Journal of the Operations Research Society of China
https://doi.org/10.1007/s40305-023-00457-5

Double-Factored Decision Theory for Markov Decision Processes with Multiple Scenarios of the Parameters

Cheng-Jun Hou[1]

Received: 16 March 2022 / Revised: 20 December 2022 / Accepted: 2 March 2023

Abstract

The double-factored decision theory for Markov decision processes with multiple scenarios of the parameters is proposed in this article. We introduce scenario belief to describe the probability distribution of scenarios in the system, and scenario expectation to formulate the expected total discounted reward of a policy. We establish a new framework named as double-factored Markov decision process (DFMDP), in which the physical state and scenario belief are shown to be the double factors serving as the sufficient statistics for the history of the decision process. Four classes of policies for the finite horizon DFMDPs are studied and it is shown that there exists a double-factored Markovian deterministic policy which is optimal among all policies. We also formulate the infinite horizon DFMDPs and present its optimality equation in this paper. An exact solution method named as double-factored backward induction for the finite horizon DFMDPs is proposed. It is utilized to find the optimal policies for the numeric examples and then compared with policies derived from other methods from the related literatures.

Keywords Dynamic programming · Markov decision process · Parameter uncertainty · Multiple scenarios of the parameters · Double-factored Markov decision process

Mathematics Subject Classification 90C39 · 90C40

This research was originally supported by the (United States) National Science Foundation (No. 1409214).

✉ Cheng-Jun Hou
chengjun.hou@gmail.com

[1] Software Research and Data Science, Amazon Robotics, North Reading, MA 01864, USA

Published online: 15 May 2023

1 Introduction

The Markov decision process (MDP) is an effective tool for modeling and decision-making in a dynamic environment, which are used in many disciplines such as robotics, automatic control, economics, manufactures and so on [1–5]. MDPs consider the intrinsic uncertainty which is realized by the uncertainty among its future transition. But the indeterminacy of the parameters (transition probability and reward) in the MDP is source of uncertainty that is not well addressed in many MDP problems. What's more, in reality it imposes significant difficulty in obtaining a precise representation of the MDP parameters. There are various reasons, such as (i) imprecise or conflicting information given by the content experts [6]; (ii) insufficient data from which to build structures for parameter estimation [7]; (iii) non-stationary transition probabilities due to insufficient and/or hidden state information [7]; and (iv) unpredictable but prevailing events impacting the system [8].

There are generally two categories of parameter uncertainty in the MDPs. Firstly, either or both of the parameters are unknown completely. For handling such problems, the theory of Bayes-adaptive Markov decision process (BAMDP) is established by utilizing the concept of Bayesian inference [9, 10]. In a BAMDP, prior information regarding the unknown parameters is represented by a parameterized distribution and Bayesian inference is used to incorporate any new information in order to update the distribution so that the problem of exploration–exploitation is addressed during learning and sampling. [11] builds a rigorous framework rooted in information theory for solving the MDPs with unknown transition probabilities. But these theories are very difficult to implement on real-world problems. Secondly, the parameters lie in a given range, in another words, there is a pre-defined uncertainty set that the parameters belong to. In this context, there is the more traditional approach of mitigating the parameter uncertainty in MDPs, known as *robust dynamic programming*. The standard robust dynamic programming is a "max–min" approach in which the decision maker seeks to find a policy that maximizes the worst-case performance when the transition probabilities are allowed to vary within an uncertainty set [6, 12–15]. One of the key results is that the max–min problem is tractable for instances that satisfy the rectangularity property [12, 13]. Essentially, rectangularity means that observing the realization of a transition probability parameter gives no information about the values of other parameters for any other state-action-time triplet. Because each parameter value for any given state-action-time triplet is independent to each other, the problem can be decomposed so that each worst-case parameter is found via an optimization problem called the inner problem. [12, 13] provide algorithms for solving the max–min problem for a variety of uncertainty sets by providing polynomial-time methods for solving the corresponding inner problem. While rectangular uncertainty sets are desirable from a computational perspective, they can give rise to policies that are overly-conservative.

Recently, there has been a line of research on mitigating the impact of the parameter uncertainty by incorporating multiple scenarios of the parameters into the solution of the MDP. The parameter uncertainty in MDPs is encoded by allowing for multiple scenarios of the rewards and transition probabilities. A scenario is one possible realization of uncertainty and may result from possible realizations of a system, or some observation of a system, or from a simulation model. So this model clearly falls into the second category of parameter uncertainty above. Each scenario is referred to as a "model" of the MDP in [4], where all models are defined on the same state space, action space, and set of decision epochs, but each model may vary in terms of its rewards and transition probabilities. Such an MDP is known as the *multi-model Markov decision process* in [4] and the *concurrent Markov decision process* in [5]. We refer to it as the *Markov decision process with multiple scenarios of the parameters* or the *multi-scenario MDP*. [4, 5, 16, 17] propose designing policies that maximize a weighted value across multiple scenarios. [18] proposes minimizing the maximum regret with respect to each scenario for the finite horizon multi-scenario MDPs. [17] and [16, 19] propose a percentile optimization approach for finite horizon and infinite horizon multi-scenario MDPs, respectively. So far, the exact solution methods for this problem have relied on mixed-integer programming (MIP) formulations where binary decision variables encode the policy and continuous variables encode the value functions for each scenario of the multi-scenario MDP [4]. Although the MIP can be used to find exact solutions, these problems are NP-hard [4, 17], and the MIP solution methods for this problem have been limited to small multi-scenario MDPs.

The invention of multi-scenario MDPs is motivated by parameters estimated with statistical uncertainty [4], sometimes the results could be dramatically different or even conflicting with each other, especially in the field of clinical/medical research, where MDPs has been successfully used to design optimal treatment and screening protocols, however, longitudinal observation data used to characterize the MDP model are often very limited due to the patient population and cost of acquisition. In this context, there is the need for designing policies that is general favorable for several parameter possibilities, known as model or scenario among related literatures. Moreover, we believe when actually implementing the policies to an individual case in practice, there is information hidden under the realization of state transitions that could be utilized to make inference on the individual itself, on how likely the individual is following each scenario. And with an initial set of weights or distribution of the scenarios, the policies is developed considering all future possibilities so that it's optimal even before any information and action being collected or implemented.

So in this article, we propose a new theoretical framework for the weighted value problem of the multi-scenario MDPs and methods solving for its exact solution. We introduce the concepts of *scenario belief* and the *scenario expectation* to formulate the expected total discounted reward of a policy, and establish a new framework named as *double-factored Markov decision process* (DFMDP). The "double-factored" here are two factors, i.e., the physical state and the scenario belief of the system, that work as sufficient statistics for the past history of observations and actions in the multi-scenario MDPs. We will demonstrate that these two factors summarize all the prior information gained up to the current stage. Thus they provide complete information for future decision-making. We analyze four classes of policies for the finite horizon DFMDPs. They are respectively *double-factored history-dependent randomized*, *double-factored history-dependent deterministic*, *double-factored Markovian randomized* and *double-factored Markovian deterministic policies*. It is shown that there exists a double-factored Markovian deterministic policy which is optimal among all policies above. An exact solution method for the double-factored Markovian deterministic policies is proposed. We also formulate the infinite horizon DFMDPs in a later section, but the focus of this paper is still on finite horizon problems. With this framework, we are able to develop algorithms that can be scaled to real-world problems, solving for the exact solutions to the identical weighted value problem of the multi-scenario MDPs, with a side benefit of keeping updating the scenario distribution based on all history observations and actions. So other than solving for the weighted value problem which is our primary goal, this framework can also be extended and utilized to solve learning or recognition problems.

Other than clinical justification where there exists multiple well-establish authorities or studies to build up the MDP parameters, in the context of E-commerce where MDP-based recommendation system has been utilized [20], there is ample amount of data to characterize the whole customer population into sub-category based on census data and shopping behaviors. When a new customer shops online, we can make tailored recommendation when we have more confidence on the sub-category the customer fits, based on his/her history reactions. Yet before the customer performs any actions, we could still make recommendation considering an initial belief or weights on all sub-categories that is benefiting the expectation of the objectives measured via either sales or click-through rate.

This article is organized as follows. In Sect. 2, the definition and notation of the multi-scenario MDPs are presented. In Sect. 3, the finite horizon and the infinite horizon DFMDPs are formulated. Four classes of policies are discussed and an exact solution method is proposed for the finite horizon DFMDPs. The computational experiments involving three sets of test instances for comparing several solution methods for the finite horizon DFMDPs are illustrated in Sect. 4. Section 5 summarizes the work presented in this article and shares plans for future research on the relevant topics.

2 Background and Notation

2.1 Standard MDP

We denote the standard MDP by a 5-elements-tuple (T, S, A, ($\boldsymbol{p}$,$\boldsymbol{r}$), s_0) where $T = \{0, 1, \cdots, Z-1\}$ is the finite discrete set of decision epoch, S is the finite state set, A is the finite action set, ($\boldsymbol{p}$,$\boldsymbol{r}$) is the *parameter pair* in which $\boldsymbol{p}$ is the $|S| \times |S| \times |A|$ transition probability matrix whose element $p^a_{s,s'}$ is the probability of ending in state $s' \in S$ if the system performs action $a \in A$ in state $s \in S$, and $\boldsymbol{r}$ is the $|S| \times |S| \times |A|$ reward matrix whose element $r^a_{s,s'}$ is the reward obtained by ending in state $s' \in S$ if the system performs action $a \in A$ in state $s \in S$, and $s_0 \in S$ is the initial state. Moreover, we assume that the *terminal reward* is denoted by $\boldsymbol{r}^0 = \left[r^0_1, \cdots, r^0_{|S|}\right]^\top$ and the *discounted factor* by $\gamma \in [0, 1]$.

