SIAM J. MATH. DATA SCI.
Vol. 4, No. 2, pp. 675–693

Speedy Categorical Distributional Reinforcement Learning and Complexity Analysis*

Markus Böck† and Clemens Heitzinger†

Abstract. In distributional reinforcement learning, the entire distribution of the return instead of just the expected return is modeled. The approach with categorical distributions as the approximation method is well-known in Q-learning, and convergence results have been established in the tabular case. In this work, speedy Q-learning is extended to categorical distributions, a finite-time analysis is performed, and probably approximately correct bounds in terms of the Cramér distance are established. It is shown that also in the distributional case the new update rule yields faster policy evaluation in comparison to the standard Q-learning one and that the sample complexity is essentially the same as the one of the value-based algorithmic counterpart. Without the need for more state-action-reward samples, one gains significantly more information about the return with categorical distributions. Even though the results do not easily extend to the case of policy control, a slight modification to the update rule yields promising numerical results.

Key words. reinforcement learning, distributional reinforcement learning, Q-learning, PAC bounds, complexity analysis

AMS subject classifications. 68Q25, 68Q32, 68T05, 68T42

DOI. 10.1137/20M1364436

1. Introduction.

Distributional reinforcement learning (DRL) is a subfield of reinforcement learning where the entire return distribution is modeled directly, rather than just the expected return. Bellemare, Dabney, and Munos [1] introduced a particular distributional framework based on categorical distributions. Rowland et al. [8] established convergence results in the tabular case for this approximation method. In 2011, Ghavamzadeh et al. [5] introduced a new variant of Q-learning [11], called speedy Q-learning (SQL), which was subject to finite-time analysis and achieved impressive experimental results. Finite-time analysis in terms of probably approximately correct (PAC) bounds was also performed for Q-learning [3].

In this work, motivated by the results of Ghavamzadeh et al. [5] and Rowland et al. [8], the SQL update rule is applied in the distributional framework, and PAC bounds for the resulting algorithm are established in terms of the Cramér distance. It is shown that the sample complexity of this algorithm is essentially the same as in the value-based case. Furthermore, the accelerated convergence of SQL in terms of the expected return also translates to the distributional case.

After presenting the theoretical background in section 2, the speedy categorical policy evaluation (SCPE) algorithm is introduced, and the main theoretical results are stated in

*Received by the editors September 4, 2020; accepted for publication (in revised form) February 18, 2022; published electronically June 6, 2022.
https://doi.org/10.1137/20M1364436

†Department of Mathematics and Geoinformation, TU Wien, Vienna, Austria (e1634838@student.tuwien.ac.at, clemens.heitzinger@tuwien.ac.at).

section 3. The corresponding proofs are given in section 4. In section 5, the problems of using the SQL update rule in the control case are discussed. Lastly, in section 6, the theoretical results in the policy evaluation case are confirmed experimentally, and it is shown that a slight modification to the update rule recovers the improved convergence in the control case.

2. Background.

2.1. DRL.

The reinforcement learning objective is formalised by a Markov decision process (MDP), i.e., a tuple $\langle \mathcal{X}, \mathcal{A}, r, p \rangle$, where $\mathcal{X}$ is a set of states and $\mathcal{A}$ is a set of actions. In this work, we only consider finite MDPs, i.e., $|\mathcal{X}| < \infty$ and $|\mathcal{A}| < \infty$. Trajectories (X_t, A_t, R_t) are obtained through the selection of actions at given states, where the probabilities of state transitions are defined by the deterministic function p and R_t are defined by the kernel r, where $r(\cdot|x, a, x')$ represents the immediate reward when transitioning from state x with action a to state x'. The Markov property requires that X_{t+1} and R_t only depend on the previous state and action (x, a), i.e.,

$$\mathbb{P}\left(R_t = s, X_{t+1} = x' | X_t = x, A_t = a, X_{t-1} = x_{t-1}, \dots\right) = r(s|x, a, x')p(x'|x, a).$$

The standard approach to reinforcement learning is to model the expected return, usually by a state-action value function Q. As the name suggests, the core of DRL is to model the entire distribution of the return directly.

For $(x, a) \in \mathcal{X} \times \mathcal{A}$, the *return* $Z^\pi(x, a)$ is the sum of discounted rewards along a trajectory following the policy π starting in state x and taking action a, i.e.,

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

The function Z^π mapping state-action pairs to random variables is called the *return distribution function*.

The usual state-action value function Q^π can be related to the return distribution function by observing that

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

Furthermore, the Bellman equation [2] can be extended to the distributional case as

$$Z^\pi(x, a) \overset{D}{=} R + \gamma Z^\pi(X', A'), \tag{2.1}$$

where $X' \sim p(\cdot|x, a)$, $A' \sim \pi(\cdot|X')$, and $R \sim r(\cdot|x, a, X')$. Here the equal sign indicates that the random variable on the left-hand side and the one on the right-hand side are identically distributed.

Let $\eta_\pi^{(x,a)}$ denote the underlying probability distribution of the random variable $Z^\pi(x, a)$, giving us a second representation

$$Z^\pi(x, a) \sim \eta_\pi^{(x,a)}$$

of return distribution functions in terms of probability measures.

