September 15, 2015
RT0966
Mathematics 35 pages

Research Report

When the optimal policy is independent of the initial state

Takayuki Osogami and Tetsuro Morimura

IBM Research - Tokyo
IBM Japan, Ltd.
NBF Toyosu Canal Front Building
6-52, Toyosu 5-chome, Koto-ku
Tokyo 135-8511, Japan

When the optimal policy is independent of the initial state

Takayuki Osogami[a,*], Tetsuro Morimura[a]

[a] *IBM Research - Tokyo*

Abstract

A Markov decision process (MDP) is a popular model of sequential decision making, but its standard objective of minimizing cumulative cost is often inadequate, for example, to avoid the possibility of large loss. Risk-sensitive objective functions and constraints have thus been proposed for MDPs. Unlike the standard MDP, however, the optimal policy for some of these MDPs can depend on the initial states, so that the optimal policy can change over time. We show that an agent can surely incur larger cumulative cost by following the latest optimal policy at every state than by following other policies. We then establish sufficient conditions on the objective function and on the constraints for the optimal policies to be consistent between the initial states. We also show when the sufficient conditions are necessary. We discuss implications of our results to the MDPs that have been studied in the literature, stating whether their optimal policies depend on the initial states.

Keywords: Dynamic programming, Markov decision processes, iterated risk measures, risk-sensitive, constraints, time-consistency

2010 MSC: 90C40

*Corresponding author

Email addresses: osogami@jp.ibm.com (Takayuki Osogami), tetsuro@jp.ibm.com (Tetsuro Morimura)

1. Introduction

A Markov decision process (MDP) is a model of sequential decision making, where the goal is to find a policy, which maps a state to an action, such that a given objective function is minimized. When the objective function is the standard expected cumulative cost, the optimal policy is known to be independent of the initial state, or where an agent starts acting. It thus suffices to find the optimal policy for an arbitrary initial state and let any agents act according to that optimal policy.

For some of the non-standard objective functions or constraints, however, it is known that the optimal policy depends on the initial state [19, 18, 25, 27, 28, 42, 65]. Recently, there has been an increasing interest in the use of non-standard objective functions [27, 33, 37, 39, 40, 41, 42, 47, 48, 49, 51, 65] or constraints [1, 6, 9, 10, 19, 18, 25, 26, 27, 28, 30, 42, 65, 66] in MDPs. For example, one of the objectives in [42] is to minimize the variance of cumulative cost under the constraint that expected cumulative cost is below a threshold. These non-standard objective functions or constraints have been introduced, because expected cumulative cost is often inadequate for example to avoid large loss or to take into account the limitations of available resources.

The dependency of the optimal policy on the initial state is rather controversial. Consider two agents who make decisions based on a common MDP. The first agent finds the optimal policy from an initial state, takes the first action, and transitions to the next state. This next state is the initial state for the second agent who starts acting when the first agent takes the second action. The state is Markovian, and the two agents are indistinguishable when the first agent takes the second action (and the second agent takes the first action). However, these two agents behave differently when they act according to respectively optimal policies.

In this paper, we study which objective function and constraints can guarantee that the optimal policy is independent of the initial state. Our results imply that the objective function should be represented by either expectation, entropic risk measure [23], or another dynamic risk measure [48] having the property that we refer to as optimality-consistency. The constraints should have the property that, if the constraints are satisfied with a policy at one moment, they will also be satisfied in the future 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. These conditions constitute our primary contributions.

The optimality-consistency of a dynamic risk measure is related to but different from time-consistency that has been studied in [5, 11, 23, 31, 53, 54, 57, 59]. Time-consistency requires that "[a]t every state of the system, optimality of our decisions should not depend on scenarios which we already know cannot happen in the future" (page 321 from [59]). This notion of time-consistency is important primarily because it guarantees that the optimal policy is the one that satisfies the Bellman equation (Equations 3.3 from [44]). However, time-consistency does not necessarily preclude the dependency of the optimal policy on the initial state. See Appendix A for further discussion on the difference between optimality-consistency and time-consistency.

The rest of the paper is organized as follows. In Section 2, we give examples of objective functions and constraints that cause the optimal policy to depend on the initial state. We will see undesirable outcomes when the agent at a state changes the policy to the one that is optimal from that state. In Section 3, we formally define the settings of our study. In Section 4, we prove sufficient conditions for the independence of the optimal policy from the initial state and discuss their necessity. In Section 5, we discuss the objective functions and constraints studied in the prior work, showing whether they cause the optimal policy to depend on the initial state. Related work is summarized in Section 6.

2. Dependency of the optimal policy on the initial state

Consider a traveler who goes from an origin, A, to a destination, C, where the travel time depends on whether the traffic is normal or busy (see Figure 1). Upon the departure, the traveler does not know the exact traffic condition but

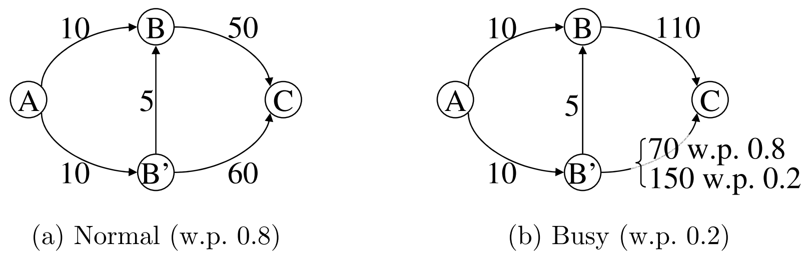


Figure 1: Travel time (a) at normal traffic and (b) (b) at busy traffic.

knows that the traffic is normal with probability 0.8 and busy otherwise. The traveler also knows 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. The path from B to C is busy if and only if the path from B′ to C is busy. The exact traffic condition becomes known when the traveler arrives at B or B′.

Using the settings of Figure 1, we discuss an MDP with constraints or a non-standard objective function, where the constraints and the objectives are with respect to the total travel time, X. The state of the MDP is the pair of the location (A, B, B′, or C) and the traffic condition (normal, busy, or yet unknown). The action of the MDP selects the next location to visit.

We first consider minimizing the expected value of total travel time, $\mathsf{E}[X]$, under the constraint that its variance, $\mathsf{Var}[X]$, is below a threshold, δ:

$$\begin{aligned} \text{minimize} \quad & \mathsf{E}[X] \\ \text{subject to} \quad & \mathsf{Var}[X] \leq \delta. \end{aligned} \tag{1}$$

Specifically, let $\delta = 360$ squared minutes in (1). This mean-variance tradeoff has been a popular criterion of optimization in the literature [42, 43].

There are five policies in our example. For each of these policies, $\mathsf{E}[X]$ and $\mathsf{Var}[X]$ are shown in Table 1 (a). Note that the indirect path from A to B′ to B (the third policy) surely takes longer than the direct path from A to B (the first policy). The optimal policy, $\pi^\star$, is the fourth policy, which suggests to first visit B', and take B′-B-C if the traffic is normal and take B′-C otherwise.

Now, consider another traveler, who starts at B' after the traffic condition

Policy	E	Var	$\mathsf{CTE}_{0.8}$
ABC	72.0	576.0	120.0
AB′C	75.2	312.9	96.0
AB′BC	77.0	576.0	125.0
$\pi^\star$: AB′BC if normal; AB′C if busy	71.2	358.5	96.0
AB′C if normal; AB′BC if busy	81.0	484.0	125.0

Table 1: The expectation (E), the variance (Var), and the conditional tail expectation ($\mathsf{CTE}_{0.8}$) of travel time from A.

Policy	E	Var	$\mathsf{CTE}_{0.8}$
B′C	60.0	0.0	60.0
B′BC	55.0	0.0	55.0

(a) Normal

Policy	E	Var	$\mathsf{CTE}_{0.8}$
B′C	86.0	1,124.0	150.0
B′BC	115.0	0.0	115.0

(b) Busy

Table 2: The expectation (E), the variance (Var), and the conditional tail expectation ($\mathsf{CTE}_{0.8}$) of the travel time from B′ when the traffic is (a) normal or (b) busy.

becomes known. Let Y be the total travel time of the second traveler. While the first traveler has already spent $c = 10$ minutes, the two travelers appear to be indistinguishable when they are at B′. In particular, the two travelers have common objective and constraint, because $\mathsf{E}[Y] = \mathsf{E}[X] - c$ and $\mathsf{Var}[Y] = \mathsf{Var}[X]$ if the they follow the same policy from B′.

The second traveler plans the travel after the traffic condition is known (see Table 2). When the traffic is normal, B′-B-C is optimal. When the traffic is busy, B′-B-C is optimal, because B′-C violates the constraint.

