Journal of Machine Learning Research 22 (2021) 1-34 Submitted 2/20; Revised 12/20; Published 1/21

Risk-Averse Learning by Temporal Difference Methods with Markov Risk Measures

Ümit Köse UEK1@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers University
Piscataway, NJ 08854, USA

Andrzej Ruszczyński RUSZ@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers University
Piscataway, NJ 08854, USA

Editor: Jan Peters

Abstract

We propose a novel reinforcement learning methodology where the system performance is evaluated by a Markov coherent dynamic risk measure with the use of linear value function approximations. We construct projected risk-averse dynamic programming equations and study their properties. We propose new risk-averse counterparts of the basic and multi-step methods of temporal differences and we prove their convergence with probability one. We also perform an empirical study on a complex transportation problem.

Keywords: Reinforcement Learning, Temporal Difference Methods, Linear Function Approximation, Dynamic Risk Measures, Stochastic Approximation

1. Introduction

The objective of this paper is to propose and analyze new risk-averse reinforcement learning methods for Markov Decision Processes (MDPs). Our goal is to combine the efficacy of the methods of temporal differences with the robustness of Markov dynamic risk measures, and to provide a rigorous mathematical analysis of the methods.

MDPs are well-known models of stochastic sequential decision problems, covered in multiple monographs (Bellman, 1957; Howard, 1960; Puterman, 1994; Bertsekas, 2017), and having countless applications. In the classical setting, the goal of an MDP is to find a control policy in a controlled Markov chain that minimizes the expected sum or the expected average of stage-wise costs over a finite or infinite horizon. Traditional MDP models, although effective for small to medium size problems, suffer from the curse of dimensionality in problems with large state spaces. Approximate dynamic programming approaches try to tackle the curse of dimensionality and provide an approximate solution of an MDP (see Powell (2011) for an overview). Many such methods, originating in (Bellman et al., 1963), involve value function approximations, where the value of a state of the Markov process is approximated by a simple, usually linear, function of some selected *features* of the state.

Reinforcement learning methods (Sutton and Barto, 1998; Powell, 2011) involve simulation or observation of a Markov process to approximate the value function and learn the corresponding policies. The first studies attempted to emulate neural networks and biological learning processes, learning by trial and error (Minsky, 1954; Farley and Clark, 1954). Some learning algorithms, such as Q-Learning (Watkins, 1989; Watkins and Dayan, 1992) and SARSA (Rummery and Niranjan, 1994), follow this idea. One of core approaches in reinforcement learning is the method of temporal differences (Sutton, 1988), known as TD(λ). It uses differences between the values of the approximate value function at successive states to improve the approximation, concurrently with the evolution of the system. TD(λ) is a continuum of algorithms depending on a parameter $\lambda \in [0, 1]$ which is used to exponentially weight past observations. Consequently, related methods such as Q(λ) (Watkins, 1989; Peng, 1993; Peng and Williams, 1994; Rummery, 1995) and SARSA(λ) were developed (Rummery and Niranjan, 1994; Rummery, 1995). The methods of temporal differences have been proven to converge in the mean (to some limit) in (Dayan, 1992) and almost surely by several studies, with different degrees of generality and precision (Peng, 1993; Dayan and Sejnowski, 1994; Tsitsiklis, 1994; Jaakkola et al., 1994). Almost sure convergence of the stochastic TD(λ) method with linear function approximation to a solution of a projected dynamic programming equation was proved by Tsitsiklis and Van Roy (1997). One of our objectives is to generalize this result to a risk-averse setting.

In the extant literature, three basic approaches to risk aversion in MDPs have been employed: utility models (Jaquette, 1973; Chung and Sobel, 1987; Fleming and Sheu, 1999; Bäuerle and Rieder, 2013; Jaśkiewicz et al., 2013), mean–variance models (White, 1988; Filar et al., 1989; Mannor and Tsitsiklis, 2013; Arlotto et al., 2014; Chen et al., 2014), and entropic (exponential) models (Howard and Matheson, 1971; Marcus et al., 1997; Bielecki et al., 1999; Coraluppi and Marcus, 1999; Di Masi and Stettner, 1999; Levitt and Ben-Israel, 2001; Bäuerle and Rieder, 2013).

Our research is rooted in the theory of dynamic measures of risk, which has been intensively developed in the last 15 years; see (Scandolo, 2003; Riedel, 2004; Roorda et al., 2005; Cheridito et al., 2006; Ruszczyński and Shapiro, 2006b; Artzner et al., 2007; Pflug and Römisch, 2007; Cheridito and Kupper, 2011) and the references therein. The main difference of this line of research from the earlier approaches is its axiomatic foundation, which, although originating from finance, has a general appeal. Ruszczyński (2010) introduced Markov dynamic risk measures, specially tailored for the MDPs. It allowed for the development of dynamic programming equations and corresponding solution methods, generalizing the well-known results for the expected value problems. Our ideas were successfully extended to undiscounted problems (Çavus and Ruszczyński, 2014a,b), partially observable and history-dependent systems (Fan and Ruszczyński, 2018b,a), average cost problems (Shen et al., 2013), and to mixed risk aversion and seeking (Lin and Marcus, 2013). The Markov risk measures are related to robust dynamic programming introduced by Nilim and El Ghaoui (2005) and analyzed by Iyengar (2005), but they do not require explicit description of the uncertainty sets for all state-control pairs. However, all these approaches require the storage and processing of the policy value functions and the optimal value function for all possible states of the system, thus limiting their applicability.

A number of works introduce models of risk into reinforcement learning: exponential utility functions (Borkar, 2001, 2002; Basu et al., 2008) and mean-variance models (Tamar

et al., 2012; Prashanth and Ghavamzadeh, 2014). A few later studies propose heuristic approaches involving specific coherent risk measures, such as CVaR in the objective or constraints (Chow and Ghavamzadeh, 2014; Chow et al., 2017; Ma et al., 2018). Simulation-based value iteration with coherent measures was analyzed by Yu et al. (2018). Risk-aware Q-learning with Markov risk measures is considered by Huang and Haskell (2017). Again, all these methods are applicable to problems with a small number od state-action pairs allowing exhaustive experimentation.

Value function approximations in the context of distributionally robust MDPs were considered by Tamar et al. (2014), who introduced the corresponding projected dynamic equation and proposed a sampling approach to value iteration and policy iteration, together with a robust version of the Least-Squares Temporal Differences method in a vector–matrix form. We advance this direction of research in the context of Markov risk measures, by establishing stronger properties of the projected equation, introducing basic and multi-step risk-averse methods of temporal differences in a stochastic approximation setting, and by providing mathematically precise convergence results. Tamar et al. (2017) study the policy gradient approach for Markov risk measures and use it in an actor-critic type algorithm. They propose projected risk-sensitive value iteration with linear function approximations and derive a formula for the gradient with respect to policy parameters, involving extensive simulation of state-control pairs and dual representation of risk. Di-Castro Shashua and Mannor (2017) augment this approach with a heuristic blend of robust dynamic programming, extended Kalman filter, and deep Q-learning, in a Bayesian setting, albeit without convergence proof. Distributed policy gradient methods with risk measures were proposed by Ma et al. (2017). Distributional reinforcement learning (Bellemare et al., 2017) avoids dealing with scalar value functions, formulates dynamic programming equations in the space of distributions, and parameterizes all possible distributions of the value functions to obtain a more tractable model. Parametrization of state-dependent probability measures is a challenge in this approach.

In this paper, we use Markov risk measures in conjunction with linear approximations of the risk-averse policy evaluation and temporal difference learning in the basic and multi-step settings. Our contributions can be summarized as follows:

- A projected risk-averse dynamic programming equation and its analysis (section 2). In particular, we prove the existence of solutions to the equation under weaker assumptions than conditions used before for a related problem in a distributionally robust setting.
- A stochastic risk-averse method of temporal differences with linear function approximation (section 3) and proof of its convergence with probability one in a simulation setting (section 4). This appears to be the first rigorous convergence proof of a temporal difference learning method with dynamic risk measures.
- A novel stochastic multistep risk-averse method of temporal differences with linear function approximation, the new projected multistep dynamic programming equation, and the analysis of its properties (section 5).
- The proof of convergence of the multistep methods in a simulation setting (section 6), involving novel techniques for analyzing the accumulation of risk.

- An empirical study illustrating the operation of the methods on an example with state space size of order 10^{427} (section 7).

The proposed methodology allows for flexible modeling of risk aversion, easy implementation, scalability to large state spaces, and enjoys mathematical convergence guarantees. We are not aware of other reinforcement learning methods involving dynamic risk measures and linear value function approximations, and having rigorous convergence proofs.

2. The Projected Risk-Averse Dynamic Programming Equation

We consider a Markov Decision Process with a finite state space $\mathcal{X} = \{1, \dots, n\}$, finite action sets $\mathcal{U}(i)$ for all $i \in \mathcal{X}$, controlled transition probabilities $P_{ij}(u)$ where $i, j \in \mathcal{X}$ and $u \in \mathcal{U}(i)$, and one-step cost function $c(i, u)$, where $i \in \mathcal{X}$ and $u \in \mathcal{U}(i)$. For a discount factor $\alpha \in (0, 1)$ and any non-anticipative policy π for determining controls $u_t \in \mathcal{U}(i_t)$, $t = 0, 1, 2, \dots$, the expected discounted cost

$$v^\pi(i) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, u_t) \,\Big|\, i_0 = i\Big],$$

is finite. It is well known that the optimal value function, $v(i) = \inf_\pi v^\pi(i)$, satisfies the *dynamic programming (Bellman) equation*:

$$v(i) = \min_{u \in \mathcal{U}(i)} \Big\{c(i, u) + \alpha \sum_{j \in \mathcal{X}} P_{ij}(u) v(j)\Big\}, \quad i \in \mathcal{X}.$$

Moreover, the Markovian policy composed of the minimizers in the Bellman equation is optimal. For every Markovian policy π, the value function associated with this policy satisfies the *policy evaluation equation*

$$v^\pi(i) = c(i, \pi(i)) + \alpha \sum_{j \in \mathcal{X}} P_{ij}(\pi(i))\, v^\pi(j), \quad i \in \mathcal{X}.$$

Viewing v^π as a vector, and defining the vector c^π with coordinates $c^\pi_i = c(i, \pi(i))$, $i \in \mathcal{X}$, and the matrix P^π with entries $P_{ij}(\pi(i))$, $i, j \in \mathcal{X}$, we can compactly write the policy evaluation equation as

$$v^\pi = c^\pi + \alpha P^\pi v^\pi. \tag{1}$$

Our plan is to use a *dynamic risk measure* to evaluate the discounted sequence of costs $Z_t = \alpha^t c(i_t, u_t)$, $t = 0, 1, 2, \dots$. Because of the need to evaluate the risk of the future costs at any time period, a dynamic risk measure (for a finite horizon T) is a sequence of conditional risk measures $\rho_{t,T}(Z_t, \dots, Z_T)$, $t = 0, \dots, T$. The fundamental property of such a nonlinear dynamic risk evaluation is *time consistency*, discussed in various forms in (Cheridito et al., 2006; Artzner et al., 2007; Cheridito and Kupper, 2011). We adopt the definition and the following discussion from (Ruszczyński, 2010): *A dynamic measure of risk is time consistent if for every $t = 0, \dots, T-1$, if $Z_t = V_t$ and $\rho_{t+1,T}(Z_{t+1}, \dots, Z_T) \le \rho_{t+1,T}(V_{t+1}, \dots, V_T)$ a.s., then $\rho_{t,T}(Z_t, \dots, Z_T) \le \rho_{t,T}(V_t, \dots, V_T)$.* Such risk measures, under normalization and translation assumptions, must have a specific recursive form: $\rho_{t,T}(Z_t, \dots, Z_T) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_T)\cdots\big)\Big)$, where each $\rho_t(\cdot)$ is a one-step conditional

risk measure. This result, generalizing the tower property of conditional expectations, is germane for of our approach.

Markov risk measures evaluate the risk of future costs in an MDP under a Markov policy in such a way that the risk is a function of the current state. This, combined with time consistency, translation, monotonicity, and normalization, implies a very specific recursive structure. Denoting by $\rho^{\pi}_{t,T}(i)$ the risk of the system starting from state i at time t, we have

$$\rho^{\pi}_{t,T}(i) = c^{\pi}_i + \alpha\sigma_i\big(P^{\pi}_i, \rho^{\pi}_{t+1,T}(\cdot)\big), \quad i \in \mathcal{X}, \quad t = 0, 1, \ldots, T-1, \tag{2}$$

with $\rho^{\pi}_{T,T}(i) = c^{\pi}_i$, $i \in \mathcal{X}$. In equation (2), the operator $\sigma : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{V} \to \mathbb{R}$, where $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$ and $\mathcal{V}$ is the space of bounded functions on $\mathcal{X}$, is a *transition risk mapping* (a Markovian version of the one-step conditional risk measure). Its first argument is the state i (which we write as a subscript). The second argument, the vector P^{π}_i, is the ith row of the matrix P^{π}: the probability distribution of the state following i under the policy π. The last argument, the function $\rho^{\pi}_{t+1,T}(\cdot)$, is the next state's value: the risk of running the system from the next state in the time interval from $t+1$ to T. A special case is the bilinear form $\sigma_i\big(P^{\pi}_i, \rho^{\pi}_{t+1,T}(\cdot)\big) = \big\langle P^{\pi}_i, \rho^{\pi}_{t+1,T}(\cdot)\big\rangle$ which is the conditional expectation of the next state's value, given the current state is i. It should be stressed that the risk evaluation procedure (2) is not an arbitrary construct, but rather the result of assumptions of normalization, monotonicity, translation, time-consistency, and the Markov property (Fan and Ruszczyński, 2018a). For recent applications of Markov risk measures in control of dynamical systems, see (Majumdar and Pavone, 2020; Sopasakis et al., 2019).

