Risk-Sensitive Reinforcement Learning

Yun Shen  
yun@ni.tu-berlin.de  
Technical University, 10587 Berlin, Germany

Michael J. Tobia  
m.tobia@uke.de  
Tobias Sommer  
t.sommer@uke.de  
University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany

Klaus Obermayer  
oby@ni.tu-berlin.de  
Technical University, 10587 Berlin, Germany, and Bernstein Center for  
Computational Neuroscience Berlin, 10115 Berlin, Germany

We derive a family of risk-sensitive reinforcement learning methods for agents, who face sequential decision-making tasks in uncertain environments. By applying a utility function to the temporal difference (TD) error, nonlinear transformations are effectively applied not only to the received rewards but also to the true transition probabilities of the underlying Markov decision process. When appropriate utility functions are chosen, the agents’ behaviors express key features of human behavior as predicted by prospect theory (Kahneman & Tversky, 1979), for example, different risk preferences for gains and losses, as well as the shape of subjective probability curves. We derive a risk-sensitive Q-learning algorithm, which is necessary for modeling human behavior when transition probabilities are unknown, and prove its convergence. As a proof of principle for the applicability of the new framework, we apply it to quantify human behavior in a sequential investment task. We find that the risk-sensitive variant provides a significantly better fit to the behavioral data and that it leads to an interpretation of the subject’s responses that is indeed consistent with prospect theory. The analysis of simultaneously measured fMRI signals shows a significant correlation of the risk-sensitive TD error with BOLD signal change in the ventral striatum. In addition we find a significant correlation of the risk-sensitive Q-values with neural activity in the striatum, cingulate cortex, and insula that is not present if standard Q-values are used.
1 Introduction

Risk arises from the uncertainties associated with future events and is inevitable since the consequences of actions are uncertain at the time a decision is made. Hence, risk has to be taken into account by the decision maker, consciously or unconsciously. An economically rational decision making rule, which is risk neutral, is to select the alternative with the highest expected reward. In the context of sequential or multistage decision-making problems, reinforcement learning (RL; Sutton & Barto, 1998) follows this line of thought. It describes how an agent ought to take actions that maximize expected cumulative rewards in an environment typically described by a Markov decision process (MDP; Puterman, 1994). RL is a well-developed model not only for human decision making but also for models of free choice in nonhumans, because similar computational structures, such as dopaminergically mediated reward prediction errors, have been identified across species (Schultz, Dayan, & Montague, 1997; Schultz, 2002).

Besides risk-neutral policies, risk-averse policies, which accept a choice with a more certain but possibly lower expected reward, are also considered economically rational (Gollier, 2004). For example, a risk-averse investor might choose to put money into a bank account with a low but guaranteed interest rate rather than into a stock with possibly high expected returns but also a chance of high losses. Conversely, risk-seeking policies, which prefer a choice with less certain but possibly a high reward, are considered economically irrational. Human agents are, however, not always economically rational (Gilboa, 2009). Behavioral studies show that humans can be risk seeking in one situation and risk averse in another (Kahneman & Tversky, 1979). The RL algorithms developed so far cannot effectively model these complicated risk preferences.

Risk-sensitive decision-making problems, in the context of MDPs, have been investigated in various fields: machine learning (Heger, 1994; Mihatsch & Neuneier, 2002), optimal control (Hernández-Hernández & Marcus, 1996), operations research (Howard & Matheson, 1972; Borkar, 2002), finance (Ruszczyński, 2010), and neuroscience (Nagengast, Braun, & Wolpert, 2010; Braun, Nagengast, & Wolpert, 2011; Niv, Edlund, Dayan, & O’Doherty, 2012). Note that the core of MDPs consists of two sets of objective quantities describing the environment: immediate rewards obtained at states by executing actions and transition probabilities for switching states when performing actions. Facing the same environment, however, different agents might have different policies, which indicates that risk is taken into account differently by different agents. Hence, to incorporate risk, which is derived from both quantities, all existing literature applies a nonlinear transformation to either the experienced reward values or the transition probabilities, or both. The former is the canonical approach in classical economics, as in expected utility theory (Gollier, 2004), while the latter originates from behavioral economics, as in subjective probability (Savage, 1972),
but is also derived from a rather recent development in mathematical finance, convex/coherent risk measures (Artzner, Delbaen, Eber, & Heath, 1999; Föllmer & Schied, 2002). For modeling human behaviors, prospect theory (Kahneman & Tversky, 1979) suggests that we should combine both approaches: human beings have different perceptions not only for the same objective amount of rewards but also the same value of the true probability. Recently, Niv et al. (2012) combined both approaches by applying piecewise linear functions (an approximation of a nonlinear transformation) to reward prediction errors that contain the information of rewards directly and the information of transition probabilities indirectly. Importantly, the reward prediction errors that incorporated experienced risk were strongly coupled to activity in the nucleus accumbens of the ventral striatum, providing a biologically based plausibility to this combined approach. In this work we show (in section 2.1) that the risk-sensitive algorithm proposed by Niv and colleagues is a special case of our general risk-sensitive RL framework.

Most of the literature in economics or engineering fields focuses on economically rational risk-averse or risk-neutral strategies, which are not always adopted by humans. The models proposed in behavioral economics, despite allowing economic irrationality, require knowledge of the true probability, which usually is not available at the outset of a learning task. In neuroscience, on the one hand, several works (Wu, Delgado, & Maloney, 2009; Preuschoff, Quartz, & Bossaerts, 2008) follow the same line as in behavioral economics and require knowledge of the true probability. Though different modified RL algorithms (Glimcher, Camerer, Fehr, & Poldrack, 2008; Symmonds, Wright, Bach, & Dolan, 2011) are applied to model human behaviors in learning tasks, the algorithms often fail to generalize across different tasks. In our previous work (Shen, Stannat, & Obermayer, 2013), we described a general framework for incorporating risk into MDPs by introducing nonlinear transformations to both rewards and transition probabilities. A risk-sensitive objective was derived and optimized by value iteration or dynamic programming. This solution, hence, does not work in learning tasks where the true transition probabilities are unknown to learning agents. For this purpose, a model-free framework for RL algorithms is to be derived in this letter, where, similar to Q-learning, the knowledge of the transition and reward model is not needed.

This letter is organized as follows. Section 2 starts with a mathematical introduction to valuation functions for measuring risk. We then specify a sufficiently rich class of valuation functions in section 2.1 and provide the intuition behind our approach by applying this class to a simple example in section 2.2. We also show that key features of prospect theory can be captured by this class of valuation functions. Restricting ourselves to the same class, we derive a general framework for risk-sensitive Q-learning algorithms and prove its convergence in section 3. In section 4, we apply this framework to quantify human behavior. We show that the risk-sensitive variant provides a significantly better fit to the behavioral data and
significant correlations are found between sequences generated by the proposed framework and changes of fMRI BOLD signals.

2 Valuation Functions and Risk Sensitivities

Suppose that we are facing choices. Each choice might yield different outcomes when events are generated by a random process. Hence, to keep generality, we model the outcome of each choice by a real-valued random variable $\{X(i), \mu(i)\}_{i \in I}$, where $I$ denotes an event space with a finite cardinality $|I|$ and $X(i) \in \mathbb{R}$ is the outcome of $i$th event with probability $\mu(i)$. We say two vectors $X \leq Y$ if $X(i) \leq Y(i)$ for all $i \in I$. Let $1$ (resp. $0$) denote the vector with all elements equal $1$ (resp. $0$). Let $\mathcal{P}$ denote the space of all possible distributions $\mu$.

Choices are made according to their outcomes. Hence, we assume that there exists a mapping $\rho : \mathbb{R}^{|I|} \times \mathcal{P} \to \mathbb{R}$ such that one prefers $(X, \mu)$ to $(Y, \nu)$ whenever $\rho(X, \mu) \geq \rho(Y, \nu)$. We assume further that $\rho$ satisfies the following axioms inspired by the risk measure theory applied in mathematical finance (Artzner et al., 1999; Föllmer & Schied, 2002). A mapping $\rho : \mathbb{R}^{|I|} \times \mathcal{P} \to \mathbb{R}$ is called a valuation function if it satisfies for each $\mu \in \mathcal{P}$,

I (monotonicity) $\rho(X, \mu) \leq \rho(Y, \mu)$, whenever $X \leq Y \in \mathbb{R}^{|I|}$

II (translation invariance) $\rho(X + y1, \mu) = \rho(X, \mu) + y$, for any $y \in \mathbb{R}$

Within the economic context, $X$ and $Y$ are outcomes of two choices. Monotonicity reflects the intuition that given the same event distribution $\mu$, if the outcome of one choice is always (for all events) higher than the outcome of another choice, the valuation of the choice must be also higher. Under the axiom of translation invariance, the sure Outcome $y1$ (equal outcome for every event) after executing decisions is considered a sure outcome before making decision. This also reflects the intuition that there is no risk if there is no uncertainty.

