Pi_{\mathcal{C}}\mathcal{T}\colon \mathcal{P}_z^{\mathcal{X}\times\mathcal{A}} \to \mathcal{P}_z^{\mathcal{X}\times\mathcal{A}}$ is not a contraction in $\overline{\ell}_2$. [2] showed that $\mathcal{T}$ is not a contraction in $\overline{w}_p$, see Section 1.2.2, and their provided counterexample also shows that the operator is not a contraction in $\overline{\ell}_2$. Thus, the proof of Lemma 3.10 would be incorrect if we simply swap $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$ for $\Pi_{\mathcal{C}}\mathcal{T}$.

ii) Let π_k be the greedy policy with respect to η_k such that $\Pi_{\mathcal{C}}\mathcal{T}\eta_k = \Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_k$. The update rule for the control case can then be rewritten as

$$\eta_{k+1} = \eta_k + \alpha_k(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_{k-1}}\eta_{k-1} - \eta_k) + (1-\alpha_k)(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_k - \Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_{k-1}}\eta_{k-1}). \quad (3.16)$$

Let's revisit the proof of Lemma 3.7. The sample update $\mathcal{D}_{k+1}[\eta_{k+1}, \eta_k]^{(x,a)}$ cannot be directly related to $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ anymore, as we have

$$\begin{aligned}
&\mathcal{D}_{k+1}[\eta_{k+1}, \eta_k]^{(x,a)} \\
&= (k+1)\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_{k+1}}\eta_{k+1}^{(x,a)} - k\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_k}\eta_k^{(x,a)} \\
&= (k+1)\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_{k+1}}(\frac{k}{k+1}\eta_k + \frac{1}{k+1}\mathcal{D}_k[\eta_k, \eta_{k-1}])^{(x,a)} - k\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_k}\eta_k^{(x,a)} \\
&= k\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_{k+1}}\eta_k^{(x,a)} + \Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_{k+1}}\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} - k\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_k}\eta_k^{(x,a)}.
\end{aligned}$$

But if $\pi_{k+1} \neq \pi_k$, this is in general not equal to $\Pi_{\mathcal{C}} \mathcal{T}_{k+1}^{\pi_{k+1}} \mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ and the stability result does not hold anymore.

iii) We can fix the stability problem by transforming (3.16) into

$$\eta_{k+1} = \eta_k + \alpha_k(\Pi_{\mathcal{C}} \mathcal{T}_k^{\pi_k} \eta_{k-1} - \eta_k) + (1 - \alpha_k)(\Pi_{\mathcal{C}} \mathcal{T}_k^{\pi_k} \eta_k - \Pi_{\mathcal{C}} \mathcal{T}_k^{\pi_k} \eta_{k-1}). \tag{3.17}$$

Unfortunately, now Lemma 3.9 does not hold as we have

$$\eta_k = \frac{1}{k}(\Pi_{\mathcal{C}} \mathcal{T}^{\pi_0} \eta_0 + (k-1)\Pi_{\mathcal{C}} \mathcal{T}^{\pi_{k-1}} \eta_{k-1} - E_{k-1}) + \frac{1}{k} \sum_{j=0}^{k-1} (j-1)(\Pi_{\mathcal{C}} \mathcal{T}^{\pi_{j-1}} \eta_{j-1} - \Pi_{\mathcal{C}} \mathcal{T}^{\pi_j} \eta_{j-1}).$$

However, if the policy is stable after a certain time, the second terms becomes small and we have $\eta_k \approx \Pi_{\mathcal{C}} \mathcal{T}^{\pi_{k-1}} \eta_{k-1}$ again.

There seems to be no straightforward way to solve the above problems. However, experimental results using the stable update rule (3.17) empirically showed the same convergence characteristics as in policy evaluation, see Section 5.1.

Chapter 4

Safe Reinforcement Learning

Thus far, we only considered the standard RL goal of maximising the expected return. In the tabular case, categorical DRL methods solving this objective are equivalent to their corresponding value-based methods in the sense that they yield the same policies, see Proposition 2.3.
However, even if the agent performs optimally with respect to the expected value, in practice, rare occurrences of large negative outcomes can still happen (consider the motivating example of Section 1.2.1).
As we model the entire distribution of the return in DRL, it seems natural to base policies on more than just the expected value. This way one would hope to achieve risk-averse behaviour of the agent. It turns out that incorporating risk measures in RL is a difficult task. Reference [8] contains a comprehensive survey about current approaches to risk aversion in RL and coined the term *safe reinforcement learning* as

> "the process of learning policies that maximise the expectation of return in problems in which it is important to ensure reasonable system performance and/or respect safety constraints during the learning and/or deployment processes."

There are two areas in RL, where we can implement risk measures: in the optimisation process and in the exploration process. In the following, only the former will be considered.

4.1 Risk-Sensitive Optimisation Criteria

Risk measures in the optimisation process can be introduced by simply replacing the optimisation objective. Three common optimisation criteria will be discussed which can be used instead of the expected reward maximisation [8].

Worst-Case Criterion. This approach aims to maximise the expected return with respect to the worst case scenario. One can write this objective as

$$\max_{\pi\in\Pi} \min_{\omega\in\Omega^\pi} \mathbb{E}_\omega \left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)\right],$$

where Π is the set of all policies and Ω^π is the set of all possible trajectories $\omega = (x_0, a_0, x_1, a_1, \dots)$ following the policy π.
One immediate drawback of this objective is the fact that very rare events with negative outcome have a big effect on the policy. So, this objective is often too pessimistic

and algorithms which compromise between the optimism of the expected return and the pessimism of the worst case return can be used instead.

