In this paper, we consider the problem of optimizing the quantiles of the cumulative rewards of Markov Decision Processes (MDP), to which we refer as Quantile Markov Decision Processes (QMDP). Traditionally, the goal of a Markov Decision Process (MDP) is to maximize expected cumulative reward over a defined horizon (possibly to be infinite). In many applications, however, a decision maker may be interested in optimizing a specific quantile of the cumulative reward instead of its expectation. (If we have some reference here, it would be good.) Our framework of QMDP provides analytical results characterizing the optimal QMDP solution and presents the algorithm for solving the QMDP. We provide analytical results characterizing the optimal QMDP solution and present the algorithms for solving the QMDP. We illustrate the model with two experiments: a grid game and a HIV optimal treatment experiment.

Key words: Markov Decision Process, Quantile Reward, Risk Profile, Medical Decision Making

1. Introduction

The problem of sequential decision making has been widely studied in the fields of operations research, management science, artificial intelligence and stochastic control. The Markov Decision Process (MDP) is one important framework for addressing such problems. In the traditional setting, an agent sequentially performs actions based on information about the current state and then obtains rewards based on the action and state. The goal of an MDP is to maximize the expected cumulative reward over a defined (finite/infinite) horizon.

In many applications, however, a decision maker may be interested in optimizing a specific quantile of the cumulative reward instead of its expectation. This quantile reflects the decision maker’s risk attitude and thus should result in different optimal strategies for different people. For example, a risk-averse physician may want to determine the optimal drug regime for a patient with the objective of maximizing the 0.10 quantile of the cumulative reward; this is the cumulative improvement in health that is expected to occur at least 90% of the time for the patient. A
company such as Amazon that provides cloud computing services might want their cloud service to be optimized at the 0.01 quantile (DeCandia et al. 2007); the company wants to provide service that satisfies 99% of its customers. On the other hand, a wealthy investor may wish to maximize the 0.90 quantile of reward; this is the return on investment that is expected to occur only 10% of the time.

People has been well aware of the risk in MDP, while not much work has been done with respect to the quantile perspective. The idea of risk-sensitive decision making in MDPs was first introduced in the 1970s when Howard and Matheson (1972) proposed the use of an exponential utility function to reflect the risk attitude: the goal of the resulting MDP is to maximize expected utility. In recent years, two independent streams of research have explored risk-sensitive MDPs. One stream focuses on optimizing the cumulative reward with a worst-case guarantee (Altman 1999, Ermon et al. 2012). This approach formulates the MDP as a constrained optimization problem with cumulative reward as the objective function and the worst-case guarantee as a constraint. The other stream of research investigates the problem of maximizing the conditional value-at-risk (CVaR) for an MDP (Carpin et al. 2016, Chow and Ghavamzadeh 2014, N. Bauerle and J. Ott 2011). The CVaR, also known as average value-at-risk (AVaR) or expected shortfall, is defined as the expectation of the loss/reward under the worst q% of cases. Although the quantile, or value-at-risk (VaR), is a more useful objective in many scenarios (A¨ıt-Sahalia and Hansen 2009, Jorion 1997), CVaR is used as an objective because it can be expressed in a one-dimensional optimization problem (Rockafellar and Uryasev 2000). In this paper, we will demonstrate that the quantile (VaR) objective also has good analytical properties that can be elegantly utilized in the context of MDP. Besides, two recent works studied the problem of “robust Markov decision processes” (Delage and Mannor 2010, Wiesemann et al. 2013). The goal is to optimize expected reward under the uncertainty of model parameters. In this paper, by looking into the quantiles, we discuss the uncertainty of the model itself, which is more essential and insightful.

To our knowledge, our paper is the first work that fully resolves the problem of optimizing quantile rewards for MDP. A few literatures are close to ours. Filar et al. (1995) proposed an algorithm to optimize the quantile of limiting average reward of infinite-horizon MDP while Ummels and Baier (2013) developed an algorithm to compute the quantile cumulative rewards for a given policy in polynomial time. One step further, Gilbert et al. (2016) addressed the quantile optimization problem for the case of deterministic reward and preference-based MDP, which was too restrictive for practical use. More importantly, a common problem for all these three works along with some other efforts is that they are too computationally costly for practical use. As we will demonstrate later in the paper, our framework and algorithms can easily scale for the problem with thousands of states and hundreds of time horizons.

To summarize, our contributions are in the following four aspects:
First, we propose a Quantile Markov Decision Process (QMDP) framework that addresses the problem of maximizing quantile reward in MDP for both finite/infinite horizon. In this QMDP framework, the optimal values and policies for all quantiles $\tau \in [0, 1]$ can be computed in a single pass of backward dynamic programming.

Second, we develop several theoretical results and algorithms for the quantiles of random variables. They can be of independent interest for other applications outside the scope of MDP.

Third, we design an efficient algorithm to solve QMDP. The complexity analysis unveils the internalized complexity associated with the QMDP and therefore concludes that no further improvement can be made in terms of computation cost.

Fourth, we conduct two experiments to demonstrate advantages of the quantile criteria, which opens the great possibilities of using QMDP in practice. Remarkably, the QMDP framework on one hand provides us quantile optimal policy which reflects the decision maker’s risk attitude, on the other hand can also serve as a tool to assess the risk for the optimal policy in traditional MDP.

2. Preliminaries and Model Setup

2.1. Markov Decision Process

The Markov Decision Process (MDP) consists of two parts (Bertsekas 1995): (1) an underlying discrete-time dynamic system, and (2) a reward function that is additive over time. A dynamic system defines the evolution of the state over time:

$$S_{t+1} = f_t(S_t, a_t, w_t), \quad t = 0, 1, ..., T - 1,$$

where $S_t$ denotes the state variable from state space $S$, $a_t$ denotes the actions/decisions at time $t$ and $w_t$ is a random variable that captures the stochasticity in the system. The reward function at time $t$, denoted by $r_t(S_t, a_t, w_t)$, accumulates over time. The total reward is

$$r_T(S_T) + \sum_{t=0}^{T-1} r_t(S_t, a_t, w_t),$$

where $r_T(S_T)$ is the terminal reward at the end of the process. The random variable $w_t \in \mathcal{W}$ determines the transition in the state space and the state $S_{t+1}$ follows a distribution $P_t(\cdot | S_t, a_t)$ that is possibly dependent on the state $S_t$ and the action $a_t$. We consider the class of policies that consist a sequence of functions $\pi = \{\mu_0, ..., \mu_{T-1}\}$ where $\mu_t$ maps historical information $h_t = (S_0, a_0, ..., S_{t-1}, a_{t-1}, S_t)$ to an admissible action $a_t \in \mathcal{A}_t \subset \mathcal{A}$. Here we use $\mathcal{A}_t$ and $\mathcal{A}$ to denote the admissible action set. The policy $\pi$ together with the function $f_t$ determines the dynamics of the process. Given an initial state $S_0$ and a policy $\pi$, we have the following expected total reward:

$$E^{\pi} \left[ r_T(S_T) + \sum_{t=0}^{T-1} r_t(S_t, a_t, w_t) \right].$$
The objective of an MDP is to choose an optimal policy in the set \( \Pi \) of all admissible policies that maximizes the expected total reward, i.e.

\[
\max_{\pi \in \Pi} E^\pi \left[ r_T(S_T) + \sum_{t=0}^{T-1} r_t(S_t, a_t, w_t) \right],
\]

(2)

where the expectation is taken with respect to \((w_0, w_1, ..., w_{T-1})\). Without loss of generality, we assume \( r_T(S_T) = 0 \) for all \( S_T \).

### 2.2. Quantile

The quantile of a random variable is defined as follow.

**Definition 1.** For \( \tau \in (0, 1) \), the \( \tau \)-quantile of a random variable \( X \) is defined as

\[
Q_\tau(X) = \inf \{ x \mid \text{Prob}(X \leq x) \geq \tau \}.
\]

For \( \tau = 0, 1 \) we define \( Q_0(X) = \inf \{ X \} \) and \( Q_1(X) = \sup \{ X \} \), respectively.

The following properties are implied by the definition.

**Lemma 1.** For a given random variable \( X \), \( Q_\tau(X) \) is a left continuous and non-decreasing function of \( \tau \). Additionally,

\[
\text{Prob}(X \leq Q_\tau(X)) \geq \tau.
\]

### 2.3. Quantile Objective and Assumptions

As its name implies, the goal of the QMDP is to maximize the \( \tau \)-quantile of the total reward,

\[
\max_{\pi \in \Pi} Q_\tau^\pi \left[ \sum_{t=0}^{T-1} r_t(S_t, a_t, w_t) \right].
\]

(3)

Here the quantile is taken with respect to the random variables \((w_0, w_1, ..., w_{T-1})\) and the superscript \( \pi \) denotes the policy we take. So far, we have discussed the finite-horizon MDP i.e. \( T < \infty \). Also, our QMDP framework extends to the infinite-horizon case and the objective is

\[
\max_{\pi \in \Pi} Q_\tau^\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t(S_t, a_t, w_t) \right],
\]

(4)

where \( \gamma \in (0, 1) \) is the discounting factor.

Throughout this paper, we will make the following three assumptions on the structure of MDP. The first assumption is quite standard for MDP and it is about the finiteness of state and action space.

---

1 Here we do not consider the effect of 0-measure set. More precisely, the definition should be \( Q_0(X) = \sup \{ D \in \mathcal{R} \mid P(X \geq D) = 1 \} \) and \( Q_1(X) = \inf \{ U \in \mathcal{R} \mid P(X \leq U) = 1 \} \).
Assumption 1 (State and Action Space). We assume the state space $S$ and the action space $A$ is finite.

