Model and Reinforcement Learning for Markov Games with Risk Preferences

Wenjie Huang,[1,2] **Pham Viet Hai,**[3] **William B. Haskell**[4]
[1]Shenzhen Research Institute of Big Data (SRIBD)
[2]Institute for Data and Decision Analysis, The Chinese University of Hong Kong, Shenzhen
[3]Department of Computer Science, School of Computing, National University of Singapore (NUS)
[4]Supply Chain and Operations Management Area, Krannert School of Management, Purdue University
wenjiehuang@cuhk.edu.cn, dcspvh@nus.edu.sg, whaskell@purdue.edu

Abstract

We motivate and propose a new model for non-cooperative Markov game which considers the interactions of risk-aware players. This model characterizes the time-consistent dynamic "risk" from *both* stochastic state transitions (inherent to the game) and randomized mixed strategies (due to all other players). An appropriate risk-aware equilibrium concept is proposed and the existence of such equilibria is demonstrated in stationary strategies by an application of Kakutani's fixed point theorem. We further propose a simulation-based Q-learning type algorithm for risk-aware equilibrium computation. This algorithm works with a special form of minimax risk measures which can naturally be written as saddle-point stochastic optimization problems, and covers many widely investigated risk measures. Finally, the almost sure convergence of this simulation-based algorithm to an equilibrium is demonstrated under some mild conditions. Our numerical experiments on a two player queuing game validate the properties of our model and algorithm, and demonstrate their worth and applicability in real life competitive decision-making.

Introduction

Markov games (a.k.a stochastic games) generalize Markov decision processes (MDPs) to the multi-player setting. In the classical case, each player seeks to minimize his expected costs. In a corresponding equilibrium, no player can decrease his expected costs by changing his strategy. We often want to compute equilibria to predict the outcome of the game and understand the behavior of the players.

In this paper, we directly account for the *risk preferences* of the players in a Markov game. Informally, risk aversion is at least weakly preferring a gamble with smaller variance when payoffs are the same. Risk-averse players give more attention to low probability but high cost events compared to risk-neutral players. Models for the risk preferences of a single agent are well established (Artzner et al. 1999; Ruszczynski and Shapiro 2006) for the static problems and (Ruszczyński 2010; Shen, Stannat, and Obermayer 2013) for the dynamic case. We extend these ideas to general sum Markov games and extend the framework of Markov risk measures (Ruszczyński 2010; Shen, Stannat, and Obermayer 2013) to the multi-agent setting. Our model specifically addresses the risk from the stochastic state transitions as well as the risk from the randomized strategies of the other players. The traditional multilinear formulation approach (Kardes, Ordonez, and Hall 2011; Aghassi and Bertsimas 2006) for computing equilibria in robust games fails in our settings, because our model has an intrinsic bilinear term due to the product of probabilities (the state transitions and mixed strategies) which leads to computational intractability. Thus, it is necessary to develop an alternative algorithm to compute equilibria.

Risk Preferences Expected utility theory (von Neumann and Morgenstern 1944; Engelmann and Steiner 2007; Thomas 2016) is a highly developed framework for modeling risk preferences. Yet, some experiments (Levin 2006) show that real human behavior may violate the independence axiom of expected utility theory. Risk measures (as developed in (Artzner et al. 1999; Ruszczynski and Shapiro 2006)) do not require the independence axiom and have favorable properties for optimization.

In the dynamic setting, (Ruszczyński 2010; Shen, Stannat, and Obermayer 2013) develop the class of Markov (a.k.a. dynamic/nested/iterated) risk measures and establish their connection to time-consistency. This class of risk measures is notable for its recursive formulation, which leads to dynamic programming equations. Practical computational schemes for solving large-scale risk-aware MDPs have been proposed, for instance, Q-learning type algorithms (Jiang and Powell 2017; Huang and Haskell 2017; 2018) and simulation-based fitted value iteration (Yu, Haskell, and Xu 2018).

Risk-sensitive/Robust Games Risk-sensitive games have already been considered in (Klompstra 2000; Ghosh, Kumar, and Pal 2016; Basu and Ghosh 2017; Bäuerle and Rieder 2017; Jose and Zhuang 2018). Risk-sensitivity refers to the specific certainty equivalent $(1/\theta)\ln(\mathbb{E}[\exp(\theta X)])$ where $\theta > 0$ is the risk sensitivity parameter. (Ghosh, Kumar, and Pal 2016; Basu and Ghosh 2017) focus on zero-sum risk-sensitive games under continuous time setting.

Robust games study ambiguity about costs and/or state transition probabilities of the game. (Aghassi and Bertsimas 2006) develop the robust equilibrium concept where each player optimizes against the worst-case expected cost over the range of model ambiguity. This paradigm is extended to Markov games in (Kardes, Ordonez, and Hall 2011), and the existence of robust Markov perfect equilibria is demonstrated. (Aghassi and Bertsimas 2006; Kardes, Ordonez, and Hall 2011) formulate robust Markov perfect equilibria as multilinear systems.

Games with risk preferences are not artificial; rather, they emerge organically from many real problems. Traffic equilibrium problems with risk-averse agents are analyzed in (Bell and Cassir 2002) with non-cooperative game theory. The preferences of risk-aware adversaries are modeled in Stackelberg security games in (Qian, Haskell, and Tambe 2015), and a computational scheme for robust defender strategies is presented.

Contributions of This Work We make three main contributions in this paper:

1. We develop a model for risk-aware Markov games where agents have time-consistent risk preferences. This model specifically addresses *both* sources of risk in a Markov game: (i) the risk from the stochastic state transitions and (ii) the risk from the randomized strategies of the other players.
2. We propose a notion of 'risk-aware' Markov perfect equilibria for this game. We show that there exist risk-aware equilibria in stationary strategies.
3. We create a practical simulation-based Q-learning type algorithm for computing risk-aware Markov perfect equilibria, and we show that it converges to an equilibrium almost surely. This algorithm is model-free and so does not require any knowledge of the true model, and thus can search for equilibria purely by observations.