Risk-Sensitive Criterion. In this approach a risk-sensitivity parameter $\beta \in \mathbb{R}$ controls the desired level of risk. $\beta = 0$ implies risk neutrality, $\beta < 0$ implies a risk-seeking and $\beta > 0$ a risk-averse behaviour.
One option is the use of exponential utility functions. Here the objective is

$$\max_{\pi \in \Pi} -\frac{1}{\beta} \log \mathbb{E}_\pi \left[\exp \left(-\beta R \right) \right] = \max_{\pi \in \Pi} -\frac{1}{\beta} \log \mathbb{E}_\pi \left[\exp \left(-\beta \sum_{t=0}^{\infty} \gamma^t R_t \right) \right],$$

where $\mathbb{E}_\pi [.]$ denotes that the returns are obtained by following the policy π. Performing a Taylor expansion yields

$$\max_{\pi \in \Pi} -\frac{1}{\beta} \log \mathbb{E}_\pi \left[\exp \left(-\beta R \right) \right] = \max_{\pi \in \Pi} \mathbb{E}_\pi [R] - \frac{\beta}{2} \mathbb{V}_\pi [R] + \mathcal{O}(\beta^2).$$

This shows that for $\beta > 0$ variability is penalized and for $\beta < 0$ variability is encouraged. Even though this method is the best analysed in literature, there are no model-free algorithms, which would be required for RL.

A second option is a weighted sum of expected return and risk measure Ψ,

$$\max_{\pi \in \Pi} \mathbb{E}_\pi [R] - \beta \Psi(R).$$

Possible risk measures could be the variance of the return or the probability of terminating in an undesirable state. As shown above, using the variance as risk measure is essentially equivalent to the exponential utility function approach. However, both methods provide different perspectives for designing algorithms and mathematical analysis.

Constrained Criterion. The expected return is maximised subject to constraints,

$$\max_{\pi \in \Pi} \mathbb{E}_\pi [R] \quad \text{subject to} \quad h_i(R) \leq \alpha_i.$$

This is equivalent to constraining the set of policies to a subset $\Gamma \subset \Pi$ and then maximising the expected value over policies in Γ. Choices for constraints could be the restriction of variance $\mathbb{V}_\pi [R] \leq \alpha$ or a minimum threshold for the expected return $\mathbb{E}_\pi [R] \geq \alpha$.

4.2 Problems of Risk-Averse Policy Iteration

The policy iteration algorithm is a theoretical improvement scheme, where one tries to create a better policy from a given policy π [23, Chapter 4]. This way one gets a chain of policies, $\pi_1 \preceq \pi_2, \preceq \pi_3 \preceq \ldots$, which converges to an optimal policy π^* for the standard RL objective. Generally, an optimal policy is one dominating every other policy, $\pi \preceq \pi^*$. In standard RL, a policy is better than another one if the expected return is greater in each state,

$$\pi \preceq \pi' \iff \mathbb{E} \left[Z^\pi(x, \pi(x)) \right] \leq \mathbb{E} \left[Z^{\pi'}(x, \pi'(x)) \right] \text{ for all } x \in \mathcal{X}. \tag{4.1}$$

For a given policy π we can make a better one π' by choosing

$$\pi'(x) \in \arg\max_{a \in \mathcal{A}} \mathbb{E} \left[Z^\pi(x, a) \right] \text{ for all } x \in \mathcal{X}. \tag{4.2}$$

Note that we simply *ordered* the actions according to their expected value and chose the best one. Furthermore, we have $\mathbb{E}\left[Z^{\pi}(x,\pi(x))\right] \leq \mathbb{E}\left[Z^{\pi}(x,\pi'(x))\right]$ for all states. This inequality means that following π' for only one decision and then π thereafter gives a better expected return. It turns out that this fact is sufficient to guarantee that the new policy π' is in indeed better.

Theorem 4.1 (Policy improvement theorem)**.** *Let π and π' be two policies such that*

$$\forall x \in \mathcal{X}: \;\; \mathbb{E}\left[Z^{\pi}(x,\pi(x))\right] \leq \mathbb{E}\left[Z^{\pi}(x,\pi'(x))\right].$$

Then $\pi \preceq \pi'$, i.e.,

$$\forall x \in \mathcal{X}: \;\; \mathbb{E}\left[Z^{\pi}(x,\pi(x))\right] \leq \mathbb{E}\left[Z^{\pi'}(x,\pi'(x))\right].$$

As we model the entire return distribution in DRL, it seems reasonable to try to establish decision rules similar to (4.2), which aim to achieve risk-averse objectives. This yields a new class of policies, which can be defined as follows [5].

Definition 4.2. *Risk-sensitive policies* are policies which depend upon more than the mean of the outcomes.

Let's consider some risk-averse decision rules:

- For the worst-case criterion, we could maximise a quantile q close to 0, i.e.,

$$\pi'(x) \in \arg\max_{a\in\mathcal{A}} F^{-1}_{Z^{\pi}(x,a)}(q). \tag{4.3}$$

- For the risk-sensitive criterion, we could simply consider the weighted sum of expected value and standard deviation with parameter $\beta > 0$, i.e.,

$$\pi'(x) \in \arg\max_{a\in\mathcal{A}} \mathbb{E}\left[Z^{\pi}(x,a)\right] - \beta\sqrt{\mathbb{V}\left[Z^{\pi}(x,a)\right]}. \tag{4.4}$$

- For the constrained criterion, we could specify a risk parameter ρ such that elements of the set $\mathcal{A}_{\rho}(x) := \{a \in \mathcal{A} : \mathbf{P}\left[Z^{\pi}(x,a) < 0\right] \leq \rho\}$ are considered safe actions at x, as they yield a negative reward with only small probability. Then we choose

$$\pi'(x) \in \begin{cases} \arg\max_{a\in\mathcal{A}_{\rho}(x)} \mathbb{E}\left[Z^{\pi}(x,a)\right] & \text{if } \mathcal{A}_{\rho}(x) \neq \emptyset, \\ \arg\min_{a\in\mathcal{A}} \mathbf{P}\left[Z^{\pi}(x,a) < 0\right] & \text{otherwise.} \end{cases} \tag{4.5}$$

For these decision rules we order policies analogously to (4.1) and (4.2). The question arises whether or not these rules result in policy improvement. For example, the last one only makes sense if there is only one final reward in a trajectory. Reference [21] established a framework that allows us to investigate this problem.

