Machine Learning, 49, 267–290, 2002

Risk-Sensitive Reinforcement Learning

OLIVER MIHATSCH oliver.mihatsch@mchp.siemens.de
RALPH NEUNEIER ralph.neuneier@mchp.siemens.de
Siemens AG, Corporate Technology, Information and Communications 4, D-81730 Munich, Germany

Editor: Satinder Singh

Abstract. Most reinforcement learning algorithms optimize the *expected* return of a Markov Decision Problem. Practice has taught us the lesson that this criterion is not always the most suitable because many applications require robust control strategies which also take into account the *variance* of the return. Classical control literature provides several techniques to deal with risk-sensitive optimization goals like the so-called *worst-case* optimality criterion exclusively focusing on risk-avoiding policies or classical *risk-sensitive control*, which transforms the returns by exponential utility functions. While the first approach is typically too restrictive, the latter suffers from the absence of an obvious way to design a corresponding model-free reinforcement learning algorithm.

Our *risk-sensitive reinforcement learning* algorithm is based on a very different philosophy. Instead of transforming the return of the process, we transform the *temporal differences* during learning. While our approach reflects important properties of the classical exponential utility framework, we avoid its serious drawbacks for learning. Based on an extended set of optimality equations we are able to formulate risk-sensitive versions of various well-known reinforcement learning algorithms which converge with probability one under the usual conditions.

Keywords: reinforcement learning, risk-sensitive control, temporal differences, dynamic programming, Bellman's equation

1. Introduction

Typical reinforcement learning algorithms optimize the *expected* return of a Markov Decision Problem. However, this is not always the most suitable optimality criterion in practice. Many applications require robust control strategies which also take into account the *variance* of the return: especially the risk that the return of a specific realization of the process happens to be considerably worse than the mean value is of a living interest for many applications.

Classical control literature provides several techniques to deal with risk-sensitive optimization goals (see Coraluppi, 1997, for a nicely written overview). One approach is the so-called *worst-case* optimality criterion which exclusively focuses on risk-avoiding policies. A policy is considered to be optimal if its worst-case return is superior. In most real world applications this approach is too restrictive because it fully takes into account very rare events (that in practice never happen). For example, consider an asset manager typically interested not only in maximizing the return of a portfolio but also in reducing its variance. If the asset manager invested according to the worst-case criterion, she would never buy any risky assets like stocks due to the positive probability of a loss. Heger (1994a) developed a reinforcement learning algorithm for the *worst-case* criterion. In practice, his algorithm

is less pessimistic than the pure worst-case criterion, because extremely rare events, which do not occur during the finite training time, will have no effect on the policy.

The second approach, which makes use of *exponential utility functions*, is the most popular one in control theory. There is also related work in the AI literature (Koenig & Simmons, 1994). The idea is to transform the cumulative returns by exponential utility functions and seek optimal policies with respect to this utility measure (Howard & Matheson, 1972). It can be shown that this kind of *risk-sensitive control* interpolates in some sense between the usual expected return and the above mentioned worst-case criterion. As we will see later, there is a close relationship of this methodology to the widely used Markowitz optimization (Elton & Gruber, 1995) to construct efficient portfolios in finance. There are several key problems which prevent this theory from being integrated into machine learning algorithms. First, optimal policies for infinite horizon discounted problems are in general not stationary. Second, it is not possible to handle non-deterministic cost structures (e.g. the return of a single trade at the stock market can not be quantified before prices are determined at the next time step). Most important, there is no obvious way to design a corresponding model-free reinforcement learning algorithm mainly because of the inappropriate structure of the corresponding optimality equations.

Our *risk-sensitive reinforcement learning* algorithm is based on a very different philosophy. Instead of transforming the cumulative return of the process as in utility theory, we transform the *temporal differences* (so-called TD-errors) which play an important role during the procedure of learning the value or Q-function. While also interpolating between the expected and the worst-case approach, we avoid the serious drawbacks of the exponential utility concept with respect to learning. Based on an extended set of optimality equations we are able to formulate risk-sensitive versions of various well-known reinforcement learning algorithms (like Q-learning, TD-learning) which provably converge under the usual conditions. Our new algorithms include the classical ones as special cases. It turns out, that the already known and widely used reinforcement algorithms only require few minor changes to transform these algorithms into risk-sensitive versions.

Our work is based on the theory of dynamic programming (Bellman & Dreyfus, 1962; Bertsekas, 1995; Puterman, 1994) and reinforcement learning (e.g. Bertsekas & Tsitsiklis, 1996, among others). We gratefully acknowledge the work on worst-case dynamic programming and its reinforcement version of Heger (1994a) and Littman and Szepesvári (1996) because they provided the solid foundation from where we started to formulate our theory. A further valuable source was the compactly written overview of classical risk-sensitive control by Coraluppi (1997).

The following part of the paper is organized in six sections. After recalling the basic results of traditional risk-neutral dynamic programming (Section 2), we review the basic properties of worst-case control (Section 3) and classical risk-sensitive control (Section 4). This constitutes the basis for our formulation of a new framework for risk-sensitive control (Section 5) which is subsequently analyzed with respect to limiting behavior and optimality. Using this theory, it is straightforward to formulate several versions of risk-sensitive reinforcement algorithms and to show that these algorithms converge under the usual assumptions (Section 6). The final section summarizes the results and poses some new challenging questions for future work. The appendix contains the necessary proofs.

The main contribution of the present paper are the following.

1. We provide a new theory of risk-sensitive control,
2. formulate reinforcement learning algorithms within this framework which require only minor changes of already known and widely used algorithms, and
3. give the corresponding convergence proofs.
4. We bypass the awkward obstacles for learning of the worst-case criterion and the exponential utility approach.

2. Classical risk-neutral control

We consider discrete time Markov decision problems with finite state and action spaces S and U, respectively. The set of admissible actions in state $i \in S$ is denoted by $U(i) \subset U$. Whenever the system is in state $i \in S$ and action $u \in U(i)$ is taken, the system moves to a successor state $j \in S$ according to the transition probabilities $p_{ij}(u)$. Each transition is associated with an immediate reward $g_{ij}(u)$. Let Π denote the set of stationary policies, i.e. each $\pi \in \Pi$ maps states into actions such that $\pi(i) \in U(i)$.

In the sequel we focus on MDPs which evolve on an infinite time horizon and where future rewards are discounted. However, this is only to simplify the presentation of the topic. Similar results will hold for finite horizon problems as well as for undiscounted MDPs with absorbing states. Moreover, our results can be easily extended to the case where the immediate rewards $g_{ij}(u)$ are random.

The commonly used risk-neutral objective is to compute (or learn) control actions u_t to be taken at time t so as to

$$\text{maximize } E\left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(u_t)\right). \tag{1}$$

Here, $\gamma \in [0, 1)$ denotes the discount factor. Furthermore, $E(\cdot)$ stands for the expectation with respect to the states i_t of the Markov process. The above performance criterion evaluates a given control strategy with respect to the expected total reward it generates while interacting with the system. However, specific realizations of the given process may significantly deviate from this mean value. Risk-neutral control does not consider the risk (or the chance) of being significantly worse (or better) than the expected mean.

We briefly review some of the most important theoretical results for risk-neutral control (see Bertsekas, 1995, Puterman, 1994).

1. The maximum of the above performance criterion (1) can be achieved within the class Π of stationary and deterministic policies.
2. Let us fix a policy $\pi \in \Pi$ and denote the *value* (*reward-to-go*) of state i under policy π by

$$\bar{J}^{\pi}(i) := E\left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(\pi(i_t)) \,\middle|\, i_0 = i\right). \tag{2}$$

If we initialize the system at state i and operate it under policy π, then $\bar{J}^\pi(i)$ is equal to the *expectation* of the sum of discounted future rewards. The value function $\bar{J}^\pi(i)$ is the unique solution of the following system of linear equations

$$\bar{J}^\pi(i) = \sum_{j \in S} p_{ij}(\pi(i))[g_{ij}(\pi(i)) + \gamma \bar{J}^\pi(j)] \quad \forall i \in S. \tag{3}$$

3. The optimal value function $\bar{J}^*(i) = \max_{\pi \in \Pi} \bar{J}^\pi(i)$ is the unique solution of *Bellman's* optimality equation

$$\bar{J}^*(i) = \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u)[g_{ij}(u) + \gamma \bar{J}^*(j)] \quad \forall i \in S. \tag{4}$$

If N is the cardinality of the state space, the Bellman equation is a system of N nonlinear equations in the N unknowns $\bar{J}^\pi(i)$. Once the optimal reward-to-go $\bar{J}^*(\cdot)$ is available, an optimal policy π^* is given by

$$\pi^*(i) = \arg \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u)[g_{ij}(u) + \gamma \bar{J}^*(j)]. \tag{5}$$

The large majority of reinforcement learning algorithms (e.g. TD(λ), Sutton, 1988 or Q-learning, Watkins, 1989) has been designed for risk-neutral control problems based on the expected cumulative reward criterion. Furthermore, there is a growing number of convergence proofs for such algorithms for both tabular and parametric representations (see Bertsekas and Tsitsiklis, 1996, for an overview).