Let $\eta\colon \mathcal{X}\times\mathcal{A}\to\mathscr{P}(\mathbb{R})$ be an arbitrary mapping to probability measures on $\mathbb{R}$. For a given policy π, the distributional Bellman operator $\mathcal{T}^\pi$ can be written in terms of cumulative distribution functions as

$$F_{\mathcal{T}^\pi\eta^{(x,a)}}(z)=\mathbb{E}\left[F_{\eta^{(X',A')}}\left(\frac{z-R}{\gamma}\right)\right] \tag{2.2}$$

due to (2.1). If η corresponds to the return distributions of π, i.e., $\eta=\eta_\pi$, it follows directly from the Bellman equation that $\mathcal{T}^\pi\eta_\pi=\eta_\pi$.

2.2. Categorical DRL. A challenge of DRL is to find appropriate methods to approximate the return distributions. Bellemare, Dabney, and Munos [1] proposed to use N fixed atoms $z_1,\dots,z_N$ or grid points and defined the set of categorical distributions as

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

As this set is not closed under the Bellman update, we have to project distributions back onto this set. Thus the categorical projection operator $\Pi_\mathcal{C}$ was introduced, which is explained in more detail later.

Rowland et al. [8] connected this operator to the Cramér distance, which is defined between two return distribution functions as

$$\begin{aligned}\bar{\ell}_2(\eta,\xi)&:=\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}}\ell_2\left(\eta^{(x,a)},\xi^{(x,a)}\right)\\&=\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left(\int_\mathbb{R}|F_{\eta^{(x,a)}}(z)-F_{\xi^{(x,a)}}(z)|^2\mathrm{d}z\right)^{1/2}.\end{aligned}$$

The key observation was that

$$\Pi_\mathcal{C}\mathcal{T}^\pi\colon\mathcal{P}_z^{\mathcal{X}\times\mathcal{A}}\to\mathcal{P}_z^{\mathcal{X}\times\mathcal{A}}$$

is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$. With this fact, the convergence of categorical policy evaluation to $\eta_\mathcal{C}$, the unique fixed point of $\Pi_\mathcal{C}\mathcal{T}^\pi$, was proven. Note that since we only approximate distributions, we have $\eta_\mathcal{C}\neq\eta_\pi$ in general. However, if the return distributions $\eta_\pi^{(x,a)}$ are supported on $[z_1,z_N]$ for all $(x,a)\in\mathcal{X}\times\mathcal{A}$, then increasing the number of atoms yields a better precision, i.e.,

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

For policy evaluation, Rowland et al. [8] considered the update rule given by the weighted sum

$$\eta_{k+1}^{(x,a)}:=(1-\alpha_k(x,a))\eta_k^{(x,a)}+\alpha_k(x,a)\Pi_\mathcal{C}\mathcal{T}_k^\pi\eta_k^{(x,a)}. \tag{2.3}$$

Here $\mathcal{T}_k^\pi$ is the stochastic Bellman operator at time k, which depends on samples $x_k'\sim p(\cdot|x,a)$ and $a_k'\sim\pi(\cdot|x_k')$ as well as the reward sample $r_k\sim r(\cdot|x,a,x_k')$ for each $(x,a)\in\mathcal{X}\times\mathcal{A}$. In terms of cumulative distribution functions, the operator can be written as

$$F_{\mathcal{T}_k^\pi\eta^{(x,a)}}(z)=F_{\eta^{(x_k',a_k')}}\left(\frac{z-r_k}{\gamma}\right), \tag{2.4}$$

which is a random variable for all $z \in \mathbb{R}$ due to (2.1), and we have

$$F_{\mathcal{T}^\pi \eta^{(x,a)}}(z) = \mathbb{E}\left[F_{\mathcal{T}_k^\pi \eta^{(x,a)}}(z)\right].$$

In the update rule, $\alpha_k(x,a)$ are stepsizes. If $\eta_\pi^{(x,a)}$ is supported on $[z_1, z_N]$ and the Robins–Monro conditions $\sum_{k=0}^\infty \alpha_k(x,a) = \infty$ and $\sum_{k=0}^\infty \alpha_k(x,a)^2 < \infty$ hold for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, then the Cramér distance to the fixed point of $\Pi_{\mathcal{C}} \mathcal{T}^\pi$ converges to zero almost surely, i.e., $\bar{\ell}_2(\eta_k, \eta_{\mathcal{C}}) \to 0$ [8, Theorem 1].

In practice, the state and action samples usually come in episodes and at time k, η_k is only updated at the current state-action pair (x_k, a_k) in the trajectory, and we have $\alpha_k(x,a) = 0$ for all $(x,a) \neq (x_k, a_k)$. This method presents the categorical distributional analogue to the temporal difference (TD) algorithm [9, 10].

