Nonparametric Return Distribution Approximation for Reinforcement Learning

Tetsuro Morimura[†] TETSURO@JP.IBM.COM
Masashi Sugiyama[‡] SUGI@CS.TITECH.AC.JP
Hisashi Kashima [††] KASHIMA@MIST.I.U-TOKYO.AC.JP
Hirotaka Hachiya[‡] HACHIYA@SG.CS.TITECH.AC.JP
Toshiyuki Tanaka[‡‡] TT@I.KYOTO.AC.JP

[†]IBM Research - Tokyo, 1623-14 Shimotsuruma, Yamato-shi, Kanagawa, 242-8502, Japan
[‡]Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
[††]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[‡‡]Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto-shi, Kyoto, 606-8501, Japan

Abstract

Standard Reinforcement Learning (RL) aims to optimize decision-making rules in terms of the expected return. However, especially for risk-management purposes, other criteria such as the expected shortfall are sometimes preferred. Here, we describe a method of approximating the *distribution* of returns, which allows us to derive various kinds of information about the returns. We first show that the Bellman equation, which is a recursive formula for the expected return, can be extended to the cumulative return distribution. Then we derive a nonparametric return distribution estimator with particle smoothing based on this extended Bellman equation. A key aspect of the proposed algorithm is to represent the recursion relation in the extended Bellman equation by a simple replacement procedure of particles associated with a state by using those of the successor state. We show that our algorithm leads to a risk-sensitive RL paradigm. The usefulness of the proposed approach is demonstrated through numerical experiments.

1. Introduction

Most reinforcement learning (RL) methods attempt to find decision-making rules that maximize the *expected* return, where the return is defined as the cumulative (discounted) total of immediate rewards. Most of the theories in RL have been developed for working with the expected return as the objective function.

However, users are sometimes interested in controlling the risk. For example, since maximizing the expected return may accept rare occurrences of large negative outcomes, some users, especially those who plan to apply a RL algorithm to practical applications, prefer a policy avoiding (or constraining) small chances of suffering a large loss (Geibel & Wysotzki, 2005; Defourny et al., 2008). On another front, some users will prefer a robust criterion against outlier events, as in cases where an estimate of the expected return will be severely affected by rare occurrences of those events (Sugiyama et al., 2009).

In this paper, we describe an approach to handling various risk-sensitive and/or robust criteria in a unified manner, where the distribution of the possible returns is approximated and then the criteria are evaluated based on the approximated distribution. In order to achieve this, our approach is to first extend the Bellman equation for conditional expectations of the returns to cover conditional *cumulative distributions* of the returns, as shown in Section 3. We next describe in Section 4 a nonparametric approach with particles to solve the Bellman equations for distributions, which can be regarded as an extension of the conventional temporal-difference learning. In Section 5, we demonstrate how to apply our approach to a risk-sensitive RL scenario, by focusing on the expected shortfall, also called conditional value-at-risk (CVaR), of returns as an example of risk-sensitive criteria. In Section 6,

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

numerical experiments show that the proposed algorithms are promising for return distribution approximation and also in a risk-sensitive RL scenario.

2. Background of Value-based RL

We briefly review the framework of value-based RL and our motivation to approximate the return distributions.

2.1. Markov Decision Process (MDP)

An RL problem is usually defined on a discrete-time Markov decision process (MDP) (Bertsekas, 1995; Sutton & Barto, 1998). The MDP is defined by the quintuplet $(\mathcal{S}, \mathcal{A}, p_{\mathrm{T}}, P_{\mathrm{R}}, \pi)$, where $\mathcal{S} \ni s$ and $\mathcal{A} \ni a$ are finite sets of states and actions, respectively. The state transition probability $p_{\mathrm{T}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a function of a state s, an action a, and the successor state s_{+1}, *i.e.*, $p_{\mathrm{T}}(s_{+1}|s, a) \triangleq \Pr(s_{+1}|s, a)$.[1] Here, we describe the state s_{+k} and the action a_{+k} as a state and an action after k time-steps from the state s and the action a, respectively. An immediate reward $r \in \mathbb{R}$ is distributed according to the reward probability $P_{\mathrm{R}} : \mathbb{R} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, which is a function of r, s, a, and s_{+1}, *i.e.*, $P_{\mathrm{R}}(r|s, a, s_{+1}) \triangleq \Pr(R \leq r|s, a, s_{+1})$.[2] The policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a probability function of a given s, which defines the decision-making rule of the learning agent, *i.e.*, $\pi(a|s) \triangleq \Pr(a|s)$. There are various policy models with a value function, such as the greedy, ε-greedy, or soft-max action selection models (Sutton & Barto, 1998).