Observe that the two travelers behave differently when they respectively behave optimally. When the traffic is found busy, the first traveler takes B′-C, and the second takes B′-B-C. Given the decision of the second traveler, it might appear that the constraint is violated for the first traveler after the traffic condition is known. This can motivate the first traveler to follow the optimal policy for the second traveler from B′. However, doing so will result in taking A-B′-B-C regardless of the traffic condition. Taking A-B′-B-C is by no means desirable, because it takes surely longer than taking A-B-C.

We now consider an MDP with a non-standard objective function (and no constraint). Specifically, our objective function is conditional tail expectation (CTE), which is also known as conditional value at risk. CTE has a parameter, α, and is defined for a random variable, Y, as follows:

$$\mathsf{CTE}_\alpha[Y] \equiv \frac{(1-\beta)\mathsf{E}[Y|Y > V_\alpha] + (\beta - \alpha)V_\alpha}{1-\alpha}, \tag{2}$$

where $V_\alpha \equiv \min\{y \mid F_Y(y) \geq \alpha\}$, F_Y is the cumulative distribution function of Y, and $\beta \equiv F_Y(V_\alpha)$. For a continuous Y, or unless Y has a mass probability at V_α, CTE is simplified to $\mathsf{CTE}_\alpha[Y] = \mathsf{E}[Y|Y > V_\alpha]$, because $\beta = \alpha$. This simplification applies to the example in this section.

The last column of Table 1 shows the values of $\mathsf{CTE}_{0.8}[X]$ for each policy. There are two equally optimal policies, which result in $\mathsf{CTE}_{0.8}[X] = 96.0$. Following any of these optimal policies, the traveler first visits B′. Consider the second traveler who starts from B′ after observing the traffic condition (see Table 2). Regardless of the traffic condition, the policy of taking B′-B-C is optimal with respect to $\mathsf{CTE}_{0.8}[Y]$ for the second traveler. The optimal policies of the two travelers are thus mutually inconsistent. This inconsistency is analogous to the situation of the two travelers who make decisions based on (1).

3. Markov decision processes

We consider the Markov decision process (MDP) having a finite number of states and a finite horizon. Let N be the number of the time steps to consider. For $\ell \in [0, N]$, let $\mathbf{S}_\ell$ be the finite set of possible states at time ℓ. We assume that the state is augmented with accumulated reward and the history of previously visited states. Then $\mathbf{S}_\ell$ and $\mathbf{S}_m$ are mutually exclusive for $\ell \neq m$, because the history of visited states is a part of the state. Let $\mathbf{S}_{n:N} \equiv \cup_{\ell\in[n,N]}\mathbf{S}_\ell$ and $\mathbf{S} \equiv \mathbf{S}_{0:N}$. A policy, π, specifies an action to take, depending on the state. We allow the action space to be continuous or have infinite number of possible actions. The transition probability function, $p^\pi(s'|s)$, specifies the probability of transitioning from $s \in \mathbf{S}_\ell$ to $s' \in \mathbf{S}_{\ell+1}$ given the action is selected according to π for $\ell \in [0, N)$. Let R_ℓ be the reward that the agent gains immediately after taking an action at time ℓ for $\ell \in [0, N)$. Because the accumulated reward is a part of the state, R_ℓ can be specified by S_ℓ and $S_{\ell+1}$ for $\ell \in [0, N)$. Because the state space is finite and the state is augmented with accumulated reward, the distribution of the immediate reward must have a finite support.

An agent can start acting at any $n \in [0, N)$. With the knowledge of the state at n, the agent finds a policy to follow from a candidate set, Π_n. Let $\Pi \equiv \Pi_0$. Then Π_n is defined from Π by limiting the domain of $\pi \in \Pi_n$ to $\mathbf{S}_{n:N-1}$. We only consider Markovian and deterministic policies (as opposed to history-dependent or stochastic policies). Namely, the action to take from a given state is selected non-probabilistically and independently of how the agent reached that state. Recall, however, that our state includes the information about the history of visited states and accumulated reward. The assumption of deterministic policy does not lose generality, because our action space can be continuous and the immediate reward, R_ℓ, can be random given S_ℓ and A_ℓ. Specifically, for any probabilistic action, we can construct a deterministic action having the same effect as that probabilistic action.

The agent follows the policy that is optimal with respect to an optimization problem of the following form:

$$\begin{aligned} &\text{maximize}_{\pi \in \Pi_n} && f_n(X^\pi(s_n)) \\ &\text{subject to} && h_n(X^\pi(s_n)) = 1, \end{aligned} \tag{3}$$

where $X^\pi(s_n)$ is the cumulative reward for the agent who starts acting from state s_n by following a policy, π. The objective function, $f_n(\cdot)$, maps $X^\pi(s_n)$ to a real number; $h_n(\cdot)$ is an indicator function that represents whether the constraints are satisfied ($h_n(\cdot) = 1$) or not ($h_n(\cdot) = 0$). For example, $h_n(\cdot) \equiv \mathbf{1}\{g_n(\cdot) \in B_n\}$, where g_n is a multidimensional function that maps $X^\pi(s_n)$ to real numbers, B_n specifies the feasible region in the codomain of g_n, and $\mathbf{1}\{C\}$ denotes the indicator random variable whose value is 1 if the random condition, C, is satisfied and 0 otherwise.

Examples of objective functions are $f_n(\cdot) = \mathsf{E}[\cdot \mid S_n]$ and $f_n(\cdot) = \mathsf{CTE}_\alpha[\cdot \mid S_n]$, which we have discussed in Section 2. A constraint that we have seen in Section 2 is $h_n(\cdot) = \mathbf{1}\{\mathsf{Var}[\cdot \mid S_n] \leq \delta\}$. Notice that our objective function, f_n, and our constraints, h_n, can depend on n.

Throughout, we consider the case where

$$X^\pi(s_n) = r(s_n) + \sum_{\ell=n}^{N-1} R_\ell, \tag{4}$$

where the state, s_n, is augmented with the accumulated reward, $r(s_n)$. The agent who starts acting from s_n can be considered to have the initial wealth of $r(s_n)$. That is, the agent at time n seeks to maximize $f_n(\sum_{\ell=0}^{N-1} R_\ell)$ instead of $f_n(\sum_{\ell=n}^{N-1} R_\ell)$. Maximizing one of these quantities is equivalent to maximizing another when f_n satisfies the following separability:

$$f_n\left(\sum_{\ell=0}^{N-1} R_\ell\right) \quad = \quad \sum_{\ell=0}^{n-1} R_\ell + f_n\left(\sum_{\ell=n}^{N-1} R_\ell\right) \tag{5}$$

The separability is, for example, satisfied when $f_n(\cdot) = \mathsf{E}[\cdot \mid S_n]$ or $f_n(\cdot) = \mathsf{CTE}_\alpha[\cdot \mid S_n]$. The separability does not hold, for example, when $f_n(\sum_{\ell=0}^{N-1} R_\ell) = \mathsf{E}\left[\mathbf{1}\left\{\sum_{\ell=0}^{N-1} R_\ell > c\right\}\right]$, where the agent seeks to maximize the probability that the cumulative reward exceeds a target value, c. When the separability does not hold, the state, $s_n \in \mathbf{S}_n$ for $n \in [1, N]$, generally needs to include additional information about $r(s_n)$ to adequately maximize $f_n(\sum_{\ell=0}^{N-1} R_\ell)$. Intuitively, if we seek to maximize the probability that our final wealth exceeds a target, the optimal action should depend on what has already been earned. In this paper, the objective function is not necessarily separable.

In summary, the MDP studied in this paper can be specified with a tuple, $\langle \mathbf{S}, \Pi, p, f, h \rangle$, where $f \equiv \{f_n \mid n \in [0, N)\}$ and $h \equiv \{h_n \mid n \in [0, N)\}$. We use $\mathtt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$ to denote the MDP with these specifications. Recall that the reward is implicitly defined via the state, which is augmented with the accumulated reward.

4. Conditions for the independence from the initial state

We study the conditions on the objective functions and the constraints of the MDP so that the optimal policy is independent of the initial state (or when the agent starts acting). We say that an MDP is *consistent* when the optimal policy is independent of the initial state:

Definition 4.1. *We say that* $\texttt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$ *is consistent if the following is satisfied. For any* $n \in [1, N)$*, if* $\pi^\star$ *is optimal with respect to the objective function* f_{n-1} *and constraints* h_{n-1} *for the agent who start acting from* $s \in \mathbf{S}_{n-1}$*, then* $\pi^\star$ *is optimal (and hence feasible), with respect to* f_n *and* h_n*, for the agent who starts acting from any* $s' \in \mathbf{S}_n$ *such that* $p^{\pi^\star}(s' \mid s) > 0$*. We say that* $\texttt{MDP}_{f,h}$ *is consistent if* $\texttt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$ *is consistent for any* $\mathbf{S}$*,* Π*, and* p*.*

We now revisit the objective and constraints in (1). In Figure 1, the optimal policy, which we find by solving the MDP upon departure, becomes infeasible for the MDP solved at intersection B if the traffic is busy. The transition probability to the busy state is strictly positive. Hence, the corresponding $\texttt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$ is indeed inconsistent in the sense of Definition 4.1.

