Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Time-Consistency of Optimization Problems

Takayuki Osogami
IBM Research - Tokyo
1623-14 Shimotsuruma, Yamato-shi
Kanagawa, Japan
osogami@jp.ibm.com

Tetsuro Morimura
IBM Research - Tokyo
1623-14 Shimotsuruma, Yamato-shi
Kanagawa, Japan
tetsuro@jp.ibm.com

Abstract

We study time-consistency of optimization problems, where we say that an optimization problem is time-consistent if its optimal solution, or the optimal policy for choosing actions, does not depend on when the optimization problem is solved. Time-consistency is a minimal requirement on an optimization problem for the decisions made based on its solution to be rational. We show that the return that we can gain by taking "optimal" actions selected by solving a time-inconsistent optimization problem can be surely dominated by that we could gain by taking "suboptimal" actions. We establish sufficient conditions on the objective function and on the constraints for an optimization problem to be time-consistent. We also show when the sufficient conditions are necessary. Our results are relevant in stochastic settings particularly when the objective function is a risk measure other than expectation or when there is a constraint on a risk measure.

Introduction

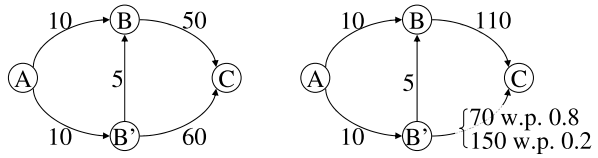
Many approaches for reinforcement learning involve repeating the process of estimating some quantity (e.g., a model or parameters) and choosing an action to take based on the estimated quantity (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998). In repeating this process, the quantity is re-estimated based on the results of the action and additional information obtained since the previous step. In theory, this repeated process as a whole can be seen as a Markov decision process (MDP). This MDP for example captures the transitions that depend on what information becomes available at each time. A decision maker using reinforcement learning at least seeks to find the optimal policy for the MDP and chooses actions based on the optimal policy. Although it is impractical to precisely estimate the transition probabilities for such an MDP, the decision maker often has an objective function and constraints clearly in mind.

In this paper, we study which objective function and constraints the decision maker should use for rational decision making (regardless of the details of the parameters of the MDP, including the definition of the state space as well as the particular values of the transition probabilities and rewards). Our study has implications beyond reinforcement learning. In general, we solve an optimization problem today and determine the actions to take today and tomorrow to maximize our benefit in short and long terms. We might solve an optimization problem tomorrow for the same purpose again, using additional information obtained since today. Are the optimal solutions today "consistent" with the optimal solutions tomorrow? When can we guarantee that they are "consistent"? In this paper, we investigate the new concept, time-consistency of an optimization problem.

Roughly speaking, if an optimization problem is not time-consistent, the optimal actions suggested by its optimal solution today can surely become suboptimal (and sometimes worst) tomorrow simply because the time has passed and a piece of uncertain information is revealed. For instance, tomorrow might be sunny or rainy. Today, the weather being uncertain, the optimal solution to an optimization problem recommends to pack our baggage so we can go picnic tomorrow. Tomorrow, the weather will turn out to be sunny or rainy. If the optimization problem is not time-consistent, the optimal solution tomorrow can recommend not go picnic no matter what the weather is, and the baggage must surely be unpacked. Following the optimal solutions today and tomorrow, respectively, is thus inferior to deciding not to go picnic today, so that we need not unpack the baggage tomorrow. Time-consistency is thus a minimal requirement for optimization problems when they are solved at multiple periods in order for the decisions made based on their optimal solutions to be always rational.

In this paper, we formally define time-consistency of optimization problems under stochastic settings (see Definition 1), which constitutes the first contribution of this paper. Although time-consistency has been studied for decision making under deterministic settings (Strotz 1956), no prior work has formally investigated the time-consistency of optimization problems in stochastic settings. We will discuss the prior work in more detail in the last section of this paper. The second contribution of this paper is a sufficient condition for an optimization problem to be time-consistent. We prove that a certain form of an optimization problem is time-consistent (see Theorem 1). We also discuss the necessity of the sufficient conditions (see Lemma 1 and Corollary 2).

Our results imply that, to guarantee that an optimiza-



(a) Normal (w.p. 0.8) (b) Busy (w.p. 0.2)

Figure 1: An example illustrating a time-inconsistent optimization-problem. (a) Travel time at normal traffic. (b) Travel time at busy traffic.

tion problem is time-consistent, its objective function should be either expectation, entropic risk measure (Foellmer and Schied 2011), or another iterated risk measure (Hardy and Wirch 2004) having the property that we refer to as optimality-consistency. Its constraints should have the property that, if the constraints are satisfied with a policy today, they will also be satisfied tomorrow with the same policy. For example, we can require that the maximum possible value of a random quantity to be below (or minimum to be above) a threshold for an optimization problem to be time-consistent.

Formal proofs are omitted in this paper but given in the associated technical report (Osogami and Morimura 2012). However, the key to the omitted formal proofs are in our definitions, which will be stated formally. We expect that the soundness of the theorems can be understood intuitively once these formal definitions are given.

