Distributional Method for Risk Averse Reinforcement Learning

Ziteng Cheng[*§], Sebastian Jaimungal[*§] and Nick Martin[‡§]

Abstract—We introduce a distributional method for learning the optimal policy in risk averse Markov decision process with finite state action spaces, latent costs, and stationary dynamics. We assume sequential observations of states, actions, and costs and assess the performance of a policy using dynamic risk measures constructed from nested Kusuoka-type conditional risk mappings. For such performance criteria, randomized policies may outperform deterministic policies, therefore, the candidate policies lie in the d-dimensional simplex where d is the cardinality of the action space. Existing risk averse reinforcement learning methods seldom concern randomized policies, naive extensions to current setting suffer from the curse of dimensionality. By exploiting certain structures embedded in the corresponding dynamic programming principle, we propose a distributional learning method for seeking the optimal policy. The conditional distribution of the value function is casted into a specific type of function, which is chosen with in mind the ease of risk averse optimization. We use a deep neural network to approximate said function, illustrate that the proposed method avoids the curse of dimensionality in the exploration phase, and explore the method's performance with a wide range of model parameters that are picked randomly.

Index Terms—risk averse, Markov decision process, reinforcement learning, deep learning

I. INTRODUCTION

Markov Decision Processes (MDPs) are a type of discrete-time stochastic control problem used for sequential decision-making in situations where costs are partially random and partially under the control of a decision maker. In risk-averse MDPs, the decision maker is concerned with the risk or variability of the outcomes beyond the expected costs. One way to incorporate risk aversion into MDPs is to use nested compositions of risk transition mappings. This approach ensures the time-consistency property and ultimately enables the use of a dynamic programming principle (DPP) to solve the corresponding sequential optimization problem. The approach is proposed in [21], where deterministic costs are considered. Both finite and infinite (required bounded costs) time horizon DPP are derived. Subsequent studies such as [23] and [8] explore infinite time horizon risk-averse DPPs with unbounded costs in different settings. [2] considers unbounded latent costs and established the corresponding finite and infinite horizon

* Deptartment of Statistical Sciences, University of Toronto, Canada {sebastian.jaimungal, ziteng.cheng}@utoronto.ca
‡ nick.martin@mail.utoronto.ca
§ Equal contribution

DPPs. More recently, [7] has developed a framework based on Kusuoka-type conditional risk mappings that also takes into account randomized actions in a risk averse manner. Depending on the type of conditional risk mappings used, randomized actions may be more preferable than deterministic actions, as illustrated in a motivating example in [7]. Other methods of incorporating risk aversion into MDPs are also available, including those discussed in [3], [8], [5], and the references therein.

The focus of this paper is on the approach of nested compositions of risk transition mappings. The main objective is to develop a reinforcement learning method that solves the infinite horizon risk-averse MDP problem presented in [7]. Specifically, the aim is to solve this problem with finite state and action spaces, deterministic latent costs, and stationary dynamics, without assuming knowledge of the controlled transition matrix or cost function. We begin by briefly reviewing some sample-based algorithms that solve risk-averse MDP problems.

For instance, [26] derives a policy gradient formula by combining the static gradient formula for coherent risk measure with the corresponding DPP. This approach is further developed in a sample-based method in [25], with its convergence analyzed in [12]. [29] proposes a family of sample-based algorithms to approximately solve problems with continuous state and action spaces. [24] presents and analyzes a risk-averse Q-learning algorithm, while [13] extends the previous Q-learning algorithm based on estimating a general minimax function with stochastic approximation, with detailed error analysis conducted in [14]. [17] studies a risk-averse temporal difference method that evaluates the value function using linear function approximations. Finally, the recent work in [9] develops an approach to address risk transition mappings induced by convex risk measures.

However, the methods mentioned above do not directly apply to the problem presented in [7], where the risk aversion also involves the randomness in the randomized actions. This is mainly because of the lack of linearity: the value function of a randomized action may not be a linear combination of the value functions of individual actions with respect to the randomizing action kernel. Naively extending the existing methods may result in a situation where we need to learn the value functions for numerous pairs of states and action kernels. Since the admissible action kernels form a d-dimensional simplex, where d is the size of the action space, the exploration task that follows may suffer from the curse of dimensionality

and demand a significant amount of data. On the other hand, the finite nature of the underlying state and action spaces suggests that we can avoid such an excessively expensive exploration task.

We propose a distributional method to address the challenges posed by the risk aversion towards randomized actions in the problem presented in [7]. The proposed method learns an auxiliary function that contains sufficient information about the value function's distribution, avoiding the curse of dimensionality and facilitating the computation of the value function defined via a risk transition mapping. We show in Theorem III.2 that the proposed method's exploration effort grows polynomially with the state and action space cardinalities. Although we initially considered deterministic latent costs, our method naturally handles random costs whose distribution depends on the current state, the realized action, and the next state. This type of random cost is seldom considered in existing literature on risk-averse reinforcement learning. We provide numerical examples that demonstrate the efficacy of the proposed method at the end of this report.

Using distributional methods to solve MDP problems that are not risk neutral has a long history (cf. [15], [27], [19], and the reference therein). More recently, a series of works including [4], [10], [28], [20], and [30] have demonstrated that distributional methods can also achieve better results in the risk neutral setting. In this broader context, our method also contributes to the understanding of the capabilities of distributional methods in solving MDP problems.

II. PRELIMINARIES

In this section, we present the set up the this paper.

A. Markov decision process

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We consider a time-homogeneous Markov decision process (MDP) with a finite state space $\mathbb{X}$ and finite action space $\mathbb{A}$. For each $k \in \mathbb{A}$, let $T^k \in \mathbb{R}^{|\mathbb{X}| \times |\mathbb{X}|}$ be a controlled transition matrix, where T_{ij}^k is the probability of transitioning to state $j \in \mathbb{X}$ at the next epoch, given the current state $i \in \mathbb{X}$ and action $k \in \mathbb{A}$. Let $\pi : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ be a stationary Markovian policy, where $\mathcal{P}(\mathbb{A})$ is the set of probability measures on $\mathbb{A}$. Since $\mathbb{A}$ is finite, $\mathcal{P}(\mathbb{A})$ is a $|\mathbb{A}|$-dimensional simplex, and for $\lambda \in \mathcal{P}(\mathbb{A})$, λ_k is the probability of action k occurring. The state-action process subject to policy π is denoted by $\{(X_t^\pi, A_t^\pi)\}_{t \geq 0}$. The MDP is associated with a bounded latent cost function $C : \mathbb{X} \times \mathbb{A} \times \mathbb{X} \to [0, c_{\max}]$, where $c_{\max} > 0$ is the upper bound of the cost. Finally, we let $\gamma \in (0, 1)$ be the discount factor.