4.1. Preliminaries

To formally study consistency, it is important to understand the objective function, f, as a dynamic risk measure (dynamic RM). In the context of the MDP, an RM, $\rho_n(\cdot)$, maps a random variable, Y, that generally depends on the states after time n to a real number given S_n. The value of $\rho_n(Y)$ is random before time n and becomes deterministic given S_n, because the value of $\rho_n(Y)$ depends on S_n, which is random before time n. When a random variable becomes deterministic at time n, we say that the random variable is $\mathcal{F}_n$-measurable. In the context of the MDP, a dynamic RM is defined as follows:

Definition 4.2. *Consider a generic Markov chain having a finite state space and a finite horizon,* N*, where the state at time* n *includes the information about the history of the states before time* n *(i.e., the state is augmented with prior* *states). Let Y be a generic $\mathcal{F}_N$-measurable random variable that can depend on the state of the Markov chain at time N. We say that $\rho \equiv \{\rho_n \mid n \in [0, N)\}$ is a dynamic RM if ρ_n maps Y to an $\mathcal{F}_n$-measurable random variable for each $n \in [0, N)$.*

In the literature of Finance, it is standard to assume that ρ_n satisfies the property of convexity or coherency [57, 31, 13, 54, 53, 5]. We do not assume these properties, because they can be undesirable for other applications. In particular, the convex or coherent RM satisfies the separability (5). Then the actions optimal with respect to these RMs are insensitive to the accumulated cost, which is undesirable for avoiding certain types of risks. Indeed, non-convex or non-coherent RMs have been considered with MDPs in the literature [41, 65].

In our MDP, X^π is an $\mathcal{F}_N$-measurable random. Because S_n is $\mathcal{F}_n$-measurable, so is $f_n(X^\pi(S_n))$. Hence, f is a dynamic RM. Analogously, h is a dynamic RM but has an additional property that its value is 0 or 1. We refer to such a dynamic RM as an indicator dynamic RM. In the following, we use $\Pr(\cdot)$ to denote the probability with respect to the probability space defined by an $\mathcal{F}_N$-measurable random variable depending on the context.

We use the following definitions to provide the conditions for consistency:

Definition 4.3. *For $n \in [1, N)$ with $N < \infty$, a dynamic RM, $\rho \equiv \{\rho_n \mid n \in [0, N)\}$, is called optimality-consistent for the n if the following is satisfied: for any $\mathcal{F}_N$-measurable random variables, Y and Z, defined on an arbitrary Markov chain on a finite state space that is augmented with prior states, 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 if ρ is optimality-consistent for any $n \in [1, N)$.*

Definition 4.4. *For $n \in [1, N)$ with $N < \infty$, a dynamic RM, $\rho \equiv \{\rho_n \mid n \in [0, N)\}$, is called non-decreasing for the n if we have $\Pr(\rho_{n-1}(Y) \leq \rho_n(Y)) = 1$ for any $\mathcal{F}_N$-measurable random variable Y defined on an arbitrary Markov chain on a finite state space that is augmented with prior states. Also, we say that ρ is non-decreasing if ρ is non-decreasing for any $n \in [1, N)$.*

To get a sense of Definition 4.3, suppose that ρ is not optimality-consistent for an n. Then one can prefer Y to 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, Z becomes surely at least as preferable as Y (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 Section 4.2. The following corollary follows from the contrapositive of the condition of Definition 4.3:

Corollary 1. *Let ρ be an optimality-consistent dynamic RM. Let Y and Z be $\mathcal{F}_N$-measurable random variables. If $\Pr(\rho_{n-1}(Y) \geq \rho_{n-1}(Z)) = 1$, then $\Pr(\rho_n(Y) > \rho_n(Z)) > 0$ or $\Pr(\rho_n(Y) \geq \rho_n(Z)) = 1$.*

4.2. Sufficient conditions

We are now ready to state a sufficient condition for $\texttt{MDP}_{f,h}$ to be consistent:

Theorem 4.1. *If f is an optimality-consistent dynamic RM and h is a non-decreasing indicator dynamic RM, then $\texttt{MDP}_{f,h}$ is consistent as defined in Definition 4.1.*

Proof. We will prove that $\texttt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$ is consistent for any $\mathbf{S}$, Π, and p if f is an optimality-consistent dynamic RM and h is a non-decreasing dynamic RM. Recall from Definition 4.1 that, for consistent $\texttt{MDP}_{f,h}(\mathbf{S}, \Pi, p)$, if π is optimal at $s_{n-1} \in \mathbf{S}_{n-1}$ and $p^\pi(s_n \mid s_{n-1}) > 0$, then π is optimal at s_n. We will prove the contrapositive of this statement. Namely, if π is not optimal at $s_n \in \mathbf{S}_n$, then π is not optimal at s_{n-1} or $p^\pi(s_n \mid s_{n-1}) = 0$. In other words, if π is not optimal from $s_n \in \mathbf{S}_n$, then π is not optimal from any s_{n-1} such that $p^\pi(s_n|s_{n-1}) > 0$. Suppose that π is not optimal from $s_n \in \mathbf{S}_n$, so that π is either infeasible or feasible but not optimal.

We first consider the case where π is infeasible from s_n. If $p^\pi(s_n|s_{n-1}) > 0$ for an $s_{n-1} \in \mathbf{S}_{n-1}$, then π must be infeasible (hence, not optimal) from the s_{n-1}; this is because h is a non-decreasing indicator dynamic RM and $h_n(X^\pi(s_n)) = 0$, so that $h_{n-1}(X^\pi(s_{n-1})) = 0$. Hence, the contrapositive of the property of the consistency holds.

The rest of the proof considers the case where π is feasible but not optimal from s_n. Then there exists an optimal policy, $\pi^\star \neq \pi$, from s_n, and we have

$$f_n(X^\pi(s_n)) \quad < \quad f_n(X^{\pi^\star}(s_n)).$$

Now, consider a policy π'. Before time n, π' is equivalent to π. At time n and after time n, π' assigns the same actions as those assigned by $\pi^\star$ if $S_n = s_n$ (namely, for all states reachable from s_n) and assigns the same actions as those assigned by π otherwise. Notice that a state $s' \in \mathbf{S}_\ell$ for $\ell \in [n+1, N)$ is reachable only from a single $s \in \mathbf{S}_n$, because a state includes the information about the history of visited states.

We will show that π is not optimal from s_{n-1} if $p^\pi(s_n|s_{n-1}) > 0$. If π is infeasible from s_{n-1}, then π is not optimal from s_{n-1}. We thus consider the case where π is feasible from s_{n-1} and establish that

$$f_{n-1}(X^\pi(s_{n-1})) \quad < \quad f_{n-1}(X^{\pi'}(s_{n-1}))$$

for the s_{n-1} such that $p^\pi(s_n|s_{n-1}) > 0$. Observe that

$$f_n(X^\pi(s_n)) \quad < \quad f_n(X^{\pi'}(s_n)), \tag{6}$$

and

$$f_n(X^\pi(s_n')) \quad = \quad f_n(X^{\pi'}(s_n')) \tag{7}$$

for all $s_n' \in \mathbf{S}_n$ such that $s_n' \neq s_n$ because of the way π' is constructed. Now, if $p^\pi(s_n|s_{n-1}) > 0$, then, by the optimality-consistency of the objective function, we must have

$$f_{n-1}(X^\pi(s_{n-1})) \quad < \quad f_{n-1}(X^{\pi'}(s_{n-1})), \tag{8}$$

which can be formally proved as follows. Let $Y \equiv X^\pi(s_{n-1})$ and $Z \equiv X^{\pi'}(s_{n-1})$. Notice that $f_n(Y)$ is a random variable that takes value $f_n(X^\pi(s_n))$ with probability $p^\pi(s_n|s_{n-1})$ for all $s_n \in \mathbf{S}_n$, and $f_n(Z)$ is a random variable having analogous properties. From the observations made with (6) and (7), we find

$\Pr(f_n(Y) \le f_n(Z)) = 1$ and $\Pr(f_n(Y) < f_n(Z)) > 0$. Hence, the optimality-consistency of the objective function implies $\Pr(f_{n-1}(Y) < f_{n-1}(Z)) > 0$. However, $\Pr(f_{n-1}(Y) < f_{n-1}(Z))$ is 0 or 1, because $f_{n-1}(X^{\pi}(s_{n-1}))$ and $f_{n-1}(X^{\pi'}(s_{n-1}))$ are deterministic. Therefore, (8) is established. □

We elaborate on the sufficient conditions in Section 4.3 and Section 4.4.