As in (Ruszczyński, 2010), we assume that for each $i \in \mathcal{X}$ and each $P^{\pi}_i \in \mathcal{P}(\mathcal{X})$, the transition risk mapping $\sigma_i(P^{\pi}_i, \cdot)$, understood as a function of its last argument, satisfies the axioms of a coherent measure of risk (Artzner et al., 1999):

Convexity: $\sigma_i(P^{\pi}_i, \alpha v + (1-\alpha)w) \le \alpha\sigma_i(P^{\pi}_i, v) + (1-\alpha)\sigma_i(P^{\pi}_i, w)$, $\forall\, \alpha \in [0,1]$, $\forall\, v, w \in \mathcal{V}$;

Monotonicity: If $v \le w$ (componentwise) then $\sigma_i(P^{\pi}_i, v) \le \sigma_i(P^{\pi}_i, w)$;

Translation equivariance: $\sigma_i(P^{\pi}_i, v + \beta\mathbb{1}) = \sigma_i(P^{\pi}_i, v) + \beta$, for all $\beta \in \mathbb{R}$;

Positive homogeneity: $\sigma_i(P^{\pi}_i, \beta v) = \beta\sigma_i(P^{\pi}_i, v)$, for all $\beta \ge 0$.

Under these conditions, one can pass to the limit with $T \to \infty$ in (2) and prove the existence of an infinite-horizon discounted risk measure:

$$\rho^{\pi}_{0,\infty}(i) = \lim_{T\to\infty} \rho^{\pi}_{0,T}(i), \quad i \in \mathcal{X}.$$

We still denote its value at state i by $v^{\pi}(i)$; it will never lead to misunderstanding. The policy value $v^{\pi}(\cdot)$ satisfies the *risk-averse policy evaluation equation*:

$$v^{\pi}(i) = c^{\pi}(i) + \alpha\sigma_i\big(P^{\pi}_i, v^{\pi}(\cdot)\big), \quad i \in \mathcal{X}.$$

We introduce the space $\mathcal{Q}$ of transition kernels on $\mathcal{X}$, define a vector-valued *transition risk operator* $\sigma : \mathcal{Q} \times \mathcal{V} \to \mathcal{V}$, with components $\sigma_i(P^{\pi}_i, \cdot)$, $i \in \mathcal{X}$, and rewrite the last equation in a way that generalizes (1):

$$v^{\pi} = c^{\pi} + \alpha\sigma(P^{\pi}, v^{\pi}). \tag{3}$$

The only difference between (1) and (3) is that the matrix P^{π} has been replaced by a convex operator σ (which still depends on P^{π}).

Coherent risk measures admit a dual representation (Ruszczyński and Shapiro, 2006a), which in our case can be stated as follows. For every $i \in \mathcal{X}$ a convex, closed and bounded set $A_i(P_i^\pi) \subset \mathcal{P}(\mathcal{X})$ (a subset of the set of probability measures on $\mathcal{X}$) exists, such that

$$\sigma_i(P_i^\pi, v) = \max_{\mu \in A_i(P_i^\pi)} \langle \mu, v \rangle, \quad v \in \mathcal{V}. \tag{4}$$

The scalar product above is the expected value with respect to μ, that is, $\langle \mu, v \rangle = \sum_{j \in \mathcal{X}} \mu_j v_j$. In the risk-neutral case, the set $A_i(P_i^\pi) = \partial \sigma_i(P_i^\pi, 0)$ contains only one element, P_i^π, but in general it is larger and has P_i^π as one of its elements, provided we always have $\sigma_i(P_i^\pi, v) \geq \langle P_i^\pi, v \rangle$. The multifunction $A : \mathcal{X} \to \mathcal{P}(\mathcal{X}) \rightrightarrows \mathcal{P}(\mathcal{X})$ is called the *risk multikernel*. Every $\mu \in A_i(P_i^\pi)$ is absolutely continuous with respect to P_i^π; in the finite-state case considered here, it means that the transitions which are impossible under P_i^π are also impossible under μ.

Example 1 *The mean–semideviation risk measure (Ogryczak and Ruszczyński, 1999) corresponds to the following transition risk mapping:*

$$\sigma_i(P_i^\pi, v) = \sum_{j \in \mathcal{X}} P_{ij}^\pi v_j + \beta \sum_{j \in \mathcal{X}} P_{ij}^\pi \max\left(0, v_j - \sum_{k \in \mathcal{X}} P_{ik}^\pi v_k\right), \quad \beta \in [0, 1].$$

It is coherent (Ruszczyński and Shapiro, 2006a), with the following dual set:

$$A_i(P_i^\pi) = \left\{ \mu \in \mathcal{P}(\mathcal{X}) : \mu_j = P_{ij}^\pi \Big(1 + \xi_j - \sum_{k \in \mathcal{X}} P_{ik}^\pi \xi_k\Big),\ 0 \leq \xi_j \leq \beta,\ j = 1, \ldots, n \right\}.$$

The use of $\beta = 0$ leads to the standard expected value MDP and $\beta \in (0, 1]$ allows for flexibility in modeling risk aversion.

By the dual representation, Markov coherent risk measures are related to *robust dynamic programming* introduced by Nilim and El Ghaoui (2005), but they do not require explicit description of the uncertainty sets for all state-control pairs; the sets are implied. Moreover, the inherent absolute continuity of the hypothetical transition probabilities with respect to the model transition probabilities P_i^π preserves the structure of the problem (the impossible transitions remain impossible).

While equation (3) can be solved by a nonsmooth Newton's method and the resulting evaluation used in a policy iteration method (Ruszczyński, 2010), all these techniques require solving linear equations with the full transition probability matrix and become impractical, when the size of the state space is very large.

An established approach to such a situation in expected value models is to assume that each state $i \in \mathcal{X}$ has a number of relevant *features* $\varphi_j(i) \in \mathbb{R}$, $j = 1, \ldots, m$, where $m \ll n$, and that the value $v^\pi(i)$ of a state can be approximated by a linear combination of its features:

$$v^\pi(i) \approx \widetilde{v}^\pi(i) = \sum_{j=1}^m r_j \varphi_j(i), \quad i \in \mathcal{X}. \tag{5}$$

From now on, we suppress the superscript π, because most of our considerations focus on evaluating a fixed policy. We define the matrix of the features of all states, namely

$$\Phi = \begin{bmatrix} \varphi^\top(1) \\ \varphi^\top(2) \\ \vdots \\ \varphi^\top(n) \end{bmatrix}.$$

Now we can write our approximation as $v \approx \tilde{v} = \Phi r$. Similar to the expected value case, if we attempt to emulate (3) with the approximate value function, we may observe that the right hand side of the equation, $c + \alpha\sigma(P, \Phi r)$, may not be represented as a linear combination of the features. Therefore, we need to project this vector on the subspace spanned by the features, range(Φ). Accordingly, we define a projection operator, $L : \mathcal{V} \to \text{range}(\Phi)$, and formulate the *projected risk-averse approximate dynamic programming equation*:

$$\Phi r = L\big(c + \alpha\sigma(P, \Phi r)\big). \tag{6}$$

It is analogous to the equation considered in the distributionally robust case in (Tamar et al., 2014, sec. 3.1).

Still following the expected value case, we assume that the Markov system under policy π is ergodic, and we denote its vector of stationary probabilities by q. We specify the projection operator using the following scalar product and the associated weighted norm: $\langle v, w\rangle_q = \sum_{i=1}^n q_i v_i w_i$, $\|w\|_q^2 = \langle w, w\rangle_q$. Then the orthogonal projection based on this norm is:

$$L(w) = \operatorname*{argmin}_{z \in \text{range}(\Phi)} ||z - w||_q, \quad w \in \mathcal{V}. \tag{7}$$

The fundamental question is the existence and uniqueness of a solution of equation (6). This can be answered by establishing the contraction mapping property of the right hand side of (6):

$$\mathcal{D}(v) = L\big(c + \alpha\sigma(P, v)\big), \quad v \in \mathcal{V}, \tag{8}$$

which would imply the existence and uniqueness of a solution of the equation

$$v = \mathcal{D}v. \tag{9}$$

Crucial in this context is the *distortion coefficient* of the risk multikernel A:

$$\varkappa = \max\left\{\frac{|m_{ij} - p_{ij}|}{p_{ij}} : m_i \in A_i(P_i^\pi),\ p_{ij} > 0,\ i, j \in \mathcal{X}\right\}.$$

By definition, $\varkappa \geq 0$, with the value 0 corresponding to the risk-neutral model. We also recall that for $p_{ij} = 0$ we always have $m_{ij} = 0$, for all $m_i \in A_i(P_i^\pi)$.

The coefficient $\varkappa$ is related to the risk premium in the transition risk mapping. For illustration, in the mean–semideviation mapping of Example 1, we always have $\varkappa \leq \beta$.

We first prove a technical lemma allowing us to deal with the nonlinearity and nondifferentiability of the transition risk operator σ.

Lemma 1 *The transition risk operator satisfies for all $w, v \in \mathcal{V}$ the inequalities:*

$$\|\sigma(P,w) - \sigma(P,v)\|_q \le \sqrt{1+\varkappa}\,\|w-v\|_q, \tag{10}$$

and

$$\|\sigma(P,w) - \sigma(P,v) - P(w-v)\|_q \le \varkappa\,\|w-v\|_q. \tag{11}$$

Proof For brevity, we omit the argument P of $\sigma(P,\cdot)$, because it is fixed. For every $i = 1,\dots,n$, by the mean value theorem for convex functions (Wegge, 1974; Hiriart-Urruty, 1980), a point $\bar v^{(i)} = (1-\theta_i)v + \theta_i w$ exists, with $\theta_i \in [0,1]$, and a subgradient $m_i \in \partial\sigma_i(\bar v^{(i)})$ exists, such that

$$\sigma_i(w) - \sigma_i(v) = \langle m_i, w - v\rangle.$$

Since the subdifferential $\partial\sigma_i(\cdot) \subseteq A_i$, we have $m_i \in A_i$. Therefore, for a matrix M having m_i, $i = 1,\dots,n$, as its rows,

$$\sigma(w) - \sigma(v) = M(w-v). \tag{12}$$

As each m_i is a probability vector, Jensen's inequality with $h = w - v$, and the equation $q^\top P = q^\top$ yield

$$\begin{aligned}\|Mh\|_q^2 = \sum_{i\in\mathcal{X}} q_i\Big(\sum_{j\in\mathcal{X}} m_{ij}h_j\Big)^2 &\le \sum_{i\in\mathcal{X}} q_i \sum_{j\in\mathcal{X}} m_{ij}h_j^2 \\ &\le (1+\varkappa)\sum_{i\in\mathcal{X}} q_i \sum_{j\in\mathcal{X}} p_{ij}h_j^2 = (1+\varkappa)\sum_{j\in\mathcal{X}} q_j h_j^2 = (1+\varkappa)\|h\|_q^2. \end{aligned}\tag{13}$$

The last two relations imply (10). In a similar way, it follows from (12) that

$$\begin{aligned}\big\|\sigma(P,w) - \sigma(P,v) - P(w-v)\big\|_q^2 = \|(M-P)h\|_q^2 &\le \sum_{i\in\mathcal{X}} q_i\Big(\sum_{j\in\mathcal{X}} |m_{ij}-p_{ij}||h_j|\Big)^2 \\ &\le \varkappa^2 \sum_{i\in\mathcal{X}} q_i\Big(\sum_{j\in\mathcal{X}} p_{ij}|h_j|\Big)^2 \le \varkappa^2 \sum_{i\in\mathcal{X}} q_i \sum_{j\in\mathcal{X}} p_{ij}|h_j|^2 = \varkappa^2\,\|w-v\|_q^2,\end{aligned}$$

which is (11). ∎

We will apply the estimate (10) right away to prove the existence and uniqueness of the solution of the risk-averse equation (9), and the estimate (11) in Theorem 8 on the multistep method.

Theorem 2 *If $\alpha\sqrt{1+\varkappa} < 1$ then the equation (9) has a unique solution v^*.*

Proof We verify that the operator (8) is a contraction mapping in the norm $\|\cdot\|_q$. The orthogonal projection L is nonexpansive. The operator P is nonexpansive in the norm $\|\cdot\|_q$ as well (this is a special case of (13) with $M = P$ and $\varkappa = 0$). The transition risk operator $\sigma(\cdot)$ multiplied by α is a contraction by Lemma 1 and our assumption on α and $\varkappa$. The assertion follows now from Banach's contraction mapping theorem. ∎

If Φ has full column rank, equation (6) has a unique fixed point as well, because only one r satisfies $v^* = \Phi r$.

We may remark here that the requirement that $\varkappa$ should be smaller, when α is larger, is related to the accumulation of risk over time. For α close to 1, many future periods contribute to the risk evaluation. As $\varkappa$ is related to the risk premium at each period, its large values might lead to overpricing of the total risk. In any case, the condition of Theorem 2 is weaker than that of (Tamar et al., 2014, Ass. 2) which requires that $\alpha(1+\varkappa) < 1$ to establish a similar result in the distributionally robust formulation.

3. The Risk-Averse Method of Temporal Differences

We propose to solve (6) by a risk-averse analog of the classical method of temporal differences (Sutton, 1988). We define v^* to be the solution of equation (9) (which exists and is unique, if $\alpha\sqrt{1+\varkappa} < 1$).

Consider the evolution of the system under policy π, resulting in a random trajectory of states i_t, $t = 0, 1, 2 \ldots$. At each time t, we have an approximation r_t of a solution of the equation (6).

The difference between the left and the right hand sides of equation (6) with coefficient values r_t and state i_t defines the *risk-averse temporal difference*:

$$d_t = \varphi^\top(i_t) r_t - c(i_t) - \alpha \sigma_{i_t}(P_{i_t}, \Phi r_t), \quad t = 0, 1, 2, \ldots. \tag{14}$$

Evidently, it cannot be easily computed or observed; this would require the evaluation of the risk $\sigma_{i_t}(P_{i_t}, v)$ and thus consideration of *all* possible transitions from state i_t. Instead, we assume that we can observe a random estimate $\widetilde{\sigma}_{i_t}(P_{i_t}, \cdot)$, such that

$$\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t) = \sigma_{i_t}(P_{i_t}, \Phi r_t) + \xi_t, \quad t = 0, 1, 2, \ldots, \tag{15}$$

with some random errors ξ_t (assumptions about $\{\xi_t\}$ will be specified in section 4). This allows us to define the observed risk-averse temporal differences,

$$\widetilde{d}_t = \varphi^\top(i_t) r_t - c(i_t) - \alpha \widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t), \quad t = 0, 1, 2, \ldots, \tag{16}$$

and to construct the *risk-averse temporal difference method* as follows:

$$r_{t+1} = r_t - \gamma_t \varphi(i_t)\, \widetilde{d}_t, \quad t = 0, 1, 2, \ldots, \tag{17}$$

with stepsizes $\gamma_t > 0$ (assumptions on the sequence $\{\gamma_t\}$ will be specified in section 4).