B. Risk averse dynamic programming

In this paper, we use the notation $L^\infty(\Omega, \mathscr{F}, \mathbb{P})$ to denote the space of bounded real-valued Borel-measurable random variables. Equality and inequality between random variables are understood in a $\mathbb{P}$-almost sure sense. Let $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra. We say $\zeta : L^\infty(\Omega, \mathscr{F}, \mathbb{P}) \to L^\infty(\Omega, \mathscr{G}, \mathbb{P})$ is a conditional risk mapping if ζ satisfies the following conditions

for any $Z, Z^1, Z^2 \in L^\infty(\Omega, \mathscr{F}, \mathbb{P})$, $Y \in L^\infty(\Omega, \mathscr{G}, \mathbb{P})$ and $\beta \geq 0$,

 (i) [Monotonicity] if $Z^1 \leq Z^2$, then $\zeta(Z^1) \leq \zeta(Z^2)$;
 (ii) [Translation equivariance] $\zeta(Y + Z) = Y + \zeta(Z)$;
 (iii) [Convexity] if $\beta \in [0, 1]$, then $\zeta(\beta Z^1 + (1 - \beta)Z^2) \leq \beta \zeta(Z^1) + (1 - \beta)\zeta(Z^2)$;
 (iv) [Positive homogeneity] $\zeta(\beta Z) = \beta \zeta(Z)$.

Some may replace β in condition (iii) (resp. (iv)) with $Y \in [0, 1]$ (resp. $Y \geq 0$), which, for the most part, does not affects the developing of the theory.

Consider $\mathscr{U}_0 = \{\emptyset, \Omega\} \subseteq \mathscr{U}_1 \subseteq \cdots \subseteq \mathscr{F}$, $\mathfrak{Z} = (Z_t)_{t \in \mathbb{N}} \subset L^\infty(\Omega, \mathscr{F}, \mathbb{P})$ and $\rho_t : L^\infty(\Omega, \mathscr{F}, \mathbb{P}) \to L^\infty(\Omega, \mathscr{U}_t, \mathbb{P})$. Suppose $|Z_t| \leq c_{\max}$ for all $t \in \mathbb{N}$. [21] proposes to use a dynamic risk measure of the form

$$\rho_{t,T}(\mathfrak{Z}) := \begin{cases} \rho_t\left(Z_t + \gamma \rho_{t+1,T}(\mathfrak{Z})\right), & t < T, \\ \rho_T(Z_T), & t = T, \end{cases}$$

$$\rho_{0,\infty}(\mathfrak{Z}) := \lim_{T \to \infty} \rho_{0,T}(\mathfrak{Z}), \quad \text{(II.1)}$$

for MDP optimization problem. It can be shown that the construction above guarantees time consistency of $(\rho_{t,\infty})_{t \in \mathbb{N}}$.[1]

In what follows, we let $\mathscr{U}_0^\pi$ be the trivial σ-algebra, $\mathscr{U}_t^\pi$ be the σ-algebra generated by $(X_1^\pi, A_1^\pi, \ldots, X_{t-1}^\pi, A_{t-1}^\pi, X_t^\pi)$, and $\mathcal{M}$ be a set of discrete probability measures with support contained by $(0, 1]$. We consider a specific type of conditional risk mapping

$$\rho_t^\pi(Z) := \sup_{\mu \in \mathcal{M}} \left\{ \int_0^1 \inf_{q \in \mathbb{R}} \left\{ q + \right.\right.$$
$$\left.\left. \xi^{-1} \int_{\mathbb{R}} (z - q)_+ P^{Z|\mathscr{U}_t^\pi}(\mathrm{d}z) \right\} \mu(\mathrm{d}\xi) \right\}, \quad \text{(II.2)}$$

where the right hand side is inspired by Kusuoka representation of law-invariant coherent risk measure (cf. [18], [22, Section 6]). We note that

$$\inf_{q \in \mathbb{R}} \left\{ q + \xi^{-1} \int_{\mathbb{R}} (z - q)_+ P^{Z|\mathscr{U}_t^\pi}(\mathrm{d}z) \right\}$$

is the conditional version of AV@R$_\xi$ under the conditional distribution of Z given $\mathscr{U}_t^\pi$. The main goal of this paper is to develop a sample-based algorithm that solves the following infinite horizon risk averse MDP optimization problem

$$\inf_\pi \rho_{0,\infty}^\pi\left(\left(C(X_t^\pi, A_t^\pi, X_{t+1}^\pi)\right)_{t \in \mathbb{N}}\right), \quad \text{(II.3)}$$

where $\rho_{0,\infty}^\pi$ is defined analogously to (II.1).

The problem (II.3) can be solved using a dynamic programming principle. Specifically, we let S be the Bellman operator acting on $v : \mathbb{X} \to \mathbb{R}$ defined as

$$Sv(i) := \inf_{\lambda \in \mathcal{P}(\mathbb{A})} \sup_{\mu \in \mathcal{M}} \left\{ \int_0^1 \inf_{q \in \mathbb{R}} \left\{ q + \right.\right.$$
$$\left.\left. \xi^{-1} \sum_{k \in \mathbb{A}} \lambda_k \sum_{j \in \mathbb{X}} T_{ij}^k \left(C(i, k, j) + \gamma v(j) - q\right)_+ \right\} \mu(\mathrm{d}\xi) \right\}.$$
$$\text{(II.4)}$$

---

[1]We do not adopt verbatim the setting from [21] for the sake of smooth transition.

We can restrict v to take values in $[0, \frac{c_{\max}}{1-\gamma}]$ due to the boundedness of the cost. This allows us to replace $q \in \mathbb{R}$ in (II.4) with $q \in [0, \frac{c_{\max}}{1-\gamma}]$. It can be shown that S is a γ-contraction, and the fixed point of S, denoted by v^*, is the optimal value function. If $\pi^* : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ attains the infimum in $Sv^*(i)$ for all $i \in \mathbb{X}$, then π^* is the optimal stationary policy, and in fact, it is also optimal among all history-dependent policies. We refer to [7] for more discussion in a general setting.