Decision epoch t is also called *time t*. The interval between two successive decision epochs is called *period*. We consider the Markov deterministic policy $\pi = (x_0, x_1, \cdots, x_{Z\text{-}1}) \in \Pi$ where x_t: $S \rightarrow A$ is the decision rule at epoch $t \in T$, which assign one action $a_t = x_t\ (s_t)$ to each state $s_t \in S$, and Π is the policy set. The goal of the decision maker is to specify a policy $\pi \in \Pi$ that maximize the expected total discounted rewards over the planning horizon [1]

$$\max_{\pi\in\Pi} \mathrm{E}_{s_0}^{\pi}\left\{\sum_{t=0}^{Z-1}\gamma^t r_{s_t}^{a_t}+\gamma^Z r_{s_Z}^0\right\}, \forall s_0\in S, \tag{1}$$

where $r_s^a = \sum_{s'\in S} p_{s,s'}^a r_{s,s'}^a$. The optimal policy of (1) can be found by the backward induction algorithm in [1].

The Bellman's equations for the infinite horizon MDP are

$$V^*(s)=\max_{a\in A}\left\{r_s^a+\gamma\sum_{s'\in \mathrm{S}} p_{s,s'}^a V^*(s')\right\}, \forall s\in S, \tag{2}$$

where $V^*(s)$ is the optimal value function of state s for the infinite horizon MDP. The optimal stationary and deterministic policy of this problem can be found by the value iteration or the policy iteration in [1].

2.2 Multi-Scenario MDP

We denote the multi-scenario MDP by a 6-elements-tuple $(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0}\rangle)$ where T, S and A are following the same definitions above, C is the finite discrete set of scenarios, $J_{\boldsymbol{p},\boldsymbol{r}} = \{(\boldsymbol{p}_1, \boldsymbol{r}_1),\cdots,(\boldsymbol{p}_{|C|}, \boldsymbol{r}_{|C|})\}$ is the *parameter pair set* in which $(\boldsymbol{p}_k, \boldsymbol{r}_k)$ is the *parameter pair of scenario* $k\in C$ where the elements in $\boldsymbol{p}_k$ are denoted by $p_{s,\mathrm{s}',k}^a$ with the same definitions as $p_{s,\mathrm{s}'}^a$ above, and the elements in $\boldsymbol{r}_k$ are denoted by $r_{s,\mathrm{s}',k}^a$ with the same definitions as $r_{s,\mathrm{s}'}^a$ above, $\langle s_0, \boldsymbol{b}_{s_0}\rangle$ is the ordered pair consisting of the initial state $s_0\in S$ and the initial probability distribution of scenarios, $\boldsymbol{b}_{s_0} = [b_{s_0,1},\cdots,b_{s_0,|C|}]^\top \in \Delta^C$, where Δ^C denotes the set of probability distributions on set C.

For easy of description, we refer to the standard MDP with the parameter pair $(\boldsymbol{p}_k, \boldsymbol{r}_k)$, $k\in C$, as the MDP k as in [4]. We use $r_{s,k}^a = \sum_{s'\in S} p_{s,s',k}^a r_{s,s',k}^a$ to denote the immediate expected reward in the MDP k when action $a \epsilon A$ is taken in state $s\in S$ and $\boldsymbol{r}_k^0 = [r_{1,k}^0,\ldots,r_{|S|,k}^0]^\top$ to denote the terminal reward in the MDP k.

Given a multi-scenario MDP $(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0}\rangle)$, the value of a policy $\pi\in\Pi$ in MDP k, for a sepcific $k\in C$, is given by its expected total discounted rewards evaluated with the parameter pair $(\boldsymbol{p}_k, \boldsymbol{r}_k)$:

$$V_k^\pi(s_0)=E_{s_0}^\pi\left[\sum_{t=0}^{Z-1}\gamma^t r_{s_t,k}^{a_t}+\gamma^Z r_{s_Z,k}^0\right]. \tag{3}$$

The weighted value of any policy $\pi\in\Pi$ in the multi-scenario MDP is defined by

$$W^\pi(s_0,b_{s_0})=\sum_{k\in C} b_{s_0,k}V_k^\pi(s_0)=\sum_{k\in C} b_{s_0,k}E_{s_0}^\pi\left[\sum_{t=0}^{Z-1}\gamma^t r_{s_t,k}^{a_t}+\gamma^Z r_{s_Z,k}^0\right], \tag{4}$$

and the weighted value problem (WVP) is defined as the problem of finding a solution to

$$W^*\left(s_0, b_{s_0}\right) = \max_{\pi \in \Pi} W^{\pi}\left(s_0, b_{s_0}\right) = \max_{\pi \in \Pi}\left\{\sum_{k \in C} b_{s_0,k} E^{\pi}_{s_0}\left[\sum_{t=0}^{Z-1} \gamma^t r^{a_t}_{s_t,k} + \gamma^Z r^0_{s_Z,k}\right]\right\}. \tag{5}$$

We also have the set of policies $\Pi^* = \{\pi\colon W^{\pi} = W^*\} \subseteq \Pi$ that achieve the maximum in (4) (see [4]).

Ref. [4] proposes two approximate methods and the MIP formulation for Eq. (5) and show that the WVP for the Markov deterministic policy class is a NP-hard problem.

3 Double-Factored Markov Decision Process

In this section, a novel framework named as double-factored Markov decision process (DFMDP) is formulated for the WVP of the multi-scenario MDP. A backward induction algorithm is proposed for solving the exact solutions of the DFMDPs.

3.1 Scenario Belief and Its Update

The weights in (5) are one's estimate of the probabilities that scenarios occur in the multi-scenario MDPs. We believe that $\boldsymbol{b}_{s_0}$ is only the initial probability distribution of the scenarios and it will change with the system evolution. In fact, system's state transitions under specific actions contain the information related to scenario probabilities. The decision makers can utilize the information to update the scenario probabilities according to the observation of state transitions when the system evolves. For that reason, we introduce the *scenario belief* $\boldsymbol{b}_t$ as follows to describe the scenario probability distribution at time t:

$$\boldsymbol{b}_t = [b_{t,1}, \cdots, b_{t,k}, \cdots, b_{t,|C|}]^\top, b_{t,k} \geqslant 0, \forall k \in C \quad \text{and} \quad \sum_{k \in C} b_{t,k} = 1.$$

In order to derive the update function for the scenario belief, we define g_t as the total available information in the multi-scenario MDP at decision epoch t. Since a_{t-1} denotes the action implemented at decision epoch $t - 1$ and s_t the state observed at decision epoch t, we have

$$g_t = (s_t, a_{t-1}, g_{t-1}). \tag{6}$$

Then the elements of the scenario belief is defined as

$$b_{t,k} = \Pr(c_k | g_t), \forall k \in C, \tag{7}$$

where c_k denotes the event that the realization of system's scenario is k. The substitution of (6) into (7) and the application of Bayes' rule yields

$$b_{t,k} = \Pr(c_k|s_t, a_{t-1}, g_{t-1}) = \frac{\Pr(s_t|c_k, a_{t-1}, g_{t-1})\Pr(c_k|a_{t-1}, g_{t-1})}{\Pr(s_t|a_{t-1}, g_{t-1})}, \forall k \in C. \quad (8)$$

In (8), the first probability in the numerator is the transition probability of scenario k, i.e., $\Pr(s_t|c_k, a_{t-1}, g_{t-1}) = p^{a_{t-1}}_{s_{t-1},s_t,k}$ since $g_{t-1} = (s_{t-1}, a_{t-1}, g_{t-2})$. The prior probability $\Pr(c_k|a_{t-1}, g_{t-1})$ is independent of a_{t-1} given g_{t-1}, since choosing for action is completely under our control at time $t-1$, so $\Pr(c_k|a_{t-1}, g_{t-1}) = \Pr(c_k|g_{t-1}) = b_{t-1,k}$. And the denominator of the equation is just the sum of the numerator over all possible value of k representing each scenario. So we have

$$b_{t,k} = \frac{b_{t-1,k}\, p^{a_{t-1}}_{s_{t-1},s_t,k}}{\sum_{k'\in C} b_{t-1,k'}\, p^{a_{t-1}}_{s_{t-1},s_t,k'}}, \forall k \in C. \quad (9)$$

We can see in (9) that $\boldsymbol{b}_t$ is a function of s_t, s_{t-1}, $\boldsymbol{b}_{t-1}$ and a_{t-1}. So we use $\boldsymbol{b}_{s_t}$ to denote the scenario belief when state s_t is observed at time t, in order to differentiate it from other states being observed, also the time index t only shows as subscript of s_t to simplify the notation. With these in mind, Eq. (9) can be rewritten as

$$b_{s_t,k} = \frac{b_{s_{t-1},k}\, p^{a_{t-1}}_{s_{t-1},s_t,k}}{\sum_{k'\in C} b_{s_{t-1},k'}\, p^{a_{t-1}}_{s_{t-1},s_t,k'}}, \forall a_{t-1} \in A, s_{t-1}, s_t \in S, k \in C. \quad (10)$$

Equation (10) can also be expressed as a function that $\boldsymbol{b}_{s_t} = \tau\big(s_{t-1}, \boldsymbol{b}_{s_{t-1}}, a_{t-1}, s_t\big)$ where $\tau(\cdot)$ is called the *scenario belief update function*. We stipulate that $\boldsymbol{b}_{s_t} = \boldsymbol{0}$ if the denominator in (10) is equal to zero, implying the situation where state s_t is unreachable at time t. Clearly $\boldsymbol{b}_{s_t} \in \Delta^C$, so Δ^C is also called *scenario belief space*.

Stating from any state $s_0 \in S$ and scenario belief $\boldsymbol{b}_{s_0} \in \Delta^C$, the scenario belief $\boldsymbol{b}_{s_t}$ in state s_t at time t can be obtained by using (10) repeatedly if s_t is observed. We denote the combination of s_t and $\boldsymbol{b}_{s_t}$ by an ordered pair $\langle s_t, \boldsymbol{b}_{s_t}\rangle$, which is referred to as the *ordered pair of state and scenario belief at time t* or the *state-belief pair* for short. It follows that $\langle s_t, \boldsymbol{b}_{s_t}\rangle \in S \times \Delta^C$ where $S \times \Delta^C$ is the Cartesian production of S and Δ^C. In addition, the scenario belief update function could be written as $\boldsymbol{b}_{s_t} = \tau\big(\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle, a_{t-1}, s_t\big)$ to highlight the state-belief pair.