Risk-aware Markov Games

In this section, we develop risk-aware Markov games. Our game consists of the following ingredients: finite set of players $\mathcal{I}$; finite set of states $\mathcal{S}$; finite set of actions $\mathcal{A}^i$ for each player $i \in \mathcal{I}$; strategy profiles $\mathcal{A} := \times_{i\in\mathcal{I}}\mathcal{A}^i$; state-action pairs $\mathcal{K} := \mathcal{S} \times \mathcal{A}$; transition kernel $P(\cdot|s,\, a) \in \mathcal{P}(\mathcal{S})$ (here $\mathcal{P}(\mathcal{S})$ denotes the distribution over $\mathcal{S}$) for all $(s,\, a) \in \mathcal{K}$, and cost functions $c^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all players $i \in \mathcal{I}$.

Each round $t \geq 0$ of the game has four steps: (i) first, all players observe the current state $s_t \in \mathcal{S}$; (ii) second, each player $i \in \mathcal{I}$ chooses $a_t^i \in \mathcal{A}^i$ (all moves are simultaneous and independent, and the corresponding strategy profile is $a_t = \left(a_t^i\right)_{i\in\mathcal{I}}$); (iii) third, each player $i \in \mathcal{I}$ realizes cost $c^i\left(s_t,\, a_t\right)$; and (iv) lastly, the state transitions to s_{t+1} according to $P\left(\cdot \mid s_t,\, a_t\right)$.

We next characterize the players' strategies. In this work, we focus on 'stationary strategies'. Stationary strategies prescribe a player the same probabilities over his actions each time the player visits a certain state, no matter what route he follows to reach that state. Stationary strategies are more prevalent than normal strategies (which rely on the entire history), due to their mathematical tractability (Vrieze 2003; Fink 1964; Kardes, Ordonez, and Hall 2011). Furthermore, the memoryless property of stationary strategies conforms to real human behavior (Vrieze 2003).

We introduce some additional notations to characterize stationary strategies x. Let $\mathcal{P}(\mathcal{A}^i)$ denote the distribution over $\mathcal{A}^i$. For each player $i \in \mathcal{I}$ and state $s \in \mathcal{S}$, $x_s^i \in \mathcal{P}\left(\mathcal{A}^i\right)$ is the mixed strategy over actions where $x_s^i\left(a^i\right)$ denotes the probability of choosing a^i at state s. We define the strategy $x^i := (x_s^i)_{s\in\mathcal{S}} \in \mathcal{X}^i := \times_{s\in\mathcal{S}}\mathcal{P}\left(\mathcal{A}^i\right)$ of player i, the multi-strategy $x := \left(x^i\right)_{i\in\mathcal{I}} \in \mathcal{X} := \times_{i\in\mathcal{I}}\mathcal{X}^i$ of all players, the complementary strategy $x^{-i} := (x^j)_{j\neq i} \in \mathcal{X}^{-i} := \times_{j\neq i}\mathcal{X}^j$, and the multi-strategy $x_s = \left(x_s^i\right)_{i\in\mathcal{I}} \in \mathcal{X}_s := \times_{i\in\mathcal{I}}\mathcal{P}\left(\mathcal{A}^i\right)$ for all players in state $s \in \mathcal{S}$. We sometimes write a multi-strategy as $x = (u^i,\, x^{-i})$ to emphasize player i's strategy u^i.

There are two sources of stochasticity in the cost sequence: the stochastic state transitions characterized by the transition kernel $P(\cdot|s,\, a)$, and the randomized mixed strategies of players characterized by x^{-i}. In this work, we consider the risk from *both* sources of stochasticity. We begin by constructing the framework for evaluating the risk of sequences of random variables. A *dynamic risk measure* is a sequence of conditional risk measures each mapping a future stream of random costs into a risk assessment at the current stage, following the definition of risk maps from (Shen, Stannat, and Obermayer 2013), and satisfying the stationary and time-consistency property of (Ruszczyński 2010, Definition 3) and (Shapiro and Pichler 2016, Definition 1). We assume each conditional risk measure satisfies three axioms: normalization, convexity, and positive homogeneity , which were originally introduced for static risk measures in the pioneering paper (Artzner et al. 1999). Here "convexity" characterizes the risk-averse behavior of players. From (Shapiro and Pichler 2016, Definition 1), a risk-aware optimal policy is *time-consistent* if, the risk of the sub-sequence of random outcome from any future stage is optimized by the resolved policy. In the Supplementary materials[1], we give explicit definitions of the above three axioms of risk measures, stationary and time-consistency risk preferences, and derivation of recursive evaluation of dynamic risk.

From (Ruszczyński 2010, Theorem 4) and (Shapiro and Pichler 2016, Proposition 4), time-consistency allows for a recursive (iterative) evaluation of risk. The infinite-horizon discounted risk for player i under multi-strategy x will be:

$$J_{s_0}^i(x^i,\, x^{-i}) := \rho^i(c^i(s_0,\, a_0) + \gamma\, \rho^i(c^i(s_1,\, a_1) + \gamma\, \rho^i\left(c^i(s_2,\, a_2) + \cdots\right))), \qquad (1)$$

where ρ^i is a one-step conditional risk measure that maps random cost from the next stage to current stage, with respect to the joint distribution of randomized mixed strategies and transition kernel. In Eq. (1), each $c^i(s_t,\, a_t)$, $t \geq 1$ is governed by the joint distribution of randomized mixed

[1] https://arxiv.org/abs/1901.04882

strategies and transition kernel

$$\times_{i\in\mathcal{I}} x^i_{s_t}(a^i_t) P(s_t|s_{t-1},\, a_{t-1}),$$

which is defined for fixed $(s_{t-1},\, a_{t-1})$ and for all s_t and a^i_t. The initial cost $c^i(s_0,\, a_0)$ is only governed by the random mixed strategies distribution $\times_{i\in\mathcal{I}} x^i_{s_0}(a^i_0)$.

The corresponding best response function for player i is:

$$\min_{x^i\in\mathcal{X}^i} J^i_{s_0}(x^i,\, x^{-i}). \tag{2}$$

Suppose we replace all ρ^i with expectation $\mathbb{E}$ in Eq. (1) which leads to $\mathbb{E}^x_s\left[\sum_{t=0}^{\infty}\gamma^t c^i\left(s_t,\, a_t\right)\right]$, where $\mathbb{E}^x_s$ denotes expectation with respect to multi-strategies x, then Problem (2) will become risk-neutral. Thus our formulation recovers the risk-neutral game as a special case.