Secondly, we require the random variable $w_t$ has a finite support, i.e. $w_t$ is a discrete random variable only taking finite possible values. Indeed, this assumption is crucial and inevitable for the framework of QMDP. Consider our goal is to compute the optimal values and policies for all the quantiles $\tau \in [0,1]$. If $w_t$ has infinite support, this will result in the reward $r_t(S_t, a_t, w_t)$ has infinite support and so does the cumulative reward. However, we have no way to store the infinite support random variable or to query its quantiles unless it has some special well-parametrized structure. Since we want to solve the QMDP problem under generic dynamic and reward functions, the assumption of finite support is necessary.

Assumption 2 (Finite Support). We assume the random variable $w_t \in \mathcal{W}$ has a finite support, i.e. $|\mathcal{W}| < \infty$.

Thirdly, for the notation simplicity in our derivation, we assume that the random variable $w_t$ can be fully recovered with the knowledge of $S_t, a_t$ and $S_{t+1}$, i.e. there exists a function $h_t$ s.t. $w_t = h_t(S_t, a_t, S_{t+1})$. This assumption asserts that there is no additional randomness other than the one that governs the state transition. It follows that the reward $r_t$ will be a function of $S_t, a_t$ and $S_{t+1}$. In practice, this assumption is well satisfied by most of the MDP applications. And we point out that the reason to impose this assumption is purely for the simplicity of derivation and we will fully remove it in section 5.

Assumption 3 (Weak Invertibility). We assume a “weak invertibility” on the function $f_t : S \times A \times \mathcal{W} \rightarrow S$ governs the dynamic system [7]. Here the “weak invertibility” means that there exists a function $h_t : S \times A \times S \rightarrow \mathcal{W}$, s.t. for any $s \in S, a \in A$ and $w \in \mathcal{W}$,

$$h_t(s, a, f_t(s, a, w)) = w.$$

Throughout this paper, we allow the dynamics $f_t(\cdot)$ and the reward function $r_t(\cdot)$ to be non-stationary and non-parametric.

3. Quantile Markov Decision Process (QMDP)
In this section, we present our main result - a dynamic programming procedure to solve the QMDP problem. In section [3.1] we derive the formulation and point out the core of the dynamic programming is an optimization problem. In section [3.2] we provide theoretical guarantees for the value function derived above and the according optimal policy.
3.1. Value Function and Dynamic Programming

We first introduce a value function for the quantile reward of the Markov decision process. Imagine the process starts from state $s$ at the timestamp $t$, we adopt the policy $\pi_t^T$ and receive cumulated reward value as

$$v_{\pi_t^T}(s, \tau) \triangleq Q_\tau \left[ \sum_{k=t}^{T-1} r_k(S_k, a_k, w_k) \bigg| S_t = s \right].$$

Here $\pi_t^T = (\mu_t, \ldots, \mu_{T-1})$ denotes the policy and the action

$$a_k = \mu_k(h_k) = \mu_k(S_t, a_t, \ldots, S_k)$$

for $k = t, \ldots, T - 1$. Since the procedure starts at time $t$, the history here also begins with $S_t$. The value function is a function of state $s$ and the quantile of interest $\tau$ and it is indexed by time $t$. And also, we notice that the value function is dependent on the policy $\pi_t^T$ we choose. The objective of QMDP is to maximize the value $v_{\pi_t^T}(s, \tau)$ by optimizing the policy $\pi_t^T$. So, we define the optimal value function as,

$$v_t(s, \tau) \triangleq \max_{\pi_t^T \in \Pi} v_{\pi_t^T}(s, \tau). \quad (5)$$

When $t = 0$, the value function $v_0(s, \tau)$ is equal to the optimal value in (3).

Then we begin constructing a dynamic programming procedure and derive the optimal value function $v_t(s, \tau)$ backward from $t = T - 1$ to $t = 0$. The key of the dynamic programming procedure is to relate the value functions $v_t(s, \tau)$ with $v_{t+1}(s, \tau)$. Intuitively, $v_{t+1}(s, \tau)$ is obtained by optimizing $\pi_{(t+1):T}$ while $v_t(s, \tau)$ by optimizing $\pi_{t:T}$. The only difference between this two is the choice of $\pi_t = \mu_t(\cdot)$. To connect them, we introduce an intermediate value function by fixing the output action of $\mu_t(s)$ to be $a$:

$$\tilde{v}_t(s, \tau, a) \triangleq \max_{\pi_{t:T} \in \Pi | \mu_t(s) = a} v_{\pi_{t:T}}(s, \tau).$$

On one hand, we can easily relate $v_t(s, \tau)$ with $\tilde{v}_t(s, \tau, a)$ by optimizing $a$,

$$v_t(s, \tau) = \max_{a} \tilde{v}_t(s, \tau, a). \quad (6)$$

On the other hand, we want to establish the relation between $\tilde{v}_t(s, \tau, a)$ and the value function $v_{t+1}(s', \tau')$ at time stamp $t + 1$.

$$\tilde{v}_t(s, \tau, a) = \max_{\pi_{t:T} \in \Pi | \mu_t(s) = a} v_{\pi_{t:T}}(s, \tau)$$

$$= \max_{\pi_{t:T} \in \Pi | \mu_t(s) = a} Q_\tau \left( \sum_{s' \in S} 1\{S_{t+1} = s'|S_t = s, a_t = a\} \left[ \sum_{k=t}^{T-1} r_k(S_k, a_k, w_k) \bigg| S_t = s, S_{t+1} = s' \right] \right). \quad (7)$$
Here the second line is obtained by introducing an indicator function for the state $S_{t+1}$. It is a summation of $|S|$ random variables where each one is associated with a specific state $s'$. To analyze each term more carefully, we have,

$$1 \{ S_{t+1} = s' | S_t = s, a_t = a \} \left[ \sum_{k=t}^{T-1} r_k(S_k, a_k, w_k) \bigg| S_t = s, S_{t+1} = s' \right]$$

$$= 1 \{ S_{t+1} = s' | S_t = s, a_t = a \} r_t(S_t, a_t, w_t) + 1 \{ S_{t+1} = s' | S_t = s, a_t = a \} \left[ \sum_{k=t+1}^{T-1} r_k(S_k, a_k, w_k) \bigg| S_t = s, S_{t+1} = s' \right].$$

The first term here is deterministic with the knowledge of $S_t$ and $S_{t+1}$ from Assumption 3. The interesting point is that the second term seems equal to the value function $v_{t+1}(s', \tau')$. It raises two questions here: First is that what should the choice of $\tau'$ be here? Notice that $v_t(s, \tau, a)$ is equal to a summation of $|S|$ random variables. Suppose we know the quantiles of each random variable, how we can relate the quantile of the summation to the quantile functions of each component in the summation? The following theorem answers these two questions and states the relations between $\tilde{v}_t(s, \tau, a)$ and $v_{t+1}(s', \tau')$.

**Theorem 1 (Value function dynamic programming).** The value functions defined in (5) is equal to the optimal value of the optimization problem $OPT(s, \tau, a, v_{t+1}('\cdot', \cdot))$. Let $S = \{s_1, ..., s_n\}$ and

$$OPT(s, \tau, a, v_{t+1}('\cdot', \cdot)) \triangleq \max_{q} \min_{i \in \{q_i \neq 1 | i = 1, 2, ..., n\}} \left[ v_{t+1}(s_i, q_i) + r_t(s, a, w_t) \right],$$

$$= \max_{q} \min_{i \in \{q_i \neq 1 | i = 1, 2, ..., n\}} \left[ v_{t+1}(s_i, q_i) + r_t(s, a, h_t(s, a, s_i)) \right],$$

subject to $\sum_{i=1}^{n} p_i q_i \leq \tau$

$q_i \in [0, 1]$, $p_i = P(S_t = s, a_t = a, S_{t+1} = s_i)$.

Here $w_t = h_t(s, a, s_{i+1}) = h_t(s, a, s_i)$ is from the Assumption 3 on invertibility. We use $v_{t+1}('\cdot', \cdot)$ to denote the value function at $t+1$ and to emphasize that it is a function of state and quantile. The decision variable here is the vector $q$. Then,

$$\tilde{v}_t(s, \tau, a) = OPT(s, \tau, a, v_{t+1}('\cdot', \cdot)).$$

The optimization problem appeared in the theorem is actually motivated from the following Lemma 2 which computes the quantile of a summation of random variables. We all know that the expectation of a summation of random variables equals to the summation of the expectations. This linearity makes it possible the backward dynamic programming in the traditional MDP. Here the
Lemma 2 plays a similar role in that it relates the quantile of the summation of random variables to the quantiles of each random variables. This result together with the optimization algorithm in next section can be of independent interest for other applications concerning quantiles of random variables.

**Lemma 2.** Consider \( n \) discrete random variables \( X_i, i = 1, \ldots, n \), (here and hereafter, by discrete random variables, we means that \( X_i \) take values on a finite set) and another \( n \) random variables \( Y_i \),

\[
\sum_{i=1}^{n} Y_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} 1(Y_i = 1) = 1.
\]

And we know that \( \{X_i\} \)'s and \( \{Y_i\} \)'s are independent. Then the quantile of the summation

\[
Q_{\tau} \left( \sum_{i=1}^{n} X_i Y_i \right),
\]

is given by the following optimization problem,

\[
\max_{\mathbf{q}} \min_{i \in \{q_i \neq 1 \mid i = 1, 2, \ldots, n\}} Q_{q_i}(X_i) \tag{9}
\]

subject to

\[
\sum_{i=1}^{n} p_i q_i \leq \tau, \quad q_i \in [0, 1], \quad p_i = P(Y_i = 1).
\]

Here \( \mathbf{q} = (q_1, \ldots, q_n) \) and \( Q_{q_i}(\cdot) \) is the \( q_i \)-quantile of \( X_i \).

The key idea for the proof of Theorem 1 is to introduce a random variable \( X_i \) such that its quantile \( Q_{\tau}(X_i) = Q_{\tau}(v_{t+1}(s_i, \tau) + r_i(s, a, h(s, a, s_i))) \) for all \( \tau \in [0, 1] \). Then the right hand side of (8) is in the same form as (9) and the Lemma 2 applies. More details can be referred to the Appendices.