Time-inconsistent optimization problem

We start by demonstrating the necessity for an optimization problem to be time-consistent. Suppose that we travel from an origin, A, to a destination, C. The travel time depends on whether the traffic is normal (Figure 1(a)) or busy (Figure 1(b)). Upon the departure, we do not know the exact traffic condition but know that the traffic is normal with probability 0.8 and busy otherwise. We also know the conditional probability distribution of the travel time given each traffic condition. For example, given that the traffic is busy, travel time from B' to C is 70 minutes with probability 0.8 and 150 minutes otherwise. Note that the path from B to C is busy if and only if the path from B' to C is busy. From A, we can go either to B or to B'. We assume that the exact traffic condition becomes known when we arrive at B or B'.

To find a strategy that leads us to C as quickly as possible, avoiding a huge delay, we consider an optimization problem of minimizing the expected travel time, X, under the constraint on the risk that its entropic risk measure (Foellmer and Schied 2011), $\mathsf{ERM}_\gamma[X] \equiv \frac{1}{\gamma} \ln \mathsf{E}\left[\exp(\gamma X)\right]$, is below a threshold, δ, where γ represents the sensitivity to riskiness:

$$\begin{array}{ll} \text{min.} & \mathsf{E}\left[X\right] \\ \text{s.t.} & \mathsf{ERM}_\gamma[X] \leq \delta. \end{array} \quad (1)$$

Specifically, let $\gamma = 0.1$ and $\delta = 130$ minutes in (1), where X denotes the total travel time from A to C. Notice that $\mathsf{ERM}_{0.1}[c] = c$ for a constant c, but the value of $\mathsf{ERM}_{0.1}[X]$ is sensitive to the distribution of X. Roughly speaking, we have $\mathsf{ERM}_{0.1}[Y] > \mathsf{ERM}_{0.1}[Z]$ for random variables Y and Z such that $\mathsf{E}[Y] = \mathsf{E}[Z]$ if Y takes a large value with small probability but Z is never large. In some sense, $\delta = 130$ represents the level of acceptable delay.

Strategy	E	$\mathsf{ERM}_{0.1}$
ABC	72.0	104.0
AB'C	75.2	127.8
AB'BC	77.0	109.0
AB'BC if normal; AB'C if busy	71.2	127.8
AB'C if busy; AB'BC if normal	81.0	109.1

(a) Evaluated at A

Strategy	E	$\mathsf{ERM}_{0.1}$
AB'C	70.0	70.0
AB'BC	65.0	65.0

(b) Evaluated at B' (normal)

Strategy	E	$\mathsf{ERM}_{0.1}$
AB'C	96.0	143.9
AB'BC	125.0	125.0

(c) Evaluated at B' (busy)

Table 1: Expectation (E) and entropic risk measure ($\mathsf{ERM}_{0.1}$) of travel time. The values are evaluated upon departure in (a). The values are evaluate upon the arrival at B' given that the traffic is found normal in (b) and busy in (c).

In our example, there are three static strategies, corresponding to three paths: A-B-C, A-B'-C, and A-B'-B-C. With the first static strategy of taking the fixed route from A to B to C, it takes 60 minutes with probability 0.8 (normal) and 120 minutes with probability 0.2 (busy). Hence, the expected travel time is 72 minutes and the entropic risk measure is approximately 104 minutes (see the first row of Table 1(a)). The expectation and the entropic risk measure with the other two strategies are shown in the second and third rows of Table 1(a). Note that the indirect path from A to B' to B surely takes longer than the direct path from A to B. Hence the third static strategy of always taking the route, A-B'-B-C, regardless of the traffic condition, is irrational.

We also consider dynamic strategies, which determine the route depending on the traffic condition. There are two dynamic strategies in our example[1]. The first dynamic strategy is to first visit B' and observe the traffic condition. If the traffic is found normal, we take the detour, B'-B-C; otherwise, we take the direct path, B'-C (the fourth row of Table 1(a)). With this dynamic strategy, the expected travel time is 71.2, which is shorter than any static strategy. The entropic risk measure is below 128, which satisfies the constraint in (1). In fact, this dynamic strategy is the optimal strategy, because the other dynamic strategy results in the expected travel time of 81 minutes (the last row of Table 1(a)).

Solving (1) upon departure, we thus find that the first dynamic strategy, π_0, is optimal and proceed to B'. Now, let us solve (1) again upon arriving at B' to verify that we indeed should follow π_0. Arriving at B', we now obtain the knowledge about the traffic condition.

First, suppose that the traffic is found normal. In this case, the total travel time with the direct path, B'-C, is 70 minutes ($\mathsf{E}[X] = \mathsf{ERM}_{0.1}[X] = 70$), including the 10 min-

[1]The policy of taking A-B-C if traffic is normal and AB'BC otherwise is invalid, because we assume that the exact traffic condition becomes known only after we reach B or B'.

utes that have already taken from A to B (see Table 1(b)). The total travel time with the detour, B'-B-C, is 65 minutes ($\mathsf{E}[X] = \mathsf{ERM}_{0.1}[X] = 65$), because B-C is normal iff B'-C is normal. Hence, taking the detour satisfies the constraint and is optimal, in agreement with π_0.

Next, suppose that the traffic is found busy. In this case, π_0 suggests that we should take the direct path, B'-C. The total travel time with the direct path is 80 minutes with probability 0.8 and 160 minutes with probability 0.2, so that the expected travel time is 96 minutes, and the entropic risk measure is over 143 minutes (see Table 1(c)). This violates the constraint in (1). After finding that the traffic is busy, the route, A-B'-C, is too risky to take in the sense of this constraint. Thus, we would have to take the detour, B'-B-C. With the detour, the total travel time is 125 minutes surely. The constraint is thus satisfied. Taking the detour is the only feasible solution and hence is optimal.