Definition 4.3. For a MDP $\langle \mathcal{X}, \mathcal{A}, R, P\rangle$ let Ω_X denote the set of all random trajectories starting in the random state X, i.e.,

$$\Omega_X = \{(X_0, A_0, X_1, A_1, \dots) : \;\; X_0 = X, \;\; A_t \in \mathcal{A}, \;\; X_{t+1} \sim P(\cdot|X_t, A_t)\}.$$

Here, A_t are also considered to be random variables with values in $\mathcal{A}$.

The preference over actions can be abstracted by considering an order over the random trajectories. That is, for each random state X we have an order relation $(\Omega_X, \preceq_X)$. For now, we only assume reflexivity and transitivity of $\preceq_X$.

For example, in standard RL the trajectory $\omega_1(x) = (x, A_{1,0}, X_{1,1}, A_{1,1}, \dots)$ is better than trajectory $\omega_2(x) = (x, A_{2,0}, X_{2,1}, A_{2,1}, \dots)$ if the expected (discounted) return is greater, i.e.,

$$\omega_1(x) \preceq_x \omega_2(x) \Leftrightarrow \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_{1,t}, A_{1,t})\right] \leq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_{2,t}, A_{2,t})\right].$$

Definition 4.4. For each (stationary) deterministic policy π and initial state x there exists one corresponding random trajectory

$$\omega^\pi(x) := (X_0, A_0, X_1, A_1, \dots), \quad X_0 = x, \;\; A_t = \pi(X_t), \;\; X_{t+1} \sim P(\cdot|X_t, A_t)$$

in Ω_x. The order relations at each state $\preceq_x$ naturally induce the order

$$\pi \preceq \pi' :\Longleftrightarrow \omega^\pi(x) \preceq_x \omega^{\pi'}(x) \text{ for all } x \in \mathcal{X}$$

on the set of deterministic policies. An optimal policy is then a policy π^* such that $\pi \preceq \pi^*$ for all policies π.

The policy iteration algorithm can now be written in terms of the orders $\preceq_x$. As previously, we want to find a new policy such that using it only for first decision and then using the old one thereafter yields a better performance.
For a given policy π choose π' such that

$$\omega^\pi(x) \preceq_x \omega^\pi(x, \pi'(x)) := (x, \pi'(x), X_1, \pi(X_1), X_2, \pi(X_2), \dots), \tag{4.6}$$

where $X_1 \sim P(\cdot|x, \pi'(x))$ and $X_{t+1} \sim P(\cdot|X_t, \pi(X_t))$.

Now we can answer the question under which conditions we have policy improvement ($\pi \preceq \pi'$) and convergence to an optimal policy [21].

Theorem 4.5. *Assume that the relations $\preceq_X$ satisfy the following conditions:*
For all $\omega_1, \omega_2, \omega_3 \in \Omega_X$

i) (reflexive) $\omega_1 \preceq_X \omega_1$

ii) (complete) either $\omega_1 \preceq_X \omega_2$ or $\omega_2 \preceq_X \omega_1$,

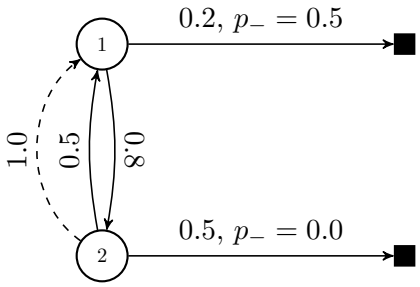
iii) (transitive) $\omega_1 \preceq_X \omega_2$ and $\omega_2 \preceq_X \omega_3$ implies $\omega_1 \preceq_X \omega_3$,

iv) (monotonicity) if $\omega_1 \preceq_X \omega_2$, then $(X', A', \omega_1) \preceq_{X'} (X', A', \omega_2)$ for all random states X' and random actions A' such that $X \sim P(\cdot|X', A')$.

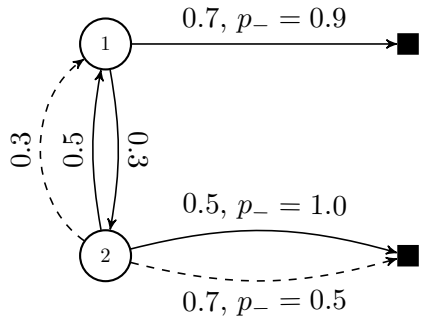
v) (countable transitivity) For a random state X let $\omega_j = (X, A_{j,1}, X_{j,2}, A_{j,2}, \dots) \in \Omega_X$ such that ω_j matches ω_0 up to state j ($A_{j,k} = A_{0,k}$ and $X_{j,k} = X_{0,k}$ for all $k \leq j$). If $\omega_1 \preceq_X \omega_2 \preceq_X \omega_3 \dots$, then $\omega_j \preceq_X \omega_0$ for all j.

Then policies obtained by (4.6) satisfy $\pi \preceq \pi'$. Furthermore, if the state and actions space is finite, the sequence converges to an optimal policy in a finite number of steps.

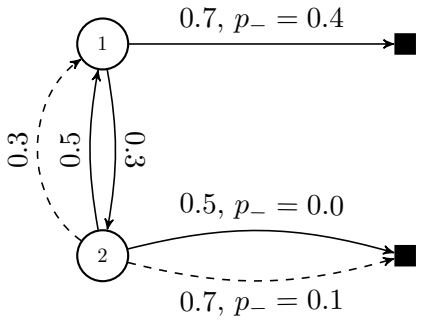
Conditions *i)* to *iii)* are standard for order relations. However, conditions *iv)* and *v)* are highly restrictive. Finding a risk-averse action ordering that guarantees improvement seems unlikely. In the following, three MDPs (found with a computer program) will be presented, which show that the rules (4.3), (4.4) and (4.5) indeed do *not* lead to global policy improvement. This conclusion seems very disappointing; however, applying the rules in practice achieved surprisingly good results, see Section 5.2.