In our setting, valuation functions are not necessarily centralized, that is, $\rho(0, \mu)$ is not necessarily $0$, since $\rho(0, \mu)$ in fact sets a reference point that can differ for different agents. However, we can centralize any valuation function by $\tilde{\rho}(X, \mu) := \rho(X, \mu) - \rho(0, \mu)$. From the two axioms, it follows that (for the proof see lemma 1 in the appendix)

$$\min_{i \in I} X_i := X \leq \tilde{\rho}(X, \mu) \leq \overline{X} := \max_{i \in I} X_i, \forall \mu \in \mathcal{P}, X \in \mathbb{R}^{|I|}. \quad (2.1)$$

$\overline{X}$ is the largest possible outcome, which represents the most optimistic prediction of the future, while $\underline{X}$ is the smallest outcome possible and the most pessimistic estimation. The centralized valuation function $\tilde{\rho}(X, \mu)$ satisfying $\tilde{\rho}(0, \mu) = 0$ can in fact be viewed as a subjective mean of the
random variable $X$, which varies from the best scenario $\bar{X}$ to the worst scenario $\underline{X}$, covering the objective mean as a special case.

To judge the risk preference induced by a certain type of valuation functions, we follow the rule that diversification should be preferred if the agent is risk averse. More specifically, suppose an agent has two possible choices, one of which leads to the future reward $(X, \mu)$ while the other leads to the future reward $(Y, \nu)$. For simplicity we assume $\mu = \nu$. If the agent diversifies—and spends only a fraction $\alpha$ of the resources on the first and the remaining amount on the second alternative—the future reward is given by $\alpha X + (1 - \alpha)Y$. If the applied valuation function is concave,

$$\rho(\alpha X + (1 - \alpha)Y, \mu) \geq \alpha \rho(X, \mu) + (1 - \alpha) \rho(Y, \mu)$$

for all $\alpha \in [0, 1]$ and $X, Y \in \mathbb{R}^{|I|}$, then the diversification should increase the (subjective) valuation. Thus, we call the agent’s behavior risk averse. Conversely, if the applied valuation function is convex, the induced risk-preference should be risk seeking.

2.1 Utility-Based Shortfall. We now introduce a class of valuation functions, the utility-based shortfall, that generalizes many important special valuation functions in the literature. Let $u : \mathbb{R} \to \mathbb{R}$ be a utility function, which is continuous and strictly increasing. The shortfall $\rho^u_{x_0}$ induced by $u$ and an acceptance level $x_0$ is then defined as

$$\rho^u_{x_0}(X, \mu) := \sup \left\{ m \in \mathbb{R} \mid \sum_{i \in I} u(X(i) - m)\mu(i) \geq x_0 \right\}. \quad (2.2)$$

It can be shown (Föllmer & Schied, 2004) that $\rho^u_{x_0}$ is a valid valuation function satisfying the axioms. The utility-based shortfall was introduced in the mathematical finance literature (Föllmer & Schied, 2004). The class of utility functions considered here will be more general than the class of utility functions typically used in finance.

Compared to the expected utility theory, the utility function in equation 2.2 is applied to the relative value $X(i) - m$ rather than to the absolute outcome $X(i)$. This reflects the intuition that human beings judge utilities usually by comparing those outcomes with a reference value, which may not be zero. The property of $u$ being convex or concave determines the risk sensitivity of $\rho^u_{x_0}$: given a concave function $u$, $\rho$ is also concave and hence risk averse (see theorem 4.61 in Föllmer & Schied, 2004). Vice versa, $\rho$ is convex (hence risk seeking) for convex $u$.

Utility-based shortfalls cover a large family of valuation functions, which have been proposed in the literature in various fields:
• For $u(x) = x$ and $x_0 = 0$, one obtains the standard expected reward 
  \[ \rho(X, \mu) = \sum_i X(i) \mu(i). \]
• For $u(x) = e^{\lambda x}$ and $x_0 = 1$, one obtains 
  \[ \rho(X, \mu) = \frac{1}{\lambda} \log \left[ \sum_i \mu(i) e^{\lambda X(i)} \right] \]
  (the so-called entropic map; see Cavazos-Cadena, 2010). Expansion with regard to $\lambda$ leads to 
  \[ \rho(X, \mu) = \mathbb{E}[X] + \lambda \text{Var}[X] + O(\lambda^2), \]
  where $\text{Var}[X]$ denotes the variance of $X$ under the distribution $\mu$. Hence, the entropic map is risk averse if $\lambda < 0$ and risk seeking if $\lambda > 0$. In neuroscience, Nagengast et al. (2010) and Braun et al. (2011) applied this type of valuation function to test risk sensitivity in human sensorimotor control.
• Mihatsch and Neuneier (2002) proposed the following setting,
  \[ u(x) = \begin{cases} 
  (1 - \kappa)x & \text{if } x > 0 \\
  (1 + \kappa)x & \text{if } x \leq 0
  \end{cases}, \]
  where $\kappa \in (-1, 1)$ controls the degree of risk sensitivity. Its sign determines the property of the utility function $u$ being convex versus concave and therefore the risk preference of $\rho$. In a recent study, Niv et al. (2012) applied this type of valuation function to quantify the risk-sensitive behavior of human subjects and interpret the measured neural signals.

When quantifying human behavior, combined convex and concave utility functions,
  \[ u_p(x) = \begin{cases} 
  k_+ x_+ & x \geq 0 \\
  -k_- (-x)_- & x < 0
  \end{cases}, \quad (2.3) \]
  are of special interest, since people tend to treat gains and losses differently and therefore have different risk preferences on the gain and loss sides. In fact, the polynomial function in equation 2.3 was used in the prospect theory (Kahneman & Tversky, 1979) to model human risk preferences; the results show that $l_+$ is usually below 1 (i.e., $u_p(x)$ is concave) and thus risk averse on gains, while $l_-$ is also below 1 and $u_p(x)$ is therefore convex and risk seeking on losses.

2.2 Utility-Based Shortfall and Prospect Theory. To illustrate the risk preferences induced by different utility functions, we consider a simple example with two events. The first event has outcome $x_1$ with probability $p$, while the other event has smaller outcome $x_2 < x_1$ with $1 - p$. Note that 
  \[ p = \frac{\mathbb{E}X - x_2}{x_1 - x_2}, \]
  where $\mathbb{E}X = px_1 + (1 - p)x_2$ denotes the risk-neutral mean.
Replacing $E_X$ with the subjective mean $\tilde{\rho}(X, p) = \rho(X, p) - \rho(0, p)$ defined in equation 2.1, we can define a subjective probability (Tversky & Kahneman, 1992) as

$$w(p) := \frac{\tilde{\rho}(X, p) - x_2}{x_1 - x_2},$$

(2.4)

which measures agents’ subjective perception of the true probability $p$.

In risk-neutral cases, $\tilde{\rho}(X, p)$ is simply the mean and $w(p) = p$. In risk-averse cases, the balance moves toward the worst scenario. Hence, the probability of the first event (with larger outcome $x_1$) is always underestimated. On the contrary, in risk-seeking cases, the probability of the first event is always overestimated. Behavioral studies show that human subjects usually overestimate low probabilities and underestimate high probabilities (Tversky & Kahneman, 1992). This can be quantified by applying mixed valuation functions $\rho$. If we apply utility-based shortfalls, it can be quantified by using mixed utility function $u$.

Let $x_1 = 1, x_2 = -1$ and the acceptance level $x_0 = 0$. Figure 1 (left) shows five different utility functions: one linear function, lin; one convex function, RS; one concave function, RA; and two mixed functions, mix1 and mix2 (for details, see the caption). The corresponding subjective probabilities are shown in Figure 1 (right). Since the function, RA is concave, the corresponding valuation function is risk averse, and therefore the probability of high-reward event is always underestimated. For the case of the convex

Figure 1: Shortfalls with different utility functions and induced subjective probabilities. (Left) Utility functions defined as follows: lin : $x$; RS : $e^x - 1$; RA : $1 - e^{-x}$; mix1: $u_p(x)$ as defined in equation 2.3 with $k_+ = 0.5, l_+ = 2, k_- = 1$ and $l_- = 2$; mix2: same as mix1 but with $k_+ = 1, l_+ = 0.5, k_- = 1.5$ and $l_- = 0.5$. (Right) Subjective probability functions calculated according to equation 2.4.
function, RS, the probability of a high-reward event is always overestimated. However, since the mix1 function is convex on $[0, \infty)$ but concave on $(-\infty, 0]$, high probabilities are underestimated while low probabilities are overestimated, which replicates very well the probability weighting function applied in prospect theory for gains (see Figure 1 in Tversky & Kahneman, 1992). Conversely, the mix2 function, which is concave on $[0, \infty)$ and convex on $(-\infty, 0]$, corresponds to the overestimation of high probabilities and the underestimation of low probabilities. This corresponds to the weighting function used for losses in prospect theory (see Figure 2 in Tversky & Kahneman, 1992).