Denote the ingredients of game $\left\{J^i_s(x^i,\, x^{-i})\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}}$ as $\{\mathcal{I},\, \mathcal{S},\, \mathcal{A},\, P,\, c,\, \rho\}$. In line with the classical definition of Markov perfect equilibrium in (Fink 1964), we now define risk-aware Markov perfect equilibrium.

Definition 1. *(Risk-aware Markov perfect equilibrium) A multi-strategy $x\in\mathcal{X}$ is a risk-aware Markov perfect equilibrium for $\{\mathcal{I},\, \mathcal{S},\, \mathcal{A},\, P,\, c,\, \rho\}$ if*

$$J^i_s(x^i,\, x^{-i}) \le J^i_s(u^i,\, x^{-i}),\, \forall s\in\mathcal{S},\, u^i\in\mathcal{X}^i,\, i\in\mathcal{I}. \tag{3}$$

In Definition 1, each player $i\in\mathcal{I}$ implements a (risk-aware) stationary best response given the stationary complementary strategy x^{-i}. It also states that x is an equilibrium if and only if no player can reduce his discounted risk by unilaterally changing his strategy.

Existence of Stationary Equilibria We prove the existence of stationary equilibira in this section. Let v^i denote player i's *value function*, which is an estimate of the discounted risk starting from the next state S'. For each player i, the *value* of the stationary strategy $x\in\mathcal{X}$ in state $s\in\mathcal{S}$ is defined to be $v^i(s) := J^i_s(x)$, and $v^i := \left(v^i(s)\right)_{s\in\mathcal{S}}$ is the entire value function for player i. The space of value functions for all players is $\mathcal{V} := \times_{i\in\mathcal{I}}\mathbb{R}^{|\mathcal{S}|}$, equipped with the supremum norm $\|v\|_\infty := \max_{s\in\mathcal{S},\, i\in\mathcal{I}} |v^i(s)|$. Eq. (1) states that each player must evaluate the stage-wise risk of random variables on $\mathcal{A}\times\mathcal{S}$, formulated as

$$c^i\left(s,\, A\right) + \gamma\, v^i\left(S'\right), \tag{4}$$

where A is the random strategy profile chosen from $\mathcal{A}$ according to x_s, and S' is the random next state visited (which first depends on x through the random choice of strategy profile a, and then depends on the transition kernel $P\left(\cdot\,|\,s,\, a\right)$ after $a\in\mathcal{A}$ is realized).

Recall that in state $s\in\mathcal{S}$, the probability that $a = (a^i)_{i\in\mathcal{I}}\in\mathcal{A}$ is chosen and then the system transitions to state $k\in\mathcal{S}$ is $\left(\times_{i\in\mathcal{I}} x^i_s\left(a^i\right)\right) P(k\,|\,s,\, a)$. The probability distribution of the strategy profile $a\in\mathcal{A}$ and next state visited $k\in\mathcal{S}$ is given by the matrix

$$\begin{aligned} &P_s\left(u^i_s,\, x^{-i}_s\right)\\ &:= \left[u^i_s\left(a^i\right)\left(\times_{j\neq i} x^j_s\left(a^j\right)\right) P(k\,|\,s,\, a)\right]_{(a,\,k)\in\mathcal{A}\times\mathcal{S}}, \end{aligned} \tag{5}$$

where we explicitly denote the dependence on the multi-strategy $x_s = \left(u^i_s,\, x^{-i}_s\right)$ in state s. For simplicity, we often write P_s instead of $P_s\left(u^i_s,\, x^{-i}_s\right)$ when it is not necessary to indicate the dependence on (u, x).

Let $C^i_s\left(v^i\right) := \left(c^i\left(s,\, A\right) + \gamma\, v^i\left(S'\right)\right)$ be the random cost-to-go for player i at state s. Based on the Fenchel-Moreau representation of risk (Föllmer and Schied 2002; Ruszczynski and Shapiro 2006; Guigues, Krätschmer, and Shapiro 2016), the convex risk of random cost-to-go denoted by $\psi^i_s(u^i_s,\, x^{-i}_s,\, v^i)$ can be computed as the worst-case expected cost-to-go

$$\begin{aligned} \psi^i_s(u^i_s,\, x^{-i}_s,\, v^i) :=& \rho^i\left(c^i(s\,,A) + \gamma\, v^i\left(S'\right)\right)\\ =& \sup_{\mu\in\mathcal{M}^i_s(P_s)} \left\{\langle\mu,\, C^i_s\left(v^i\right)\rangle - b^i_s(\mu)\right\}, \end{aligned}$$

where $\left\{\mathcal{M}^i_s(P_s)\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}} \subset \mathcal{P}(\mathcal{A}\times\mathcal{S})$ is the risk envelope of ρ^i that depends on P_s, and $\left\{b^i_s\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}} : \mathcal{P}\left(\mathcal{A}\times\mathcal{S}\right)\to\mathbb{R}$ are convex functions satisfying $\inf_{\mu\in\mathcal{P}(\mathcal{A}\times\mathcal{S})} b^i_s(\mu) = 0$ for all $i\in\mathcal{I}$ and $s\in\mathcal{S}$. To connect to risk-neutral games, we can just choose all $\mathcal{M}^i_s(P_s)$ to be singletons $\{P_s\left(u^i_s,\, x^{-i}_s\right)\}$ and $b^i_s(\mu) = 0$ for all $\mu\in\mathcal{M}^i_s(P_s)$, $i\in\mathcal{I}$, and $s\in\mathcal{S}$.

We next introduce further assumptions on ρ^i, $\left\{\mathcal{M}^i_s(P_s)\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}}$, and $\left\{b^i_s\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}}$, that will lead to the existence of stationary equilibria.

Assumption 1. *(i) All ρ^i are law invariant, $\rho^i(X) = \rho^i(Y)$ for all $X =_D Y$, where $=_D$ denotes equality in distribution.*

(ii) $\left\{\mathcal{M}^i_s(P_s)\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}} \subset \mathcal{P}\left(\mathcal{A}\times\mathcal{S}\right)$ is a collection of set-valued mappings where $\mathcal{M}^i_s(P_s)$ are closed and polyhedral convex for all P_s. Explicitly, there exists $M\ge 1$ linear constraints and $[M] := \{1, 2, ..., M\}$. Then $\mathcal{M}^i_s(P_s)$ is defined as:

$$\left\{\mu\in\mathbb{R}^{|\mathcal{A}||\mathcal{S}|} :\ \begin{array}{c} A^i_{s,\,m}\,\mu + f_m(P_s)\ge h^i_{s,\,m}, m\in[M],\\ e^T\mu = 1,\\ \mu\ge 0, \end{array}\right\} \tag{6}$$

where $A^i_{s,\,m}$ are matrices, f_m are linear functions in P_s and $h^i_{s,\,m}$ are constants.