(a) Worst-Case Criterion



(b) Variance Penalized Criterion



(c) Probability Constrained Case

Figure 4.1: Three MDPs with random state transitions (edge weights denote transition probability). In state 2 there are two possible actions (a_1 solid and a_2 dashed line). When transitioning to a terminal state rewards are either -1 or $+1$, otherwise 0. p_- is the probability of the negative reward.

Consider the MDP in Figure 4.1a and decision rule (4.3) with $q = 0.2$. We assume no discounting, $\gamma = 1$. As rewards are only given when transitioning to a terminal state, we can calculate the probability of return -1 when choosing action a_1 in state 2 by solving $p_1 = 0.2 \cdot 0.5 + 0.8p_2$, $p_2 = 0.5p_1$. We have $p_1 = 1/6$ and $p_2 = 1/12$. Recall that $F_Z^{-1}(q) = \inf\{t : F_Z(t) \geq q\}$ and as both values are smaller than 0.2, the q-quantiles at both states are equal to 1.0. So from the point of state 2 both actions look equivalent by rule (4.3). But it is clear that choosing a_2 yields $p_1 = p_2 = 0.5$, so both q-quantiles are equal to -1.0. Even though both actions looked equivalent, we got worse performance and policy improvement does not hold.

Now lets consider the MDP in Figure 4.1b and rule (4.4) with $\beta = 1$ and discount factor $\gamma = 0.9$. One can calculate the expected return and the variance of the return by solving a system of linear equations, see [22]. The calculations are omitted for the sake of brevity. When choosing a_1 we have expected returns $v_1 \approx -0.7911$ and $v_2 \approx -0.856$ with standard deviations $s_1 \approx 0.5341$ and $s_2 \approx 0.3692$. When choosing a_2 for one decision and then a_1 thereafter, we have an expected return of $\bar{v}_2 = 0.7 \cdot 0.0 + 0.3\gamma v_1 \approx -0.2136$ with standard deviation $\bar{s}_2 = (0.7 \cdot (1.0^2 + 0.0^2) + 0.3\gamma^2(s_1^2 + v_1^2) - \bar{v}_2^2)^{1/2} \approx 0.9358$ in state 2 (formula for

variance of mixture distribution; transitioning to terminal state has expected return 0.0 with variance 1.0). By rule (4.4) we would choose a_2. However, choosing a_2 every time yields $v_1 \approx -0.604$ with $s_1 \approx 0.7482$, which is not an improvement.

Lastly, consider the MDP in Figure 4.1c and rule (4.5) with $\rho = 0.2$. We assume $\gamma = 0.9$. Again, as rewards are only given when transitioning to a terminal state, we can calculate the probabilities of a negative reward (choosing a_1) by solving $p_1 = 0.7 \cdot 0.4 + 0.3p_2$, $p_2 = 0.5p_1$. We get $p_1 \approx 0.3294$ and $p_2 \approx 0.1647$. Solving for the expected return gives $v_1 \approx 0.313$ and $v_2 \approx 0.6409$. By rule (4.5) we would choose a_2 as $0.7 \cdot 0.1 + 0.3 \cdot p_1 \leq \rho$ and $0.7 \cdot 0.8 + 0.3\gamma v_1 \geq v_2$ (0.8 is the expected reward when transitioning to the bottom right terminal state). But choosing a_2 yields $p_1 \approx 0.3308$, which means that we just got a riskier policy, thus no global improvement.

4.3 Current Approaches to Risk-Averse DRL

Modelling the return distribution provides many possibilities to incorporate risk measures in the learning process. As great advances in both the theoretical and practical side of DRL were made in the last three years, including risk-aversion strategies in DRL is subject of active research.

There are approaches that make use of risk-averse action selection. Reference [5] approximates the quantile function with a neural network. Actions are then selected to maximise the expected value distorted by a continuous monotonic function $\beta\colon [0,1] \to [0,1]$,

$$\pi'(x) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_\beta \left[Z^\pi(x,a) \right] = \arg\max_{a \in \mathcal{A}} \int_{\mathbb{R}} z \frac{\partial}{\partial z} (\beta \circ F_{Z^\pi(x,a)})(z) \mathrm{d}z.$$

The function β is called *distortion risk measure* and it is known that this decision rule is equivalent to maximising the expected value of some utility function.
Recent work [12] uses the *conditional value at risk (CVAR)*, a popular risk measure in finance which is loosely speaking the expected value of the tail of a distribution,

$$\pi'(x) \in \arg\max_{a \in \mathcal{A}} \mathrm{CVAR}_\alpha(\tilde{Z}^\pi(x,a)) = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[\tilde{Z}^\pi(x,a) \,\middle|\, \tilde{Z}^\pi(x,a) \leq F^{-1}_{\tilde{Z}^\pi(x,a)}(\alpha) \right].$$

Before applying the CVAR_α, the return distribution is transformed to accelerate convergence, denoted by $\tilde{Z}^\pi(x,a)$. It is worth mentioning that also categorical distributions were used as approximation method.
In both mentioned papers sophisticated algorithms were developed which achieved good experimental results. However, these methods should be subject to the same scepticism as discussed in the precious section and theoretical guarantees are much needed.

Another recent work [15] considers risk-aversion as a secondary objective. When actions appear equivalent under the expected return, the one is chosen which stochastically dominates the others (in the second order),

$$Z^\pi(x,a) \preceq_{(2)} Z^\pi(x,\pi'(x)) \iff \forall \alpha \in \mathbb{R} : \int_{-\infty}^{\alpha} F_{Z^\pi(x,a)}(z)\mathrm{d}z \leq \int_{-\infty}^{\alpha} F_{Z^\pi(x,\pi'(x))}(z)\mathrm{d}z.$$

An implication of this choice is that $Z^\pi(x,\pi'(x))$ has the lowest variance amongst its contenders. Even though it is argued that equivalence in the expected outcome is frequent in financial optimisation, the applicability in other domains is questionable.