A related algorithm has been heuristically proposed by Di-Castro Shashua and Mannor (2017) in the context of Q-learning with robust dynamic programming. It uses the set of all possible next states that may follow i_t, in order to estimate the worst distribution from the uncertainty set.

Before proceeding to the detailed proof of convergence of the method (17) in the stochastic case, we analyze its deterministic model, in which the errors ξ_t are ignored and the updates of the sequence $\{r_t\}$ are averaged over all states (with the distribution q). Using the matrix $Q = \mathrm{diag}(q)$, we define the operator:

$$U(r) = \mathbb{E}_{i\sim q}\big[\varphi(i)\big(\varphi^\top(i) r - c(i) - \alpha\sigma_i(P_i, \Phi r)\big)\big] = \Phi^\top Q\big[\Phi r - c - \alpha\, \sigma(P, \Phi r)\big]. \tag{18}$$

The deterministic analog of (16)–(17) reads:

$$\bar{r}_{t+1} = \bar{r}_t - \gamma\, U(r_t), \quad t = 0, 1, 2, \ldots, \quad \gamma > 0. \tag{19}$$

By the definition of the projection operator L, a point r^* is a solution of (6) if and only if

$$r^* = \operatorname*{argmin}_r \frac{1}{2}\big\|\Phi r - \big(c + \alpha\sigma(P, \Phi r^*)\big)\big\|_q^2.$$

This occurs if and only if r^* is a zero of $U(\cdot)$ and thus supports our idea of using the method (16)–(17).

Theorem 3 *If $\alpha\sqrt{1+\varkappa} < 1$, then $\gamma_0 > 0$ exists, such that for all $\gamma \in (0, \gamma_0)$ the algorithm (19) generates a sequence $\{\bar{r}_t\}$ convergent to a point r^* such that $U(r^*) = 0$.*

Proof We shall show that for sufficiently small $\gamma > 0$ the operator $I - \gamma U$ is nonexpansive (a contraction, if Φ has full column rank). For arbitrary r' and r'', we have

$$\begin{aligned}\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 &= \|r' - r''\|^2 \\ &\quad - 2\gamma\langle r' - r'', \Phi^\top Q\Phi(r' - r'')\rangle + 2\gamma\alpha\big\langle r' - r'', \Phi^\top Q\big[\sigma(P, \Phi r') - \sigma(P, \Phi r'')\big]\big\rangle \\ &\quad + \gamma^2\Big\|\Phi^\top Q\Phi(r' - r'') - \alpha\Phi^\top Q\big[\sigma(P, \Phi r') - \sigma(P, \Phi r'')\big]\Big\|^2.\end{aligned}$$

The last term (with γ^2) can be bounded by $\gamma^2 C\|\Phi(r' - r'')\|_q^2$ where C is some constant. Then

$$\begin{aligned}\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 &\le \|r' - r''\|^2 - 2\gamma\|\Phi(r' - r'')\|_q^2 \\ &\quad + 2\gamma\alpha\big\langle\Phi(r' - r''), \sigma(\Phi r') - \sigma(\Phi r'')\big\rangle_q + \gamma^2 C\|\Phi(r' - r'')\|_q^2.\end{aligned}$$

The scalar product can be bounded by (10), and thus

$$\begin{aligned}&\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 \\ &\quad \le \|r' - r''\|^2 - 2\gamma\|\Phi(r' - r'')\|_q^2 + 2\gamma\alpha\sqrt{1+\varkappa}\|\Phi(r' - r'')\|_q^2 + \gamma^2 C\|\Phi(r' - r'')\|_q^2 \\ &\quad = \|r' - r''\|^2 - 2\gamma\left(1 - \alpha\sqrt{1+\varkappa} + \frac{\gamma C}{2}\right)\|\Phi(r' - r'')\|_q^2.\end{aligned}$$

Since $\alpha\sqrt{1+\varkappa} < 1$, then using $0 < \gamma < 2(1 - \alpha\sqrt{1+\varkappa})/C$, we have

$$\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 \le \|r' - r''\|^2 - \gamma\beta\|\Phi(r' - r'')\|_q^2, \tag{20}$$

with some $\beta > 0$. In particular, setting $r' = \bar{r}_t$ and $r'' = r^*$ for a solution r^* of (6), we obtain the following relation between the successive iterates of the method (19):

$$\|\bar{r}_{t+1} - r^*\|^2 \le \|\bar{r}_t - r^*\|^2 - \gamma\beta\|\Phi(\bar{r}_t - r^*)\|_q^2. \tag{21}$$

This immediately proves that the sequence $\{\bar{r}_t\}$ is bounded and $\Phi\bar{r}_t \to \Phi r^*$. Every accumulation point $\hat{r}$ of $\{\bar{r}_t\}$ must be then a solution of equation (6). Substituting this accumulation point for r^* in the last inequality, we conclude that $\bar{r}_t \to \hat{r}$. ■

If Φ has full column rank, the solution r^* is unique, because substituting another solution for $\bar{r}_t$ in (19) we obtain $\bar{r}_{t+1} = \bar{r}_t$, which leads to a contradiction in (21).

4. Convergence of the Risk-Averse Method of Temporal Differences

We shall use the following result on convergence of deterministic nonmonotonic algorithms (Nurminski, 1972).

Theorem 4 *Let $Y^* \subset \mathbb{R}^m$. Suppose $\{r_t\} \subset \mathbb{R}^m$ is a bounded sequence which satisfies the following assumptions:*

A) If a subsequence $\{r_t\}_{t\in\mathcal{K}}$ converges to $r' \in Y^*$, then $\|r_{t+1} - r_t\| \to 0$, as $t \to \infty$, $t \in \mathcal{K}$;

B) If a subsequence $\{r_t\}_{t\in\mathcal{K}}$ converged to $r' \notin Y^*$, then $\varepsilon_0 > 0$ would exist such that for all $\varepsilon \in (0, \varepsilon_0]$ and for all $k \in \mathcal{K}$, the index $s(t,\varepsilon) = \min\{\ell \geq k : \|r_\ell - r_t\| > \varepsilon\}$ would be finite;

C) A continuous function $W : \mathbb{R}^m \to \mathbb{R}$ exists such that if $\{r_t\}_{t\in\mathcal{K}}$ converged to $r' \notin Y^*$ then $\varepsilon_1 > 0$ would exist such that for all $\varepsilon \in (0, \varepsilon_1]$ we would have

$$\limsup_{t\in\mathcal{K}} W(r_{s(t,\varepsilon)}) < W(r'),$$

where $s(t,\varepsilon)$ is defined in B;

D) The set $\{W(r) : r \in Y^*\}$ does not contain any segment of nonzero length.

Then the sequence $\{W(r_t)\}$ is convergent and all limit points of the sequence $\{r_t\}$ belong to Y^*.

The set of solutions of equation (6) is defined as

$$Y^* = \{r \in \mathbb{R}^m : \Phi r = v^*\},$$

where v^* is the unique solution of (9), provided $\alpha\sqrt{1+\varkappa} < 1$. We shall show that sequence generated by the method (17) converges to Y^*, under the above-mentioned condition and some additional conditions on the stepsizes $\{\gamma_t\}$ and errors $\{\xi_t\}$.

We define $\mathcal{F}_t$ to be the σ-algebra generated by $\{i_0, r_0, \ldots, i_t, r_t\}$, $t = 0, 1, \ldots$, and make the following assumptions about the stepsize and error sequences. The stepsizes may be random.

Assumption 1 *The sequence $\{\gamma_t\}$ is adapted to the filtration $\{\mathcal{F}_t\}$ and such that*

(i) $\gamma_t > 0$, $t = 0, 1, \ldots,$ *a.s.;*

(ii) $\sum_{t=0}^{\infty} \gamma_t = \infty$ *a.s.;*

(iii) $\mathbb{E}\sum_{t=0}^{\infty} \gamma_t^2 < \infty$;

(iv) *For any $\varepsilon > 0$,* $\lim\limits_{t_0\to\infty} \sup\limits_{\{T:\sum_{t=t_0}^{T}\gamma_t\leq\varepsilon\}} \sum\limits_{t=t_0}^{T} |\gamma_t - \gamma_{t+1}| = 0$ *a.s..*

Assumption 2 *The sequence of errors $\{\xi_t\}_{t\geq 1}$ satisfies for $t = 0, 1, 2 \ldots$ the conditions*

(i) $\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = 0$ *a.s.;*

(ii) $\mathbb{E}[\|\xi_t\|^2 \,|\, \mathcal{F}_t] \leq C(1 + \|r_t\|^2)$ *a.s., with some constant $C > 0$.*

First, we establish an important implication of the ergodicity of the chain. We write as e_i the ith unit vector in $\mathbb{R}^n$. The following technical result, using the Poisson equation method invented by Métivier and Priouret (1987), will help us to deal with the Markovian dependence of the temporal differences.

Lemma 5 *If the chain $\{i_t\}$ is ergodic with stationary distribution q and Assumption 1 is satisfied, then*

$$\lim_{T\to\infty} \frac{\sum_{t=0}^T \gamma_t(e_{i_t} - q)}{\sum_{t=0}^T \gamma_t} = 0, \quad a.s., \tag{22}$$

and for any $\varepsilon > 0$,

$$\lim_{t_0\to\infty} \sup_{T\geq t_0} \frac{\sum_{t=t_0}^T \gamma_t(e_{i_t} - q)}{\max\left(\varepsilon, \sum_{t=t_0}^T \gamma_t\right)} = 0, \quad a.s.. \tag{23}$$

Proof Due to the ergodicity of the chain, the vectors

$$\nu(i) = \mathbb{E}\left[\sum_{t=0}^\infty (e_{i_t} - q) \,\middle|\, i_0 = i\right], \quad i \in \mathcal{X},$$

are finite and satisfy the *Poisson equation*

$$\nu(i) = e_i - q + \sum_{j\in\mathcal{X}} P_{ij}\nu(j), \quad i \in \mathcal{X}. \tag{24}$$

Consider the sums $\sum_{t=0}^T \gamma_t(e_{i_t} - q)$. By the Poisson equation,

$$e_{i_t} - q = \nu(i_t) - \sum_{j\in\mathcal{X}} P_{i_t j}\nu(j) = \big[\nu(i_t) - \nu(i_{t+1})\big] + \Big[\nu(i_{t+1}) - \sum_{j\in\mathcal{X}} P_{i_t j}\nu(j)\Big]. \tag{25}$$

We consider the two components of the right hand side of (25), marked with brackets, separately. Due to Assumption 1, (i)—(iii), the series

$$\sum_{t=1}^\infty \gamma_t\Big[\nu(i_{t+1}) - \sum_{j\in\mathcal{X}} P_{i_t j}\nu(j)\Big] = \sum_{t=1}^\infty \gamma_t\big(\nu(i_{t+1}) - \mathbb{E}[\nu(i_{t+1}) \,|\, \mathcal{F}_t]\big)$$

is a convergent martingale. Therefore,

$$\lim_{T\to\infty} \frac{\sum_{t=0}^T \gamma_t\big(\nu(i_{t+1}) - \mathbb{E}_t[\nu(i_{t+1})]\big)}{\sum_{t=0}^T \gamma_t} = 0, \quad \text{a.s.}$$

We now focus on the sums

$$\sum_{t=0}^T \gamma_t\big[\nu(i_t) - \nu(i_{t+1})\big] = \gamma_0\nu(i_0) + \sum_{t=1}^T (\gamma_t - \gamma_{t-1})\nu(i_t) - \gamma_T\nu(i_{T+1}).$$

Assumption 1(iv) and (Ruszczyński and Syski, 1983, Lem. A.3) imply (22)–(23). ■

We can now prove the convergence of the stochastic method.