We have seen that, by solving (1) at every intersection and always following the (latest) optimal solution, we end up in taking the route, A-B'-B-C, regardless of the traffic condition. Recall that the total travel time with the route, A-B'-B-C, is surely longer than that with the route, A-B-C. Specifically, the direct path from A to B takes 10 minutes, while the indirect path from A to B' to B takes 15 minutes. This means that solving (1) at multiple epochs can lead to an irrational strategy (e.g., taking A-B'-B-C). With this irrational strategy we incur the cost that is surely larger than the cost that we incur with another strategy (e.g., taking A-B-C). In this sense, we say that (1) is not time-consistent[2]. This motivates us to formally characterize the conditions, not only on the constraints but also on the objective function, for an optimization problem to be time-consistent.

Time-consistency of optimization problems

In this section, we define time-consistency of an optimization problem. Throughout, let $\mathbb{Z}_I$ denote the set of integers in the interval I, and $\mathbb{Z}_{[a,\infty]} \equiv \mathbb{Z}_{[a,\infty)} \cup \{\infty\}$ for $a < \infty$. Suppose that a decision maker tries to maximize worthiness of a random quantity, X, while keeping riskiness associated with X at an acceptable level. The value of X will be determined in future, on or before time N, but is random before that. To achieve the goal, he solves an optimization problem at time 0 to select the policy, π_0, that determines the action to take for each state at each time $n \in \mathbb{Z}_{[0,N)}$. At a future time $\ell \in \mathbb{Z}_{[1,N)}$, he might solve an optimization problem that can depend on the state, S_ℓ, at time ℓ, where S_ℓ can include his belief, conditions of the environment, and other information available to him by time ℓ. We want the policy selected at time ℓ to be consistent with π_0.

We allow N to be either finite or infinite. There are multiple ways to interpret X when $N = \infty$. For example, we can see X as a random quantity, X_τ, that is determined at a stopping time, τ, when S_τ becomes a particular state, where τ is finite almost surely but unbounded.

[2]Time-inconsistency of the optimization problem (1) does not contradict the claim that ERM is a time-consistent risk measure. Time-consistency is defined for a risk measure in the prior work but for a process of optimization problems in this paper.

Process of optimization problem

More formally, for $n \in \mathbb{Z}_{[0,N)}$, let $\mathtt{P}_n(s)$ be the optimization problem that a decision maker solves at time n under the condition that $S_n = s_n \in \mathbf{S}_n$, where $\mathbf{S}_n$ is the set of the possible states at time n. Specifically, we consider the optimization problem, $\mathtt{P}_n(s_n)$, of the following form:

$$\mathtt{P}_n(s_n): \begin{array}{ll} \max_{\pi\in\Pi_n} & f_n(X^\pi(s_n)) \\ \text{s.t.} & g_n(X^\pi(s_n)) \in B_n, \end{array} \tag{2}$$

where $X^\pi(s_n)$ denotes the conditional random variable, X, given that $S_n = s_n$ and π is the policy used on and after time n. The decision maker seeks to find the optimal policy from the candidate set, Π_n, where $|\Pi_n| \geq 1$. A policy, $\pi_n \in \Pi_n$, determines an action, a_ℓ, to take for each state, $s_\ell \in \mathbf{S}_\ell$, at each time, $\ell \in \mathbb{Z}_{[n,N)}$. For two distinct policies, $\pi_n, \pi'_n \in \Pi_n$, we assume that the actions taken with π_n and those with $\pi'_n \neq \pi_n$ differ for at least one state, $s_\ell \in \mathbf{S}_\ell, \ell \in \mathbb{Z}_{[n,N)}$. Throughout, we consider only deterministic policies, where the action to take from a given state is selected non-probabilistically. However, we allow a candidate action to be equivalent to a probabilistic action that randomly selects one among multiple candidates.

In (2), the objective function, f_n, maps $X^\pi(s_n)$ to a real number. For example, setting $f_n(X^\pi(s_n)) = \mathsf{E}[X^\pi(s_n)]$, the decision maker can select the optimal policy from Π_n that maximizes the expected value of the random quantity, X, given that $S_n = s_n$.

The constraint in (2) specifies the acceptable riskiness at time n. Here, g_n is a multidimensional function that maps $X^\pi(s_n)$ to real numbers, and B_n specifies the feasible region in the codomain of g_n. For example, the constraint could be specified with an ERM such that

$$\mathsf{ERM}_\gamma[-X^\pi(s_n)] \quad \leq \quad b_n,$$

where b_n denotes the upper bound of the acceptable risk. Here, the value of ERM represents a magnitude of a loss, so that the negative sign is appended to the reward X.