We will see in the following section that the advantage of using the utility-based shortfall is that we can derive iterating learning algorithms for estimating the subjective valuations, whereas it is difficult to derive such algorithms in the framework of prospect theory.

3 Risk-Sensitive Reinforcement Learning

A Markov decision process (see Puterman, 1994),

$$\mathcal{M} = \{S, (A, A(s), s \in S), P, (r, P_r)\},$$

consists of a state space $S$; admissible action spaces $A(s) \subset A$ at $s \in S$; a transition kernel $P(s'|s, a)$, which denotes the transition probability moving from one state $s$ to another state $s'$ by executing action $a$; and a reward function $r$ with its distribution $P_r$. In order to model random rewards, we assume that the reward function has the form\(^1\)

$$r(s, a, \varepsilon) : S \times A \times E \rightarrow \mathbb{R}.$$  

$E$ denotes the noise space with distribution $P_r(\varepsilon|s, a)$; given $(s, a)$, $r(s, a, \varepsilon)$ is a random variable with values drawn from $P_r(\cdot|s, a)$. Let $R(s, a)$ be the random reward gained at $(s, a)$, which follows the distribution $P_r(\cdot|s, a)$. The random state (resp. action) at time $t$ is denoted by $S_t$ (resp. $A_t$). Finally, we assume that all sets $S, A, E$ are finite.

A Markov policy $\pi = [\pi_0, \pi_1, \ldots]$ consists of a sequence of single-step Markov policies at times $t = 0, 1, \ldots$, where $\pi_t(A_t = a|S_t = s)$ denotes the probability of choosing action $a$ at state $s$. Let $\Pi$ be the set of all Markov policies. The optimal policy within a time horizon $T$ is obtained by maximizing

---

\(^1\)In standard MDPs, it is sufficient (Puterman, 1994) to consider the deterministic reward function $\bar{r}(s, a) := \sum_{\varepsilon \in E} r(s, a, \varepsilon)P_r(\varepsilon|s, a)$, that is, the mean reward at each $(s, a)$-pair. In risk-sensitive cases, random rewards cause also risk and uncertainties. Hence, we keep the generality by using random rewards.
the expectation of the discounted cumulative rewards,

\[ J(T, \pi, s) := \max_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=0}^{T} \gamma^t R(S_t, A_t) \mid S_0 = s, \pi \right], \tag{3.1} \]

where \( s \in S \) denotes the initial state and \( \gamma \in [0, 1) \) the discount factor. Expanding the sum leads to

\[
J(T, \pi, s) = \mathbb{E}_{S_0 = s}^{T_0}[R(S_0, A_0)] + \gamma \mathbb{E}_{S_1}^{T_1}[R(S_1, A_1)]
+ \cdots + \gamma \mathbb{E}_{S_T}^{T_T}[R(S_T, A_T)] \ldots. \tag{3.2}
\]

We now generalize the conditional expectation \( \mathbb{E}_{\pi}^{s} \) to represent the valuation functions considered in section 2. Let \( K := \{(s, a) \mid s \in S, a \in A(s)\} \) be the set of all admissible state-action pairs. Let

\[
I = S \times \mathbb{E} \quad \text{and} \quad \mu_{s, a}(s', \varepsilon) = \mathcal{P}(s' \mid s, a) \mathcal{P}_r(\varepsilon \mid s, a). \tag{3.3}
\]

A mapping \( \mathcal{U}(X, \mu \mid s, a) : \mathbb{R}^{|I|} \times \mathcal{P} \times K \rightarrow \mathbb{R} \) is called a valuation map, if for each \((s, a) \in K\), \( \mathcal{U}(\cdot \mid s, a) \) is a valuation function on \( \mathbb{R}^{|I|} \times \mathcal{P} \). Let \( \mathcal{U}_{s,a}(X, \mu) \) be a short notation of \( \mathcal{U}(X, \mu \mid s, a) \), and let

\[
\mathcal{U}_{s}^{\pi}(X, \mu) := \sum_{a \in A(s)} \pi(a \mid s) \mathcal{U}(X, \mu \mid s, a)
\]

be the valuation map averaged over all actions. Since \( \mu = \mu_{s, a} \) for each \((s, a) \in K\), we will omit \( \mu \) in \( \mathcal{U} \) in the following. When we replace the conditional expectation \( \mathbb{E}_{\pi}^{s} \) with \( \mathcal{U}_{s}^{\pi} \) in equation 3.2, the risk-sensitive objective becomes

\[
\tilde{J}_T(\pi, s) := \mathcal{U}_{s_0}^{T_0}[R(S_0, A_0)] + \gamma \mathcal{U}_{s_1}^{T_1}[R(S_1, A_1)]
+ \cdots + \gamma \mathcal{U}_{s_T}^{T_T}[R(S_T, A_T)] \ldots. \tag{3.4}
\]

The optimal policy is then given by \( \max_{\pi \in \Pi} \tilde{J}_T(\pi, s) \). For an infinite-horizon problem, we obtain

\[
\max_{\pi \in \Pi} \tilde{J}_T(\pi, s) := \lim_{T \to \infty} \tilde{J}_T(\pi, s), \tag{3.5}
\]

using the same line of argument.

The optimization problem for finite-stage objective function \( \tilde{J}_T \) can be solved by a generalized dynamic programming (Bertsekas &
Tsitsiklis, 1996), while the one defined in equation 3.5 requires the solution to the risk-sensitive Bellman equation:

$$V^*(s) = \max_{a \in A(s)} \mathcal{U}_{s,a}(R(s, a) + \gamma V^*).$$  \hspace{1cm} (3.6)

The latter is a consequence of the following theorem:

**Theorem 1** (theorem 5.5 in Shen et al., 2013). $V^*(s) = \max_\pi \bar{J}(\pi, s)$ holds for all $s \in S$, whenever $V^*$ satisfies equation 3.6. Furthermore, a deterministic policy $\pi^*$ is optimal if $\pi^*(s) = \arg\max_{a \in A(s)} \mathcal{U}_{s,a}(R + \gamma V^*)$.

Define $Q^*(s, a) := \mathcal{U}_{s,a}(R + \gamma V^*)$. Then equation 3.6 becomes

$$Q^*(s, a) = \mathcal{U}_{s,a} \left( R(s, a) + \gamma \max_{a \in A(s')} Q^*(s', a) \right). \forall (s, a) \in K. \hspace{1cm} (3.7)$$

To carry out value iteration algorithms, the MDP $M$ must be known a priori. In many real-life situations, however, the transition probabilities are unknown, as well as the outcome of an action before its execution. Therefore, an agent has to explore the environment while gradually improving its policy. We now derive RL-type algorithms for estimating Q-values of general valuation maps based on the utility-based shortfall, which do not require knowledge of the reward and transition model.

**Proposition 1** (see Proposition 4.104 in Föllmer & Schied, 2004). Let $\rho^u$ be a shortfall defined in equation 2.2, where $u$ is continuous and strictly increasing. Then the following statements are equivalent: (1) $\rho^u_{x_0}(X) = m^*$ and (2) $\mathbb{E}^\mu[u(X - m^*)] = x_0$.

For the proof, see the appendix.

Consider the valuation map induced by the utility-based shortfall,$^2$

$$\mathcal{U}_{s,a}(X) = \sup\{m \in \mathbb{R} | \mathbb{E}^{\mu_{s,a}}[u(X - m)] \geq x_0\},$$

where $\mu_{s,a}$ is defined in equation 3.3. If $\mathcal{U}_{s,a}(X) = m^*(s, a)$ exists, proposition 1 ensures that $m^*(s, a)$ is the unique solution to

$$\mathbb{E}^{\mu_{s,a}}[u(X - m^*(s, a))] = x_0.$$

$^2$In principle, we can apply different utility functions $u$ and acceptance levels $x_0$ at different $(s, a)$-pairs. However, for simplicity, we drop their dependence on $(s, a)$.
Let \( X = R + \gamma V^* \). Then \( m^*(s, a) \) corresponds to the optimal Q-value \( Q^*(s, a) \) defined in equation 3.7, which is equivalent to

\[
\sum_{s' \in S, \varepsilon \in E} P(s'|s, a)P_r(\varepsilon|s, a)u\left(r(s, a, \varepsilon) + \gamma \max_{a' \in A(s')} Q^*(s', a') - Q^*(s, a)\right)
= x_0, \forall (s, a) \in K.
\]  

(3.8)

Let \( \{s_t, a_t, s_{t+1}, r_t\} \) be the sequence of states, chosen actions, successive states, and received rewards. Analogous to the standard Q-learning algorithm, we consider the following iterative procedure,

\[
Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \times \left[u\left(r_t + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)\right) - x_0\right],
\]  

(3.9)

where \( \alpha_t \geq 0 \) denotes a learning rate function that satisfies \( \alpha_t(s, a) > 0 \) only if \( (s, a) \) is updated at time \( t \), that is, \( (s, a) = (s_t, a_t) \). In other words, for all \( (s, a) \) that are not visited at time \( t \), \( \alpha_t(s, a) = 0 \) and their Q-values are not updated. Consider utility functions \( u \) with the following properties.