3. Worst-case control

Alternatively, Heger (1994a) and Littman and Szepesvári (1996) proposed learning methods for a performance criterion which exclusively focuses on risk-avoiding policies: the so-called *worst-case* or *minimax* criterion. This approach is also discussed at length in the control literature (Basar & Bernhard, 1995; Coraluppi & Marcus, 1999; Coraluppi, 1997).

The objective is to learn optimal control actions u_t so as to

$$\text{maximize} \inf_{\substack{i_0, i_1, \ldots \\ p(i_0, i_1, \ldots) > 0}} \left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(u_t) \right). \tag{6}$$

This criterion evaluates a given control strategy with respect to its "worst case scenario", even though it may be very unlikely. The probability with which each trajectory $i_0, i_1, \ldots$ occurs is significant only to the extent that it is zero or nonzero. This approach is too pessimistic for most practical applications which usually results in a very low average performance.

To facilitate the understanding of the following sections, we briefly summarize the main results of worst-case control theory (see Coraluppi, 1997).

1. Unfortunately, the maximization of the worst-case criterion usually involves *time-dependent* policies. Stationary policies are suboptimal in general. However, if we restrict the optimization to the class of stationary policies, we end up at similar optimality equations compared with the risk-neutral approach.
2. Let us fix a *stationary* policy $\pi \in \Pi$ and denote the worst-case *value* (*reward-to-go*) of state i under policy π by

$$\underline{J}^{\pi}(i) := \inf_{\substack{i_0, i_1, \ldots \\ p(i_0, i_1, \ldots) > 0}} \left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(\pi(i_t)) \,\middle|\, i_0 = i \right). \tag{7}$$

   If we initialize the system at state i and operate it under policy π, then $\underline{J}^{\pi}(i)$ is equal to the worst possible sum of discounted future rewards. The value function $\underline{J}^{\pi}(i)$ is the unique solution of the following system of nonlinear equations:

$$\underline{J}^{\pi}(i) = \min_{\substack{j \in S \\ p_{ij}(\pi(i)) > 0}} [g_{ij}(\pi(i)) + \gamma \underline{J}^{\pi}(j)] \quad \forall i \in S. \tag{8}$$

3. The optimal value function $\underline{J}^{*}(i) = \max_{\pi \in \Pi} \underline{J}^{\pi}(i)$ is the unique solution of the optimality equation

$$\underline{J}^{*}(i) = \max_{u \in U(i)} \min_{\substack{j \in S \\ p_{ij}(u) > 0}} [g_{ij}(u) + \gamma \underline{J}^{*}(j)] \quad \forall i \in S. \tag{9}$$

   If N is the cardinality of the state space, the Bellman equation is system of N equations in the N unknowns $\underline{J}^{\pi}(i)$. Once the optimal reward-to-go $\underline{J}^{*}(\cdot)$ is available, an optimal policy π^{*} in the worst-case sense is given by

$$\underline{\pi}^{*}(i) = \arg \max_{u \in U(i)} \min_{\substack{j \in S \\ p_{ij}(u) > 0}} [g_{ij}(u) + \gamma \underline{J}^{*}(j)]. \tag{10}$$

Heger (1994a) presented a Q-learning algorithm for worst-case control. Unfortunately, his approach is restricted to tabular value function representations. There is no obvious way to extent his algorithm to parametric representations, because one has to guarantee that the worst-case Q-function is never underestimated during learning.

4. Risk-sensitive control based on exponential utilities

The most popular approach to incorporate risk sensitivity into objective functions makes use of utility theory. The main idea is to transform the cumulative returns by appropriate utility functions and seek optimal policies with respect to this utility measure (Howard & Matheson, 1972; Pratt, 1964; Coraluppi, 1997; Koenig & Simmons, 1994). More specifically, we consider exponential utility functions of the form $\exp(\beta z)$, where the parameter β controls

the desired risk-sensitivity. In this context, the objective function takes the form

$$\text{maximize } \frac{1}{\beta} \log E\left(\exp\left(\beta \sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(u_t)\right)\right). \tag{11}$$

The utility approach is motivated and justified by the general axiomatic foundation of the utility theory (von Neumann & Morgenstern, 1953). We note the following important properties of the exponential utility criterion (Coraluppi & Marcus, 1999; Coraluppi, 1997). It turns out, that risk-sensitive control based on the exponential utility criterion contains the risk-neutral control as a special and worst-case control as a limiting case.

1. A straightforward Taylor expansion of the exp and log terms of Eq. (11) yields

$$\frac{1}{\beta} \log E(\exp(\beta Z)) = E(Z) + \frac{\beta}{2} Var(Z) + \mathcal{O}(\beta^2).$$

   Thus, the objective (11) reduces to the risk-neutral objective (1) for $\beta \to 0$. Variability is penalized if $\beta < 0$ and enforced otherwise. Therefore, the objective is risk-averse for $\beta < 0$ and risk-seeking for $\beta > 0$. This is closely related to the Markowitz (Elton & Gruber, 1995) approach used in finance for the construction of efficient portfolios.
2. In the case of *finite horizon* MDPs the exponential utility objective (11) converges to the worst-case objective (6) in the large risk limit $\beta \to -\infty$. No such result is available for *infinite horizon* problems.

Unfortunately, there are several key problems which prevent this classical risk-sensitive control theory from being integrated into machine learning algorithms.

1. Optimal policies are time-dependent in general. This complicates the computation of optimal controls in the infinite horizon setting.
2. There is no effective methodology to handle problems with a nondeterministic reward structure. Optimality equations in the spirit of Eqs. (4) and (9) only hold if $g_{ij}(u)$ does not depend on j, which is a serious restriction for many real world problems.
3. Even if we restrict ourselves to the case of deterministic rewards $g_i(u)$, the optimality equations do not give rise to model-free reinforcement learning algorithms in the spirit of TD(0) or Q-learning. The structure of the corresponding optimality equations is of the form

$$\tilde{J}^{\pi}_{\beta,t}(i) = \gamma^t g_i(\pi(i)) + \frac{1}{\beta} \log \sum_{j \in S} p_{ij}(\pi(i)) \exp\left(\beta \tilde{J}^{\pi}_{\beta,t+1}(j)\right) \quad \forall i \in S. \tag{12}$$

   Here, $\tilde{J}^{\pi}_{\beta,t}(i)$ denotes the reward-to-go in the risk-sensitive sense when the system starts from state i at time t and follows policy π. A model-free (TD(0)-like) learning algorithm usually relies on single sample unbiased estimates of the right hand side of the above equation, which are not available in this case due to the log-term.

The exponential utility approach constitutes the most popular and best analyzed risk-sensitive control framework in the literature, but there remain serious drawbacks which prevent the formulation of corresponding reinforcement learning algorithms. Therefore, we present a different framework which directly leads to a risk-sensitive learning methodology.

5. A new framework for risk-sensitive control

We now formulate our risk-sensitive control framework. Instead of transforming the total return of the process as in the exponential utility approach, we will transform the *temporal differences* that occur during learning. This approach directly leads to a new family of corresponding optimality equations.

Let $\kappa \in (-1, 1)$ be a scalar parameter which we use to specify the desired risk-sensitivity. We define the transformation function

$$\mathcal{X}^\kappa : x \mapsto \begin{cases} (1-\kappa)x & \text{if } x > 0, \\ (1+\kappa)x & \text{otherwise.} \end{cases} \tag{13}$$

Now, let us fix a stationary policy π and define the corresponding value function J_κ^π implicitly as the solution of the system of (defining) equations (compare Eq. (3))

$$0 = \sum_{j \in S} p_{ij}(\pi(i)) \mathcal{X}^\kappa \left(g_{ij}(\pi(i)) + \gamma J_\kappa^\pi(j) - J_\kappa^\pi(i) \right). \tag{14}$$

(We will shortly present a theorem assuring that a unique solution of the above equation exists). Even though $J_\kappa^\pi(i)$ is defined implicitly rather than explicitly (compare Eqs. (2) and (7)), we interpret it as the risk-sensitive *reward-to-go* if we start at state i and follow policy π. A stationary policy π^* is considered to be optimal in the risk-sensitive sense if

$$J_\kappa^{\pi^*}(i) \geq J_\kappa^\pi(i) \quad \forall \pi \in \Pi, i \in S.$$

Note that the defining equation (14) reduces to the risk-neutral policy evaluation equation. (3), if $\kappa = 0$. Our framework therefore contains the risk-neutral criterion as a special case. If we choose κ to be positive, then we overweight negative temporal differences

$$g_{ij}(\pi(i)) + \gamma J_\kappa^\pi(j) - J_\kappa^\pi(i) < 0$$

with respect to positive ones. Loosely speaking, we overweight transitions to successor states where the immediate return $g_{ij}(u)$ happened to be smaller than in the average. On the other hand, we underweight transitions to states that promise a higher return than in the average. In other words, the objective function is risk-avoiding if $\kappa > 0$ and risk-seeking if $\kappa < 0$. Furthermore, the risk-sensitive value function J_κ^π converges towards the minimax value function $\underline{J}^\pi$ in the large risk limit $\kappa \to 1$. Analogously, the risk-seeking limit $\kappa \to -1$

constitutes a very optimistic measure of the process where we assume that, for all possible next states, the one that happens is the one that is the best for us. The following theorem states this more formally.