III. DISTRIBUTIONAL METHOD FOR RISK-AVERSE LEARNING

In this section, we introduce a novel concept called g-values. We then propose a learning method based on g-values and establish a convergence result under suitable conditions, as stated in Theorem III.2. Finally, we provide a detailed description of the algorithm for implementing the method.

A. g-value

In view of (II.4), we define the Q-value as

$$Q(i, \lambda) := \sup_{\mu \in \mathcal{M}} \int_{[0,1]} \inf_{q \in \mathbb{R}} \Bigg\{ q + \xi^{-1} \sum_{k \in \mathbb{A}} \lambda_k$$
$$\sum_{j \in \mathbb{X}} T_{ij}^k \Big(C(i,k,j) + \gamma v^*(j) - q \Big)_+ \Bigg\} \mu(\mathrm{d}\xi), \quad \text{(III.1)}$$

and derive the following equation for Q-learning

$$Q(i, \lambda) = \sup_{\mu \in \mathcal{M}} \int_{[0,1]} \inf_{q \in \mathbb{R}} \Bigg\{ q + \xi^{-1} \sum_{k \in \mathbb{A}} \lambda_k$$
$$\sum_{j \in \mathbb{X}} T_{ij}^k \Big(C(i,k,j) + \gamma \inf_{\lambda \in \mathcal{P}(\mathbb{A})} Q(j,\lambda) - q \Big)_+ \Bigg\} \mu(\mathrm{d}\xi).$$
$$\text{(III.2)}$$

However, learning the Q function on a fine grid of $\mathbb{X} \times \mathcal{P}(\mathbb{A})$ turns out to be excessively expensive. Therefore, instead of continuing with (III.2), we propose to learn the following g-value[2][3]

$$g(i, \lambda, q) := \sum_{k \in \mathbb{A}} \lambda_k \sum_{j \in \mathbb{X}} T_{ij}^k \big(C(i,k,j) + \gamma v^*(j) - q \big)_+ .$$
$$\text{(III.3)}$$

Such g-value has an advantage of being linear in λ, which helps mitigates the cost of exploration. Moreover, it is worth noting that $q \mapsto g(i, \lambda, q)$ is non-increasing and 1-Lipschitz for any $(i, \lambda) \in \mathbb{X} \times \mathcal{P}(\mathbb{A})$, which will be useful in future analysis. Furthermore, we argue that g-value is aligned with

---

[2]By using g to express trapezoidal shaped functions and invoking dominated convergence (cf. [6, Theorem 2.8.1]) and monotone class theorem (cf. [6, Theorem 1.9.3 (ii)]), it can be shown that the function $q \mapsto \mathbb{E}((Z - q)_+)$, $q \in \mathbb{R}$, characterizes the distribution of Z.

[3]One may derive an equation for g-value in analogous to (III.2), but such equation needs not leads to a contraction in general.

our goal of solving (II.3), since by the aforementioned DPP and (III.3), we have

$$v^*(i) = \inf_{\lambda \in \mathcal{P}(\mathbb{A})} \sup_{\mu \in \mathcal{M}}$$
$$\Bigg\{ \int_{(0,1]} \inf_{q \in [0, \frac{c_{\max}}{1-\gamma}]} \Big\{ q + \xi^{-1} g(i, \lambda, q) \Big\} \mu(\mathrm{d}\xi) \Bigg\}.$$

This formula shows that the g-value is a crucial ingredient in our approach for solving (II.3).

B. Theoretical foundation

Suppose that we have observed the running states, actions and costs subject to some exploration policy upto time $t_{\max}$, resulting in a set of data $\{(x_t, a_t, x_{t+1}, c_t)\}_{t=1}^{t_{\max}-1}$, where $c_t = C(x_t, a_t, x_{t+1})$. In order to approximate g, we employ a parameterized model $f_\theta : \mathbb{X} \times \mathbb{A} \times \mathbb{R} \to \mathbb{R}$, where $\theta \in \Theta$ is the parameter, and $f_\theta(i, k, q)$ is designated to approximate $g(i, \delta_k, q)$, where δ_k is the Dirac measure on k. In view of (III.3), $g(i, \lambda, q)$ can be approximated by $\sum_{k \in \mathbb{A}} \lambda_k f_\theta(i, k, q)$. We use $\hat{\theta}$ to denote the estimate of the optimal parameter (if exists). In view of the bounded cost and positive discount factor, we approximate v^* with $\hat{v} : \mathbb{X} \to [0, \frac{c_{\max}}{1-\gamma}]$. Heuristically, we want to update $\hat{\theta}$ and $\hat{v}$ recursively in the following way

$$\begin{cases} \hat{\theta}_{n+1} \in \arg\min_{\theta \in \Theta} \sup_{(i,k,q) \in \mathbb{X} \times \mathbb{A} \times [0, \frac{c_{\max}}{1-\gamma}]} \\ \quad \Big(f_\theta(i,k,q) - \sum_{j \in \mathbb{X}} \hat{T}_{ij}^k \big(C(i,k,j) + \gamma \hat{v}_n(j) - q \big)_+ \Big)^2, \\ \hat{v}_{n+1}(i) = \inf_{\lambda \in \mathcal{P}(\mathbb{A})} \sup_{\mu \in \mathcal{M}} \int_{(0,1]} \\ \quad \inf_{q \in [0, \frac{c_{\max}}{1-\gamma}]} \Big\{ q + v\xi^{-1} \sum_{k \in \mathbb{A}} \lambda_k f_{\hat{\theta}_{n+1}}(i,k,q) \Big\} \mu(\mathrm{d}\xi), \end{cases}$$
$$\text{(III.4)}$$

where we define

$$\hat{T}_{ij}^k := \frac{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(x_t, a_t)}.$$