(iii) All $\left\{b^i_s\right\}_{s\in\mathcal{S},\, i\in\mathcal{I}}$ are convex and Lipschitz continuous.

Formulation (6) explains how $\mathcal{M}^i_s(P_s)$ depends on P_s. In addition, if f_m depends linearly on P_s, then f_m also depends linearly on u^i_s and x^{-i}_s by definition of P_s in Eq. (5). In computational terms, this assumption is close to (Kardes, Ordonez, and Hall 2011) which assumes polyhedral uncertainty sets for the transition probabilities in its robust Markov game model. This assumption also corresponds to the one in (Ferris and Philpott 2018) about representation of agent risk preferences.

Example 1. *Conditional value-at-risk (CVaR) is a widely investigated coherent risk measure that computes the conditional expectation of random losses exceeding a threshold with probability α.*

CVaR can be constructed from system (6) when we choose $M = 1$, $A^i_{s,m} = -e$, $f_m(P_s) = P_s/(1-\alpha^i)$, and $h^i_{s,m} = 0$ with $m = 1$.

The best response function v_* corresponding to a risk-aware Markov perfect equilibrium, for all $s \in \mathcal{S}$, $i \in \mathcal{I}$, satisfies

$$v_*^i(s) = \min_{u_s^i \in \mathcal{P}(\mathcal{A}^i)} J_s^i(u^i, x^{-i}) = \min_{u_s^i \in \mathcal{P}(\mathcal{A}^i)} \psi_s^i(u_s^i, x_s^{-i}, v_*^i), \quad (7)$$

$$x_s^i \in \arg\min_{u^i \in \mathcal{X}^i} J_s^i(u^i, x^{-i}), \quad (8)$$

and v_*^i may not be unique. In the mapping $C_s^i\left(v^i\right)$ on $\mathcal{A} \times \mathcal{S}$, the players control the *distribution* on $\mathcal{P}(\mathcal{A} \times \mathcal{S})$ through their mixed strategies. Eqs. (7) - (8) together simply restate Eq. (3). However, Eqs. (7) - (8) give a computational recipe that can be encoded into an operator on multi-strategies. We define this operator Φ on $\mathcal{X}$:

$$\Phi(x) := \Big\{\tilde{q} \in \mathcal{X} : \tilde{q}_s^i \in \arg\min_{u_s^i \in \mathcal{P}(\mathcal{A}^i)} \psi_s^i(u_s^i, x_s^{-i}, v_*^i), \; v_*^i(s) = \min_{u_s^i \in \mathcal{P}(\mathcal{A}^i)} \psi_s^i(u_s^i, x_s^{-i}, v_*^i), \forall s \in \mathcal{S}, i \in \mathcal{I}\Big\}. \quad (9)$$

This operator returns the set of strategies for every player that are best responses to all other players' strategies.

The following Theorem 1 briefly describes the existence of stationary strategies with detailed proof in the Supplementary materials.

Theorem 1. *Suppose Assumption 1 holds, then the game* $\{\mathcal{I}, \mathcal{S}, \mathcal{A}, P, c, \rho\}$ *has an equilibrium in stationary strategies.*

Our proof of existence of Theorem 1 draws from (Fink 1964; Kardes, Ordonez, and Hall 2011). The main idea is to show that Φ is a nonempty, closed, and convex subset of $\mathcal{X}$, and that Φ is upper semicontinuous. Then, we apply Kakutani's fixed point theorem to show that this correspondence Φ has a fixed point which coincides with a risk-aware Markov perfect equilibrium.

A Q-Learning Algorithm

We propose a simulation-based and asynchronous algorithm for computing equilibria of the risk-aware game $\{\mathcal{I}, \mathcal{S}, \mathcal{A}, P, c, \rho\}$, called Risk-aware Nash Q-learning (RaNashQL). This algorithm does not require a model for the cost functions $\left\{c^i\right\}_{i \in \mathcal{I}}$ or the transition kernel P, nor does not it require prior knowledge on $\mathcal{S}$. The algorithm has an outer-inner loop structure, where the risk estimation is performed in the inner loop and the equilibrium estimation is performed in the outer loop.

In each iteration of RaQL, a collection of Q-values for each player for all strategy profiles, is generated. The one-shot game formed by the collection of Q-values is called a *stage game*. We will later formulate stage game explicitly. The outer-inner loop structure follows (Jiang and Powell 2017; Huang and Haskell 2017; 2018) where multiple "stochastic approximation instances" for both risk estimation and Q-value updates are "pasted" together. We show that the Nash equilibria mapping for stage games is non-expansive, and both the risk estimation error and equilibrium estimation error are bounded by the gap between the estimated Q-value and the Q-value under the equilibrium. These two conditions allow us to prove the convergence of the algorithm using the theory of stochastic approximation, as shown in (Even-Dar and Mansour 2004).

For this section, we assume that our risk measures $\left\{\rho^i\right\}$ have a special form as stochastic saddle-point problems to facilitate computation. Define a probability space $(\Omega, \mathcal{F}, P)$ and the space of essentially bounded random variables $\mathcal{L} = L_\infty(\Omega, \mathcal{F}, P)$.