For policy control, we change $\mathcal{T}_k^\pi$ in (2.3) to the stochastic optimality operator $\mathcal{T}_k$ given by

$$F_{\mathcal{T}_k \eta^{(x,a)}}(z) = F_{\eta^{(x'_k, a_k^*)}}\left(\frac{z - r_k}{\gamma}\right), \quad a_k^* = \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_{Z \sim \eta_k^{(x'_k, a)}}[Z], \tag{2.5}$$

and obtain the categorical distributional equivalent to Q-learning [11, 10]. In this control case, convergence was only established with the additional assumption that a unique optimal policy exists [8, Theorem 2].

2.3. SQL and sample complexities. For a sample x, a, r, x', where $x' \sim p(\cdot|x,a)$, the Q-learning [11] update rule reads

$$Q_{k+1}(x,a) = (1 - \alpha_k(x,a)) Q_k(x,a) + \alpha_k(x,a)\left(r + \gamma \max_{a \in \mathcal{A}} Q_k(x', a)\right). \tag{2.6}$$

Let Q^* denote the unique optimal value function. Further, assume that the rewards are bounded by $R_{\max}$. For $\gamma < 1$ and $\beta := \frac{1}{1-\gamma}$, let $V_{\max} := \beta R_{\max}$ be the maximal attainable return.

If the updates with (2.6) are performed *synchronously*, that is, at each time step k all state-action pairs $(x,a) \in \mathcal{X} \times \mathcal{A}$ are updated, and we have polynomial learning rates

$$\alpha_k = \frac{1}{(k+1)^\omega}, \quad \frac{1}{2} < \omega < 1,$$

then for a finite state-action space $n = |\mathcal{X} \times \mathcal{A}|$ and for $\gamma < 1$, the following finite-time behavior is known [3]: with probability at least $1 - \delta$, the inequality

$$\|Q^* - Q_T\|_\infty \leq \epsilon \tag{2.7}$$

holds for

$$T \geq C\left(\left(\frac{\beta^4 R_{\max}^2 \log \frac{n\beta^2 R_{\max}}{\delta\epsilon}}{\epsilon^2}\right)^{\frac{1}{\omega}} + \left(\beta \log \frac{\beta R_{\max}}{\epsilon}\right)^{\frac{1}{1-\omega}}\right)$$

and for some constant $C > 0$.

Following the reasoning of Even-Dar and Mansour [3] and Ghavamzadeh et al. [5], if γ is close to 1, β becomes the dominant term and the bound is optimized for $\omega = 4/5$, yielding a complexity of

$$\mathcal{O}\left(\left(\frac{\beta^4 R_{\max}^2 \log \frac{n\beta^2 R_{\max}}{\delta\epsilon}}{\epsilon^2}\right)^{5/4}\right) = \tilde{\mathcal{O}}(\beta^5/\epsilon^{2.5}),$$

since $g = \tilde{\mathcal{O}}(f) \iff g \leq C_1 f \log^{C_2}(f)$ for some constants $C_1, C_2 > 0$.

The bound in probability and the derived sample complexity also hold for the evaluation of the value function Q^π for an arbitrary policy π.

Ghavamzadeh et al. [5] introduced a faster variant of Q-learning and named it SQL. They defined the update rule

(2.8)
$$Q_{k+1}(x,a) := (1-\alpha_k)Q_k(x,a) + \alpha_k \mathcal{T}_k Q_{k-1}(x,a) + (1-\alpha_k)(\mathcal{T}_k Q_k(x,a) - \mathcal{T}_k Q_{k-1}(x,a))$$

based on two previous time steps instead of just one, where

$$\mathcal{T}_k Q(x,a) = r + \gamma \max_{a \in \mathcal{A}} Q(x',a)$$

and the learning rate is $\alpha_k = \frac{1}{k+1}$. The key difference to Q-learning is that SQL uses a more aggressive learning rate for the third term. Changing it to $\alpha_k(\mathcal{T}_k Q_k(x,a) - \mathcal{T}_k Q_{k-1}(x,a))$ would be equivalent to Q-learning. The difference seems small; however, it yields faster convergence, i.e., it can be shown that the inequality

$$\|Q^* - Q_T\|_\infty \leq 2V_{\max}\beta\left(\frac{\gamma}{T} + \sqrt{\frac{2\log\frac{2n}{\delta}}{T}}\right) \tag{2.9}$$

holds with probability $1-\delta$, and thus for

$$T := \frac{11.66\beta^2 V_{\max}^2 \log\frac{2n}{\delta}}{\epsilon^2}.$$

we have

$$\|Q^* - Q_T\|_\infty \leq \epsilon.$$

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

3. SCPE.

In the following, the update rule of SQL is extended to categorical distributions in the policy evaluation case. We chose to extend SQL to distributions, rather than standard Q-learning, because it yields faster convergence. However, it is worth mentioning that the main idea of the proof is also applicable if one uses (2.3) and (2.6).

In order to translate SQL to categorical distributions, we combine (2.8) and (2.3) for the evaluation of a fixed policy π into the update formula

$$\eta_{k+1}^{(x,a)} = \eta_k^{(x,a)} + \alpha_k\left(\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_{k-1}^{(x,a)} - \eta_k^{(x,a)}\right) + (1-\alpha_k)\left(\Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_k^{(x,a)} - \Pi_{\mathcal{C}}\mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}\right), \tag{3.1}$$

where we start with two initial return distribution functions $\eta_0 = \eta_{-1} \in \mathcal{P}_z$. We again use the learning rate $\alpha_k := \frac{1}{k+1}$.