**Assumption 1.** (i) The utility function \( u \) is strictly increasing, and there exists some \( y_0 \in \mathbb{R} \) such that \( u(y_0) = x_0 \). (ii) There exist positive constants \( \epsilon, L \) such that \( 0 < \epsilon \leq \frac{u(x) - u(y)}{x - y} \leq L \), for all \( x \neq y \in \mathbb{R} \).

Then the following theorem holds (for the proof, see section A.1):

**Theorem 2.** Suppose assumption 1 holds. Consider the generalized Q-learning algorithm stated in equation 3.9. If the nonnegative learning rates \( \alpha_t(s, a) \) satisfy

\[
\sum_{t=0}^{\infty} \alpha_t(s, a) = \infty \text{ and } \sum_{t=0}^{\infty} \alpha_t^2(s, a) < \infty, \forall (s, a) \in K,
\]  

(3.10)

then \( Q_t(s, a) \) converges to \( Q^*(s, a) \) for all \( (s, a) \in K \) with probability 1.

The assumption in equation 3.10 requires in fact that all possible state-action pairs be visited infinitely often. Otherwise the first sum in equation 3.10 would be bounded by the setting of the learning rate function \( \alpha_t(s, a) \). It means that, similar to the standard Q-learning, the agent has to explore the whole state-action space for gathering sufficient information about the environment. Hence, it cannot take too greedy a policy in the learning procedure before the state-action space is well explored. We call
A typical policy, which is widely applied in RL literature as well as in models of human reward-based learning, is given by

$$a_t \sim p(a_t|s_t) := \frac{e^{\beta Q_t(s_t,a_t)}}{\sum_a e^{\beta Q_t(s_t,a)}}.$$  \hspace{1cm} (3.11)

where $\beta \in [0, \infty)$ controls how greedy the policy should be. In section A.4, we prove that under some technical assumptions on the transition kernel of the underlying MDP, this policy is always proper. A widely used setting satisfying both conditions in equation 3.10 is to let $\alpha_t(s,a) := 1/N_t(s,a)$, where $N_t(s,a)$ counts the number of times of visiting the state-action pair $(s,a)$ up to time $t$ and is updated trial-by-trial. This leads to the learning procedure shown in algorithm 1 (see also Figure 2).

The expression

$$TD_t := r_t + \gamma \max_a Q_{t}(s_{t+1}, a) - Q_t(s,a)$$

inside the utility function of equation 3.9 corresponds to the standard temporal difference (TD) error. Comparing equation 3.9 with the standard Q-learning algorithm, we find that the nonlinear utility function is applied to the TD error (see Figure 2). This induces nonlinear transformation not only of the true rewards but also of the true transition probabilities, as shown in section 2.1. By applying an S-shape utility function, which is partially convex and partially concave, we can replicate key effects of prospect theory without the explicit introduction of a probability-weighting function.

Assumption 1ii seems to exclude several important types of utility functions. The exponential function $u(x) = e^x$ and the polynomial function $u(x) = x^p$, $p > 0$, for example, do not satisfy the global Lipschitz condition required in assumption 1ii. This problem can be solved by a truncation.
Figure 2: Illustration of risk-sensitive Q-learning (see algorithm 1). The value function $Q(s, a)$ quantifies the current subjective evaluation of each state-action pair $(s, a)$. The next action is then randomly chosen according to a proper policy (e.g., equation 3.11), which is based on the current values of $Q$. After interacting with the environment, the agent obtains the reward $r$ and moves to the successor $s'$. The value function $Q(s, a)$ is then updated by the rule given in equation 3.9. This procedure continues until some stopping criterion is satisfied.

when $x$ is very large and by an approximation when $x$ is very close to 0. For more details see sections A.2 and A.3.

4 Modeling Human Risk-Sensitive Decision Making

4.1 Experiment. Subjects were told that they are influential stockbrokers, whose task is to invest into a fictive stock market (Tobia et al., 2014). At every trial (see Figure 3a) subjects had to decide how much ($a = 0, 1, 2, \text{ or } 3 \text{ EUR}$) to invest in a particular stock. After the investment, subjects first saw the change of the stock price and then were informed how much money they earned or lost. The received reward was proportional to the investment. The different trials, however, were not independent of each other (see Figure 3b). The sequential investment game consisted of seven states, each one coming with a different set of contingencies, and subjects were transferred from one state to the next depending on the amount of money they invested. For high investments, transitions followed the path labeled “risk seeking” (RS in Figure 3b). For low investments, transitions followed the path labeled “risk averse” (RA in Figure 3b). After three decisions, subjects were always transferred back to the initial state, and the reward, which was accumulated during this round, was shown. State information was available to the subjects throughout every trial (see Figure 3a). Altogether, 30 subjects (young healthy adults) experienced 80 rounds of the three-decision sequence.

Formally, the sequential investment game can be considered as an MDP with seven states and four actions (see Figure 3b). Depending on the strategy
Figure 3: The sequential investment paradigm. The paradigm is an implementation of a Markov decision process with seven states and four possible actions (decisions to take) at every state. (a) Every decision (trial) consists of a choice phase (3 s), during which an action (invest 0, 1, 2, or 3 euros) must be taken by adjusting the scale bar on the screen, an anticipation phase (.5 s); an outcome phase (2–5 s), where the development of the stock price and the reward (wins and loses) are revealed; an evaluation phase (2–5 s), where it reveals the maximal possible reward that could have been obtained for the (in hindsight) best possible action; and a transition phase (2.7 s), where subjects are informed about the possible successor states and the specific transition that will occur. The intervals of the outcome and evaluation phase are jittered for improved fMRI analysis. State information is provided by different patterns, the black field provides stock price information during the anticipation phase, and the white field provides the reward and the maximal possible reward of this trial. After each round (three trials), the total reward of this round is shown to subjects. (b) Structure of the underlying Markov decision process. The seven states are indicated by numbered circles; arrows denote the possible transitions. The labels RS and RA indicate the transitions caused by the two risk-seeking (investment of 2 or 3 euros) and the two risk-averse (investment of 0 or 1 euro) actions. Bigaussian distributions with a standard deviation of 5 are used to generate the random price changes of the stocks. Panels next to the states provide information about the means (top row) and the probabilities (center row) of every component. M (bottom row) denotes the mean price change. The reward received equals the price change multiplied by the amount of money the subject invests. The right-most panels provide the total expected rewards (EV) and the standard deviations (SD) for all possible state sequences (path 1 to path 4) under the assumption that every sequence of actions consistent with a particular sequence of states is chosen with equal probability.

of the subjects, there are four possible paths, each composed of three states. The total expected return for each path, averaged over all policies consistent with it, is shown in the right panels of Figure 3b (EV). Path 1 provides the largest expected return per round (EV = 90), while path 4 leads to an average
loss of $-9.75$. Hence, to follow the on-average highest rewarded path 1, subjects have to take risky actions (investing 2 or 3 euros at each state). Always taking conservative actions (investing 0 or 1 euros) results in path 4 and a high on-average loss. Since the standard deviation of the return $R$ of each state equals $\text{std}(R) = a \times C$, where $a$ denotes the action (investment) the subject takes and $C$ denotes the price change, the higher the investment, the higher the risk. Path 1 therefore has the highest standard deviation (SD = 14.9) of the total average reward, whereas the standard deviation of path 4 is the smallest (SD = 6.9). Path 3 provides a trade-off option: it has slightly lower expected value (EV = 52.25) than path 1 but comes with a lower risk (SD = 12.3). Hence, the paradigm is suitable for observing and quantifying the risk-sensitive behavior of subjects.

4.2 Risk-Sensitive Model of Human Behavior. Figure 4 summarizes the strategies chosen by the 30 subjects. Seventeen subjects mainly chose path 1, which provided them high rewards. Six subjects chose path 4, which gave very low rewards. The remaining seven subjects show no significant preference among all four paths, and the rewards they received are an average between the rewards received by the other two groups. The optimal policy for maximizing the expected reward is the policy that follows path 1. The results shown in Figure 4, however, indicate that the standard model fails to explain the behavior of more than 40% of the subjects.
Table 1: Parameters for the Two Branches $x \geq 0$ (Left) and $x < 0$ (Right) of the Polynomial Utility Function $u(x)$ (Equation 4.2), Its Shape and the Induced Risk Preference.