By putting together Theorem 1 together with (6), we establish the relation between \( v_t(s, \tau) \) and \( v_{t+1}(s', \tau') \) and build the foundation for a backward dynamic programming to compute optimal value functions. Importantly, the algorithm is to derive the whole value functions, i.e., at time stamp \( t \), the output we obtain is the function \( v_t(\cdot, \cdot) \) rather than its evaluation at some specific \( s \) and \( \tau \). In the following, we will state that the value functions we obtain from the backward dynamic programming is equal to the optimal value for the QMDP and it defines the optimal policy accordingly.

### 3.2. Optimal Value and Optimal Policy

In this section, we derive the optimal value and the implied policy based on the backward dynamic programming routine in Theorem 1. The procedure of computing value functions is illustrated in Figure 1.
Theorem 2 (Optimal Value Function). Let \( v_T(s, \tau) = 0 \) for all \( s \in S \) and \( \tau \in [0,1] \). Iteratively, we compute

\[
v_t(s, \tau) = \max_a \text{OPT}(s, \tau, a, v_{t+1}(\cdot, \cdot)),
\]

for \( t = T - 1, \ldots, 0 \), then we have

\[
v_0(s, \tau) = \max_{\pi \in \Pi} Q^\pi \sum_{k=0}^{T-1} r_k(S_k, a_k, w_k).
\]

Theorem 3 (Optimal Policy). We augment the state \( S_t \) with a quantile \( \tau_t \) to assist the execution of the optimal policy. At the initial state \( s_0 \) and \( \tau_0 = \tau \), we define our initial policy function as,

\[
\pi_0 : \mu_0(s_0, \tau_0) = \arg \max_a \bar{v}_0(s_0, \tau_0, a).
\]

At time stamp \( t \), we execute the output of \( \mu_t \) and then the process arrives on state \( S_{t+1} \). Let \( q^* \) be the solution to the optimization problem \( \text{OPT}(S_t, \mu_t(S_t, \tau_t), \tau_t, v_{t+1}(\cdot, \cdot)) \). Here \( v_{t+1}(\cdot, \cdot) \) is computed as in Theorem 2. The term \( \tau_{t+1} \) is assigned as

\[
\tau_{t+1} = q^*_i,
\]

for the specific \( i \) that \( S_{t+1} = s_i \). Then we define \( \pi_{t+1} \) as

\[
\pi_{t+1} : \mu_{t+1}(S_{t+1}, \tau_{t+1}) = \arg \max_a \bar{v}_{t+1}(S_{t+1}, \tau_{t+1}, a).
\]

The policy \( \pi = (\pi_0, \ldots, \pi_T) \) defined above is the optimal policy for the objective and obtains the optimal value \( v_0(s_0, \tau_0) \).
The above two theorems can be easily implied from Theorem 1 and the proofs are deferred to the appendices. Theorem 2 verifies that the value function computed via backward dynamic programming is equal to the optimal quantile value. For Theorem 3, it is worth to notice that unlike the case of normal MDP, the optimal policy \( \pi_t \) for QMDP is a function of the history \( h_t = (S_0, a_0, ..., S_t) \) instead of simply the current state \( S_t \). The point of surprise is that all the history \( h_t \) is summarize into the quantile level \( \tau_t \). In other words, \( \tau_t \) is a function (although not explicit) of the history and plays a role like the “summary statistics”. The theorem tells us that the optimal policy \( \pi_t \) is a function of only the current state \( S_t \) and the “summary statistics” \( \tau_t \). Intuitively, this augmented quantile level \( \tau_t \) reflects the historical performance of the MDP. Higher quantile level will encourage more aggressive policy in the remaining periods while lower quantile level will encourage conservative moves. For example, if we start with \( \tau_0 = \tau = 0.5 \) which means our ultimate goal is to maximize the median cumulative reward over 0 to \( T \), then at some time stamp \( t \) in between, if we have already achieved pretty good reward, i.e., a large \( \sum_{k=0}^{t} r_k \), the augmented quantile level \( \tau_t \) will decrease to some value smaller than 0.5 accordingly. This will drive us to take relatively conservative moves in the future, and vise versa.

4. Algorithm and Computation

In this section, we present the algorithms for solving QMDP and discuss its computational aspect. As mentioned earlier, the key for computing the value function is to solve the optimization problem \( OPT(s, \tau, a, v_{t+1}(.)) \). In the following, we will first provide an efficient algorithm for solving the \( OPT \) problem and then analyze its complexity.

4.1. Algorithm for Solving the Optimization Problem \( OPT(s, \tau, a) \)

We formulate \( OPT(s, \tau, a, v_{t+1}(.,\cdot)) \) in a more general way as follow:

\[
OPT \triangleq \max_q \min_{i \in \{1, 2, ..., n\}} \sum_{i=1}^{n} p_i q_i \leq \tau,
\]

subject to

\[
\sum_{i=1}^{n} p_i = 1,
\]

\[q_i \in [0, 1], \quad \text{for} \quad i = 1, ..., n.
\]

Here \( \tau \) and \( p_i \)'s are known parameters as well as the function \( g : \{1, ..., n\} \times [0, 1] \rightarrow \mathcal{R} \). The decision variable is \( q = (q_1, ..., q_n) \). Furthermore, we assume \( g(i, \cdot) \) to be a left continuous and piecewise constant function with finite breakpoints for all \( i \). The variable \( i \) refers to the state back in the QMDP settings. We will show later that these assumptions will be satisfied for value functions of
QMDP with finite-state space and discrete rewards. Therefore, we can represent and encode each function \( g(i, \cdot) \) with a set of breakpoint-value pairs

\[
\left\{ \left( b_i^{(1)}, v_i^{(1)} \right), \ldots, \left( b_i^{(n_i)}, v_i^{(n_i)} \right) \right\}
\]

where \( n_i \) is the number of the pairs. Then we have

\[
g(i, x) = \begin{cases} 
  v_i^{(1)}, & \text{for } x \in \left[ b_i^{(1)}, b_i^{(2)} \right], \\
  v_i^{(k)}, & \text{for } x \in \left( b_i^{(k)}, b_i^{(k+1)} \right] \text{ and } k = 2, \ldots, n_i.
\end{cases}
\]

Here we define \( b_i^{(1)} = 0 \) and \( b_i^{(n_i+1)} = 1 \) for all \( i \).

**Algorithm 1** Algorithm for the optimization problem \([10]\)

1: **Input:** \( \{ p_i, g(i, \cdot) \mid i = 1, \ldots, n \} \)
2: **Initialize** \( k_1 = \ldots = k_n = 1, \tau_{tmp} = 0. \)
3: Let \( u_i = g(i, 0) = v_i^{(1)}, i \in S = \{ 1, \ldots, n \} \) and \( u = \min_{i \in S} u_i. \)
4: Let \( f(0) = u. \)
5: while \( S \) is non-empty do
6: \( S_0 = \arg\min_{i \in S} u_i \)
7: \( \tau_{new} = \tau_{tmp} \)
8: for \( i \in S_0 \) do
9: if \( k_i = n_i \) then
10: \( \tau_{new} = \tau_{new} + p_i \left( 1 - b_i^{(k_i)} \right) \)
11: \( S = S \setminus \{ i \} \)
12: else
13: \( \tau_{new} = \tau_{new} + p_i \left( b_i^{(k_i+1)} - b_i^{(k_i)} \right) \)
14: \( k_i = k_i + 1 \)
15: Update \( u_i = v_i^{(k_i)} \)
16: Let \( f(\tau) = u \) for \( \tau \in (\tau_{tmp}, \tau_{new}] \).
17: Update \( \tau_{tmp} = \tau_{new} \)
18: Update \( u = \min_{i \in S} u_i \)
19: **Return** \( f(\cdot) \)

We provide the details for solving \( OPT \) in Algorithm \( \square \). The idea is quite straightforward: we start with \( q_i = 0 \) for all \( i \) and gradually increase the specific \( q_i \) that has the smallest value of \( g(i, q_i) \) until the constraint \( \sum_{i=1}^{n} p_i q_i \leq \tau \) is violated. The \( g(i, q_i) \) that has smallest value is exactly
Figure 2  Step-by-step execution of Algorithm 1 with $n = 3$ sample $g(\cdot, \cdot)$ functions: numbers inside and along the block represent the values and breakpoints of the input functions $g(i, \cdot)$. The shaded regions reflect the progress of the algorithm. In the end, the output is $f$.

the bottleneck for the objective value. By increasing the corresponding $q_i$, we keep improving the objective value. The output of the algorithm $f(\cdot)$ restores the optimal values of OPT as a function of $\tau \in [0, 1]$. We illustrate the execution of the algorithm for a toy example of $n = 3$ in Figure 2.

4.2. Algorithm for Solving QMDP

In this subsection, we summarize the previous results and provide the algorithm for solving QMDP as Algorithm 2. It is obtained by putting together the Algorithm 1 with Theorem 2 and 3.

4.3. Complexity Analysis and Approximation

The computation cost of QMDP (Algorithm 2) is mostly concentrated on the part of computing the value functions. First, it is easy to show that all the value functions are piecewise constant functions. This is because when the input functions of $OPT$ are piecewise constant, the $OPT$ procedure will output a piecewise constant function as well. Also, we can see that the complexity of Algorithm 1 is linear of the number of breakpoints for its output functions. Based on these facts, we have the following corollary.

**Corollary 1.** When the rewards are integers and bounded, $|r_t| \leq R$ for all $t$, then the complexity for computing value functions for QMDP is $O(\text{AST} \cdot \max(RT, S))$. Here $T$ is the length of time horizon. $A = |A|$ and $S = |S|$ are the numbers of action and state space respectively.