It is straightforward to see that (3.1) can be rewritten as the 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 $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ is a probability measure. In general, it is a finite signed measure, and thus we also do not know if the recursively defined $\eta_k^{(x,a)}$ are indeed probability measures. The consideration of this problem makes up a substantial part of the analysis in section 4.

In the following analysis, we only consider the synchronous version of policy evaluation, which is shown as pseudocode in Algorithm 3.1. Like the finite-time analysis of SQL and Q-learning, it can also be extended to the asynchronous case, where we consider a policy with finite covering time.

In order to formulate the main result below, we collect the following assumptions.

Assumption 1. The state-action space is finite with $n := |\mathcal{X} \times \mathcal{A}|$ elements. The categorical distribution $\eta_{\mathcal{C}}$ is the unique fixed point of $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$. The rewards are bounded by $R_{\max} > 0$. The discount factor γ is smaller than 1, and we let $\bar{\beta} := \frac{1}{1-\sqrt{\gamma}}$. Let $V_{\max} := \frac{1}{1-\gamma}R_{\max}$ be the maximal attainable return. For the N fixed atoms we assume $z_1 = -V_{\max}$ and $z_N = V_{\max}$. Lastly, the two initial return distribution functions are equal, i.e., $\eta_{-1} = \eta_0$, and the η_k are obtained by update rule (3.1).

The main result is the following.

Algorithm 3.1. Synchronous SCPE

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 $x_k' \sim p(\cdot|x,a)$, $a_k' \sim \pi(\cdot|x_k')$, $r_k \sim r(\cdot|x,a,x_k')$
    $\mathcal{T}_k^{\pi}\eta_k^{(x,a)} \leftarrow \sum_{i=1}^N p_{k,i}^{(x_k',a_k')}\delta_{r_k+\gamma z_i}$ # 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

Theorem 3.1. *Under Assumption* 1, *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 with probability at least $1-\delta$.

We give two corollaries.

Corollary 3.2. *Under Assumption* 1, *for any* $0 < \epsilon \leq \sqrt{V_{\max}}$, *the inequality* $\|\eta_{\mathcal{C}} - \eta_T\|_{\bar{\ell}_2} \leq \epsilon$ *holds with probability at least* $1-\delta$ *after*

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

steps of SCPE.

Corollary 3.3. *Under Assumption* 1, η_T *converges to* $\eta_{\mathcal{C}}$ *almost surely in* $\bar{\ell}_2$.

The proofs are deferred to section 4.

Corollary 3.2 leads to following complexity analysis. For each time step k, we sweep over the entire state-action space. Therefore, after T iterations, $3nT$ samples are available in total (reward, next state, and next action in each time step). 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 N of atoms only contributes to the logarithmic factor. Thus, increasing the accuracy of the distribution approximation causes only a small 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 power of β stems from the fact that a different metric was used. To see how the difference in expected values and the Cramér distance relate, consider two measures μ, ν supported on $[z_1, z_N]$. Then,

$$\begin{aligned}
&|\mathbb{E}_{Z_\mu\sim\mu}[Z_\mu] - \mathbb{E}_{Z_\nu\sim\nu}[Z_\nu]| \\
&\quad= \left|\int_0^\infty (1-F_\mu(z)) - (1-F_\nu(z))\mathrm{d}z - \int_{-\infty}^0 F_\mu(z) - F_\nu(z)\mathrm{d}z\right| \\
&\quad\leq \int_{\mathbb{R}} |F_\mu(z) - F_\nu(z)|\mathrm{d}z \\
&\quad\leq \|F_\mu - F_\nu\|_2 \left\|\mathbb{1}_{[z_1,z_N]}\right\|_2 \\
&\quad= (z_N - z_1)^{1/2}\|F_\mu - F_\nu\|_2 = \sqrt{2V_{\max}}\,\ell_2(\mu,\nu).
\end{aligned}$$

This inequality precisely captures the relationship of inequality (2.9) and Theorem 3.1. As $V_{\max} = \beta R_{\max}$ this also explains the difference in the power of β in the sample complexities.

It is quite an interesting result that the sample complexity remains the same when switching to distributions. One *does not* need more samples when modeling the entire distribution. Further, the sample complexity is independent of the number of atoms—the precision with which the return distributions are modeled. However, the computational complexity $\tilde{\mathcal{O}}(nN\beta^3/\epsilon^2)$ is higher, of course, and a table with nN elements is needed to store the return distributions.

4. Analysis. The analysis follows the outline of Ghavamzadeh et al. [5]. 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 Kallenberg [6].