Theorem 1 (*limiting behavior*)**.** *For each* $\kappa \in (-1, 1)$ *there is a unique solution* J_κ^π *of the defining system of equation* (14)*. Thus, the risk-sensitive value function is well defined. We have the limiting properties for all* $i \in S$

$$J_0^\pi(i) = \bar{J}^\pi(i) = E\left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(\pi(i_t)) \,\middle|\, i_0 = i\right), \tag{15}$$

$$\lim_{\kappa \to 1} J_\kappa^\pi(i) = \underline{J}^\pi(i) = \inf_{\substack{i_0, i_1, \ldots \\ p(i_0, i_1, \ldots) > 0}} \left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(\pi(i_t)) \,\middle|\, i_0 = i\right), \tag{16}$$

$$\lim_{\kappa \to -1} J_\kappa^\pi(i) = \sup_{\substack{i_0, i_1, \ldots \\ p(i_0, i_1, \ldots) > 0}} \left(\sum_{t=0}^{\infty} \gamma^t g_{i_t i_{t+1}}(\pi(i_t)) \,\middle|\, i_0 = i\right). \tag{17}$$

All proofs will be provided in the appendix.

The theorem shows that the new risk-sensitive control framework and the exponential utility approach share the same limiting behavior (compare Section 4). Both methods interpolate between the risk-neutral and the worst-case criterion. However, the flavor of the new framework is different because the value of a given policy is defined implicitly as the solution of a certain equation rather than explicitly in the spirit of Eqs. (2) and (7). This "inconvenience" turns out to be advantageous for learning.

In order to shed some light on the behavior of the risk-sensitive value function J_κ^π at intermediate values of κ, we study the following simple

Example 1. Consider the simple 2-state MDP given by figure 1. At state 0 we have two possible control options, "stay" or "move". If we choose to "stay", than we receive an immediate reward of 0 for sure. In contrast, the action "move" will bring us to state 1, where we have the chance to collect future rewards of 1 at subsequent time steps. However, there is a (small) loss probability of θ that we will have to pay the cost $\rho \geq 0$ and end up at state 0, again.

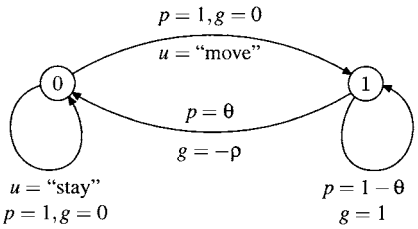

Figure 1. A simple 2-state MDP. Transitions are indicated by arrows. The arrow labels u, p, g stand for the corresponding control action, transition probability, and immediate reward, respectively.

After some simple but lengthy computation which we leave to the reader we get the following risk-sensitive value function J_κ^π as a solution of the defining equation (14):

$$J_\kappa^{\text{stay}}(0) = 0, \tag{18}$$

$$J_\kappa^{\text{move}}(0) = \frac{\gamma}{1-\gamma} \cdot \frac{(1-\theta)(1-\kappa) - \rho\theta(1+\kappa)}{(1-\theta)(1-\kappa) + (1+\gamma)\theta(1+\kappa)}. \tag{19}$$

We conclude that the policy "move" is optimal, if $J_\kappa^{\text{move}}(0) \geq 0$, i.e. if

$$\rho \leq \frac{1-\theta}{\theta} \cdot \frac{1-\kappa}{1+\kappa}. \tag{20}$$

This amounts to the following: it is optimal to "move", if the cost ρ does not exceed a threshold which is monotonically decreasing with respect to both, the loss probability θ and the risk parameter κ. In the extreme cases $\theta = 1$ (losses are inevitable), or $\kappa = 1$ (worst-case optimality criterion), "moving" is suboptimal, unless the cost ρ vanishes. Conversely, if the losses are impossible ($\theta = 0$), or if we use an extremely risk-seeking optimality criterion ($\kappa = -1$), than the above threshold becomes unbounded, i.e. "moving" becomes the optimal action for all finite costs ρ.

Our next theorem provides some useful estimates, in order to gain deeper insight into the nature of the risk-sensitive criterion for intermediate risk parameters κ. For sake of brevity we will restrict these considerations to the more interesting risk-averse case $\kappa \geq 0$. Analogous estimates hold for $\kappa \leq 0$.

We need some additional notation. Let $\bar{T}^\pi$ and $\underline{T}^\pi$ denote the dynamic programming operators corresponding to the risk-neutral and worst-case criterion, respectively. In particular, both operators acting on the space of value function are defined as

$$\begin{aligned} \bar{T}^\pi[J](i) &= \sum_{j \in S} p_{ij}(\pi(i))(g_{ij}(\pi(i)) + \gamma J(j)), \\ \underline{T}^\pi[J](i) &= \min_{\substack{j \in S \\ p_{ij}(\pi(i))>0}} (g_{ij}(\pi(i)) + \gamma J(j)). \end{aligned}$$

Using this shorthand notation, we can rewrite the policy evaluation equations (3) and (8) as $\bar{T}^\pi[\bar{J}^\pi] = \bar{J}^\pi$ and $\underline{T}^\pi[\underline{J}^\pi] = \underline{J}^\pi$, respectively. For each state i we define $N_\kappa^\pi(i)$ to be the set of all feasible successor states j under policy π for which the corresponding temporal difference (i.e. the argument of $\mathcal{X}^\kappa(\cdot)$ in Eq. (14)) attains its minimum:

$$\begin{aligned} N_\kappa^\pi(i) := \Bigg\{ j \in S \Bigg| & p_{ij}(\pi(i)) > 0 \text{ and } g_{ij}(\pi(i)) + J_\kappa^\pi(j) \\ & = \min_{\substack{j' \in S \\ p_{ij}(\pi(i))>0}} \left[g_{ij'}(\pi(i)) + J_\kappa^\pi(j') \right] \Bigg\}. \end{aligned}$$

Furthermore, let

$$p_\kappa^\pi(i) = \sum_{j \in N_\kappa^\pi(i)} p_{ij}(\pi(i))$$

be the probability of transitions to such successor states. The quantity $p_\kappa^\pi(i)$ can be interpreted as the probability of worst-case transitions from state i under policy π, where the future rewards are estimated by $J_\kappa^\pi(\cdot)$. The following theorem gives some estimates which clarify how changes of the the risk parameter κ affect the risk-sensitive value function.

Theorem 2 (*limiting behavior (cont'd)*)**.** *Let $\pi \in \Pi$ be a policy. For each $\kappa \in [0, 1)$ and state $i \in S$ we have*

$$\underline{J}^\pi(i) \le J_\kappa^\pi(i) \le \bar{J}^\pi(i). \tag{21}$$

Furthermore, there are quantities $\xi_\kappa^\pi(i)$ satisfying

$$0 < p_\kappa^\pi(i) \le \xi_\kappa^\pi(i) \le 1,$$

such that for each state $i \in S$

$$\begin{aligned} 0 \le \bar{T}^\pi\big[J_\kappa^\pi\big](i) - J_\kappa^\pi(i) &= \frac{2\kappa\xi_\kappa^\pi(i)}{1-\kappa}\big(J_\kappa^\pi(i) - \underline{T}^\pi\big[J_\kappa^\pi\big](i)\big) \\ &\le \frac{2\kappa\xi_\kappa^\pi(i)}{1-\kappa}(\bar{J}^\pi(i) - \underline{J}^\pi(i)), \end{aligned} \tag{22}$$

$$\begin{aligned} 0 \le J_\kappa^\pi(i) - \underline{T}^\pi\big[J_\kappa^\pi\big](i) &= \frac{1-\kappa}{2\kappa\xi_\kappa^\pi(i)}\big(\bar{T}^\pi\big[J_\kappa^\pi\big](i) - J_\kappa^\pi(i)\big) \\ &\le \frac{1-\kappa}{2\kappa\xi_\kappa^\pi(i)}(\bar{J}^\pi(i) - \underline{J}^\pi(i)). \end{aligned} \tag{23}$$

The above inequalities involve a number of interesting quantities.

1. The residuals $\bar{T}^\pi[J_\kappa^\pi](i) - J_\kappa^\pi(i)$ and $J_\kappa^\pi(i) - \underline{T}^\pi[J_\kappa^\pi](i)$ tell us how far the risk-sensitive value $J_\kappa^\pi(i)$ is away from its risk-neutral and worst-case counterparts $\bar{J}^\pi$ and $\underline{J}^\pi$, respectively.
2. Both residuals are bounded above by terms which depend on the difference $\bar{J}^\pi(i) - \underline{J}^\pi(i)$ between the expected and the worst-case cumulative rewards under policy π. This quantity measures the amount of risk inherent in our MDP at hand.
3. The quantity $\xi_\kappa^\pi(i)$ is by definition an upper bound on the probability $p_\kappa^\pi(i)$ of worst-case transitions. A higher probability $p_\kappa^\pi(i)$ of worst-case transitions leads to a smaller worst-case residual $J_\kappa^\pi(i) - \underline{T}^\pi[J_\kappa^\pi](i)$ (Eq. (23)) which implies that the risk-averse reward-to-go J_κ^π is closer to the worst-case value function $\underline{J}^\pi$.

The following theorems provide extensions of two standard results of the dynamic programming theory within the risk-sensitive setting.