It is well-known that $\hat{T}_{ij}^k$ is the MLE of the transition probability (cf. [1]). Note that $\sum_{j \in \mathbb{X}} \hat{T}_{ij}^k \big(c_t + \gamma \hat{v}(x_{t+1}) - q \big)_+$ is a convex and 1-Lipschitz function of q that falls within the range $[0, \frac{c_{\max}}{1-\gamma}]$. Therefore, although the objective involves the supremum over an uncountable set, updating $\hat{\theta}$ is not infeasible. However, such an update requires knowledge of C. To circumvent this requirement, we observe that for q fixed,

$$\sum_{t=1}^{T-1} \big(y_{x_t a_t} - (c_t + \gamma \hat{v}(x_{t+1}) - q)_+ \big)^2$$
$$= \sum_{(i,k) \in \mathbb{X} \times \mathbb{A}} \sum_{t=1}^{T-1} \mathbb{1}_{(i,k)}(x_t, a_t) \sum_{j \in \mathbb{X}} \mathbb{1}_j(x_{t+1})$$
$$\big(y_{ik} - (C(i,k,j) + \gamma \hat{v}(j) - q)_+ \big)^2, \quad \text{(III.5)}$$

as a function of $(y_{ik})_{(i,k)\in\mathbb{X}\times\mathbb{A}}$, attains the infimum if

$$y_{ik} = \sum_{j\in\mathbb{X}} \hat{T}_{ij}^k \big(C(i,k,j) + \gamma\hat{v}(j) - q\big)_+, \quad (i,k)\in\mathbb{X}\times\mathbb{A}. \tag{III.6}$$

We can then use the following updating scheme as an alternative

$$\begin{cases} \hat{\theta}_{n+1} \in \underset{\theta\in\Theta}{\arg\min} \underset{q\in[0,\frac{c_{\max}}{1-\gamma}]}{\sup} \\ \displaystyle\sum_{t=1}^{T-1} \Big(f_\theta(x_t,a_t,q) - \big(c_t + \gamma\hat{v}_n(x_{t+1}) - q\big)_+\Big)^2, \\ \hat{v}_{n+1}(i) = \underset{\lambda\in\mathcal{P}(\mathbb{A})}{\inf}\underset{\mu\in\mathcal{M}}{\sup} \\ \displaystyle\int_{(0,1]}\underset{q\in[0,\frac{c_{\max}}{1-\gamma}]}{\inf}\Big\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k f_{\hat{\theta}_{n+1}}(i,a,q)\Big\}\mu(\mathrm{d}\xi). \end{cases} \tag{III.7}$$

In order to obtain a convergence result, we make the following technical assumption.

Assumption III.1. *Let* $c_{max} > 0$, $\ell \in \mathbb{N}$, $b,\varepsilon_e \in (0,1)$, $\varepsilon_\theta,\varepsilon_v > 0$ *be absolute constants. We assume that*

(i) *the range of the cost function C is contained by* $[0,c_{max}]$;
(ii) $\sup_{\mu\in\mathcal{M}}\mu([0,b]) = 0$;
(iii) $\{(X_t^\pi, A_t^\pi)\}_{t=1}^T$ *is subject to an exploration policy π such that*

$$\mathbb{P}\left(\sum_{r=t+1}^{t+\ell}\mathbb{1}_{(i,k)}(X_r^\pi,A_r^\pi) \geq 1\,\Big|\,\mathcal{F}_t^\pi\right) > \varepsilon_e,$$

for any $(t,i,k) \in \mathbb{N}\times\mathbb{X}\times\mathbb{A}$, *where* $\mathcal{F}_t^\pi :=\sigma(X_1^\pi, A_1^\pi, \ldots, X_t^\pi, A_t^\pi)$;
(iv) *regardless of the data and $\hat{v}:\mathbb{X}\to[0,\frac{c_{max}}{1-\gamma}]$, we always find $\hat{\theta}_{new}\in\Theta$ and $\hat{v}_{new}:\mathbb{X}\to[0,\frac{c_{max}}{1-\gamma}]$ such that, for all* $(i,k,q)\in\mathbb{X}\times\mathbb{A}\times[0,\frac{c_{max}}{1-\gamma}]$,

$$\left| f_{\hat{\theta}_{new}}(i,k,q) - \sum_{j\in\mathbb{X}}\hat{T}_{ij}^k\big(C(i,k,j)+\gamma\hat{v}(x_{t+1})-q\big)_+\right|$$
$$\leq \varepsilon_\theta, \tag{III.8}$$

and

$$\underset{i\in\mathbb{X}}{\sup}\left|\hat{v}_{new}(i) - \underset{\lambda\in\mathcal{P}(\mathbb{A})}{\inf}\underset{\mu\in\mathcal{M}}{\sup}\int_0^1 \right.$$
$$\left.\underset{q\in[0,\frac{c_{max}}{1-\gamma}]}{\inf}\Big\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k f_{\hat{\theta}_{new}}(i,k,q)\Big\}\mu(\mathrm{d}\xi)\right| \leq \varepsilon_v. \tag{III.9}$$

Condition (i) and (ii) follows automatically from the setting above; these conditions are included in the assumption for the sake of easy navigation. Condition (iii) is a version of parallel sampling model (PSM). PSM was originally introduced in [16] and is commonly used in reinforcement learning literature as an exploration policy that achieves perfect exploration (cf. [11]). Condition (iv) regards the accuracy of the update. In particular, (III.8) corresponds to the computation of $\hat{\theta}_{n+1}$ in

(III.7). Based on the separability of the objective illustrated in (III.5), the convexity of $\big(y_{ik} - (C(i,k,j)+\gamma\hat{v}(j)-q)_+\big)^2$ in y_{ik}, and the observed good behavior of $q\mapsto\sum_{j\in\mathbb{X}}\hat{T}_{ij}^k\big(c_t + \gamma\hat{v}(x_{t+1}) - q\big)_+$, we consider (III.8) reasonable.

Below is our main result. The proof is deferred to the appendix.

Theorem III.2. *Suppose Assumption III.1. Let* $t_{\max} > \ell$. *Given data* $\{(x_t,a_t,x_{t+1},c_t)\}_{t=1}^{t_{\max}-1}$ *and an arbitrary* $\hat{v}_0 : \mathbb{X}\to[0,\frac{c_{max}}{1-\gamma}]$, *we compute* $\{(\hat{\theta}_n,\hat{v}_n)\}_{n\in\mathbb{N}}$ *according to (III.7), approximately as in Assumption III.1 (iv). Then, for any* $\varepsilon\in(0,1]$, *there is a probability of at least*

$$1 - 3|\mathbb{X}|^2|\mathbb{A}|\left(e^{-\frac{\varepsilon_e^2}{4}\lfloor\frac{t_{\max}-1}{\ell}\rfloor} + e^{-\frac{\varepsilon^2\varepsilon_e^2}{8\ell}\lfloor\frac{t_{\max}-1}{\ell}\rfloor}\right)$$

that

$$\|\hat{v}_n - v^*\|_\infty \leq \gamma^n\|\hat{v}_0 - v^*\|_\infty + \frac{b^{-1}c_{\max}\varepsilon}{(1-\gamma)^2} + \frac{b^{-1}\varepsilon_\theta + \varepsilon_v}{1-\gamma}.$$

for all $n\in\mathbb{N}$.