Observe that the optimization problems that the decision maker is solving can be seen as a stochastic process, $\mathtt{POP}(X, \mathbf{S}, \mathbf{p})$, which we refer to as a Process of Optimization Problems (POP):

$$\mathtt{POP}(X, \mathbf{S}, \mathbf{p}): \begin{array}{ll} \max_{\pi\in\Pi} & f(X^\pi) \\ \text{s.t.} & g(X^\pi) \in B. \end{array} \tag{3}$$

Here, X^π denotes the conditional random variable of X given that policy π is used. For simplicity, we assume that the initial state S_0 is known to be $s_0 \in \mathbf{S}_0$ (i.e., $|\mathbf{S}_0| = 1$), but it is trivial to extend our results to the case with $|\mathbf{S}_0| > 1$. For $n \in \mathbb{Z}_{[0,N)}$, the decision maker solves $\mathtt{P}_n(s)$ at time n under the condition that $S_n = s$. Note that the optimal policy at time n depends on how the state transitions after time n (specifically, the set of transition probabilities,

$$\mathbf{p}^{(n)} \equiv \left\{ p^\pi_\ell(s'|s) \;\middle|\; \begin{array}{l} s \in \mathbf{S}_\ell, s' \in \mathbf{S}_{\ell+1}, \\ \ell \in \mathbb{Z}_{[n,N)}, \pi \in \Pi_n \end{array} \right\} \tag{4}$$

where $p^\pi_\ell(s'|s)$ denotes the probability that $S_{\ell+1} = s'$ under the condition that $S_\ell = s$ and actions follow π). Namely, the decision maker assumes (4) when he solves $\mathtt{P}_n(s_n)$ for

$s_n \in \mathbf{S}_n$. In (3), $\mathbf{S} \equiv \{\mathbf{S}_n \mid n \in \mathbb{Z}_{[0,N)}\}$ represents the state space, and $\mathbf{p} = \mathbf{p}^{(0)}$ represents the sets of transition probabilities that the decision maker uses in solving the optimization problems.

We assume that the state includes all of the information about the history of prior states and prior actions. Under this assumption, the state transition diagram has a tree structure. Then we limit Π to be the set of Markovian policies, with which the action to take at a state, s, depends only on s (conditionally independent of the history of the prior states and prior actions given s). Notice however that, for our purposes, this assumption is not as limiting as it might appear at first sight. Consider a general MDP, where a state can be reached with different histories. It is straightforward to expand the state space of such a general MDP in such a way that each state in the expanded state space includes all of the information about the history. The MDP with the expanded state space is equivalent to the original MDP, so that the original MDP is time-consistent if and only if the MDP with the expanded state space is time-consistent. We assume that the decision maker knows the state at any time, so that he solves $\mathtt{P}_n(s)$ if the state is $s \in \mathbf{S}_n$ at time $n \in \mathbb{Z}_{[0,N)}$. Notice, however, that the state might just represent the belief of the decision maker, and in that case he only knows what he believes and the relevant history.

Time-consistent process of optimization problems

To determine whether the optimization problems that a decision maker is solving lead to contradicting decisions over time, we define time-consistency of a POP. Consider a verifier who determines whether the optimization problem is time-consistent. We primarily consider the case where the verifier knows none of X, $\mathbf{S}$, and $\mathbf{p}$ that the decision maker is using. This case is relevant for example when the decision maker estimates $\mathbf{p}$, but his estimation is unknown to her. If the verifier does not know $\mathbf{p}$, then she does not know the distribution of X, because X depends on $\mathbf{p}$. In fact, there might be multiple decision makers who solve $P(X, \mathbf{S}, \mathbf{p})$ but with different $\mathbf{p}$. Because the verifier knows nothing about $\mathbf{p}$, any assumption about $\mathbf{p}$ should seem reasonable to the verifier. Alternatively, it might be the case that a decision maker chooses the optimal policy for the MDP, where he estimates $\mathbf{p}$. Depending on how $\mathbf{p}$ is estimated, the decision maker solves the MDP having varying $\mathbf{p}$. The verifier wants to know whether the MDP is time-consistent before the decision maker estimates $\mathbf{p}$. Also, the verifier might not know $\mathbf{S}$ that the decision maker defines by analogous reasons.

At time 0, the decision maker finds the optimal policy, $\pi_0^\star$, for the MDP that starts from s_0. At time 0, $\pi_0^\star$ is most appealing to him, because the constraint, $g_0(X^{\pi_0^\star}) \in B_0$, is satisfied, and the value of the objective function cannot be made greater than $f_0(X^{\pi_0^\star})$ by any feasible policy $\pi \in \Pi_0$. Notice that $\pi_0^\star$ can be used to determine the action that he should take for any $s_n \in \mathbf{S}_n$ at any time $n \in \mathbb{Z}_{[0,N)}$. However, he might solve an optimization problem at a future time to find the optimal policy at that time. Our expectation is that $\pi_0^\star$ continues to be one of the most appealing policies to him at any time $n \in \mathbb{Z}_{[1,N)}$, if the associated POP is time-consistent. This leads us to formally define time-consistency of the POP as follows:

Definition 1 *We say that* $\mathtt{POP}(X, \mathbf{S}, \mathbf{p})$ *is time-consistent if the following property is satisfied. For any* $n \in \mathbb{Z}_{[1,N)}$*, if* $\pi^\star$ *is optimal from* $s \in \mathbf{S}_{n-1}$ *(i.e.,* $\pi^\star$ *is an optimal solution to* $\mathtt{P}_{n-1}(s)$*), then* $\pi^\star$ *is optimal from any* $s' \in \mathbf{S}_n$ *such that* $p_{n-1}^{\pi^\star}(s' \mid s) > 0$*. We say that* $\mathtt{POP}$ *is time-consistent if* $\mathtt{POP}(X, \mathbf{S}, \mathbf{p})$ *is time-consistent for any* X*,* $\mathbf{S}$*, and* $\mathbf{p}$*.*

Observe that Definition 1 matches with our intuition about optimizing a standard MDP, where the optimal policy found at time 0 is optimal at any time $n \in \mathbb{Z}_{[0,N)}$. We now revisit the optimization problem studied with Figure 1. The optimal policy π_0, which we find by solving the optimization problem upon departure, becomes infeasible (hence not optimal) for the optimization problem solved at intersection B if the traffic is busy. The transition probability to the busy state is 0.2, which is strictly positive. Hence, this optimization problem is indeed time-inconsistent in the sense of Definition 1.