<table>
<thead>
<tr>
<th>Branch $x \geq 0$</th>
<th>Shape</th>
<th>Risk Preference</th>
<th>Branch $x &lt; 0$</th>
<th>Shape</th>
<th>Risk Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; l_+ &lt; 1$</td>
<td>Concave</td>
<td>Risk Averse</td>
<td>$0 &lt; l_- &lt; 1$</td>
<td>Convex</td>
<td>Risk Seeking</td>
</tr>
<tr>
<td>$l_+ = 1$</td>
<td>Linear</td>
<td>Risk Neutral</td>
<td>$l_- = 1$</td>
<td>Linear</td>
<td>Risk Neutral</td>
</tr>
<tr>
<td>$l_+ &gt; 1$</td>
<td>Convex</td>
<td>Risk Seeking</td>
<td>$l_- &gt; 1$</td>
<td>Concave</td>
<td>Risk Averse</td>
</tr>
</tbody>
</table>

We now quantify subjects’ behavior by applying three classes of Q-learning algorithm: standard Q-learning, the risk-sensitive Q-learning (RSQL) method described by algorithm 1, and an expected utility (EU) algorithm with the following update rule,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(u(r_t) - x_0 + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right),$$

(4.1)

where the nonlinear transformation is applied to the reward $r_t$ directly. The latter is a straightforward extension of expected utility theory. Risk sensitivity is implemented via the nonlinear transformation of the true reward $r_t$. For both risk-sensitive Q-learning methods (RSQL and EU), we set the reference level $x_0 = 0$ and consider the family of polynomial mixed utility functions

$$u(x) = \begin{cases} k_+ x_+^l & x \geq 0 \\ -k_- (-x)^l & x < 0 \end{cases}.$$  

(4.2)

The parameters $k_\pm > 0$ and $l_\pm > 0$ quantify the risk preferences separately for wins and losses (see Table 1). Hence, there are four parameters for $u$ that have to be determined from the data. For all three classes, actions are generated according to the softmax policy (see equation 3.11) which is a proper policy for the paradigm (for the proof see section A.4), and the learning rate $\alpha$ is set constant across trials.

For RSQL, the learning rate is absorbed by the coefficients $k_\pm$. Hence, six parameters $\{\beta, \gamma, k_\pm, l_\pm\} =: \theta$ have to be determined. Standard Q-learning corresponds to the choice $l_\pm = 1$ and $k_\pm = \alpha$. The risk-sensitive model applied by Niv et al. (2012) is also a special case of the RSQL framework and corresponds to $l_\pm = 1$. For the EU algorithm, there are seven parameters, $\{\alpha, \beta, \gamma, k_\pm, l_\pm\} =: \theta$, which have to be fitted to the data. $l_\pm = 1$ and $k_\pm = 1$ again correspond to the standard Q-learning method.
Parameters were determined subject-wise by maximizing the log likelihood of the subjects’ action sequences,

$$\max_{\theta} L(\theta) := \sum_{t=1}^{T} \log p(a_t|s_t, \theta) = \sum_{t=1}^{T} \log \frac{e^{\beta Q(s_t, a_t|\theta)}}{\sum_{a} e^{\beta Q(s_t, a|\theta)}}$$

(4.3)

where $Q(s, a|\theta)$ indicates the dependence of the Q-values on the model parameters $\theta$. Since RSQL/EU and the standard Q-learning are nested model classes, we apply the Bayesian information criterion (BIC; see Ghosh, Delampady, & Samanta, 2006),

$$B := -2L + k \log(n),$$

for model selection. $L$ denotes the log likelihood, equation 4.3. $k$ and $n$ are the number of parameters and trials, respectively.

To compare results, we report relative BIC scores, $\Delta B := B - B_Q$, where $B$ is the BIC score of the candidate model and $B_Q$ is the BIC score of the standard Q-learning model. We obtain

$$\Delta B = -500.14 \quad \text{for RSQL},$$

$$\Delta B = -23.10 \quad \text{for EU}.$$

The more negative the relative BIC score is, the better the model fits data. Hence, the RSQL algorithm provides a significantly better explanation for the behavioral data than the EU algorithm and standard Q-learning. In the following, we discuss only the results obtained with the RSQL model.

Figure 5 shows the distribution of best-fitting values for the two parameters $l_{\pm}$, which quantify the risk preferences of the individual subjects. We conclude (see Table 1) that most of the subjects are risk averse for positive and risk seeking for negative TD errors. The result is consistent with previous studies from the economics literature (see Tversky & Kahneman, 1992).

After determining the parameters $\{k_{\pm}, l_{\pm}\}$ for the utility functions, we perform an analysis similar to that discussed in section 2.2. Given an observed reward sequence $r_{i}^{N}_{i=1}$, the empirical subjective mean $m_{sub}$ is obtained by solving the following equation:

$$\frac{1}{N} \sum_{i=1}^{N} u(r_i - m_{sub}) = 0.$$

If subjects are risk neutral, then $u(x) = x$, and $m_{sub} = m_{emp} = \frac{1}{N} \sum_{i=1}^{N} r_i$ is simply the empirical mean. Following the idea of prospect theory, we define
Figure 5: Distribution of values for the shape parameters $l_+$ (left) and $l_-$ (right) for the RSQL model.

Figure 6: Distribution of normalized subjective probabilities, $\Delta p$, equation 4.4, for the different subject groups defined in Figure 4.

A normalized subjective probability $\Delta p$:

$$\Delta p := \frac{m_{sub} - \min_i r_i}{\max_i r_i - \min_i r_i} - \frac{m_{emp} - \min_i r_i}{\max_i r_i - \min_i r_i} = \frac{m_{sub} - m_{emp}}{\max_i r_i - \min_i r_i}.$$  

(4.4)

If $\Delta p$ is positive, the probability of rewards is overestimated and the induced policy is therefore risk seeking. If $\Delta p$ is negative, the probability of rewards is underestimated, and the policy is risk averse. Figure 6 summarizes the distribution of normalized subjective probabilities for subjects from the path 1, path 4 and random groups of Figure 4. For subjects within
Figure 7: Modulation of the fMRI BOLD signal by TD errors (a) and by Q-values (b) generated by the RSQL model with best-fitting parameters. The data are shown whole-brain-corrected to \( p < .05 \) (voxel-wise \( p < .001 \) and minimum 125 voxels). The color bar indicates the \( t \)-value ranging from 0 to the maximal value. The cross indicates location of strongest modulation for TD errors (in the left ventral striatum (−14, 8−16)) and for Q-values (in the right ventral striatum (14, 8−4)). However, it is remarkable that for both TD errors and Q-values, modulations in the left and right ventral striatum are almost equally strong with a slight difference.

Figure 7a shows that the sequence of TD errors for the RSQL model (with the best-fitting parameters) positively modulated the BOLD signal in the subcallosal gyrus extending into the ventral striatum (−14, 8−16).

4.3 fMRI Results. Functional magnetic resonance imaging (fMRI) data were simultaneously recorded while subjects played the sequential investment game. The analysis of the data was conducted in SPM8 (Wellcome Department of Cognitive Neurology, London, UK; details of the magnetic resonance protocol and data processing are presented in appendix B). The sequences of Q-values for the action chosen at each state were used as parametric modulators during the choice phase, and temporal difference (TD) errors were used at the outcome phase (see Figure 3a).

Figure 7a shows that the sequence of TD errors for the RSQL model (with the best-fitting parameters) positively modulated the BOLD signal in the subcallosal gyrus extending into the ventral striatum (−14, 8−16).
(marked by the cross in Figure 7a), the anterior cingulate cortex (8 48 6), and the visual cortex (−8 −92 16; z = 7.9). The modulation of the BOLD signal in the ventral striatum is consistent with previous experimental findings (Schultz, 2002; O’Doherty, 2004) and supports the primary assertion of computational models that reward-based learning occurs when expectations (here, expectations of “subjective” quantities) are violated (Sutton & Barto, 1998).

Figure 7b shows the results for the sequence of Q-values for the RSQL model (with best-fitting parameters), which correspond to the subjective (risk-sensitive) expected value of the reward for each discrete choice. Several large clusters of voxels in cortical and subcortical structures were significantly modulated by the Q-values at the moment of choice. The sign of this modulation was negative. The peak of this negative modulation occurred in the left anterior insula (−36 12 −2; z = 4.6), with strong modulation also in the bilateral ventral striatum (14 8 −4, marked by the cross in Figure 7b; −16 4 0) and the cingulate cortex (4 16 28). The reward prediction error processed by the ventral striatum (and other regions noted above) would not be computable in the absence of an expectation, and as such, this activation is important for substantiating the plausibility for the RSQL model of learning and choice. Sequences of Q-values obtained with standard Q-learning (with best-fitting parameters), on the other hand, failed to predict any changes in brain activity even at a liberal statistical threshold of \( p < .01 \) (uncorrected). This lack of neural activity for the standard Q model, in combination with the significant activation for our RSQL, supports the hypothesis that some assessment of risk is induced and influences valuation. Whereas the areas modulated by Q-values differ from what has been reported in other studies (i.e., the ventromedial prefrontal cortex as in Gläscher, Hampton, & O’Doherty, 2009), it does overlap with the representation of TD errors. Furthermore, the opposing signs of the correlated neural activity suggest that a neural mismatch of signals in the ventral striatum between Q-value and TD errors may underlie the mechanism by which values are learned.