The proof of this corollary is straightforward: When the reward is integer and bounded by $R$, the cumulated reward is bounded by $RT$. So any value function has at most $RT$ breakpoints, which means each call of $OPT$ will induce at most $O(RT)$ complexity. Along with that, it will have a $O(S)$ read and write complexity. So each iteration has $O(\max(RT, S))$ complexity. Since there are $AST$ iterations in total, the overall complexity is $O(\text{AST} \cdot \max(RT, S))$. The case is more complicated with respect to non-integer rewards, where we have no simple way to bound the number of breakpoints for the value functions. In other words, the value function can be exponentially complicated with the procedure of backward dynamic programming, so that the cost
Algorithm 2 Algorithm for Solving QMDP

1: **Input:** Transition probabilities \( P(S_t, a, S_{t+1}) \), reward function \( r_t(s_t, a_t, w_t) \), time horizon \( T \).

2: **Computing Value Functions:**

3: **Initialize** Let \( S = \{s_1, ..., s_n\} \), \( v_T(s_i, \tau) = 0 \) for all \( i = 1, ..., n \) and \( \tau \in [0, 1] \).

4: **for** \( t = T - 1, ..., 0 \) **do**

5: **for** \( i = 1, ..., n \) **do**

6: **for** \( a \in A_t \) **do**

7: \( p_{tmp}(s_j) = P(S_t = s_i, a, S_{t+1} = s_j) \) for \( j = 1, ..., n \)

8: \( v_{tmp}(s_j, \tau) = v_{t+1}(S_{t+1} = s_j, \tau) + r_t \) for \( j = 1, ..., n \) and \( \tau \in [0, 1] \)

9: \( \tilde{v}(s_i, \tau, a) = OPT(\{p_{tmp}(s_j)\}_{j=1}^n, \{v_{tmp}(s_j, \tau)\}_{j=1}^n) \)

10: \( v_t(s_i, \tau) = \max_a \tilde{v}(s_i, \tau, a) \)

11: **Output:** \( \{v_t(s, \tau), \tilde{v}(s, \tau, a)\}_{t=0}^{T-1} \) for all \( s \in S, a \in A \) and \( \tau \in [0, 1] \)

12: **Execution:**

13: **Initialize** \( S_0 = s \) and our goal is to maximize \( \tau \) quantile. Let \( R = 0 \) and \( \tau_0 = \tau \).

14: **for** \( t = 0, ..., T - 1 \) **do**

15: Take action \( a_t = \arg \max_a \tilde{v}_t(S_t, \tau_t, a) \)

16: Transit from \( S_t \) to \( S_{t+1} = s_j \) for some \( j \in \{1, ..., n\} \)

17: \( R = R + r_t(s_t, a_t, s_{t+1}) \)

18: Let \( q^* \) be the optimizer of \( OPT(S_t, a_t, v_{t+1}('', '') \)

19: Update \( \tau_{t+1} = q^*_j \)

20: **Output:** Cumulative Reward \( R \)

To restore the value function will also grow exponentially. To prevent this explosion, we provide two solutions: (1) to truncate the rewards to integers; (2) to do approximation for the value functions. For the first solution, if we still want to preserve the computation precision, we can scale up the rewards before the truncation. For the second solution, by doing approximation, we only restore the function value at \( N \) uniform breakpoints. The number \( N \) is up to the choice of user and can be as large as 10000, which means that we restore only and all the quantiles values with an interval of 0.0001.

From the above analysis, we can observe that the bottleneck for the complexity of our algorithm lies in the complexity of value function. In traditional MDP, the value function is a function of state \( s \) and time stamp \( t \). However, in QMDP, for each given \( s \) and \( t \), we need to compute and memorize the optimal values for all the quantiles so as to derive that for time \( t - 1 \), which is inevitable. Therefore, there is not much room for improvement on this complexity upper bound under a generic setting.
5. Extensions

In this section, we discuss two extensions for the QMDP framework. First we establish the results of optimal value and policy for infinite-horizon case in section 5.1. Then we discuss how to relax the Assumption 3 in section 5.2.

5.1. Infinite-Horizon QMDP

For the infinite-horizon QMDP, the objective is
\[
\max_{\pi \in \Pi} Q_{\tau}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t (S_t, a_t, w_t) \right].
\]

Here \( r_t = r(S_t, a_t, w_t) \) is stationary and \( \gamma \in (0, 1) \) is the discounting factor. The policy \( \pi = \{\mu_t\}_{t=0}^\infty \) consists a sequence of decision functions and \( \mu_t \) maps the historical information \( h_t = (S_0, a_0, ..., S_{t-1}, a_{t-1}, S_t) \) to an admissible action \( a_t \in \mathcal{A}_t \subset \mathcal{A}. \) Let the value function
\[
v(s, \tau) \triangleq \max_{\pi \in \Pi} Q_{\tau}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t (S_t, a_t, w_t) \big| S_0 = s \right].
\]

(11)

Similar to the infinite-horizon MDP, we propose a value iteration procedure to compute the QMDP value function. The result is formally stated as Theorem 4. We use \( k \) to denote the iteration number here to distinguish it with the index notation \( t \) in Theorem 2 which is the time stamp for backward dynamic programming.

**Theorem 4 (Infinite-horizon Optimal Value Function).** We propose a value iteration procedure as:
\[
v^{(0)}(s, \tau) = 0,
\]
\[
\tilde{v}^{(k+1)}(s, \tau, a) = OPT(s, \tau, a, \gamma v^{(k)}(\cdot, \cdot)),
\]
\[
v^{(k+1)}(s, \tau) = \max_a \tilde{v}^{(k+1)}(s, \tau, a),
\]

Then we have,
\[
\lim_{k \to \infty} v^{(k)}(s, \tau) = v(s, \tau),
\]
for any \( s \in \mathcal{S} \) and \( \tau \in [0, 1] \). Furthermore, since the function \( v(s, \tau) \) is a monotonous function for \( \tau \), the convergence is uniform.

The key to the proof of the theorem is to show that the \( OPT \) procedure, as an operator, features the same contractive mapping property as the Bellman operator in traditional MDP. And the contraction rate is simply the discounting factor \( \gamma \). The detailed proof can be referred in Appendices. Based on the optimal value function, we have the following result on the optimal policy.
**Theorem 5 (Infinite-horizon Optimal Policy).** Let $v(\cdot, \cdot)$ be the optimal value function as in Theorem 4 and

$$
\hat{v}(s, \tau, a) \triangleq \text{OPT}(s, \tau, a, \gamma v(\cdot, \cdot)).
$$

We augment the state $S_t$ with a quantile $\tau_t$ to assist the execution of the optimal policy. At the initial state $s_0$ and $\tau_0 = \tau$, we define our initial policy function as,

$$
\mu_0(s_0, \tau_0) = \arg \max_a \hat{v}(s_0, \tau_0, a).
$$

At time stamp $t$, we execute $\pi_t$ and then arrive on state $S_{t+1}$. Let $q^*$ be the solution to the optimization problem $\text{OPT}(S_t, \mu_t(S_t, \tau_t), \tau_t, \gamma v(\cdot, \cdot))$. The term $\tau_{t+1}$ is defined as

$$
\tau_{t+1} = q^*_i,
$$

for the specific $i$ that $S_{t+1} = s_i$, and $\pi_{t+1}$ is defined as

$$
\mu_{t+1}(S_{t+1}, \tau_{t+1}) = \arg \max_a \hat{v}(S_{t+1}, \tau_{t+1}, a).
$$

The policy $\pi = \{\mu_t\}_{t=0}^\infty$ defined above is the optimal policy for the objective $\text{11}$ and obtains the optimal value $v(s_0, \tau_0)$.

The value iteration procedure appears similar to the backward dynamic programming procedure for the finite-horizon case. This is because we can always interpret the finite-horizon reward as an approximation for the infinite-horizon reward by truncating off the reward after time $T$.

5.2. Relaxation of Assumption

**Algorithm 3** Algorithm for Computing Quantiles of the Sum of Random Variable

1: **Input**: Two random variables and their probability-value pairs:
2: $X_1 : \{(p_1, a_1), ..., (p_n, a_n)\}$ and $X_2 : \{(q_1, b_1), ..., (q_m, b_m)\}$
3: **Initialize** Let $D = \{\}$.
4: **for** $i = 1, ..., n$ **do**
5:  **for** $j = 1, ..., m$ **do**
6:    $D = D \cup \{(p_i q_j, a_i + b_j)\}$
7: **Merge** pairs in $D$ with the same value by:
8: Removing two pairs with same value and appending a new pair with the same value but the probability as the sum of the two removed probabilities.
9: **Output**: $D$: the set of probability-value pairs to represent the distribution of $X_1 + X_2$. 
We mentioned earlier in the model setup that Assumption 3 is introduced for the purpose of the simplicity in mathematical derivation. In this subsection, we will discuss the removal of the assumption with the help of another interesting lemma about the quantiles of random variables. In Assumption 3, we require that the reward function has some invertibility, in other words, the reward is a function of $S_t, a_t$ and $S_{t+1}$. Under this, recall that the input for the $OPT$ procedure is

$$v_{t+1}(s_i, q_i) + r_t(s, a, w_t) = v_{t+1}(s_i, q_i) + r_t(s, a, h_t(s, a, s_i)).$$

The assumption will render the second part as a deterministic constant (in that it is a function of $S_t = s, a_t = a$ and $S_{t+1} = s_i$). Therefore, from $v_{t+1}(\cdot, \cdot)$, we can easily obtain the input for $OPT$ by increasing it with the constant. However, when the assumption is not true, the reward $r_t(s, a, w_t)$ is a random variable and no longer a constant determined by $S_t, a_t$ and $S_{t+1}$. Theorem 1 and its proof tells us that this difference does not matter as long as we can compute the $\tau$-th quantile of the sum

$$v_{t+1}(s_i, q_i) + r_t(s, a, w_t)$$

for any $\tau \in [0, 1]$. As in the proof of Theorem 1, we introduce a random variable $X_{s_i}$, s.t.

$$Q_\tau(X_{s_i}) = v_{t+1}(s_i, \tau)$$

for any $\tau \in [0, 1]$. Let $W_{s_i} = r_t(s, a, w_t)$. It is easy to see that given $S_{t+1} = s_i$ (or conditional on $S_{t+1}$), $X_{s_i}$ and $W_{s_i}$ is independent. Then, what left is to come up with an algorithm that takes as input the quantile functions of two independent random variables and outputs the quantile function of their summations. For discrete random variables, this indeed can be done efficiently. First, let the input be two random variables $X_1$ and $X_2$, which are represented by two sets of probability-value pairs $\{(p_1, a_1), \ldots, (p_n, a_n)\}$ and $\{(q_1, b_1), \ldots, (q_m, b_m)\}$, i.e. $P(X_1 = a_i) = p_i$ and $P(X_2 = b_i) = q_i$. The idea is that when $X_1$ and $X_2$ are independent, we can easily compute the distribution of their sum and therefore the quantile function as Algorithm 3.