Sometimes it is advisable to assume that ε_e is proportional to $(|\mathbb{X}||\mathbb{A}|)^{-1}$. In order to maintain the same level of accuracy (in terms of the probability bound), we need to set $t_{\max}\propto|\mathbb{X}|^2|\mathbb{A}|^2\log(|\mathbb{X}|^2|\mathbb{A}|)$. In this case, the effort required for exploration only needs to grow polynomially as $|\mathbb{X}||\mathbb{A}|$ increases.

C. Implementation

In our algorithm, we use a deep neural network for f_θ. We let $m_{\mathrm{grid}}\in\mathbb{N}$ and $(q_1,\ldots,q_{m_{\mathrm{grid}}})$ be a pre-selected grid on $[0,\frac{c_{\max}}{1-\gamma}]$. We are given data $\{(x_t,a_t,x_{t+1},c_t)\}_{t=1}^{t_{\max}-1}$, and an a priori guess $\hat{v}$ of the value function.

Instead of following strictly (III.7), we consider the updating procedure below

$$\hat{\theta}_{n+1}\in\underset{\theta\in\Theta}{\arg\min}$$
$$\sum_{m=1}^{m_{\mathrm{grid}}}\sum_{t=1}^{T-1}\left(f_\theta(x_t,a_t,q_m) - \big(c_t + \gamma\hat{v}_n(x_{t+1}) - q_m\big)_+\right)^2$$
$$+ \beta\psi(\theta), \tag{III.10}$$

where ψ is a penalization for ensuring monotonicity on $q\mapsto f_\theta(i,k,q)$, defined as

$$\psi(\theta) := \sum_{(i,k)\in\mathbb{X}\times\mathbb{A}}\sum_{m=1}^{m_{\mathrm{grid}}-1}\big(f_\theta(i,k,q_{m+1}) - f_\theta(i,k,q_m)\big)_+,$$

and $\beta\geq 0$ is the regularization parameter. We use stochastic gradient descent for $\arg\min_{\theta\in\Theta}$. After (III.10) is done, we perform

$$\hat{v}_{n+1}(i) = \underset{\lambda\in\mathcal{P}(\mathbb{A})}{\inf}\underset{\mu\in\mathcal{M}}{\max}$$
$$\int_{(0,1]}\underset{q\in[0,\frac{c_{\max}}{1-\gamma}]}{\inf}\Big\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k f_{\hat{\theta}_{n+1}}(i,a,q)\Big\}\mu(\mathrm{d}\xi), \tag{III.11}$$

where we recall that $\mathcal{M}$ is a finite set of discrete probabilities on $[0,1]$, and thus the $\int_{(0,1]}$ is in fact a finite sum. We use gradient descent with random initialization for $\inf_{q\in[0,\frac{c_{\max}}{1-\gamma}]}$, and random search for $\inf_{\lambda\in\mathcal{P}(\mathbb{A})}$. After (III.11) is done, we may return to (III.10) for next round of update. In order to obtain an approximated optimal policy $\hat{\pi}$, we should record the approximated minimizor of $\inf_{\lambda\in\mathcal{P}(\mathbb{A})}$ for each i.

We summarize the implementation in Algorithm (III-C). We point out that, Algorithm (III-C) can be integrate asynchronously into a larger implementation that involves running data.

Algorithm 1 Distributional Method for Risk-Averse RL

Input: data $\{(x_t, a_t, x_{t+1}, c_t)\}_{t=1}^{t_{\max}-1}$, aprior guess $\hat{v}$
Output: $\hat{v}, \hat{\pi}$
 1: **repeat**
 2:     Randomly initialize θ, or use previous θ if given
 3:     **while** find θ **do**
 4:       Stochastic gradient descent according to (III.10)
 5:     **end while**
 6:     **while** update $\hat{v}$ and $\hat{\pi}$ **do**
 7:       Random search according to (III.11)
         Record the best $\lambda\in\mathcal{P}(\mathbb{X})$ for each $i\in\mathbb{X}$
 8:     **end while**
 9: **until** Convergence
10: Update $\hat{\pi}(i)$ according to the record
11: **return** $\hat{v}, \hat{\pi}$

IV. Numeric experiments

In this section, we present numerical experiments to validate the performance of Algorithm III-C. We consider a state-action space with $|\mathbb{X}| = |\mathbb{A}| = 4$, and a discount factor of $\gamma = 0.3$. We use the following $\mathcal{M}$ for the conditional risk mapping (II.2):

$$\mathcal{M} = \big\{ 0.2\delta_{0.2} + 0.8\delta_1, \delta_{0.5}, 0.1\delta_{0.05} + 0.5\delta_{0.4} + 0.6\delta_{0.6}, \\ 0.5\delta_{0.3} + 0.5\delta_{0.8} \big\},$$

where δ denotes the Dirac measure. The transition matrices used in the experiment are randomly generated. Although our algorithm was introduced for deterministic latent costs, it also handles random costs without requiring significant modifications. We test the algorithm with various random costs, such as $\mathrm{Beta}(\alpha,\beta)$ with $\alpha,\beta : \mathbb{X}\times\mathbb{A}\times\mathbb{X}\to(0,\infty)$ depending on the current state, the realized action, and next state. We assume the knowledge of $[0, c_{\max}]$, and set $(q_1,\dots,q_{m_{\mathrm{grid}}})$ as a uniform partition of $[0, c_{\max}]$ with $m_{\mathrm{grid}} = 100$. We set $t_{\max} = 10000$ and sample according to some randomly picked stationary policy. We then compute $\hat{v}$ and $\hat{\pi}$ using Algorithm III-C. To ensure accuracy when updating $\hat{v}$ and $\hat{\pi}$, we perform a thorough random search. However, we conjecture that there is a certain structure that we can take advantage of in learning $\hat{v}$ and $\hat{\pi}$, and the computation cost does not grow exponentially as $|\mathbb{A}|$ increases. In Figure 1, we plot the relative errors of $\hat{v}$ for each $i\in\mathbb{X}$ in 10 different experiments. The benchmark in each experiment is computed using brute force search.
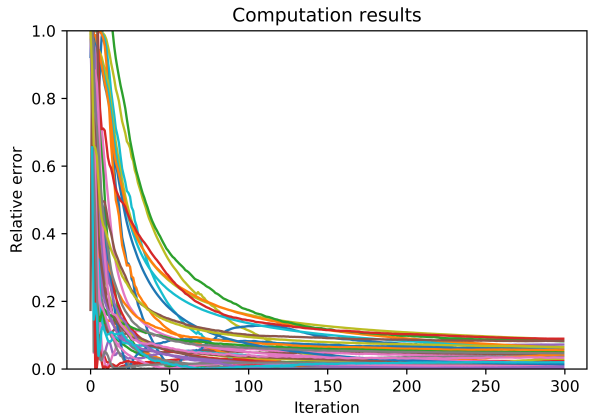