The important feature of (10) is that the calculation of $\boldsymbol{b}_{s_t}$ in state s_t at time t requires only the state-belief pair $\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle$ at time $t-1$; thus, $\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle$ summarizes all the information gained prior to time $t-1$ and represents a sufficient statistic for the complete past history of the process g_{t-1}. This important feature sets the foundation of the double-factored decision theory.

3.2 Scenario Expectation and Expected Total Reward

Starting with any $\langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$ and using (10) repeatedly, all state-belief pairs $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ and actions a_t for $t = 0, 1,\cdots, Z$ will form the *transition tree* of the system as shown in Fig. 1. Within the tree, there are usually $|A|$ actions behind each state-belief pair and $|S|$ state-belief pairs behind each action. Each $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ usually has $|A| \cdot |S|$ successors $\langle s_{t+1}, \boldsymbol{b}_{s_{t+1}} \rangle = \langle s_{t+1}, \tau(\langle s_t, \boldsymbol{b}_{s_t} \rangle, a_t, s_{t+1}) \rangle$. And apart from $\langle s_0, \boldsymbol{b}_{s_0} \rangle$, each $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ has one unique predecessor.

There are two types of special state-belief pairs in the transition tree. If $\boldsymbol{b}_{s_t} = \boldsymbol{e}_0$ in state s_t at some decision epoch t where $\boldsymbol{e}_0$ is a $|C|$-dimensional zero vector, the corresponding state-belief pair $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ is referred to as the *pseudo state-belief pair*. The reason for the appearance of $\langle s_t, \boldsymbol{e}_0 \rangle$ is that state s_t is unreachable. Thus, a pseudo-state-belief pair is a "leaf" in the tree and it no longer has any successor. If $\boldsymbol{b}_{s_t} = \boldsymbol{e}_k$ in state s_t at some decision epoch t where $\boldsymbol{e}_k$ is a $|C|$-dimensional unit vector with its kth element being one, the corresponding state-belief pair $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ is referred to as the *degenerate state-belief pair*.

Any path $(\langle s_0, \boldsymbol{b}_{s_0} \rangle, a_0, \langle s_1, \boldsymbol{b}_{s_1} \rangle, \cdots, a_{t-1}, \langle s_t, \boldsymbol{b}_{s_t} \rangle)$ along the transition tree will be the possible realization when the system evolves. Thus the history for $\left(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0} \rangle\right)$ is defined as the sequence of state-belief pairs and actions, i.e., $h_t = \left(\langle s_0, \boldsymbol{b}_{s_0} \rangle, a_0, \langle s_1, \boldsymbol{b}_{s_1} \rangle, \cdots, a_{t-1}, \langle s_t, \boldsymbol{b}_{s_t} \rangle\right)$, which is referred to as the *history up to time t from* $\langle s_0, \boldsymbol{b}_{s_0} \rangle$ *onward*, or the *history up to time t* for short. As a convention, $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ denotes the state-belief pair at time t when the history is h_t. The history h_t follows the recursion $h_t = \left(h_{t-1}, a_{t-1}, \langle s_t, \boldsymbol{b}_{s_t} \rangle\right)$ with $h_0 = \langle s_0, \boldsymbol{b}_{s_0} \rangle$. We let H_t denote the set of all possible history up to time t and call it *history set up to time t from* $\langle s_0, \boldsymbol{b}_{s_0} \rangle$ *onward*, or *history set up to time t* for short.

Similar to the standard MDPs, there are four classes of decision rules x_t for the DFMDPs. *History-dependent randomized decision rules* map the history set H_t into the probability distribution set Δ^A on action set A, i.e., $x_t : H_t \to \Delta^A$ or $\mu_{h_t}(\cdot) = x_t(h_t)$, $\mu(\cdot) \in \Delta^A$, $h_t \in H_t$. *History-dependent deterministic decision rules* map the history set H_t into the action set A, i.e., $x_t : H_t \to A$ or $a_t = x_t(h_t), h_t \in H_t$.

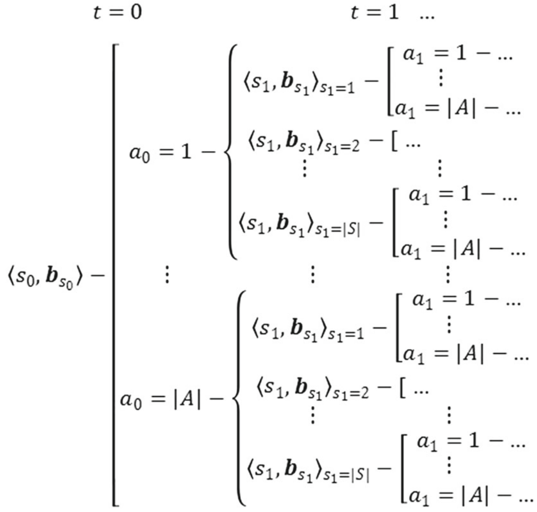
$t = 0$
$t = 1$ ...
$\langle s_0, \boldsymbol{b}_{s_0} \rangle$ −
$a_0 = 1$ −
$\langle s_1, \boldsymbol{b}_{s_1} \rangle_{s_1=1}$ −
$a_1 = 1$ − ...
$a_1 = |A|$ − ...
$\langle s_1, \boldsymbol{b}_{s_1} \rangle_{s_1=2}$ − [...
$\langle s_1, \boldsymbol{b}_{s_1} \rangle_{s_1=|S|}$ −
$a_0 = |A|$ −

Fig. 1 The transition tree

Double-factored Markovian randomized decision rules map $S \times \Delta^C$ into Δ^A, i.e., $x_t : S \times \Delta^C \to \Delta^A$ or $\mu_{\langle s_t, \boldsymbol{b}_{s_t} \rangle}(\cdot) = x_t\big(s_t, \boldsymbol{b}_{s_t}\big), \mu(\cdot) \in \Delta^A, \langle s_t, \boldsymbol{b}_{s_t} \rangle \in S \times \Delta^C$. *Double-factored Markovian deterministic decision rules* map $S \times \Delta^C$ into A, i.e., $x_t : S \times \Delta^C \to A$ or $a_t = x_t\big(s_t, \boldsymbol{b}_{s_t}\big), \langle s_t, \boldsymbol{b}_{s_t} \rangle \in S \times \Delta^C$. A double-factored Markovian deterministic decision rule specifies the action choice only considering the fact that the system is in state s_t with belief $\boldsymbol{b}_{s_t}$ at time t. We will prove later that it is Markovian (memoryless) because it depends on the previous history only through the current state-belief pair of the system, i.e., $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ is a sufficient statistic of the past history as mentioned above.

The policies composed of the above four classes of decision rules, $\pi = (x_0, x_1, \cdots, x_{Z-1})$, are referred to as *double-factored history-dependent randomized policies*, *double-factored history-dependent deterministic policies*, *double-factored Markovian randomized policies* and *double-factored Markovian deterministic policies*, respectively. And the sets of all policies of these classes are denoted by Π^{DHR}, Π^{DHD}, Π^{DMR}, and Π^{DMD}, respectively.

The relationship between the various classes of policies is as follows: $\Pi^{\text{DMD}} \subset \Pi^{\text{DMR}} \subset \Pi^{\text{DHR}}$ and $\Pi^{\text{DMD}} \subset \Pi^{\text{DHD}} \subset \Pi^{\text{DHR}}$. We will show later with a series of theorems that the optimal double-factored Markovian deterministic policies are the best among all classes of policies for the finite horizon DFMDPs.

We next use a one-period decision-making problem to introduce the concept of the *scenario expectations*. In a one-period problem, $Z = 1$ and $T = \{0\}$. We assume that whenever the system is in state s_1 at the end of this period, the decision maker receives a terminal reward $V(s_1)$, where V is a definite real-valued function on S. Suppose the system sits in state s_0 with scenario belief $\boldsymbol{b}_{s_0}$ at the start of the period. The decision maker aims to select an action $a \in A$ to maximize the expected total discounted reward. Suppose he chooses a deterministic policy $\pi = (x_0)$, which selects action $a \in A$ at initial decision epoch.

In this problem, the expected reward for the MDP k is

$$V_k^a(s_0) = r_{s_0,k}^a + \gamma \sum_{s_1 \in S} p_{s_0,s_1,k}^a V(s_1). \tag{11}$$

Since the scenario realization of the system is uncertain before the decision-making, we can consider the problem from the perspective of expectations over possible scenarios and define the expected total discounted reward as follows:

$$V^a\big(s_0, \boldsymbol{b}_{s_0}\big) = \sum_{k \in C} b_{s_0,k} V_k^a(s_0) = \sum_{k \in C} b_{s_0,k} \left[r_{s_0,k}^a + \gamma \sum_{s_1 \in S} p_{s_0,s_1,k}^a V(s_1) \right]. \tag{12}$$

Letting $\overline{r}_{s_0}^a = \sum_{k \in C} b_{s_0,k} r_{s_0,k}^a$ and $\overline{p}_{s_0,s_1}^a = \sum_{k \in C} b_{s_0,k} p_{s_0,s_1,k}^a$, (12) is rewritten as

$$V^a\big(s_0, \boldsymbol{b}_{s_0}\big) = \overline{r}_{s_0}^a + \gamma \sum_{s_1 \in S} \overline{p}_{s_0,s_1}^a V(s_1). \tag{13}$$

The optimal action a^* is obtained by finding the maximum of (13) over all actions in set A:

$$a^*\left(s_0, \boldsymbol{b}_{s_0}\right) \in \underset{a\in A}{\operatorname{argmax}}\left\{\overline{r}^a_{s_0} + \gamma \sum_{s_1\in S} \overline{p}^a_{s_0,s_1} V(s_1)\right\}.$$

If the decision maker uses a randomized policy with probability $\mu(a)$ to select actions in state s_0, the expected total discounted reward equals $\sum_{a\in A} \mu(a)\left\{\overline{r}^a_{s_0} + \gamma \sum_{s_1\in S} \overline{p}^a_{s_0,s_1} V(s_1)\right\}$, where $\sum_{a\in A} \mu(a) = 1$ and $\mu(a) \geqslant 0$ for $a \in A$. Since

$$\max_{\mu\in\Delta^A}\left\{\sum_{a\in A} \mu(a)\left[\overline{r}^a_{s_0} + \gamma \sum_{s_1\in S} \overline{p}^a_{s_0,s_1} V(s_1)\right]\right\} = \max_{a\in A}\left\{\overline{r}^a_{s_0} + \gamma \sum_{s_1\in S} \overline{p}^a_{s_0,s_1} V(s_1)\right\},$$

obviously we cannot obtain a larger expected reward in state s_0 by means of randomized policies.