4.4 Discussion. We applied the risk-sensitive Q-learning (RSQL) method to quantify human behavior in a sequential investment game and investigated the correlation of the predicted TD- and Q-values with the neural signals measured by fMRI.

We first showed that the standard Q-learning algorithm cannot explain the behavior of a large number of subjects in the task. Applying RSQL generated a significantly better fit and also outperformed the expected utility algorithm. The risk sensitivity revealed by the best-fitting parameters is consistent with the studies in behavioral economics, that is, subjects are risk averse for positive while risk seeking for negative TD errors. Finally, the relative subjective probabilities provide a good explanation as to why
some subjects have conservative policies: they underestimate the true probabilities of gaining rewards.

The fMRI results showed that the TD sequence generated by our model has a significant correlation with the activity in the subcallosal gyrus extending into the ventral striatum. The sequence of Q-values has a significant correlation with the activities in the left anterior insula. Previous studies (see chapter 23 of Glimcher et al., 2008, and Symmonds et al., 2011) suggest that higher-order statistics of outcomes, such as variance (the second order) and skewness (the third order), are encoded in human brains separately and the individual integration of these risk metrics induces the corresponding risk-sensitivity. Our results indicate, however, that risk-sensitivity can be simply induced (and therefore encoded) by a nonlinear transformation of TD errors; no additional neural representation of higher-order statistics is needed.

5 Summary

We applied a family of valuation functions, the utility-based shortfall, to the general framework of risk-sensitive Markov decision processes and derived a risk-sensitive Q-learning algorithm. We proved that the proposed algorithm converges to the optimal policy corresponding to the risk-sensitive objective. By applying S-shape utility functions, we show that key features predicted by prospect theory can be replicated using the proposed algorithm. Hence, the novel Q-learning algorithm provides a good candidate model for human risk-sensitive sequential decision-making procedures in learning tasks where mixed risk-preferences are shown in behavioral studies. We applied the algorithm to model human behaviors in a sequential investment game. The results showed that the new algorithm fit the data significantly better than the standard Q-learning and the expected utility model. The analysis of fMRI data shows a significant correlation of the risk-sensitive TD error with the BOLD signal change in the ventral striatum and also a significant correlation of the risk-sensitive Q-values with neural activity in the striatum, cingulate cortex, and insula, which is not present if standard Q-values are applied.

Some technical extensions are possible within our general risk-sensitive reinforcement learning (RL) framework. First, the Q-learning algorithm derived in this letter can be regarded as a special type of RL algorithms, TD(0). It can be extended to other types of RL algorithms like SARSA and TD(λ) for λ ≠ 0. Second, in our previous work (Shen et al., 2013), we also provided a framework for the average case. Hence, RL algorithms for the average case can also be derived similar to the discounted case considered in this letter. Finally, the algorithm in its current form applies to MDPs with finite state spaces only. It can be extended for MDPs with continuous state spaces by applying function approximation technique.
Appendix A: Mathematical Proofs

The sup-norm is defined as $|X|_\infty := \max_{i \in I} |X(i)|$, where $X = [X(i)]_{i \in I}$ can be considered as a $|I|$-dimensional vector.

**Lemma 1.** Let $\rho$ be valuation function on $\mathbb{R}^{|I|} \times \mathcal{P}$ and $\tilde{\rho}(X, \mu) := \rho(X, \mu) - \rho(0, \mu)$. Then the following inequality holds:

$$\min_{i \in I} X_i =: X \leq \tilde{\rho}(X, \mu) \leq \max_{i \in I} X_i, \forall \mu \in \mathcal{P}, X \in \mathbb{R}^{|I|}.$$

**Proof.** By $X \leq X_i \leq X$, $\forall i \in I$ and monotonicity of valuation functions, we obtain

$$\rho(X, \mu) \leq \rho(X_i, \mu) \leq \rho(0, \mu) + X_i,$$

Due to the translation invariance, we then have

$$\rho(X, \mu) = \rho(0, \mu) + X,$$

which immediately implies that

$$\min_{i \in I} X_i =: X \leq \rho(X, \mu) \leq \rho(0, \mu) + X, \forall \mu \in \mathcal{P}, X \in \mathbb{R}^{|I|}.$$

**Proof of Proposition 1.** (ii) $\Rightarrow$ (i). By definition, $m^* \leq \rho^u_{x_0}(X)$. For any $\epsilon > 0$, since $u$ is strictly increasing, we have $u(X(i) - m^* - \epsilon) < u(X(\omega) - m^*), \forall i \in I$, which implies $\mathbb{E}u(X - m^* - \epsilon) < \mathbb{E}u(X - m^*) = x_0$. Hence, $m^* = \rho^u_{x_0}(X)$.

(i) $\Rightarrow$ (ii). By definition we have $\mathbb{E}u(X - m^*) \geq x_0$. Assume that $\mathbb{E}u(X - m^*) > x_0$. By the continuity of $u$, there exists an $\epsilon > 0$ such that $\mathbb{E}u(X - m^* - \epsilon) > x_0$, which implies $\rho^u_{x_0}(X) \geq m^* + \epsilon > m^*$ and hence contradicts (i). Thus, (ii) holds.

**A.1 Proofs for Risk-Sensitive Q-Learning.** Before proving the risk-sensitive Q-learning, we consider a more general update rule,

$$q_{t+1}(i) = (1 - \alpha_t(i))q_t(i) + \alpha_t(i) \left[ (Hq_t)(i) + w_t(i) \right],$$

(A.1)

where $q_t \in \mathbb{R}^d$, $H : \mathbb{R}^d \to \mathbb{R}^d$ is an operator, $w_t$ denotes some random noise term, and $\alpha_t$ is learning rate with the understanding that $\alpha_t(i) = 0$ if $q(i)$
is not updated at time \( t \). Denote by \( \mathcal{F}_t \) the history of the algorithm up to time \( t \),

\[
\mathcal{F}_t = \{q_0(i), \ldots, q_t(i), w_0(i), \ldots, w_{t-1}(i), \alpha_0(i), \ldots, \alpha_t(i), i = 1, \ldots, t\}.
\]

We restate the following proposition:

**Proposition 2.** (proposition 4.4, in Bertsekas & Tsitsiklis, 1996). Let \( q_t \) be the sequence generated by iteration A.1. We assume the following

a. The learning rates \( \alpha_t(i) \) are nonnegative and satisfy

\[
\sum_{t=0}^{\infty} \alpha_t(i) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2(i) = \infty, \forall i.
\]

b. The noise terms \( w_t(i) \) satisfy (i) for every \( i \) and \( t \), \( \mathbb{E}[w_t(i)|\mathcal{F}_t] = 0; \) (ii) Given some norm \( \| \cdot \| \) on \( \mathbb{R}^d \), there exist constants \( A \) and \( B \) such that \( \mathbb{E}[w_t^2(i)|\mathcal{F}_t] \leq A + B\|q_t\|^2 \).

c. The mapping \( H \) is a contraction under sup-norm.

Then \( q_t \) converges to the unique solution \( q^* \) of the equation \( Hq^* = q^* \) with probability 1.

To apply proposition 2, we first reformulate the Q-learning rule, equation 3.9, in a different form,

\[
q_{t+1}(s, a) = \left(1 - \frac{\alpha_t(s, a)}{\alpha}\right) q_t(s, a) + \frac{\alpha_t(s, a)}{\alpha} \left[\alpha u(d_t) - x_0 + q_t(s, a)\right]
\]

where \( \alpha \) denotes an arbitrary constant such that \( \alpha \in (0, \min(L^{-1}, 1)] \). Recall that \( L \) is defined in assumption 1. For simplicity, we define \( \tilde{u}(x) := u(x) - x_0 \), \( d_t := r_t + \gamma \max_a q_t(s_{t+1}, a) - q_t(s, a) \) and set

\[
(Hq_t)(s, a) = \alpha \mathbb{E}_{s',a} \tilde{u}(r_t + \gamma \max_a q_t(s_{t+1}, a) - q_t(s, a)) + q_t(s, a).
\] (A.2)

\[
w_t(s, a) = \alpha \tilde{u}(d_t) - \alpha \mathbb{E}_{s,a} \tilde{u}(r_t + \gamma \max_a q_t(s_{t+1}, a) - q_t(s, a)).
\] (A.3)

More explicitly, \( Hq \) is defined as

\[
(Hq)(s, a) = \alpha \sum_{s', \epsilon} \tilde{P}(s', \epsilon|s,a) \tilde{u}\left(r(s, a, \epsilon) + \gamma \max_{a'} q(s', a') - q(s, a)\right)
\]

\[
+ q(s, a).
\]
where \( \tilde{P}(s', \epsilon | s, a) := P(s'|s, a)P_{r}(\epsilon | s, a) \). We assume the size of the space \( K \) is \( d \).