**Corollary 2.** The complexity of Algorithm 3 is $O(mn)$, where $n$ and $m$ are the number of breakpoints for the quantile function of $X_1$ and $X_2$ respectively. In the context of QMDP, the complexity is upper bounded by $O(R^2T)$. Therefore, the complexity bound for QMDP with arbitrary reward function will be $O(\max(\text{AST} \cdot \text{max}(R^2T, S)))$. Here $T$ is the length of time horizon. $A = |A|$ and $S = |S|$ are the numbers of action and state space respectively.

With relaxing the Assumption 3 we need to compute the sum of two random variables every time for the input for $OPT$. This operation brings in an order of $O(R)$ in the complexity. However, since the overall complexity is at most quadratic with respect to all the variables, the algorithm is still efficient and scalable.
6. Experiments

We present two examples to demonstrate the model of QMDP, one for an MDP game and one for HIV optimal treatment.\footnote{The code for all the experiments in this section will be released for reproduction. We will also release a general-purpose QMDP class object for the usage in other applications, wrapping up all the details of OPT optimization and value function derivation.}

6.1. A MDP Game

6.1.1. Background and Model Formulation

We construct a toy game and perform the simulations to demonstrate the QMDP time complexity with respect to the state size, time horizon and reward structure. In this game, a player is moving along a chain and keeps receiving rewards dependent on his location. He takes the action to stay or to move and his goal is to maximize expected cumulative reward over time horizon. We formulate the game in the language of MDP as follows:

- **Time Horizon**: We assume there are \( T \) decision periods.
- **State**: We denote the state by \( S_t, t = 0, \ldots, T \) where \( S_t \in S = \{1, \ldots, n\} \).
- **Action**: At each time \( t \), the player takes an action \( a_t \in A = \{\text{Stay, Move}\} \).
- **Transition Probability**:
  - When \( a_t = \text{Stay} \), the player will stay on his location with probability 1.
  - When \( a_t = \text{Move} \), the player will move randomly to one of its neighbors. When the player starts from the end of the chain, then he moves to his single neighbor with probability 1.
- **Rewards**: When the player stays at state \( i \) at the beginning of a time period, he receives a reward \( R_i \).

Figure 3 illustrates the model of this toy example, in which the arrows represent the possible moving directions.

![An illustration of the toy game model](image)

6.1.2. Results

First we discuss the computation complexity with respect to several key model characteristics such as time horizon \( T \), state size \( n \) and maximum reward \( R_{\text{max}} \). We run \( 10^5 \) simulation trials in total (on a laptop with 2.8 GHz Intel Core i7) and plot the average computation time in the Figure 4. Each time the transition probabilities are randomly generated and the rewards are randomly sampled integers no greater than \( R_{\text{max}} \).
Figure 4  Runtime of QMDP on the toy game. Base model parameters: time horizon $T = 20$, state space size $n = 10$, max reward $R_{max} = 20$. For each experiment, we change one single parameter and monitor the running time.

We observe that the running time is quadratic in time horizon $T$ of the model. The graph in the middle showcases that the running time is linear of state size $n$. For the plot on the right, the runtime grows linear and then fluctuate after the $R_{max}$ reaches certain level. This does not contradict the complexity analysis in Corollary 1: the quadratic complexity over $R_{max}$ is an upper bound but not necessarily to be tight for every trial. Overall speaking, the algorithm is efficient and can elegantly handle large-scale problems.

Figure 5  Value Functions: We randomly initialize three toy games with $T = 20$, $S_0 = 5$, $n = 15$ and $R_{max} = 8$. The grey dashed lines are the cumulative density function for simulations with the execution of optimal MDP policy. The dark red lines are the optimal quantile rewards computed via QMDP. The difference between the two lines demonstrate the MDP risk and the space for quantile improvement.

Quantile optimal value tells us the maximal quantile cumulative reward that is achievable for an MDP. While the traditional MDP solution outputs an optimal policy $\pi^*$ that maximized the
expectation of cumulative reward, it does not necessarily maximize the quantiles. In Figure 5, we demonstrate with three random trials the sub-optimality of $\pi^*$ in terms of quantiles. First we compute the $\pi^*$ using the dynamic programming for traditional MDP. Then we follow $\pi^*$ to 20,000 simulation trials and plot their cumulative rewards as the red lines in Figure 5 (in cumulative density function). A point $(q, r_0)$ on the red line means that $\pi^*$ can achieve at least $r_0$ cumulative reward with probability $1 - q$. The black lines are the plots for optimal value function of QMDP. And a point $(q, r_1)$ on the black lines means the optimal $q$ quantile reward is $r_1$, i.e., there exists a policy $\pi$ s.t. it can achieve at least $r_1$ cumulative rewards with probability $1 - q$. The gap between the red line and black line tells the space of improvement in terms of quantile. For the plot on the left and right, the difference is not too large uniformly for all the quantiles. This means that the policy $\pi^*$, although not necessarily achieve quantile optimality, is quite stable and robust. However, for the plot in the middle, we can see that there is a large space of improvement for lower quantiles, which means that there exists certain quantile policy that can improve greatly the bad cases.

On one hand, the comparison of QMDP and the cumulative density function of traditional MDP assesses the risk with respect to the policy $\pi^*$. On the other hand, it gives us a sense whether we want to adopt some quantile policies as for the example in the middle. Also, we want to point out that, due to the nature of this toy game, the traditional MDP optimal policy $\pi^*$ is also optimal for many quantiles. But this is always not the case, like for the asset selling and machine replacement examples.

6.2. Antiretroviral Therapy (ART) Optimal Initiation Time

6.2.1. Background and Model Formulation  HIV is a serious health and development challenge and around 37 millions people are living with HIV around the world. Considerable resources have been devoted to expanding treatment of HIV. Effective antiretroviral therapy (ART), which suppresses an individual’s HIV viremia, reduces HIV associated morbidity and mortality for treated individuals (Tanser et al. 2013) and thus transform HIV into a chronic disease. However, the debate around optimal time to initiate ART has lasted for decades due to the side effects. One major concern is the increased cardiac risk during the treatment process. The delay of ART causes the patients to sacrifice immunological benefits from treating HIV in the near future but shields the patients from the cardiac mortality. Therefore, a good medical decision making tool will help identify the optimal time to initiate ART. Negoescu et al. (2012) approached the problem by constructing a traditional MDP model and thus identified the optimal policy that maximized the expected cumulative quality adjusted life expectancy. However, in the era of Patient-centered care, patient preference should be included in the measurement of health care quality. According to Fraenkel et al. (2003), patient’s risk attitude is associated with his or her treatment preference.
QMDP bridges the gap between the traditional MDP and patient’s risk attitude that is imposed by setting the quantile threshold in QMDP.

The formulation of ART optimal initiation time problem is straightforward and is similar to the MDP formulation.

- **Time Horizon**: We assume the patient will be assessed at each time period $t$ and $t \in \{0, 1, 2, \cdots, T\}$

- **State**: We characterize the state of a HIV-infected patient at time $t$ as $S_t = (c_t, y_t, d_t)$. The state is a function of the patient’s CD4 cell count ($c_t$), age ($y_t$) and ART treatment duration ($d_t$). We divide the continuous CD4 cell counts into $L$ bins $C = \{C_1, C_2, \cdots, C_L\}$. In addition, we create an absorbing state for death, $D$. For age, we have $y_t \in [Y_0, Y_N]$, where $Y_0$ is the starting age and $Y_N$ is the terminal age of the patient. For treatment duration, $d_t \geq 0$ with 0 denotes the delay of ART. Once the patient has started ART, $d_t$ will increase by one unit after each time step.

- **Action**: At each time $t$, the patient can take an action $a_t \in \{W, Rx\}$, where $W$ represents waiting for another period and $Rx$ means starting ART treatment immediately.

- **Transition Probability**: The transition probability $P_k(S_t, a_t, S_{t+1})$ depends on the patient’s current state $S_t$, the action $a_t$ at time $t$ and the state $S_{t+1}$ at time $t+1$. There are two basic types of transitions: the transition probability between CD4 counts $C$ level to another CD4 counts level and the transition from CD4 counts level $C$ to terminal state death $D$.

- **Rewards**: There are 2 types of reward: immediate reward and terminal reward. First, immediate reward ($R^I$) is measured as the quality adjusted life years (QALYs) the patient experienced from $S_k$ to $S_{t+1}$ ($S_t \in C, S_{t+1} \in C, D$). To simplify the calculation, we assume patient will die uniformly from $t$ to $t+1$ and the resulted immediate reward would be halved. Second, the terminal reward ($R^E$) is cumulative QALYs from the patient who past the terminal age ($Y_N$)

### 6.2.2. Results

We model HIV infected women from United States from 20 years old to 90 years old and they will be presented with 2 choices every 6 months: starting ART immediately or delaying ART for another 6 months. To obtain the cumulative QALYs after the terminal age, we perform a cohort simulation that utilizes the same model parameters including transition probabilities and utility. All model parameters are showcased in Appendix B.