Theorem 6 *Suppose the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)$ satisfy* (15)*, Assumptions 1 and 2 are satisfied, and $\alpha\sqrt{1+\varkappa} < 1$. If the sequence $\{r_t\}$ is bounded with probability 1, then every accumulation point of the sequence $\{r_t\}$ is an element of Y^* with probability 1.*

Proof We use the global Lyapunov function:

$$W(r) = \min_{r^* \in Y^*} \|r - r^*\|^2. \tag{26}$$

The direction used in (17) at step t can be represented as

$$\varphi(i_t)\widetilde{d}_t = U(r_t) + \Delta_t, \tag{27}$$

with the operator $U(\cdot)$ defined in (18), and

$$\Delta_t = -\alpha\xi_t\varphi(i_t) + \Phi^\top \operatorname{diag}\big(e_{i_t} - q\big)\big[\Phi r_t - c - \alpha\sigma(P, \Phi r_t)\big]. \tag{28}$$

Our intention is to verify the conditions of Theorem 4 for almost all paths of the sequence $\{r_t\}$. For this purpose, we estimate the decrease of the function (26) in iteration t. For any $r^* \in Y^*$ we have:

$$\|r_{t+1} - r^*\|^2 = \|r_t - \gamma_t U(r_t) - r^*\|^2 - 2\gamma_t\langle\Delta_t, r_t - \gamma_t U(r_t) - r^*\rangle + \gamma_t^2\|\Delta_t\|^2.$$

The term involving $U(r_t)$ was estimated in the derivation of (21). We obtain the inequality

$$\|r_{t+1} - r^*\|^2 \le \|r_t - r^*\|^2 - 2\gamma_t(1 - \alpha\sqrt{1+\varkappa})\|\Phi(r_t - r^*)\|_q^2 - 2\gamma_t\langle\Delta_t, r_t - \gamma_t U(r_t) - r^*\rangle + C\gamma_t^2. \tag{29}$$

Now we can verify the conditions of Theorem 4 for almost all paths of the sequence $\{r_t\}$. *Condition A.* Due to the boundedness of $\{r_t\}$ the sequence $\{U(r_t)\}$ is bounded as well. In view of (27), it is sufficient to verify that $\gamma_t\xi_t \to 0$. By Assumption 2(i), the sequence

$$S_T = \sum_{t=0}^{T-1} \gamma_t\xi_t, \quad T = 1, 2, \ldots, \tag{30}$$

is a martingale. Consider its quadratic variation process $\big\{\langle S\rangle_T\big\}$ recursively given by

$$\langle S\rangle_{T+1} = \langle S\rangle_T + \mathbb{E}\big[S_{T+1}^2 - S_T^2 \,\big|\, \mathcal{F}_T\big] = \langle S\rangle_T + \gamma_t^2\mathbb{E}\big[\|\xi_t\|^2 \,\big|\, \mathcal{F}_t\big], \quad T = 1, 2, \ldots.$$

Due to Assumption 2(ii), $\langle S\rangle_{T+1} \le C\sum_{t=0}^T \gamma_t^2(1 + \|r_t\|^2)$. By virtue of Assumption 1(iii) and the martingale convergence theorem, the series $\sum_{t=0}^\infty \gamma_t^2$ is convergent a.s.. By the boundedness of $\{r_t\}$, $\lim_{T\to\infty}\big\{\langle S\rangle_T\big\} < \infty$ with probability 1. This implies that the martingale $\{S_T\}$ is convergent a.s. (see, *e. g.*, Neveu (1975)), which yields $\lim_{t\to\infty}\gamma_t\xi_t = 0$ a.s.. *Condition B.* Suppose $r_k \to r' \notin Y^*$ for $k \in \mathcal{K}$ (on a certain path ω). If B were false, then for all $\varepsilon_0 > 0$ we could find $\varepsilon \in (0, \varepsilon_0]$ and $k \in \mathcal{K}$ such that $\|r_t - r_k\| \le \varepsilon$ for all $t \ge k$. Then for all $k_0 \in \mathcal{K}$, $k_0 \ge k$, we have $\|r_t - r_{k_0}\| \le 2\varepsilon$ for all $t \ge k_0$. Since r' is not optimal, we can choose $\varepsilon_0 > 0$ small enough, $k_0 \in \mathcal{K}$ large enough, and $\delta > 0$ small enough, so that $\|\Phi(r_t - r^*)\|_q^2 > \delta$ for all $t \ge k_0$. Then (29) for $T \ge k_0$ yields

$$\begin{aligned}\|r_T - r^*\|^2 &\le \|r_{k_0} - r^*\|^2 \\ &+ \left(-\delta(1 - \alpha\sqrt{1+\varkappa}) + \frac{\sum_{t=k_0}^{T-1}\gamma_t\langle\Delta_t, r_t - \gamma_t U(r_t) - r^*\rangle}{\sum_{t=k_0}^{T-1}\gamma_t} + C\frac{\sum_{t=k_0}^{T-1}\gamma_t^2}{\sum_{t=k_0}^{T-1}\gamma_t}\right)\sum_{t=k_0}^{T-1}\gamma_t.\end{aligned} \tag{31}$$

We fix $r^* = \text{Proj}_{Y^*}(r_{k_0})$ and estimate the growth of the sums involving Δ_t. We write $\Delta_t = \Delta_t^{(1)} + \Delta_t^{(2)}$, where, in view of (28),

$$\Delta_t^{(1)} = -\alpha\xi_t\varphi(i_t), \quad \Delta_t^{(2)} = \Phi^\top \text{diag}\,\big(e_{i_t} - q\big)\big[\Phi r_t - c - \alpha\sigma(P, \Phi r_t)\big].$$

Since (30) is a convergent martingale and the terms $\langle \varphi(i_t), r_t - \gamma_t U(r_t) - r^*\rangle$ are bounded and $\mathcal{F}_t$-measurable, we have

$$\lim_{T\to\infty} \frac{\left|\sum_{t=k_0}^{T-1} \gamma_t \langle \Delta_t^{(1)}, r_t - \gamma_t U(r_t) - r^*\rangle\right|}{\sum_{t=k_0}^{T-1}\gamma_t} = 0.$$

To deal with the sum involving $\Delta_t^{(2)}$, observe that $\|r_t - r_{k_0}\| \le 2\varepsilon_0$ and that we can choose k_0 large enough so that $\gamma_t \le O(\varepsilon_0)$. Then

$$\begin{aligned}\langle \Delta_t^{(2)}, r_t - \gamma_t U(r_t) - r^*\rangle = \\ \Big\langle \text{diag}\,\big(e_{i_t} - q\big)\big[\Phi r_{k_0} - c - \alpha\sigma(P, \Phi r_{k_0})\big], \Phi(r_{k_0} - r^*)\Big\rangle + h_t = \langle e_{i_t} - q, w\rangle + h_t, \quad (32)\end{aligned}$$

where $|h_t| \le C\varepsilon_0$ and w is a fixed vector (depending on k_0 only). It follows that

$$\left|\sum_{t=k_0}^{T-1} \gamma_t \langle \Delta_t^{(2)}, r_t - \gamma_t U(r_t) - r^*\rangle\right| \le C\left\|\sum_{t=k_0}^{T-1}\gamma_t(e_{i_t} - q)\right\| + C\varepsilon_0\sum_{t=k_0}^{T-1}\gamma_t. \quad (33)$$

Dividing both sides of (33) by $\sum_{t=k_0}^{T-1}\gamma_t$ and using (22), we see that we can choose $\varepsilon > 0$ small enough and $k_0 \in \mathcal{K}$ large enough, so that the entire expression in parentheses in (31) is smaller than $-\delta(1 - \alpha\sqrt{1+\varkappa})/2$, if T is large enough. But this yields $\|r_T - r^*\| \to -\infty$, as $T \to \infty$, a contradiction. Therefore, Condition B is satisfied.

Condition C. The inequality (31) remains valid for $T = s(k_0, \varepsilon)$. By the definition of $s(k_0, \varepsilon)$,

$$\Big\|\sum_{t=k_0}^{T-1}\gamma_t\varphi(i_t)(d_t + \xi_t)\Big\| \ge \varepsilon.$$

By the convergence of (30), and the boundedness of $\{d_t\}$, a constant $C > 0$ exists such that for all sufficiently large k_0 and sufficiently small ε, we have

$$\sum_{t=k_0}^{T-1}\gamma_t \ge \varepsilon/C.$$

Using (23), by a similar argument as in the analysis of Condition B, we can choose $\varepsilon_1 > 0$ small enough so that for all $k_0 \in \mathcal{K}$ large enough, the entire expression in parentheses in (31) is smaller than $-\delta(1 - \alpha\sqrt{1+\varkappa})/2$. Therefore, for all $\varepsilon \in (0, \varepsilon_1]$ and all sufficiently large $k_0 \in \mathcal{K}$

$$\|r_{s(k_0,\varepsilon)} - r^*\|^2 \le \|r_{k_0} - r^*\|^2 - \frac{\delta(1 - \alpha\sqrt{1+\varkappa})\varepsilon}{2C}.$$

We fix $r^* = \mathrm{Proj}_{Y^*}(r_{k_0})$ on the right hand side, and obtain for the merit function (26) the inequality

$$W(r_{s(k_0,\varepsilon)}) \le \|r_{s(k_0,\varepsilon)} - r^*\|^2 \le W(r_{k_0}) - \frac{\delta(1-\alpha\sqrt{1+\varkappa})\varepsilon}{2C}.$$

Now, the limit with respect to $k_0 \to \infty$, $k_0 \in \mathcal{K}$, proves Condition C.
Condition D is satisfied trivially, because $W(r^*) \equiv 0$ for $r^* \in Y^*$. ■

The only question remaining is the boundedness of the sequence $\{r_t\}$. It is a common issue in the analysis of stochastic approximation algorithms (Kushner and Yin, 2003, §5.1). In our case, no additional conditions and analysis are needed, because our Lyapunov function (26) is the squared distance to the optimal set. Therefore, a simple algorithmic modification: the projection on a bounded set Y intersecting with $\{r \in \mathbb{R}^m : \Phi r = v^*\}$, is sufficient to guarantee boundedness. The modified method (17) reads:

$$r_{t+1} = \mathrm{Proj}_Y\big(r_t - \gamma_t \varphi(i_t)\,\widetilde{d}_t\big), \quad t = 0, 1, 2, \ldots. \tag{34}$$

Now, $Y^* = \{r \in Y : \Phi r = v^*\}$ and we require that this set is nonempty. This modification does not affect our analysis in any meaningful way, because the projection is nonexpansive. In the proof of Theorem 3, we use the inequality

$$\big\|\mathrm{Proj}_Y(r' - \gamma U(r')) - \mathrm{Proj}_Y(r'' - \gamma U(r''))\big\|^2 \le \big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2$$

and proceed as before. In the proof of Theorem 6, we start from

$$\|r_{t+1} - r^*\|^2 = \big\|\mathrm{Proj}_Y\big(r_t - \gamma_t(U(r_t) + \Delta_t)\big) - r^*\big\|^2 \le \big\|r_t - \gamma_t(U(r_t) + \Delta_t) - r^*\big\|^2,$$

and then continue in the same way as before. We did not include projection into the method originally, because it obscures the presentation. In practice, we have not yet encountered any need for it.

If the stepsizes $\{\gamma_t\}$ are deterministic, we can guarantee the boudedness without using projections, by employing the ODE method and the scaling technique developed by Borkar and Meyn (2000).

Theorem 7 *Suppose the stepsizes $\{\gamma_t\}$ are deterministic and satisfy the deterministic version of Assumption 1, the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)$ satisfy* (15) *and Assumption 2, the matrix Φ has full column rank, and $\alpha\sqrt{1+\varkappa} < 1$. Then the sequence $\{r_t\}$ is bounded with probability 1.*

Proof We apply (Borkar and Meyn, 2000, Thm. 2.1) (a closely related result in (Kushner and Yin, 2003, Thm. 5.4.1) could also be used). The ODE corresponding to (17) has the form $\dot{r}(t) = -U(r(t))$. Due to the positive homogeneity of $\sigma(P, \cdot)$, the asymptotic "fluid" operator has the form

$$U_\infty(r) = \lim_{M\to\infty} \frac{U(Mr)}{M} = \Phi^\top Q\big[\Phi r - \alpha\,\sigma(P, \Phi r)\big].$$

It is Lipschitz continuous, by Lemma 1. The rescaled ODE is $\dot{r}(t) = -U_\infty(r(t))$. The origin is its only equilibrium, and it is globally stable. This can be established by using

the Lyapunov function $L_\infty(r) = \frac{1}{2}\|r\|^2$. Indeed, applying (10) with $w = \Phi r$ and $v = 0$, we obtain

$$\begin{aligned}\dot{L}_\infty(r) &= -\langle r, U_\infty(r)\rangle = -\big\langle \Phi r, \Phi r - \alpha\,\sigma(P, \Phi r)\big\rangle_q \\ &= -\big\|\Phi r\big\|_q^2 + \alpha\big\|\Phi r\big\|_q\big\|\sigma(P,\Phi r)\big\|_q \le \big(-1 + \alpha\sqrt{1+\varkappa}\big)\big\|\Phi r\big\|_q^2 < 0,\end{aligned}$$

unless $r = 0$. The convergence of interpolated trajectories to the solutions of the corresponding ODE can be established as in (Borkar and Meyn, 2000, Thm. 2.1), with the additional use of our Lemma 5. This part of the analysis repeats much of the proof of Theorem 6, where we deal with the errors Δ_t. ■

5. The Multistep Risk-Averse Method of Temporal Differences

In the method discussed so far, the approximation coefficients $\{r_t\}$ are corrected by moving in the direction of the last observed feature vector $\varphi(i_t)$. Alternatively, we may use the weighted averages of all previous feature observations, where the highest weight is given to the most recent observation and the weights decrease exponentially as we look into the past. This is the core idea of the well-known TD(λ) algorithm (Sutton, 1988). We generalize it to the risk-averse case.

For a fixed policy π, we refer to v^π as v, and to P^π as P, for simplicity. The multistep risk-averse method of temporal differences carries out the following iterations:

$$z_t = \lambda\alpha z_{t-1} + \varphi(i_t), \quad t = 0, 1, 2, \dots, \tag{35}$$

$$r_{t+1} = r_t - \gamma_t z_t \widetilde{d}_t, \quad t = 0, 1, 2, \dots \tag{36}$$

where $\lambda \in [0, 1]$, and $\widetilde{d}_t$ is given by (16). For simplicity, z_{-1} is assumed to be the zero vector. In the risk-neutral case, when $\sigma_{i_t}(P_{i_t}, \Phi r_t) = P_{i_t}\Phi r_t$, the method reduces to the classical TD(λ).

Our convergence analysis will use some ideas from the analysis in the previous two sections, albeit in a form adapted to the version with exponentially averaged features. However, contrary to the expected value setting, the method (35)–(36) will converge to a solution of an equation different from (9), but still relevant for our problem.