4.3. Remarks on optimality-consistent objective functions

First, we remark that expectation and entropic risk measure (ERM) can be shown to be optimality-consistent. Here, the ERM of a random variable, X, is defined with the parameter of risk-sensitivity, γ, such that $\mathsf{ERM}_\gamma[X] \equiv \frac{1}{\gamma} \ln \mathsf{E}[\exp(\gamma\, X)]$ [23]. In fact, optimality-consistency can be shown for a class of iterated RMs:

Definition 4.5. *Consider a generic $\mathcal{F}_N$-measurable random variable, Y. We say that a dynamic RM, $\rho \equiv \{\rho_n \mid n \in [0, N)\}$, is an iterated RM if $\rho_n[Y] = \bar\rho_n[\rho_{n+1}[Y]]$ for each $n \in [0, N)$, 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.*

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, ERM is an iterated RM with $\bar\rho_n[\cdot] = \mathsf{ERM}_\gamma[\cdot|S_n]$. Iterated conditional tail expectation (ICTE) studied in [13, 31, 48] is an iterated RM with $\bar\rho_n[\cdot] = \mathsf{CTE}_\alpha[\cdot|S_n]$. The following corollary can be proved formally:

Corollary 2. *An iterated RM, as defined in Definition 4.5, is optimality-consistent for a particular $n \in [1, N)$ if $\Pr(\bar\rho_{n-1}[V] < \bar\rho_{n-1}[W]) > 0$ for any $\mathcal{F}_n$-measurable random variable, V and W, such that $\Pr(V \le W) = 1$ and $\Pr(V < W) > 0$.*

Proof. By Definition 4.3, it suffices to show that, for an iterated RM that satisfies the conditions of the corollary, we have $\Pr(\rho_{n-1}[Y] < \rho_{n-1}[Z]) > 0$ for any $\mathcal{F}_N$-measurable random variables, Y and Z, that satisfy

$$\Pr(\rho_n[Y] \le \rho_n[Z]) = 1 \quad \text{and} \quad \Pr(\rho_n[Y] < \rho_n[Z]) > 0. \tag{9}$$

By the definition of ρ, we have

$$\rho_{n-1}[Z] - \rho_{n-1}[Y] = \bar{\rho}_{n-1}[\rho_n[Z]] - \bar{\rho}_{n-1}[\rho_n[Y]]. \tag{10}$$

Observe that $V = \rho_n[Y]$ and $W = \rho_n[Z]$ satisfy the conditions of the corollary by (9). Therefore, $\Pr(\rho_{n-1}[Y] < \rho_{n-1}[Z]) > 0$, which completes the proof. $\square$

Corollary 2 allows us to check whether a given iterated RM, ρ, is optimality-consistent by studying the properties of $\bar{\rho}_n$ for $n \in [0, N)$. For example, an iterated RM defined for a random variable, V, with

$$\bar{\rho}_n(V) \quad \equiv \quad \eta\, \mathsf{E}[V|S_n] + (1-\eta)\, \mathsf{CTE}_\alpha[V|S_n] \tag{11}$$

is optimality-consistent for $\alpha, \eta \in (0,1)$. One can expect that minimizing this iterated RM of cumulative cost, X, leads to balancing between minimizing expected cumulative cost and minimizing the riskiness of large loss. The MDP of minimizing the iterated RM defined with (11) can also be shown to be equivalent to a robust MDP of minimizing the expectation when the parameters of the MDP have uncertainties (see [49]).

An easy way to verify the conditions of Corollary 2 is to demonstrate that $\bar{\rho}_{n-1}(V)$ can be expressed as

$$\bar{\rho}_{n-1}(V) = \int_{x\in\mathbb{R}} u(x, F_V(x))\, dF_V(x), \tag{12}$$

where $u(\cdot,\cdot)$ is monotonically increasing with respect to its first argument, and F_V is the cumulative distribution function of an $\mathcal{F}_n$-measurable random variable, V. Let $F_V^{-1}(q) = \min\{x \mid F_V(x) = q\}$. For the V and W defined in Corollary 2, we have

$$\bar{\rho}_{n-1}(W) = \int_0^1 u(F_W^{-1}(q), q)\, dq > \int_0^1 u(F_V^{-1}(q), q)\, dq = \bar{\rho}_{n-1}(Y),$$

where the inequality holds, because u is monotonically increasing with respect to its first argument, and $\Pr(V \le W) = 1$ and $\Pr(V < W) > 0$ (i.e., $F_W^{-1}(q) \ge F_V^{-1}(q)$ for $0 \le q \le 1$, and there exists $q_0 < q_1$ such that $F_W^{-1}(q) > F_V^{-1}(q)$ for $q_0 \le q \le q_1$).

For example, $\mathsf{E}[V \mid S_n] = \int_{x\in\mathbb{R}} x\, dF_V(x)$, so that $u(x, F_V(x)) = x$ is monotonically increasing with x. For brevity, here, we do not explicitly denote that F_V is conditional on S_n, which is clear from the context. Also, the $\bar{\rho}_n(V)$ defined with (11) can be expressed as $\bar{\rho}_n(V) = \int_{x\in\mathbb{R}} u(x, F_V(x))\, dF_V(x)$ by defining

$$u(x, F_V(x)) \quad \equiv \quad \begin{cases} \eta\, x & \text{if } F_V(x) < \beta \\ \eta\, x + \frac{(1-\eta)(\beta-\alpha)}{(1-\alpha)(\beta-\alpha^-)} & \text{if } F_V(x) = \beta \\ \eta\, x + \frac{1-\eta}{1-\alpha}\, x & \text{if } F_V(x) > \beta, \end{cases} \tag{13}$$

where α and β are as defined for (2) but now with respect to F_V; also, $\alpha^- \equiv \sup_{F_V(x)<\beta} F_V(x)$. Observe that $u(x, F_V(x))$ of (13) is monotonically increasing with respect to x for $0 < \eta \leq 1$. By letting $\eta = 0$ in (11), we have $\bar{\rho}_n(V) = \mathsf{CTE}_\alpha[V \mid S_n]$. Also, $u(x, F_V(x))$ stays constant for $F(x) < \beta$ when $\eta = 0$, which agrees with the fact that ICTE is not optimality-consistent.

The iterated RM that satisfies the conditions of Corollary 2 is not only optimality-consistent but also time-consistent. That is, the optimal policy satisfies the Bellman equation. The Bellman equation can be verified by showing that the optimal policy from an arbitrary $s_n \in \mathbf{S}_n$ can be found by first finding optimal policies from each of $s_{n+1} \in \mathbf{S}_{n+1}$ for $n \in [0, N)$. Formally,

$$\rho_n\left(X^{\pi^\star_{s_n}}(s_n)\right) = \max_a \bar{\rho}_{n+1}\left(\rho_{n+1}\left(X^{\pi^\star_{S_{n+1}}}(S_{n+1})\right) \mid S_n = s_n, A(s_n) = a\right), \tag{14}$$

where $\pi^\star_{s_n}$ is the optimal policy from $s_n \in \mathbf{S}_n$, and $A(s_n)$ denotes the action selected at s_n. Here we use $\bar{\rho}_{n+1}(\cdot \mid S_n = s_n, A(s_n) = a)$ to denote that $\bar{\rho}_{n+1}$ is calculated given that $S_n = s_n$ and $A(s_n) = a$. Now, suppose that we follow a suboptimal policy, $\pi'_{s_{n+1}}$, at s_{n+1} that we can reach with positive probability from s_n by taking the action, a. Then, because we have

$$\rho_{n+1}\left(X^{\pi^\star_{s_{n+1}}}(s_{n+1})\right) \quad > \quad \rho_{n+1}\left(X^{\pi'_{s_{n+1}}}(s_{n+1})\right),$$

Corollary 2 implies that the value of $\bar{\rho}_{n+1}$ in (14) decreases by replacing $\pi^\star_{s_{n+1}}$ with $\pi'_{s_{n+1}}$. Here, notice that the value of $\Pr(\bar{\rho}_n[V] < \bar{\rho}_n[W])$ is 0 or 1 when $s_n \in \mathbf{S}_n$ is given, so that $\Pr(\bar{\rho}_n[V] < \bar{\rho}_n[W]) > 0$ implies $\bar{\rho}_n[V(s_n)] < \bar{\rho}_n[W(s_n)]$, where $V(s_n)$ denotes the conditional V given $S_n = s_n$, and $W(s_n)$ is defined analogously. The above argument is summarized in the following proposition:

Proposition 1. *The Bellman equation (14) holds for the iterated RM that satisfies the conditions of Corollary 2.*

We remark that an iterated RM, ρ, does not have the property, $\rho_n(\cdot) = \bar{\rho}_n(\cdot)$, except for special cases such as, for all n, $\bar{\rho}_n[\cdot] = \mathsf{E}[\cdot|S_n]$, $\bar{\rho}_n[\cdot] = \mathsf{ERM}_\gamma[\cdot|S_n]$, $\bar{\rho}_n[\cdot] = \min[\cdot|S_n]$, or $\bar{\rho}_n[\cdot] = \max[\cdot|S_n]$, where $\max[X]$ (respectively, $\min[X]$) denotes the maximum (respectively, minimum) value that X can take with positive probability. 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, $N - n$, when N is finite. A key implication of Corollary 2 and (12) is that there is a large class of iterated RMs having optimality-consistency.