2.2. From expected returns to return distributions

Let us define the return as the following time-discounted cumulative reward with a discount rate $\gamma \in [0, 1)$,

$$\eta \triangleq \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r_{+t}, \tag{1}$$

where r_{+0} is the original r. Alternatively, the return could be defined as $\eta \triangleq \sum_{t=0}^{T} \gamma^t r_{+t}$ with a bounded T and $\gamma \in [0, 1]$ in an episodic problem. The return is observed by the learning agent with (infinite) time delay and usually is a random variable E, reflecting randomness due to p_{T}, P_{R}, and π. Once the policy is fixed, the MDP is regarded as a Markov chain $\mathrm{M}(\pi) \triangleq \{\mathcal{S}, \mathcal{A}, p_{\mathrm{T}}, P_{\mathrm{R}}, \pi, \gamma\}$. We write the (conditional) cumulative distribution functions of the returns as $P^{\pi}_{\mathrm{E}} : \mathbb{R} \times \mathcal{S} \times \mathcal{A} \times \mathcal{M} \to [0, 1]$ and $\bar{P}^{\pi}_{\mathrm{E}} : \mathbb{R} \times \mathcal{S} \times \mathcal{M} \to [0, 1]$, where $\mathcal{M}$ is a family of Markov chains,

$$P^{\pi}_{\mathrm{E}}(\eta \,|\, s, a) \triangleq \Pr(E \leq \eta \,|\, s, a, \mathrm{M}(\pi)),$$
$$\bar{P}^{\pi}_{\mathrm{E}}(\eta \,|\, s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s) \Pr(E \leq \eta \,|\, s, a, \mathrm{M}(\pi)).$$

A statistic for the return from a state (or a state-action pair) is called a value function of the state (or the state-action pair). The objective of RL is formalized as maximizing the value function over the states. Although there would be various return-statistics to be considered for the value function, in most value-based RL problems, the expected return conditioned on a state s or a state-action pair (s, a) is used for the value function, such as

$$V^{\pi}(s) \triangleq \mathbb{E}^{\pi}[\eta \,|\, s],$$
$$Q^{\pi}(s, a) \triangleq \mathbb{E}^{\pi}[\eta \,|\, s, a], \tag{2}$$

where $\mathbb{E}^{\pi}[\cdot]$ denotes the expectation over the Markov chain $\mathrm{M}(\pi)$. The function $Q^{\pi}(s, a)$ is called the state-action value or Q-value function and often serves a basis of the policy. However, decision making that relies only on information on the expected return will be insufficient for the control of the risk. These value functions are often inappropriate for the risk-sensitive criteria, as described in Section 1. In addition, the expected return is not robust against outliers, *i.e.*, these value functions can be severely affected by rare occurrences of large noises that could be included in the reward or state observations (Sugiyama et al., 2009).

Altogether, the major drawback of the ordinary RL approach follows from the fact that the approach ignores all information about the return except for the expectations. In actuality, if we have an approximation for the return distribution, this gives us access to a lot of information about the return and allows us to handle various risk-sensitivity or robustness criteria for the return. This is why we focus on the approximation of the return distribution.

2.3. Related work involving the return distribution

There are several approaches to estimating other statistics for the return distribution beyond the expected return. In the seminal paper of Dearden et al. (1998), a recursive formula for the second moment of the return is shown when the reward is defined by a deterministic function of the state-action

[1] Although to be precise it should be $\Pr(S_{+1}\!=\!s_{+1}|S\!=\!s, A\!=\!a)$ for the random variables S_{+1}, S, and A, we often write $\Pr(s_{+1}|s, a)$ for simplicity. The same simplification is used for other distributions.

[2] While $r(s, a, s_{+1})$ may often be deterministic, we assume r to be stochastic, in order to consider a more general case of RL. All results presented here are applicable to the deterministic reward case as well.

pair, $\mathbb{E}^\pi\{\eta^2|s\} = r^2 + 2\gamma V^\pi(s_{+1}) + \gamma^2\mathbb{E}^\pi\{\eta^2|s_{+1}\}$. They developed a Bayesian learning method in which the return distribution is estimated with the normal-gamma distribution. However, this approach sometimes requires numerical integration and requires extensive computation time, while our proposed algorithm does not require any numerical integration. As a similar line to this work, some *mean-variance* model-free RL algorithms were developed by Sato et al. (2001), Engel et al. (2005), and Ghavamzadeh & Engel (2007), and were successful in the variance penalized MDPs. However, these mean-variance-based algorithms assume that the return distribution is Gaussian, which may not be true in practice. In contrast, our nonparametric approach using particles has many degrees of freedom as regards the return distribution model.

3. Distributional Bellman Equation for Cumulative Return Distribution

We derive a Bellman-type recursive formula for the return distribution, comparing it with the ordinary Bellman equation for the Q-value function. From the definitions of the return (Eq. (1)) and the Q-value function (Eq. (2)), the following equation is derived (Sutton & Barto, 1998),

$$Q^\pi(s,a) \triangleq \sum_{s_{+1}\in\mathcal{S}} p_{\mathrm{T}}(s_{+1}|s,a)\Big\{\int_{r\in\mathbb{R}} r\, dP_{\mathrm{R}}(r|s,a,s_{+1}) + \gamma \sum_{a_{+1}\in\mathcal{A}} \pi(a_{+1}|s_{+1})Q^\pi(s_{+1},a_{+1})\Big\}.$$

This equation is well-known as the Bellman equation for the Q-value function. Similarly, we can derive a Bellman-type recursive formula for the conditional cumulative distribution of the returns given a state-action pair, which we call the (cumulative) *distributional* Bellman equation for the returns.