We start from a heuristic analysis of a deterministic counterpart of the method, to extract its drift. In the next section, we make all approximations precise, but we believe that this introduction is useful to decipher our detailed approach to follow. By direct calculation,

$$z_t = \sum_{k=0}^{t} (\lambda\alpha)^{t-k}\varphi(i_k), \tag{37}$$

and thus

$$z_t d_t = \Phi^\top \sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top \big(\Phi r_t - c - \alpha\sigma(P, \Phi r_t)\big).$$

Heuristically assuming that $r_t \approx r'$, we focus on the operator acting on the expected temporal differences. As each of the observed feature vectors $\varphi(i_k)$ affects all succeeding steps

of the method, via the filter (35), we need to study the cumulative effect of many steps. We look, therefore, at the sums

$$G_T = \mathbb{E}\Big[\sum_{t=0}^{T}\gamma_t\sum_{k=0}^{t}(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top\Big].$$

Changing the order of summation and using the fact that $\{(\lambda\alpha)^{t-k}\}_{t\geq k}$ diminishes very fast, as compared to $\{\gamma_t\}_{t\geq k}$, we get

$$G_T = \mathbb{E}\Big[\sum_{k=0}^{T}\sum_{t=k}^{T}\gamma_t(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top\Big] \approx \mathbb{E}\Big[\sum_{k=0}^{T}\gamma_k\sum_{t=k}^{T}(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top\Big].$$

Therefore

$$\begin{aligned}
G_T &\approx \mathbb{E}\Big[\sum_{k=0}^{T}\gamma_k\sum_{t=k}^{T}(\lambda\alpha)^{t-k}e_{i_k}\mathbb{E}\big[e_{i_t}^\top \,\big|\, \mathcal{F}_k\big]\Big] = \mathbb{E}\Big[\sum_{k=0}^{T}\gamma_k\sum_{t=k}^{T}(\lambda\alpha)^{t-k}e_{i_k}e_{i_k}^\top P^{t-k}\Big]\\
&= \sum_{k=0}^{T}\gamma_k\mathbb{E}[\mathrm{diag}(e_{i_k})]\sum_{t=k}^{T}(\lambda\alpha)^{t-k}P^{t-k} \approx \sum_{k=0}^{T}\gamma_k\mathbb{E}[\mathrm{diag}(e_{i_k})]\sum_{t=k}^{\infty}(\lambda\alpha)^{t-k}P^{t-k}\\
&\approx Q\sum_{k=0}^{T}\gamma_k\sum_{t=k}^{\infty}(\lambda\alpha)^{t-k}P^{t-k}.
\end{aligned}$$

The last approximations are possible because $\lambda\alpha \in [0,1)$ and $\mathbb{E}\big[\,\mathrm{diag}(e_{i_k})\big] \to q$ at an exponential rate.

We now define the multistep transition matrix,

$$\overline{P} = (1-\lambda\alpha)\sum_{\ell=0}^{\infty}(\lambda\alpha)^\ell P^\ell. \tag{38}$$

By construction, $\overline{P} \in \overline{\mathrm{conv}}\{I, P, P^2, \dots\}$. With these approximations, we can simply write

$$G_T \approx \frac{1}{1-\lambda\alpha}Q\overline{P}\sum_{k=0}^{T}\gamma_k.$$

Define the operator

$$\overline{U}(r) = \Phi^\top Q\overline{P}\big[\Phi r - c - \alpha\,\sigma(P,\Phi r)\big], \tag{39}$$

and consider the following deterministic counterpart of (35)–(36), with $\overline{\gamma} \sim \gamma_t/(1-\lambda\alpha)$:

$$r_{r+1} = r_t - \overline{\gamma}\,\overline{U}(r_t), \quad t = 0,1,2,\dots, \quad \overline{\gamma} > 0. \tag{40}$$

Our intention is to show that for sufficiently small $\overline{\gamma}$ the method (40) converges to a point r^* such that $\overline{U}(r^*) = 0$. Such a point is also a solution of the following *projected multistep risk-averse dynamic programming equation*:

$$L\overline{P}\Phi r = L\overline{P}\big(c + \alpha\sigma(P,\Phi r)\big), \tag{41}$$

where L is the projection operator defined in (7). The solutions of (41) differ from the solutions of (6), unlike in the risk-neutral case ($\varkappa = 0$). If we replace $\overline{P}$ with I, (41) reduces to (6).

Theorem 8 *If Φ has full column rank and $\alpha(1+\varkappa)<1$, then $\overline{\gamma}_0>0$ exists, such that for all $\overline{\gamma}\in(0,\overline{\gamma}_0)$ the algorithm (40) generates a sequence $\{r_t\}$ convergent to a unique solution r^* of the equation $\overline{U}(r^*)=0$.*

Proof We show that for sufficiently small $\overline{\gamma}>0$ the operator $I-\overline{\gamma}\,\overline{U}$ is a contraction mapping. For two arbitrary points r' and r'' we have

$$\begin{aligned}\Big\|\big(r'-\overline{\gamma}\,\overline{U}(r')\big)-\big(r''-\overline{\gamma}\,\overline{U}(r'')\big)\Big\|^2 \\ = \|r'-r''\|^2 \;+2\overline{\gamma}\Big\langle r'-r'',\Phi^\top Q\overline{P}\big[-\Phi(r'-r'')+\alpha\sigma(P,\Phi r')-\alpha\sigma(P,\Phi r'')\big]\Big\rangle \\ +\overline{\gamma}^2\Big\|\Phi^\top Q\overline{P}\Phi(r'-r'')-\alpha\Phi^\top Q\overline{P}\big[\sigma(P,\Phi r')-\sigma(P,\Phi r'')\big]\Big\|^2. \quad (42)\end{aligned}$$

We focus on the scalar product in the middle of the right hand side of (42):

$$\begin{aligned}\Big\langle\Phi(r'-r''),\overline{P}\big[-\Phi(r'-r'')+\alpha\sigma(P,\Phi r')-\alpha\sigma(P,\Phi r'')\big]\Big\rangle_q \\ =\Big\langle\Phi(r'-r''),\overline{P}\big[-\Phi(r'-r'')+\alpha P\Phi(r'-r'')\big]\Big\rangle_q \\ +\alpha\Big\langle\Phi(r'-r''),\overline{P}\big[\sigma(P,\Phi r')-\sigma(P,\Phi r'')-P\Phi(r'-r'')\big]\Big\rangle_q. \quad (43)\end{aligned}$$

Setting $h=\Phi(r'-r'')$, we can estimate the first (quadratic) term on the right hand side of (43) by a calculation borrowed from (Tsitsiklis and Van Roy, 1997, Lem. 8):

$$\begin{aligned}\Big\langle h,\overline{P}\big[-h+\alpha Ph\big]\Big\rangle_q &= (1-\alpha\lambda)\Big\langle h,\sum_{\ell=0}^{\infty}(\alpha\lambda)^\ell P^\ell\big[-h+\alpha Ph\big]\Big\rangle_q \\ &= (1-\alpha\lambda)(1-\lambda)\Big\langle h,\sum_{k=0}^{\infty}\lambda^k\sum_{\ell=0}^{k}\alpha^\ell P^\ell\big[-h+\alpha Ph\big]\Big\rangle_q \\ &= (1-\alpha\lambda)(1-\lambda)\Big\langle h,\sum_{k=0}^{\infty}\lambda^k\big[\alpha^{k+1}P^{k+1}h-h\big]\Big\rangle_q \\ &= (1-\alpha\lambda)\Big\langle h,(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\alpha^{k+1}P^{k+1}h-h\Big\rangle_q \\ &= (1-\alpha\lambda)\Big\langle h,\frac{\alpha(1-\lambda)}{1-\alpha\lambda}P\overline{P}h-h\Big\rangle_q \le(\alpha-1)\|h\|_q^2.\end{aligned}$$

The last inequality is due to the fact that both P and $\overline{P}$ are nonexpansive in $\|\cdot\|_q$.

The second (nonsmooth) term on the right hand side of (43) can be estimated by (11), again with the use of the nonexpansiveness of $\overline{P}$:

$$\Big\langle\Phi(r'-r''),\overline{P}\big[\sigma(P,\Phi r')-\sigma(P,\Phi r'')-P\Phi(r'-r'')\big]\Big\rangle_q\le\varkappa\big\|\Phi(r'-r'')\big\|_q^2.$$

The last term on the right hand side of (42) (with γ^2) can be bounded by $\gamma^2\overline{C}\|\Phi(r'-r'')\|_q^2$, where $\overline{C}$ is some constant. Integrating all these estimates into (42), we obtain the inequality

$$\big\|(I-\overline{\gamma}\,\overline{U})(r')-(I-\overline{\gamma}\,\overline{U})(r'')\big\|^2\le\|r'-r''\|^2-2\overline{\gamma}\Big(1-\alpha(1+\varkappa)-\frac{\overline{\gamma}\,\overline{C}}{2}\Big)\big\|\Phi(r'-r'')\big\|_q^2.$$

If $\alpha(1+\varkappa)<1$, then using $0<\overline{\gamma}<2(1-\alpha(1+\varkappa))/\overline{C}$, we obtain:

$$\left\|(I-\overline{\gamma}\,\overline{U})(r')-(I-\overline{\gamma}\,\overline{U})(r'')\right\|^2 \le \|r'-r''\|^2-\overline{\gamma}\beta\|\Phi(r'-r'')\|_q^2, \tag{44}$$

with some $\beta>0$. Since the columns of Φ are linearly independent, the operator $I-\overline{\gamma}\overline{U}$ is a contraction mapping. This implies the assertion of the theorem. In particular, setting $r'=\bar{r}_t$ and $r''=r^*$ for a solution r^* of (41), we obtain the following relation between successive iterates of the method (40):

$$\|r_{t+1}-r^*\|^2 \le \|r_t-r^*\|^2-\overline{\gamma}\beta\|\Phi(r_t-r^*)\|_q^2. \tag{45}$$

■

6. Convergence of the Risk-Averse Multistep Method

We now carry out a detailed analysis of the stochastic method (35)–(36).

Lemma 9 *For any array of uniformly bounded random variables* $\{A_{k,t}\}_{k\ge0,\,t\ge0}$

$$\lim_{T\to\infty}\frac{\sum_{k=0}^T\sum_{t=k}^T\gamma_t(\lambda\alpha)^{t-k}A_{k,t}-\sum_{k=0}^T\gamma_k\sum_{t=k}^\infty(\lambda\alpha)^{t-k}A_{k,t}}{\sum_{k=0}^T\gamma_k}=0 \quad a.s..$$

Proof Changing the order of summation twice, we obtain

$$\begin{aligned}
&\sum_{k=0}^T\sum_{t=k+1}^T|\gamma_t-\gamma_k|(\lambda\alpha)^{t-k}\le\sum_{k=0}^T\sum_{t=k+1}^T\sum_{\ell=k+1}^t|\gamma_\ell-\gamma_{\ell-1}|(\lambda\alpha)^{t-k}\\
&\le\frac{1}{1-\lambda\alpha}\sum_{k=0}^T\sum_{\ell=k+1}^T|\gamma_\ell-\gamma_{\ell-1}|(\lambda\alpha)^{\ell-k}=\frac{1}{1-\lambda\alpha}\sum_{\ell=1}^T|\gamma_\ell-\gamma_{\ell-1}|\sum_{k=0}^{\ell-1}(\lambda\alpha)^{\ell-k}\\
&\le\frac{\lambda\alpha}{(1-\lambda\alpha)^2}\sum_{\ell=1}^T|\gamma_\ell-\gamma_{\ell-1}|.
\end{aligned}$$

Therefore, with C being the uniform bound on $\|A_{k,t}\|$ and $\gamma_k^{\max}=\max_{t\ge k}\gamma_t$, we obtain

$$\begin{aligned}
&\left\|\sum_{k=0}^T\sum_{t=k}^T\gamma_t(\lambda\alpha)^{t-k}A_{k,t}-\sum_{k=0}^T\gamma_k\sum_{t=k}^\infty(\lambda\alpha)^{t-k}A_{k,t}\right\|\\
&\le C\sum_{k=0}^T\sum_{t=k+1}^T|\gamma_t-\gamma_k|(\lambda\alpha)^{t-k}+C\sum_{k=0}^T\sum_{t=T+1}^\infty\gamma_t(\lambda\alpha)^{t-k}\\
&\le\frac{C\lambda\alpha}{(1-\lambda\alpha)^2}\sum_{\ell=1}^T|\gamma_\ell-\gamma_{\ell-1}|+\frac{C\gamma_{T+1}^{\max}\lambda\alpha}{(1-\lambda\alpha)^2}.
\end{aligned}$$

Assumption 1(iv) and (Ruszczyński and Syski, 1983, Lem. A.3) imply the assertion. ■

We need another auxiliary result, extending Lemma 5 to our case.

Lemma 10

$$\lim_{T\to\infty} \frac{\sum_{t=0}^T \gamma_t \Big(\sum_{k=0}^t (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top - \frac{1}{1-\lambda\alpha} Q\overline{P} \Big)}{\sum_{t=0}^T \gamma_t} = 0 \quad a.s., \tag{46}$$

and for any $\varepsilon > 0$,

$$\lim_{t_0\to\infty} \sup_{T\geq t_0} \frac{\sum_{t=t_0}^T \gamma_t \Big(\sum_{k=0}^t (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top - \frac{1}{1-\lambda\alpha} Q\overline{P} \Big)}{\max\Big(\varepsilon, \sum_{t=t_0}^T \gamma_t\Big)} = 0 \quad a.s.. \tag{47}$$

Proof Consider the sums appearing in the numerators of (46) and (47):

$$\sum_{t=t_0}^T \gamma_t \sum_{k=0}^t (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top = \sum_{k=t_0}^T e_{i_k} \sum_{t=k}^T (\lambda\alpha)^{t-k} \gamma_t e_{i_t}^\top, \quad T = 1, 2, \ldots.$$

In view of Lemma 9, it is sufficient to consider the sums

$$S_{t_0,T} = \sum_{k=t_0}^T \gamma_k \, e_{i_k} \sum_{t=k}^\infty (\lambda\alpha)^{t-k} e_{i_t}^\top, \quad T = 1, 2, \ldots.$$

We transform the inner sum and change the order of summation:

$$\begin{aligned}\sum_{t=k}^\infty (\lambda\alpha)^{t-k} e_{i_t}^\top &= \sum_{t=k}^\infty (\lambda\alpha)^{t-k} \Big\{ \sum_{\ell=k+1}^t \big[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \big] + e_{i_k}^\top P^{t-k} \Big\} \\ &= \sum_{t=k}^\infty (\lambda\alpha)^{t-k} e_{i_k}^\top P^{t-k} + \sum_{\ell=k+1}^\infty \sum_{t=\ell}^\infty (\lambda\alpha)^{t-k} \big[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \big].\end{aligned}$$