We generalize the idea of scenario expectations above to the decision horizon of $\left(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0}\rangle\right)$ and use it to define the expected total discounted reward generated by a policy. Let $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ be the state-belief pair in state s_t at decision epoch t. We define

$$\begin{aligned} \overline{r}^{a_t}_{s_t} &= \sum_{k\in C} b_{s_t,k} r^{a_t}_{s_t,k}, \forall a_t \in A, \\ \overline{p}^{a_t}_{s_t,s_{t+1}} &= \sum_{k\in C} b_{s_t,k} p^{a_t}_{s_t,s_{t+1},k}, \forall a_t \in A, s_{t+1} \in S, \end{aligned} \tag{14}$$

where $\overline{r}^{a_t}_{s_t}$ is called the *scenario expectation reward* and $\overline{p}^{a_t}_{s_t,s_{t+1}}$ the *scenario expectation transition probability* in state s_t at decision epoch t. Let $V^{\pi}(s_0, \boldsymbol{b}_{s_0})$ represent the expected total discounted reward over the decision-making horizon if policy π is used and the system has an initial state-belief pair $\langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$. For $\pi \in \Pi^{\mathrm{DHR}}$, it is defined by

$$V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right) = E^{\pi}_{\langle s_0,\boldsymbol{b}_{s_0}\rangle}\left\{\sum_{t=0}^{Z-1} \gamma^t \overline{r}^{a_t}_{s_t} + \gamma^Z \overline{r}^0_{s_Z}\right\}, \tag{15}$$

where $\overline{r}^0_{s_Z} = \sum_{k\in C} b_{s_Z,k} r^0_{s_Z,k}$. For $\pi \in \Pi^{\mathrm{DHD}}$ and $\pi \in \Pi^{\mathrm{DMD}}$, the expected total discounted reward can be respectively expressed as

$$V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right) = E^{\pi}_{\langle s_0,\boldsymbol{b}_{s_0}\rangle}\left\{\sum_{t=0}^{Z-1} \gamma^t \overline{r}^{x_t(h_t)}_{s_t} + \gamma^Z \overline{r}^0_{s_Z}\right\}, \tag{16}$$

and

$$V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right) = E^{\pi}_{\langle s_0, \boldsymbol{b}_{s_0}\rangle}\left\{\sum_{t=0}^{Z-1}\gamma^{t}\overline{r}_{s_t}^{x_t(s_t,\boldsymbol{b}_{s_t})} + \gamma^{Z}\overline{r}_{s_Z}^{0}\right\}. \tag{17}$$

It will be shown later by Theorem 7 that $V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right)$ equals $W^{\pi}(s_0, \boldsymbol{b}_{s_0})$ in (4).

3.3 Finite Horizon DFMDP

According to the above theory, the optimization problem for a finite horizon DFMDP is expressed as

$$V^{*}\left(s_0, \boldsymbol{b}_{s_0}\right) = \max_{\pi\in\Pi^{\mathrm{DHR}}} V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right), \forall\, \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^{C}. \tag{18}$$

Now we provide a recursive algorithm to evaluate $V^{\pi}(s_0, \boldsymbol{b}_{s_0})$. Let $U_t^{\pi} : H_t^{\pi} \to \mathrm{R}$ denote the expected total discounted reward obtained by using a fixed policy π at decision epoch $t, t+1, \cdots, Z$, where H_t^{π} is the history set following the policy π up to time t and $H_t^{\pi} = H_t$ for $\pi \in \Pi^{\mathrm{DHR}}$. If the history at decision epoch t is $h_t \in H_t^{\pi}$, then we define $U_t^{\pi}(t < Z)$ by

$$U_t^{\pi}(h_t) = E^{\pi}_{h_t}\left\{\sum_{n=t}^{Z-1}\gamma^{n-t}\overline{r}_{s_n}^{a_n} + \gamma^{Z-t}\overline{r}_{s_Z}^{0}\right\}. \tag{19}$$

To simplify the notation, assume that a deterministic $\pi \in \Pi^{\mathrm{DHD}}$ has been specified. In practice, we will not need to evaluate randomized policies, because subsequent results establish that deterministic policies are optimal under the expected total reward criteria. For a given $\langle s_0, \boldsymbol{b}_{s_0}\rangle$, the corresponding H_t^{π} is defined by

$$\begin{aligned} &H_0^{\pi} = \{h_0\} = \left\{\langle s_0, \boldsymbol{b}_{s_0}\rangle\right\},\\ &H_t^{\pi} = \left\{h_t : h_t = \left(h_{t-1}, x_{t-1}(h_{t-1}), \langle s_t, \tau\left(\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle, x_{t-1}(h_{t-1}), s_t\right)\rangle\right), \forall\, h_{t-1} \in H_{t-1}^{\pi}, \forall\, s_t \in S\right\}. \end{aligned} \tag{20}$$

It is obvious that $H_t^{\pi} \subset H_t$ when $|A|> 1$. If the history at decision epoch t is $h_t \in H_t^{\pi}$, U_t^{π} is expressed by

$$U_t^{\pi}(h_t) = E^{\pi}_{h_t}\left\{\sum_{n=t}^{Z-1}\gamma^{n-t}\overline{r}_{s_n}^{x_n(h_n)} + \gamma^{Z-t}\overline{r}_{s_Z}^{0}\right\}. \tag{21}$$

```
The Policy Evaluation Algorithm (for fixed $\pi \in \Pi^{\mathrm{DHD}}$)
Step 1 For a given $\langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$ and a fixed $\pi \in \Pi^{\mathrm{DHD}}$, create history sets $H_t^\pi$ for $t = 0, 1, \cdots, Z$ by (20).
Step 2 Let $t = Z$ and $U_Z^\pi(h_Z) = \bar{r}_{s_Z}^0$ for all $h_Z = \left(h_{Z-1}, x_{Z-1}(h_{Z-1}), \langle s_Z, \boldsymbol{b}_{s_Z}\rangle\right) \in H_Z^\pi$.
Step 3 If $t = 0$, stop; otherwise go to step 4.
Step 4 Substitute $t-1$ for $t$ and compute $U_t^\pi(h_t)$ for each $h_t = \left(h_{t-1}, x_{t-1}(h_{t-1}), \langle s_t, \boldsymbol{b}_{s_t}\rangle\right) \in H_t^\pi$ by
    $U_t^\pi(h_t) = \bar{r}_{s_t}^{x_t(h_t)} + \gamma \sum_{s_{t+1}\in S} \bar{p}_{s_t,s_{t+1}}^{x_t(h_t)} U_{t+1}^\pi\left(h_t, x_t(h_t), \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle\right),$    (22)
    where $\langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle = \langle s_{t+1}, \tau(\langle s_t, \boldsymbol{b}_{s_t}\rangle, x_t(h_t), s_{t+1})\rangle$.
Step 5 Return to 3.
```

The following theorem guarantees that the expected value $U_0^\pi(h_0)$ generated by the algorithm above is equal to $V^\pi\left(s_0, \boldsymbol{b}_{s_0}\right)$ in (16).

Theorem 1 *For any given $h_0 = \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$ and a fixed policy $\pi \in \Pi^{\mathrm{DHD}}$, suppose U_t^π, $t \in T$, has been generated by the policy evaluation algorithm above. Then, for all $t \leqslant Z$, (21) holds and $V^\pi\left(s_0, \boldsymbol{b}_{s_0}\right)$ in (16) equals $U_0^\pi(h_0)$ for any $h_0 = \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$.*

For ease of reading, we defer all theorem proofs to "Appendix".

To generalize the algorithm to randomized policies, it would require an additional summation in (22) to account for the probability distribution of the action at decision epoch t under decision rule x_t as follows:

$$U_t^\pi(h_t) = \sum_{a\in A} \mu_{h_t}(a) \left\{ \bar{r}_{s_t}^{x_t(h_t)} + \gamma \sum_{s_{t+1}\in S} \bar{p}_{s_t,s_{t+1}}^{x_t(h_t)} U_{t+1}^\pi\left(h_t, x_t(h_t), \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle\right) \right\}. \tag{23}$$

Theorem 1 shall be extended to $\pi \in \Pi^{\mathrm{DHR}}$ as follows.

Theorem 2 *For any given $h_0 = \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$ and a fixed policy $\pi \in \Pi^{\mathrm{DHR}}$, suppose U_t^π, $t \in T$, has been generated by the policy evaluation algorithm with (23) replacing (22). Then,* for all *$t \leqslant Z$, (19) holds and $V^\pi\left(s_0, \boldsymbol{b}_{s_0}\right)$ in (15) equals $U_0^\pi(h_0)$ for any $h_0 = \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$.*

Now let

$$U_t^*(h_t) = \max_{\pi\in\Pi^{\mathrm{DHR}}} U_t^\pi(h_t).$$

It indicates the maximum over all policies of the expected total discounted reward from decision epoch t onward when the history up to time t is h_t.