Proposition 1 *The (cumulative) distributional Bellman equation for the return η is given as*

$$\begin{aligned} &P_{\mathrm{E}}^\pi(\eta|s,a) \\ &= \sum_{s_{+1}\in\mathcal{S}}\sum_{a_{+1}\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a_{+1}|s_{+1}) \\ &\quad \times \int_{r\in\mathbb{R}} P_{\mathrm{E}}^\pi\Big(\frac{\eta-r}{\gamma}\,|\,s_{+1},a_{+1}\Big)dP_{\mathrm{R}}(r|s,a,s_{+1}). \end{aligned} \tag{3}$$

Proof: Here we use a strict notation only for the return distribution such as $\Pr(E\le\eta|s,a) := P_{\mathrm{E}}^\pi(\eta|s,a)$. From the definition of the return in Eq. (1), the recursive form with respect to the return is

$$\eta = r + \gamma\eta_{+1}.$$

The random variables $R = r$ and $E_{+1} = \eta_{+1}$ in the above equation are conditionally independent given the successor state s_{+1}, since these are distributed as $r \sim P_{\mathrm{R}}(r|s,a,s_{+1})$ and $\eta_{+1} \sim P_{\mathrm{E}}^\pi(\eta_{+1}|s_{+1},a_{+1})$, respectively, where "$x \sim P_X(x)$" means that x is distributed according to a probability function $P_X(x)$. Therefore, by considering the convolution integral with respect to r, this proposition (Eq. (3)) is proved as

$$\begin{aligned} &\Pr(E\le\eta\,|\,s,a) \\ &= \Pr(R+\gamma E_{+1}\le\eta\,|\,s,a) \\ &= \sum_{s_{+1}\in\mathcal{S}}\sum_{a_{+1}\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a_{+1}|s_{+1}) \\ &\qquad \times \int_{r\in\mathbb{R}} \Pr(\gamma E_{+1}\le\eta-r\,|\,s_{+1})dP_{\mathrm{R}}(r|s,a,s_{+1}) \\ &= \sum_{s_{+1}\in\mathcal{S}}\sum_{a_{+1}\in\mathcal{A}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a_{+1}|s_{+1}) \\ &\qquad \times \int_{r\in\mathbb{R}} \Pr\Big(E_{+1}\le\frac{\eta-r}{\gamma}\,|\,s_{+1}\Big)dP_{\mathrm{R}}(r|s,a,s_{+1}). \end{aligned}$$

□

The distributional Bellman equation for the conditional return distribution given a *state-action pair* (Eq. (3)) is directly extended to that given a *state* as

$$\begin{aligned} \bar{P}_{\mathrm{E}}^\pi(\eta|s) &= \sum_{a\in\mathcal{A}}\sum_{s_{+1}\in\mathcal{S}} \pi(a|s)p_{\mathrm{T}}(s_{+1}|s,a) \\ &\quad \times \int_{r\in\mathbb{R}} \bar{P}_{\mathrm{E}}^\pi\Big(\frac{\eta-r}{\gamma}\,|\,s_{+1}\Big)dP_{\mathrm{R}}(r|s,a,s_{+1}). \end{aligned} \tag{4}$$

These conditional return distributions are in principle evaluated by solving the distributional Bellman equations, just as the (ordinary) Bellman equation gives the Q-value function (*i.e.*, the conditional expected return given a state-action pair) as its solution.

4. Return Distribution Approximation with Nonparametric Model

To approximately solve the distributional Bellman equations, we propose a nonparametric method based on a particle smoothing (Doucet et al., 2001), called the return distribution particle smoother (RDPS), in which particles are used to approximate the return distributions. An eligibility trace technique for RDPS is also discussed.

Note that, in this section, we will discuss an approximation only for the conditional return distribution given a *state*. However, these results are directly generalized and applied to the case of the conditional return distribution given a *state-action pair*.

4.1. General view of return distribution approximation with distributional Bellman equation

A possible approach for approximating the return distribution would be to use a Monte Carlo sampling technique. However, this would require an enormous number of samples and have high computational costs since each observation of the return has multiple time delays. As an alternative, the distributional Bellman equation as given in Section 3 illustrates the connection between the neighboring-time return distributions and gives the return distributions as its solution.

For simplicity, the right-hand side of Eq. (4) is denoted as $\Pi\bar{P}^{\pi}_{\mathrm{E}}$ with the distributional Bellman operator Π as

$$\Pi\bar{P}^{\pi}_{\mathrm{E}}(\eta|s) \triangleq \sum_{a\in\mathcal{A}}\sum_{s_{+1}\in\mathcal{S}}\pi(a|s)p_{\mathrm{T}}(s_{+1}|s,a) \times \int_{R}\bar{P}^{\pi}_{\mathrm{E}}\Big(\frac{\eta-r}{\gamma}\,|\,s_{+1}\Big)dP_{\mathrm{R}}(r|s,a,s_{+1}).$$

The distributional Bellman equations are expressed as $\bar{P}^{\pi}_{\mathrm{E}} = \Pi\bar{P}^{\pi}_{\mathrm{E}}$. When a probability function $F(\eta|s)$ satisfies the distributional Bellman equations for all of the states, the function F is the solution of the equations and is equivalent to the conditional cumulative distribution function of the return given a state. However, it is hard to deal with the distributional Bellman equation in practice due to its numerous functional degrees of freedom. To address this problem, we approximate a solution in a nonparametric way using particles.