Fig. 1. Relative errors of all value functions in 10 experiments

References

[1] T. W. Anderson and L. A. Goodman, "Statistical inference about markov chains," *The Annals of Mathematical Statistics*, vol. 28, no. 1, p. 89–110, 1957.

[2] N. Bäuerle and A. Glauner, "Markov decision processes with iterated coherent risk measures," *European Journal of Operational Research*, vol. 296, no. 3, pp. 953–966, 2022.

[3] N. Bäuerle and U. Rieder, "More risk-sensitive markov decision processes," *Mathematics of Operations Research*, vol. 39, no. 1, pp. 105–120, 2013.

[4] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," *Proceedings of Machine Learning Research*, vol. 70, pp. 449–458, 2017.

[5] T. R. Bielecki, T. Chen, and I. Cialenco, "Risk-sensitive markov decision problems under model uncertainty: finite time horizon case," *arXiv:2104.06915*, 2021.

[6] V. I. Bogachev, *Measure Theory Volume I*. Springer-Verlag Berlin Heidelberg, 2007.

[7] Z. Cheng and S. Jaimungal, "Markov decision processes with kusuoka-type conditional risk mappings," *Preprint*, 2022.

[8] S. Chu and Y. Zhang, "Markov decision processes with iterated coherent risk measures," *International Journal of Control*, vol. 88, no. 11, pp. 2286–2293, 2014.

[9] A. Coache and S. Jaimungal, "Reinforcement learning with convex risk measure," *Preprint*, 2022.

[10] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," *Proceedings of Machine Learning Research*, vol. 80, pp. 1104–1131, 2018.

[11] E. Even-Dar and Y. Mansour, "Learning rates for q-learning," *Journal of Machine Learning Research*, vol. 5, pp. 1–25, 2003.

[12] A. Huang, L. Leiqi, Z. C. Lipton, and K. Azizzadenesheli, "On the convergence and optimality of policy gradient for markov coherent risk," *Preprint*, 2021.

[13] W. Huang and W. B. Haskell, "Risk-aware q-learning for markov decision processes," *IEEE 56th Annual Conference on Decision and Control*, 2017.

[14] ——, "Stochastic approximation for risk-aware markov decision processes," *IEEE 56th Annual Conference on Decision and Control*, vol. 66, no. 3, pp. 1314–1320, 2021.

[15] S. C. Jaquette, "Markov decision processes with a new optimality criterion: Discrete time," *The Annals of Statistics*, vol. 1, no. 3, pp. 496–505, 1973.

[16] M. K. and S. P. Singh, "Finite-sample convergence rates for q-learning and indirect algorithms," *Advances in Neural Information Processing Systems*, vol. 11, p. 996–1002, 1999.

[17] U. Köse and A. Ruszczyński, "Risk-averse learning by temporal difference methods with markov risk measures," *Journal of Machine Learning Research*, vol. 22, no. 38, pp. 1–34, 2021.

[18] S. Kusuoka, "On law invariant coherent risk measures," *Advances in Mathematical Economics*, p. 83–95, 2000.

[19] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka, "Nonparametric return distribution approximation for reinforcement learning," *Proceedings of the 27th International Conference on Machine Learning*, pp. 799–806, 2010.

[20] T. Nguyen-Tang, S. Gupta, and S. Venkatesh, "Distributional reinforcement learning via moment matching," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 9144–9152, 2021.

[21] A. Ruszczyński, "Risk-averse dynamic programming for markov decision processes," *Mathematical Programming, Series B*, vol. 125, pp. 235–261, 2010.

[22] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. Springer, 2021.

[23] Y. Shen, W. Stannat, and K. Obermayer, "Risk-sensitive markov control processes," *SIAM Journal on Control and Optimization*, vol. 51, no. 5, pp. 3652—-3672, 2013.

[24] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," *Neural Computation*, vol. 26, no. 7, pp. 1298–1328, 2014.

[25] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Sequential decision making with coherent risk," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3323–3338, 2017.

[26] ——, "Policy gradient for coherent risk measures," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 1468–1476.

[27] D. J. White, "Mean, variance, and probabilistic criteria in finite markov decision processes: a review," *Journal of Optimization Theory and Applications*, vol. 56, no. 1, pp. 1–29, 1988.

[28] D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu, "Fully parameterized quantile function for distributional reinforcement learning," *33rd Conference on Neural Information Processing Systems*, pp. 6190–6199, 2019.

[29] P. Yu, W. B. Haskell, and H. Xu, "Approximate value iteration for risk-aware markov decision processes," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 3135–3142, 2018.

[30] F. Zhou, Z. Zhu, Q. Kuang, and L. Zhang, "Non-decreasing quantile function network with efficient exploration for distributional reinforcement learning," *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3455–3461, 2021.

We fix $t_{\max} > 1$ and $\pi : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ for the remainder of this section. Firstly, we will introduce the contraction property of S.

Lemma A.1. *For any* $v, v' : \mathbb{X} \to \mathbb{R}$, $\|Sv - Sv'\|_\infty \le \gamma \|v - v'\|_\infty$.

Proof. This is an immediate consequence of [7, Lemma 3.3]. $\qquad\square$

The proof of Theorem III.2 is also dependent on the following two technical lemmas.