The *optimality equations* for the finite horizon DFMDP are given by

$$U_t(h_t) = \max_{a_t\in A} \left\{ \bar{r}_{s_t}^{a_t} + \gamma \sum_{s_{t+1}\in S} \bar{p}_{s_t,s_{t+1}}^{a_t} U_{t+1}\left(h_t, a_t, \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle\right) \right\} \tag{24}$$

for $t \in T$ and $h_t = \big(h_{t-1}, a_{t-1}, \langle s_t, \boldsymbol{b}_{s_t}\rangle\big) \in H_t$, where $\langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle = \langle s_{t+1}, \tau(\langle s_t, \boldsymbol{b}_{s_t}\rangle, a_t, s_{t+1})\rangle$. For $t = Z$ and $h_Z = \big(h_{Z-1}, a_{Z-1}, \langle s_Z, \boldsymbol{b}_{s_Z}\rangle\big) \in H_Z$, we add the *boundary condition*

$$U_Z(h_Z) = \overline{r}^{0}_{s_Z}. \tag{25}$$

Before stating more theorems, we introduce the following lemma 1.

Lemma 1 Let *f be a real-valued function on an arbitrary finite discrete set A* and $\mu()$ *be a probability distribution on A. Then*

$$\sup_{a \in A} f(a) \geqslant \sum_{a \in A} \mu(a) f(a).$$

The following theorem summarizes the optimality properties of solutions to the optimality equations.

Theorem 3 Let *H_t be the history up to time t from any $h_0 = \langle s_0, \boldsymbol{b}_{s_0}\rangle$ onward. Suppose U_t are solutions of* (24) *and* (25) *for $t = 0,1,\cdots, Z$. Then*
(i) *$U_t(h_t) = U^*_t(h_t)$ for all $h_t \epsilon H_t$, $t = 0,1,\cdots, Z$, and*
(ii) *$V^*\big(s_0, \boldsymbol{b}_{s_0}\big) = U_0(h_0)$.*

The following result shows how to use the optimality equations to find the optimal policy and to verify its optimality.

Theorem 4 *Suppose U^*_t, $t = 0, 1, \cdots, Z$, are solutions of* (24) *and* (25), *and policy $\pi^* = \big(x^*_0, x^*_1, \ldots, x^*_{Z-1}\big) \in \Pi^{\mathrm{DHD}}$ satisfies*

$$\begin{aligned} &\overline{r}^{x^*_t(h_t)}_{s_t} + \gamma \sum_{s_{t+1} \in S} \overline{p}^{x^*_t(h_t)}_{s_t, s_{t+1}} U^*_{t+1}\big(h_t, x^*_t(h_t), \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle\big) \\ &= \max_{a_t \in A} \left\{ \overline{r}^{a_t}_{s_t} + \gamma \sum_{s_{t+1} \in S} \overline{p}^{a_t}_{s_t, s_{t+1}} U^*_{t+1}\big(h_t, a_t, \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}}\rangle\big) \right\} \end{aligned} \tag{26}$$

for $t \in T$. Then

(i) for $t = 0,1,\cdots, Z$,

$$U^{\pi^*}_t(h_t) = U^*_t(h_t), \forall\, h_t \in H_t.$$

(ii) *π^* is the optimal policy, and*

$$V^{\pi^*}\big(s_0, \boldsymbol{b}_{s_0}\big) = V^*\big(s_0, \boldsymbol{b}_{s_0}\big), \forall \langle s_0, \boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C.$$

Note that we have restricted attention to double-factored history-dependent deterministic policies in Theorem 4. This is because if there existed a double-factored history-dependent randomized policy which satisfied the obvious generalization of (26), as a result of Lemma 1, we could find a deterministic policy which satisfies (26).

Equation (26) can be written as

$$x_t^*(h_t) \in \underset{a_t \in A}{\operatorname{argmax}} \left\{ \overline{r}_{s_t}^{a_t} + \gamma \sum_{s_{t+1} \in S} \overline{p}_{s_t, s_{t+1}}^{a_t} U_{t+1}^* \left(h_t, a_t, \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}} \rangle \right) \right\}, t \in T. \tag{27}$$

The optimal policy π^* derived from (27) in Theorem 4 exists because of the finite set S and A. So we obtain directly Theorem 5.

Theorem 5 *There exists an optimal double-factored history-dependent deterministic policy for finite state set and action set.*

We next show that there exists an optimal policy which is double-factored Markovian and deterministic.

Theorem 6 Let $U_t^*, t = 0, 1, \cdots, Z$, *be the solutions of* (24) *and* (25). *Then*

(i) for $t = 0, 1, \cdots, Z$, $U_t^*(h_t)$ depends on h_t only through $\langle s_t, \boldsymbol{b}_{s_t} \rangle$, i.e., $U_t^*(h_t) = U_t^*(s_t, \boldsymbol{b}_{s_t})$;

(ii) *there exists an optimal policy which is double-factored Markovian and deterministic when both S* and *A are finite.*

Theorem 6 shows that there exists a double-factored Markovian deterministic policy that is optimal among all classes of policies. Furthermore, it follows from (A7) in the "Appendix" that there are the optimality equations in terms of the optimal double-factored value function, $U_t^*(s_t, \boldsymbol{b}_{s_t})$:

$$U_t^*(s_t, \boldsymbol{b}_{s_t}) = \max_{a_t \in A} \left\{ \sum_{k \in C} b_{s_t,k} r_{s_t,k}^{a_t} + \gamma \sum_{s_{t+1} \in S} \sum_{k \in C} b_{s_t,k} p_{s_t,s_{t+1},k}^{a_t} U_{t+1}^* \left(s_{t+1}, \tau \left(\langle s_t, \boldsymbol{b}_{s_t} \rangle, a_t, s_{t+1} \right) \right) \right\}. \tag{28}$$

So we have established that

$$\begin{aligned} V^* \left(s_0, \boldsymbol{b}_{s_0} \right) &= \max_{\pi \in \Pi^{\mathrm{DHR}}} V^\pi \left(s_0, \boldsymbol{b}_{s_0} \right) = \max_{\pi \in \Pi^{\mathrm{DHD}}} V^\pi \left(s_0, \boldsymbol{b}_{s_0} \right) \\ &= \max_{\pi \in \Pi^{\mathrm{DMD}}} V^\pi \left(s_0, \boldsymbol{b}_{s_0} \right), \forall \langle s_0, \boldsymbol{b}_{s_0} \rangle \in S \times \Delta^C, \end{aligned} \tag{29}$$

where the expected total discounted rewards $V^\pi(s_0, \boldsymbol{b}_{s_0})$ generated by policies $\pi \in \Pi^{\mathrm{DHR}}$, $\pi \in \Pi^{\mathrm{DHD}}$, and $\pi \in \Pi^{\mathrm{DMD}}$ are expressed by (15), (16) and (17), respectively.

The following theorem establishes the relationship between the WVP in [4] and our finite horizon DFMDP problem.

Theorem 7 *The finite horizon DFMDP is equivalent to the WVP of multi-scenario MDP.*

Thus a DFMDP can also be denoted by the 6-elements-tuple $\left(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0}\rangle\right)$ according to the equivalence in Theorem 7.

Now we present the *double-factored backward induction algorithm* (DFBI) for the exact solutions of the finite horizon DFMDPs based on the above theory.

The Double-Factored Backward Induction Algorithm

```
Step 1 For a given $\langle s_0,\boldsymbol{b}_{s_0}\rangle \in S \times \Delta^C$, compute all $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ by (10) and place them into set $\Psi_t$ for $t = 0,1, \cdots, Z$.
Step 2 Set $t = Z$ and
    $U_Z^*\left(s_Z, \boldsymbol{b}_{s_Z}\right) = \bar{r}_{s_Z}^0, \qquad \forall\langle s_Z, \boldsymbol{b}_{s_Z}\rangle \in \Psi_t.$
Step 3 If $t = 0$ go to step 6; otherwise go to step 4.
Step 4 Substitute $t - 1$ for $t$, compute $U_t^*(s_t, \boldsymbol{b}_{s_t})$ for each $\langle s_t, \boldsymbol{b}_{s_t}\rangle \in \Psi_t$ by
    $U_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = \max_{a_t\in A}\left\{\sum_{k\in C} b_{s_t,k} r_{s_t,k}^{a_t} + \gamma \sum_{s_{t+1}\in S}\sum_{k\in C} b_{s_t,k} p_{s_t,s_{t+1},k}^{a_t} U_{t+1}^*\left(s_{t+1}, \tau\left(\langle s_t, \boldsymbol{b}_{s_t}\rangle, a_t, s_{t+1}\right)\right)\right\}$
    and establish the optimal action set $A_t^*\left(s_t, \boldsymbol{b}_{s_t}\right)$ for $\langle s_t, \boldsymbol{b}_{s_t}\rangle$:
    $A_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = \operatorname{argmax}_{a_t\in A}\left\{\sum_{k\in C} b_{s_t,k} r_{s_t,k}^{a_t} + \gamma \sum_{s_{t+1}\in S}\sum_{k\in C} b_{s_t,k} p_{s_t,s_{t+1},k}^{a_t} U_{t+1}^*\left(s_{t+1}, \tau\left(\langle s_t, \boldsymbol{b}_{s_t}\rangle, a_t, s_{t+1}\right)\right)\right\}.$
Step 5 Return to step 3.
Step 6 Establish the optimal policy to be
    $\pi^* = \left(x_0^*\left(s_0, \boldsymbol{b}_{s_0}\right), x_1^*\left(s_1, \boldsymbol{b}_{s_1}\right), \ldots, x_{Z-1}^*\left(s_{Z-1}, \boldsymbol{b}_{s_{Z-1}}\right)\right); x_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) \in A_t^*\left(s_t, \boldsymbol{b}_{s_t}\right), \forall\langle s_t, \boldsymbol{b}_{s_t}\rangle \in \Psi_t^*, t \in T,$
    where the sets $\Psi_t^*$ are defined as follows:
    $\Psi_0^* = \left\{\langle s_0, \boldsymbol{b}_{s_0}\rangle\right\},$
    $\Psi_t^* = \left\{\langle s_t, \boldsymbol{b}_{s_t}\rangle : \langle s_t, \boldsymbol{b}_{s_t}\rangle = \langle s_t, \tau\left(\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle, x_{t-1}^*(s_{t-1}, \boldsymbol{b}_{s_{t-1}}), s_t\right)\rangle, \forall\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}}\rangle \in \Psi_{t-1}^*, \forall s_t \in S\right\},$
        $t = 1, \ldots., Z - 1.$
```

The DFBI is an exact solution method for the finite horizon DFMDPs. We know from the transition tree in Fig. 1 that if all state-belief pairs $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ for $t = 0,1, \cdots, Z$ are reachable, $|\Psi_t| = (|A|\cdot|S|)^t$. So the time complexity of the algorithm is $O((|A|\cdot|S|\cdot|C|)^Z)$ and it increases exponentially with the number of decision epochs.