4.2. Return distribution particle smoother (RDPS) algorithm

We suppose the usage of look-up table for the states (or the state-action pairs). Each entry of the table has a number of particles[3], $\boldsymbol{v}_s = \{v_{s,1},\ldots,v_{s,K}\}$, where the subscripts s and k of $v_{s,k}$ denotes the state and the identifier, respectively. Each particle $v_{s,k}\in\mathbb{R}$ represents a return η from the state s. For simplicity, we also suppose each state has the same number K of particles. The set $\boldsymbol{V}_K \in \mathbb{R}^{|\mathcal{S}|\times K}$ is the complete set of particles $\{\boldsymbol{v}_s\}$, where $|\mathcal{S}|$ denotes the number of states. The return distribution $\bar{P}^{\pi}_{\mathrm{E}}(\eta|s)$ is approximated by a distribution of the particles $\boldsymbol{v}_s$, *i.e.*, an estimate of the cumulative probability distribution function of the return defined as

$$\hat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K) \triangleq \frac{1}{K}\sum_{k=1}^{K} I(v_{s,k}\le\eta), \tag{5}$$

where I is the indicator function, equal to 1 if $v_{s,k}\le\eta$ and otherwise 0.

A key aspect of the proposed algorithm is to represent the recursion relation in the distributional Bellman equation by extending the conventional temporal-difference learning to a particle smoothing approach, where a simple replacement procedure of particles associated with a state with those of the successor state is executed. We call this approach the Return Distribution Particle Smoother (RDPS).

Given a state s, one can generate (a, s_{+1}, r) according to the policy π, the state transition probability p_{T}, and the reward probability P_{R}. The quantity $r+\gamma\eta_{+1}$, defined from a pair (r,η_{+1}) with η_{+1} sampled from $\bar{P}^{\pi}_{\mathrm{E}}(\eta_{+1}|s_{+1})$, can be regarded as following (or being drawn from) the distribution $\Pi\bar{P}^{\pi}_{\mathrm{E}}(\eta|s)$. Thus, with independent paired samples $(r^{(1)},\eta^{(1)}_{+1}),\ldots,(r^{(N)},\eta^{(N)}_{+1})$ given a state s, this limit holds;

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} I(r^{(n)}+\gamma\eta^{(n)}_{+1}\le\eta) = \Pi\bar{P}^{\pi}_{\mathrm{E}}(\eta|s).$$

Therefore, for satisfying the distribution Bellman equation $\bar{P}^{\pi}_{\mathrm{E}} = \Pi\bar{P}^{\pi}_{\mathrm{E}}$ (Eq. (4)), the value $r+\gamma\eta_{+1}$ of the paired samples (r,η_{+1}) also has to be distributed according to the distribution $\bar{P}^{\pi}_{\mathrm{E}}(\eta|s)$,

$$r+\gamma\eta_{+1}\sim\bar{P}^{\pi}_{\mathrm{E}}(\eta|s).$$

This result suggests that an adjustment of the approximated return distribution is to allow $\hat{\bar{P}}_{\mathrm{E}}(\eta|s)$ to explain the sample (r,η_{+1}) appropriately. More specifically in the particle case, (some of) the particles $\boldsymbol{v}_s$ need to explain or contain $\{r+\gamma\boldsymbol{v}_{s_{+1}}\} = \{(r+\gamma v_{s_{+1},1}),\ldots,(r+\gamma v_{s_{+1},K})\}$ to some extent. One of the straightforward approaches would be to replace a particle $v_{s,k}$ randomly chosen from $\boldsymbol{v}_s$ by a value $r+\gamma v_{s_{+1},k'}$ as defined by the observed reward and a particle of the successor state. This approach has the desirable property of using the Kolmogorov–Smirnov statistic (distance) $D_{\mathrm{KS}}\{P(x),Q(x)\}$ for a measure of the difference of the two cumulative distribution functions $P(x)$ and $Q(x)$, (Feller, 1948),

$$D_{\mathrm{KS}}\{P(x),Q(x)\} \triangleq \sup_x\big|P(x)-Q(x)\big|.$$

Proposition 2 *Let $\boldsymbol{V}_K=\{v_{s,k}\}$ be a complete set of the particles and let all values of the particles, except for the particles of the state s, be fixed. Also let $D^{*}_{\mathrm{KS}}(s,\boldsymbol{V}_K)$ be a Kolmogorov–Smirnov statistic, $D_{\mathrm{KS}}\{\hat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K),\Pi\hat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)\}$, for which the following replacement is iterated a sufficient number of times, with the right-to-left substitution operator* $:=$,

$$v_{s,k} := r+\gamma v_{s_{+1},l}, \tag{6}$$

[3] However, our approach can be extended to continuous state space or feature vector space with neighborhood or discretization approach, *i.e.*, as long as particles around a state can be gathered, our approach will be applicable.

where k and l are integers independently drawn from the uniform distribution $\mathrm{U}(1,\mathrm{K})$ generating an integer from 1 to K, and r and s_{+1} are drawn from the model distributions P_{R} and p_{T}, respectively. Then there are bounds for the asymptotic mean and variance of $D^*_{\mathrm{KS}}(s,K)$ in terms of the particle number K as

$$\mathbb{E}\left[\lim_{K\to\infty}\sqrt{K}D^*_{\mathrm{KS}}(s,K)\right] \leq \sqrt{\frac{\pi}{2}}\ln(2),$$
$$\mathbb{V}\left[\lim_{K\to\infty}\sqrt{K}D^*_{\mathrm{KS}}(s,K)\right] \leq \frac{1}{12}\pi(\pi-6\ln^2(2)),$$

where $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote the expectation and the variance, respectively.