Assumption 2. *(Stochastic saddle-point problem) For all* $i \in \mathcal{I}$*,*

$$\rho^i(X) = \min_{y \in \mathcal{Y}^i} \max_{z \in \mathcal{Z}^i} \mathbb{E}_P\left[G^i(X, y, z)\right], \forall X \in \mathcal{L}, \quad (10)$$

where: (i) $\mathcal{Y}^i \subset \mathbb{R}^{d_1}$ *and* $\mathcal{Z}^i \subset \mathbb{R}^{d_2}$ *are compact and convex with diameters* $D_{\mathcal{Y}}$ *and* $D_{\mathcal{Z}}$*, respectively. (ii)* G^i *is Lipschitz continuous on* $\mathcal{L} \times \mathcal{Y}^i \times \mathcal{Z}^i$ *with constant* $K_G > 1$*. (iii)* G *is convex in* $y \in \mathcal{Y}^i$ *and concave in* $z \in \mathcal{Z}^i$*. (iv) The subgradients of* G *on* y *and* z *are Borel measurable and uniformly bounded for all* $X \in \mathcal{L}$*.*

In (Huang and Haskell 2018, Theorem 3.2), conditions on G^i are given to ensure that the corresponding minimax structure (10) is a convex risk measure. Some examples of the functions G^i are shown in the Supplementary materials such that the corresponding risk-aware Markov perfect equilibria exist. For instance, CVaR can be written as:

$$\text{CVaR}_{\alpha^i}(X) := \min_{\eta \in \mathbb{R}} \left\{\eta + \frac{1}{1-\alpha^i}\mathbb{E}\left[\max\left\{X - \eta, 0\right\}\right]\right\}, \quad (11)$$

where $\alpha^i \in [0, 1)$ is the risk tolerance for player i.

Risk-aware Nash Q-learning Algorithm RaNashQL is updated based on future equilibrium costs (which depend on all players). In contrast, single-agent Q-learning updates are only based on the player's own costs. Thus, to predict equilibrium losses, every player must maintain and update a model for all other player's costs and their risk assessments, which follows the settings in (Hu and Wellman 2003).

For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $i \in \mathcal{I}$,

$$Q_*^i(s, a) := \min_{y \in \mathcal{Y}^i} \max_{z \in \mathcal{Z}^i} \mathbb{E}_{P(\cdot \mid s, a)}\big\{ G^i\left(c^i(s, A) + \gamma\, v_*^i(S), y, z\right)\big\}, \quad (12)$$

denotes the Q-values corresponding to a stationary equilibrium and its best response function v_*. In the case of multiple equilibria, different Nash strategy profiles may have different equilibrium Q-values, so the pair (v_*^i, Q_*^i) may not be unique.

In a multi-agent Q-learning algorithm, the agents play a sequence of stage games where the payoffs are the current Q-values. In each state $s \in \mathcal{S}$, the corresponding stage game is the collection $(Q^i(s))_{i \in \mathcal{I}}$, where $Q^i(s) := \{Q^i(s, a) : a \in \mathcal{A}\}$ is the array of Q-values for player i for all strategy profiles. Let x_s be a Nash equilibrium of the stage game $(Q^i(s))_{i \in \mathcal{I}}$, then the corresponding *Nash Q-value* for all $i \in \mathcal{I}$ is denoted:

$$Nash^i(Q^j(s))_{j \in \mathcal{I}} := \sum_{a \in \mathcal{A}} \left(\times_{j \in \mathcal{I}} x_s^j\left(a^j\right)\right) Q^i(s, a),$$

which gives each player's corresponding expected cost in state $s \in \mathcal{S}$ (with respect to the Q-values) under x_s.

RaNashQL builds upon the algorithm in (Hu and Wellman 2003) for the risk-aware case. Figure 1 illustrates how players interact with others and update their equilibrium estimation through RaQL. Each player chooses an action based on

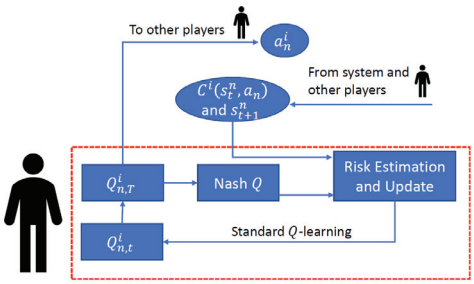

Figure 1: Illustration of RaQL

a Nash equilibrium of their current Q-values, observed cost, other players' actions, and then the new state in each iteration. The Q-values follow a stochastic approximation-type update as in standard Q-learning.

Algorithm 1 Risk-aware Nash Q-learning

(Step 0) Initialize: Let $n = 1$, and $t = 1$, get the initial state s_1. Let the learning agent be indexed by i. For all $s \in \mathcal{S}$ and $a^i \in \mathcal{A}^i$, $i \in \mathcal{I}$, let $Q_{n,t}^i(s,\, a) = 0$.
For $n = 1,\, ...,\, N$ **do**
(Step 1) Choose a_n^i based on the exploration policy π. Observe the actions and costs for all players, then observe a new state;
  For $t = 1,\, ...,\, T$ **do**
  (Step 2) Compute the Nash Q-value; Compute the risk-aware cost-to-go for all players;
  (Step 3) Update each $Q_{n,t}^i$, $i \in \mathcal{I}$ using stochastic approximation;
  (Step 4) Stochastic approximation of risk measure by SASP;
  end for
end for
Return Approximated Q-value $Q_{N,T}^i$, $i \in \mathcal{I}$.

The steps of RaNashQL are summarized in Algorithm 1, which contains N and T number of iterations for outer and inner loops, respectively. In Step 4, we use the stochastic approximation for saddle-point problems (SASP) algorithm, (Nemirovski and Rubinstein 2005, Algorithm 2.1). Classical stochastic approximation may result in extremely slow convergence for degenerate objectives (i.e. when the objective has a singular Hessian). However, the SASP algorithm with a properly chosen parameter preserves a "reasonable" (close to $O(n^{-1/2})$) convergence rate, even when the objective is non-smooth and/or degenerate. Thus, SASP is a robust choice for solving problem (10). The extended formulations from Steps (0)-(4) in Algorithm 1 are given in the Supplementary materials.

Game 1	*Left*	*Right*
Up	0, 1	10, 7
Down	7, 10	11, 8

Game 2	*Left*	*Right*
Up	5, 5	10, 4
Down	4, 10	8, 8

Game 3	*Left*	*Right*
Up	0, 1	10, 9
Down	7, 10	8, 8

Table 1: Examples of $\mathcal{I}'$-mixed point

Almost Sure Convergence Let $\{Q_{n,T}\}_{i\in\mathcal{I}}$ be the Q-value estimations at iteration n and T (the end of each inner loop after the risk estimation has been done) from Algorithm 1. We would like to demonstrate the almost sure convergence of $Q_{n,T}^i$ to the risk-aware equilibrium Q-values Q_*^i for all players. (Hu and Wellman 2003) introduce two conditions on the Nash equilibria of all the stage games that lead to almost sure convergence, a *global optimal* point when every player receives his lowest cost at this point, and a *saddle* point when each agent would receive a lower cost when at least one of the other players deviates. We found a special type of Nash equilibria that we call an $\mathcal{I}'$-mixed point, which builds on (Hu and Wellman 2003), and plays a major role in our convergence analysis.