Another popular approach to safe RL are policy gradient methods. Here we directly approximate the policy with a differentiable function, π_θ, and use stochastic gradient ascent/descent methods to move towards a policy, which achieves (local) optimal performance for a certain objective.
Reference [1] describes a template for solving risk-constrained optimisation objectives of form

$$\max_\theta \mathbb{E}\left[Z^{\pi_\theta}(x_0)\right] \quad \text{subject to} \quad \Psi(Z^{\pi_\theta}(x_0)) \leq \alpha$$

for an initial state x_0, return distribution $Z^{\pi_\theta}(x_0) = Z^{\pi_\theta}(x_0, \pi_\theta(x_0))$ and arbitrary risk measure Ψ (like variance or CVAR). This objective can be brought in a relaxed form

$$\max_\lambda \max_\theta \left(\mathbb{E}\left[Z^{\pi_\theta}(x_0)\right] + \lambda(\Psi(Z^{\pi_\theta}(x_0)) - \alpha)\right)$$

using a Lagrangian approach. For such problems convergence to an local optimal policy can be established; however, estimating the policy gradient of the risk measure is challenging. In the recent paper [20] the CVAR is estimated and a method called *sample-based distributional policy gradients (SDPG)* is employed in an actor-critic policy gradient algorithm. Again, experimental results indicate the advantages of this method but convergence results remain to be established.

In summary, there are many perspectives on safe DRL. Even though the proposed methods are theoretically sound and yield impressive empirical results, there seems to be lack of convergence results. It will be interesting to follow the development of the mathematical theory in the future.

Chapter 5

Experimental Results

5.1 Combination Lock

Consider the combination lock environment [9]. Here we have a set of 500 states x_i, which are arranged in a chain. In each state we can choose between two actions `LEFT` and `RIGHT`, see Figure 5.1. Choosing `RIGHT` takes us to state x_{i+1} but yields a reward of -0.01. Taking `LEFT` brings us to a previous state with probability $p(x_k|x_i, \texttt{LEFT}) \propto \frac{1}{i-k}$ and yields reward 0. Transitioning to the goal state x_{500} gives $+15$ reward.

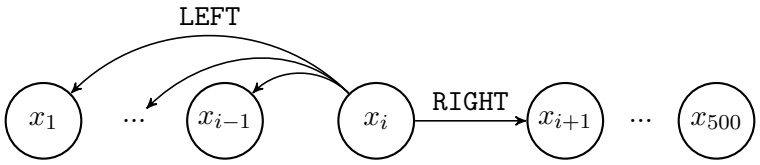
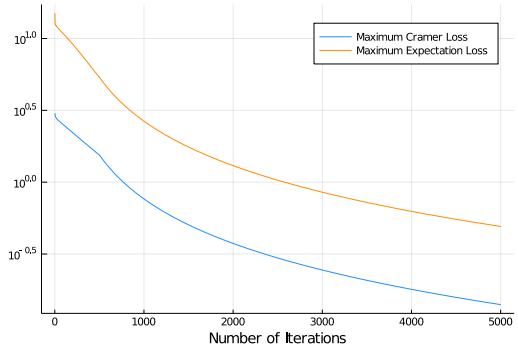


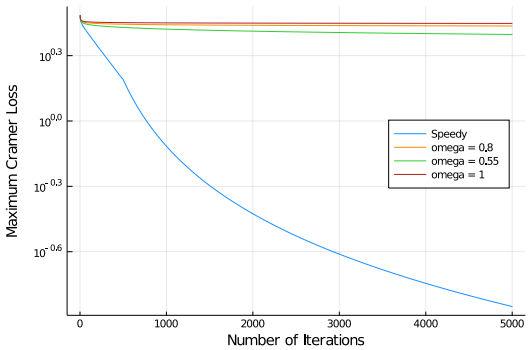
Figure 5.1: Combination lock environment

The action `RIGHT` brings us closer to the goal state but yields a negative reward, whereas the action `LEFT` has no immediate negative reward but moves us further from x_{500}. The rewards are setup such that choosing `RIGHT` in all states is the optimal policy. This makes an interesting control problem because the long chain has to be essentially solved right to left. It is also a good benchmark for policy evaluation because the trajectories are long and when choosing `LEFT` there are a lot of possible successor states. For this reasons, the complexity analysis of Section 3.2 is tested on this environment.

For the experiment, the SCPE algorithm (Algorithm 2) with a discount factor of $\gamma = 0.999$ ($\bar{\beta} \approx 2000$) and 51 equally spaced atoms between -10 and 15 was used. For comparison, the standard Q-learning update rule (2.3) with a polynomial learning rate $1/(k+1)^\omega$, $\omega \in \{0.55, 0.8, 1\}$ was also tested in a synchronous fashion. The experiment was repeated 10 times for randomly generated initial distributions and the results were averaged. An accurate estimation $\hat{\eta}_\mathcal{C}$ of $\eta_\mathcal{C}$ was obtained by performing 50 000 iterations.
In Figure 5.2a the convergence in the Cramér distance is compared to the convergence in expectation. You can see $\overline{\ell}_2(\eta_k, \hat{\eta}_\mathcal{C})$ in blue and $\sup_{(x,a)} |\mathbb{E}_{Z\sim\eta_k^{(x,a)}}[Z] - \mathbb{E}_{Z\sim\hat{\eta}_\mathcal{C}^{(x,a)}}[Z]|$ in orange. The convergence in the Cramér distance is as fast if not faster than the convergence in expectation which is backed up by the theoretical results in 5.1.
Furthermore, there is a great performance benefit if one chooses the SCPE update rule over the standard Q-learning update rule, see Figure 5.2b. The plot closely resembles the results of [9].