Definition 4.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and define*

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

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 $X\colon (\Omega, \mathcal{A}, \mathbb{P}) \to (M, \mathcal{M})$, $\omega \mapsto X_\omega$, 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

$$F_{\mathbb{E}[X]}(z) = \mathbb{E}[X]((-\infty, z]) = \mathbb{E}[X(-\infty, z]] = \mathbb{E}[F_X(z)]. \tag{4.1}$$

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

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

4.1. 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 4.2. *Let $\mathcal{L}$ be the set of finite signed Borel measures*

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

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

$$(a\mu + b\nu)(A) := a\mu(A) + b\nu(A), \quad \mu, \nu \in \mathcal{L},\ a, b \in \mathbb{R},\ A \text{ a measurable set}. \tag{4.2}$$

Equation (4.2) immediately implies that

$$F_{a\mu+b\nu} = aF_\mu + bF_\nu. \tag{4.3}$$

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

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

The categorical projection operator $\Pi_{\mathcal{C}}$ *can be easily applied to elements of* $\mathcal{L}$ *by defining*

$$\Pi_{\mathcal{C}} \colon \mathcal{L} \to \mathcal{L}_z, \quad F_{\Pi_{\mathcal{C}}\nu}(z_i) = \frac{1}{z_{i+1} - z_i} \int_{z_i}^{z_{i+1}} F_\nu(z)\mathrm{d}z, \quad F_{\Pi_{\mathcal{C}}\nu}(z_N) = \lim_{z\to\infty} F_\nu(z). \tag{4.4}$$

From (4.4) and (4.3), it is not difficult to see that $\Pi_{\mathcal{C}} \colon \mathcal{L} \to \mathcal{L}_z$ is a linear projection. Furthermore, from characterisation (2.4) and (4.3) it follows that also $\mathcal{T}_k^\pi \colon \mathcal{L}^{\mathcal{X}\times\mathcal{A}} \to \mathcal{L}^{\mathcal{X}\times\mathcal{A}}$ is a linear mapping.

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

Lemma 4.3. *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 an element of $\mathscr{P}(\mathcal{P}_z)$, since $\eta_0^{(x,a)} \in \mathcal{P}_z$. Of course, $\eta_0^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ also (interpreted as a random measure which takes $\eta_0^{(x,a)}$ with probability 1).

Assume that $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}$ and $\eta_k^{(x,a)}$ are random probability measures. To show 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

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

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 $\eta_{k+1}^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$. ∎

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

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

The expected update is given by

$$\mathcal{D}[\eta_k, \eta_{k-1}]^{(x,a)} := \mathbb{E}\left[\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \,\middle|\, \mathcal{F}_{k-1}\right] \overset{(2.4)}{=} k\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_k^{(x,a)} - (k-1)\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_{k-1}^{(x,a)}.$$

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

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

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

$$\eta_{k+1}^{(x,a)} = \frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}\left(\mathcal{D}[\eta_k, \eta_{k-1}]^{(x,a)} - \epsilon_k^{(x,a)}\right). \tag{4.5}$$

It is not immediately clear how one can turn the errors into a martingale. The following lemma shows that we have to look at the cumulative distribution function at each atom. Lemma 4.3 and Lemma 4.4 are the core results that allow us to extend the analysis of SQL [5] to categorical distributions. One can extend the result (2.7) from Even-Dar and Mansour [3] in a similar fashion.

Lemma 4.4. *The inclusions $\epsilon_k^{(x,a)} \in \mathscr{P}(\mathcal{L}_z)$ and $E_k^{(x,a)} \in \mathscr{P}(\mathcal{L}_z)$ hold 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 martingale difference sequence, i.e.,*

$$\text{for all } k \geq 0, \quad \mathbb{E}\left[F_{\epsilon_k^{(x,a)}}(z_i) \,\middle|\, \mathcal{F}_{k-1}\right] = 0 \;\wedge\; \left|F_{\epsilon_k^{(x,a)}}(z_i)\right| \leq 1. \tag{4.6}$$

Proof. By Lemma 4.3, $\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \in \mathscr{P}(\mathcal{P}_z)$ holds. It follows from (4.1) 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 $\mathcal{P}_z$. Therefore 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. ∎

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

Lemma 4.5. *For all $k \geq 1$, the equality*

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

holds.

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 $k \geq 1$. The definitions of $\mathcal{D}[\eta_k, \eta_{k-1}]$ and E_k imply

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

which concludes the proof. ∎

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

$$\|\nu\|_{\ell_2} := \left(\sum_{i=1}^{N-1}(z_{i+1} - z_i)F_\nu(z_i)^2 + F_\nu(z_N)^2\right)^{1/2} \tag{4.7}$$

for all $\nu \in \mathcal{L}_z$. 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 $\bar{\ell}_2$.

Further we define the norm

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

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

$$\ell_2(\mu,\nu) = \|\mu - \nu\|_{\ell_2} \leq \sqrt{2V_{\max}}\,\|\mu - \nu\|_{\ell_\infty} \leq \sqrt{2V_{\max}} \tag{4.8}$$

hold for all μ and $\nu \in \mathcal{P}_z$. 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 4.4), $F_{\epsilon_k^{(x,a)}}(z_N) = 0$ holds, and thus $F_{E_k^{(x,a)}}(z_N) = 0$ also. The inequality

$$\|E_k\|_{\bar{\ell}_2} \leq \sqrt{2V_{\max}}\,\|E_k\|_{\bar{\ell}_\infty} \tag{4.9}$$

follows from (4.7).

Lemma 4.6. *For all $k \geq 1$, the inequalities*

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

hold.