We can thus write $S_{t_0,T} = S^{(1)}_{t_0,T} + S^{(2)}_{t_0,T}$, where, using (38), we have

$$S^{(1)}_{t_0,T} = \sum_{k=t_0}^T \gamma_k \, e_{i_k} e_{i_k}^\top \sum_{t=k}^\infty (\lambda\alpha)^{t-k} P^{t-k} = \frac{1}{1-\lambda\alpha} \sum_{k=t_0}^T \gamma_k \, \mathrm{diag}(e_{i_k}) \overline{P},$$

and

$$\begin{aligned} S^{(2)}_{t_0,T} &= \sum_{k=t_0}^T \gamma_k \, e_{i_k} \sum_{\ell=k+1}^\infty \sum_{t=\ell}^\infty (\lambda\alpha)^{t-k} \big[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \big] \\ &= \frac{1}{1-\lambda\alpha} \sum_{k=t_0}^T \gamma_k \, e_{i_k} \sum_{\ell=k+1}^\infty (\lambda\alpha)^{\ell-k} \big[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P \big] \overline{P} \\ &= \frac{1}{1-\lambda\alpha} \sum_{k=t_0}^T e_{i_k} \sum_{\ell=k+1}^\infty \gamma_{\ell-1} \, (\lambda\alpha)^{\ell-k} \big[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P \big] \overline{P} \\ &\quad + \frac{1}{1-\lambda\alpha} \sum_{k=t_0}^T e_{i_k} \sum_{\ell=k+1}^\infty (\lambda\alpha)^{\ell-k} (\gamma_k - \gamma_{\ell-1}) \big[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P \big] \overline{P} = S^{(2,1)}_{t_0,T} + S^{(2,2)}_{t_0,T}. \end{aligned} \tag{48}$$

Both double sums in (48) will be analyzed separately. Since $\mathbb{E}\big[e_{i\ell}^\top \,\big|\, \mathcal{F}_{\ell-1}\big] = e_{i_{\ell-1}}^\top P$, the first one, after changing the order of summation, becomes a martingale adapted to $\{\mathcal{F}_\ell\}$:

$$\begin{aligned}\lim_{T\to\infty} S_{t_0,T}^{(2,1)} &= \frac{1}{1-\lambda\alpha}\sum_{k=t_0}^{\infty} e_{i_k} \sum_{\ell=k+1}^{\infty} \gamma_{\ell-1}\,(\lambda\alpha)^{\ell-k}\big[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P\big]\overline{P} \\ &= \frac{1}{1-\lambda\alpha}\sum_{\ell=t_0+1}^{\infty} \gamma_{\ell-1}\Big(\sum_{k=0}^{\ell-1} e_{i_k}(\lambda\alpha)^{\ell-k}\Big)\big[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P\big]\overline{P}.\end{aligned}$$

Due to Assumption 1(iii), the limit is finite a.s. and thus

$$\lim_{T\to\infty} \frac{S_{0,T}^{(2,1)}}{\sum_{t=0}^T \gamma_t} = 0 \quad \text{a.s.} \quad \text{and} \quad \lim_{t_0\to\infty} \sup_{T\ge t_0} \frac{S_{t_0,T}^{(2,1)}}{\max\Big(\varepsilon, \sum_{t=t_0}^T \gamma_t\Big)} = 0 \quad \text{a.s..} \tag{49}$$

We now estimate the second double sum on the right hand side of (48). A constant C exists, such that uniformly for all T and with probability one,

$$\begin{aligned}\big\|S_{t_0,T}^{(2,2)}\big\| &\le C\sum_{k=t_0}^{T}\sum_{\ell=k+2}^{\infty}(\lambda\alpha)^{\ell-k}|\gamma_k-\gamma_{\ell-1}| \\ &\le C\sum_{k=t_0}^{T}\sum_{\ell=k}^{k+T-1}(\lambda\alpha)^{\ell-k+2}|\gamma_k-\gamma_{\ell+1}| + C\sum_{k=t_0}^{T}\gamma_k\sum_{\ell=k+T}^{\infty}(\lambda\alpha)^{\ell-k+2} \\ &\le C\sum_{k=t_0}^{T}\sum_{\ell=k}^{k+T-1}(\lambda\alpha)^{\ell-k+2}\sum_{j=k}^{l}|\gamma_j-\gamma_{j+1}| + \frac{C(\lambda\alpha)^{T+2}}{1-\lambda\alpha}\sum_{k=t_0}^{T}\gamma_k \\ &\le \frac{C}{1-\lambda\alpha}\sum_{k=t_0}^{T}\sum_{j=k}^{k+T-1}|\gamma_j-\gamma_{j+1}|(\lambda\alpha)^{j-k+2} + \frac{C(\lambda\alpha)^{T+2}}{1-\lambda\alpha}\sum_{k=t_0}^{T}\gamma_k \\ &\le \frac{C}{1-\lambda\alpha}\sum_{j=t_0}^{2T-1}\sum_{k=t_0}^{j}|\gamma_j-\gamma_{j+1}|(\lambda\alpha)^{j-k+2} + \frac{C(\lambda\alpha)^{T+2}}{1-\lambda\alpha}\sum_{k=t_0}^{T}\gamma_k \\ &\le \frac{C(\lambda\alpha)^2}{(1-\lambda\alpha)^2}\sum_{j=t_0}^{2T-1}|\gamma_j-\gamma_{j+1}| + \frac{C(\lambda\alpha)^{T+2}}{1-\lambda\alpha}\sum_{k=t_0}^{T}\gamma_k.\end{aligned}$$

In view of Assumption 1(iv), we conclude that the relations (49) are true for $S_{t_0,T}^{(2,2)}$ as well. It follows that $S_{t_0,T}^{(2)}$ in (48) is negligible as compared to $\sum_{t=t_0}^T \gamma_t$. We can thus apply Lemma 5 to $S_{t_0,T}^{(1)} - \frac{1}{1-\lambda\alpha}\sum_{k=t_0}^T \gamma_k \operatorname{diag}(q)\overline{P}$, and obtain both assertions. ■

Now we can follow the arguments of section 4 and establish the convergence of the stochastic multistep method.

Theorem 11 *Assume that $\alpha(1+\varkappa)<1$, the sequence $\{r_t\}$ is bounded with probability 1, and the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)$ satisfy* (15)*. Then, with probability 1, every accumulation point of the sequence $\{r_t\}$ generated by* (35)–(36) *is a solution of* (41)*.*

Proof We represent the direction used in (36) at step t as

$$z_t\widetilde{d}_t = \frac{1}{1-\lambda\alpha}\overline{U}_t(r_t) + \Delta_t^{(1)} + \Delta_t^{(2)},$$

with the operator $\overline{U}_t(\cdot)$ defined in (39), and

$$\begin{aligned}\Delta_t^{(1)} &= -\alpha z_t \xi_t,\\ \Delta_t^{(2)} &= z_t d_t - \frac{1}{1-\lambda\alpha}\overline{U}_t(r_t).\end{aligned}$$

For any r^* solving (41), with $\overline{\gamma}_t = \gamma_t/(1-\lambda\alpha)$, we have

$$\big\|r_{t+1}-r^*\big\|^2 = \big\|r_t-\overline{\gamma}_t\overline{U}_t(r_t)-r^*\big\|^2 - 2\overline{\gamma}_t\langle\Delta_t^{(1)}+\Delta_t^{(2)}, r_t-\overline{\gamma}_t\overline{U}_t(r_t)-r^*\rangle + \overline{\gamma}_t^2\big\|\Delta_t^{(1)}+\Delta_t^{(2)}\big\|^2.$$

Our intention is to verify the conditions of Theorem 4 for almost all paths of the sequence $\{r_t\}$.

Condition A. The sequence $\{z_t\}$ is bounded by construction. Since the series (30) is a convergent martingale, we conclude that $\lim_{t\to\infty}\gamma_t z_t\widetilde{d}_t = 0$.

Conditions B and C: We follow the proof of Theorem 6. The deterministic term involving $\overline{U}_t(r_t)$ can be estimated as in (45):

$$\big\|r_t-\overline{\gamma}_t\overline{U}_t(r_t)-r^*\big\|^2 \le \big\|r_t-r^*\big\|^2 - 2\overline{\gamma}_t\big(1-\alpha(1+\varkappa)\big)\big\|\Phi(r_t-r^*)\big\|_q^2 + C\overline{\gamma}_t^2.$$

Since $\{z_t\}$ and $\{r_t\}$ are bounded, Assumptions 1 and 2 imply that $\sum_{t=0}^{\infty}\overline{\gamma}_t\langle\Delta_t^{(1)}, r_t - \overline{\gamma}_t\overline{U}_t(r_t)-r^*\rangle$ is a convergent martingale.

To analyze the second error term, $\Delta_t^{(2)}$, we observe that for a vector e_{i_k} having 1 at position i_k and zero otherwise, the formula (37) yields

$$\begin{aligned}z_t d_t &= \sum_{k=0}^{t}(\lambda\alpha)^{t-k}\varphi(i_k)\big(\varphi^\top(i_t)r_t - c(i_t) - \alpha\sigma_{i_t}(P_{i_t},\Phi r_t)\big)\\ &= \Phi^\top\Big(\sum_{k=0}^{t}(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top\Big)\big(\Phi r_t - c - \alpha\sigma(P,\Phi r_t)\big).\end{aligned}$$

Subtracting (39), we obtain

$$\Delta_t^{(2)} = \Phi^\top\Big(\sum_{k=0}^{t}(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top - \frac{1}{1-\lambda\alpha}Q\overline{P}_t\Big)\big[\Phi r_t - c - \alpha\sigma(P,\Phi r_t)\big].$$

By virtue of Lemma 10, for any $\varepsilon > 0$,

$$\lim_{T\to\infty}\frac{\sum_{t=0}^{T}\gamma_t\Delta_t^{(2)}}{\sum_{t=0}^{T}\gamma_t} = 0, \qquad \lim_{t_0\to\infty}\sup_{T\ge t_0}\frac{\sum_{t=t_0}^{T}\gamma_t\Delta_t^{(2)}}{\max\Big(\varepsilon, \sum_{t=t_0}^{T}\gamma_t\Big)} = 0 \quad \text{a.s.}.$$

The remaining analysis is the same as in the proof of Theorem 6. We obtain an inequality corresponding to (31):

$$\|r_T - r^*\|^2 \le \|r_{k_0} - r^*\|^2 + \left(-\delta(1-\alpha(1+\varkappa)) + \frac{\sum_{t=k_0}^{T-1} \gamma_t \langle \Delta_t^{(1)} + \Delta_t^{(2)}, r_t - \overline{\gamma}_t \overline{U}(r_t) - r^* \rangle}{\sum_{t=k_0}^{T-1} \overline{\gamma}_t} + C \frac{\sum_{t=k_0}^{T-1} \overline{\gamma}_t^2}{\sum_{t=k_0}^{T-1} \overline{\gamma}_t} \right) \sum_{t=k_0}^{T-1} \overline{\gamma}_t,$$

with $\delta > 0$. This allows us to verify the conditions of Theorem 4 and prove our assertion following the last steps of the proof of Theorem 6 verbatim. ■

We remark that the convergence condition for the multistep method: $\alpha(1+\varkappa) < 1$, is stronger that the condition for the basic method: $\alpha\sqrt{1+\varkappa} < 1$. Mathematically, it follows from the need to estimate (43) with the use of (11). Because of the multistep transition matrix $\overline{P}$, additional accumulation of risk premiums over time is introduced.

Again, as in the case of the basic method, discussed in section 4, the boundedness of the sequence $\{r_k\}$ is not an issue of concern, because it can be guaranteed by projection on a bounded set Y. The modified method has the following form:

$$r_{t+1} = \mathrm{Proj}_Y\big(r_t - \gamma_t z_t \widetilde{d}_t\big), \quad t = 0, 1, 2, \ldots. \tag{50}$$

We just need Y to have a nonempty intersection Y^* with the set of solutions of (41). Due to the nonexpansiveness of the projection operator, all our proofs remain unchanged with this modification, as discussed at the end of section 4.

We can also use (Borkar and Meyn, 2000, Thm. 2.1) to establish the boundedness a.s. of the sequence $\{r_t\}$.

Theorem 12 *Suppose the stepsizes $\{\gamma_t\}$ are deterministic and satisfy the deterministic version of Assumption 1, the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)$ satisfy* (15) *and Assumption 2, the matrix Φ has full column rank, and $\alpha(1+\varkappa) < 1$. Then the sequence $\{r_t\}$ is bounded with probability 1.*

The proof mimics the proof of Theorem 7, with the ODE $\dot{r}(t) = -\overline{U}(r(t))$ and the re-scaled ODE $\dot{r}(t) = -\overline{U}_\infty(r(t))$, where $\overline{U}_\infty(r) = \Phi^\top Q \overline{P}\big[\Phi r - \alpha\, \sigma(P, \Phi r)\big]$.

7. Empirical Illustration

7.1 Risk estimation

We first discuss the issue of obtaining stochastic estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \cdot)$ satisfying (15) and Assumption 2:

$$\mathbb{E}\big[\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)\big|\, \mathcal{F}_t\big] = \sigma_{i_t}(P_{i_t}, \Phi r_t), \quad t = 0, 1, 2, \ldots, \tag{51}$$

In the expected value case, where $\sigma_{i_t}(P_{i_t}, \Phi r_t) = P_{i_t} \Phi r_t = \mathbb{E}\big[\varphi^\top(i_{t+1}) r_t \,\big|\, \mathcal{F}_t\big]$, we could just use the approximation value at the next state observed, $\varphi^\top(i_{t+1}) r_t$, as the stochastic estimate of the expected value function. However, due to the nonlinearity of a risk measure with respect to the probability measure P_{i_t}, such a straightforward approach is no longer possible in a risk-averse setting.