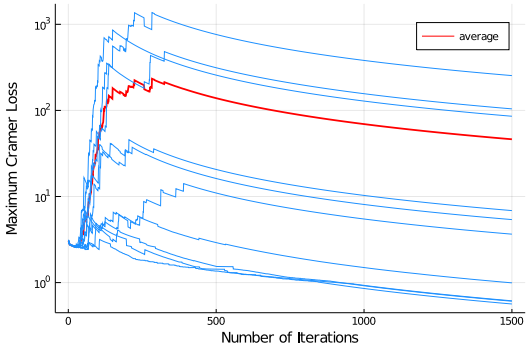
(a) Convergence in Cramér distance vs. Convergence in expectation



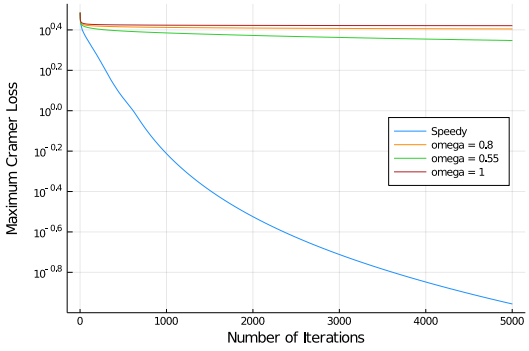
(b) SCPE and polynomial learning rates

Figure 5.2: Policy evaluation in the Combination Lock environment.

The same experiment was repeated for the control case. In Figure 5.3a you can see the instability issue for the unchanged update rule as discussed in Section 3.3. However, using the adjusted update rule (3.17) yields the same convergence characteristics as policy evaluation, compare Figure 5.2b and 5.3b.



(a) Instabilities of unchanged update rule



(b) Similar convergence to policy evaluation

Figure 5.3: Policy control in the Combination Lock environment.

5.2 Gridworld with Lake

Let's consider the gridworld environment from the introduction again. Recall that the environment only accepts the agents action 90% of the times, otherwise the agent will be placed in a random direction. If the agent lands on a blue field, the lake, there is a 1% chance of drowning and receiving reward -100. The agent starts at cell S and the objective is to reach cell G, which yields $+100$ reward. There is no reward otherwise. Lastly, we have a discount factor of $\gamma = 0.95$.

The rewards, the drowning probability and the discount factor are carefully chosen such that the optimal policy with respect to the expected return goes straight through the lake. Equipped with the theoretical advantages and limitations of DRL, we like to try to formulate an algorithm, which yields safer policies that avoid the lake.

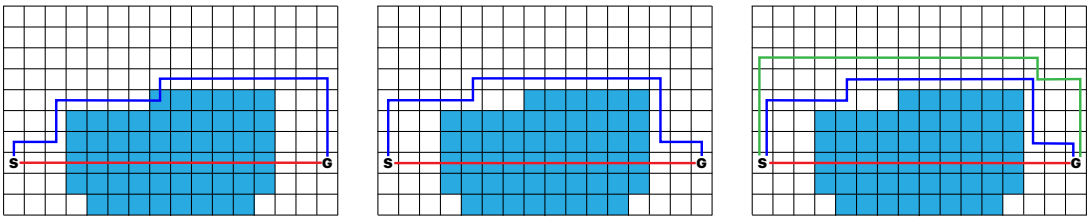
The resulting Algorithm 3 is essentially distributional Q-learning with two key differences. Firstly, we allow a more generic action selection, where the decision rules of Section 4.2 will be implemented. Secondly, we apply the algorithm in a synchronous fashion. That is,

we loop over the entire state-action space. This way we neglect the exploration process, which allows us to compare the different decision rules on a fair basis.

Algorithm 3 Synchronous Risk-Averse Policy Control

Require: $\eta^{(x,a)} = \sum_{i=1}^{N} p_i^{(x,a)} \delta_{z_i}$ for fixed atoms $z_1, \dots z_N$
Input: discount factor γ, number of iterations T, initial guess η_0
$\eta \leftarrow \eta_0$
for $k \in 0, \dots, T-1$ **do**
    for $(x, a) \in \mathcal{X} \times \mathcal{A}$ **do**
        Sample $x'_k \sim p(\cdot|x, a)$, $r_k \sim R(x, a)$
        $a^* \leftarrow \text{greedy}(\eta, x'_k)$ # Risk-averse action selection
        $\mathcal{T}_k^{\pi} \eta^{(x,a)} \leftarrow \sum_{i=1}^{N} p_i^{(x'_k, a^*)} \delta_{r_k + \gamma z_i}$ # Bellman update
        $\eta^{(x,a)} \leftarrow (1 - \alpha_k)\eta^{(x,a)} + \alpha_k \Pi_{\mathcal{C}} \mathcal{T}_k^{\pi} \eta^{(x,a)}$ # Update η
    end for
end for

For the experiment, 51 equally spaced atoms on the interval $[-100, 100]$ were used. The initial guess η_0 was generated randomly. Lastly, a total of $T = 250$ iterations were performed with stepsize parameters $\alpha_k = (1 + \lfloor k/32 \rfloor)^{-1}$ (chosen through testing).



(a) Worst-Case Criterion (b) Risk-Sensitive Criterion (c) Constrained Criterion

Figure 5.4: Resulting policies by performing risk-averse policy control in the gridworld.

Write $Z(x, a) \sim \eta^{(x,a)}$. In Figure 5.4a the actions were selected with $q \in \{0.1, 0.01\}$ by

$$\pi'(x) \in \arg\max_{a \in \mathcal{A}} F^{-1}_{Z(x,a)}(q),$$

in Figure 5.4b with $\beta \in \{0.1, 0.5\}$ by

$$\pi'(x) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[Z(x, a)\right] - \beta \sqrt{\mathbb{V}\left[Z(x, a)\right]},$$

and in Figure 5.4c with $\rho \in \{0.1, 0.01, 10^{-5}\}$ by

$$\pi'(x) \in \begin{cases} \arg\max_{a \in \mathcal{A}_\rho(x)} \mathbb{E}\left[Z(x, a)\right] & \text{if } \mathcal{A}_\rho(x) = \{a : \mathbf{P}\left[Z(x, a) < 0\right] \leq \rho\} \neq \emptyset, \\ \arg\min_{a \in \mathcal{A}} \mathbf{P}\left[Z(x, a) < 0\right] & \text{otherwise.} \end{cases}$$

For parameters $q = 0.1$ and $\rho = 0.1$ the main trajectory of the resulting policy (red) goes straight through the lake. This makes sense as going through the lake has approximately 9.5% chance of drowning. $\beta = 0.1$ also has the same result. The choice of $q = 0.01$, $\beta = 0.5$ and $\rho = 0.01$ resulted in the blue paths. Here the objective of avoiding the lake was successfully achieved (except at one cell in Figure 5.4a). Only with the constrained

criterion an even more risk-averse policy was obtained. The choice of $\rho = 10^{-5}$ resulted in an even greater detour around the lake (green trajectory in Figure 5.4c).
In summary, even though improvement in each step is not guaranteed (see Section 4.2), risk-sensitive policies were achieved quite robustly with each decision rule.

