2017 IEEE 56th Annual Conference on Decision and Control (CDC)
December 12-15, 2017, Melbourne, Australia

Risk-aware Q-Learning for Markov Decision Processes

Wenjie Huang[†] and William B. Haskell[‡]

Abstract— **We are interested in developing reinforcement learning algorithm to tackle risk-aware sequential decision making problems. The model we investigate is a discounted infinite-horizon Markov decision processes with finite state and action spaces. Our algorithm is based on estimating a general minimax function with stochastic approximation, and we show that several risk measures fall within this form. We derive finite-time bounds for this algorithm by combining stochastic approximation with the theories of risk-aware dynamic programming. Finally, we present extensions to several variations of risk measures.**

I. Introduction

The analysis of systems such as inventory control, financial market, waste-to-energy plants, and computer networks is becoming increasingly more difficult because of the growing complexity of these systems, especially when we want guarantees on the *risk* of these systems. Risk-aware optimization offers stronger and more reliable performance guarantees than the risk-neutral case, and furthermore it reflects the risk preferences of the decision maker.

In this paper, we synthesize the work on risk-aware sequential optimization with the stochastic approximation technique for risk-aware Markov decision processes. In many cases, the model underlying the Markov decision process is not known, and we can only observe the trajectory of states, actions, and rewards. Q-learning is designed specifically for this setting. As our contribution, we develop a new class of asynchronous Q-learning algorithms for risk-aware Markov decision processes. In particular, we focus on infinite-horizon problems with general minimax risk measures.

Mathematically, a risk measure a.k.a. risk function is a mapping from random variables to scalars. Given a probability space, a risk function should have certain properties, in particular, monotonicity and translation equivariance (see [1]). In recent years attention has turned towards convex and coherent risk measures. A monotone risk measure satisfying convexity is called convex risk measure and a convex risk measure also satisfying positive homogeneity is called coherent risk measure. In [2], a theory of convex analysis and optimization theory is developed for general convex risk measures. Several specific examples of convex and coherent risk measures along with their representation are given in [3, Chapter 6], including mean-deviation, average value-at-risk, and expected utility. The most widely investigated and most common risk measure is Conditional value-at-risk (CVaR). An optimization perspective for CVaR is developed in [4], which reveals that CVaR has many desirable properties for stochastic optimization. In [5], stochastic approximation is used to estimate CVaR in data-driven optimization problems. Moreover, in [6], stochastic interior-point algorithms are created for constrained stochastic optimization problems and cast in the framework of stochastic approximation. The most famous representation result for law-variant coherent risk measures would be the Kusuoka representation (see [7] for example), which shows that any law-invariant coherent risk measure can be represented as the supremum of mixtures of average value-at-risk measures. There are several other important classes of risk measures such as: optimized certainty equivalent [8], spectral measures of risk [9], and distortion risk measures [10].

We are especially interested in sequential decision-making problems. Markov decision processes (MDPs) introduced by Bellman in [11] provide a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of decision makers. In [12], modern formulations for MDPs and their computational techniques are presented in detail. In the area of reinforcement learning, there is a specific technique called Q-learning introduced in [13]. The idea of Q-learning is to use the observed state transitions and costs to approach the optimal policy (so that exact knowledge of the MDP is not needed). In [14], a thorough convergence proof of the Q-learning algorithm is given which combines stochastic approximation with the theory of parallel asynchronous algorithms (see [15] for more details on stochastic approximation). A similar technique is applied to the risk-sensitive control of finite MDPs problem in [16], where the optimal control policy converge to desired solutions with probability one.

Risk-aware MDPs have been studied from several perspectives. In [17], the authors provide convex analytic approach to risk-aware MDPs. In most cases, it can be shown that the problem can be formulated as infinite-dimensional linear program (LP) in occupation measures when the state space is augmented. They also provide a discretization method and finite approximations for solving the resulting LPs. Dynamic programming equations are developed for a wide class of risk-aware MDPs in [18], and value iteration and policy iteration algorithms are given. In [19], the authors specially investigate the problem of minimizing the Average-Value-at-Risk of the discounted cost over a finite and an infinite horizon which is generated by a MDP, and show that this problem can be reduced to an ordinary MDP. In [20], both risk and modeling errors are taken into account in MDP framework for risk-sensitive and robust decision

[†]Wenjie Huang (wenjie_huang@u.nus.edu) and [‡]William B. Haskell (isehwb@nus.edu.sg) are with Department of Industrial Systems Engineering & Management, National University of Singapore

978-1-5090-2873-3/17/$31.00 ©2017 IEEE

making. Moreover, an approximate value-iteration algorithm is presented to solve the problem with error guarantees. In [21], a specific class of risk measures called quantile based risk measures are proposed for MDP and a simulation-based approximate dynamic programming (ADP) algorithm is developed for the resulting risk-aware problem. This paper emphasizes importance sampling, to direct samples toward the *risky region* as the ADP algorithm progresses, which will increase the efficiency and accuracy of function approximation. In [22], the theory of risk-sensitive MDPs is developed, where the Markovian type of risk measures which depends only on the current state, rather than on the whole history in [18]. In [23], a risk-sensitive reinforcement learning algorithm based on utility function has been investigated. The Q-learning algorithm proposed in this work are able to tackle more general risk measure formulation than [23], and our algorithm is both model-free and distribution-free compared with above risk-aware sequential optimization techniques proposed in [17], [18], [20]–[22] which rely on given transition probability distribution or simulation.