Theorem 3 (*policy improvement*). *Let $\kappa \in (-1, 1)$. Let π and π' be stationary policies such that*

$$\pi'(i) = \arg\max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa} \left(g_{ij}(u) + \gamma J_{\kappa}^{\pi}(j) - J_{\kappa}^{\pi}(i) \right). \tag{24}$$

Then we have

$$J_{\kappa}^{\pi'}(i) \geq J_{\kappa}^{\pi}(i) \quad \forall i \in S. \tag{25}$$

Furthermore, if π is not optimal, than strict inequality holds in the above equation for at least one state $i \in S$.

The theorem shows that we can improve a given policy π in the risk-sensitive sense by a simple maximization process involving the value function J_{κ}^{π}. This constitutes the theoretical justification for a risk-sensitive counterpart of the well-known policy iteration algorithm.

Now, we show that optimal stationary policies exist in the risk-sensitive sense. The corresponding optimal value function is unique and can be obtained as the solution of a risk-sensitive counterpart to Bellman's optimality equation.

Theorem 4 (*optimal policies*). *For each $\kappa \in (-1, 1)$ there is a unique optimal value function*

$$J_{\kappa}^{*}(i) = \max_{\pi \in \Pi} J_{\kappa}^{\pi}(i) \quad \forall i \in S,$$

which satisfies the optimality equation

$$0 = \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_{\kappa}^{*}(j) - J_{\kappa}^{*}(i)) \quad \forall i \in S. \tag{26}$$

Furthermore, a policy π^* is optimal if and only if

$$\pi^*(i) = \arg\max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_{\kappa}^{*}(j) - J_{\kappa}^{*}(i)). \tag{27}$$

Now, we introduce the concept of Q-functions in the risk-sensitive context. It turns out that the important risk-neutral results also carry over to the risk-sensitive case. Given the risk parameter κ we define the Q-functions $Q_{\kappa}^{\pi}(i, u)$, i.e. the value of applying action u in state i and following policy π thereafter, as the solution of the following system of equations.

$$0 = \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa} \left(g_{ij}(u) + \gamma J_{\kappa}^{\pi}(j) - Q_{\kappa}^{\pi}(i, u) \right) \quad \forall i \in S, u \in U(i) \tag{28}$$

Our definition relies on the fact that a unique solution of Eq. (28) exists.

Theorem 5 (*Q-function*). *For each $\kappa \in (-1, 1)$ there is a unique solution Q_{κ}^{π} of the defining system of equations* (28). *Thus, the Q-function Q_{κ}^{π} is well defined.*

The intuitive property $Q^\pi_\kappa(i, \pi(i)) = J^\pi_\kappa(i)$ directly follows from the above uniqueness statement, since by definition both quantities satisfy Eq. (28) for actions $u = \pi(i)$ (also compare Eq. (14)).

Given the optimal value function J^*_κ the optimal Q-function Q^*_κ, i.e. the value of applying action u in state i and following an optimal policy thereafter is defined as the solution of

$$0 = \sum_{j \in S} p_{ij}(u) \mathcal{X}^\kappa (g_{ij}(u) + \gamma J^*_\kappa(j) - Q^*_\kappa(i, u)) \quad \forall i \in S, u \in U(i) \tag{29}$$

which is essentially the same equation as (28) except that J^π_κ has been changed to J^*_κ. We have the following risk-sensitive optimality equations for optimal Q-functions.

Theorem 6 (*optimal Q-function*)**.** *The optimal Q-function Q^*_κ is the unique solution of the optimality equation*

$$0 = \sum_{j \in S} p_{ij}(u) \mathcal{X}^\kappa \Big(g_{ij}(u) + \gamma \max_{v \in U(j)} Q^*_\kappa(j, v) - Q^*_\kappa(i, u) \Big) \quad \forall i \in S, u \in U(i). \tag{30}$$

Furthermore, a policy π^* is optimal if and only if

$$\pi^*(i) = \arg \max_{u \in U(i)} Q^*_\kappa(i, u). \tag{31}$$

We finish this section with a couple of remarks concerning some aspects of the theorem proofs. Analogous to the classical dynamic programming theory, the proofs rely on the *contraction property* of certain operators which act on value functions or Q-functions and which are closely related to various forms of Bellman's equation (cf. Appendix A.1, Lemma 2). The contraction property, which can be proven using the fact that the derivative of the transformation function $\mathcal{X}^\kappa$ is bounded, holds for all interior risk parameters $|\kappa| < 1$. The contraction is strongest for the risk-neutral operator ($\kappa = 0$) and diminishes with increasing risk sensitivity ($|\kappa| \to 1$). This will affect the convergence rate of the risk-sensitive learning algorithms (to be defined in the next section). Convergence will slow down with increasing risk sensitivity.

6. Risk-sensitive reinforcement learning

We now formulate risk-sensitive versions of two well-known reinforcement learning algorithms, TD(0) (Sutton, 1988) and Q-learning (Watkins, 1989). This task turns out to be relatively straightforward, since the defining equation (14) for J^π_κ as well as the risk-sensitive optimality equation (30) for Q-functions are both of the form $E(\ldots) = 0$. Note that the classical exponential utility approach (Eq. (12)) does not share this appealing property.

First, let us consider *risk-sensitive* TD(0). We fix a policy π and a risk parameter κ. Let $\hat{J}$ be a (tabular) approximation of the risk-sensitive value function J^π_κ. Our goal is to formulate a simulation based stochastic algorithm to tune the approximation $\hat{J}$ such that it converges to J^π_κ.

Let $(i_0, i_1, i_2, \ldots)$ be the sequence of states that we obtain while interacting with the system and let $\hat{J}_t$ denote the value function approximation available after the t-th time step. The risk-sensitive TD(0) algorithm updates $\hat{J}_t$ according to

$$\begin{aligned} \hat{J}_t(i) &= \hat{J}_{t-1}(i) + \sigma_{t-1}(i)\mathcal{X}^\kappa\left(g_{i_{t-1}i_t}(\pi(i_{t-1})) + \gamma \hat{J}_{t-1}(i_t) - \hat{J}_{t-1}(i_{t-1})\right), \\ \sigma_{t-1}(i_{t-1}) &= \sigma_{t-1} > 0, \\ \sigma_{t-1}(i) &= 0 \quad \text{if } i \neq i_{t-1}, \end{aligned} \tag{32}$$

where the stepsizes $\sigma_{t-1}(i)$ are defined to be nonzero only for the current state i_{t-1}.

Our update rule (32) is very similar to the original risk-neutral TD(0). The only difference between them is in the weighting function $\mathcal{X}^\kappa$ that transforms the temporal differences $g_{i_{t-1}i_t}(\pi(i_{t-1})) + \gamma \hat{J}_{t-1}(i_t) - \hat{J}_{\kappa,t-1}(i_{t-1})$. For $\kappa = 0$, the risk-sensitive TD(0) reduces to the classical risk-neutral one.

We have the following important convergence result saying that our risk-sensitive version of TD(0) converges under the same generic conditions than the original risk-neutral one.

Theorem 7 (*convergence*). *Let $\kappa \in (-1, 1)$. Consider the risk-sensitive* TD(0)*-algorithm, as described by equation* (32). *If the stepsizes $\sigma_t(i)$ are nonnegative and satisfy*

$$\sum_{t=0}^{\infty} \sigma_t(i) = \infty \quad \sum_{t=0}^{\infty} (\sigma_t(i))^2 < \infty \quad \forall i \in S$$

then $\hat{J}_t(i)$ converges to $J_\kappa^\pi(i)$ for all i, with probability 1.

The stepsize conditions imply that we need to visit each state infinitely often during learning.

Let us now consider *risk-sensitive Q-learning*. With $\hat{Q}_\kappa(i, u)$ we denote a tabular approximation of the optimal Q-function Q_κ^* corresponding to a specified risk parameter κ. The risk-sensitive Q-learning algorithm tunes $\hat{Q}_\kappa(i, u)$ so as to converge to the optimal Q-function Q_κ^*.

Let $(i_0, u_0, i_1, u_1, \ldots)$ be the sequence of states and actions which we encounter while interacting with the system and let $\hat{Q}_t$ denote the current approximation available after time step t. Then, we update $\hat{Q}_t$ at each time step according to

$$\begin{aligned} \hat{Q}_t(i, u) &= \hat{Q}_{t-1}(i, u) + \sigma_{t-1}(i, u)\mathcal{X}^\kappa\Big(g_{i_{t-1}i_t}(u_{t-1}) \\ &\qquad + \gamma \max_{v \in U(i_t)} \hat{Q}_{t-1}(i_t, v) - \hat{Q}_{t-1}(i_{t-1}, u_{t-1})\Big), \\ \sigma_{t-1}(i_{t-1}, u_{t-1}) &= \sigma_{t-1} > 0, \\ \sigma_{t-1}(i, u) &= 0 \quad \text{if } (i, u) \neq (i_{t-1}, u_{t-1}). \end{aligned} \tag{33}$$

Again, the only difference from traditional risk-neutral Q-learning is in the transformation $\mathcal{X}^\kappa$ that weights positive and negative temporal differences appropriately. Setting $\kappa = 0$, we recover Watkins' original risk-neutral algorithm.

Risk-sensitive Q-learning converges under the usual generic conditions.