**Lemma 2.** Suppose that assumption 1 holds and \( 0 < \alpha \leq \min(L^{-1}, 1) \). Then there exists a real number \( \bar{\alpha} \in [0, 1) \) such that for all \( q, q' \in \mathbb{R}^{d} \), \( \| Hq - Hq' \|_{\infty} \leq \bar{\alpha} \| q - q' \|_{\infty} \).

**Proof.** Define \( v(s) := \max_a q(s, a) \) and \( v'(s) := \max_a q'(s, a) \). Thus,

\[
|v(s) - v(s)| \leq \max_{(s, a) \in K} |q(s, a) - q'(s, a)| = \|q - q'\|_{\infty}.
\]

By assumption 1ii and the monotonicity of \( \tilde{u} \), there exists a \( \xi_{(x,y)} \in [\epsilon, L] \) such that \( \tilde{u}(x) - \tilde{u}(y) = \xi_{(x,y)}(x - y) \). Analogously, we obtain

\[
(Hq)(s, a) - (Hq')(s, a) = \sum_{s', \epsilon} \tilde{P}(s', \epsilon | s, a) \left\{ \alpha \xi_{(s,a,\epsilon,s',q,q')} [\gamma v(s') - \gamma v'(s') - q(s, a) + q'(s, a)] + (q(s, a) - q'(s, a)) \right\}
\]

\[
= \alpha \gamma \sum_{s', \epsilon} \tilde{P}(s', \epsilon | s, a) \xi_{(s,a,\epsilon,s',q,q')} [v(s') - v'(s')] + (1 - \alpha) \sum_{s', \epsilon} \tilde{P}(s', \epsilon | s, a) \xi_{(s,a,\epsilon,s',q,q')} [q(s, a) - q'(s, a)]
\]

\[
\leq \left( 1 - \alpha (1 - \gamma) \right) \sum_{s', \epsilon} \tilde{P}(s', \epsilon | s, a) \xi_{(s,a,\epsilon,s',q,q')} \|q - q'\|_{\infty}
\]

\[
\leq (1 - \alpha (1 - \gamma) \epsilon) \|q - q'\|_{\infty}.
\]

Hence, \( \bar{\alpha} = 1 - \alpha (1 - \gamma) \epsilon \) is the required constant.

**Proof of Theorem 2.** Obviously, condition a in proposition 2 is satisfied, and condition c holds due to lemma 2. It remains to check condition b. \( \mathbb{E}[w_{1}(s, a)|F_{t}] = 0 \) holds by its definition in equation A.3. Next we prove (ii). \( \mathbb{E}[w_{2}^{2}(s, a)|F_{t}] = \alpha^{2} \mathbb{E} \left[ (\tilde{u}(d_{t}))^{2}|F_{t} \right] - \alpha^{2} (\mathbb{E} [\tilde{u}(d_{t})|F_{t}])^{2} \leq \alpha^{2} \mathbb{E} \left[ (\tilde{u}(d_{t}))^{2}|F_{t} \right] \).

Let \( \bar{R} \) be the upper bound for \( r_{t} \). Then \( |d_{t}| \leq \bar{R} + 2 \|q_{t}\|_{\infty} \), which implies that \( |\tilde{u}(d_{t}) - \tilde{u}(0)| \leq L(\bar{R} + 2 \|q_{t}\|_{\infty}) \) due to assumption 1ii. Hence, \( |\tilde{u}(d_{t})| \leq \bar{R} + \frac{2L}{\alpha^{2}} \).
\(|\tilde{u}(0)| + L(\bar{R} + 2\|q_i\|_\infty)\). On the other hand, since
\[
(|\tilde{u}(0)| + L\bar{R} + 2L\|q_i\|_\infty)^2 \leq 2(|\tilde{u}(0)| + L\bar{R})^2 + 8L^2\|q_i\|_\infty^2.
\]
we have \(\alpha^2 \mathbb{E}(d_t^2 | \mathcal{F}_t) \leq 2\alpha^2 (|\tilde{u}(0)| + L\bar{R})^2 + 8\alpha^2 L^2\|q_i\|_\infty^2\). Hence, condition b holds.

**A.2 Truncated Algorithms with Weaker Assumptions.** Some functions, like \(u(x) = e^x\) and \(u(x) = x^p, p > 0\), do not satisfy the global Lipschitz condition required in assumption 1ii. In actual applications, however, we can relax the assumption to assume that the Lipschitz condition holds locally within a “sufficiently large” subset. Lemma 4 states such a subset, provided the upper bound of absolute value of rewards is known.

**Assumption 2.** The reward function \(r(s, a, \epsilon)\) is bounded under sup-norm, that is,
\[
\bar{R} := \sup_{(s, a) \in \mathcal{K}, \epsilon \in \mathcal{E}} |r(s, a, \epsilon)| < \infty.
\]

Define an operator \(T : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}\) as
\[
T_s(V) = \max_{a \in \mathcal{A}(s)} \mathcal{U}_{s,a}(R(s, a) + \gamma V).
\]

**Lemma 3.** (see lemma 5.4 in Shen et al., 2013) \(T\) is a contracting map under sup-norm:
\[
|T(V) - T(V')|_\infty \leq \gamma \|V - V'\|_\infty, \forall V, V' \in \mathbb{R}^{|S|}.
\]

**Lemma 4.** Under assumptions 1i and 2, applying the valuation map in equation 3.8, the solution \(Q^*\) satisfies \(\frac{-\bar{R} - y_0}{1-\gamma} \leq Q^*(s, a) \leq \frac{\bar{R} - y_0}{1-\gamma}, \forall (s, a) \in \mathcal{K}\).

**Proof.** By assumption, \(u^{-1}(x_0)\) exists. Since \(u\) is strictly increasing, we have \(\mathcal{U}_{s,a}(0) = \sup \{m \in \mathbb{R} | u(-m) \geq x_0\} = -u^{-1}(x_0)\). Hence, together with equation 2.1, we obtain for all \((s, a) \in \mathcal{K}\),
\[
-u^{-1}(x_0) - \bar{R} = \mathcal{U}_{s,a}(0) - \bar{R} \leq \mathcal{U}_{s,a}(R) \leq \mathcal{U}_{s,a}(0) + \bar{R} = -u^{-1}(x_0) + \bar{R}.
\]
Note that lemma 3 implies that $V^* = T^\infty(V_0)$ for any $V_0 \in \mathbb{R}^{|S|}$. Without loss of generality, we start from $V_0 = 0$. Define $u := -u^{-1}(x_0) - \bar{R}$ and $\bar{u} := -u^{-1}(x_0) + \bar{R}$. Hence, we have $u \leq T(0) = \max_a U_{s,a}(R) \leq \bar{u}$, which implies

$$T^2(0) = \max_a U_{s,a}(R + \gamma T(0)) \leq \max_a U_{s,a}(R) + \gamma \bar{u} \leq (1 + \gamma)\bar{u},$$

and

$$T^2(0) = \max_a U_{s,a}(R + \gamma T(0)) \geq \max_a U_{s,a}(R) + \gamma u \geq (1 + \gamma)u.$$

Repeating the above procedure, we obtain $(1 + \gamma + \cdots + \gamma^{n-1})u \leq T^n(0) \leq (1 + \gamma + \cdots + \gamma^{n-1})\bar{u}$. Hence, $\frac{u}{1 - \gamma} \leq V^* = T^\infty(0) \leq \frac{\bar{u}}{1 - \gamma}$. By the definition of $Q^*$, the above inequalities hold for $Q^*$ as well.

Define

$$\bar{x} := y_0 - \frac{2\bar{R}}{1 - \gamma} \quad \text{and} \quad \bar{x} := y_0 + \frac{2\bar{R}}{1 - \gamma}. \quad (A.4)$$

Given lemma 4, we can truncate the utility function $u$ outside the interval $[\bar{x}, \bar{x}]$ as

$$u'(x) = \begin{cases} 
    u(\bar{x}) + \epsilon(x - \bar{x}), & x \in (-\infty, \bar{x}) \\
    u(x), & x \in [\bar{x}, \bar{x}] \\
    u(\bar{x}) + \epsilon(x - \bar{x}), & x \in (\bar{x}, \infty)
\end{cases} \quad (A.5)$$

**Theorem 3.** Suppose that assumption 1i and 2 hold. Assume further that there exist positive constants $\epsilon, L \in \mathbb{R}^+$ such that $0 < \epsilon \leq \frac{u(x) - u(y)}{x - y} \leq L$, for all $x \neq y \in [\bar{x}, \bar{x}]$, where $\bar{x}, \bar{x}$ are defined in equation A.4. Then the unique solution $Q^*_1$ to equation 3.8 with $u$ and the unique solution $Q^*_2$ to equation 3.8 with $u'$ are identical.