II. Preliminaries

This section introduces preliminary concepts and notations to be used throughout the paper.

A. Risk measure

Define a certain probability space $(\Omega, \mathcal{F}, P_0)$, where Ω is a sample space, $\mathcal{F}$ is a $\sigma - algebra$ on Ω, and P_0 is a probability measure on $(\Omega, \mathcal{F})$. We concern throughout with random variables in $\mathcal{L} = L_\infty(\Omega, \mathcal{F}, P_0)$, the space of essentially bounded $\mathcal{F}$-measurable functions. If let $\mathcal{X}$ denote the linear space of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}$. For $X, Y \in \mathcal{X}$, the notation $Y \succeq X$ means that $Y(\omega) \geq X(\omega)$ for all $\omega \in \Omega$.

We first define a *risk function* following [2]. It is a function ρ which assigns to an uncertain random variable X a real value $\rho(X)$. A risk function is a mapping $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ if $\rho(0)$ is finite and if ρ satisfies the following conditions for all $X, Y \in \mathcal{X}$. The following four properties of risk functions are important throughout our analysis:

(A1) Monotonicity: If $Y \succeq X$, then $\rho(X) \geq \rho(Y)$.

(A2) Translation Invaraince: If $z \in \mathbb{R}$, then $\rho(X + z) = \rho(X) - z$.

(A3) Convexity: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$, for $0 \leq \lambda \leq 1$.

(A4) Positive Homogeneity: If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda\rho(X)$.

Any real valued function $\rho : \mathcal{X} \to \mathbb{R}$ satisfying conditions (A1)-(A3) is called *convex measures of risk*. Any convex risk measure satisfies condition (A4) is called *coherent risk measure*. We consider minimax risk measure in this paper. *Minimax risk measure* is a range of risk measure investigated in [24]–[27]. To construct, we define set $\mathcal{Y} \subseteq \mathbb{R}^q$ and $\mathcal{Z} \subseteq \mathbb{R}^p$, $p, q > 0$ to be compact sets and let function $G : \mathcal{Z} \times \mathcal{Y} \times \mathcal{L} \to \mathbb{R}$ be a loss function depends on primal decision variable z, dual decision variable y and random variable X. We assume in our paper that function G is sub-differentiable on its support. In addition, we assume G is convex on $\mathcal{Z}$ and concave on $\mathcal{Y}$, i.e., it is a saddle function. The general minimax risk measure can be formulated as

$$\rho_M(X) = \min_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}} \mathbb{E}\left[G(z, y, X)\right]. \tag{1}$$

In this paper, we use a primal-dual stochastic approximation method to estimate (1), and we show that a wide class of risk measures fit into this form.

B. Risk-aware MDPs

In [18], a dynamic programming model for risk-aware MDP problem is developed. We first introduce a typical representation for a discrete time MDP:

$$(\mathcal{S}, \mathcal{A}, P, c).$$

Both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite. Let $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(\mathcal{A})$ denote the spaces of probability measures over $\mathcal{S}$ and $\mathcal{A}$, respectively. The set of feasible state-action pairs is $\mathbb{K} := \{(s, a) \in \mathcal{S} \times \mathcal{A} : a \in A(s)\}$. The transition law P governs the system evolution, $P(\cdot|s, a) \in \mathcal{P}(\mathcal{A})$ for all $(s, a) \in \mathbb{K}, i.e., P(j|s, a)$ for $j \in \mathcal{S}$ is the probability of visiting the state j next given the current state-action pair (s, a). Finally $c : \mathbb{K} \to \mathbb{R}$ is the cost associated with each state-action pair. Let Π denote the class of stationary deterministic Markov policies, i.e., mappings $\pi : \mathcal{S} \to \mathcal{A}$, and we assume that Π only contains feasible policies that correspond to the set $\mathbb{K}$. The state and action at time t are denoted as s_t and a_t, respectively. The general risk-aware infinite-horizon MDP is then

$$\min_{\pi \in \Pi} \{\rho(c(s_0, a_0) + \gamma\rho(c(s_1, a_1) + \gamma\rho(c(s_2, a_2) + \cdots)))\},$$

and the risk measure ρ conforming to the above structure is called *dynamic risk measure* with time-consistent property defined in [18]. Based on [18, Theorem 4], the infinite-horizon risk-aware MDP problem has an optimal stationary Markov policy and the optimal value can be computed via:

$$v(s) = \min_{a \in \mathcal{A}} \left\{c(s, a) + \gamma\rho_{X\sim P(\cdot|s,a)}(v(X))\right\}, \; s \in \mathcal{S}. \tag{2}$$

Here $\rho_{X\sim P(\cdot|s,a)}(v(X))$ denotes the risk measure of value function $v(X)$ with respect to the transition probability $P(\cdot|s, a)$ given state-action pair (s, a). Based on [13], we can define the Q-values for (2) as state-action value version, by defining Q-value as:

$$Q(s, a) = c(s, a) + \gamma\rho_{X\sim P(\cdot|s,a)}(v(X)), \tag{3}$$

where $v(X) = \min_{a' \in \mathcal{A}} Q(X, a')$. In other words, Q-value stores the risk measure for executing action a at state s. This reformulation sets up Q-learning. As the main contribution of our paper, we are interested in solving following risk-aware optimality equation in data driven way.

$$Q^*(s, a) = c(s, a) + \gamma\rho_{X\sim P(\cdot|s,a)}(v^*(X)), \tag{4}$$

where $Q^*(s, a)$ to be optimal Q-value associated with (s, a), and the corresponding optimal value function to be $v^*(X) = \min_{a' \in \mathcal{A}} Q^*(X, a')$.