Theorem 8 (*convergence*). *Let* $\kappa \in (-1, 1)$. *Consider the risk-sensitive Q-learning algorithm, as described by Eq.* (33). *If the stepsizes* $\sigma_t(i, u)$ *are nonnegative and satisfy*

$$\sum_{t=0}^{\infty} \sigma_t(i,u) = \infty \quad \sum_{t=0}^{\infty} (\sigma_t(i,u))^2 < \infty \quad \forall u \in U(i), i \in S$$

then $\hat{Q}_t(i, u)$ *converges to* $Q^*_\kappa(i, u)$ *for all* i *and* u, *with probability* 1.

Finally, we discuss risk-sensitive learning algorithms which involve parameterized function approximators for the value functions. This is done using functions $\tilde{J}(i; r)$ and $\tilde{Q}(i, u; w)$ which, given a state i, or, given a state-action pair (i, u), produce approximations to $J^\pi_\kappa(i)$ or $Q^*_\kappa(i, u)$, respectively. The approximation functions involve parameters r and w, respectively, and may be implemented by neural networks, feature mappings or other architectures. Within this context, the risk-sensitive TD(0)-algorithm takes the following form

$$\begin{aligned} r_t &= r_{t-1} + \sigma_t \mathcal{X}^\kappa(d_t) \nabla_r \tilde{J}(i_{t-1}; r_{t-1}), \quad \text{with} \\ d_t &= g_{i_{t-1} i_t}(\pi(i_{t-1})) + \gamma \tilde{J}(i_t; r_{t-1}) - \tilde{J}(i_{t-1}; r_{t-1}), \end{aligned} \tag{34}$$

where r_t denotes the parameter available after the t-th time step. Similarly, risk-sensitive Q-learning updates the corresponding parameter vector w according to

$$\begin{aligned} w_t &= w_{t-1} + \sigma_t \mathcal{X}^\kappa(d_t) \nabla_w \tilde{Q}(i_{t-1}, u_{t-1}; w_{t-1}), \quad \text{with} \\ d_t &= g_{i_{t-1} i_t}(u_{t-1}) + \gamma \max_{u \in U(i_t)} \tilde{Q}(i_t, u; w_{t-1}) - \tilde{Q}(i_{t-1}, u_{t-1}; w_{t-1}). \end{aligned} \tag{35}$$

Convergence results for reinforcement learning methods involving function approximation are generally much more difficult to obtain. Even in the classical risk-neutral case there are only few results available which usually cover special classes of problems. (TD(λ) for linear function approximators, Tsitsiklis & Van Roy, 1997, Q-learning using averagers, Gordon, 1995, Q-learning for optimal stopping problems, Tsitsiklis & Van Roy, 1999). Despite this lack of performance guarantees for many interesting cases, the risk-neutral variants of TD(λ) and Q-learning have been found to perform well in a variety of contexts (Singh & Bertsekas, 1997; Zhang & Dietterich, 1996; Marbach, Mihatsch, & Tsitsiklis, 2000; Neuneier, 1998). There is already some evidence that our risk-sensitive version of the algorithm will perform similarly in practice. In Neuneier and Mihatsch (2000), we successfully applied risk-sensitive reinforcement learning involving neural network architectures to the real world task of allocating funds to the German stock index DAX. For a broad range of risk parameters κ our algorithm converged to value functions which led to well-behaving policies of high performance. We could demonstrate the different degrees of risk-sensitivity of these policies. Nevertheless, the formal proofs extending the aforementioned risk-neutral convergence results to the risk-sensitive setting have to be done.

We finish with some remarks about the practical application of the above algorithms. As we already mentioned (see end of Section 5) the convergence speed will slow down with increasing risk sensitivity. This suggests the following learning strategy. Start learning with a small risk parameter κ (or even with $\kappa = 0$) and let the algorithm run until convergence. Then, use the result as initial point for a series of subsequent runs with increasing risk

sensitivity, where the resulting value function of each such run serves as initial guess for the following one.

Doing so, one obtains a series of near-optimal policies for several degrees of risk-sensitivity, from which we can learn quite a lot about our MDP at hand. The sensitivity of the policies with respect to changes in the risk parameter κ is an indicator for the amount of risk inherent in our MDP (see Theorem 2 and the discussion thereafter). The larger the amount of risk, i.e. the larger the likelihood of worst-case transitions, the more sensitive the policies depend on the risk parameter κ.

There is another interesting aspect of the proposed learning strategy which is worth mentioning. By letting $\kappa \to 1$ in the above procedure we obtain an algorithm which solves the worst-case optimization problem. Compared with Heger's (1994a) existing algorithm for this kind of problem, two advantages come into mind. First, the initial value function does not need to have any particular form, and, second, there is a natural way to extend this algorithm to involve function approximation.

Another interesting method is to start learning with $\kappa = 0$ and increase (decrease) the risk parameter continuously during learning, instead of changing it between subsequent runs with constant κ. Even though this algorithm is not covered by our convergence theorems, we expect it to perform well in practice. A more careful investigation of this approach remains to be done.

7. Conclusion and future work

We developed a new theory of risk-sensitive control which immediately leads to various risk-sensitive reinforcement algorithms. We showed that these new learning algorithms converge to optimal policies in the risk-sensitive sense and thereby circumvent the awkward obstacles of the worst-case criterion and the exponential utility approach. Fortunately, it turns out, that the already known and widely used reinforcement algorithms only require few minor changes to transform these algorithms into risk-sensitive versions (cf. Eqs. (32–35)).

Further analysis should address the following.

1. The rather technical questions how to extend our results to risk-sensitive versions of other reinforcement algorithms like Sarsa or TD(λ) for $\lambda \neq 0$.
2. The extension of the new framework to undiscounted and finite horizon MDPs.
3. Convergence results for risk-sensitive algorithms involving function approximation.
4. Convergence results for risk-sensitive algorithms which change the risk parameter κ during learning.
5. The extension of our results to more general transformations of the TD-error other than $\mathcal{X}^\kappa$ (Eq. (13)), which may be useful in certain contexts.

Appendix

A.1. Some preliminary lemmas

We need the following basic

Lemma 1. *Let* $\kappa \in (-1, 1)$. *To each pair of real numbers* a, b *there is a* $\xi_{(a,b,\kappa)} \in [1 - |\kappa|, 1 + |\kappa|]$ *such that*

$$\mathcal{X}^\kappa(a) - \mathcal{X}^\kappa(b) = \xi_{(a,b,\kappa)}(a - b).$$

Proof: The lemma is a simple extension of the mean-value theorem, which can be found in any analysis textbook, to piecewise differentiable mappings. □

In order to make the exposition of the subsequent proofs more concise, we define operators $\mathcal{T}^\pi_{\alpha,\kappa}, \mathcal{T}_{\alpha,\kappa}, \mathcal{M}_{\alpha,\kappa}, \mathcal{N}_{\alpha,\kappa}$ acting on the space of value functions and Q-functions, respectively, as

$$\begin{aligned}
\mathcal{T}^\pi_{\alpha,\kappa}[J](i) &:= J(i) + \alpha \sum_{j\in S} p_{ij}(\pi(i))\mathcal{X}^\kappa(g_{ij}(\pi(i)) + \gamma J(j) - J(i)),\\
\mathcal{T}_{\alpha,\kappa}[J](i) &:= J(i) + \alpha \max_{u\in U(i)} \sum_{j\in S} p_{ij}(u)\mathcal{X}^\kappa(g_{ij}(u) + \gamma J(j) - J(i)),\\
\mathcal{M}_{\alpha,\kappa}[J, Q](i, u) &:= Q(i, u) + \alpha \sum_{j\in S} p_{ij}(u)\mathcal{X}^\kappa(g_{ij}(u) + \gamma J(j) - Q(i, u)),\\
\mathcal{N}_{\alpha,\kappa}[Q](i, u) &:= Q(i, u) + \alpha \sum_{j\in S} p_{ij}(u)\mathcal{X}^\kappa\Big(g_{ij}(u) + \gamma \max_{v\in U(j)} Q(j, v) - Q(i, u)\Big).
\end{aligned}$$

The real value α denotes an arbitrary positive "stepsize". All operators, including the previously defined $\bar{T}^\pi$ and $\underline{T}^\pi$ turn out to be contraction mappings with respect to the maximum norm $|J| := \max_{i\in S}|J(i)|$ and $|Q| := \max_{i\in S, u\in U(i)}|Q(i, u)|$, respectively, provided that α is small enough.

Lemma 2. *Let* $\kappa \in (-1, 1)$, $0 \le \gamma < 1$ *and* $0 < \alpha < (1 + |\kappa|)^{-1}$. *For all value functions* J_1, J_2 *and* Q*-functions* Q_1, Q_2 *we have*

$$\left|\mathcal{T}^\pi_{\alpha,\kappa}[J_1] - \mathcal{T}^\pi_{\alpha,\kappa}[J_2]\right| \le \rho|J_1 - J_2|, \tag{36}$$
$$|\mathcal{T}_{\alpha,\kappa}[J_1] - \mathcal{T}_{\alpha,\kappa}[J_2]| \le \rho|J_1 - J_2|, \tag{37}$$
$$|\mathcal{M}_{\alpha,\kappa}[J_1, Q_1] - \mathcal{M}_{\alpha,\kappa}[J_1, Q_2]| \le \rho|Q_1 - Q_2|, \tag{38}$$
$$|\mathcal{N}_{\alpha,\kappa}[Q_1] - \mathcal{N}_{\alpha,\kappa}[Q_2]| \le \rho|Q_1 - Q_2|, \tag{39}$$
$$|\bar{T}^\pi[J_1] - \bar{T}^\pi[J_2]| \le \gamma|J_1 - J_2|, \tag{40}$$
$$|\underline{T}^\pi[J_1] - \underline{T}^\pi[J_2]| \le \gamma|J_1 - J_2|, \tag{41}$$

where $\rho = (1 - \alpha(1 - |\kappa|)(1 - \gamma)) \in (0, 1)$. *Thus, all operators are contraction mappings.*

Proof: The contraction property of $\underline{T}^\pi$ and $\bar{T}^\pi$ is a standard result, which has been proven elsewhere (Bertsekas, 1995; Heger, 1994b). Here, we show only the inequality (37). The other properties (36, 38, 39) can be proven similarly.