Conditions for time-consistency

In this section, we provide conditions that the objective function and the constraints should satisfy so that a POP is time-consistent. To formally state the definition of time-consistency, it is important to understand the objective function, f_n, in (2) as a dynamic risk measure (RM). Let Y be a random variable, and let $\rho_n(Y)$ be the value of the RM of Y evaluated at time n. Note that $\rho_n(Y)$ is random before time n and becomes deterministic at time n, because $\rho_n(Y)$ depends on the state that is random before time n. In this sense, $\rho_n(Y)$ is called $\mathcal{F}_n$-measurable, which can be understood more precisely with measure theory. Formally, a dynamic RM is defined as follows:

Definition 2 *Consider a filtered probability space,* $(\Omega, \mathcal{F}, P)$*, such that* $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_N = \mathcal{F}$*, where* $N \in \mathbb{Z}_{[1,\infty]}$*, and, if* $N = \infty$*,* $\mathcal{F}_\infty$ *is defined as the* σ*-field generated by* $\cup_{\ell=0}^{\infty} \mathcal{F}_\ell$*. Let* Y *be an* $\mathcal{F}$*-measurable random variable. We say that* ρ *is a dynamic RM if* $\rho_\ell(Y)$ *is* $\mathcal{F}_\ell$*-measurable for each* $\ell \in \mathbb{Z}_{[0,N]}$*.*

In our context, X^π is random before time N but becomes deterministic by time N, so that X^π is $\mathcal{F}_N$-measurable. Because S_n is random before time n, the value of the objective function, $f_n(X^\pi(S_n))$, is random before it becomes deterministic at time n (i.e., $\mathcal{F}_n$-measurable). Hence, the objective function is a dynamic RM. Next, let $h(X^\pi) \equiv \mathsf{I}\{g(\cdot) \in B\}$, where $\mathsf{I}\{\cdot\}$ is an indicator random variable. Observe that $h(X^\pi)$ is a dynamic RM, because $g_n(X^\pi)$ and (the random variables that define) B_n are $\mathcal{F}_n$-measurable.

We use the following definitions to provide a sufficient condition for an POP to be time-consistent:

Definition 3 *A dynamic RM,* ρ*, is called optimality-consistent if following condition is satisfied for any* $\mathcal{F}$*-measurable random variables* Y, Z *and for any* $n \in \mathbb{Z}_{[1,N)}$*: if* $\Pr(\rho_n(Y) \leq \rho_n(Z)) = 1$ *and* $\Pr(\rho_n(Y) < \rho_n(Z)) > 0$*, then* $\Pr(\rho_{n-1}(Y) < \rho_{n-1}(Z)) > 0$*. Also, we say that* ρ *is optimality-consistent for a particular* n *if the above condition is satisfied for the* n*.*

Definition 4 *Let Y be an $\mathcal{F}$-measurable random variable. A dynamic RM, ρ, is called non-decreasing if* $\Pr(\rho_{n-1}(Y) \leq \rho_n(Y)) = 1$ *for any* $n \in \mathbb{Z}_{[1,N)}$. *Also, we say that ρ is non-decreasing for a particular n if the above property is satisfied for the n.*

To get a sense of Definition 3, suppose that ρ is not optimality-consistent for an n. Then one can choose Z at time $n-1$ (i.e., $Pr(\rho_{n-1}(Y) \geq \rho_{n-1}(Z)) = 1$) despite the fact that, at time n, Y becomes surely at least as good as Z (i.e., $Pr(\rho_n(Y) \leq \rho_n(Z)) = 1$) and sometimes better than Z (i.e., $Pr(\rho_n(Y) < \rho_n(Z)) > 0$). This will be formalized in the following.

Sufficient condition

We are now ready to state a sufficient condition for a POP to be time-consistent (see the associated technical report (Osogami and Morimura 2012) for a formal proof):

Theorem 1 *If f is an optimality-consistent dynamic RM and $h(\cdot) \equiv \mathsf{I}\{g(\cdot) \in B\}$ is a non-decreasing dynamic RM, then* POP *as defined with (3) is time-consistent.*

We elaborate on the sufficient conditions in the following. First, we remark that expectation and entropic risk measure can be shown to be optimality-consistent. In fact, optimality-consistency can be shown for a class of iterated RMs that have been studied for example in Hardy and Wirch (2004):

Definition 5 *Consider a filtered probability space, $(\Omega, \mathcal{F}, P)$, such that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_N = \mathcal{F}$, where $N \in \mathbb{Z}_{[1,\infty)}$. Let Y be an $\mathcal{F}$-measurable random variable. We say that ρ is an iterated RM if $\rho_N[Y] = Y$ and $\rho_n[Y] = \bar{\rho}_n[\rho_{n+1}[Y]]$ for $n \in \mathbb{Z}_{[0,N)}$, where $\bar{\rho}_n$ is a conditional RM that maps an $\mathcal{F}_{n+1}$-measurable random variable to an $\mathcal{F}_n$-measurable random variable, for $n \in [0, N)$.*