Definition 2. *Let $(C^i)_{i\in\mathcal{I}}$ denote the expected cost of all players as a function of the multi-strategy $x \in \mathcal{X}$. A multi-strategy $x \in \mathcal{X}$ is a $\mathcal{I}'$-mixed point of $\left(C^i\right)_{i\in\mathcal{I}}$ if: (i) it is a Nash equilibrium and (ii) there exists an index of players $\mathcal{I}' \subseteq \mathcal{I}$ such that: $C^i\left(x\right) \leq C^i\left(x'\right)$, $\forall x' \in \mathcal{X}$, $i \in \mathcal{I}'$, and $C^i\left(x^i,\, x^{-i}\right) \leq C^i\left(x^i,\, u^{-i}\right)$, $\forall u^{-i} \in \mathcal{X}^{-i}$, $i \in \mathcal{I}\backslash\mathcal{I}'$.*

Our definition of '$\mathcal{I}'$-mixed point' combines both notions of global optimal point and saddle point. From Definition 2, a subset of players $\mathcal{I}' \subseteq \mathcal{I}$ minimizes their expected costs at x. The rest of the players $\mathcal{I}\backslash\mathcal{I}'$ each would receive a lower expected cost when at least one of the other players deviates. An example of an $\mathcal{I}'$-mixed point in a one shot game follows.

Example 2. *Player 1 has choices* Up *and* Down*, and Player 2 has choices* Left *and* Right*. Player 1's loss is the first entry in each cell, and Player 2's are the second. The first game has a unique Nash equilibrium (Up, Left), which is a global optimal point. The second game also has a unique Nash equilibrium (Down, Right), which is a saddle-point. The third game has two Nash equilibrium: a global optimum (Up, Left), and a mixed point (Down, Right). In equilibrium (Down, Right), Player 1 receives a lower cost if Player 2 deviates, while Player 2 receives a higher cost if Player 1 deviates.*

We now introduce the following additional assumptions for our analysis of RaNashQL.

Assumption 3. *One of the following holds for all stage games $(Q_{n,T}^i(s))_{i\in\mathcal{I}}$ for all n and $s \in \mathcal{S}$ in Algorithm 1.*

(i) Every $(Q_{n,T}^i(s))_{i\in\mathcal{I}}$ for all n and $s \in \mathcal{S}$ has a global optimal point.

(ii) Every $(Q_{n,T}^i(s))_{i\in\mathcal{I}}$ for all n and $s \in \mathcal{S}$ has a saddle point.

(iii) For any two stage games Q, $\tilde{Q} \in (Q^i_{n,T}(s))_{i\in\mathcal{I}}$ *for all* n *and* $s \in \mathcal{S}$*, we suppose* Q_1 *has a* $\mathcal{I}_1$*-mixed point* x *and* Q_2 *has a* $\mathcal{I}_2$*-mixed point* $\tilde{x}$*. Then: For* $i \in \mathcal{I}_1 \cup (\mathcal{I}\backslash\mathcal{I}_2)$*, then* $Q^i(x) \geq \tilde{Q}^i(\tilde{x})$*; For* $i \in \mathcal{I}_2 \cup (\mathcal{I}\backslash\mathcal{I}_1)$*, then* $Q^i(x) \leq \tilde{Q}^i(\tilde{x})$.

Compared with (Hu and Wellman 2003, Assumption 3), Assumption 3(iii) enables wider application of RaNashQL. In particular, even the indices $\mathcal{I}_1$ and $\mathcal{I}_2$ of all the stage games may differ across iterations. Next we list further standard assumptions on exploration in RaNashQL and its asynchronous updates.

Assumption 4. *(i) The exploration policy* π *is* $\varepsilon-$*greedy, meaning with probability* $\varepsilon \in (0, 1)$*, action* a^i *is chosen uniformly from* $\mathcal{A}^i$*, and with probability* $1-\varepsilon$*, action* a^i *is drawn from* $\mathcal{A}^i$ *according to* x^i_s *which is the equilibrium of the stage game* $\{Q^i(s)\}_{i\in\mathcal{I}}$*; (ii) a single state-action pair is updated when it is observed in each iteration.*

By the Extended Borel-Cantelli Lemma (Breiman 1992), the algorithm satisfying Assumption 4(i) will visit every state-action pair infinitely often with probability one.

Theorem 2. *Suppose Assumptions 3 and 4 hold. For any* $T \geq 1$*, Algorithm 1 generates sequences* $\left\{Q^i_{n,T}\right\}_{n\geq 1}$ *such that* $Q^i_{n,T} \to Q^i_*$ *almost surely as* $n \to \infty$ *for all* $i \in \mathcal{I}$.

The proof sketch of Theorem 2 is listed as follows, with the details presented in the Supplementary materials. (i) Show that all $\mathcal{I}'$-mixed points of a stage game have equal value, and the property also holds for global optimal points and saddle points. Consequently, from (Hu and Wellman 2003), the mapping from Q-values to Nash equilibrium (of the stage games) is non-expansive. (ii) Show that the Hausdorff distance between the subdifferentials of the estimated risk on $\mathcal{Y}^i$ and $\mathcal{Z}^i$ (corresponding to Eq. (10)), is bounded by a function of $\|Q^i_{n-1,T} - Q^i_*\|_2$. (iii) Show that the duality gaps of all the saddle point estimation problems are bounded by a function of $\|Q^i_{n-1,T} - Q^i_*\|_2$. (iv) If the conditions in (i)-(iii) hold, then $Q^i_{n,T}$ from RaNashQL are a well-behaved stochastic approximation sequence (Even-Dar and Mansour 2004, Definition 7) that converges to Q^i_* with probability one.