We have

$$
\begin{aligned}
&|\mathcal{T}_{\alpha,\kappa}[J_1](i) - \mathcal{T}_{\alpha,\kappa}[J_2](i)| \\
&\quad = \left| \max_{u\in U(i)} \left(J_1(i) + \alpha \sum_{j\in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_1(j) - J_1(i)) \right) \right. \\
&\qquad \left. - \max_{u\in U(i)} \left(J_2(i) + \alpha \sum_{j\in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_2(j) - J_2(i)) \right) \right| \\
&\quad \le \max_{u\in U(i)} \left| J_1(i) - J_2(i) + \alpha \sum_{j\in S} p_{ij}(u) \left(\mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_1(j) - J_1(i)) \right. \right. \\
&\qquad \left. \left. - \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_2(j) - J_2(i)) \right) \right|
\end{aligned}
$$

With the help of Lemma 1, we can get rid of the function $\mathcal{X}^{\kappa}$.

$$
\begin{aligned}
|\mathcal{T}_{\alpha,\kappa}[J_1](i) - \mathcal{T}_{\alpha,\kappa}[J_2](i)| &\le \max_{u\in U(i)} \left| J_1(i) - J_2(i) \right. \\
&\quad \left. + \alpha \sum_{j\in S} p_{ij}(u) \xi_{(i,j,u,J_1,J_2)} (\gamma (J_1(j) - J_2(j)) - (J_1(i) - J_2(i))) \right| \\
&= \max_{u\in U(i)} \left| \left(1 - \alpha \sum_{j\in S} p_{ij}(u) \xi_{(i,j,u,J_1,J_2)} \right) (J_1(i) - J_2(i)) \right. \\
&\quad \left. + \alpha\gamma \sum_{j\in S} p_{ij}(\pi(i)) \xi_{(i,j,u,J_1,J_2)} (J_1(j) - J_2(j)) \right|.
\end{aligned}
$$

We note that $1 - \alpha \sum_{j\in S} p_{ij}(\pi(i)) \xi_{(i,j,J_1,J_2)} > 0$, since $0 < \alpha \le (1 + |\kappa|)^{-1}$ (by hypothesis) and $\xi_{(i,j,u,J_1,J_2)} \in [1 - |\kappa|, 1 + |\kappa|]$ (Lemma 1). Thus,

$$
\begin{aligned}
|\mathcal{T}_{\alpha,\kappa}[J_1](i) - \mathcal{T}_{\alpha,\kappa}[J_2](i)| &\le \max_{u\in U(i)} \left(1 - \alpha(1-\gamma) \sum_{j\in S} p_{ij}(u) \xi_{(i,j,u,J_1,J_2)} \right) |J_1 - J_2| \\
&\le (1 - \alpha(1-\gamma)(1 - |\kappa|)) |J_1 - J_2|. \qquad \square
\end{aligned}
$$

The subsequent proofs will make use of the monotonicity property of the operators $\mathcal{T}^{\pi}_{\alpha,\kappa}$, $\mathcal{T}_{\alpha,\kappa}$, $\underline{T}^{\pi}$ and $\bar{T}^{\pi}$.

Lemma 3 (*monotonicity*). *Let J_1 and J_2 be arbitrary value functions such that $J_1(i) \ge J_2(i)$ for all states $i \in S$. Let $0 < \alpha < (1 + |\kappa|)^{-1}$. Then*

$$\mathcal{T}^{\pi}_{\alpha,\kappa}[J_1](i) \ge \mathcal{T}^{\pi}_{\alpha,\kappa}[J_2](i), \tag{42}$$

$$\mathcal{T}_{\alpha,\kappa}[J_1](i) \ge \mathcal{T}_{\alpha,\kappa}[J_2](i), \tag{43}$$

$$\bar{T}^{\pi}[J_1](i) \geq \bar{T}^{\pi}[J_2](i), \tag{44}$$

$$\underline{T}^{\pi}[J_1](i) \geq \underline{T}^{\pi}[J_2](i) \tag{45}$$

for all $i \in S$.

Proof: The monotonicity lemmas for the standard dynamic programming operators $\bar{T}^{\pi}$ and $\underline{T}^{\pi}$ are well-known results (Bertsekas, 1995; Heger, 1994b). Here, we only show the monotonicity of the risk-sensitive operator $\mathcal{T}_{\alpha,\kappa}$. The proof for $\mathcal{T}^{\pi}_{\alpha,\kappa}$ is similar.

We have

$$\begin{aligned}
&\mathcal{T}_{\alpha,\kappa}[J_1](i) - \mathcal{T}_{\alpha,\kappa}[J_2](i) \\
&\quad = J_1(i) - J_2(i) + \alpha \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_1(j) - J_1(i)) \\
&\qquad - \alpha \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^{\kappa}(g_{ij}(u) + \gamma J_2(j) - J_2(i)) \\
&\quad \geq J_1(i) - J_2(i) + \alpha \sum_{j \in S} p_{ij}(\tilde{u}) (\mathcal{X}^{\kappa}(g_{ij}(\tilde{u}) + \gamma J_1(j) - J_1(i)) \\
&\qquad - \mathcal{X}^{\kappa}(g_{ij}(\tilde{u}) + \gamma J_2(j) - J_2(i))
\end{aligned}$$

where $\tilde{u}$ is an action for which the second max-operator attains its maximum. Using Lemma 1 to get rid of the functions $\mathcal{X}^{\kappa}$, we obtain

$$\begin{aligned}
\mathcal{T}_{\alpha,\kappa}[J_1](i) - \mathcal{T}_{\alpha,\kappa}[J_2](i) \geq &\left(1 - \alpha \sum_{j \in S} p_{ij}(\tilde{u}) \xi_{(i,j,J_1,J_2)}\right) (J_1(i) - J_2(i)) \\
&+ \alpha\gamma \sum_{j \in S} p_{ij}(\tilde{u}) \xi_{(i,j,J_1,J_2)} (J_1(j) - J_2(j)) \geq 0.
\end{aligned}$$

We note that all quantities in the above equation are nonnegative and the result follows. □

A.2. Proof of Theorems 1 and 2 (limiting behavior)

First, we will prove the existence and uniqueness statement of Theorem 1. Then, we will proceed with Theorem 2. The rest of Theorem 1 will be a simple conclusion thereof.

Since $\mathcal{T}^{\pi}_{\alpha,\kappa}$ is a contraction mapping with respect to the maximum norm (see Lemma 2, Eq. (36)), there is unique fixed point J^{π}_{κ} satisfying $\mathcal{T}^{\pi}_{\alpha,\kappa}[J^{\pi}_{\kappa}] = J^{\pi}_{\kappa}$ which is exactly the existence and uniqueness result of Theorem 1.

We proceed with Theorem 2. Let $\kappa \in [0, 1)$. The defining equation (14) for J^{π}_{κ} can be rewritten in terms of the operators $\bar{T}^{\pi}$ and $\bar{T}^{\pi}$

$$\begin{aligned}
0 = &(1 - \kappa)\left(\bar{T}^{\pi}\left[J^{\pi}_{\kappa}\right](i) - J^{\pi}_{\kappa}(i)\right) \\
&+ 2\kappa \sum_{j \in \tilde{S}} p_{ij}(\pi(i)) \Big[\underbrace{g_{ij}(\pi(i)) + \gamma J^{\pi}_{\kappa}(j) - J^{\pi}_{\kappa}(i)}_{\leq 0}\Big],
\end{aligned} \tag{46}$$

where the summation runs over all successor states j with nonpositive temporal differences. The sum in the above equation is bounded above and below.

$$\begin{aligned}\underline{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) - J_{\kappa}^{\pi}(i) &\leq \sum_{j\in\tilde{S}} p_{ij}(\pi(i))\big[g_{ij}(\pi(i)) + \gamma J_{\kappa}^{\pi}(j) - J_{\kappa}^{\pi}(i)\big] \\ &\leq p_{\kappa}^{\pi}(i)\big(\underline{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) - J_{\kappa}^{\pi}(i)\big)\end{aligned}$$

Thus, there are quantities $\xi_{\kappa}^{\pi}(i)$ with $p_{\kappa}^{\pi}(i) \leq \xi_{\kappa}^{\pi}(i) \leq 1$ such that Eq. (46) takes the form

$$0 = (1-\kappa)\big(\underbrace{\bar{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) - J_{\kappa}^{\pi}(i)}_{\geq 0}\big) + 2\kappa\xi_{\kappa}^{\pi}(i)\big(\underbrace{\underline{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) - J_{\kappa}^{\pi}(i)}_{\leq 0}\big). \tag{47}$$

The second term is nonpositive which, in turn, implies that that the first one has to be nonnegative. In other words, we have

$$J_{\kappa}^{\pi}(i) \leq \bar{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i), \tag{48}$$

$$J_{\kappa}^{\pi}(i) \geq \underline{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i). \tag{49}$$

Equations (48) and (49) together with (47) establish the proof of the left part of the chain of inequalities (22) and (23) given in Theorem 2.