4.4. Remarks on non-decreasing constraints

Examples of the constraints that make $\texttt{MDP}_{f,h}$ consistent include $\max[X^\pi] \le \delta$ and $\min[X^\pi] \ge \delta$. 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, $\mathbf{1}\{\max(X^\pi) \le \delta\}$ is non-decreasing. Analogously, $\mathbf{1}\{\min(X^\pi) \ge \delta\}$ is non-decreasing.

We have seen with Figure 1 that the MDP with (1) is not consistent. We can now understand that the inconsistency is due to the constraint in (1), because the objective function in (1) is optimality-consistent. Observe that $\mathsf{Var}[X^{\pi^\star}]$ increases from 312.9 upon departure to 1124.0 at B′ if the traffic is found busy at B′. Hence, $\mathbf{1}\{\mathsf{Var}[\cdot \mid S_n] \le 360\}$ is not non-decreasing.

A way to modify (1) into an consistent $\texttt{MDP}_{f,h}$ is to incorporate the constraints that might need to satisfy in the future:

$$\begin{aligned} \text{min.} \quad & \mathsf{E}[X] \\ \text{s.t.} \quad & \mathsf{Var}[X \mid S_\ell = s_\ell] \le \delta, \forall s_\ell \in \mathbf{S}_\ell, \forall \ell \in [0, N), \end{aligned} \tag{15}$$

Then $\pi^\star$ becomes infeasible for the optimization problem to be solved at the time of the departure, which resolves the issue of the inconsistency. This construction of non-decreasing constraints can be applied in the following general settings, where recall that $\mathbf{1}\{\cdot\}$ denotes an indicator random variable:

Corollary 3. *Let h be an indicator dynamic RM that is not necessarily non-decreasing. Let X be an $\mathcal{F}_N$-measurable random variable and $X(S_\ell)$ be the conditional X given S_ℓ, the state at time ℓ, for $\ell \in [0, N]$. Then h' such that*

$$h'_n(X(S_n)) \quad \equiv \quad \mathbf{1}\left\{\mathsf{E}\left[\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n, N]\right\} \mid S_n\right] = 1\right\}$$

for each $n \in [0, N)$ is an indicator dynamic RM and non-decreasing.

Proof. Observe that h' is an indicator dynamic RM, because $h'_n(X(S_n))$ is either 0 or 1 and becomes deterministic at time n for each $n \in [0, N]$ (to see why $h'_n(X(S_n))$ is deterministic at time n, notice that $h'_n(X(s_n)) = 1$ iff $h'_\ell(X(s_\ell))$ for any state s_ℓ reachable from s_n for $\ell \in (n, N]$). It suffices to show, for $n \in [1, N]$, that

$$\mathsf{E}\left[h'_n(X(S_n)) \mid S_{n-1} = s_{n-1}\right] = 1. \tag{16}$$

under the condition that $h'_{n-1}(X(s_{n-1})) = 1$ for $s_{n-1} \in \mathbf{S}_{n-1}$.

Because $h'_{n-1}(X(s_{n-1})) = 1$, we have

$$\mathsf{E}\left[\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n-1, N]\right\} \mid S_{n-1} = s_{n-1}\right] \quad = \quad 1,$$

by the definition of h'. This implies

$$\mathsf{E}\left[\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n, N]\right\} \mid S_{n-1} = s_{n-1}\right] \quad = \quad 1, \tag{17}$$

because

$$\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n-1, N]\right\} \;\leq\; \mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n, N]\right\} \;\leq\; 1.$$

By the recursive property of expectation, we thus have from (17) that

$$\mathsf{E}\left[\mathsf{E}\left[\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n-1, N]\right\} \mid S_n\right] \mid S_{n-1} = s_{n-1}\right] \quad = \quad 1.$$

Because the value of the inner expectation is in $[0, 1]$, it must be 1 with probability one. Then this implies $h'(X(S_n)) = 1$. Formally, we have

$$\mathsf{E}\left[\mathbf{1}\left\{h_\ell(X(S_\ell)) = 1, \forall \ell \in [n-1, N]\right\} \mid S_n\right] = 1 \;\Rightarrow\; h'(X(S_n)) = 1$$

This establishes (16). $\square$

4.5. Necessary conditions

Next, we study necessity of the sufficient condition provided in Theorem 4.1:

Lemma 4.1. *If* $\texttt{MDP}_{f,h}$ *is consistent for any optimality consistent* f*, then* h *must be non-decreasing. If* $\texttt{MDP}_{f,h}$ *is consistent for any non-decreasing* h*, then* f *must be optimality-consistent.*

Proof. The proof consists of two parts. In Part I, we will prove that h must be non-decreasing if $\texttt{MDP}_{f,h}$ is consistent for $f \equiv 0$ (i.e., any feasible policy is optimal). In Part II, we will prove that f must be optimality-consistent if $\texttt{MDP}_{f,h}$ is consistent for $h \equiv 1$ (i.e., no constraints).

Part I:. It suffices to construct an $\texttt{MDP}_{f\equiv 0,h}(\mathbf{S}, \Pi, p)$ that is not consistent, for every h that is not non-decreasing. Because h is not non-decreasing, Definition 4.4 implies that there exist $N \in [0, \infty)$, a Markov chain on a finite state space, $\tilde{\mathbf{S}}$, that is augmented with prior states, and an $\mathcal{F}_N$-measurable random variable, Y, such that we have $h_{n-1}(Y) = 1$ and $h_n(Y) = 0$ for an $n \in [1, N)$ with non-zero probability. Let $\tilde{p}$ be the transition probability function of that Markov chain, where $\tilde{p}(\tilde{s}_n \mid \tilde{s}_{n-1})$ denotes the probability of transitioning from $\tilde{s}_{n-1}$ to $\tilde{s}_n$ for $(\tilde{s}_{n-1}, \tilde{s}_n) \in \tilde{\mathbf{S}}_{n-1} \times \tilde{\mathbf{S}}_n$, $n \in [1, N]$.

From the Markov chain on $\tilde{\mathbf{S}}$ with transition probability function, $\tilde{p}$, we can construct the $\texttt{MDP}_{f\equiv 0,h}(\mathbf{S}, \Pi, p)$ that is not consistent. We first augment the state with cumulative reward with a default policy, π, such that

$$\mathbf{S}_n = \{(\tilde{s}_n, 0) \mid \tilde{s}_n \in \tilde{\mathbf{S}}_n\}, \forall n < N \tag{18}$$

$$\mathbf{S}_N = \{(\tilde{s}_N, Y(\tilde{s}_N)) \mid \tilde{s}_N \in \tilde{\mathbf{S}}_N\}, \tag{19}$$

where $Y(\tilde{s}_N)$ denotes the value of Y given that the state of the Markov chain at time N is $\tilde{s}_N$. Let the transition probability, $p^\pi \equiv q$, with π be such that

$$p^\pi((\tilde{s}_n, 0) \mid (\tilde{s}_{n-1}, 0)) = \tilde{p}(\tilde{s}_n \mid \tilde{s}_{n-1}), \forall (\tilde{s}_{n-1}, \tilde{s}_n) \in \tilde{\mathbf{S}}_{n-1} \times \tilde{\mathbf{S}}_n, n < N \tag{20}$$

$$p^\pi((\tilde{s}_N, Y(\tilde{s}_N)) \mid (\tilde{s}_{N-1}, 0)) = \tilde{p}(\tilde{s}_N \mid \tilde{s}_{N-1}), \forall (\tilde{s}_{N-1}, \tilde{s}_N) \in \tilde{\mathbf{S}}_{N-1} \times \tilde{\mathbf{S}}_N \tag{21}$$

The policy π is then feasible (and hence optimal) before time n but can become infeasible at time n. The $\texttt{MDP}_{f\equiv 0,h}(\mathbf{S},\Pi,p)$ thus constructed is hence not consistent.