III. The Risk-aware Q-learning Algorithm

As our main contribution, we combine risk-aware MDPs with stochastic approximation to solve risk-aware sequential decision-making problems online. We call this type of algorithm *Risk-aware Q-learning (RaQL)*. We would illustrate this algorithm in this section.

First we extend (1) to the Markovian setting through the recursive relationship (2) and (3). We define the sets $\mathcal{Y} := \{y | y : \mathcal{S} \times \mathcal{A} \to \mathscr{Y}\}$, $\mathscr{Y} \subseteq \mathbb{R}^q$ and $\mathcal{Z} := \{z | z : \mathcal{S} \times \mathcal{A} \to \mathscr{Z}\}$, $\mathscr{Z} \subseteq \mathbb{R}^p$, and both $\mathscr{Y}$ and $\mathscr{Z}$ are compact sets. Let P' be a fixed probability measure on $(\mathcal{S}, \mathscr{B}(\mathcal{S}))$, and construct $\mathcal{V} = \mathcal{L}_\infty(\mathcal{S}, \mathscr{B}(\mathcal{S}), P')$. Let function $G : \mathscr{Z} \times \mathscr{Y} \times \mathcal{V} \to \mathbb{R}$ be a loss function depending on primal decision variable z, dual decision variable y, and value function. Then $\forall (s, a) \in \mathbb{K}$, the minimax risk measure defined for Q-values is:

$$Q(s, a) = \quad c(s, a) + \gamma \min_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}} \{\mathbb{E}_{X \sim P(\cdot|s,a)}[\quad (5)$$

$$G(z(s, a), y(s, a), v(X))]\}, \quad (6)$$

where $v(X) = \min_{a' \in \mathcal{A}} Q(X, a')$.
We need an estimator to deal with (1), and we note that Problem (1) naturally leads to an stochastic gradient descent type algorithm. Similar algorithms have been investigated in [21], [28]–[30]. For our MDP, we define $z^*(s, a)$ and $y^*(s, a)$ to be the optimal solutions for (5)-(6). Using (4) allows us to write

$$Q^*(s, a) = \quad \mathbb{E}_{X \sim P(\cdot|s,a)}[c(s, a) \quad (7)$$

$$+\gamma G(z^*(s, a), y^*(s, a), v^*(X))]. \quad (8)$$

We first discuss the necessary definitions and assumptions for our algorithm, most of which are standard in stochastic approximation. Define the following subgradients:

$$\begin{aligned}
&g_t^{n-1} \in \partial_z \mathbb{E}\left[G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v^*)\right], \\
&h_t^{n-1} \in \partial_y \mathbb{E}\left[G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v^*)\right], \\
&g_t(Q_{t+1}^{n-1}) = \partial_z G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v_{t+1}^{n-1}), \\
&g_t(Q^*) = \partial_z G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v^*).
\end{aligned}$$

Additionally, we define

$$\begin{aligned}
&h_t(Q_{t+1}^{n-1}) = \partial_y G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v_{t+1}^{n-1}), \\
&h_t(Q^*) = \partial_y G(z_t^{n-1}(s, a), y_t^{n-1}(s, a), v^*).
\end{aligned}$$

Assumption 1: The followings hold: (i) (Bounded cost) For all $(s, a) \in \mathbb{K}$, there exists a constant $C_{\max}$ such that $c(s, a) \leq C_{\max}$.

(ii) (ε-greedy policy) The RaQL algorithm is defined on a new probability space $(\Omega, \mathcal{G}, P)$, where

$$\mathcal{G} = \sigma\{(s_t^n, a_t^n), n \geq 0, t \leq T\},$$

and the history of the algorithms is

$$\mathcal{G}_t^n = \left\{\sigma\left\{(s_\tau^i, a_\tau^i), i < n, \tau \leq T\right\} \cup \left\{(s_\tau^i, a_\tau^i), \tau \leq t\right\}\right\},$$

for $t \leq T$ and $n \geq 1$, with $\mathcal{G}_t^0 = \{\emptyset, \Omega\}$ for all $t \leq T$. The resulting filtration obeys $\mathcal{G}_t^n \subseteq \mathcal{G}_{t+1}^n$ for $t \leq T - 1$ and $\mathcal{G}_T^n \subseteq \mathcal{G}_0^{n+1}$. Suppose that there exists an $\varepsilon > 0$ such that for all $(s, a) \in \mathbb{K}$ and $t \leq T$, the state sampling policy satisfies

$$\mathbb{P}\left((s_t^n, a_t^n) = (s, a) | \mathcal{G}_{t-1}^n\right) \geq \varepsilon,$$

and

$$\mathbb{P}\left((s_0^n, a_0^n) = (s, a) | \mathcal{G}_T^{n-1}\right) \geq \varepsilon,$$

which is an exploration requirement.