(Huang and Haskell 2018, Theorem 4.7) shows that the single-agent version of RaNashQL has complexity

$$\Omega\left(\left(S\,A\ln(S\,A/\delta\epsilon)/\epsilon^2\right)^{1/\beta} + (\ln(\sqrt{S\,A}/\epsilon))^{1/(1-\beta)}\right), \tag{13}$$

with probability $1-\delta$, where S and A denote the cardinality of state and actions spaces and $\beta \in (0, 1]$ is the learning rate. In the multi-agent case, our conjecture is to replace A with $|\mathcal{A}|$ in the term (13) to get a rough estimate of the time complexity of RaNashQL. However, the explicit complexity bound is difficult to derive and remains for future research. In RaNashQL, there are multiple Q-values being updated in each iteration for each state, and their relationships are complex (they are linked by the solutions of a stage game, since each stage game may yield multiple Nash equilibria).

In the Supplementary materials, we also discuss (i) methods for computing Nash equilibria of stage games involving two or more players; (ii) a rule for choosing a unique Nash equilibrium of stage games from multiple choices; (iii) the storage space requirement of RaNashQL.

A Queuing Control Application

We apply our techniques to the single server exponential queuing system from (Kardes, Ordonez, and Hall 2011). In this packet switched network, it is service provider's (denoted as "SP" latter in the tables) benefit to increase the amount of packets processed in the system. However, such an increase may result in an increase in packets' waiting times in the buffer (called latency), and routers (denoted as "R" latter in the tables) are used to reduce packets' waiting times. Thus, the game arises because the service provider and router choose their service rates to achieve competing objectives.

The state space $\mathcal{S}$ represents the maximum number (30 in these experiments) of packets allowed in the system. We assume that the time until the admission of a new packet and the next service completion are both exponentially distributed. Therefore, the number of packets in the system can be modeled as a birth and death process with fixed state transition probabilities. In the Supplementary materials, we provide the explicit formulation of cost functions, state transition probabilities, as well as other parameter settings. We suppose that each player has the same two available actions (service rates) in every state. CVaR is the risk measure for both players in all the experiments. The player's risk preferences are obtained by setting α^i for $i = 1, 2$, and we allow $\alpha^1 \neq \alpha^2$.

Experiment I (RaNashQL vs. Nash Q-learning) We compare RaNashQL with Nash Q-learning in (Hu and Wellman 2003) in terms of their convergence rates. Given any precision $\epsilon > 0$, we record the iteration count n until the convergence criterion $\|Q^i_{n,T} - Q^i_*\|_2 \leq \epsilon$ is satisfied. Figure 2 (top) reveals that RaNashQL is more computationally expensive than Nash Q-learning. Table 2 shows the discounted cost under equilibrium by simulation (1000 samples). The first table reveals that incorporating risk will help the service provider reduce its mean cost, while increase the mean cost of the router. The second table shows that incorporating risk will help to reduce the overall cost to the entire system with only a slightly higher variance.

The first part of Table 3 shows that the mean cost of service provider (-44.31) is lower than that under the risk-neutral Markov perfect equilibrium (-22.22), and the mean cost of router (59.64) is lower than that under the risk-aware Markov perfect equilibrium (37.48). This result shows that incorporating risk preference can help decision makers reach a new equilibrium that further reduces his mean cost compared to cases where both players are either risk-neutral or risk-aware. Similar phenomena can also be shown in the second part of Table 3. In the final part of Table 3, we construct a new two-player one-shot game where the risk preferences (risk-neutral and risk-aware) are the actions and the expected value from simulation will be outcome of the game. We find that a equilibrium is attained for this game when the router

Player	Method	Mean	Variance	5%-CVaR	10%-CVaR
SP	Neutral	-22.22	$1.4736e-06$	-22.22	-22.22
	CVaR	-77.78	407.84	-69.34	-68.26
R	Neutral	37.48	7.32	37.94	38.18
	CVaR	83.68	491.20	86.03	87.54

Method	Mean	5%-CVaR	10%-CVaR
Neutral	15.26	15.72	15.96
CVaR	5.9	16.69	19.28

Table 2: Simulation (Constructing CVaR with $\alpha^1 = \alpha^2 = 0.1$)

Player	Method	Mean	Variance	5%-CVaR	10%-CVaR
SP	CVaR	-44.31	266.06	-43.38	-42.70
R	Neutral	59.64	316.71	61.18	62.77

Player	Method	Mean	Variance	5%-CVaR	10%-CVaR
SP	CVaR	-54.76	26.05	-54.71	-54.67
R	Neutral	70.56	31.03	71.56	71.81

		Router	
		Risk-neutral	Risk-aware
Service Provider	Risk-neutral	$(-22.22, 37.48)$	$(-54.76, 70.56)$
	Risk-aware	$(-44.44, 59.64)$	$(-77.78, 83.68)$

Table 3: Simulation (Constructing CVaR with $\alpha^1 = 0.95$, $\alpha^2 = 0.1$ for the first table, and $\alpha^1 = 0.1$, $\alpha^2 = 0.95$ for the second)

is risk-neutral and the service provider is risk-aware. This one-shot game demonstrates that the router should be risk-neutral when service provider is risk-aware, in order to reduce his expected cost.

In the Supplementary materials, we further explain the reason for the increase in variance in risk-aware games in Table 2 which is counter-intuitive.

Experiment II (RaNashQL vs. Multilinear System) In this experiment, we consider a special case where the risk only comes from state transitions (this setting is basically a risk-aware interpretation of (Kardes, Ordonez, and Hall 2011)). In this case, we can compute the risk-aware Markov equilibrium "exactly" using a multilinear system and interior point algorithm as detailed in the Supplementary materials. We evaluate performance in terms of the relative error

$$\frac{\sqrt{\sum_{s\in\mathcal{S}}\left(Nash^i(Q^j_{n,T}(s))_{j\in\mathcal{I}} - v^i_*(s)\right)^2}}{\sqrt{\sum_{s\in\mathcal{S}} v^i_*(s)^2}},\ n \le N,$$

where v^i_* is the value function corresponding to the equilibrium solved by multilinear system. The Supplementary materials confirm that the service provider's strategy produced by RaNashQL converges almost surely to the one produced by multilinear system. From the Supplementary materials, interior point algorithm finds a local optimum with 10471.975 seconds, and RaNashQL has relative error lower than 25% with 5122.657 seconds. Thus, our approach possesses superior computational performance compared to an interior point algorithm for solving multilinear systems.