Proof Sketch: It can be proved that, after a sufficient time iterations for the replacement of Eq. (6), each particle value of the state s is a return drawn from $\Pi\widehat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)$. Therefore, $D^*_{\mathrm{KS}}(s,\boldsymbol{V}_K)$ can be regarded as the Kolmogorov–Smirnov statistic between an empirical distribution of K samples independently drawn from $\Pi\widehat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)$ and the (hypothesized) distribution $\Pi\widehat{\bar{P}}_{\mathrm{E}}(\eta|s;\boldsymbol{V}_K)$. With the results in Kolmogoroff (1941) and Feller (1948), the limit, $\lim_{K\to\infty}\Pr(D^*_{\mathrm{KS}}(k,\boldsymbol{V}_K)\leq z/\sqrt{K})\geq\Phi(z)$, holds for $z\geq 0$, and the limit is equal to zero for $z<0$, where

$$\Phi(z)\triangleq 1-2\sum_{\upsilon=1}^{\infty}(-1)^{\upsilon-1}\exp(-2\upsilon^2z^2).$$

The first and second order moments of $z\sim\Phi(z)$ are given as

$$\begin{aligned}\int_{z\in[0,\infty]}z\,d\Phi(z) &= \sum_{\upsilon=1}^{\infty}\int_{z\in[0,\infty]}8(-1)^{\upsilon-1}\upsilon^2z^2\exp(-2\upsilon^2z^2)dz\\ &= \sum_{\upsilon=1}^{\infty}\frac{\pi^{\frac{1}{2}}(-1)^{\upsilon-1}}{2^{\frac{3}{2}}\upsilon} = \sqrt{\frac{\pi}{2}}\ln(2),\\ \int_{z\in[0,\infty]}z^2\,d\Phi(z) &= \sum_{\upsilon=1}^{\infty}\frac{(-1)^{\upsilon-1}}{2\upsilon^2} = \frac{\pi^2}{12},\end{aligned}$$

respectively. Thus, Proposition 2 is proved. □

Proposition 2 indicates that the Kolmogorov–Smirnov statistic $D_{\mathrm{KS}}\{\widehat{\bar{P}}_{\mathrm{E}},\Pi\widehat{\bar{P}}_{\mathrm{E}}\}$ will be smaller as the number of the particles is increased. While the number of replaced particles at each iteration in Proposition 2 is assumed to be one, it will be possible to accelerate the convergence speed for learning the particles by increasing the number of replaced particles in each iteration. Here we introduce a learning rate parameter $\alpha\in[0,1]$, which defines the number of particles, N, as

$$N=\lceil\alpha K\rceil,$$

where $\lceil x\rceil\triangleq\min\{n\in\mathbb{Z}|n\geq x\}$. The resulting online algorithm, termed RDPS, is described in Table 1.

Table 1. Online algorithm for nonparametrically approximating conditional return distribution given a state

```
Return Distribution Particle Smoother (RDPS) algorithm
Given
  • a policy: $\pi(a|s)$
  • a number of particles for each state: $K$
  • a particle updating rate $\alpha \in [0,1]$
  • a discount rate for the return: $\gamma \in [0,1)$
Set
  • initial values of all particles: $\boldsymbol{V}_K \in \mathcal{R}^{|\mathcal{S}|\times K}$
  • an initial state: $s_0 \in \{1,\dots,|\mathcal{S}|\}$ $(\sim \Pr(s_0))$
For $t = 0$ to $T$ do
  (Interaction with environment)
  • choose and execute action $a_t \sim \pi(a|s)$
  • observe following state $s_{t+1}$ and reward $r_t$
  (Update particles)
  For $n = 1$ to $\lceil\alpha K\rceil$ do
    • choose index for updated particle: $p \sim \mathrm{U}(1,K)$
    • choose index for targeted particle: $q \sim \mathrm{U}(1,K)$
    • update particle of $s_t$: $v_{s_t,p} := r_t + \gamma v_{s_{t+1},q}$
  End
End
```

4.3. Eligibility trace technique for RDPS

Here, we extend the one-step distributional Bellman equation (Eq. (4)) to the multi-step t version,

$$\begin{aligned}&\bar{P}^{\pi}_{\mathrm{E}}(\eta|s)\\ &=\sum_{a,s_{+1},\dots,a_{+t-1},s_{+t}}\pi(a|s)p_{\mathrm{T}}(s_{+1}|s,a)\cdots\pi(a_{+t-1}|s_{+t-1})\\ &\qquad\qquad\times p_{\mathrm{T}}(s_{+t}|s_{+t-1},a_{+t-1})\\ &\quad\times\int_{r,\dots,r_{+t-1}}\bar{P}^{\pi}_{\mathrm{E}}\Big(\frac{\eta-\sum_{k=0}^{t-1}\gamma^kr_{+k}}{\gamma^t}\mid s_{+t}\Big)\\ &\quad\times dP_{\mathrm{R}}(r|s,a,s_{+1})\cdots dP_{\mathrm{R}}(r_{+t-1}|s_{+t-1},a_{+t-1},s_{+t})\\ &\triangleq\Pi^t\bar{P}^{\pi}_{\mathrm{E}}(\eta|s)\end{aligned}\tag{7}$$