Notice that $\mathsf{E}[\cdot|S_n] = \mathsf{E}[\mathsf{E}[\cdot|S_{n+1}]|S_n]$, so that expectation is an iterated RM, where $\bar{\rho}_n[\cdot] = \mathsf{E}[\cdot|S_n]$. Likewise, entropic risk measure is an iterated RM, where $\bar{\rho}_n[\cdot] = \mathsf{ERM}_\gamma[\cdot|S_n]$. The following corollary can be proved formally:

Corollary 1 *An iterated RM, as defined in Definition 5, is optimality-consistent for a particular n if the following property is satisfied: for any $\mathcal{F}$-measurable random variables, Y and D, such that* $\Pr(D \geq 0) = 1$ *and* $\Pr(D > 0) > 0$, *we have* $\Pr(\bar{\rho}_n[Y + D] > \bar{\rho}_n[Y]) > 0$.

Corollary 1 allows us to check whether a given iterated RM, ρ, is optimality-consistent by studying the properties of $\bar{\rho}_n$ for $n \in \mathbb{Z}_{[0,N)}$. For example, an iterated RM defined with $\bar{\rho}_n(X) \equiv (1-\beta)\,\mathsf{E}[X|S_n] - \beta\,\mathsf{CTE}_\alpha[-X|S_n]$ is optimality-consistent for $\alpha, \beta \in (0,1)$, where $\mathsf{CTE}_\alpha[-X|S_n]$ denotes the conditional tail expectation (also known as conditional value at risk (Artzner et al. 1999; Rockafellar and Uryasev 2000)), having parameter α, of $-X$ given S_n. One can expect that maximizing this iterated RM leads to balancing between maximizing expectation and minimizing riskiness. An easy way to verify the conditions of Corollary 1 is to demonstrate that $\rho_n(X)$ can be expressed as

$$\bar{\rho}_n(X) = \int_{x \in \mathbb{R}} u(x, F_X(x))\, dF_X(x), \tag{5}$$

where $u(\cdot,\cdot)$ is monotonically increasing with respect to its first argument, and F_X is the cumulative distribution function of X. For the D defined in Corollary 1, we have

$$\bar{\rho}_n(Y + D) = \int_0^1 u(F^{-1}_{Y+D}(q), q)\, dq > \int_0^1 u(F^{-1}_Y(q), q)\, dq = \bar{\rho}_n(Y).$$

We remark that an iterated RM, ρ, does not have the property, $\rho_n(\cdot) = \bar{\rho}_n(\cdot)$, unless $\bar{\rho}_n$ is $\mathsf{E}[\cdot|S_n]$ for all $n \in \mathbb{Z}[0,N)$ or $\mathsf{ERM}_\gamma[\cdot|S_n]$ for all $n \in \mathbb{Z}[0,N)$. If the objective function is an iterated RM, then $\rho_n(\cdot)$ is maximized at each time n, where the number of conditional RMs $(\bar{\rho}_n, \ldots, \bar{\rho}_{N-1})$ used to define ρ_n depends on the remaining time, $N - n$. A key implication of Corollary 1 and (5) is that there is a large class of iterated RMs having optimality-consistency.

Simple examples of the constraints that make a POP time-consistent include $\max(X^\pi) \leq \delta$ and $\min(X^\pi) \geq \delta$, where $\max(X^\pi)$ (respectively, $\min(X^\pi)$) denote the maximum (respectively, minimum) value that X^π can take with positive probability. Notice that $\max(X^\pi)$ is non-increasing over time for any sample path, because we obtain more information about (the maximum possible value of) X^π as time passes. Therefore, $\mathsf{I}\{\max(X^\pi) \leq \delta\}$ is non-decreasing. Analogously, $\mathsf{I}\{\min(X^\pi) \geq \delta\}$ is non-decreasing.

We have seen with Figure 1 that (1) is not necessarily time-consistent. We can now see that the time-inconsistency is due to the constraint in (1), because the objective function in (1) is optimality-consistent. Observe that $\mathsf{ERM}_{0.1}[X^{\pi_0}]$ increases from 127.8 upon departure to 143.9 at intersection B' if the traffic is found busy at B'. Hence, $\mathsf{I}\{\mathsf{ERM}_{0.1}[\cdot] \leq 130\}$ is not non-decreasing.

A way to modify (1) into a time-consistent POP is to incorporate the constraints that might need to satisfy in future:

$$\begin{array}{ll} \min. & \mathsf{E}[X] \\ \text{s.t.} & \mathsf{ERM}_\gamma[X \mid S_\ell = s_\ell] \leq \delta, \\ & \forall s_\ell \in \mathbf{S}, \forall \ell \in \mathbb{Z}_{[0,N)}, \end{array} \tag{6}$$

Then π_0 becomes infeasible for the optimization problem to be solved on Day 0, which resolves the issue of the time-inconsistency.