Figure [6] shows the optimal actions for the HIV infected women to follow at different ages. We consider three QMDP models: with different quantile threshold ($\tau = 0.2, 0.5, 0.8$). As $\tau$ increases,
the patient becomes less risk averse. There are several interesting observations can be made from the figure. To start with, policies obtained from difference instances of QMDP model exhibit the similar patterns that is showcased in original MDP paper Negoescu et al. (2012). Patient will tends to choose to delay ART to next time period when the patients are older or their CD4 counts are high. In both cases, the reduced HIV associated morbidity and mortality from initiating ART is outweighed by the induced cardiac risks. When patients are in the later stages of their lives, the induced cardiac risks are substantial because of the higher baseline cardiac risks at older ages. Moreover, when patients’ CD4 counts is high, patient are relatively healthy and additional benefits from starting ART is diminishing. Furthermore, different risk attitudes of patients will cause different treatment preferences. As shown in Figure 6, patients will tend to start ART when the patient become less risk averse since patient are more willing to embrace the elevated cardiac risks in order to enjoy the immunological benefits from treating HIV with ART. Aside from the general patterns that matches the expectation, instabilities exist in the QMDP policies, especially within the region where action switch are made. Such effects stems from the non-convexity in the transition probabilities, which requires further research on the necessary conditions that eliminates such instabilities. Overall, QMDP exhibits similar properties in terms of balancing various risks that can be also derived from traditional MDP and more importantly it demonstrates an elegant way to incorporate patients’ risk attitudes that will contribute to the provisioning of patient centered care.

7. Conclusions, Limitations, and Future Directions
In this paper, we present a novel quantile framework for Markov Decision Process in which the objective is to maximize the quantiles of the cumulative rewards. Several theoretical results on quantiles of random variables are presented and they contribute to an efficient algorithm that solve the QMDP. We present three experiments and analyze the results from the perspective of...
optimal value, optimal policy and computation cost. In the first place, we recommend people to run QMDP for all the MDP applications as a risk diagnose. On one hand, it can provide us an alternative and safe strategies for MDP models with high risk; on the other hand, it meets the need of quantile criteria in many MDP application. Some open questions are: What is the cause of the risk of the MDP? As the toy game experiment in Section 6.1, the traditional MDP optimal policy also provides good quantile rewards for some trials, while the policy performs rather bad in certain quantiles in other cases. The reason that causes this discrepancy deserves more understanding. Also, the quantile criteria naturally extends to the MDP with a utility-function criteria for which our goal is to optimize the utility of the cumulative reward. The good news is that the quantile criteria is invariant for a monotonous utility function. So the QMDP provides a good alternative solution for the MDPs with utility function objective that is hard to optimize. As for future work, we plan to apply the QMDP framework on more practical problems as a tool to understand the underlying risk and to generate quantile policies reflecting risk attitude.
Appendix A: Proof of lemmas and theorems

A.1. Proof of Theorem \(2\)

We first introduce several lemmas.

**Lemma 3.** Consider \(n\) random variables \(Y_i\) for \(i = 1, ..., n\), s.t.
\[
\sum_{i=1}^{n} Y_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} 1(Y_i = 1) = 1
\]
and \(n\) random variables \(X_i\), and \(Y_i \perp X_i\), we have
\[
P\left(\sum_{i=1}^{n} X_i Y_i \geq C\right) = \sum_{i=1}^{n} P(Y_i = 1) P(X_i \geq C),
\]
\[
P\left(\sum_{i=1}^{n} X_i Y_i > C\right) = \sum_{i=1}^{n} P(Y_i = 1) P(X_i > C),
\]
\[
P\left(\sum_{i=1}^{n} X_i Y_i \leq C\right) = \sum_{i=1}^{n} P(Y_i = 1) P(X_i \leq C),
\]
\[
P\left(\sum_{i=1}^{n} X_i Y_i < C\right) = \sum_{i=1}^{n} P(Y_i = 1) P(X_i < C),
\]
for any \(C \in \mathbb{R}\).

[Proof of Lemma 3] We show the first equation and the rest are similar.
\[
P\left(\sum_{i=1}^{n} X_i Y_i \geq C\right) = \sum_{i=1}^{n} P\left(\sum_{i=1}^{n} Y_i = 1, X_i Y_i \geq C\right)
\]
\[
= \sum_{i=1}^{n} P(Y_i = 1, X_i \geq C)
\]
\[
= \sum_{i=1}^{n} P(Y_i = 1) P(X_i \geq C)
\]

**Lemma 4.** For a discrete random variable \(X\) and \(C \in \mathbb{R}\), let
\[
q = P(X < C),
\]
then
\[
Q_{q+\epsilon}(X) \geq C
\]
for any \(\epsilon > 0\). Here the \(Q_\tau(\cdot)\) is the quantile function as in Definition \(7\).

[Proof for Lemma 4] For any \(C_0 < C\), \(P(X \leq C_0) \leq q < q + \epsilon\). Consider the definition of the quantile, we finish the proof.

With the help of the above two lemmas, we begin the proof of the lemma \(2\) namely for computing the quantile of the summation of random variables.
Proof for Lemma 2. On one hand, we show that
\[ Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right) \geq \max_q \min_{i \in \{q_i \neq 1\mid i = 1, 2, \ldots, n\}} h_i(q_i). \] (12)

Let
\[ f(q) = \min_{i \in \{q_i \neq 1\mid i = 1, 2, \ldots, n\}} h_i(q_i). \]

To show (12), we only need to show for any feasible \( q = (q_1, \ldots, q_n) \),
\[ Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right) \geq f(q). \] (13)

By saying \( q \) is feasible, we mean \( q = (q_1, \ldots, q_n) \) is subject to
\[ \sum_{i=1}^{n} p_i q_i \leq \tau, \]
\[ q_i \in [0, 1], \quad p_i = P(Y_i = 1). \]

We show (13) by contradiction. If
\[ Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right) < f(q), \]
then there exist \( \epsilon > 0 \),
\[ Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right) \leq f(q) - \epsilon. \]

By the definition of quantile, this implies,
\[ P \left( \sum_{i=1}^{n} X_i Y_i \leq f(q) - \epsilon \right) \geq \tau. \]

However, from Lemma 3
\[ P \left( \sum_{i=1}^{n} X_i Y_i \leq f(q) - \epsilon \right) = p_i P \left( \sum_{i=1}^{n} X_i \leq f(q) - \epsilon \right) \leq p_i P \left( \sum_{i=1}^{n} X_i \leq h(q_i) - \epsilon \right) < \sum_{i=1}^{n} p_i q_i \leq \tau. \]

Here the inequality in the middle of the second line is strict because of the definition of \( h(q_i) \). So the above leads to contradiction, which means the inequality (13) is true.

On the other hand, we prove by construction that there exists \( q' = (q'_1, \ldots, q'_n) \), s.t.
\[ Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right) \leq \min_{i \in \{q'_i \neq 1\mid i = 1, 2, \ldots, n\}} h_i(q'_i). \] (14)

Let
\[ C = Q_\tau \left( \sum_{i=1}^{n} X_i Y_i \right). \]

By the definition of quantiles, we know that
\[ P \left( \sum_{i=1}^{n} X_i Y_i \leq C \right) \geq \tau \]
and
\[ P \left( \sum_{i=1}^{n} X_i Y_i \leq C - \epsilon \right) < \tau \]
(15)
for any \( \epsilon > 0 \). First, we take \( \bar{q}_i = P(X_i < C) \). We know that
\[ \sum_{i=1}^{n} p_i \bar{q}_i = \sum_{i=1}^{n} p_i P(X_i < C) = P \left( \sum_{i=1}^{n} Y_i X_i < C \right) \leq \tau. \]
We first prove that the inequality in the end will not hold, i.e.
\[ \sum_{i=1}^{n} p_i \bar{q}_i < \tau. \]
If
\[ \sum_{i=1}^{n} p_i \bar{q}_i = \tau, \]
it follows that
\[ P \left( \sum_{i=1}^{n} Y_i X_i = C \right) = P \left( \sum_{i=1}^{n} Y_i X_i \leq C \right) - P \left( \sum_{i=1}^{n} Y_i X_i < C \right) = 0. \]
Thus, \( P(X_i = C) = 0 \) for all \( i \). Consider the discreteness of random variable \( X_i \)'s, there must exist \( C_0 \leq C \), s.t.
\[ P[X_i \in (C_0, C)] = 0 \]
for all \( i \). As a result,
\[ P \left( \sum_{i=1}^{n} Y_i X_i \leq C_0 \right) = P \left( \sum_{i=1}^{n} Y_i X_i \leq C \right) - P \left( \sum_{i=1}^{n} Y_i X_i \in (C_0, C) \right) - P \left( \sum_{i=1}^{n} Y_i X_i = C \right) \]
\[ = P \left( \sum_{i=1}^{n} Y_i X_i \leq C \right) = \tau. \]
But this contradicts (15). So we have
\[ \sum_{i=1}^{n} p_i \bar{q}_i < \tau. \]
With this strict inequality, let
\[ \tau_0 = \sum_{i=1}^{n} p_i \bar{q}_i, \]
\[ \epsilon = \frac{\tau - \tau_0}{n}, \]
and
\[ q'_i = \min\{\bar{q}_i + \epsilon, 1\}. \]
From Lemma 4, we know that
\[ h_i(q'_i) = Q_{q'_i}(X_i) \geq C, \]
for \( q'_i \neq 1 \). Therefore, (14) follows.
A.2. Proof of Theorem 1, 2 and 3