Based on Eq. (7), we can use the particles of the various successor states, $\boldsymbol{v}_{s_{+1}},\dots,\boldsymbol{v}_{s_{+t}}$, to update the particles of the current state, $\boldsymbol{v}_s$. We simply employ the eligibility trace technique for RL with an eligibility decay rate $\lambda\in[0,1)$ (Sutton & Barto, 1998) to use a set of multi-time-step particles, instead of the one-time-step particles $\boldsymbol{v}_{s_{+1}}$, where the target distribution for updating the particles $\boldsymbol{v}_s$ is

$$\lim_{T\to\infty}(1-\lambda)\sum_{t=1}^{T}\lambda^{t-1}\Pi^t\widehat{\bar{P}}_{\mathrm{E}}(\eta|s_{+t};\boldsymbol{V}_K),$$

for a nonepisodic task. In the episodic task case, with $\lambda\in[0,1]$, the target distribution is

$$(1-\lambda)\sum_{t=1}^{T}\lambda^{t-1}\Pi^t\widehat{\bar{P}}_{\mathrm{E}}(\eta|s_{+t};\boldsymbol{V}_K)+\lambda^{T-1}\widehat{\bar{P}}_{\mathrm{E}}(\eta|s_{+T};\boldsymbol{V}_K).$$

This can be implemented easily by carrying over some of the particles chosen for the update to the next successor state. This ratio of carrying particles over is λ. This carrying-over process is repeated until there is no particle to be carried over or an absorbing state is reached in the episodic task. The proposed RDPS algorithm could be regarded as an extension of the temporal-difference learning with eligibility traces, TD(λ), for approximating the return distribution.

5. Using the Approximated Return Distribution for Risk-Sensitive RL

Now we have a means to approximately evaluate the return distributions, so we can formulate RL algorithms with any criterion defined on the basis of a return distribution. As a representative example, in this section we develop an explicit formulations of a SARSA-learning-type approach with Conditional Value at Risk (CVaR) (Rockafellar & Uryasev, 2000; Kashima, 2007) as the evaluation criterion, on the basis of the particle distributions for the approximation.

Since the *upper*-tail CVaR (CVaR$^+$) of the return for the state-action value function is defined with $c \in [0,1]$ as

$$Q^{\pi}_{\mathrm{CVaR}^+}(s,a;c) \triangleq \mathbb{E}^{\pi}\left[\eta \,|\, P^{\pi}_{\mathrm{E}}(\eta|s,a) \geq 1-c\right],$$

the policy based on this criterion will take a risk. In contrast, a policy based on the *lower*-tail CVaR, $Q^{\pi}_{\mathrm{CVaR}^-}(s,a;c) \triangleq \mathbb{E}^{\pi}[\eta \,|\, \lim_{\varepsilon\to 0^-} P^{\pi}_{\mathrm{E}}(\eta+\varepsilon|s,a) \leq c]$, will be risk averse. It is known that these risk-taking and risk-aversion strategies lead to the exploration and exploitation (robust) behaviors, respectively (Bagnell, 2004).

From Eq. (5), an estimate of CVaR with RDPS can easily be computed as, for $c \in (0,1]$,

$$\widehat{Q}_{\mathrm{CVaR}^+}(s,a;c) = \frac{1}{\lceil cK \rceil}\sum_{k=1}^{K} v_{s,k}\, I(\widehat{P}_{\mathrm{E}}(v_{s,k}|s,a) \geq 1-c),$$

or $\widehat{Q}_{\mathrm{CVaR}^+}(s,a;c=0) = \max_k\{v_{s,k}\}$. One possible approach is to use the CVaR$^+$ of the return for the behavior policy for an efficient exploitation, while the target policy (also called the estimation policy) is to maximize the expected returns. In this scenario, the off-policy learning, in which the importance sampling is often used to adjust the learning rate (Precup et al., 2000), is suitable. Our algorithm uses the importance sampling to condition the number of updated particles, *i.e.*, learning rate α. This resulting algorithm, which we call the *distributional*-SARSA-with-CVaR (or *d*-SARSA with CVaR) algorithm, can provide a practical risk-sensitive RL algorithm.

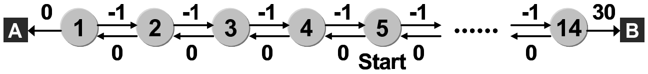

Figure 1. Task setting: There are fourteen (normal) states and two special states A and B. State 5 is a start state. Values near by arrows are rewards.

Note that, when $c = 1$, then CVaR is equivalent to the expected value, $Q^{\pi}_{\mathrm{CVaR}^+}(s,a;c=1) = Q^{\pi}(s,a)$. Accordingly, if c_t of the CVaR$^+$ at a time-step t is scheduled as monotonically increasing to 1 over time, the initial policy in learning tends to take risks and then drive exploitation, while the final policy tries to maximize the ordinary value function. Thus, this scheduling of c will be useful if a big opportunity for improvement is hidden in a distant state. However, this type of agent will probably often be trapped in states that have heavy noise in the reward. It will be effectual to combine the RDPS with the CVaR approach and the other exploitation RL algorithm such as E^3 or R-MAX (Kearns & Singh, 2002; Brafman & Tennenholtz, 2003).