5.3 Sepsis Treatment

Sepsis is the third leading cause of death worlwide and there have been efforts to find optimal sepsis treatment strategies with RL methods [13]. Having detailed information about the consequences of treatment decision is imperative in medical problems. Therefore, the use of DRL methods is of great benefit. While the methodology of [13] met criticism [4], we will follow their setup assuming it is valid and focus on the advantages of using distributional algorithms.

Sepsis patient information was gained by querying the MIMIC-III (Medical Information Mart for Intensive Care III) database. After preprocessing the data, a k-means clustering was performed, resulting in 800 unique patient states. A simulator was created from the clustered dataset which was used as environment for policy evaluation and control. Treatment comprises the administration of intravenous fluids and vasopressors, which was discretized into 25 possible actions. The recovery of a patient is rewarded by $+100$ and the death of a patient penalized with -100.

As the simulator is modelled from patient state trajectories following the treatment decisions of clinicians, choosing a random action in the simulator corresponds to a clinician's decision. Performing speedy policy evaluation for random actions with Algorithm 2 gives interesting insights into the consequences of treatment decisions. Again, 51 equally spaced atoms on the interval $[-100, 100]$ were used and 1000 iterations were performed. As we care more about the overall outcome, rather than immediate rewards, the discount factor was set to $\gamma = 0.99$.
Results show that less than 1% of the clinicians' actions predominantly lead to a negative outcome (above 90% chance of the patient dying). Also about 1% of the actions had a highly bimodal return distribution. These are actions, where a positive outcome is almost as likely as a negative outcome. However, the majority of actions ($\approx 67\%$) lead predominantly to recovery but had also a not negligible probability (5% to 20%) of a negative outcome. With a recovery probability over 95% around every seventh action ($\approx 14\%$) could be considered safe. In Figure 5.5 examples of the various discussed return distributions can be seen.

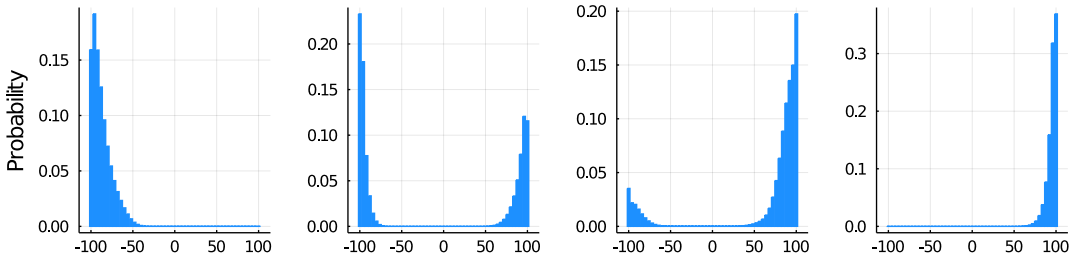


Figure 5.5: Examples of return distributions when following the clinicians' policy. From left to right: negative outcome, bimodal distribution, slightly bimodal distribution, positive outcome.

Performing policy evaluation and policy improvement with respect to the expected return iteratively yielded an optimal policy π^*. In Figure 5.6 you can see the approximated return distributions at each initial state choosing actions according to π^*. Quite surprisingly, it was possible to select actions such that from every initial state the recovery of the patient is guaranteed with very little variance in the treatment length. Almost for every initial state the return distribution is right skewed corresponding to fast patient recovery. There are very few distributions which spike at zero. They correspond to erroneous states in the simulator for which the agent had virtually no possibility of leaving the state and thus remained in the state forever and never got any reward. From this result is clear that one could not do any better using risk-sensitive policies.

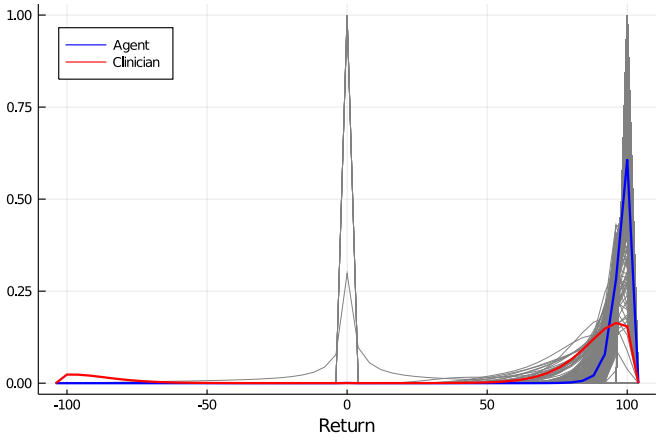


Figure 5.6: Return distributions of the resulting agent for all initial states in gray and average return distribution (weighted by the number of occurrence as initial state) in blue. Average return distribution of the clinicians' policy in red for comparison.

However, this performance was only achieved on the data the agent was trained on. Therefore, a second simulator was created for evaluation purposes with patient state trajectories unseen by the agent. Sadly, due to insufficient data and the problem of finding a good clustering, the second simulator had very different state transition probabilities compared to the first one and following the policy π^* lead to many infinite episodes. In these episodes the agent remained in the same state when choosing the action according to π^*, never reaching a terminal state. To solve this problem, a stochastic policy was derived from the return distributions of the deterministic one. This stochastic policy chooses the best action 50% of the times, the second best 25% of the times and so on. The worst action gets the remaining probability. The stochastic policy was more robust towards different clusterings of the data and "kicked" the agent out of such looping states.