(iii) For all $(s, a) \in \mathbb{K}$ and $t \leq T$, there exist constant $L_\mathcal{Z}, L_\mathcal{Y} > 0$ that

$$|\mathbb{E}\left[g_t(Q_{t+1}^{n-1}) - g_t(Q_{t+1}^*)\right]| \leq L_\mathcal{Z} \|Q_{t+1}^{n-1} - Q_{t+1}^*\|_\infty,$$

and

$$\begin{aligned}
&|g_t^{n-1} - \partial_z \mathbb{E}\left[G(z_t^*(s, a), y_t^*(s, a), v^*)\right]| \\
&\leq L_\mathcal{Z} \|z_t^{n-1} - z_t^*\|_\infty.
\end{aligned}$$

Similarly, we have

$$|\mathbb{E}\left[h_t(Q_{t+1}^{n-1}) - h_t(Q_{t+1}^*)\right]| \leq L_\mathcal{Y} \|Q_{t+1}^{n-1} - Q_{t+1}^*\|_\infty,$$

and

$$\begin{aligned}
&|h_t^{n-1} - \partial_y \mathbb{E}\left[G(z_t^*(s, a), y_t^*(s, a), v^*)\right]| \\
&\leq L_\mathcal{Y} \|y_t^{n-1} - y_t^*\|_\infty,
\end{aligned}$$

and we assume that the these subgradients are bounded i.e., $\|\partial_z G(z, y, v)\| \leq \kappa_z$ and $\|\partial_y G(z, y, v)\| \leq \kappa_y$ for $(z, y, v) \in \mathcal{Z} \times \mathcal{Y} \times \mathcal{V}$.

(iv) The risk-aware functions $G : \mathcal{Z} \times \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ is Lipschitz continuous and there exists constant $L_\Phi > 0$, i.e, for all $x, x' \in \mathcal{Z} \times \mathcal{Y} \times \mathcal{V}$, satisfying $|G(x) - G(x')| \leq L_\Phi \|x - x'\|_1$.

Assumption 1 (ii) guarantees, by the Extended Borel-Cantelli Lemma in [31], that we will visit every state infinitely often with probability one. In particular, for an ε-greedy sampling policy (i.e., explore with probability ε, follow current policy otherwise), this assumption holds. Assumption 1 (iii) bounds the sub-gradients of function G both on z and y, and requires the sub-gradients and it corresponding expected value to be lipschitz continuous on Q, z and y respectively. Assumption 1 (iv) states that G is the Lipschitz continuous on its support.

The structure of RaQL are illustrated in Algorithm 1.

We illustrate the detail procedures of Step 1-3 of Algorithm 1 as follows:
Step 1: Determine the action using ε-greedy policy. With probability ε, choose an action a_t^n at random from $\mathcal{A}$. With probability $1 - \varepsilon$, choose the greedy-policy using $a_t^n = \arg\min_{a \in \mathcal{A}} Q_t^{n-1}(s_t^n, a)$. Observe a new state s_{t+1}^n. Compute

$$\begin{aligned}
&\hat{q}_t^n(s_t^n, a_t^n) = c(s_t^n, a_t^n) \\
&+\gamma G\left(z_t^{n-1}(s_t^n, a_t^n), y_t^{n-1}(s_t^n, a_t^n), v_{t+1}^{n-1}(s_{t+1}^n)\right),
\end{aligned}$$

where $v_{t+1}^{n-1}(s_{t+1}^n) = \min_{a \in \mathcal{A}} Q_{t+1}^{n-1}(s_{t+1}^n, a)$.
Step 2: Compute sub-gradients

$$\begin{aligned}
&\nabla_z G\left(z_t^{n-1}(s_t^n, a_t^n), y_t^{n-1}(s_t^n, a_t^n), v_{t+1}^{n-1}(s_{t+1}^n)\right), \\
&\nabla_y G\left(z_t^{n-1}(s_t^n, a_t^n), y_t^{n-1}(s_t^n, a_t^n), v_{t+1}^{n-1}(s_{t+1}^n)\right),
\end{aligned}$$

Algorithm 1 RaQL with general minimax risk measure

```
Initialize an approximation $Q^0(s, a)$ for all states $s \in \mathcal{S}$ and decision $a \in \mathcal{A}$
Initialize learning rate $k$ ,and stepsize rule $\theta_t, \lambda_t > 0$ for $t \leq T$
Initialize $(z^0(s, a), y^0(s, a)) \in \mathcal{Z} \times \mathcal{Y}$
Set $Q_T^n = 0$ for all $n$
for $n = 1, 2, ...$ do
    Randomly choose an initial state $s_0^n \in \mathcal{S}$
    for $t = 0, 1, ..., T-1$ do
        Step 1. Choose action based on $\varepsilon$ policy;
        Step 2. Online risk measure estimation by stochastic approximation (w.r.t inner loops $t$);
        Step 3. Update Q-value by standard Q-learning algorithm (w.r.t outer loops $n$);
    end for
    For all $(s, a) \in \mathbb{K}$, update
        $z^n(s, a) = z^{n-1}(s, a), \ y^n(s, a) = y^{n-1}(s, a).$
end for
return  Approximated $Q$-value $\{Q^n\}$.
```

update by primal-dual structure

$$
\begin{aligned}
z_{t+1}^{n-1}(s, a) = \Pi_{\mathcal{Z}} \{ & z_t^{n-1}(s, a) - \theta_{t,k}^n \\
& \times \nabla_z G \left(z_t^{n-1}(s_t^n, a_t^n), y_t^{n-1}(s_t^n, a_t^n), v_{t+1}^{n-1}(s_{t+1}^n) \right) \},
\end{aligned}
$$