The rest follows from standard arguments. We recall that repeated application of $\bar{T}^{\pi}$ and $\underline{T}^{\pi}$ to an arbitrary value function J results in a sequence of value functions converging towards $\bar{J}^{\pi}$ and $\underline{J}^{\pi}$, respectively. Applying repeatedly the operators $\bar{T}^{\pi}$ and $\underline{T}^{\pi}$ on both sides of Eqs. (48) and (49), respectively, and using the Monotonicity Lemma 3, we obtain

$$\begin{aligned}J_{\kappa}^{\pi}(i) &\leq \bar{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) \leq \cdots \leq (\bar{T}^{\pi})^{k}\big[J_{\kappa}^{\pi}\big](i) \\ &\leq \lim_{N\to\infty} (\bar{T}^{\pi})^{N}\big[J_{\kappa}^{\pi}\big](i) = \bar{J}^{\pi}(i),\end{aligned} \tag{50}$$

$$\begin{aligned}J_{\kappa}^{\pi}(i) &\geq \underline{T}^{\pi}\big[J_{\kappa}^{\pi}\big](i) \leq \cdots \leq (\underline{T}^{\pi})^{k}\big[J_{\kappa}^{\pi}\big](i) \\ &\geq \lim_{N\to\infty} (\underline{T}^{\pi})^{N}\big[J_{\kappa}^{\pi}\big](i) = \underline{J}^{\pi}(i).\end{aligned} \tag{51}$$

This proves Eq. (21) and the remaining parts of Eqs. (22) and (23) which completes Theorem 2.

Next, we finish the proof of Theorem 1. The limiting property (15) is a direct consequence of the fact that Eq. (14) reduces to $\bar{T}^{\pi}[J_0^{\pi}] = J_0^{\pi}$ if $\kappa = 0$. For $\kappa \to 1$ we have

$$\lim_{\kappa\to 1}\big(J_{\kappa}^{\pi}(i) - \underline{T}\big[J_{\kappa}^{\pi}\big](i)\big) = 0 \quad \forall i \in S,$$

which is a consequence of Eq. (23). (Note that $\xi_{\kappa}^{\pi}(i)$ can not come arbitrary close to zero, since we are dealing with a finite state space). In combination with the following standard result (Heger, 1994b)

$$|J(i) - \underline{J}^{\pi}(i)| \leq \frac{\gamma}{1-\gamma} \max_{j\in S} |\underline{T}^{\pi}[J](j) - J(j)| \quad \forall \text{ value functions } J, i \in S$$

the risk-averse limiting property (16) follows.

We omit the analogous proof for the risk-seeking limit (17), since it does not involve any further ideas. □

A.3. Proof of Theorem 4 (optimal policies)

The contraction property of the operator $\mathcal{T}_{\alpha,\kappa}$ (Lemma 2) implies that there is a unique value function J_κ^* satisfying $\mathcal{T}_{\alpha,\kappa}[J_\kappa^*] = J_\kappa^*$ which is a shorthand notation of Eq. (26).

Let π be a policy with an associated value function not worse than J_κ^*

$$J_\kappa^\pi(i) \geq J_\kappa^*(i) \quad \forall i \in S.$$

Applying repeatedly the operator $\mathcal{T}_{\alpha,\kappa}$ to both sides of the inequality and using the monotonicity of $\mathcal{T}_{\alpha,\kappa}$ (Lemma 3), we obtain the chain of inequalities

$$\begin{aligned} J_\kappa^*(i) &= \lim_{N\to\infty} (\mathcal{T}_{\alpha,\kappa})^N \left[J_\kappa^\pi\right](i) \geq \mathcal{T}_{\alpha,\kappa}\left[J_\kappa^\pi\right](i) \\ &\geq \mathcal{T}_{\alpha,\kappa}^\pi \left[J_\kappa^\pi\right](i) = J_\kappa^\pi(i) \geq J_\kappa^*(i), \end{aligned}$$

where $J_\kappa^* = \lim_{N\to\infty}(\mathcal{T}_{\alpha,\kappa})^N[J_\kappa^\pi]$ is implied by the contraction property of $\mathcal{T}_{\alpha,\kappa}$ and $\mathcal{T}_{\alpha,\kappa}[J_\kappa^\pi](i) \geq \mathcal{T}_{\alpha,\kappa}^\pi[J_\kappa^\pi](i)$ follows immediately from the definition of the operators. Thus, $J_\kappa^\pi = J_\kappa^*$ and $\mathcal{T}_{\alpha,\kappa}[J_\kappa^\pi] = \mathcal{T}_{\alpha,\kappa}^\pi[J_\kappa^\pi]$ showing that no value function can be larger than J_κ^* and that optimal policies fulfill Eq. (27). □

A.4. Proof of Theorem 3 (policy improvement)

Using operator notation we can rewrite Eq. (24) as

$$\mathcal{T}_{\alpha,\kappa}^{\pi'}\left[J_\kappa^\pi\right](i) \geq \mathcal{T}_{\alpha,\kappa}^\pi\left[J_\kappa^\pi\right](i) \overset{(14)}{=} J_\kappa^\pi(i) \quad \forall i \in S.$$

Applying repeatedly $\mathcal{T}_{\alpha,\kappa}^{\pi'}$ on both sides of the above inequality and using the Monotonicity Lemma 3, we obtain

$$J_\kappa^{\pi'}(i) = \lim_{N\to\infty} \left(\mathcal{T}_{\alpha,\kappa}^{\pi'}\right)^N \left[J_\kappa^\pi\right](i) \geq \cdots \geq \mathcal{T}_{\alpha,\kappa}^{\pi'}\left[J_\kappa^\pi\right](i) \geq J_\kappa^\pi(i)$$

proving Eq. (25).

If $J_\kappa^{\pi'} = J_\kappa^\pi$, then from the preceding relation it follows that $J_\kappa^\pi = \mathcal{T}_{\alpha,\kappa}^{\pi'}[J_\kappa^\pi]$ and since by hypothesis we have $\mathcal{T}_{\alpha,\kappa}^{\pi'}[J_\kappa^\pi] = \mathcal{T}_{\alpha,\kappa}[J_\kappa^\pi]$, we obtain $J_\kappa^\pi = \mathcal{T}_{\alpha,\kappa}[J_\kappa^\pi]$ implying that $J_\kappa^\pi = J_\kappa^*$ by Theorem 4. Thus, π must be optimal. It follows that if π is not optimal, then $J_\kappa^{\pi'}(i) > J_\kappa^\pi(i)$ for some state i. □

A.5. Proof of Theorem 5 (Q-function)

Theorem 5 is a direct consequence of the contraction property of the operator $\mathcal{M}[\cdot,\cdot]$ with respect to its second argument (Lemma 1, Eq. (38)) which ensures that there is a unique fixed point Q_κ^π satisfying $\mathcal{M}[J_\kappa^\pi, Q_\kappa^\pi] = Q_\kappa^\pi$. □

A.6. *Proof of Theorem 6 (Optimal Q-function)*

We recall that J_κ^* and Q_κ^* are the unique solutions of (see Eqs. (26) and (29))

$$0 = \sum_{j \in S} p_{ij}(u) \mathcal{X}^\kappa(g_{ij}(u) + \gamma J_\kappa^*(j) - Q_\kappa^*(i,u)) \quad \forall i \in S, u \in U(i), \tag{52}$$

$$0 = \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) \mathcal{X}^\kappa(g_{ij}(u) + \gamma J_\kappa^*(j) - J_\kappa^*(i)) \quad \forall i \in S. \tag{53}$$

Building the difference of the above equations we obtain the inequality

$$0 \leq \sum_{j \in S} p_{ij}(u)(\mathcal{X}^\kappa(g_{ij}(u) + \gamma J_\kappa^*(j) - Q_\kappa^*(i,u)) - \mathcal{X}^\kappa(g_{ij}(u) + \gamma J_\kappa^*(j) - J_\kappa^*(i)))$$

The above inequality is true for all actions $u \in U(i)$. We conclude from Theorem 4 (Eq. (27)) that equality holds if and only if u is optimal, i.e. if the maximum in Eq. (53) is attained. Again, we can get rid of the transformation function $\mathcal{X}^\kappa$ with the help of Lemma 1. Some further algebra leads us to

$$0 \leq \left(\sum_{j \in S} p_{ij}(u) \xi_{(i,j,u,Q_\kappa^*,J_\kappa^*)} \right) (J_\kappa^*(i) - Q_\kappa^*(i,u)).$$

As before, equality holds if and only if u is optimal. Since the first factor in the above inequality is strictly positive, we conclude that

$$J_\kappa^*(i) \geq Q_\kappa^*(i,u) \quad \forall i \in S, u \in U(i)$$
$$J_\kappa^*(i) = Q_\kappa^*(i,u^*) \quad \forall i \in S \text{ for all optimal actions } u^*,$$

which proves Eq. (31). Inserting the property $J_\kappa^*(i) = \max_{u \in U(i)} Q_\kappa^*(i,u)$ into Eq. (52), we see that $Q_\kappa^*(i,u)$ solves the optimality equation (30). The contraction property (39) of the operator $\mathcal{N}_{\alpha,\kappa}$ ensures that $Q_\kappa^*(i,u)$ is the unique solution thereof. □

A.7. *Proof of Theorems 7 and 8 (TD(0)- and Q-learning)*

The TD(0)-algorithm (32) can be viewed as a special case of Q-learning (33) where the sets $U(i)$ of admissible actions are singletons, i.e. $U(i) = \{\pi(i)\}$. Thus, we only have to prove the convergence of risk-sensitive Q-learning (Theorem 8).