Proof. Again, this is proved by induction. We use the fact that $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$, substitute the equality from Lemma 4.5, and apply the norm inequality (4.8).

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 $k+1$, since

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

which concludes the proof. ∎

4.4. Step 4: Bounding the error in probability. Applying the Hoeffding–Azuma inequality is the crucial step in proving Theorem 3.1.

Lemma 4.7 (maximal Hoeffding–Azuma inequality [5]). *Let $\mathcal{V} := \{V_1, \dots, V_T\}$ be a martingale 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*

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

holds.

Lemma 4.8. *For all $\epsilon > 0$ and all time steps T, the inequality*

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

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 4.4, $V_j = F_{\epsilon_j^{(x,a)}}(z_i)$, $j = 0,\dots,T$, is a martingale 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

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

By taking the union over all atoms, we have

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

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

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

which concludes the proof. ∎

4.5. Step 5: Concluding the Proof of Theorem 3.1.

Proof of Theorem 3.1. By Lemma 4.6 and inequality (4.9), 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 4.8 the inequality

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

holds. Setting δ as above and solving for ϵ yields

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

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

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

which concludes the proof of the theorem. ∎

Proof of Corollary 3.2. Define

$$T := \frac{C\bar{\beta}^2 V_{\max}\log\frac{2nN}{\delta}}{\epsilon^2},$$
$$t := \frac{\bar{\beta}^2 V_{\max}\log\frac{2nN}{\delta}}{\epsilon^2} \ge 1,$$

implying

$$\frac{1}{t} \le \frac{1}{\sqrt{t}}.$$

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

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

Proof of Corollary 3.3. After rearranging, we have

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

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

$$\mathbb{P}\left(\bar{\ell}_2(\eta_{\mathcal{C}},\eta_T)>\epsilon\right) \le Cq^T, \quad C>0, \quad 0<q<1.$$

Therefore $\sum_{T=0}^{\infty}\mathbb{P}\left(\bar{\ell}_2(\eta_{\mathcal{C}},\eta_T)>\epsilon\right) < \infty$, and by the Borel–Cantelli lemma we have almost sure convergence. ∎

5. Policy control. Unfortunately, the analysis cannot be easily extended to categorical distributions in the control case. There are several reasons.

First, the Bellman optimality operator $\mathcal{T}$ is not a contraction in $\bar{\ell}_2$. Bellemare et al. [1] provided a counterexample for the Wasserstein distance that also works for the Cramér distance. Therefore Lemma 4.6 does not hold if $\mathcal{T}^\pi$ is replaced by $\mathcal{T}$.

Nevertheless, we consider the update rule for the control case,

$$\eta_{k+1} := \eta_k + \alpha_k(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_{k-1}}\eta_{k-1} - \eta_k) + (1-\alpha_k)\left(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_k - \Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_{k-1}}\eta_{k-1}\right). \tag{5.1}$$

Here π_k denotes the greedy policy w.r.t. the expected values of η_k, and it holds that $\mathcal{T}_k^{\pi_k}\eta_k = \mathcal{T}_k\eta_k$.

It can be shown that such update rules produce the same expected values as their value-based algorithmic counterpart [7]. Therefore, we can be sure that $Q_k(x,a) := \mathbb{E}_{Z\sim\eta_k^{(x,a)}}[Z]$ converges to the unique optimal value function Q^*, because Q_k satisfies (2.8).

If we assume a unique optimal policy π^*, then Q_k comes close enough to Q^* such that $\pi_k = \pi^*$ after some time, and it remains to evaluate the return distributions of π^* for which convergence holds. This is the reasoning Rowland et al. [8] used to prove their control theorem.

This approach does not work in the present control case, as the update rule (5.1) does not necessarily yield probability measures anymore, which can be seen as follows by revisiting the proof of Lemma 4.3 and calculating

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

for the control case. But if $\pi_{k+1} \neq \pi_k$, this is not equal to $\Pi_{\mathcal{C}}\mathcal{T}_{k+1}^{\pi_{k+1}}\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)}$ in general, and hence $\frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}\mathcal{D}_k[\eta_k,\eta_{k-1}]^{(x,a)}$ is not necessarily an element of $\mathscr{P}(\mathcal{P}_z)$, meaning that we now obtain signed measures in the general case.

In order to fix this problem, one could alter the update rule to become

$$\eta_{k+1} := \eta_k + \alpha_k(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_{k-1} - \eta_k) + (1-\alpha_k)(\Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_k - \Pi_{\mathcal{C}}\mathcal{T}_k^{\pi_k}\eta_{k-1}). \tag{5.2}$$

With this changed definition, Lemma 4.3 holds again, but we run into different problems. The first problem is that Lemma 4.5 does not hold any more, as we now have

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

But if the policies do not change anymore after time step T, the summands are zero for $k > T$, and the second term becomes small as k tends to infinity. Thus, $\eta_k \approx \Pi_{\mathcal{C}}\mathcal{T}\eta_{k-1}$ holds again, which indicates this problem to be minor. However, it leads to the second, more serious, problem, namely, showing that the expected values of $\eta_k^{(x,a)}$ obtained by (5.2) still converge to the optimal value function Q^*.