Part II:. Now, we will construct an $\texttt{MDP}_{f,h\equiv 1}(\mathbf{S},\Pi,p)$ that is not consistent, for every f that is not optimality-consistent. Because f is not optimality-consistent, Definition 4.3 implies that there exist $N \in [0,\infty)$, a Markov chain on a finite state space, $\tilde{\mathbf{S}}$, that is augmented with prior states, and $\mathcal{F}_N$-measurable random variables, Y_1 and Y_2, such that, for an $n \in [1,N)$, we have $f_{n-1}(Y_1) \geq f_{n-1}(Y_2)$ and $f_n(Y_1) \leq f_n(Y_2)$ with probability one, and $f_n(Y_1) < f_n(Y_2)$ with non-zero probability. Let $\tilde{p}$ be the transition probability function of that Markov chain.

Analogously to Part I, we can construct the $\texttt{MDP}_{f,h\equiv 1}(\mathbf{S},\Pi,p)$ that is not consistent. Let

$$\begin{aligned}\mathbf{S}_n &= \{(\tilde{s}_n,0) \mid \tilde{s}_n \in \tilde{\mathbf{S}}_n\}, \forall n < N && (22)\\ \mathbf{S}_N &= \{(\tilde{s}_N, Y_1(\tilde{s}_N)) \mid \tilde{s}_N \in \tilde{\mathbf{S}}_N\} \cup \{(\tilde{s}_N, Y_2(\tilde{s}_N)) \mid \tilde{s}_N \in \tilde{\mathbf{S}}_N\}, && (23)\end{aligned}$$

where $Y_1(\tilde{s}_N)$ and $Y_1(\tilde{s}_N)$ are defined analogously to $Y(\tilde{s}_N)$ in Part I.

Consider two policies, π_1 and π_2, and define their transition probabilities, $p^{\pi_1} \equiv q$ and $p^{\pi_2} \equiv q$, respectively as follows:

$$p^{\pi_i}((\tilde{s}_n,0) \mid (\tilde{s}_{n-1},0)) = \tilde{p}(\tilde{s}_n \mid \tilde{s}_{n-1}), \forall (\tilde{s}_{n-1},\tilde{s}_n) \in \tilde{\mathbf{S}}_{n-1} \times \tilde{\mathbf{S}}_n, n < N, \quad (24)$$

$$p^{\pi_i}((\tilde{s}_N, Y_i(\tilde{s}_N)) \mid (\tilde{s}_{N-1},0)) = \tilde{p}(\tilde{s}_N \mid \tilde{s}_{N-1}), \forall (\tilde{s}_{N-1},\tilde{s}_N) \in \tilde{\mathbf{S}}_{N-1} \times \tilde{\mathbf{S}}_N \quad (25)$$

for $i = 1,2$. The policy π_1 is then optimal before time n ($f_{n-1}(Y_1) \geq f_{n-1}(Y_2)$ with probability one) but can become suboptimal at time n ($f_n(Y_1) < f_n(Y_2)$ with non-zero probability). The $\texttt{MDP}_{f,h\equiv 1}(\mathbf{S},\Pi,p)$ thus constructed is hence not consistent. □

For a limited class of objective functions, we can establish the necessary and sufficient condition for $\texttt{MDP}_{f,h}$ to be consistent. Specifically, the following corollary follows from the results in Section 4.

Corollary 4. *Suppose that there exists ω such that $-\infty < \omega < f_n(X)$ for any $\mathcal{F}_N$-measurable random variable, X. If we say that any infeasible policy is optimal when there is no feasible policy, then $\mathtt{MDP}_{f,h}$ is consistent if and only if $f'(\cdot) \equiv \{(f_n(\cdot) - \omega)\, h_n(\cdot) \mid n \in [0, N]\}$ is optimality-consistent.*

Proof. By Theorem 4.1 and Part II of the proof for Lemma 4.1, $\mathtt{MDP}_{f',1}$ is consistent if and only if f' is optimality-consistent. Hence, the corollary can be established by showing that $\mathtt{MDP}_{f,h}$ and $\mathtt{MDP}_{f',1}$ are equivalent with respect to the optimality of a policy. Consider two policies, π_1 and π_2. Without loss of generality, we assume that $f'_n(X^{\pi_1}(s_n)) \geq f'_n(X^{\pi_2}(s_n))$ at $s_n \in \mathbf{S}_n$. By observing the following four cases, we can conclude that the optimality of a policy in the two MDPs is consistent with each other:

Case 1 Both policies are infeasible:

$$f'_n(X^{\pi_1}(s_n)) = f'_n(X^{\pi_2}(s_n)) = 0 \Leftrightarrow h_n(X^{\pi_1}(s_n)) = h_n(X^{\pi_2}(s_n)) = 0.$$

Case 2 One policy is infeasible:

$$f'_n(X^{\pi_1}(s_n)) > f'_n(X^{\pi_2}(s_n)) = 0 \Leftrightarrow h_n(X^{\pi_1}(s_n)) = 1 \text{ and } h_n(X^{\pi_2}(s_n)) = 0.$$

Case 3 Two policies are feasible and equally good:

$$\begin{aligned} & f'_n(X^{\pi_1}(s_n)) = f'_n(X^{\pi_2}(s_n)) > 0 \\ & \quad \Leftrightarrow \quad h_n(X^{\pi_1}(s_n)) = h_n(X^{\pi_2}(s_n)) = 1 \text{ and } f_n(X^{\pi_1}(s_n)) = f_n(X^{\pi_2}(s_n)). \end{aligned}$$

Case 4 Two policies are feasible, and one is better than the other:

$$\begin{aligned} & f'_n(X^{\pi_1}(s_n)) > f'_n(X^{\pi_2}(s_n)) > 0 \\ & \quad \Leftrightarrow \quad h_n(X^{\pi_1}(s_n)) = h_n(X^{\pi_2}(s_n)) = 1 \text{ and } f_n(X^{\pi_1}(s_n)) > f_n(X^{\pi_2}(s_n)). \end{aligned}$$

□

5. Implications to the previously studied MDPs

Here, we discuss implications of the results in the prior sections to the MDPs that have been studied in the literature. Although a popular objective is to minimize expected cumulative cost, there exists a significant amount of the work on those MDPs that are sensitive to risk or require to satisfy constraints to avoid huge loss. In Section 5.1, we review the MDPs that have both risk-sensitive objectives and constraints. We review the MDPs having constraints in Section 5.2 and those having risk-sensitive objectives in Section 5.3. We will see what objective functions are optimality-consistent and what constraints have the non-decreasing property.

5.1. Risk-sensitive Markov decision processes with constraints

The tradeoff between the expected value and the variance of the cumulative reward over a finite horizon is studied in [42]. Specifically, the objective is to minimize the variance, which is not optimality-consistent. The constraint requires that the expected cumulative reward is above a threshold, which does not have the non-decreasing property. Hence, the corresponding $\mathtt{MDP}_{f,h}$ is not consistent. In addition, "Bellman's principle of optimality does not hold" [42].

This tradeoff is also studied for the reward at the steady state [15, 36, 60]. Considering only the reward at the steady state is out of the scope of this paper. However, when the MDP is a uni-chain, there is a unique distribution of the steady state. In this case, there is a unique policy that is optimal with respect to the reward at the steady state. The unique optimal policy found at one moment stays optimal in the future, because the steady state stays unchanged.

5.2. Markov decision processes with constraints

MDPs with constraints, or constrained MDPs, have been studied extensively in the literature. These include MDPs with multiple criteria, where one of the criteria is used to define the objective function, and others are used for constraints. Our results apply to such setting with multiple criteria as well.

In this section, we identify whether the previously studied constraints have the non-decreasing property.

Altman [3] studies constrained MDPs that require to minimize the expect cumulative cost of one type, while keeping the expected cumulative costs of other types below thresholds. Altman's class of constrained MDPs has also been studied in [19, 20, 21, 27, 28]. In general, these constraints do not have the non-decreasing property; i.e., the $\mathtt{MDP}_{f,h}$ defined with the constrained MDP of Altman's class is not necessarily consistent. It has been pointed out that Bellman's principle of optimality is not necessarily satisfied for a constrained MDP of Altman's class [32, 34, 55, 58], where counter-examples have been constructed for the case where the constrained MDP is a multi-chain over an infinite horizon.

The constraints of Altman's class do not have to be with respect to the expected cumulative cost. The constraint that requires the cumulative cost to be below a predefined level with high probability is studied in [18, 19, 65]. Geibel studies analogous constraints on reward instead of cost [28]. Fulkerson et al. [25] study the constraint that the probability of reaching the goal must be above a desired level. Teichteil-Königsbuch [63] expresses the constrains with Probabilistic Real Time Computation Tree Logic, which is popular in model checking. In general, these constraints do not have the non-decreasing property.