When actually implementing the method, there are several approaches listed below to reduce the computational complexity such that the algorithm would be practicable to solve real-world problems.

(i) For any $\langle s_t, \boldsymbol{b}_{s_t}\rangle$, $t = 1, \cdots, Z$, in step 1, if $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ equals its elder $\langle s_{t'}, \boldsymbol{b}_{s_{t'}}\rangle$, $t' < t$, or its brother $\langle s_t, \boldsymbol{b}_{s_t}\rangle'$, its sons can be directly copied from ones of $\langle s_{t'}, \boldsymbol{b}_{s_{t'}}\rangle$ or $\langle s_t, \boldsymbol{b}_{s_t}\rangle'$ to reduce the computation. If $\langle s_t, \boldsymbol{b}_{s_t}\rangle$ is a pseudo-state-belief pair, no computation for its sons is required.

(ii) For $U_t^*(s_t, \boldsymbol{b}_{s_t})$ and $A_t^*(s_t, \boldsymbol{b}_{s_t})$ of $\langle s_t, \boldsymbol{b}_{s_t}\rangle \in \Psi_t$ in step 4, since the same state-belief pairs in set Ψ_t have the same U_t^* and A_t^* according to Theorem 6, copies of U_t^* and A_t^* are directly used to avoid repetitive computations. When programming the algorithm, let $U_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = 0$ and $A_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = \varnothing$ when $\langle s_t, \boldsymbol{b}_{s_t}\rangle = \langle s_t, \boldsymbol{e}_0\rangle$, and $U_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = U_{t,k}^*(s_t)$ and $A_t^*\left(s_t, \boldsymbol{b}_{s_t}\right) = A_{t,k}^*(s_t)$ when $\langle s_t, \boldsymbol{b}_{s_t}\rangle = \langle s_t, \boldsymbol{e}_k\rangle$, $k \in C$, where $U_{t,k}^*(s_t)$ and $A_{t,k}^*(s_t)$ are the optimal value function and the optimal action set of MDP k at time t.

(iii) When implementing the DFBI, we only need to store the set Ψ_t, as well as the parent-son relationship between $\langle s_t, \boldsymbol{b}_{s_t} \rangle \in \Psi_t$ and $\langle s_{t+1}, \boldsymbol{b}_{s_{t+1}} \rangle \in \Psi_{t+1}$ in order to reduce the storage demand of the algorithm.

There are three types of scenarios illustrated in [16]. They can be described by three sets of parameter pair respectively:

(I) For certain transition probabilities and uncertain rewards, $J_{\boldsymbol{p},\boldsymbol{r}} = J_{\boldsymbol{p}_0,\boldsymbol{r}} = \{(\boldsymbol{p}_0, \boldsymbol{r}_1), \cdots, (\boldsymbol{p}_0, \boldsymbol{r}_{|C|})\}$.

(II) For uncertain transition probabilities and certain rewards, $J_{\boldsymbol{p},\boldsymbol{r}} = J_{\boldsymbol{p},\boldsymbol{r}_0} = \{(\boldsymbol{p}_1, \boldsymbol{r}_0), \cdots, (\boldsymbol{p}_{|C|}, \boldsymbol{r}_0)\}$.

(III) For uncertain transition probabilities and rewards, $J_{\boldsymbol{p},\boldsymbol{r}} = \{(\boldsymbol{p}_1, \boldsymbol{r}_1), \cdots, (\boldsymbol{p}_{|C|}, \boldsymbol{r}_{|C|})\}$.

Actually, both the type-I and the type-II scenarios are the special cases of the type-III scenarios.

The following theorem once again establishes the relationship between the DFMDPs and the standard MDPs.

Theorem 8 *The DFMDP* $\left(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0} \rangle\right)$ *with the type-I of the multiple scenarios, i.e.,* $J_{\boldsymbol{p},\boldsymbol{r}} = \left\{(\boldsymbol{p}_0, \boldsymbol{r}_1), \cdots, (\boldsymbol{p}_0, \boldsymbol{r}_{|C|})\right\}$, *is equivalent to the standard MDP with parameter pair* $(\boldsymbol{p}_0, \overline{\boldsymbol{r}})$ *where* $\overline{\boldsymbol{r}} = \sum_{k \in C} b_{s_0,k} \boldsymbol{r}_k$.

The Weight-Select-Update (WSU) approximation algorithm in [4] is also a backward induction algorithm. Now we show that the WSU can solve exactly the DFMDPs with the type-I scenarios.

With the notations in this article, the procedure of the WSU is as follows:

The Weight-Select-Update Algorithm

```
Input: a multi-scenario MDP
Let $U_{Z,k}(s_Z) = r^0_{s_Z,k}, \forall\ k \in C$
$t \leftarrow Z-1$
while $t \geq 0$ do
    for every state $s_t \in S$ do
        $x_t(s_t) \leftarrow \operatorname*{argmax}_{a_t \in A} \left\{ \sum_{k \in C} b_{s_0,k} \left( r^{a_t}_{s_t,k} + \sum_{s_{t+1} \in S} b_{s_t,k} p^{a_t}_{s_t,s_{t+1},k} U_{t+1,k}(s_{t+1}) \right) \right\}.$   (30)
    end for
    for every model $k \in C$ do
        $U_{t,k}(s_t) \leftarrow r^{x_t(s_t)}_{s_t,k} + \sum_{s_{t+1} \in S} b_{s_t,k} p^{x_t(s_t)}_{s_t,s_{t+1},k} U_{t+1,k}(s_{t+1}).$   (31)
    end for
    $t \leftarrow t-1$
end while
Output: $\pi = (x_0, \cdots, x_{Z-1})$
```

There is $J_{\boldsymbol{p},\boldsymbol{r}} = \{(\boldsymbol{p}_0, \boldsymbol{r}_1), \cdots, (\boldsymbol{p}_0, \boldsymbol{r}_{|C|})\}$ in the $(T, S, A, C, J_{\boldsymbol{p},\boldsymbol{r}}, \langle s_0, \boldsymbol{b}_{s_0}\rangle)$ with the type-I scenarios. Let $U_t(s_t) = \sum_{k \in C} b_{s_0,k} U_{t,k}(s_t)$, $t = 0, 1, \cdots, Z$. With parameters $\boldsymbol{p}_0$ and $\overline{\boldsymbol{r}}$, (30) and (31) in the above procedure become as follow:

$$x_t(s_t) \leftarrow \underset{a_t \in A}{\operatorname{argmax}} \left\{ \overline{r}_{s_t}^{a_t} + \sum_{s_{t+1} \in S} p_{s_t, s_{t+1}, 0}^{a_t} U_{t+1}(s_{t+1}) \right\}, \tag{30$'$}$$

and

$$\begin{aligned} &U_Z(s_Z) = \overline{r}_{s_Z}^{0}, \\ &U_t(s_t) \leftarrow \overline{r}_{s_t}^{x_t(s_t)} + \sum_{s_{t+1} \in S} p_{s_t, s_{t+1}, 0}^{x_t(s_t)} U_{t+1}(s_{t+1}). \end{aligned} \tag{31$'$}$$

The procedure with (30′) and (31′) is just for the solutions of the standard MDP with parameter pair ($\boldsymbol{p}_0$,$\overline{\boldsymbol{r}}$).

The mean value problem (MVP) heuristic is another approximation algorithm for the DFMDPs. The MVP is a simple problem in which all parameters take on their expected values. For the DFMDPs, MVP corresponds to the case where all transition probabilities and rewards are weighted as follows:

$$\tilde{r}_s^a = \sum_{k \in C} b_{s_0,k} r_{s,k}^a, \; \tilde{p}_{s,s'}^a = \sum_{k \in C} b_{s_0,k} p_{s,s',k}^a, \; \forall\, s, s' \in S, a \in A.$$

That is, the MVP is a standard MDP problem with the parameter pair ($\widetilde{\boldsymbol{p}}$, $\widetilde{\boldsymbol{r}}$).

By the means of computational experiment in Sect. 4, we will compare the solutions by the MVP and the WSU approximation methods with the solutions by our method. Furthermore, we can obtain the following corollary based on the definition of the MVP and the WSU methods above.

Corollary *Both the MVP and the WSU methods are the exact solution methods for DFMDPs with type-I scenarios.*

3.4 Infinite Horizon DFMDP

Now we formulate the infinite horizon DFMDPs. We assume that the state set, the action set and the scenario set are finite, the transition probabilities and the rewards for each scenario are stationary (time homogeneous), the rewards of each scenario are bounded. The discounted factor γ satisfies that $0 < \gamma < 1$.We consider the stationary and deterministic policy $\pi = (x(s, \boldsymbol{b}_s), x(s, \boldsymbol{b}_s), \cdots)$, where the decision rule x: $S \times C \rightarrow A$ specifies the action to be taken at any state-belief pair.