and

$$
\begin{aligned}
y_{t+1}^{n-1}(s, a) = \Pi_{\mathcal{Y}} \{ & y_t^{n-1}(s, a) + \theta_{t,k}^n \\
& \times \nabla_y G \left(z_t^{n-1}(s_t^n, a_t^n), y_t^{n-1}(s_t^n, a_t^n), v_{t+1}^{n-1}(s_{t+1}^n) \right) \},
\end{aligned}
$$

where the Euclidean projection operator to a set $\mathcal{X}$ is given by the usual definition: $\Pi_{\mathcal{X}}(y) = \arg\min_{z \in \mathcal{X}} \|y - x\|_2^2$.
Step 3: For all $(s, a) \in \mathbb{K}$, update Q_t^n and using

$$
\begin{aligned}
Q_t^n(s, a) = \quad & \left(1 - \lambda_{t,k}^n(s, a)\right) Q_t^{n-1}(s, a) \\
& + \lambda_{t,k}^n(s, a) \cdot \hat{q}_t^n(s, a),
\end{aligned}
$$

with $k \in (1/2, 1)$ called polynomial learning rate, and $k = 1$ linear learning rate.

In the above steps, for all $(s, a) \in \mathbb{K}$ and $t \leq T$, the step-sizes for stochastic approximation of risk measure satisfy a.s. that: $\sum_{n=1}^{\infty} \theta_t^n(s, a) = \infty$, and $\sum_{n=1}^{\infty} \theta_t^n(s, a)^2 < \infty$, and learning rate of Q-function satisfies a.s. that $\sum_{n=1}^{\infty} \lambda_t^n(s, a) = \infty$, and $\sum_{n=1}^{\infty} \lambda_t^n(s, a)^2 < \infty$. We assume that both the step-sizes $\theta_{t,k}^n(s, a)$ and learning rate $\lambda_{t,k}^n(s, a)$ are deterministic harmonic sequences satisfying $\theta_{t,k}^n(s, a) = \frac{\theta_t}{n^k} 1_{\{(s,a)=(s_t^n, a_t^n)\}}$, and $\lambda_{t,k}^n(s, a) = \frac{\lambda_t}{n^k} 1_{\{(s,a)=(s_t^n, a_t^n)\}}$, where $\theta_t, \lambda_t > 0$ are deterministic time-dependent constants.

Above statements represent the asynchronous nature of the algorithm, sending the step-size to zero whenever a state is not visited.

In our RaQL algorithm, we avoid overestimation by keeping track of mutually dependent approximations $\{Q^n\}$, $\{z^n\}$, and $\{y^n\}$ to the optimal values Q^*, z^*, and y^* in outer and inner loops respectively, where the approximate Q-values are updated in the outer loop and the risk is approximated in the inner loop by stochastic approximation on $\{z^n\}$ and $\{y^n\}$.

The following theorem presents the finite-time upper bounds on the convergence rates of RaQL algorithm. Our results only exhibit polynomial learning rates in this section.

Theorem 1: (Convergence Rate) Suppose an ε-greedy sampling policy is used, and the conditions of Assumption 1 hold for asynchronous RaQL with polynomial learning rate i.e., $k \in (1/2, 1)$. Let N and T be the number of iteration for outer and inner loops, respectively. Define $\tilde{\varepsilon}$ as any positive constant. Suppose

$$
\begin{aligned}
N = \Omega \Bigg(& \left(\frac{V_{\max}^2 \ln(V_{\max} |S||A| L / \delta \beta(T) \tilde{\varepsilon})}{\beta(T) \tilde{\varepsilon}^2 (1-\varepsilon)^{1+3k}} \right)^{1/k} \\
& + \left(\frac{1}{(1-\varepsilon)\beta(T)} \ln \frac{V_{\max}}{\tilde{\varepsilon}} \right)^{\frac{1}{1-k}} \Bigg),
\end{aligned}
$$

where

$$
\beta(T) = \left\{ 1 - \gamma L_\Phi - \left[1 + (\kappa_z^2 + \kappa_y^2) \theta_T^2 \right] / (1-\varepsilon)^k T^k \right\} / 2
$$

and $V_{\max} = C_{\max}/(1-\gamma)$, then Q_T^N generated by Algorithm 1 satisfies $\mathbb{E}\left[\|Q_T^N - Q^*\|_2 \right] \leq \tilde{\varepsilon}$ with probability $1 - \delta$.

Proof: Define $0 < \beta(t) < 1$, $t \leq T$ and denote $V_{\max} = C_{\max}/(1-\gamma)$. We define a sequence of value D, such that $D_1 = V_{\max}$ and $D_{m+1} = (1 - \beta(t)) D_m$ for $m \geq 1$. Clearly the sequence D_m converges to zero since all $\beta(t) \in [0, 1]$, and for every m there exists some time τ_m such that for any $n \geq \tau_m$ we have $\|Q^n - Q^*\| \leq D_m$. Then, we can decompose the problem into two processes:

$$
Z_t^{n+1,\tau}(s, a) = (1 - \lambda_{t,k}^n(s, a)) Z_{t,\tau}(s, a) + \lambda_{t,k}^n(s, a) \varepsilon_{t+1}^{n,\tau}
$$