Statistical estimation of measures of risk is a challenging problem, for which, so far, only solutions in special cases have been found (Dentcheva et al., 2017). To mitigate this problem, we propose to use a special class of transition risk mappings which are very convenient for statistical estimation. For a given transition risk mapping $\overline{\sigma}_i(P_i, v)$, we sample N conditionally independent transitions from the state i, resulting in states $j^1, \ldots, j^N$. This sample set defines a random empirical distribution, $P_i^N = \frac{1}{N}\sum_{k=1}^N e_{j^k}$, where e_j is the jth unit vector in $\mathbb{R}^n$. Since the sample set is finite, we can calculate the plug-in risk measure estimate,

$$\widetilde{\sigma}_i^N(P_i, v) = \overline{\sigma}_i(P_i^N, v), \tag{52}$$

by a closed-form expression. One can verify directly from the definition that the resulting *sample-based transition risk mapping*

$$\sigma_i^N(P_i, v) = \mathbb{E}\big[\overline{\sigma}_i(P_i^N, v)\big],$$

satisfies all conditions of a transition risk mapping of section 2, if $\overline{\sigma}_i(\cdot,\cdot)$ does. The expectation above is over all possible N-samples. Therefore, if we treat $\sigma_i^N(\cdot,\cdot)$ as the "true" risk measure that we want to estimate, the plug-in formula (52) satisfies (15) and Assumption 2. In fact, for a broad class of measures of risk $\overline{\sigma}_i(P_i, v)$, we have a central limit result: $\overline{\sigma}_i(P_i^N, v)$ is convergent to $\overline{\sigma}_i(P_i, v)$ at the rate $1/\sqrt{N}$, and the error has an approximately normal distribution (Dentcheva et al., 2017). However, we do not rely on this result here, because we work with fixed N. In our experiments, the sample size $N = 4$ turned out to be sufficient for generating a random observation of stage-wise risk, and even $N = 2$ would work well, because these observations are implicitly averaged over many iterations of a recursive stochastic algorithm.

7.2 Example

We apply the risk-averse methods of temporal differences to a version of a transportation problem discussed by Powell and Topaloglu (2006). We have $K = 200$ vehicles at $M = 50$ locations. At each time period t, a stochastic demand D_{ijt} for transportation from location i to location j occurs, $i, j = 1, \ldots, M$, $t = 1, 2, \ldots$. The demand arrays D_t in different time periods are independent and drawn from a truncated normal distribution: $D_{ijt} = \big\lfloor \max(0, \mathcal{N}(0, s_{ij}))\big\rfloor$. The vehicles available at location i may be used to satisfy this demand. They may also be moved empty. The state x_t of the system at time t is the M-dimensional integer vector containing the numbers of vehicles at each location. The size of the state space is $\binom{K+M-1}{M-1} \sim 10^{427}$.

For simplicity, we assume that a vehicle can carry a unit demand, and the total demand at the location i at time t can be satisfied only if $x_t(i) \geq \sum_{j=1}^M D_{ijt}$; otherwise, the demand may be only partially satisfied and the excess demand is lost. One can relocate the vehicles empty or loaded, and we denote the cost of moving a vehicle empty from location i to location j as c_{ij}^e. Since we stay in a cost minimization setting, we also denote the net negative profit of moving a vehicle loaded from location i to location j as c_{ij}^ℓ. Let u_{ijt}^e be the number of vehicles moved empty from location i to location j at time t and u_{ijt}^ℓ be the number of vehicles that are moved loaded. For notational simplicity, let us refer to the

combination of u_t^e and u_t^ℓ as u_t and denote:

$$c^\top u_t = \sum_{i,j=1}^{M} \big(c_{ij}^e u_{ijt}^e + c_{ij}^\ell u_{ijt}^\ell\big).$$

In this problem, the control u_t is decided *after* state x_t and demand D_t have been observed. The next state is a linear function of x_t and u_t:

$$x_{t+1} = x_t - Au_t,$$

where A can be written in an explicit way by counting the outgoing and incoming vehicles.

We denote by $\mathcal{U}(x_t, D_t)$ the set of decisions that can be taken at state x_t under demand D_t. Our approach allows us to evaluate a look-ahead policy defined by a simple linear programming problem:

$$u_t^\pi(x_t, D_t) = \operatorname*{argmin}_{u \in \mathcal{U}(x_t, D_t)} \Big\{c^\top u + \alpha \pi^\top (x_t - Au)\Big\}. \tag{53}$$

Here, π is the vector of approximate next-state values fully defining the policy. In our case, the immediate cost $c^\top u_t$ depends on D_t, and thus the risk-averse policy evaluation equation (3) has the following form:

$$v^\pi(x) = \sigma\Big(P, c^\top u^\pi(x, D) + \alpha v^\pi\big(x - Au^\pi(x, D)\big)\Big),$$

with P denoting the distribution of the demand. Our objective is to evaluate the policy π and to improve it. As the size of the state space is astronomical, we resort to linear approximations of form (5), using the state x as the feature vector:

$$\widetilde{v}^\pi(x_t) = r^\top x_t. \tag{54}$$

The approximate risk-averse dynamic programming equation (6) takes on the form:

$$r^\top x = L\sigma\Big(P, c^\top u^\pi(x, D) + \alpha r^\top \big(x - Au^\pi(x, D)\big)\Big), \tag{55}$$

where L is the weighted projection on the space of linear functions of x (we never use L explicitly).

In fact, we can combine the learning and policy improvement in one process, known as the *optimistic approach*, in which we always use the current r_t as the vector π defining the policy.

7.3 Results

We tested the risk-averse and the risk-neutral TD(λ) methods under the same long simulated sequence of demand vectors. At every time t, we sampled $N = 4$ instances of the demand vectors, and for each instance, we computed the best decisions by (53), and the resulting states. Then we computed the empirical risk measure (52) of the approximate value of the next state, and we used it in the observed temporal difference calculation (16):

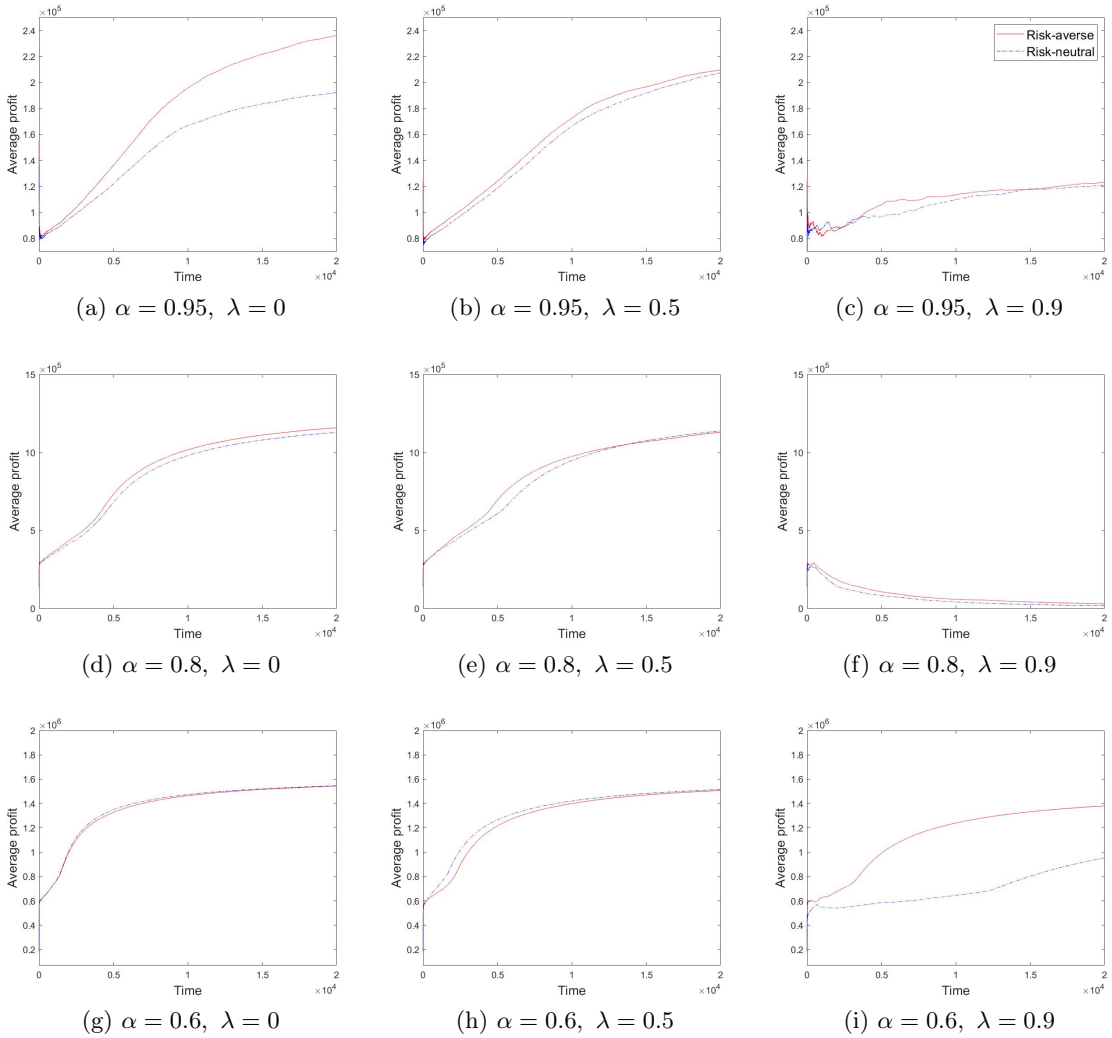
$$\widetilde{d}_t = r_t^\top x_t - \alpha\overline{\sigma}\Big(P^N, c^\top u^{r_t}(x_t, D) + \alpha r_t^\top \big(x_t - Au^{r_t}(x_t, D)\big)\Big). \tag{56}$$

We used the mean–semideviation risk measure of Example 1 as $\overline{\sigma}(\cdot,\cdot)$. It can be calculated in closed form for an empirical distribution P^N with observed transition costs $w^{(1)},\dots,w^{(N)}$:

$$\overline{\sigma}(P^N,v)=\mu+\beta\frac{1}{N}\sum_{j=1}^{N}\max(0,w^{(j)}-\mu),\quad \mu=\frac{1}{N}\sum_{j=1}^{N}w^{(j)},\quad \beta\in[0,1].$$

We used $\beta=1$ and $N=4$. The stepsize was constant and equal to $\gamma=0.0001$. In the expected value model ($\beta=0$), we also used $N=4$ observations per stage, and we averaged them, to make the comparison fair. The choice of $N=4$ was due to the use of a four-core computer, on which the N transitions could be simulated and analyzed in parallel.

We compared the performance of the risk-averse and risk-neutral TD(λ) algorithms for $\lambda=0$, 0.5, and 0.9, and $\alpha=0.95$, 0.8, and 0.6, in terms of average profit per stage, on a trajectory with 20,000 decision stages. The results are depicted in Figure 1. We observe



(a) $\alpha=0.95,\ \lambda=0$ (b) $\alpha=0.95,\ \lambda=0.5$ (c) $\alpha=0.95,\ \lambda=0.9$

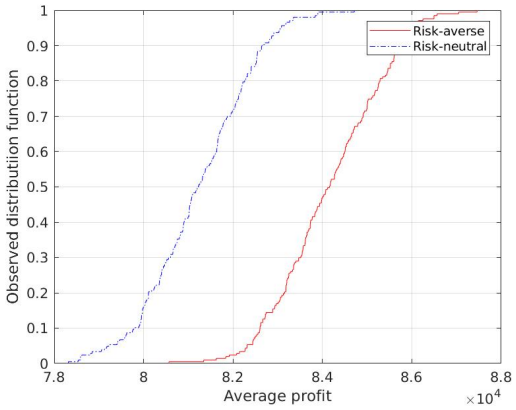
(d) $\alpha=0.8,\ \lambda=0$ (e) $\alpha=0.8,\ \lambda=0.5$ (f) $\alpha=0.8,\ \lambda=0.9$

(g) $\alpha=0.6,\ \lambda=0$ (h) $\alpha=0.6,\ \lambda=0.5$ (i) $\alpha=0.6,\ \lambda=0.9$

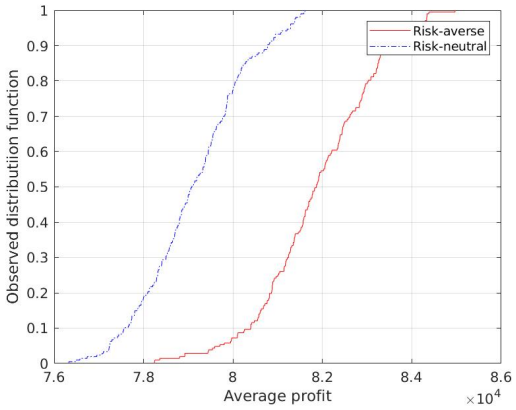
Figure 1: Evolution of the average profit per stage.

that in many cases the risk-averse algorithms outperform their risk-neutral counterparts in terms of the average profit in the long run. A large value of λ appears to be problematic and leads to the failure of the optimistic method in one case. Our results illustrate that the risk-averse method not only has solid theoretical guarantees but can also operate in an efficient way on a nontrivial problem.

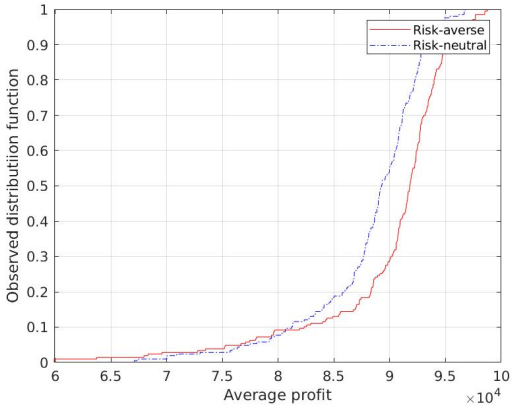
In addition to these results, we used 207 distinct trajectories, each with 200 decision stages, to compare the performance of the risk-averse and risk-neutral algorithms at the early training stages in terms of profit per stage. Figure 2 shows the empirical distribution function of the profit per stage of the risk-averse and risk-neutral algorithms at $t = 200$, for $\alpha = 0.95$ and $\lambda = 0$, 0.5, and 0.9.



(a) $\lambda = 0$



(b) $\lambda = 0.5$



(c) $\lambda = 0.9$

Figure 2: Empirical distribution of the average profit at $t = 200$.