Under the above assumptions, the optimality equations (28) can also be written as

$$V_n(s_n, \boldsymbol{b}_{s_n}) = \max_{a_n \in A} \left\{ \sum_{k \in C} b_{s_n,k} r_{s_n,k}^{a_n} + \gamma \sum_{s_{n+1} \in S} \sum_{k \in C} b_{s_n,k} p_{s_n,s_{n+1},k}^{a_n} V_{n+1}\big(s_{n+1}, \tau\big(\langle s_n, \boldsymbol{b}_{s_n} \rangle, a_n, s_{n+1}\big)\big) \right\}. \tag{32}$$

Passing to the limit in (32) suggests that equations of the following form will characterize values and optimal policies for the infinite horizon DFMDPs:

$$V(s, \boldsymbol{b}_s) = \max_{a \in A} \left\{ \sum_{k \in C} b_{s,k} r_{s,k}^{a} + \gamma \sum_{s' \in S} \sum_{k \in C} b_{s,k} p_{s,s',k}^{a} V\big(s', \tau\big(\langle s, \boldsymbol{b}_s \rangle, a, s'\big)\big) \right\}, \tag{33}$$

where $V(s, \boldsymbol{b}_s)$ denotes the optimal value function for $\langle s, \boldsymbol{b}_s \rangle$. The system of equations (33) are called the *optimality equations* or *Bellman equations* for the infinite horizon DFMDPs.

The Bellman equations (33) can also be rewritten in the value function mapping form. Let $\mathcal{V}$ be the space of real-valued bounded functions $V\colon S \times C \to \mathbb{R}$, we have $\eta\colon S \times \Delta^C \times A \times \mathcal{V} \to \mathbb{R}$ defined as

$$\eta(s, \boldsymbol{b}_s, a, V) = \sum_{k \in C} b_{s,k} r_{s,k}^{a} + \gamma \sum_{s' \in S} \sum_{k \in C} b_{s,k} p_{s,s',k}^{a} V\big(s', \tau\big(\langle s, \boldsymbol{b}_s \rangle, a, s'\big)\big).$$

Now by defining the value function mapping $H\colon \mathcal{V} \to \mathcal{V}$ as $HV(s, \boldsymbol{b}_s) = \max_{a \in A} \eta(s, \boldsymbol{b}_s, a, V)$, the Bellman equations (33) can be written as $V = HV$.

Theorem 9 *Let H be the value function mapping defined above, then H is an isotone mapping and a contraction under the supremum norm* $||V|| = \sup_{\langle s, \boldsymbol{b}_s \rangle \in S \times \Delta^C} |V(s, \boldsymbol{b}_s)|$.

Since H is a contraction mapping, there exists a unique $V^* \in \mathcal{V}$ such that $V^* = HV^*$ by Barnach's fixed-point theorem [1]. And for any $V_0 \in \mathcal{V}$, the sequence $\{V_n\}$ defined below converges to V^* (see [1]):

$$V_n = HV_{n-1} = H^n V_0.$$

The theory above is the base of developing algorithms solving the infinite horizon DFMDPs. To limit the scope of this paper, the algorithms and case studies for the infinite horizon DFMDPs would be presented by our follow-up articles.

4 Computational Experiments

The computational experiments involving three sets of test instances for comparing solution methods for finite horizon DFMDPs considering run-time and solution quality are illustrated in this section. The first set of experiments is based on a randomly generated multi-scenario MDP. The second set of experiments is based on a multi-scenario MDP for determining the most cost-effective human immunodeficiency virus (HIV) treatment policy which has been used for pedagogical purposes in the medical decision-making literature [21, 22]. The third set of experiments is an illustrative instance of the scenario recognition problem.

As benchmarks, other than solving the problems with the DFBI algorithm, we will also generate solutions from the WSU, MVP methods and the MIP formulation. We use V_Z^* to denote the optimal value obtained by the DFBI, and $\widetilde{V}_Z^*$ to denote the optimal value obtained from either WSU or MVP for the multi-scenario MDP with Z decision epochs. Then let

$$\text{Gap} = \frac{V_Z^* - \widetilde{V}_Z^*}{V_Z^*} \times 100\%.$$

All solution methods are implemented using MATLAB 2017a.

4.1 Random Instance

To generate the random instances, firstly the number of states, actions, scenarios, and decision epochs for the problem need to be defined. Then, the scenario parameters are randomly sampled. We sample the rewards from uniform distribution and the transition probabilities from Dirichlet distribution. The initial beliefs are uninformed priors on the scenarios.

Let $|S| = |A| = |C| = 4$ and $Z = 4$. The rewards for all scenarios are randomly generated from uniform distribution between 0 and 1. The transition probabilities are randomly generated from Dirichlet distribution which is characterized by $|S|$ parameters $(\rho\alpha_1, \rho\alpha_2, \cdots, \rho\alpha_{|S|})$, where $(\alpha_1, \alpha_2, \cdots, \alpha_{|S|})$ with $\alpha_i > 0\ \forall\ i$, is the *base measure* of the distribution and $\rho > 0$ is the *concentration parameter*. Then we repeat the process, generating 30 instances for $\rho = 1, 10, 100$, respectively. Table 1 demonstrates the run-time of the four methods: the MVP, WSU, MIP and DFBI. We find that the MVP and WSU methods are able to solve these instances more quickly (under 0.05 CPU seconds for each instance) and the DFBI also does fairly quickly (under 5 CPU seconds for each instance) with the exact solutions. The MIP takes much more run-time to solve for the exact solutions among these instances.

We evaluate the optimality gap of MVP and WSU methods considering three type of the multiple scenarios. Table 2 demonstrates some summary statistics for the optimality gap of MVP and WSU methods, obtained from these 50 instances for each type of scenarios and each concentration parameter. For the type-I scenarios, the results conforms with the Corollary in Sect. 3.3 and both the MVP and the WSU find the exact solutions. For fixed ρ, the optimality gaps of the MVP and WSU methods for the type-III scenarios are much greater than the ones for the type-II scenarios. For both the type-II and type-III scenarios, the greater the value of ρ, the smaller the optimality gaps of the MVP and WSU methods. This is because the greater value of ρ corresponds to the smaller variance for transition matrix across scenarios, which approximates to the problem with the type-I scenarios.

Table 1 The solution time (CPU seconds) of the MVP, WSU, MIP and DFBI methods on random DFMDP test instances

Concentration parameter ρ	MVP		WSU		MIP		DFBI	
	Average	Maximum	Average	Maximum	Average	Maximum	Average	Maximum
1	< 0.05	< 0.05	< 0.05	< 0.05	42.57	127.43	3.62	4.70
10	< 0.05	< 0.05	< 0.05	< 0.05	39.50	94.02	3.85	4.35
100	< 0.05	< 0.05	< 0.05	< 0.05	36.71	75.66	4.08	4.88

Table 2 The optimality gap of MVP and WSU methods on random DFMDPs with three types of multiple scenarios

Types of multiple scenarios	Concentration parameter ρ	Optimality Gap of MVP/%		Optimality Gap of WSU/%	
		Average	Maximum	Average	Maximum
Type-I	1, 10, 100	0.00	0.00	0.00	0.00
Type-II	1	1.97	3.12	2.03	3.12
	10	0.09	0.57	0.07	0.43
	100	< 0.01	< 0.01	< 0.01	< 0.01
Type-III	1	10.33	23.84	6.84	15.31
	10	2.88	6.91	2.27	5.28
	100	0.39	1.61	0.36	1.40

4.2 Instance of Medical Decision-Making

An MDP for determining the optimal timing of treatments for HIV is considered. In the MDP, HIV is characterized according to 4 health states: Mild, Moderate, Severe, or Dead. The decision maker can choose to start the patient on one of three treatments: Treatment **A**, Treatment **B**, and Treatment **C**. Treatment **A** is the least effective but also the least expensive while Treatment **C** is the most effective but comes at the highest cost. A summary table of parameter values for this MDP as well as some sampling distributions for each parameter is provided in [21]. In our experiments, we sampled two scenarios of the transition probabilities from the Dirichlet distribution in [21] to simulate findings coming from different clinical studies. They are listed below.

$$\boldsymbol{p}_1^{\mathbf{A}} = \begin{bmatrix} 0.710 & 0.209 & 0.070 & 0.011 \\ 0 & 0.581 & 0.400 & 0.019 \\ 0 & 0 & 0.739 & 0.261 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{p}_1^{\mathbf{B}} = \begin{bmatrix} 0.790 & 0.151 & 0.051 & 0.008 \\ 0 & 0.697 & 0.290 & 0.013 \\ 0 & 0 & 0.811 & 0.189 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{p}_1^{\mathbf{C}} = \begin{bmatrix} 0.898 & 0.074 & 0.025 & 0.003 \\ 0 & 0.852 & 0.142 & 0.006 \\ 0 & 0 & 0.908 & 0.092 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{p}_2^{\mathbf{A}} = \begin{bmatrix} 0.733 & 0.198 & 0.064 & 0.005 \\ 0 & 0.582 & 0.408 & 0.010 \\ 0 & 0 & 0.753 & 0.247 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{p}_2^{\mathbf{B}} = \begin{bmatrix} 0.806 & 0.143 & 0.046 & 0.005 \\ 0 & 0.697 & 0.296 & 0.007 \\ 0 & 0 & 0.821 & 0.179 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{p}_2^{\mathbf{C}} = \begin{bmatrix} 0.862 & 0.102 & 0.033 & 0.003 \\ 0 & 0.784 & 0.211 & 0.005 \\ 0 & 0 & 0.872 & 0.128 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The data related to the rewards is taken directly from [21, 22] and they are the same for two scenarios. Let $S = \{\text{Mild, Moderate, Severe}\}$, $A = \{\mathbf{A, B, C}\}$, $Z = 9$, and $\gamma = 1$. With these settings, we obtain the optimal policies, π_k^* for MDP k ($k = 1, 2$) and π^* for the DFMDP. We consider that the number of patients eligible for either scenario is balanced, thus $\boldsymbol{b}_{s_0} = [0.5{,}0.5]^\top$. Then we can achieve the weighted value (i.e., the average *net benefit* in [21, 22]) for each policy, i.e., $V^\pi\left(s_0, b_{s_0}\right), \pi = \pi_1^*, \pi_2^*, \pi^*$ listed in Table 3.

Table 3 shows that for the patients starting in the mild health state, the average net benefit of policy π^* is greater than that for policies $\pi_k^*, k = 1, 2$ by 0.44% and 3.05% respectively. For the patients starting in the moderate health state, the average net benefit of policy π^* is greater than that for policies $\pi_k^*, k = 1, 2$ by 0.24% and 12.38% respectively. For the patients starting in the severe health state, the average net benefits of policies π^*, π_1^*, π_2^* are all zero.