We plan to apply the following result concerning stochastic iterative algorithms (see Bertsekas & Tsitsiklis, 1996, Proposition 4.4, p. 156) which we state here without proof.

Theorem 9. *Let $r_t \in \mathbf{R}^m$ be the sequence generated by the iteration algorithm*

$$r_{t+1}(i) = (1 - \sigma_t(i)) r_t(i) + \sigma_t(i)((Hr_t)(i) + \omega_t(i).) \quad \forall i \in 1, \ldots, m \tag{54}$$

We assume the following.

1. *The stepsizes* $\sigma_t(i)$ *are nonnegative and satisfy*

$$\sum_{t=0}^{\infty} \sigma_t(i) = \infty \quad \sum_{t=0}^{\infty} (\sigma_t(i))^2 < \infty \quad \forall i \in 1, \ldots, m \tag{55}$$

2. *The noise terms* $\omega_t(i)$ *satisfy*
   (a) *For every* i *and* t, *we have* $E[\omega_t(i)|\mathcal{F}_t] = 0$, *where* $\mathcal{F}_t$ *denotes the history of the process up to (and including) time step* t.
   (b) *Given any norm* $\|\cdot\|$ *on* $\mathbf{R}^m$ *there exist constants* A *and* B *such that*

$$E\left[\omega_t^2(i)\middle|\mathcal{F}_t\right] \leq A + B\|r_t\|^2 \quad \forall i, t$$

3. *The mapping* H *is a maximum norm contraction.*

Then, r_t *converges the unique solution* r^* *of the equation* $Hr^* = r^*$, *with probability* 1.

In order to prove the convergence of risk-sensitive Q-learning, we need to verify all the required assumptions in the above theorem. To this effect, we reformulate the Q-learning rule (33) in a slightly different way

$$\hat{Q}_t(i,u) = \left(1 - \frac{\sigma_{t-1}(i)}{\alpha}\right)\hat{Q}_{t-1}(i,u) + \frac{\sigma_{t-1}(i,u)}{\alpha}(\alpha\mathcal{X}^{\kappa}(d_{t-1}) + \hat{Q}_{t-1}(i,u)),$$

where

$$d_{t-1} = g_{i_{t-1}i_t}(u_{t-1}) + \gamma \max_{u \in U(i_t)} \hat{Q}_{t-1}(i_t, u;) - \hat{Q}_{t-1}(i_{t-1}, u_{t-1}),$$

and where α denotes an arbitrary scalar which we choose such that $\alpha \in (0, (1 + |\kappa|)^{-1})$. With the help of the operator $\mathcal{N}_{\alpha,\kappa}$ we get

$$\hat{Q}_t(i,u) = \left(1 - \frac{\sigma_{t-1}(i)}{\alpha}\right)\hat{Q}_{t-1}(i,u) + \frac{\sigma_{t-1}(i,u)}{\alpha}(\mathcal{N}_{\alpha,\kappa}[\hat{Q}_{t-1}](i,u) + \omega_{t-1}(i,u))$$

where $\omega_{t-1}(i,u)$ is defined as

$$\omega_{t-1}(i,u) = \alpha\mathcal{X}^{\kappa}(d_{t-1}) + \hat{Q}_{t-1}(i,u) - \mathcal{N}_{\alpha,\kappa}[\hat{Q}_{t-1}](i,u).$$

Our update rule is now exactly of the form (54) and Theorem 9 can be applied easily. The mapping $\mathcal{N}_{\alpha,\kappa}$ (playing the role of H) is a maximum norm contraction by Lemma 2. The stepsize condition (55) is obviously fulfilled. The condition $E[\omega_t(i)|\mathcal{F}_t] = 0$ is a straightforward consequence of the definition of $\omega_t(i)$. It remains to show that $E[\omega_t^2(i)|\mathcal{F}_t] \leq A + B\|\hat{Q}_t\|^2$ with respect to some norm $\|\cdot\|$.

Since the variance of any random variable is no larger than its mean square, $E[\omega_t^2(i) \mid \mathcal{F}_t]$ is bounded above by

$$E[(\mathcal{X}^{\kappa}(d_t))^2 \mid \mathcal{F}_t].$$

Let G be an upper bound for $g_{ij}(u)$. Then $|d_t| \leq G + 2\|\hat{Q}_t\|$ implying that $\mathcal{X}^\kappa(d_t) \leq (1 + |\kappa|)(G + 2\|\hat{Q}_t\|)$. The desired result follows with the help of $(G + 2\|\hat{Q}_t\|)^2 \leq 2G^2 + 8\|\hat{Q}_t\|^2$.

By Theorem 9 the sequence $\hat{Q}_t$ converges to the unique solution Q_κ^* of $\mathcal{N}_{\alpha,\kappa}[Q_\kappa^*] = Q_\kappa^*$ which is the optimal Q-function by Theorem 6. □

References

Basar, T. S., & Bernhard, P. (1995). *H^∞-optimal control and related minimax design problems: A dynamic game approach* (2nd edn.). Boston: Birkhäuser.

Bellman, R. E., & Dreyfus, S. E. (1962). *Applied dynamic programming*. Princeton: Princeton University Press.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control* (Vol. 2.). Belmont, MA: Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Coraluppi, S. (1997). Optimal control of Markov decision processes for performance and Robustness. Ph.D. Thesis, University of Maryland.

Coraluppi, S. P., & Marcus, S. I. (1999). Risk-sensitive, minimax and mixed risk-neutral/minimax control of Markov decision processes. In W. M. McEarney, G. G. Yin, & Q. Zhang (Eds.), *Stochastic analysis, control, optimization and applications* (pp. 21–40). Boston: Birkhäuser.

Elton, E. J., & Gruber, M. J. (1995). *Modern portfolio theory and investment analysis*. New York: John Wiley & Sons.

Gordon, G. J. (1995). Stable function approximation in dynamic programming. In A. Prieditis, & S. J. Russel (Eds.), *Machine Learning: Proceedings of the Twelfth International Conference* (pp. 261–268). San Francisco: Morgan Kaufmann Publishers.

Heger, M. (1994a). Consideration of risk and reinforcement learning. In W. W. Cohen, & H. Hirsh (Eds.), *Machine Learning: Proceedings of the Eleventh International Conference* (pp. 105–111). San Francisco: Morgan Kaufmann Publishers.

Heger, M. (1994b). Risk and reinforcement learning: Concepts and dynamic programming. Technical Report, Zentrum für Kognitionswissenschaften, Universität Bremen, Germany.

Howard, R. A., & Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management Science, 18:7*, 356–369.

Koenig, S., & Simmons, R. G. (1994). Risk-sensitive planning with probabilistic decision graphs. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning* (KR) (pp. 363–373).

Littman, M. L., & Szepesvári, C. (1996). A generalized reinforcement-learning model: Convergence and applications. In L. Saitta (Ed.), *Machine Learning: Proceedings of the Thirteenth International Conference* (pp. 310–318). San Francisco: Morgan Kaufman Publishers.

Marbach, P., Mihatsch, O., & Tsitsiklis. J. N. (2000). Call admission control and routing in integrated services networks using neuro-dynamic programming. *IEEE Journal on Selected Areas in Communications, 18:2*, 197–208.

Neuneier, R. (1998). Enhancing Q-learning for optimal asset allocation. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (Vol. 10). Cambridge, MA: The MIT Press.

Neuneier, R., & Mihatsch, O. (2000). Risk-averse asset allocation using reinforcement learning. In: *Proceedings of the Seventh International Conference on Forecasting Financial Markets: Advances for Exchange Rates, Interest Rates and Asset Management*.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica, 32*, 122–136.

Puterman, M. L. (1994). *Markov decision processes*. New York: John Wiley & Sons.

Singh, S., & Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in neural information processing systems* (Vol. 9, pp. 974–980). Cambridge, MA: The MIT Press.

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

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

Tsitsiklis, J. N., & Van Roy, B. (1999). Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives. *IEEE Transactions on Automatic Control, 44:10*, 1840–1851.

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior* (3rd edn.). Princeton University Press.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Ph.D. Thesis, University of Cambridge, England.

Zhang, W., & Dietterich, T. G. (1996). High-performance job-shop scheduling with a time-delay TD(λ) network. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 1024–1030). Cambridge, MA: The MIT Press.

Received March 16, 1999
Revised July 26, 2000
Accepted July 26, 2000
Final manuscript July 26, 2000