Lemma A.2. *For any* $(i, j, k) \in \mathbb{X} \times \mathbb{A} \times \mathbb{X}$, $\varepsilon \in (0, 1)$ *and integer* $N < \varepsilon_e \lfloor \frac{t_{\max} - 1}{\ell} \rfloor$, *we have*

$$\mathbb{P}\left(\left| \frac{\sum_{t=1}^{t_{max}-1} \mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(x_t, a_t)} - T_{ij}^k \right| > \varepsilon \right) \le \exp\left(-\frac{(N - \varepsilon_e \lfloor \frac{t_{\max}-1}{\ell} \rfloor)^2}{\lfloor \frac{t_{\max}-1}{\ell} \rfloor} \right) + 2\exp\left(-\frac{\varepsilon^2 N^2}{2 t_{\max}} \right).$$

Proof. To start with note that

$$\left\{ \left| \frac{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(X_t, X_t, X_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_t, X_t)} - T_{ij}^k \right| \ge \varepsilon \right\}$$

$$\subseteq \left\{ \sum_{r=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi) < N \right\} \cup \left(\left\{ \sum_{r=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi) \ge N \right\} \cap \left\{ \left| \frac{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(X_t, X_t, X_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_t, X_t)} - T_{ij}^k \right| \ge \varepsilon \right\} \right)$$

$$\subseteq \left\{ \sum_{r=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi) < N \right\} \cup \left\{ \left| \sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(X_t, X_t, X_{t+1}) - T_{ij}^k \sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_t, X_t) \right| \ge \varepsilon N \right\}.$$

Therefore,

$$\mathbb{P}\left(\left| \frac{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(x_t, a_t)} - T_{ij}^k \right| > \varepsilon \right)$$

$$\le \mathbb{P}\left(\sum_{r=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi) < N \right) + \mathbb{P}\left(\left| \sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(X_t, X_t, X_{t+1}) - T_{ij}^k \sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_t, X_t) \right| \ge \varepsilon N \right). \quad \text{(A.1)}$$

In order to investigate the first term in right hand side of (A.1), we introduce an auxiliary process. For $\iota = 1, \ldots, \lfloor \frac{t_{\max}-1}{\ell} \rfloor$, we let

$$L_\iota := \sum_{t=1}^{\ell\iota} \mathbb{1}_{(i,k)}(X_t, A_t) - \varepsilon_e \iota.$$

Note that $(L_\iota)_{\iota=1}^{\lfloor \frac{t_{\max}-1}{\ell} \rfloor}$ is a sub-martingale under the filtration $(\mathscr{F}_{\ell\iota}^\pi)_{r=1}^{\lfloor \frac{t_{\max}-1}{\ell} \rfloor}$. Indeed, by the Markov property of $\{(X_t^\pi, A_t^\pi)\}_{t \in \mathbb{N}}$, we have

$$\mathbb{E}\left(L_{\iota+1} \big| \mathscr{F}_{\ell\iota}^\pi \right) = L_\iota + \mathbb{E}\left(\sum_{t=\ell\iota+1}^{\ell(\iota+1)} \mathbb{1}_{(i,k)}(X_t, A_t) - \varepsilon_e \big| \mathscr{F}_{\ell\iota}^\pi \right) \ge L_\iota,$$

where we have used Assumption III.1 (iii) in the last equality. Then, by Azuma's inequality, for $N < \varepsilon_e \lfloor \frac{t_{\max}-1}{\ell} \rfloor$,

$$\mathbb{P}\left(\sum_{r=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi) < N \right) \le \mathbb{P}\left(L_{\lfloor \frac{t_{\max}-1}{\ell} \rfloor} \le N - \varepsilon_e \left\lfloor \frac{t_{\max}-1}{\ell} \right\rfloor \right)$$

$$\le \exp\left(-\frac{(N - \varepsilon_e \lfloor \frac{t_{\max}-1}{\ell} \rfloor)^2}{\lfloor \frac{t_{\max}-1}{\ell} \rfloor} \right). \quad \text{(A.2)}$$

Regarding the second term in (A.1), we define $M_1^{ikj} := 0$ and

$$M_t^{ikj} := \sum_{r=1}^{t-1} \mathbb{1}_{(i,k,j)}(X_r^\pi, A_r^\pi, X_{r+1}^\pi) - T_{ij}^k \sum_{r=1}^{t-1} \mathbb{1}_{(i,k)}(X_r^\pi, A_r^\pi), \quad t \ge 2.$$

Note that $(M_t^{ikj})_{t\in\mathbb{N}}$ is a $(\mathcal{F}_t^\pi)_{t\in\mathbb{N}}$-martingale:

$$\mathbb{E}\big(M_{t+1}^{ikj}\big|\mathcal{F}_t^\pi\big) = M_t^{ikj} + \mathbb{E}\bigg(\mathbb{1}_{(i,k,j)}(X_t^\pi, A_t^\pi, X_{t+1}^\pi) - T_{ij}^k\mathbb{1}_{(i,k)}(X_t^\pi, A_t^\pi)\bigg|\mathcal{F}_t^\pi\bigg)$$

$$= M_t^{ikj} + \mathbb{E}\bigg(\mathbb{1}_{(i,k,j)}(X_t^\pi, A_t^\pi, X_{t+1}^\pi) - T_{ij}^k\mathbb{1}_{(i,k)}(X_t^\pi, A_t^\pi)\bigg|\sigma(X_t^\pi, A_t^\pi)\bigg) = M_t^{ikj},$$

where we have used the Markov property of $\{(X_t^\pi, A_t^\pi)\}_{t\in\mathbb{N}}$ in the second line. It follows from Azuma's inequality that

$$\mathbb{P}\big(\big|M_{t_{\max}}^{ijk}\big| \geq \varepsilon N\big) \leq \exp\left(-\frac{\varepsilon^2 N^2}{2t_{\max}}\right). \tag{A.3}$$

Finally, by combining (A.1), (A.2) and (A.3), we complete the proof. □

Lemma A.3. *Let v^* be the fixed point of S defined in* (II.4). *Let $\hat{v}$ and $\hat{v}_{new}$ be introduced as in Assumption* III.1 *(iv). Then,*

$$\|\hat{v}_{new} - v^*\|_\infty \leq \gamma\|\hat{v} - v^*\|_\infty + b^{-1}\bigg(\sup_{(i,k,j)\in\mathbb{X}\times\mathbb{A}\times\mathbb{A}}|\hat{T}_{ij}^k - T_{ij}^k|\frac{c_{\max}}{1-\gamma} + \varepsilon_\theta\bigg) + \varepsilon_v$$