The optimal policy for the constrained MDPs of Altman's class cannot, in general, be found with dynamic programming, as is suggested by the violation of Bellman's principle of optimality. Geibel [28] thus studies four approaches for optimization. One of his approaches is to strengthen the constraints in such a way that the original constraints must be satisfied from potential future states. This approach is equivalent to our modification of (1) into (15) or that in Corollary 3. This approach is also suggested in [26]. As we have discussed in Section 4, the strengthened constraints have the non-decreasing property.

There also exists work that uses hard constraints that are analogous to the strengthened constraints in [28]. For example, the constraint that the probability of terminating in undesirable state is below a threshold *for all states* is studied in [29, 30]. It is required that the constraint must be satisfied *for* *every sample path* in [32, 55]. These hard constraints directly imply the non-decreasing property. The above-mentioned [25] also study the hard constraint that requires that the probability of reaching the goal is 1 (i.e. for every sample path). In this case, the constraint has the non-decreasing property.

Another class of constraints having the non-decreasing property is studied in [66]. They require that the maximum possible total cost is below a threshold. This is equivalent to $\mathbf{1}\{\max(X^\pi) \leq \delta\}$ that we have studied in Section 4.2 and hence has the non-decreasing property.

There also exists work that places constraints on policies. For example, Dolgov and Durfee [19] limit the search space to the set of deterministic policies when the optimal policy is in general randomized. This type of a "constraint" is not considered to be a constraint in this paper. That "constraint" simply defines the set, Π, of candidate policies, and the Π does not change over time. Therefore, the corresponding $\mathtt{MDP}_{f,h}$ is consistent unless there are other constraints. Abe et al. [1] study "constraints" that require that the expected cost with respect to a given probability distribution over states and the probability distribution of actions specified by a policy is below a prespecified level. That "constraint" is also considered to define the set of candidate policies, which stays unchanged over time. Hence, the corresponding $\mathtt{MDP}_{f,h}$ is consistent.

There are other "constraints" that are studied as constraints but do not fall into the class of our constraints. These "constraints" also directly limit the space of possible policies. For example, temporal constraints and precedence constraints are studied in [9, 10]. A temporal constraint requires that an action must be executed in a given time window. A precedence constraint requires that some actions must be completed before an action can be taken. These "constraints" can be taken into account with the definitions of states, actions, and transitions. Hence the $\mathtt{MDP}_{f,h}$ having these "constraints" is consistent unless there are other constraints. Becker et al. [6] consider soft temporal constraints, where taking an action makes the cost of other actions low or high. These soft constraints can also be taken into account with the definitions of states, action, and transitions, and are not considered to be constraints in this paper.

5.3. Risk-sensitive Markov decision processes

Expected utility is widely considered to be the objective function for rational decision making [56] and so is well studied as the objectives of MDPs [7]. The standard utility function is an exponential function [12, 16, 35, 37, 39]. Piecewise-linear utility function is studied in [41]. Geibel [27] minimizes the probability of being absorbed into a fatal state, which can be represented as the expectation of an indicator random variable (i.e., an expected utility). Xu and Mannor [65] study probabilistic goal of maximizing the probability that the total reward exceeds a given threshold, which again can be represented as the expectation of an indicator random variable. One can show that the expected utility is generally optimality-consistent. Notice that, in [41, 65], the state space is augmented so that an action can depend on the cumulative reward that is obtained by the time the action is taken.

Researchers have also investigated objective functions that cannot be represented as expected utility. We have already seen an example, variance, in Section 5.1. Kawai [36] minimizes variance without constraints. White [64] surveys MDPs, where "principle of optimality fails" (Page 4 from [64]) or "no stationary optimal policy exists" (Page 4 from [64]). These statements suggest that their objective functions are not optimality-consistent (or not time-consistent) or their constraints do not have the non-decreasing property.

The worst possible cumulative cost is minimized in [17, 33, 40]. This is the case where (backward) dynamic programming can find the optimal policy that stays optimal over time, even though the objective function is not optimality-consistent. Figure 2 shows an MDP, represented by an AND/OR graph [44], that illustrates this point. There are two candidate policies: one chooses action a_{20} from state s_2, and the other chooses a_{21}. The cost, C, is associated with the actions from s_1 and s_2. Either policy is optimal from s_0, because the worst possible cumulative cost is 4, regardless of the action from s_2. However, if we transition to s_2, which can happen with probability 0.5, the policy of choosing a_{20} becomes suboptimal. Dynamic programming will, however, find the policy that chooses a_{21} from s_2. However, there can be a wide range of algorithms,

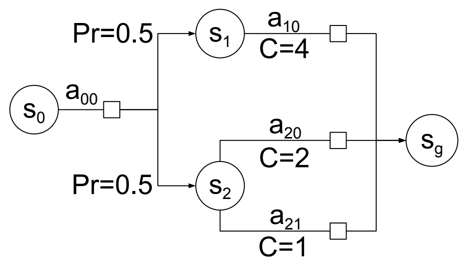


Figure 2: An example of an MDP, represented as an AND/OR graph, that illustrates that worst possible cumulative cost is not optimality-consistent.

including policy iteration, for MDPs [44], and some of these algorithms might find the policy that is optimal at one moment but will become suboptimal.

Recently, Ruszczyński [57] studies dynamic programming for an MDP whose objective is a Markov RM, a particular iterated RM. Osogami [48] studies dynamic programming for an MDP whose objective is a particular iterated RM when the future cost is discounted. Petrik and Subramanian [51] studies an MDP whose objective is a particular iterated RM for the case where the state space and the action space are continuous. However, these iterated RMs require to satisfy conditions that are not needed for optimality-consistency and no constraints are considered in [57, 48, 51].

As we have already seen at the end of Section 5.1, some of the prior work study the performance at the steady state. For example, Filar et al. [22] studies the expected reward minus the variance of the reward at the steady state. Although the objective function is not optimality-consistent, the optimal policy stays optimal as long as the MDP is a uni-chain.

There also exists a large body of the literature on risk-sensitive reinforcement learning [4, 14, 45, 46, 47, 50] where objective functions are not explicitly given, but learning algorithms are designed with the hope that learning agents can avoid large loss.

6. Conclusion

As we have seen in Section 5, risk-sensitive objective functions and constraints have been studied extensively in the literature of MDPs, where minimizing cumulative cost is sometimes found inadequate to avoid large loss or severe damage. However, implications of the use of these risk-sensitive objective functions and constraints have not been well understood. We show that the optimal policy can depend on the initial state and thus can change over time, and following the latest optimal policy at every time step can lead to poor results. This is in contrast to the standard MDP, where the optimal policy at one moment is guaranteed to stay optimal over time (i.e., the standard MDP is consistent). The nonstationarity of the optimal policy has been reported for example in [32, 34, 42, 55, 58, 64] for particular risk-sensitive objective functions and constraints, but our systematic study is new.

To formally study the stationarity of the optimal policies over time, we have defined the consistency of $\mathtt{MDP}_{f,h}$. We have provided the sufficient conditions for $\mathtt{MDP}_{f,h}$ to be consistent (Theorem 4.1). Namely, $\mathtt{MDP}_{f,h}$ is consistent if the objective function is an optimality-consistent dynamic RM and the constraints are given by a non-decreasing indicator dynamic RM. We have shown the necessity of these sufficient conditions (Lemma 4.1).

The consistency of optimality of a plan has been discussed primarily in deterministic settings. In particular, Strotz [62] shows that future cost should be discounted exponentially for the consistency of optimality. Sozou [61] provides an argument that hyperbolic discounting can be made consistent when uncertainty is involved. There is large body of the literature on how to plan when the optimal plan changes over time [52]. Such planning is important to understand how humans make decisions [24] but leads to suboptimal decisions [38].

Optimality-consistency is closely related to but different from the time-consistency that has been studied in [5, 8, 11, 23, 31, 53, 54, 57, 59]. In the context of MDPs, time-consistency is important primarily because the optimal policy then satisfies the Bellman equation. In general, time-consistency does not imply optimal-consistency, and vice versa. In particular, we have shown that ICTE studied in [31, 48, 49] is time-consistent but not optimality-consistent.

We have also developed specific classes of objective functions and constraints that one can use to define a consistent $\mathtt{MDP}_{f,h}$. In particular, the iterated RMs that satisfy the conditions of Corollary 2 are optimality-consistent. These iterated RMs are also shown to be time-consistent (Proposition 1). Our results thus provide a strong incentive to choose an objective function from this class of iterated RMs. We have established a general method for converting the constraints that do not have the non-decreasing property into the one that we can use to construct a consistent $\mathtt{MDP}_{f,h}$ (Corollary 3). Such constructed constraints are generally stronger than the original constraints and prevents the optimal policy to become infeasible by making a policy that becomes infeasible to be infeasible from the beginning.