A.2.1. Proof of Theorem 1

From the definition of $\tilde{v}_t(s, \tau, a)$, we know that

$$
\tilde{v}_t(s, \tau, a) = \max_{\{s_{t+1} \in \Pi_{\mu_t}(s) = a\}} v_t^{s_{t+1}, \tau}(s, \tau)
$$

$$
= \max_{\{s_{t+1} \in \Pi_{\mu_t}(s) = a\}} Q_{\tau} \left( \sum_{s' \in S} 1\{s_{t+1} = s'| s_t = s, a_t = a\} \left[ \sum_{k=t}^{T-1} r_k(S_k, a_k, w_k)\right|_{S_t = s, S_{t+1} = s'} \right)
$$

Without loss of generality, we assume that our policy $\pi \in \Pi_t \triangleq \{\pi_t, s \in \Pi| \mu_t(s) = a\}$ is a function of $(S_t = s, \tau, S_{t+1})$. As we will see shortly, the optimal policy will be indeed determined by $(S_t = s, \tau, S_{t+1})$. Since $s$ and $\tau$ are fixed in our discussion here, we denote the policy as $\pi_{\tau'}$ to reflect its dependence on $S_{t+1} = s'$. Consequently, the cumulative loss

$$
\left[ \sum_{k=t}^{T-1} r_k(S_k, a_k, w_k)\right|_{S_t = s, S_{t+1} = s'} = r_t(S_t, a_t, w_t) + \left[ \sum_{k=t+1}^{T-1} r_k(S_k, a_k, w_k)\right|_{S_t = s, S_{t+1} = s'}
$$

Here the first term $r_t(S_t, a_t, w_t)$ is deterministic with the knowledge of $S_t$, $a$ and $S_{t+1}$ while the second term is a random variable dependent on the state $s'$ and the corresponding policy $\pi_{\tau'}$. Let

$$
X_t^{s_{t+1}} \triangleq r_t(S_t, a_t, w_t) + \left[ \sum_{k=t+1}^{T-1} r_k(S_k, a_k, w_k)\right|_{S_t = s, S_{t+1} = s'}\right].
$$

The subscripts indicate that the random variable is dependent on the state of $S_{t+1}$ and the policy thereafter.

Let state space $S = \{s_1, ..., s_n\}$. The definition of the value function $v_{t+1}(s', \tau')$ tells us that

$$
Q_{\tau'}(X_t^{s_{t+1}}) \leq r_t(S_t, a_t, h_t(S_t, a_t, s_t)) + v_{t+1}(s_t, \tau')
$$

for any $i = 1, ..., n$ and $\tau' \in [0, 1]$. Consider the fact the right hand side is a non-decreasing function with respect to $\tau'$, so we can introduce a random variable $Z_i$ s.t.

$$
Q_{\tau'}(Z_i) = r_t(S_t, a_t, h_t(S_t, a_t, s_t)) + v_{t+1}(s_t, \tau')
$$

for any $\tau' \in [0, 1]$. Let $Y_t = 1\{s_{t+1} = s_t | S_t = s, a_t = a\}$. Then, we have

$$
\tilde{v}_t(s, \tau, a) = \max_{\{s_{t+1} \in \Pi_{\mu_t}(s) = a\}} Q_{\tau} \left( \sum_{i=1}^{n} Y_i X_t^{s_{i+1}} \right)
$$

and from Lemma 2 we know

$$
OPT(s, \tau, v_{t+1}(\cdot, \cdot)) = Q_{\tau} \left( \sum_{i=1}^{n} Y_i Z_i \right).
$$

The relation between $X_t^{s_{i+1}}$ and $Z_i$ is that for any $\tau' \in [0, 1]$

$$
Q_{\tau'}(X_t^{s_{i+1}}) \leq Q_{\tau'}(Z_i),
$$

and for each $i$ and every $\tau'$ there exists a policy $\pi_{s_{i+1}, \tau'}$ that makes the equality holds. First we show that

$$
\tilde{v}_t(s, \tau, a) \leq OPT(s, \tau, v_{t+1}(\cdot, \cdot)).
$$
This only requires for any \( (\pi_1, \ldots, \pi_n) \),

\[
Q_\pi \left( \sum_{i=1}^{n} Y_i X_{s_i}^{\pi_i} \right) \leq \text{OPT} (s, \tau, a, v_{t+1}(\cdot, \cdot)) = Q_\tau \left( \sum_{i=1}^{n} Y_i Z_i \right).
\]

And this is obviously true because the relation between \( X_{s_i}^{\pi_i} \) and \( Z_i \). Then what left is to show that there exists \((\tilde{\pi}_1, ..., \tilde{\pi}_n)\) such that

\[
Q_\tau \left( \sum_{i=1}^{n} Y_i X_{s_i}^{\tilde{\pi}_i} \right) \geq Q_\tau \left( \sum_{i=1}^{n} Y_i Z_i \right).
\]

Let \( q^* = (q_1^*, ..., q_n^*) \) be the optimal solution for the optimization problem \( \text{OPT} (s, \tau, a, v_{t+1}(\cdot, \cdot)) \), and we take \( \tilde{\pi}_i = \pi_i^* q_i^* \) such that

\[
Q_{q_i^*} (X_{s_i}^{\pi_i^* q_i^*}) = Q_{q_i^*} (Z_i).
\]

Therefore,

\[
Q_\tau \left( \sum_{i=1}^{n} Y_i X_{s_i}^{\tilde{\pi}_i} \right) \geq \min_{\{q_i^* \neq 1\} \ldots n} Q_{q_i^*} (X_{s_i}^{\pi_i^* q_i^*}) = \min_{\{q_i^* \neq 1\} \ldots n} Q_{q_i^*} (Z_i) = Q_\tau \left( \sum_{i=1}^{n} Y_i Z_i \right)
\]

Here the first inequality is from Lemma 2 and the last one is from the optimality of \( q \). Thus, we have

\[
\bar{v}_t (s, \tau, a) \geq \text{OPT} (s, \tau, a, v_{t+1}(\cdot, \cdot)).
\]

By combining the above two aspects, we finish the proof.

A.2.2. Proof of Theorem 2

This theorem is easily implied from Theorem 1 with the fact that

\[
v_t (s, \tau) = \max_a \bar{v}_t (s, \tau, a) = \max_a \text{OPT} (s, \tau, a, v_{t+1}(\cdot, \cdot)).
\]

A.2.3. Proof of Theorem 3

We prove by backward induction that the defined policy will can achieve the value function \( v_k (S_k, \tau_k) \). When \( k = T \), the result is trivial. Assume the result is true for \( k = t + 1 \), i.e. for any state \( s' \) and quantile \( \tau' \), there is a policy \( \pi_{(t+1), T}^{s', \tau'} \) such that under this policy, we can achieve a \( \tau' \)-quantile reward \( v_{t+1} (s', \tau') \). Then for \( k = t \), if we want to maximize the \( \tau \)-quantile reward with the state \( S_t = s \), then by solving the optimization problem \( \text{OPT} (s, a, \tau, v_{t+1}(\cdot, \cdot)) \) where \( a = \arg \max v_t (s, \tau) \), we obtain the optimizer \( q^* = (q_1^*, ..., q_n^*) \). From the last part of the proof for Theorem 1 we know that by take the policy \( \pi_{(t+1), T}^{s', q^*} \), we can make to the \( \tau \)-quantile reward \( v_t (s, \tau) \). Thus, we finish the proof for the optimal policy by this induction argument.

A.3. Proofs of Theorem 4 and 5

A.3.1. Proof of Theorem 4

To show the convergence of the value iteration algorithm for QMDP, we need to prove the convergence to the optimal as in the traditional MDP case (Section 1.2 & 1.6, Volume II. (Bertsekas 1995)). The idea is to interpret the \( \text{OPT} \) optimization procedure as a contraction mapping and to verify that the optimal value function is the fixed point of the mapping. For a value function \( v : S \times [0, 1] \rightarrow R \), let \( L \) be the operator

\[
L v = \max_a \text{OPT} (s, \tau, a, \gamma v (\cdot, \cdot)).
\]
Here $Lv$ is a function of $(s, \tau)$ which are the input for $OPT$ on the right hand side. Then we have

$$v^{(k+1)} = Lv^{(k)}.$$  

First, we show that the optimal value function $v(s, \tau)$ is the fixed-point of the operator $L$, namely $v = Lv$. By regarding

$$\sum_{t=1}^{\infty} \gamma^t r_t(s_t, a_t, w_t)$$

as

$$\sum_{t=k+1}^{T-1} \gamma^t r_t(s_t, a_t, w_t)$$

in the proof of Theorem 2, we can utilize the same argument to prove $v(\cdot, \cdot)$ as the fixed point. Then what is left is to show that the operator $L$ is a contraction mapping. We consider the $\infty$-norm as the norm in the space of functions, $\|f\|_\infty = \sup_x |f(x)|$. Then we need to show, for any two value functions $v_1(\cdot, \cdot)$ and $v_2(\cdot, \cdot)$,

$$\|Lv_1 - Lv_2\|_\infty \leq \gamma \|v_1 - v_2\|_\infty,$$

where the contraction rate is the discounting factor $\gamma$. Indeed, it is sufficient to show

$$|OPT(s, \tau, a, \gamma v_1(\cdot, \cdot)) - OPT(s, \tau, a, \gamma v_2(\cdot, \cdot))| \leq \gamma \|v_1 - v_2\|_\infty,$$  \hspace{1cm} (16)$$

for any $(s, \tau)$. Without loss of generality, we assume $OPT(s, \tau, a, \gamma v_1(\cdot, \cdot)) \geq OPT(s, \tau, a, \gamma v_2(\cdot, \cdot))$. Let $q^*$ be the optimizer of first $OPT$, i.e.,

$$q^* = \arg \max_q \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_1(s_i, q_i) + r(s_i, a, h(s_i, s_i)),$$