**Proof.** Both uniquenesses are due to theorem 1 and proposition 1. By lemma 4, $\frac{-R - y_0}{1 - \gamma} \leq Q^*_i(s, a) \leq \frac{-\bar{R} - y_0}{1 - \gamma}$ hold for all $(s, a) \in K$ and $i = 1, 2$. Hence, we have for both $i = 1, 2$ and for all $(s, a), (s', a') \in K, \epsilon \in E,$

$$y_0 - \frac{2\bar{R}}{1 - \gamma} \leq r(s, a, \epsilon) + \gamma Q^*_i(s', a') - Q^*_i(s, a) \leq y_0 + \frac{2\bar{R}}{1 - \gamma}.$$

Since $u$ and $u'$ are identical within the set $[\bar{x}, \bar{x}]$, $Q^*_1(s, a) = Q^*_2(s, a)$ for all $(s, a) \in K$.

Now we state the risk-sensitive Q-learning algorithm with truncation.
A.3 Heuristics for Polynomial Utility Functions. So far we have relaxed the assumption for utility functions to locally Lipschitz. However, some functions of interest are not even locally Lipschitz. For instance, the function $u(x) = x^p$, $p \in (0, 1)$ is not Lipschitz at the area close to 0. We suggest two types of approximation to avoid this problem:

1. Approximate $u$ by $u^\varphi(x) = (x + \varphi)^p - \varphi^p$ with some positive $\varphi$.
2. Approximate $u$ close to 0 by a linear function, that is,

$$u^\varphi(x) = \begin{cases} 
  u(x) & x \geq \varphi \\
  xu(\varphi) - \varphi^p & x \in [0, \varphi) 
\end{cases}.$$ 

In both cases, $\varphi$ should be set very close to 0.

The assumption in theorem 3 and assumption 1ii requires the strictly positive lower bound $\epsilon$. This causes problems when applying $u(x) = x^p$, $p > 1$ at the area close to 0. We can again apply the two approximation schemes to overcome the problem by selecting small $\varphi$. In section 4, for both $p > 1$ and $p \in (0, 1)$, we apply the second scheme to ensure assumption 1.

A.4 Softmax Policy. Recall that we call a policy proper if, under such a policy, every state is visited infinitely often. In this section, we show that under some technical assumptions, the softmax policy (see equation 3.11) is proper. A policy $\pi = [\pi_0, \pi_1, \ldots]$ is deterministic if for all state $s$ and $t$, there exists an action $a \in A(s)$ such that $\pi_t(a|s) = 1$. Under one policy $\pi$, the $n$-step transition probability $P^\pi(S_n = s'|S_0 = s)$ for some $s, s' \in S$ can be calculated as

$$P^\pi(S_n = s'|S_0 = s) = \sum_{S_1, S_2, \ldots, S_{n-1}} P^\pi_0(S_1|s)P^\pi_1(S_2|S_1) \ldots P^\pi_{n-1}(s'|S_{n-1}),$$

Algorithm 2: Q-Learning with Truncation.
initialize $Q(s, a) = 0$ and $N(s, a) = 0$ for all $s, a$.
for $t = 1$ to $T$ do
  at state $s_t$, choose action $a_t$ randomly using a proper policy (e.g., equation 3.11);
  observe date $(s_t, a_t, r_t, s_{t+1})$;
  $N(s_t, a_t) \leftarrow N(s_t, a_t) + 1$ and set learning rate: $\alpha_t := 1/N(s_t, a_t)$;
  update $Q$ as in equation 3.9;
  truncate $Q$ as in equation A.5, where $\bar{x}$ and $\underline{x}$ are defined in equation A.4.
end for
where $P^\pi(y|x) := \sum_a P(y|x,a)\pi(a|x)$ and $P$ is the transition kernel of the underlying MDP.

**Proposition 3.** Assume that the state and action space are finite and the assumptions required by theorem 2 hold. Assume further that for each $s, s' \in S$, there exist a deterministic policy $\pi_d$, $n \in \mathbb{N}$ and a positive $\epsilon > 0$ such that $P^{\pi_d}(S_n = s'|S_0 = s) > \epsilon$. Then the softmax policy stated in equation 3.11 is proper.

**Proof.** Due to the contraction property of $Q$ (see lemma 2), $\{Q_t\}$ is uniformly bounded with regard to $t$. Let $\pi_s = [\pi_0, \pi_1, \ldots]$ be a softmax policy associated with $\{Q_t\}$. Then, by the definition of softmax policies (see equation 3.11), there exists a positive $\epsilon_0 > 0$ such that $\pi_t(a|s) \geq \epsilon_0$ holds for each $(s, a) \in K$ and $t \in \mathbb{N}$. It implies that for each $s, s' \in S$,

$$P^{\pi_d}(S_n = s'|S_0 = s) \geq \epsilon_0^n P^{\pi_d}(S_n = s'|S_0 = s),$$

for any deterministic policy $\pi_d$. Then by the assumption of this proposition, we obtain that for each $s, s' \in S$, $P^{\pi_d}(S_n = s'|S_0 = s) \geq \epsilon_0^n > 0$. It implies that each state will be visited infinitely often.

The MDP applied in the behavioral experiment in section 4 satisfies the above assumptions, since for each $s, s' \in S$, there exists a deterministic policy $\pi_d$ such that $P^{\pi_d}(S_n = s'|S_0 = s) = 1, n \leq 4$, no matter which initial state $s$ we start with.

**Appendix B: Magnetic Resonance Protocol and Data Processing**

Magnetic resonance (MR) images were acquired with a 3T whole-body MR system (Magnetom TIM Trio, Siemens Healthcare) using a 32-channel receive-only head coil. Structural MRI were acquired with a T1 weighted magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a voxel resolution of $1 \times 1 \times 1$ mm$^3$, coronal orientation, phase encoding in left-right direction, FoV = 192 $\times$ 256 mm, 240 slices, 1100 ms inversion time, TE = 2.98 ms, TR = 2300 ms, and 90 flip angle. Functional MRI time series were recorded using a T2* GRAPPA EPI sequence with TR = 2380 ms, TE = 25 ms, anterior-posterior phase encode, 40 slices acquired in descending (noninterleaved) axial plane with $2 \times 2 \times 2$ mm$^3$ voxels (204 $\times$ 204 mm FoV; skip factor = .5), with an acquisition time of approximately 8 minutes per scanning run.

Structural and functional magnetic resonance image analyzes were conducted in SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Anatomical images were segmented and transformed to the Montreal Neurological Institute (MNI) standard space, and a group average T1 custom anatomical template image was generated using DARTEL. Functional images were corrected for slice-timing acquisition offsets, realigned, and corrected for the interaction of motion and distortion using unwarp tool.
box, co-registered to anatomical images and transformed to MNI space using DARTEL, and finally smoothed with an 8 mm FWHM isotropic gaussian kernel.

Functional images were analyzed using the general linear model (GLM) implemented in SPM8. First-level analyses included onset regressors for each stimulus event excluding the anticipation phase (see Figure 3a), and a set of parametric modulators corresponding to trial-specific task outcome variables and computational model parameters. Trial-specific task outcome variables (and their corresponding stimulus event) include the choice value of the investment (choice phase) and the total value of rewards (gains or losses) over each round (corresponding to a multtrial feedback event). Model-derived parametric modulators included the time series of Q-values for the selected action (choice phase), TD (outcome phase). Reward value was not modeled as a parametric modulator because the TD error time series and trial-by-trial reward values were strongly correlated (all rs > .7; ps < .001). The configuration of the first-level GLM regressors for the standard Q-learning model was identical to that employed in the risk-sensitive Q-learning model. All regressors were convolved with a canonical hemodynamic response function. Prior to model estimation, coincident parametric modulators were serially orthogonalized as implemented in SPM (i.e., the Q-value regressor was orthogonalized with respect to the choice value regressor). In addition, we included a set of regressors for each participant to censor EPI images with large head movement–related spikes in the global mean. These first-level beta values were averaged across participants and tested against zero with a $t$-test. Monte Carlo simulations determined that a cluster of more than 125 contiguous voxels with a single-voxel threshold of $p < .001$ achieved a corrected $p$-value of .05.

Acknowledgments

Thanks to Wendelin Böhmer, Rong Guo, and Maziar Hashemi-Nezhad for useful discussions and suggestions and to the anonymous referee for helpful comments. The work of Y.S. and K.O. was supported by the BMBF (Bernsteinfokus Lernen TP1), 01GQ0911, and the work of M.J.T. and T.S. was supported by the BMBF (Bernsteinfokus Lernen TP2), 01GQ0912.

References


Received October 16, 2013; accepted January 23, 2014.