An interesting future direction is to numerically investigate the impact of the inconsistency of optimal policies in real tasks. How often does the optimal policy change? How much do we lose (or gain) by following the initial optimal policy after it becomes suboptimal or by following the latest optimal policy at every step? How much can we gain by making decisions based on a consistent $\mathtt{MDP}_{f,h}$ relative to the one without consistency? Our results do not yet provide quantitative answers to these questions.

Acknowledgments

A part of this research was supported by JST, CREST.

References

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

[2] G. Ainslie. *Breakdown of Will*. Cambridge University Press, Cambridge, UK, first edition, 2001.

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

[4] C. W. Anderson, P. M. Young, M. R. Buehner, J. N. Knight, K. A. Bush, and D. C. Hittle. Robust reinforcement learning control using integral quadratic constraints for recurrent neural networks. *IEEE Transactions on Neural Networks*, 18(4):993–1002, 2007.

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

[6] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research*, 22(1):423–455, 2004.

[7] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientifc, 1995.

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

[9] A. Beynier and A.-I. Mouaddib. A polynomial algorithm for decentralized Markov decision processes with temporal constraints. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005)*, pages 963–969, 2005.

[10] A. Beynier and A.-I. Mouaddib. An iterative algorithm for solving constrained decentralized Markov decision processes. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, pages 1089–1094, 2006.

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

[12] V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311, 2002.

[13] P. Boyle, M. Hardy, and T. Vorst. Life after VaR. *The Journal of Derivatives*, 13:48–55, 2005.

[14] P. Campos and T. Langlois. Abalean: A risk-sensitive approach to self-play learning in Abalone. In *Proceedings of the 14th European Conference on Machine Learning (ECML 2003)*, pages 35–46, 2003.

[15] K.-J. Chung. Mean-variance tradeoffs in an undiscounted MDP: The unichain case. *Operations Research*, 42(1):184–188, 1994.

[16] K.-J. Chung and M. Solbel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

[17] S. Coraluppi and S. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35:301–309, 1999.

[18] D. Dolgov and E. H. Durfee. Approximating optimal policies for agents with limited execution resources. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1107–1112, 2003.

[19] D. Dolgov and E. H. Durfee. Constructing optimal policies for agents with constrained architectures. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2003)*, pages 974–975, 2003.

[20] E. A. Feinberg and A. Shwartz. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21:922–945, 1996.

[21] E. A. Feinberg and A. Shwartz. Constrained dynamic programming with two discount factors: Applications and an algorithm. *IEEE Transactions on Automatic Control*, 44(3):628–631, 1999.

[22] J. A. Filar, L. C. M. Kallenberg, and H.-M. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

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

[24] S. Frederick, G. Loewenstein, and T. O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40:351–401, 2002.

[25] M. S. Fulkerson, M. L. Littman, and G. A. Krim. Speeding safely: Multi-criteria optimization in probabilistic planning. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI 97)*, July 1997.

[26] Z. Gabor, Z. Kalmar, and C. Szepesvari. Multi-criteria reinforcement learning. In *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, pages 197–205, 1998.

[27] P. Geibel. Reinforcement learning with bounded risk. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 162–169, 2001.

[28] P. Geibel. Reinforcement learning for MDPs with constraints. In *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*, pages 646–653, September 2006.

[29] P. Geibel. *Risk-Sensitive Approaches for Reinforcement Learning*. Shaker-Verlag, 2006.

[30] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24(1):81–108, 2005.

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

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

[33] M. Heger. Consideration of risk in reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML 1994)*, pages 105–111, 1994.

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

[35] R. Howard and J. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18:356–369, 1972.

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

[37] S. Koenig and R. G. Simmons. Risk-sensitive planning with probabilistic decision graphs. In *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning (KR 1994)*, pages 363–373, 1994.

[38] F. E. Kydland and E. C. Prescott. Rules rather than discretion: The inconsistency of optimal plans. *The Journal of Political Economy*, 85:473–492, 1977.

[39] Y. Lin, R. Goodwin, and S. Koenig. Risk-averse auction agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2003)*, pages 353–360, 2003.

[40] M. L. Littman and C. Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 310–318, 1996.

[41] Y. Liu and S. Koenig. Risk-sensitive planning with one-switch utility functions: Value iteration. In *Proceedings of the 20th AAAI Conference on Artificial Intelligence*, pages 993–999, 2005.

[42] S. Mannor and J. N. Tsitsiklis. Mean-variance optimization in Markov decision processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[43] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[44] Mausam and A. Kolobov. *Planning with Markov Decision Processes: An AI Perspective.* Morgan & Claypool Publishers, 2012.

[45] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290, 2002.

[46] J. Morimoto and K. Doya. Robust reinfocement learning. *Neural Computation*, 17(2):335–359, 2005.

[47] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 368–375, 2010.

[48] T. Osogami. Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 567–574, July 2011.

[49] T. Osogami. Robustness and risk-sensitivity in markov decision processes. In *Advances in Neural Information Processing Systems 25*, pages 233–241, December 2012.

[50] T. J. Perkins. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 3:803–832, 2002.

[51] M. Petrik and D. Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, pages 805–814, 2012.

[52] R. A. Pollak. Consistent planning. *The Review of Economic Studies*, 35:201–208, 1968.

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

[54] B. Roorda, J. M. Schumacher, and J. Engwerda. Coherent acceptability measures in multiperiod models. *Mathematical Finance*, 15:589–612, 2005.

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

[56] S. J. Russell and D. Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2(1):575–609, 1995.

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

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

[59] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lecture Notes on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[60] M. J. Sobel. Mean-variance tradeoffs in an undiscounted MDP. *Operations Research*, 42(1):175–183, 1994.

[61] P. D. Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society B Biological Sciences*, 265(1409):2015–2020, 1998.

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

[63] F. Teichteil-Königsbuch. Path-constrained Markov decision processes: Bridging the gap between probabilistic model-checking and decision-theoretic planning. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 744–749, August 2012.

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

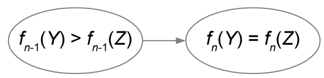
[65] H. Xu and S. Mannor. Probabilistic goal Markov decision processes. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 2046–2052, 2011.

[66] W.-L. Yeow, C.-K. Tham, and W.-C. Wong. Hard constrained semi-Markov decision processes. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, pages 549–554, 2006.

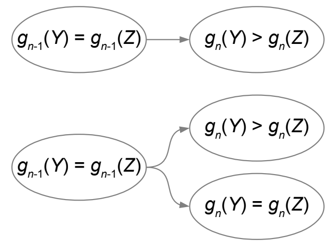
Appendix A. Appendix: Optimality-consistency and time-consistency

Our definition of optimality-consistency (Definition 4.3) is different from time-consistency that has been studied in the literature [5, 11, 23, 31, 53, 54, 57, 59]. In our context, a time-consistent dynamic RM can be defined as follows:

Definition Appendix A.1. *A dynamic RM, $\rho \equiv \{\rho_n \mid n \in [0, N)\}$, is called time-consistent if the following is satisfied for any $n \in [1, N)$:* $\Pr(\rho_n(Y) \leq \rho_n(Z)) = 1$ *implies* $\Pr(\rho_{n-1}(Y) \leq \rho_{n-1}(Z)) = 1$ *for any $\mathcal{F}_N$-measurable random variables, Y and Z.*



(a) Optimality-consistent dynamic RM



(b) Time-consistent dynamic RM

Figure A.3: Characteristic transitions with (a) an optimality-consistent dynamic RM and (b) a time-consistent dynamic RM.

The time-consistency of a dynamic RM does not imply the optimality-consistency of the dynamic RM, and vice versa. For example, iterated conditional tail expectation studied in [31, 48, 13] is a dynamic RM that is time-consistent but not optimality-consistent.

Figure A.3 illustrates the difference between optimality-consistency and time-consistency. Specifically, Figure A.3 (a) illustrates the transition that characterizes optimality-consistency. Here, $f_{n-1}(Y) > f_{n-1}(Z)$ for an optimality-consistent dynamic RM, f, and random variables, Y and Z, at a state. From that state, we can transition to a state with $f_n(Y) = f_n(Z)$ with probability 1. This transition is not allowed when f is time-consistent. Figure A.3 (b) shows two of the transitions that are allowed with a time-consistent dynamic RM, g. These transitions are not allowed when g is optimality-consistent.

When the transition of Figure A.3 (a) is possible, the Bellman equation (Equations 3.3 from [44]) is violated. The Bellman equation allows one to find the optimal policy for a finite-horizon MDP ($N < \infty$) through backward induction, which implies that we cannot have $f_{n-1}(Y) > f_{n-1}(Z)$ if $f_n(Y) = f_n(Z)$ surely. If time-consistency is also desirable, one can explicitly require that the objective function should be both optimality-consistent and time-consistent, which we further discuss in Section 4.3 (in particular, see Proposition 1).