Here $q^* = (q_1^*, \ldots, q_n^*)$. Moreover, the second part the reward is $r(\cdot)$ instead of $r_t(\cdot)$ because in the infinite-horizon case the reward is stationary. And also, the reward part is the same for the two $OPT$s in (16). Then,

$$OPT(s, \tau, a, \gamma v_1(\cdot, \cdot)) = \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_1(s_i, q_i^*) + r(s_i, a, h(s_i, s_i)),$$

and by definition,

$$OPT(s, \tau, a, \gamma v_2(\cdot, \cdot)) \geq \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_2(s_i, q_i^*) + r(s_i, a, h(s_i, s_i)).$$

By combining these two results, we have

$$OPT(s, \tau, a, \gamma v_1(\cdot, \cdot)) - OPT(s, \tau, a, \gamma v_2(\cdot, \cdot)) \leq \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_1(s_i, q_i^*) + r(s_i, a, h(s_i, s_i))$$

$$- \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_2(s_i, q_i^*) + r(s_i, a, h(s_i, s_i))$$

$$= \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_1(s_i, q_i^*) - \min_{i \in \{q_i \neq 1\}|i=1,2,\ldots,n} \gamma v_2(s_i, q_i^*)$$

$$\leq \gamma \|v_1 - v_2\|_\infty.$$

Consider the arbitrariness of $(s, \tau)$, we have

$$\|Lv_1 - Lv_2\|_\infty \leq \gamma \|v_1 - v_2\|_\infty.$$
A.3.2. Proof of Theorem 5

The proof of the optimal policy is exactly the same as the case in finite-horizon case. The derivation of the value function implies the existence of the optimal policy to achieve the respect optimal value. By assuming this existence for time stamp \( t = 1 \), the optimal policy at \( t = 0 \) will be as stated in the theorem.

Appendix B: ART Optimal Initiation Time Model Parameter

| Variable                                         | Base Value | Source         |
|-------------------------------------------------|------------|----------------|
| General Population Death Rate                   | Varied by Age | WHO (2016)    |
| Mortality Multiplier for Cardiovascular Disease | 2          | WHO (2016)    |
| CD4 decrease every 6 months without ART         | 35.25      | J.W. et al. (1997) |
| CD4 Increase every 6 months on ART by treatment duration \( d_t \) | 100 (\( d_t \leq 6 \)) | Egger et al. (2002) |
|                                                  | 50 (6 < \( d_t \leq 12 \)) |                |
|                                                  | 40 (12 < \( d_t \leq 18 \)) |                |
|                                                  | 40 (18 < \( d_t \leq 24 \)) |                |
|                                                  | 25 (24 < \( d_t \leq 30 \)) |                |
|                                                  | 20 (30 < \( d_t \leq 36 \)) |                |
|                                                  | 20 (36 < \( d_t \leq 42 \)) |                |
|                                                  | 0 (\( d_t > 42 \)) |                |
| 6 month HIV death probability by CD4 levels \( c_k \) without ART | 0.1618 (\( c_k \leq 50 \)) | Egger et al. (2002) |
|                                                  | 0.0692 (50 < \( c_k \leq 100 \)) |                |
|                                                  | 0.0549 (100 < \( c_k \leq 200 \)) |                |
|                                                  | 0.0428 (200 < \( c_k \leq 300 \)) |                |
|                                                  | 0.0348 (300 < \( c_k \leq 400 \)) |                |
|                                                  | 0.0295 (400 < \( c_k \leq 500 \)) |                |
|                                                  | 0.0186 (\( c_k > 500 \)) |                |
| 6 month HIV death probability by CD4 levels \( c_k \) on ART | 0.1356 (\( c_k \leq 50 \)) | Egger et al. (2002) |
|                                                  | 0.0472 (50 < \( c_k \leq 100 \)) |                |
|                                                  | 0.0201 (100 < \( c_k \leq 200 \)) |                |
|                                                  | 0.0103 (200 < \( c_k \leq 300 \)) |                |
|                                                  | 0.0076 (300 < \( c_k \leq 400 \)) |                |
|                                                  | 0.0076 (400 < \( c_k \leq 500 \)) |                |
|                                                  | 0.0045 (\( c_k > 500 \)) |                |
| Utility for HIV infected patient without ART by CD4 levels \( c_k \) | 0.82 (\( c_k \leq 50 \)) | Negoescu et al. (2012) |
|                                                  | 0.83 (50 < \( c_k \leq 100 \)) |                |
|                                                  | 0.84 (100 < \( c_k \leq 200 \)) |                |
|                                                  | 0.85 (200 < \( c_k \leq 300 \)) |                |
|                                                  | 0.86 (300 < \( c_k \leq 400 \)) |                |
|                                                  | 0.87 (400 < \( c_k \leq 500 \)) |                |
|                                                  | 0.88 (\( c_k > 500 \)) |                |
| Utility for HIV infected patient on ART by CD4 levels \( c_k \) | 0.72 (\( c_k \leq 50 \)) | Negoescu et al. (2012) |
|                                                  | 0.75 (50 < \( c_k \leq 100 \)) |                |
|                                                  | 0.78 (100 < \( c_k \leq 200 \)) |                |
|                                                  | 0.81 (200 < \( c_k \leq 300 \)) |                |
|                                                  | 0.84 (300 < \( c_k \leq 400 \)) |                |
|                                                  | 0.87 (400 < \( c_k \leq 500 \)) |                |
|                                                  | 0.90 (\( c_k > 500 \)) |                |
References
Aït-Sahalia, Y., L.P. Hansen. 2009. *Handbook of Financial Econometrics: Tools and Techniques*, vol. 1. Elsevier.

Altman, E. 1999. *Constrained Markov Decision Processes*. CRC Press.

Bertsekas, D.P. 1995. *Dynamic Programming and Optimal Control*, vol. 1. Athena Scientific, Belmont, MA.

Carpín, S., Y. L. Chow, M. Pavone. 2016. Risk aversion in finite Markov decision processes using total cost criteria and average value at risk. *2016 IEEE Int. Conf. Robot. Autom.*. 335–342. doi:10.1109/ICRA.2016.7487152.

Chow, Y., M. Ghavamzadeh. 2014. Algorithms for CVaR optimization in MDPs. *Adv. Neural Inf. Process Syst.*. 3509–3517.

DeCandia, G., D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels. 2007. Dynamo: Amazon’s highly available key-value store. *SIGOPS Oper. Syst. Rev.* 41(6) 205–220. doi:10.1145/1323293.1294281. URL http://doi.acm.org/10.1145/1323293.1294281.

Delage, E., S. Mannor. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Oper. Res.* 58(1) 203–213.

Egger, M., M. May, G. Chne, A.N. Phillips, B. Ledegerber, F. Dabis, D. Costagliola, A. D’Arminio Monforte, F. de Wolf, P. Reiss, J.D. Lundgren, A.C. Justice, S. Staszewski, C. Leport, R.S. Hogg, C.A. Sabin, M.J. Gill, B. Salzberger, J.A.C. Sterne. 2002. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *Lancet* 360(9327) 119–129. doi:https://doi.org/10.1016/S0140-6736(02)09411-4. URL http://www.sciencedirect.com/science/article/pii/S0140673602094114.

Ermon, S., C. Gomes, B. Selman, A. Vladimirsky. 2012. Probabilistic planning with non-linear utility functions and worst-case guarantees. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 965–972.

Filar, J.A., D. Krass, K.W. Ross. 1995. Percentile performance criteria for limiting average Markov decision processes. *IEEE Trans. Autom. Control* 40(1) 2–10.

Fraenkel, L, Jr S T Bogardus, D R Wittink. 2003. Risk-attitude and patient treatment preferences. *Lupus* 12(5) 370–376. doi:10.1191/0961203303lu371oa. URL https://doi.org/10.1191/0961203303lu371oa PMID: 12765300.

Gilbert, H., P. Weng, Y. Xu. 2016. Optimizing quantiles in preference-based Markov decision processes. *ArXiv e-prints*.

Howard, R.A., J.E. Matheson. 1972. Risk-sensitive Markov decision processes. *Manage. Sci.* 18(7) 356–369.

Jorion, P. 1997. *Value at Risk*. McGraw-Hill, New York.
J.W., Mellors, Munoz A., Giorgi J.V., et al. 1997. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann. Intern. Med.* **126**(12) 946–954. doi:10.7326/0003-4819-126-12-199706150-00003. URL [http://dx.doi.org/10.7326/0003-4819-126-12-199706150-00003](http://dx.doi.org/10.7326/0003-4819-126-12-199706150-00003).

N. Bauerle and J. Ott. 2011. Markov decision processes with average-value-at-risk criteria. *Math. Methods Oper. Res.* **74**(3) 361–379. doi:10.1007/s00186-011-0367-0. URL [http://dx.doi.org/10.1007/s00186-011-0367-0](http://dx.doi.org/10.1007/s00186-011-0367-0).

Negoescu, D.M., D.K. Owens, M.L. Brandeau, E. Bendavid. 2012. Balancing immunological benefits and cardiovascular risks of antiretroviral therapy: when is immediate treatment optimal? *Clin. Infect. Dis.* **55**(10) 1392–1399. doi:10.1093/cid/cis731. URL [http://dx.doi.org/10.1093/cid/cis731](http://dx.doi.org/10.1093/cid/cis731).

Rockafellar, R.T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *J. Risk* **2** 21–42.

Tanser, F, T. Bärnighausen, E. Grapsa, J. Zaidi, M.L. Newell. 2013. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science* **339**(6122) 966–971.

Ummels, M., C. Baier. 2013. Computing quantiles in Markov reward models. *International Conference on Foundations of Software Science and Computational Structures*. Springer, 353–368.

WHO, World Health Organization. 2016. WHO Life Table. URL [http://http://www.who.int/gho/mortality_burden_disease/life_tables/life_tables/en/](http://http://www.who.int/gho/mortality_burden_disease/life_tables/life_tables/en/).

Wiesemann, W., D. Kuhn, B. Rustem. 2013. Robust Markov decision processes. *Math. of O.R.* **38**(1) 153–183.