6. Numerical Experiments

6.1. Task setting of 14-state MDP

In this section, we investigated the performances of the new algorithms through a simple, but very illustrative, MDP. Fig. 1 shows the settings of this MDP, which is based on (Sutton & Barto, 1998) and (Sutton et al., 2009). There are fourteen (normal) states $\{1,2,\ldots,14\}$ and two special states $\{A,B\}$. Each normal state has two actions that cause left- and right-transitions to neighboring states. The rewards in each left- or right-transitions were zero or -1, respectively, except for a transition into the right special state B, for which a large reward of $+30$ was used.

For an episodic task, the special states are absorbing (terminal) states. An episode begins in state 5 and continues until an absorbing state is reached. When the absorbing state was reached, a new episode restarts. In a nonepisodic case, both the special states were identical to the start state, *i.e.*, upon arrival to A or B, the agent is teleported to the state '5'. Episodes began in the state 5 and continued until a simulation run was terminated.

6.2. Return distribution approximation

Here, we used a fixed policy that chose the left- or right-transition action randomly with equal probability, *i.e.*, $\pi(a|s) = 0.5$, $^{\forall}a$, $^{\forall}s$. This type of MDP setting can be regarded as a random-walk problem, which is known as a useful RL benchmark to assess the prac-

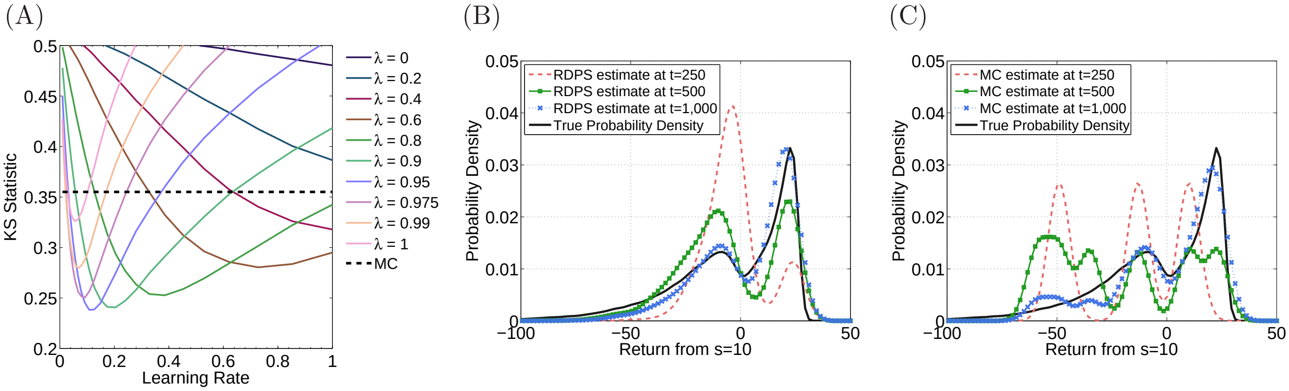

Figure 2. Evaluations of the RDPS algorithm with 1,000 particles compared with the Monte-Carlo estimation in the episodic task: (A) Average Kolmogorov-Smirnov (KS) statistics of the state $s = 10$ at 500 time-steps over 500 independent simulation runs, (B) and (C) Typical estimates at 250, 500, and 1,000 time-steps, where the particles or the instances of the returns are converted to the probability density with the normal kernel density estimator.

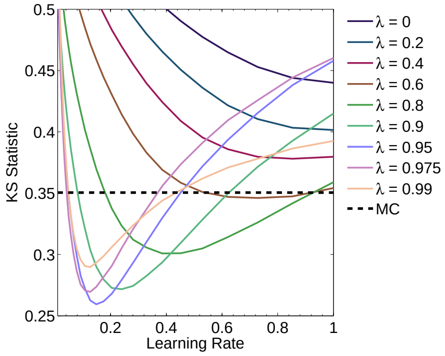

Figure 3. Average Kolmogorov-Smirnov statistics of the state $s=10$ in the non-episodic task at 500 time-steps over 500 independent simulation runs.

tical utility of RL algorithms for value function regression, and which has been used often (Sutton & Barto, 1998; Sutton et al., 2009).

We confirmed that the return distribution approximator of the RDPS with $\lambda \simeq 0.95$ worked well for both episodic and nonepisodic tasks, as shown in Figs. 2 and 3. The effect of the eligibility trace is very similar to the TD(λ) algorithm of (Sutton & Barto, 1998). Therefore, our proposed RDPS algorithm can be regarded as a natural extension of the TD(λ) learning from the expected return to the return distribution.

6.3. Policy learning

Here, we assess the performance of the (off-policy) d-SARSA-with-CVaR algorithm. We also used the (normal) d-SARSA and LSTD-SARSA[4] as the baseline algorithms, the Q-value functions of which are estimated by the RDPS and the LSTDQ(λ) (Lagoudakis & Parr, 2003), respectively. The ε-greedy selection method was used for the policy (Sutton & Barto, 1998), where the action with the highest value is selected with probability $\varepsilon + (1 - \varepsilon)/2$ and one of the other actions is selected with probability $(1-\varepsilon)/2$. The ε and the discounted rate γ were set to 0.1 and 0.98, respectively. The c parameter of RDPS with CVaR depended on the time t as $c_t = \min\{t/3000, 1\}$.