We have

$$
\begin{aligned}
& Y_t^{n+1,\tau}(s, a) \leq (1 - \lambda_{t,k}^n(s, a)) Y_{t,\tau}(s, a) \\
& + \lambda_{t,k}^n(s, a) L_\Phi \left(\gamma D_m + \|z_{t+1}^{n-1}(s, a) - z^*(s, a)\|_\infty \right. \\
& \left. + \|y_{t+1}^{n-1}(s, a) - y^*(s, a)\|_\infty \right).
\end{aligned}
$$

Then we have, for all $\forall (s, a) \in \mathbb{K}$, $Q_t^n(s, a) - Q^*(s, a) \leq |Z_t^{n,\tau}(s, a)| + Y_t^{n,\tau}(s, a)$. For any $n \in [\tau_{m+1}, \tau_{m+2}]$ and $1 \leq l \leq n$, we have that $Z_t^{l,\tau_m}(s, a)$ is a martingale sequence, satisfying

$$
|Z_t^{l,\tau_m}(s, a) - Z_t^{l-1,\tau_m}(s, a)| \leq \frac{1}{(1-\varepsilon)^k \tau_m^k} V_{\max}.
$$

Additionally, we obtain that, under the ε-greedy sampling policy,

$$
\begin{aligned}
& \mathbb{E}\left[\|z_{t+1}^{n-1} - z^*\|_2^2 + \|y_{t+1}^{n-1} - y^*\|_2^2 \right] \\
& \leq \frac{1}{(1-\varepsilon)^k t^k} \left[1 + (\kappa_z^2 + \kappa_y^2) \theta_t^2 \right] D_m.
\end{aligned}
$$

Thus assume that for any $n \geq \tau_m$ we have $Y^{n+1,\tau_m} \leq D_m$. Then, for any $n \geq \tau_{m+1}$ we have

$$
\begin{aligned}
& Y^{n+1,\tau_m} \leq \\
& D_m (\gamma L_\Phi + \frac{1}{(1-\varepsilon)^k t^k} \left[1 + (\kappa_z^2 + \kappa_y^2) \theta_t \right] + \frac{2}{e} \beta(t)).
\end{aligned}
$$

Futher, by applying Azuma's inequality to $Z_t^{n,\tau_m}(s,\, a)$, we have, for $t \leq T$,

$$\mathbb{P}\left[\forall m \in [1,\, L],\, \forall n \in [\tau_{m+1},\, \tau_{m+2}],\right.$$
$$\left.\forall s,\, a:\, |Z_t^{n,\tau_m}(s,\, a)| \leq \tilde{\varepsilon}\right] \geq 1-\delta,$$

where $\tau_0 = \Theta\left(\left(\frac{V_{\max}^2 \ln(V_{\max}|S||A|L/\delta\beta(t)\tilde{\varepsilon})}{\beta(t)\tilde{\varepsilon}^2(1-\varepsilon)^{1+3k}}\right)^{1/k}\right)$.

Given the following fact that : let

$$a_{m+1} = a_m + \frac{1}{1-\varepsilon} a_m^k = a_0 + \sum_{i=0}^{m} \frac{1}{1-\varepsilon} a_i^k.$$

Then for any constant $k \in (0,\, 1)$,

$$a_m = O\left((a_0^{1-k} + \frac{1}{1-\varepsilon} m)^{\frac{1}{1-k}} \right)$$
$$= O\left(a_0 + (\frac{1}{1-\varepsilon} m)^{\frac{1}{1-k}} \right).$$

We then have the desired result. ■

From above theorem, we derive the convergence rate with respect to the ε-greedy policy, and we use an outer-inner loop structure in the algorithm, so the quality of risk estimation increases with number of inner iterations t in $\beta(t)$.

IV. Examples

Based on our previous arguments, we can show that several classes of risk measures will work with our RaQL algorithm. We start with optimized certainty equivalent (OCE). Given random variable $\mathcal{W} \in \mathcal{L}$, OCE is defined in [8] by the formula $S_u(\mathcal{W}) = \sup_{\eta \in \mathbb{R}} \{\eta + \mathbb{E}u(\mathcal{W}-\eta)\}$, where u is a normalized concave utility function which is differentiable with bounded derivative $\partial u(\cdot)$. To capture costs, we use the risk measure $\rho(\mathcal{W}) = -S_u(-\mathcal{W})$. Based on [8, Proposition 2.1], the expression of OCE is equivalent to $S_u(\mathcal{W}) = \sup_{\eta \in [\mathcal{W}_{\min}, \mathcal{W}_{\max}]} \{\eta + \mathbb{E}u(\mathcal{W}-\eta)\}$, where $[\mathcal{W}_{\min}, \mathcal{W}_{\max}]$ is the support of random variable $\mathcal{W}$. In our MDP, given a current state-action pair $(s,\, a) \in \mathbb{K}$, we define $\eta : \mathcal{S} \times \mathcal{A} \to [0,\, C_{\max}]$, and mapping $G : (0,\, C_{\max}] \times \mathcal{V} \to \mathbb{R}$, via $G(\eta(s,\, a),\, v(X)) = \eta(s,\, a) - u\,(\eta(s,\, a) - v(X))$, to obtain

$$\rho_{X\sim P(\cdot|s,a)}(v(X)) = \max_{\eta(s,\,a)\in(0,\,C_{\max}]} \mathbb{E}_{X\sim P(\cdot|s,a)}[$$
$$G(\eta(s,\, a),\, v(X))].$$

Since function G is obviously convex in η, and above problem with OCE can be treated as a special case for our algorithm with only primal variables.