In order to test the performance of the resulting policy, 10 000 patient state trajectories were simulated. With a mean return of 74.38 and a recovery rate of 91.88 the agent outperformed the clinicians' policy with a mean return of 63.54 and a recovery rate of 86.16 on both metrics. However, this is not the optimal performance achieved in the training process. This is solely owed to the patient state representation by the clustering. Experimenting with different state representation techniques remains for future work.

Chapter 6

Discussion

In many papers the experimental results of learning the return distribution with neural networks when optimising for the expected return were celebrated. In this work, a particular DRL framework, modelling the return with categorical distributions, was reviewed. Instead of considering function approximation methods, the focus was on tabular methods and it was argued that the power of DRL lies in the ability of making decision based on more than just the mean of outcomes.

In Chapter 3 it was shown that Q-learning-like tabular categorical DRL algorithms have essentially the same sample complexity as their standard RL counterparts. This means that we are able to gain significantly more information about the return from the same number of state transitions observed by the agent. This novel theoretical result was confirmed empirically in Section 5.1.

In Chapter 4 we concluded that even though the DRL framework allows many new possibilities to incorporate risk measures in the learning process, unfortunately, simple policy iteration approaches can not guarantee improvement and may lead to unintuitive policies. A short, non-exhaustive survey showed that there is a very active research community around risk-averse DRL; however, theoretical results are needed.

Unimpressed by the negative results about the aforementioned risk-averse policy iteration methods, they were tested on a gridworld toy problem in Section 5.2. The objective of detouring around a lake was achieved quite robustly, whereas standard RL methods would go straight through the lake and risk drowning.

Furthermore, the application of DRL was considered in the high-stakes problem of sepsis treatment. Policy control with respect to the expected return lead to an optimal treatment policy with zero patient mortality on the training environment. Thus, there was no room for improvement by using risk-sensitive policies. However, this conclusion could only be made by having access to the approximate return distributions. Additionally, analysing the return distributions gave detailed insights into the consequences of clinicians' treatment decisions.

Finally, distributional reinforcement learning is a very young research field and it will be exciting to follow future developments.

Bibliography

[1] P. L. A. and M. C. Fu, *Risk-sensitive reinforcement learning: A constrained optimization viewpoint*, CoRR, abs/1810.09126 (2018).

[2] M. G. Bellemare, W. Dabney, and R. Munos, *A distributional perspective on reinforcement learning*, in Proceedings of the 34th International Conference on Machine Learning – Volume 70, ICML'17, JMLR.org, 2017, p. 449–458.

[3] R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.

[4] R. J. Bradford, C. J. Sangwin, S. P. Shashikumar, and S. Nemati, *Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care?*, CoRR, abs/1902.03271 (2019).

[5] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, *Implicit quantile networks for distributional reinforcement learning*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, pp. 1096–1105.

[6] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, *Distributional reinforcement learning with quantile regression*, CoRR, abs/1710.10044 (2017).

[7] E. Even-Dar and Y. Mansour, *Learning rates for Q-learning*, J. Mach. Learn. Res., 5 (2004), p. 1–25.

[8] J. García, Fern, and o Fernández, *A comprehensive survey on safe reinforcement learning*, Journal of Machine Learning Research, 16 (2015), pp. 1437–1480.

[9] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos, *Reinforcement learning with a near optimal rate of convergence*, INRIA, 00636615v2 (2011).

[10] ——, *Speedy Q-learning*, in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, Inc., 2011, pp. 2411–2419.

[11] O. Kallenberg, *Random Measures, Theory and Applications*, Springer, 2017.

[12] R. Keramati, C. Dann, A. Tamkin, and E. Brunskill, *Being optimistic to be conservative: Quickly learning a cvar policy*, arXiv, abs/1911.01546 (2020).

[13] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, *The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care*, Nature Medicine, 24 (2018), pp. 1716–1720.

[14] C. Lyle, P. S. Castro, and M. G. Bellemare, *A comparative analysis of expected and distributional reinforcement learning*, CoRR, abs/1901.11084 (2019).

[15] J. D. Martin, M. Lyskawinski, X. Li, and B. Englot, *Stochastically dominant distributional reinforcement learning*, arXiv, abs/1905.07318 (2019).

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, *Human-level control through deep reinforcement learning*, Nature, 518 (2015), pp. 529–533.

[17] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka, *Nonparametric return distribution approximation for reinforcement learning*, in Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Madison, WI, USA, 2010, Omnipress, p. 799–806.

[18] ——, *Parametric return density estimation for reinforcement learning*, in Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI'10, Arlington, Virginia, USA, 2010, AUAI Press, p. 368–375.

[19] M. Rowland, M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh, *An analysis of categorical distributional reinforcement learning*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, A. Storkey and F. Perez-Cruz, eds., vol. 84 of Proceedings of Machine Learning Research, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018, PMLR, pp. 29–37.

[20] R. Singh, Q. Zhang, and Y. Chen, *Improving robustness via risk averse distributional reinforcement learning*, arXiv, abs/2005.00585 (2020).

[21] M. Sobel, *Ordinal dynamic programming*, Management Science, 21 (1975), pp. 967–975.

[22] M. J. Sobel, *The variance of discounted markov decision processes*, Journal of Applied Probability, 19 (1982), p. 794–802.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, second ed., 2018.

[24] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Mach. Learn., 16 (1994), p. 185–202.

[25] C. J. C. H. Watkins and P. Dayan, *Q-learning*, in Machine Learning, 1992, pp. 279–292.

Statutory Declaration

I herewith declare that I wrote this thesis and performed the associated research myself, using only literature cited in this volume. If text passages from sources are used literally, they are marked as such.
I confirm that this work is original and has not been submitted elsewhere for any examination, nor is it currently under consideration for a thesis elsewhere.

Vienna, August 29, 2020