Because ERM is optimality-consistent, the following POP is time-consistent for any parameter α:

$$\begin{array}{ll} \min. & \mathsf{ERM}_\alpha[X] \\ \text{s.t.} & \mathsf{ERM}_\gamma[X \mid S_\ell = s_\ell] \leq \delta, \\ & \forall s_\ell \in \mathbf{S}, \forall \ell \in \mathbb{Z}_{[0,N)}, \end{array} \tag{7}$$

Conditional tail expectation is not optimality-consistent, so that the following POP is not time-consistent:

$$\begin{array}{ll} \min. & \mathsf{CTE}_\alpha[X] \\ \text{s.t.} & \mathsf{ERM}_\gamma[X \mid S_\ell = s_\ell] \leq \delta, \\ & \forall s_\ell \in \mathbf{S}, \forall \ell \in \mathbb{Z}_{[0,N)}. \end{array} \tag{8}$$

Necessary conditions

Next, we study necessity of the sufficient condition provided in Theorem 1. We can show that the sufficient condition on

the constraints is necessary for POP to be time-consistent for any objective function. Also, the sufficient condition on the objective function is necessary for POP to be time-consistent for any constraints (see the associated technical report (Osogami and Morimura 2012) for a formal proof):

Lemma 1 *If* POP *is time-consistent for any objective function f, then $h_n(\cdot) \equiv \mathsf{I}\{g_n(\cdot) \in B_n\}$ must be non-decreasing. If* POP *is time-consistent for any constraints $g(\cdot) \in B$, then the objective function f must be optimality-consistent.*

The above results lead to the following necessary and sufficient condition:

Corollary 2 *Suppose that there exists $\gamma \equiv \min_{n \in \mathbb{Z}_{[0,N)}, s_n \in \mathbf{S}_n} f_n(X^\pi(s_n))$. Then* POP *is time-consistent iff $(f(\cdot) - \gamma)\,\mathsf{I}\{g(\cdot) \in B\}$ is optimality-consistent.*

Notice that our results also applies to the case where the distribution of X, $\mathbf{S}$, and $\mathbf{p}$, which the decision maker is using in solving the optimization problems, are known to a verifier, because $\mathtt{POP}(X, \mathbf{S}, \mathbf{p})$ is time-consistent if POP is time-consistent. The decision maker might want to verify that the POP associated with the problem of finding the optimal policy for the MDP is time-consistent. Because the decision maker is a verifier, the verifier knows the distribution of X, $\mathbf{S}$, and $\mathbf{p}$ that the decision maker is using.

Related work and discussion

Time-consistency was first discussed in deterministic settings regarding how future cost should be discounted. In particular, Strotz (1956) shows that exponential discounting is necessary for time-consistency. Exponential discounting is thus standard for decision making with discounted expected utility (R. E. Lucas 1978). Bewley (1987) essentially shows that an optimization problem is time-consistent if the objective function is a discounted expected cumulative reward. On the other hand, we give more general sufficient conditions and discuss their necessity. Bewley (1987), however, considers the settings that are different from ours. In Bewley (1987), a decision maker is not certain about probability distributions, and his preference cannot be expressed in a total order (i.e., Knightian decision theory). Our results do not cover the Knightian decision theory, and it is an interesting direction to extend our results to the Knightian settings.

The necessity of time-consistency for rational decision making is illuminated by the following quote from Ainslie (2001) (p.30-31): "if a hyperbolic discounter engaged in trade with someone who used an exponential curve, she'd soon be relieved of her money. Ms. Exponential could buy Ms. Hyperbolic's winter coat cheaply every spring, for instance, because the distance to the next winter would depress Ms. H's evaluation of it more than Ms. E's. Ms. E could then sell the coat back to Ms. H every fall when the approach of winter sent Ms. H's valuation of it into a high spike." That is, a time-consistent decision maker (Ms. E) can squeeze an infinite amount of money out of a time-inconsistent decision maker (Ms. H). Analogous arguments apply to time-consistency in the stochastic settings studied in this paper.

Time-consistency of the RM has also been discussed in the literature (Artzner et al. 2007; Hardy and Wirch 2004; Riedel 2004; Boda and Filar 2005; Foellmer and Schied 2011), where an RM is used for a bank to determine the amount of the capital to be reserved. Time-consistency is widely considered to be a necessary property of such an RM. We remark that E and ERM_γ are both time-consistent RMs. In this paper, however, we have seen undesirable outcomes when a decision maker seeks to minimize expected loss (or equivalently to maximize expected profit) when there is a regulation that requires that the decision maker keep the ERM of the loss below a threshold. Namely, we find that an optimization problem is not necessarily time-consistent (as formally defined in Definition 1) even if the functions that constitute its objective and constraints are time-consistent (as defined in Artzner et al. (2007)).

Although there has not been unified discussion about the time-consistency of optimization problems in stochastic settings, time-inconsistency has been reported for several models of MDPs. In the prior work, when optimization of an MDP is time-inconsistent, it has been stated in various ways, including "the optimal policy changes over time," "the optimal policy is nonstationary," or "the principle of optimality is not satisfied," which will be detailed in the following.