As a remark, Conditional value at risk (CVaR), the most widely investigated coherent risk measure, is a special case of OCE corresponding to the utility function $u(t) = -(1-\alpha)\max\{0, -t\}$ with $\alpha \in [0,\, 1)$.

The absolute semi-deviation is a type of mean-risk measure. As defined in [3], the absolute semi-deviation of the random variable $\mathcal{W} \in \mathcal{L}$ is

$$\rho(\mathcal{W}) = \mathbb{E}[\mathcal{W}] + r\mathbb{E}\left[(\mathcal{W} - \mathbb{E}([\mathcal{W}])_+\right],$$

with weight coefficient $r \in [0,\, 1]$. We then define

$$G(\eta,\, \phi,\, W) = W + r\,[W-\eta]_+ + r\phi\,[\eta - W]$$

to obtain $\rho(\mathcal{W}) = \min_{\eta\in\mathbb{R}} \max_{\phi\in[0,\,1]} \mathbb{E}\,[G(\eta,\, \phi,\, W)]$. For our MDP, we define $\eta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the mapping $G : \mathbb{R} \times [0,\, 1] \times \mathcal{V} \to \mathbb{R}$, via

$$G(\eta(s,\, a), \phi(s,\, a)\, v(X)) = \quad v(X) + r\,[v(X) - \eta(s,\, a)]^+$$
$$+ r\phi(s,\, a)\,(\eta(s,\, a) - v(X))\,,$$

to obtain

$$\rho_{X\sim P(\cdot|s,a)}(v(X)) = \min_{\eta\in\mathbb{R}} \max_{\phi\in[0,\,1]} \mathbb{E}_{X\sim P(\cdot|s,a)}[$$
$$G(\eta(s,\, a), \phi(s,\, a)\, v(X))].$$

As in [32], [33], we consider functionally coherent risk measures given by a discretization $\{\alpha_i\}_{i=0}^m \subset [0,\, 1)$ so ordered that $0 \leq \alpha_0 < \alpha_1 < \cdots < \alpha_m \leq 1$. Then, we define $\eta : \mathcal{S} \times \mathcal{A} \to [0,\, \eta_{\max}]^m$ and $p : \mathcal{S} \times \mathcal{A} \to \mathfrak{M}$, and define the mapping $H : [0,\, \eta_{\max}]^m \times \mathfrak{M} \times \mathcal{V} \to \mathbb{R}$ as

$$H(\eta(s,\, a),\, p(s,\, a),\, v(X)) := \sum_{i=1}^{m} p_i(s,\, a)\,\big(\eta^i(s,\, a)$$
$$+ (1-\alpha_i)^{-1}\left[v(X) - \eta^i(s,\, a)\right]^+\big),$$

where $\eta^i(s,\, a)$ represents the η value of state-action pair $(s,\, \alpha)$ with respect to the confidence level α_i, $i = 1, 2, ..., m$. The resulting risk measure is

$$\rho_{X\sim P(\cdot|s,a)}(v(X)) = \min_{\eta(s,\,a)\in[0,\,\eta_{\max}]^m} \max_{p(s,\,a)\in\mathfrak{M}} \{$$
$$\mathbb{E}_{X\sim P(\cdot|s,a)}[$$
$$H(\eta(s,\, a),\, p(s,\, a),\, v(X))]\}.$$

The function H is convex on η and concave in p, so we are able to simultaneously perform gradient descent in the primal variables $\eta(s,\, a)$ (corresponding to minimization) and gradient ascent in the dual variables $p(s,\, a)$ (corresponding to maximization).

V. Numerical Experiments

In this section, we conduct numerical experiments to validate the performance of our RaQL algorithm with different risk measures. We evaluate our algorithm in terms of the relative error $\|Q^n - Q^*\|/\|Q^*\|$, $n \leq N$, where Q^* denotes the optimal Q-value derived by traditional risk-aware dynamic programming (RaDP) as proposed in [18], and Q^n denotes the Q-value given by RaQL algorithm in nth iteration. We experiment on an MDP with random cost, discounter factor $\gamma = 0.5$, and with five states and five actions. We set both the number of outer and inner iterations to be $N = T = 100$. In those experiments, we use linear learning rate i.e., $k = 1$ in both $\theta_{t,k}^n(s,\, a)$ and $\lambda_{t,k}^n(s,\, a)$ for all $t \leq T, n \leq N$ and $(s,\, a) \in S$.

- *Experiment I (Conditional value-at-risk)*: Set quantile level $\alpha = 0.1$; RaQL algorithm terminates in 0.263s, while RaDP terminates in 97.365s.
- *Experiment II (Optimized certainty equivalent)*: Choose exponential Utility function i.e., $u(t) = 1 - \exp(-t)$, $t \in \mathbb{R}$; RaQL algorithm terminates in 0.235s, while RaDP terminates in 38.248s.

- *Experiment III (Absolute semi-deviation)*: Set $r = 0.5$; RaQL algorithm terminates in 0.371s, while RaDP terminates in 216.3412s.

We see that our RaQL algorithm converges almost surely to the optimal Q-value, and is fast compared to traditional dynamic programming. The Fig.1. illustrates the convergence performance of RaQL for these different risk measures.