Experiment III (Computational Complexity Conjecture) In this experiment, we explore the conjecture on the computational complexity of RaNashQL. Given a fixed ϵ, we could compute the complexity conjecture through formulation (13). Figure 2 (bottom) shows that the relative errors of service provider and router under computed complexity conjecture are bounded by ϵ. Thus we derive a potential heuristic for the computational complexity of solving a general sum game given the size of the game. In other words, each practitioner can estimate the upper bound of total complexity in computing the $\epsilon-$ equilibrium through this conjecture.

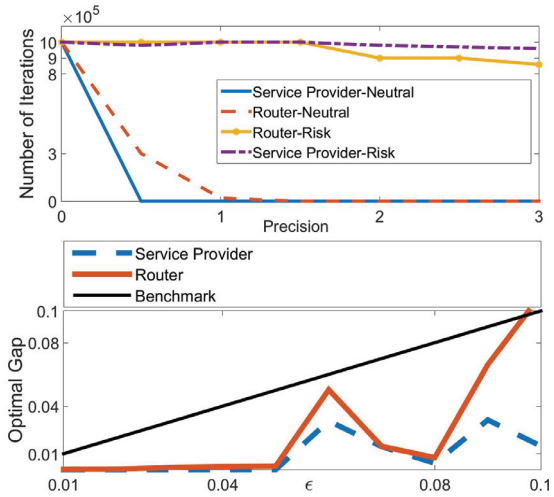

Figure 2: Computational Complexity

Conclusion

In this paper, we propose a model and simulation-based algorithm for non-cooperative Markov games with time-consistent risk-aware players. This work has made the following contributions: (i) The model characterizes the "risk" from both the stochastic state transitions and the randomized strategies of the other players. (ii) We define risk-aware Markov perfect equilibrium and prove its existence in stationary strategies. (iii) We show that our algorithm converges to risk-aware Markov perfect equilibrium almost surely. (iv) From a queuing control numerical example, we find that risk-aware Markov games will reach new equilibria other than risk-neutral ones (this is the equilibrium shifting phenomenon). Moreover, the variance is increased for risk-aware Markov games, which is contrary to the variance reduction property of risk-aware optimization for single agents. The sum of expected cost over all players is reduced in risk-aware Markov game, compared to risk-neutral ones. In future research, we seek to improve the scalability of our framework for large-scale Markov games.

Acknowledgments

This work is supported by SRIBD International Postdoctoral Fellowship and by the NUS Young Investigator Award

"Practical Considerations for Large-Scale Competitive Decision Making".

References

Aghassi, M., and Bertsimas, D. 2006. Robust game theory. *Mathematical Programming* 107(1-2):231–273.

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

Basu, A., and Ghosh, M. K. 2017. Nonzero-sum risk-sensitive stochastic games on a countable state space. *Mathematics of Operations Research* 43(2):516–532.

Bäuerle, N., and Rieder, U. 2017. Zero-sum risk-sensitive stochastic games. *Stochastic Processes and their Applications* 127(2):622–642.

Bell, M. G., and Cassir, C. 2002. Risk-averse user equilibrium traffic assignment: an application of game theory. *Transportation Research Part B: Methodological* 36(8):671–681.

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

Engelmann, D., and Steiner, J. 2007. The effects of risk preferences in mixed-strategy equilibria of $2\times$ 2 games. *Games and Economic Behavior* 60(2):381–388.

Even-Dar, E., and Mansour, Y. 2004. Learning rates for q-learning. *The Journal of Machine Learning Research* 5:1–25.

Ferris, M., and Philpott, A. 2018. Dynamic risked equilibrium.

Fink, A. M. 1964. Equilibrium in a stochastic n-person game. *Journal of science of the hiroshima university, series ai (mathematics)* 28(1):89–93.

Föllmer, H., and Schied, A. 2002. Convex measures of risk and trading constraints. *Finance and stochastics* 6(4):429–447.

Ghosh, M. K.; Kumar, K. S.; and Pal, C. 2016. Zero-sum risk-sensitive stochastic games for continuous time markov chains. *Stochastic Analysis and Applications* 34(5):835–851.

Guigues, V.; Krätschmer, V.; and Shapiro, A. 2016. Statistical inference and hypotheses testing of risk averse stochastic programs. *arXiv preprint arXiv:1603.07384*.

Hu, J., and Wellman, M. P. 2003. Nash q-learning for general-sum stochastic games. *Journal of machine learning research* 4(Nov):1039–1069.

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

Huang, W., and Haskell, W. B. 2018. Stochastic approximation for risk-aware markov decision processes. *arXiv preprint arXiv:1805.04238*.

Jiang, D. R., and Powell, W. B. 2017. Risk-averse approximate dynamic programming with quantile-based risk measures. *Mathematics of Operations Research* 43(2):554–579.

Jose, V. R. R., and Zhuang, J. 2018. Incorporating risk preferences in stochastic noncooperative games. *IISE Transactions* 50(1):1–13.

Kardes, E.; Ordonez, F.; and Hall, R. W. 2011. Discounted robust stochastic games and an application to queueing control. *Operations Research* 59(2):365–382.

Klompstra, M. B. 2000. Nash equilibria in risk-sensitive dynamic games. *IEEE Transactions on Automatic Control* 45(7):1397–1401.

Levin, J. 2006. Choice under uncertainty. *Lecture Notes*.

Nemirovski, A., and Rubinstein, R. 2005. An efficient stochastic approximation algorithm for stochastic saddle point problems. *Modeling Uncertainty* 156–184.

Qian, Y.; Haskell, W. B.; and Tambe, M. 2015. Robust strategy against unknown risk-averse attackers in security games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1341–1349. International Foundation for Autonomous Agents and Multiagent Systems.

Ruszczynski, A., and Shapiro, A. 2006. Optimization of convex risk functions. *Mathematics of operations research* 31(3):433–452.

Ruszczyński, A. 2010. Risk-averse dynamic programming for markov decision processes. *Mathematical programming* 125(2):235–261.

Shapiro, A., and Pichler, A. 2016. Time and dynamic consistency of risk averse stochastic programs. *optimization-online.org*.

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

Thomas, P. 2016. Measuring risk-aversion: The challenge. *Measurement* 79:285–301.

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Vrieze, O. 2003. Stochastic games and stationary strategies. In *Stochastic Games and Applications*. Springer. 37–50.

Yu, P.; Haskell, W. B.; and Xu, H. 2018. Approximate value iteration for risk-aware markov decision processes. *IEEE Transactions on Automatic Control*.