The results demonstrate that in the early stages of learning ($t = 200$), the time-average profit of the risk-averse algorithm is more likely to be higher than that of the risk-neutral algorithm, and the difference is very pronounced for lower values of λ. The first order stochastic dominance relation between empirical distributions appears to exist. In that case,

by the consistency with first order stochastic dominance, for every law invariant monotonic risk measure, our risk-averse solution will be preferred (Shapiro et al., 2014, Thm. 6.50).

Although risk-averse methods aim at optimizing the dynamic risk measure, rather than the expected value, they may outperform the expected value model also in expectation. This may be due to the fact that the use of risk measures makes the method less sensitive to the imperfections of the value function approximation.

To gain more insight into this issue, we carried out a small-scale experiment with $K = 12$ vehicles and $M = 3$ locations (the first three locations of the large-scale experiment). In this case, the size of the state space is 91 and the calculation of the policy value functions $v^\pi_{\rm RN}(\cdot)$ and $v^\pi_{\rm RA}(\cdot)$ in the risk-neutral and the risk-averse models was possible. As a reference policy we chose the myopic policy, corresponding to $\pi = 0$ in (53). Then we carried out the approximate policy evaluation by the method of temporal differences, in both cases with the linear architecture, by using the observed temporal differences (56). In parallel to that, we calculated the "full" policy evaluation sequence $V_t(\cdot)$, $t = 0, 1, 2 \dots$, by the corresponding method of temporal differences as well:

$$\widetilde{d}^{\rm full}_t = V_t(x_t) - \alpha\overline{\sigma}\Big(P^N, c^\top u^{r_t}(x_t, D) + \alpha V_t\big(x_t - Au^{r_t}(x_t, D)\big)\Big),$$
$$V_{t+1}(x_t) = V_t(x_t) - \gamma\widetilde{d}^{\rm full}_t.$$

For the risk neutral model we proceeded in the same way, just $\sigma(\cdot, \cdot)$ was replaced by the expected value with respect to P^N.

In this way, in one simulation run of the myopic policy we learned four evaluations of this policy: two full 91-dimensional vectors for the risk-neutral and the risk-averse case, and two linear approximations for both cases. Figure 3 depicts the evolution of the mean-squared errors between the full and linear policy evaluation in the risk-neutral case (the top graph) and the full and linear policy evaluation in the risk-averse case (the bottom graph). The evaluations are learned over time, so the graphs stabilize half-way through the run. As we can see, both approximations are rather crude, and the final error is twice as large for the risk-neutral case, as it is for the risk averse case. The final values of the coefficients of the

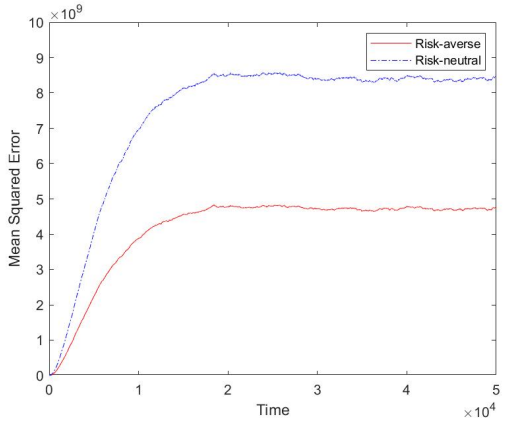


Figure 3: The mean-square errors of the policy value approximations.

linear value function approximations are denoted by $r_{\rm RN}$ and $r_{\rm RA}$, for the risk-neutral and the risk-averse case, respectively.

Finally, we compare long simulations of three policies: the initial myopic policy with $\pi = 0$ in (53), the new risk-neutral policy with $\pi = r_{\mathrm{RN}}$, and the new risk-averse policy with $\pi = r_{\mathrm{RA}}$ (this was the first step of the policy iteration method for both settings). Figure 4 depicts the average profits over time of the three policies. It appears that the linear

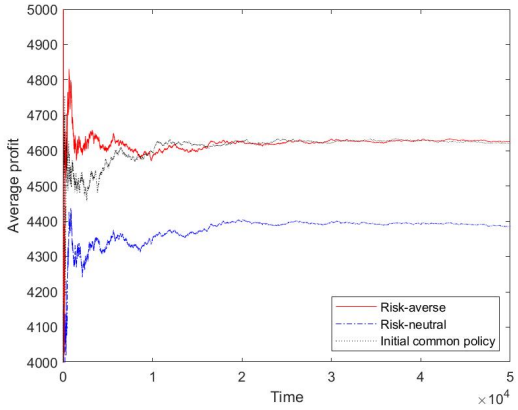


Figure 4: Performance of the three policies.

approximation of the value function is unsuitable in the risk-neutral case, and the myopic policy is better than the policy based on this approximation. In the risk-averse case, a small improvement in performance over the myopic policy is observed (even in the expected value and certainly in risk). Apparently, some robustness of the risk-averse method with respect to the model imperfections exists in our example.

Undoubtedly, very serious theoretical challenges arise from our preliminary observations. We plan to address them in our future research.

Acknowledgments

This work was partially supported by the National Science Foundation Award DMS-1907522. The authors thank six anonymous reviewers for providing a valuable feedback that helped them eliminate some inaccuracies and improve the manuscript. They also acknowledge the Office of Advanced Research Computing (http://oarc.rutgers.edu) at Rutgers, The State University of New Jersey, for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

References

A. Arlotto, N. Gans, and J. M. Steele. Markov decision problems where means bound variances. *Operations Research*, 62(4):864–875, 2014.

P. Artzner, F. Delbaen, J.-M. Eber, and D Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, 152:5–22, 2007.

A. Basu, T. Bhattacharyya, and V. S. Borkar. A learning algorithm for risk-sensitive cost. *Mathematics of Operations Research*, 33(4):880–898, 2008.

N. Bäuerle and U. Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2013.

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458, 2017.

R. Bellman, R. Kalaba, and B. Kotkin. Polynomial approximation – a new computational technique in dynamic programming. *Math. Comp.*, 17(8):155–161, 1963.

R. E. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6 (5):679–684, 1957.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 4 edition, 2017.

T. Bielecki, D. Hernández-Hernández, and S. R. Pliska. Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. *Mathematical Methods of Operations Research*, 50(2):167–188, 1999.

V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27 (2):294–311, 2002.

V. S. Borkar and S. P. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

V.S. Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339 – 346, 2001.

Ö. Çavus and A. Ruszczyński. Computational methods for risk-averse undiscounted transient Markov models. *Operations Research*, 62(2):401–417, 2014a.

Ö. Çavus and A. Ruszczyński. Risk-averse control of undiscounted transient Markov models. *SIAM Journal on Control and Optimization*, 52(6):3935–3966, 2014b.

Z. Chen, G. Li, and Y. Zhao. Time-consistent investment policies in Markovian markets: a case of mean-variance analysis. *J. Econom. Dynam. Control*, 40:293–316, 2014.

P. Cheridito and M. Kupper. Composition of time-consistent dynamic monetary risk measures in discrete time. *International Journal of Theoretical and Applied Finance*, 14(01): 137–162, 2011.

P. Cheridito, F. Delbaen, and M. Kupper. Dynamic monetary risk measures for bounded discrete-time processes. *Electronic Journal of Probability*, 11:57–106, 2006.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in neural information processing systems*, pages 3509–3517, 2014.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1): 6070–6120, 2017.

K. J. Chung and M. J. Sobel. Discounted MDPs: distribution functions and exponential utility maximization. *SIAM*, 25:49–62, 1987.

S. P. Coraluppi and S. I. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35(2):301–309, 1999.

P. Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.

P. Dayan and T. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 14: 295–301, 1994.

D. Dentcheva, S. Penev, and A. Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

S. Di-Castro Shashua and S. Mannor. Deep robust Kalman filter. *arXiv preprint arXiv:1703.02310*, 2017.

G. B. Di Masi and Ł. Stettner. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM J. Control Optim.*, 38(1):61–78, 1999.

J. Fan and A. Ruszczyński. Process-based risk measures and risk-averse control of discrete-time systems. *Mathematical Programming*, pages 1–28, 2018a.

J. Fan and A. Ruszczyński. Risk measurement and risk-averse control of partially observable discrete-time Markov systems. *Mathematical Methods of Operations Research*, 88: 161–184, 2018b.

B. G. Farley and W. A. Clark. Simulation of self-organizing systems by digital computer. *IRE Transactions on Information Theory*, 4:76–84, 1954.

J. A. Filar, L. C. M. Kallenberg, and H.-M. Lee. Variance-penalized Markov decision processes. *Math. Oper. Res.*, 14(1):147–161, 1989.

W. H. Fleming and S. J. Sheu. Optimal long term growth rate of expected utility of wealth. *The Annals of Applied Probability*, 9:871–903, 1999.

J. B. Hiriart-Urruty. Mean value theorems in nonsmooth analysis. *Numerical Functional Analysis and Optimization*, 2(1):1–30, 1980.

R. A. Howard. *Dynamic Programming and Markov Processes*. John Wiley & Sons, 1960.

R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Sci.*, 18:356–369, 1971.

W. Huang and W. B. Haskell. Risk-aware Q-learning for Markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE, 2017.

G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.

S. C. Jaquette. Markov decision processes with a new optimality criterion: discrete time. *Ann. Statist.*, 1:496–505, 1973.

A. Jaśkiewicz, J. Matkowski, and A. S. Nowak. Persistently optimal policies in stochastic dynamic programming with generalized discounting. *Mathematics of Operations Research*, 38(1):108–121, 2013.

H. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, New York, 2003.

S. Levitt and A. Ben-Israel. On modeling risk in Markov decision processes. In *Optimization and Related Topics (Ballarat/Melbourne, 1999)*, volume 47 of *Appl. Optim.*, pages 27–40. Kluwer Acad. Publ., Dordrecht, 2001.

K. Lin and S. I. Marcus. Dynamic programming with non-convex risk-sensitive measures. In *American Control Conference (ACC), 2013*, pages 6778–6783. IEEE, 2013.

W.-J. Ma, D. Dentcheva, and M. M. Zavlanos. Risk-averse sensor planning using distributed policy gradient. In *2017 American Control Conference (ACC)*, pages 4839–4844. IEEE, 2017.

W.-J. Ma, C. Oh, Y. Liu, D. Dentcheva, and M. M. Zavlanos. Risk-averse access point selection in wireless communication networks. *IEEE Transactions on Control of Network Systems*, 6(1):24–36, 2018.

A. Majumdar and M. Pavone. How should a robot assess risk? Towards an axiomatic theory of risk in robotics. In *Robotics Research*, pages 75–84. Springer, 2020.

S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean-variance optimization in Markov decision processes. *European J. Oper. Res.*, 231(3):645–653, 2013.

S. I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernández, S. Coraluppi, and P. Fard. Risk sensitive Markov decision processes. In *Systems and Control in the Twenty-First Century (St. Louis, MO, 1996)*, volume 22 of *Progr. Systems Control Theory*, pages 263–279. Birkhäuser, Boston, MA, 1997.

M. Métivier and P. Priouret. Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and Related Fields*, 74 (3):403–428, 1987.

M. L. Minsky. *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem*. PhD thesis, Princeton University, 1954.

J. Neveu. *Discrete-Parameter Martingales*. North-Holland, Amsterdam, 1975.

A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

E. A. Nurminski. Convergence conditions for nonlinear programming methods. *Kibernetika (Kiev)*, (6):79–81, 1972.

W. Ogryczak and A. Ruszczyński. From stochastic dominance to mean–risk models: semideviations as risk measures. *European Journal of Operational Research*, 116:33–50, 1999.

J. Peng. *Efficient Dynamic Programming-Based Learning for Control*. PhD thesis, Northeastern University, 1993.

J. Peng and R. J. Williams. Incremental multi-step Q-learning. *Proceedings of the Eleventh International Conference on Machine Learning*, pages 226–232, 1994.

G.Ch. Pflug and W. Römisch. *Modeling, Measuring and Managing Risk*. World Scientific, Singapore, 2007.

W. B. Powell. *Approximate Dynamic Programming - Solving the Curses of Dimensionality*. Wiley, 2011.

W. B. Powell and H. Topaloglu. Approximate dynamic programming for large-scale resource allocation problems. In *Models, Methods, and Applications for Innovative Decision Making*, pages 123–147. INFORMS, 2006.

L. A. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive reinforcement learning. *CoRR*, abs/1403.6530, 2014. URL `http://arxiv.org/abs/1403.6530`.

M. L. Puterman. *Markov Decision Processes*. Wiley, 1994.

F. Riedel. Dynamic coherent risk measures. *Stochastic Processes and Their Applications*, 112:185–200, 2004.

B. Roorda, J. M. Schumacher, and J. Engwerda. Coherent acceptability measures in multiperiod models. *Mathematical Finance*, 15(4):589–612, 2005.

G. A. Rummery. *Problem Solving with Reinforcement Learning*. PhD thesis, Cambridge University, 1995.

G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical report, Engineering Department, Cambridge University, 1994.

A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Math. Program.*, 125(2, Ser. B):235–261, 2010.

A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006a.

A. Ruszczyński and A. Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561, 2006b.

A. Ruszczyński and W. Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12):1097–1105, 1983.

G. Scandolo. *Risk Measures in a Dynamic Setting*. PhD thesis, Università degli Studi di Milano, Milan, Italy, 2003.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2014.

Y. Shen, W. Stannat, and K. Obermayer. Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

P. Sopasakis, D. Herceg, A. Bemporad, and P. Patrinos. Risk-averse model predictive control. *Automatica*, 100:281–288, 2019.

R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396, 2012.

A. Tamar, S. Mannor, and H. Xu. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2017.

J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Fransactions on Automatic Control*, 42(5):674–690, 1997.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

C. J. C. H. Watkins and P. Dayan. Q - learning. *Machine Learning*, 8(3-4):279–292, 1992.

L. L. Wegge. Mean value theorem for convex functions. *Journal of Mathematical Economics*, 1(2):207–208, 1974.

D. J. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: a review. *J. Optim. Theory Appl.*, 56(1):1–29, 1988.

P. Yu, W. B. Haskell, and H. Xu. Approximate value iteration for risk-aware Markov decision processes. *IEEE Transactions on Automatic Control*, 63(9):3135–3142, 2018.