Fig. 4 shows the average performances in the episodic task over 300 independent simulation runs. This result indicates that the d-SARSA-with-CVaR algorithm as a risk-sensitive approach based on the RDPS algorithm can handle risk as discussed in Section 5.

7. Conclusion

We proposed an approach for approximating the *distribution* of returns, which allows us to handle various criteria including risk-sensitivities in a unified manner. The Bellman-type recursive formulas for the conditional cumulative probability distributions of the returns were derived. We proposed a nonparametric method for approximating the return distributions with particles. We also presented a risk-sensitive SARSA algorithm utilizing CVaR to explore effectively. The numerical experiments on simple MDPs indicated that the proposed algorithms are promising for the return distribution approximation and also in the risk-sensitive RL scenario.

Analyses of the proposed algorithms especially in terms of their convergences and empirical studies with some more challenging domains will be necessary to more deeply understand the properties and efficiency of our proposed approach to nonparametrically approximating the return distributions.

[4] Since it is better if all of the tested algorithm have on the same setting (ε-greedy policy), we apply the LSTD-SARSA, instead of the LSPI (Lagoudakis & Parr, 2003).

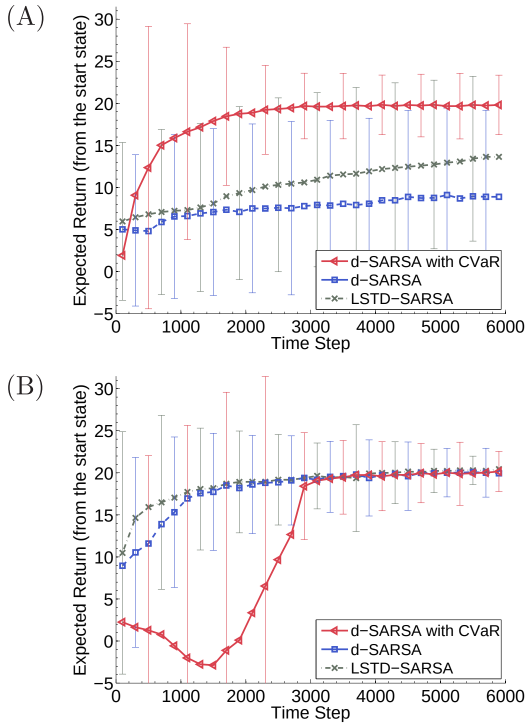

Figure 4. Average performances in the episodic task over 300 independent simulation runs: Error bar represents the standard deviation. (A) Deterministic reward setting, (B) Stochastic reward setting, where the reward of the transition to the absorbing state A is drawn from the normal distribution $N(\mu=0, \sigma=10)$ and the all other rewards are set same as those in the deterministic setting.

Acknowledgments

This work was supported by the PREDICT of the Ministry of Internal Affairs and Communications of Japan and also by the FIRST Program.

References

Bagnell, J. A. *Learning Decisions: Robustness, Uncertainty, and Approximation.* PhD thesis, Robotics Institute, Carnegie Mellon University, 2004.

Bertsekas, D. P. *Dynamic Programming and Optimal Control, Volumes 1 and 2.* Athena Scientific, 1995.

Brafman, R. I. and Tennenholtz, M. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2003.

Dearden, R., Friedman, N., and Russell, S. Bayesian Q-learning. In *National Conference on Artificial Intelligence*, pp. 761–768, 1998.

Defourny, B., Ernst, D., and Wehenkel, L. Risk-aware decision making and dynamic programming. In *NIPS 2008 Workshop on Model Uncertainty and Risk in RL*, 2008.

Doucet, A., de Freitas, N., and Gordon, N. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

Engel, Y., Mannor, S., and Meir, R. Reinforcement learning with Gaussian processes. In *International Conference on Machine Learning*, pp. 201–208, 2005.

Feller, W. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, 19(2):177–189, 1948.

Geibel, P. and Wysotzki, F. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Ghavamzadeh, M. and Engel, Y. Bayesian actor-critic algorithms. In *International Conference on Machine Learning*, pp. 297–304, 2007.

Kashima, H. Risk-sensitive learning via minimization of empirical conditional value-at-risk. *IEICE Transaction on Information and Systems*, E90-D (12):2043–2052, 2007.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49 (2-3):209–232, 2002.

Kolmogoroff, A. Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461–463, 1941.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Precup, D., Sutton, R.S., and Singh, S. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine learning*, 2000.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Physical Review E*, 2(3): 21–41, 2000.

Sato, M., Kimura, H., and Kobayahi, S. TD algorithm for the variance of return and mean-variance reinforcement learning. *The IEICE Transactions on Information and Systems (Japanese Edition)*, 16(3): 353–362, 2001.

Sugiyama, M., Hachiya, H., Kashima, H., and Morimura, T. Least absolute policy iteration for robust value function approximation. In *IEEE International Conference on Robotics and Automation*, pp. 699–704, 2009.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning.* MIT Press, 1998.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *International Conference on Machine Learning*, pp. 993–1000, 2009.