A constrained MDP requires to minimize the expect cost of one type, while keeping the expected cost of another type below a threshold (Altman 1999). It has been pointed out that Bellman's principle of optimality is not necessarily satisfied by an optimal policy of a constrained MDP, i.e., the constrained MDP is not necessarily time-consistent (Haviv 1996; Henig 1984; Ross and Varadarajan 1989; Sennott 1993), where counter-examples have been constructed for the case where the constrained MDP is a multi-chain over an infinite horizon. For particular constrained MDPs, however, it has been shown that Bellman's principle of optimality holds, i.e., they are time-consistent (Abe et al. 2010; Haviv 1996; Ross and Varadarajan 1989). Our results shed light on the time-consistency and time-inconsistency of these constraint MDPs: one can easily verify that the constraints in the general constrained MDP do not necessarily have the non-decreasing property, while they do in Abe et al. (2010); Haviv (1996); Ross and Varadarajan (1989). For example, in Haviv (1996); Ross and Varadarajan (1989), the constraint must be satisfied for every sample path, which directly implies the non-decreasing property.

There also exists a large body of the literature that studies the MDPs whose objective functions are not optimality-consistent. For example, variance is not optimality-consistent, so that the problem of minimizing the variance of an MDP, studied for example in Kawai (1987), is not time-consistent. White surveys MDPs, where "principle of optimality fails" or "no stationary optimal policy exists" (White 1988). One should be warned that making decisions based on these MDPs lead to contradicting decisions over time, which in turn can result in being deprived of an infinite amount of wealth by a time-consistent decision maker. We have shown that there is a large class of iterated RMs that are optimality-consistent, which can be used to formulate time-consistent optimization problems that can be used by

decision makers who are risk-sensitive and rational.

Observe that time-consistency is related to dynamic programming. Namely, if a POP is time-consistent, the corresponding optimal policy can be found with dynamic programming. Because the optimal policy today is also optimal tomorrow if the POP is time-consistent, we can find the optimal policy today by first finding the optimal policy for each possible state of tomorrow and then finding the optimal action today based on the optimal policies for tomorrow. Therefore, a time-consistent POP is attractive from computational point of view, besides it leads to rational decision making.

The objective of the standard risk-sensitive MDPs is expected exponential utility (or equivalently entropic risk measure), so that these MDPs are time-consistent. Ruszczyński (2010) studies dynamic programming for an MDP whose objective is a Markov RM, a particular iterated RM. Osogami (2011) studies dynamic programming for an MDP whose objective is a particular iterated RM when the future cost is discounted. However, these iterated RMs require to satisfy conditions that are not needed for optimality-consistency and no constraints are considered in Ruszczyński (2010) and Osogami (2011).

Acknowledgments This work was supported by "Promotion program for Reducing global Environmental loaD through ICT innovation (PREDICT)" of the Ministry of Internal Affairs and Communications, Japan.

References

Abe, N.; Melville, P.; Pendus, C.; Reddy, C.; Jensen, D.; Thomas, V. P.; Bennett, J. J.; Anderson, G. F.; Cooley, B. R.; Kowalczyk, M.; Domick, M.; and Gardinier, T. 2010. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the ACM KDD 2010*, 75–84.

Ainslie, G. 2001. *Breakdown of Will*. Cambridge, UK: Cambridge University Press, first edition.

Altman, E. 1999. *Constrained Markov Decision Processes*. Chapman and Hall/CRC.

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

Artzner, P.; Delbaen, F.; Eber, J.-M.; Heath, D.; and Ku, H. 2007. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research* 152:5–22.

Bertsekas, D. P., and Tsitsiklis, J. N. 1996. *Neuro-Dynamic Programming*. Athena Scientific.

Bewley, T. F. 1987. Knightian decision theory, Part II: Intertemporal problems. Cowles Foundation Discussion Paper No. 845, Yale University.

Boda, K., and Filar, J. A. 2005. Time consistent dynamic risk measures. *Mathematics of Operations Research* 63(1):169–186.

Foellmer, H., and Schied, A. 2011. *Stochastic Finance: An Introduction in Discrete Time*. Berlin, Germany: De Gruyter, third revised and extended edition.

Hardy, M. R., and Wirch, J. L. 2004. The iterated CTE: A dynamic risk measure. *North American Actuarial Journal* 8:62–75.

Haviv, M. 1996. On constrained Markov decision processes. *Operations Research Letters* 19:25–28.

Henig, M. 1984. Optimal paths in graphs with stochastic or multidimensional weights. *Communications of the ACM* 26:670–676.

Kawai, H. 1987. A variance minimization problem for a Markov decision process. *European Journal of Operational Research* 31:140–145.

Osogami, T., and Morimura, T. 2012. Time-consistency of optimization problems (2nd edition). Technical Report RT0942, IBM Research - Tokyo.

Osogami, T. 2011. Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In *Proceedings of the UAI 2011*, 567–574.

R. E. Lucas, J. 1978. Asset prices in an exchange economy. *Econometrica* 46:1429–1445.

Riedel, F. 2004. Dynamic coherent risk measures. *Stochastic Processes and their Applications* 112:185–200.

Rockafellar, R. T., and Uryasev, S. 2000. Optimization of conditional value-at-risk. *The Journal of Risk* 2:21–41.

Ross, K. W., and Varadarajan, R. 1989. Markov decision processes with sample-path constraints: The communicating case. *Operations Research* 37:780–790.

Ruszczyński, A. 2010. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming* 125:235–261.

Sennott, L. I. 1993. Constrained average cost Markov decision chains. *Probability in the Engineering and Informational Sciences* 7:69–83.

Strotz, R. H. 1956. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies* 23:165–180.

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

White, D. J. 1988. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications* 56:1–29.