Nevertheless, this adjusted update rule shows good experimental results, as discussed in the next section.

6. Experimental results.

6.1. Combination-lock. Consider the combination-lock environment [4]. 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` or `right`; see Figure 1. Choosing `right` takes the agent to state x_{i+1} but yields a reward of -0.01. Taking `left` takes the agent 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 the reward $+15$.

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 set up such that choosing `right` in all states is the unique 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.

In the experiment, $\gamma = 0.999$ and 51 equally spaced atoms or grid points between -10 and 15 were chosen. The SCPE algorithm was run 10 times for 5000 iterations with random initial distributions (51 random numbers were drawn independently from the uniform distribution $[0, 1]$ for all (x, a) and then divided by their sum to form probabilities). For comparison, the TD update rule (2.3) with polynomial learning rates $\omega \in \{0.55, 0.8, 1\}$ was tested. The limiting return distribution $\eta_{\mathcal{C}}$ was estimated by performing SCPE for 50 000 iterations, denoted by $\hat{\eta}_{\mathcal{C}}$.

In Figure 2(a), the maximum Cramér distance $\overline{\ell}_2(\eta_k, \hat{\eta}_{\mathcal{C}})$ to the estimated limiting return distribution function and the maximum absolute distance of the corresponding expected returns, averaged over the 10 runs, are shown. This confirms that indeed about the same sample complexity holds in both cases.

In Figure 2(b) the clear performance benefit of the speedy update rule (3.1) over the TD one (2.3) is visible. This plot resembles the results of Ghavamzadeh et al. [4].

The same experiment was performed in the control case. The instability problem of using the unadjusted update rule (5.1) is illustrated in Figure 2(a). Here, measures with negative probabilities were indeed produced. Using the adjusted update rule (5.2) yields almost exactly the same performance improvements as in the policy evaluation case; see Figures 2(b) and 3(b).

6.2. Gridworld. In order to put the adjusted update rule (5.2) to the test, we investigated the convergence of the expected values to the optimal value function Q^* in an environment with multiple optimal policies. We consider an $n \times n$ gridworld, where the agent can move up, down, left, and right. If the agent moves to the cell with coordinates $(x, y) \in \{1, \dots, n\}^2$, it receives reward $\pm(|(n - x + 1) - y| + 1)$ with equal probability. Only at the goal cell (n, n) is the positive reward n always obtained. Figure 4 shows an overview of this environment.

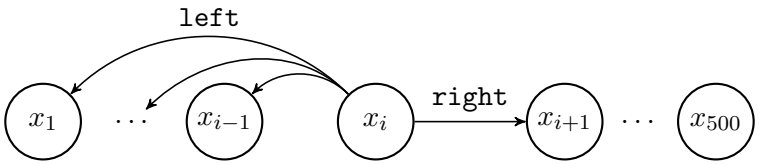
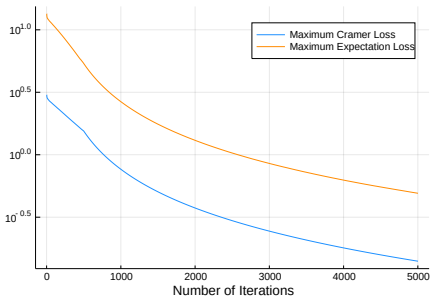
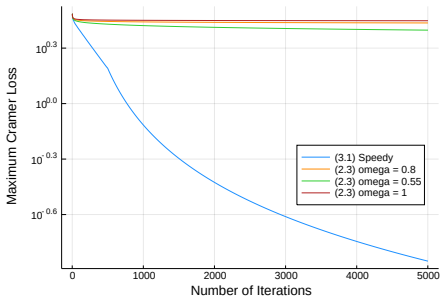