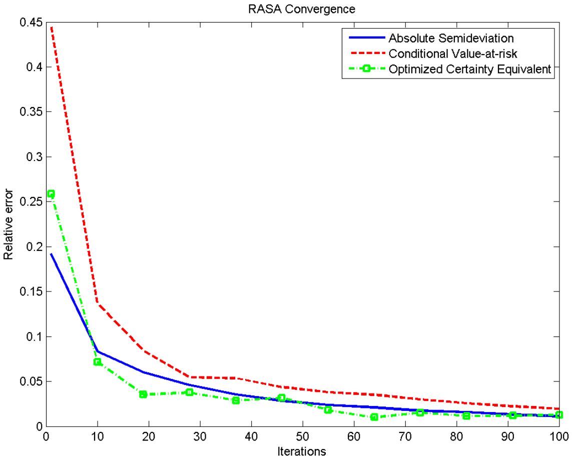

Fig. 1. Numerical experiment results

VI. Conclusion

We developed a new algorithm for risk-aware sequential decision making called Risk-aware Q-learning (RaQL). We show its convergence to the optimal Q-value with probability one and derive finite-time bounds. We also demonstrate that many common risk measures can be used within this framework. Our preliminary experimental performance is promising, as it shows fast convergence of the algorithm. In future research, we will extend our RaQL algorithm to problems with continuous state and action spaces by function approximation techniques. We will also explore methods for speeding up the risk estimation for each state.

VII. Acknowledgment

This research is funded by the National Research Foundation (NRF), Prime Ministers Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

References

[1] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[2] A. Ruszczynski and A. Shapiro, "Optimization of convex risk functions," *Mathematics of operations research*, vol. 31, no. 3, pp. 433–452, 2006.

[3] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming Modeling and Theory*. Society for Industrial and Applied Mathematics and the Mathematical Programming Society, 2009.

[4] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[5] O. A. Bardou, N. Frikha, and G. Pags, "Computation of var and cvar using stochastic approximations and unconstrained importance sampling."

[6] P. Carbonetto, M. Schmidt, and N. D. Freitas, "An interior-point stochastic approximation method and an l1-regularized delta rule," in *Advances in neural information processing systems*, 2009, pp. 233–240.

[7] A. Shapiro, "On kusuoka representation of law invariant risk measures," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 142–152, 2013.

[8] A. Ben-Tal and M. Teboulle, "An old-new concept of convex risk measures: The optimized certainty equivalent," *Mathematical Finance*, vol. 17, no. 3, pp. 449–476, 2007.

[9] C. Acerbi, "Spectral measures of risk: a coherent representation of subjective risk aversion," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1505–1518, 2002.

[10] D. Bertsimas and D. B. Brown, "Constructing uncertainty sets for robust linear optimization," *Operations research*, vol. 57, no. 6, pp. 1483–1495, 2009.

[11] R. Bellman, "A markovian decision process," DTIC Document, Tech. Rep., 1957.

[12] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

[13] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[14] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.

[15] V. S. Borkar *et al.*, "Stochastic approximation," *Cambridge Books*, 2008.

[16] V. S. Borkar, "Q-learning for risk-sensitive control," *Mathematics of operations research*, vol. 27, no. 2, pp. 294–311, 2002.

[17] W. B. Haskell and R. Jain, "A convex analytic approach to risk-aware markov decision processes," *SIAM Journal on Control and Optimization*, vol. 53, no. 3, pp. 1569–1598, 2015.

[18] A. Ruszczyński, "Risk-averse dynamic programming for markov decision processes," *Mathematical programming*, vol. 125, no. 2, pp. 235–261, 2010.

[19] N. Bäuerle and J. Ott, "Markov decision processes with average-value-at-risk criteria," *Mathematical Methods of Operations Research*, vol. 74, no. 3, pp. 361–379, 2011.

[20] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar optimization approach," in *Advances in Neural Information Processing Systems*, 2015, pp. 1522–1530.

[21] D. R. Jiang and W. B. Powell, "Risk-averse approximate dynamic programming for dynamic quantile-based risk measures," *arXiv preprint arXiv:1509.01920*, 2015.

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

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

[24] A. Shapiro and A. Kleywegt, "Minimax analysis of stochastic problems," *Optimization Methods and Software*, vol. 17, no. 3, pp. 523–542, 2002.

[25] A. Shapiro and S. Ahmed, "On a class of minimax stochastic programs," *SIAM Journal on Optimization*, vol. 14, no. 4, pp. 1237–1249, 2004.

[26] A. Shapiro, "On duality theory of convex semi-infinite programming," *Optimization*, vol. 54, no. 6, pp. 535–543, 2005.

[27] ——, "Minimax and risk averse multistage stochastic programming," *European Journal of Operational Research*, vol. 219, no. 3, pp. 719–726, 2012.

[28] M. Mahdavi, T. Yang, and R. Jin, "Online stochastic optimization with multiple objectives," *arXiv preprint arXiv:1211.6013*, 2012.

[29] W. B. Haskell, H. Xu, Q. Chao, and Y. Zhiyue, "Online risk-aware optimization," 2016.

[30] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, 2011.

[31] L. Breiman, "Probability, volume 7 of classics in applied mathematics," *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA*, 1992.

[32] N. Noyan and G. Rudolf, "Kusuoka representations of coherent risk measures in finite probability spaces," 2012.

[33] ——, "Optimization with multivariate conditional value-at-risk constraints," *Operations Research*, vol. 61, no. 4, pp. 990–1013, 2013.