Proof. To start with, by (III.9) and the fact that $v^* = Sv^*$,

$$\|\hat{v}_{new} - v^*\|_\infty \leq \sup_{i\in\mathbb{X}}\bigg|\inf_{\lambda\in\mathcal{P}(\mathbb{A})}\sup_{\mu\in\mathcal{M}}\bigg\{\int_{(0,1]}\inf_{q\in[0,\frac{c_{\max}}{1-\gamma}]}\bigg\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k f_{\hat{\theta}_{new}}(i,k,q)\bigg\}\mu(\mathrm{d}\xi)\bigg\} - Sv^*(i)\bigg| + \varepsilon_v.$$

Then, by Assumption III.1 (ii), (III.8) and Lemma A.1,

$$\|\hat{v}_{new} - v^*\|_\infty \leq \sup_{i\in\mathbb{X}}\bigg|\inf_{\lambda\in\mathcal{P}(\mathbb{A})}\sup_{\mu\in\mathcal{M}}\bigg\{\int_{(0,1]}\inf_{q\in[0,\frac{c_{\max}}{1-\gamma}]}\bigg\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k\sum_{j\in\mathbb{X}}\hat{T}_{ij}^k\big(c_t + \gamma\hat{v}(x_{t+1}) - q\big)_+\bigg\}\mu(\mathrm{d}\xi)\bigg\} - Sv^*(i)\bigg|$$

$$+ b^{-1}\varepsilon_\theta + \varepsilon_v$$

$$\leq \sup_{i\in\mathbb{X}}\bigg|\inf_{\lambda\in\mathcal{P}(\mathbb{A})}\sup_{\mu\in\mathcal{M}}\bigg\{\int_{(0,1]}\inf_{q\in[0,\frac{c_{\max}}{1-\gamma}]}\bigg\{q + \xi^{-1}\sum_{k\in\mathbb{A}}\lambda_k\sum_{j\in\mathbb{X}}\hat{T}_{ij}^k\big(c_t + \gamma\hat{v}(x_{t+1}) - q\big)_+\bigg\}\mu(\mathrm{d}\xi)\bigg\} - S\hat{v}(i)\bigg|$$

$$+ \|S\hat{v} - Sv^*\|_\infty + b^{-1}\varepsilon_\theta + \varepsilon_v$$

$$\leq \gamma\|\hat{v} - v^*\|_\infty + b^{-1}\bigg(\sup_{(i,k,j)\in\mathbb{X}\times\mathbb{A}\times\mathbb{A}}|\hat{T}_{ij}^k - T_{ij}^k|\frac{c_{\max}}{1-\gamma} + \varepsilon_\theta\bigg) + \varepsilon_v.$$

The proof is complete. □

We are now in position to prove Theorem III.2.

Proof of Theorem III.2. We first simplify Lemma A.2 by letting $N = \lceil\frac{1}{2}\varepsilon_e\lfloor\frac{t_{\max}-1}{\ell}\rfloor\rceil$

$$\mathbb{P}\bigg(\bigg|\frac{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k)}(x_t, a_t)} - T_{ij}^k\bigg| > \varepsilon\bigg) \leq \exp\bigg(-\frac{\lfloor\frac{1}{2}\varepsilon_e\lfloor\frac{t_{\max}-1}{\ell}\rfloor\rfloor^2}{\lfloor\frac{t_{\max}-1}{\ell}\rfloor}\bigg) + 2\exp\bigg(-\frac{\varepsilon^2\varepsilon_e^2\lfloor\frac{t_{\max}-1}{\ell}\rfloor^2}{8t_{\max}}\bigg)$$

$$\leq e^{\varepsilon_e}\exp\bigg(-\frac{\varepsilon_e^2}{4}\bigg\lfloor\frac{t_{\max}-1}{\ell}\bigg\rfloor\bigg) + 2\exp\bigg(-\frac{\varepsilon^2\varepsilon_e^2(\frac{t_{\max}}{\ell}-1)\lfloor\frac{t_{\max}-1}{\ell}\rfloor}{8t_{\max}}\bigg)$$

$$\leq 3\exp\bigg(-\frac{\varepsilon_e^2}{4}\bigg\lfloor\frac{t_{\max}-1}{\ell}\bigg\rfloor\bigg) + 3\exp\bigg(-\frac{\varepsilon^2\varepsilon_e^2}{8\ell}\bigg\lfloor\frac{t_{\max}-1}{\ell}\bigg\rfloor\bigg),$$

where we have used the fact that $\lfloor\frac{t_{\max}-1}{\ell}\rfloor \geq \frac{t_{\max}}{\ell} - 1 \geq 0$ in the second line. Consequently,

$$\mathbb{P}\bigg(\bigg|\frac{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k)}(x_t, a_t)} - T_{ij}^k\bigg| \leq \varepsilon, (i,k,j) \in \mathbb{X}\times\mathbb{A}\times\mathbb{X}\bigg)$$

$$\geq 1 - \sum_{(i,j,k)\in\mathbb{X}\times\mathbb{A}\times\mathbb{X}}\mathbb{P}\bigg(\bigg|\frac{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k,j)}(x_t, a_t, x_{t+1})}{\sum_{t=1}^{t_{\max}-1}\mathbb{1}_{(i,k)}(x_t, a_t)} - T_{ij}^k\bigg| > \varepsilon, (i,k,j) \in \mathbb{X}\times\mathbb{A}\times\mathbb{X}\bigg)$$

$$\geq 1 - 3|\mathbb{X}|^2|\mathbb{A}|\bigg(\exp\bigg(-\frac{\varepsilon_e^2}{4}\bigg\lfloor\frac{t_{\max}-1}{\ell}\bigg\rfloor\bigg) + \exp\bigg(-\frac{\varepsilon^2\varepsilon_e^2}{8\ell}\bigg\lfloor\frac{t_{\max}-1}{\ell}\bigg\rfloor\bigg)\bigg)$$

Finally, under the realization that $\left| \frac{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k,j)}(x_t,a_t,x_{t+1})}{\sum_{t=1}^{t_{\max}-1} \mathbb{1}_{(i,k)}(x_t,a_t)} - T_{ij}^k \right| \leq \varepsilon$ for all $(i,k,j) \in \mathbb{X} \times \mathbb{A} \times \mathbb{X}$, invoking (A.3) iteratively, we yield

$$\|\hat{v}_n - v^*\|_\infty \leq \gamma^n \|\hat{v}_0 - v^*\| + \frac{b^{-1}}{1-\gamma}\left(\frac{c_{\max}}{1-\gamma}\varepsilon + \varepsilon_\theta\right) + \frac{\varepsilon_v}{1-\gamma},$$

which completes the proof. $\qquad\square$