Figure 1. *Combination-lock environment.*

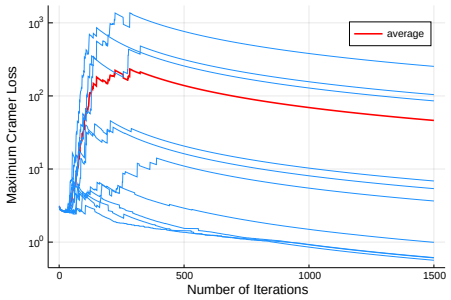


(a) Convergence in Cramér distance vs. convergence in expectation

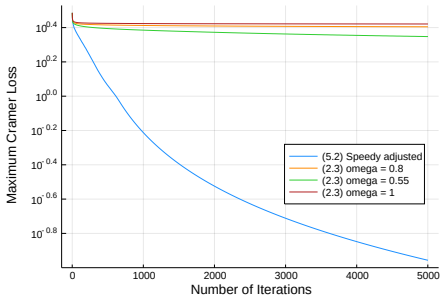


(b) SCPE and polynomial learning rates

Figure 2. *Policy evaluation in the combination-lock environment.*



(a) Instability of unchanged algorithm



(b) Improved convergence for adjusted algorithm

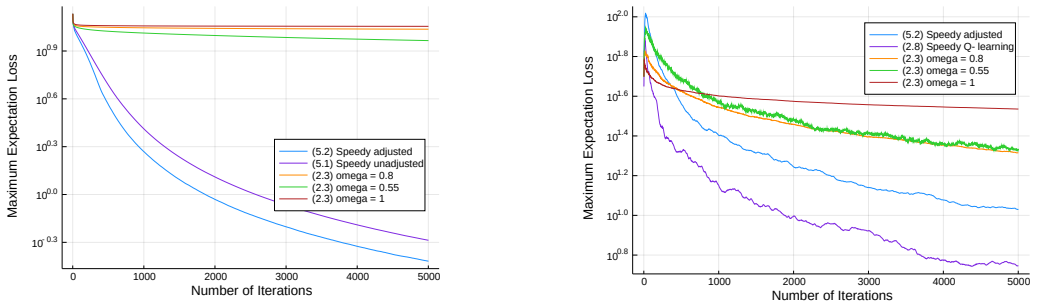
Figure 3. *Q-learning in the combination-lock environment.*

$\pm n$	$\pm(n-1)$	$\dots$	± 3	± 2	± 1
$\pm(n-1)$	$\pm(n-2)$	$\dots$	± 2	± 1	± 2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
± 2	± 1	$\dots$	$\pm(n-3)$	$\pm(n-2)$	$\pm(n-1)$
± 1	± 2	$\dots$	$\pm(n-2)$	$\pm(n-1)$	n

Figure 4. *Gridworld with rewards given at each cell.*

The difficulty for the agent is to recognize that wandering around in the environment gives an expected return of zero and that the optimal strategy is to reach the goal cell as quickly as possible. This can be done along multiple paths in the grid, and lucky immediate rewards causes the agent to often change direction.

For this environment we used the same experiment setup as in section 6.1 with $n = 25$ and $\gamma = 0.9$. While it was possible to compute the expected values simply from the categorical distributions for update rules (2.3) and (5.2), this was not the case for rule (5.1) in the gridworld environment. The unadjusted update rule (5.1) led to such instabilities that the signed probabilities under- and overflowed the 64 bit double value range. For this reason, we



(a) Combination-Lock (b) Gridworld

Figure 5. *Comparison of the adjusted* (5.2) *versus the unadjusted* (5.1) *update rule by the maximum absolute difference to the theoretical optimal value function.*

directly used the update rule for Q values (2.8) in this case, where the initial Q-tables were uniformly sampled from $[-n, n]$.

In Figure 5, the maximum absolute difference to the theoretical optimal value function Q^* is shown for both environments. For the combination-lock environment, the expected values were obtained from the distributional updates of the control experiment of section 2; for the gridworld, the expected values were obtained as described in the last paragraph. While the adjusted update rule is slightly better than the unadjusted one in the combination-lock environment, it lags behind in the gridworld example.

Both the adjusted (5.2) and the unadjusted (5.1) SQL update rules are convincingly faster than the standard Q-learning update rule in both environments. Further, the greedy policies changed up to around time step 600 in the combination-lock environment, whereas in the gridworld they change in over 85% of the steps up to the last iteration. This suggests that the adjusted update rule is still robust under frequent policy changes, but they may slow down convergence.

7. Conclusions.

In this paper, SQL was extended from the value-based case to categorical distributions. For evaluating a fixed policy, PAC bounds in terms of the Cramér distance were established. This led to the observation that even though the computational and space complexity scale linearly in the number of atoms N, the sample complexity scales only logarithmically in N. Thus, switching from standard reinforcement learning to DRL or increasing the accuracy of the distribution approximation yields only a small penalty in terms of transition samples needed. An application in two simple environments confirmed the theoretical results empirically.

The reasons as to why the finite-time analysis could not be easily extended to the case of policy control were stated. Experiments showed that a slight modification to the update rule results in the same performance improvements as in the policy evaluation case. An in-depth analysis of this adjusted updated rule remains for future work.

REFERENCES

[1] M. G. Bellemare, W. Dabney, and R. Munos, *A distributional perspective on reinforcement learning*, in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 449–458.

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
[3] E. Even-Dar and Y. Mansour, *Learning rates for Q-learning*, J. Mach. Learn. Res., 5 (2004), p. 1–25.
[4] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos, *Reinforcement Learning with a Near Optimal Rate of Convergence*, Technical report, INRIA, Oct. 2011, ID 00636615v2.
[5] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos, *Speedy Q-learning*, in Proceedings of the Conference on Neural Information Processing Systems, 2011, pp. 2411–2419.
[6] O. Kallenberg, *Random Measures, Theory and Applications*, Springer, Cham, 2017.
[7] C. Lyle, P. S. Castro, and M. G. Bellemare, *A Comparative Analysis of Expected and Distributional Reinforcement Learning*, preprint, arXiv:1901.11084 [cs.LG], 2019.
[8] 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, 2018, pp. 29–37.
[9] R. Sutton, *Learning to predict by the method of temporal differences*, Mach. Learn., 3 (1988), pp. 9–44.
[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, second ed., The MIT Press, Cambridge, MA, 2018.
[11] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Mach. Learn., 8 (1992), pp. 279–292.