This instance illustrates that it is difficult to obtain the best average gain by using the optimal policy of a single MDP under multiple scenarios of the parameters, but the optimal policy from DFMDP can do.

4.3 Variant of Medical Instance

A illustrative instance of the scenario recognition problem is given here. Assume that there are two authoritative findings with different transition probabilities in the previous medical instance. We represent them with scenario $k = 1, 2$ respectively as follows.

$$\boldsymbol{p}_1^{\mathbf{A}} = \begin{bmatrix} 0.750 & 0.150 & 0.090 & 0.010 \\ 0 & 0.600 & 0.380 & 0.020 \\ 0 & 0 & 0.750 & 0.250 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{p}_1^{\mathbf{B}} = \begin{bmatrix} 0.812 & 0.112 & 0.068 & 0.008 \\ 0 & 0.700 & 0.285 & 0.015 \\ 0 & 0 & 0.812 & 0.188 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{p}_1^{\mathbf{C}} = \begin{bmatrix} 0.875 & 0.075 & 0.045 & 0.005 \\ 0 & 0.800 & 0.190 & 0.010 \\ 0 & 0 & 0.875 & 0.125 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

Table 3 The average net benefits $V^\pi\left(s_0, \boldsymbol{b}_{s_0}\right), \pi = \pi_1^*, \pi_2^*, \pi^*$ (Unit: US$)

s_0	$V^{\pi_1^*}\left(s_0, \boldsymbol{b}_{s_0}\right)$	$V^{\pi_2^*}\left(s_0, \boldsymbol{b}_{s_0}\right)$	$V^{\pi^*}\left(s_0, \boldsymbol{b}_{s_0}\right)$
Mild	163 574	159 436	164 298
Moderate	52 040	46 419	52 167
Severe	0	0	0

$$\boldsymbol{p}_2^{\mathbf{A}} = \begin{bmatrix} 0.787 & 0.145 & 0.068 & 0 \\ 0 & 0.742 & 0.245 & 0.013 \\ 0 & 0 & 0.771 & 0.229 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{p}_2^{\mathbf{B}} = \begin{bmatrix} 0.842 & 0.108 & 0.050 & 0 \\ 0 & 0.809 & 0.182 & 0.009 \\ 0 & 0 & 0.830 & 0.170 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{p}_2^{\mathbf{C}} = \begin{bmatrix} 0.898 & 0.102 & 0 & 0 \\ 0.120 & 0.751 & 0.129 & 0 \\ 0 & 0.084 & 0.796 & 0.120 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where the transition probabilities of scenario 1 come from [22], and ones of scenario 2 are hypothetical. The data related to the rewards is taken directly from [21, 22] and the same for two scenarios. Let $S = \{1, 2, 3\} = \{\text{Mild, Moderate, Severe}\}$, $A = \{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$, $Z = 8$, and $\gamma = 1$.

If the patient exhibits transition that shows similar pattern as scenario 1, we would infer the patient belongs to patient group I, similarly for patient group II. If we don't have any prior knowledge of the population percentage for these two groups, we would just assume an equal possibility, i.e., $\boldsymbol{b}_{s_0} = [0.5, 0.5]^\top$. Now let's start with a patient currently in the mild health state, and apply the optimal treatment strategy while recognizing which group this patient belongs to. The optimal treatment policy for the patients starting in the mild health state ($s_0 = 1$) can be found by using the DFBI and is demonstrated in Fig. 2.

For this patient X with his/her current health state being in mild, our first-year treatment plan is **C**. And at the end of year-1, if his/her health station remains at mild or transitions to moderate, then we continue providing treatment plan **C** for the patient. If the health state becomes severe, the optimal treatment in year-2 is $x_1^*(3, \mathbf{b}_3) = x_1^*(3, \boldsymbol{e}_1) = x_{1,1}^*(3) = \mathbf{A}$. The state-belief pair $\langle 3, \boldsymbol{e}_1 \rangle$ implies that patient X belongs to group-I rather than group-II, since $p_{1,3,2}^{\boldsymbol{C}} = 0$ implies that it's impossible for group-II patient's health state transitioning from mild to severe. Similarly, for another patient Y if his/her health state becomes mild at the end of year-2, after two consecutive treatment plan **C**, the optimal treatment in year 3 is $x_2^*(1, \boldsymbol{b}_1) = x_2^*(1, \boldsymbol{e}_2) = x_{2,2}^*(1) = \mathbf{C}$, and the state-belief pair $\langle 1, \boldsymbol{e}_2 \rangle$ implies that patient Y belongs to group-II.

This is only an illustrative example with very limited scale. In order to apply this idea to real-world scenario recognition problems, the method needs to be combined with

$$x_0^*(s_0, \boldsymbol{b}_{s_0}) \qquad x_1^*(s_1, \boldsymbol{b}_{s_1}) \qquad x_2^*(s_2, \boldsymbol{b}_{s_2}) \qquad \cdots$$

$$x_0^*(1, [0.5, 0.5]^{\mathrm{T}}) = \mathbf{C} - \left\{ \begin{array}{l} x_1^*(1, [0.494, 0.506]^{\mathrm{T}}) = \mathbf{C} - \left\{ \begin{array}{ll} x_2^*(1, [0.487, 0.513]^{\mathrm{T}}) = \mathbf{C} & \cdots \\ x_2^*(2, [0.417, 0.583]^{\mathrm{T}}) = \mathbf{C} & \cdots \\ x_2^*(3, \boldsymbol{e}_1) = \mathbf{A} & \cdots \end{array} \right. \\ x_1^*(2, [0.424, 0.576]^{\mathrm{T}}) = \mathbf{C} - \left\{ \begin{array}{ll} x_2^*(1, \boldsymbol{e}_2) = \mathbf{C} & \cdots \\ x_2^*(2, [0.439, 0.561]^{\mathrm{T}}) = \mathbf{C} & \cdots \\ x_2^*(3, [0.520, 0.480]^{\mathrm{T}}) = \mathbf{A} & \cdots \end{array} \right. \\ x_1^*(3, \boldsymbol{e}_1) = \mathbf{A} \qquad\qquad \cdots \end{array} \right.$$

Fig. 2 The treatment policy of patients starting in the mild health state

algorithms dealing with infinite horizon DFMDPs. This will be one of our focusing directions for future works.

5 Conclusions and Future Works

For the weighted value problem of MDPs with multiple scenarios of the parameters, we introduce the concept of scenario belief to indicate the probability distribution that the scenarios realize in the system and derive its update at every decision epoch based on Bayesian rule. We formulate the expected total discounted reward of a policy by adding the expectation over the scenario beliefs, on top of the usual expectation over the intrinsic state transition uncertainty, and establish a new framework named as the DFMDPs. We show that the usage of state-belief pair is the sufficient statistics of the past history and contains the complete information required for decision-making. We discuss four classes of policies in the finite horizon DFMDPs and prove that there exists a double-factored Markovian deterministic policy which is optimal among all classes of policies. The double-factored backward induction algorithm is proposed. It is an efficient exact solution method for the double-factored Markovian deterministic policies in the finite horizon DFMDPs. We also show that the optimality equation for the infinite horizon DFMDPs is an isotone mapping and a contraction under the supremum norm. This ensures the existence of solutions of the optimality equation. Our work enriches the theories of MDPs and expands the application scope of MDPs.

Future work will focus on the following items:

(i) The restrictions on the state set and action set can be relaxed such that DFMDP models apply to more real-world problems.
(ii) The efficient solution methods for the infinite horizon DFMDPs are further studied.
(iii) More efficient offline and online algorithms can be invented to solve the large-size DFMDPs for real-world applications.
(iv) Investigating a learning-based approach in the framework of DFMDP is one of the directions for future exploration.

Acknowledgements The author is grateful to the editors and reviews for their dedicated inputs and supports during the review process. The author would like to thank his advisor Prof. Theodore Allen from Systems Engineering of The Ohio State University, for his continuous support and guidance. The author also wants to express his sincere appreciation to Prof. Nick Hall from Fisher Business School of The Ohio State University for his valuable suggestions on this work.

Author Contributions C.-J. Hou contributed to all the elements involved in this paper, including framework conceptualization, algorithm design, mathematical proof, and numeric validation.

Competing Interests The author declares that there are no competing interests related to the research work and the writing of this paper.

Appendix

Proof of Theorem 1

For any given $h_0 = \langle s_0, \boldsymbol{b}_{s_0} \rangle \in S \times \Delta^C$ and a fixed $\pi \epsilon \Pi^{\mathrm{DHD}}$, the history sets H_t^π for $t = 0,1, \cdots, Z$ are determined by step 1 in the policy evaluation algorithm. Equation (22) shows that when the history up to time t is $h_t \in H_t^\pi$, the expected value of policy π at decision epoch t, $t + 1$, ···, Z is equal to the reward $\overline{r}_{s_t}^{x_t(h_t)}$ received by selecting action $x_t(h_t)$ plus the expected total discounted reward over the remaining periods. The second term contains the product of $\overline{p}_{s_t,s_{t+1}}^{x_t(h_t)}$ the probability of state transiting from s_t to s_{t+1} when action $x_t(h_t)$ is performed at decision epoch t, and $U_{\mathrm{t+1}}^\pi$ the expected total discounted reward obtained by applying π at decision epoch $t + 1$, ···, Z when the history up to time $t + 1$ is $h_{t+1} = \left(h_t, x_t(h_t), \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}} \rangle\right)$. Summing over all possible states s_{t+1} gives the desired expectation expressed in terms of $U_{\mathrm{t+1}}^\pi$. So Eq. (22) can be written as below

$$U_t^\pi(h_t) = \overline{r}_{s_t}^{x_t(h_t)} + \gamma E_{h_t}^\pi \left\{ U_{t+1}^\pi \left(h_t, x_t(h_t), \langle s_{t+1}, \boldsymbol{b}_{s_{t+1}} \rangle\right) \right\}. \tag{A1}$$

The rest of the proof is by backward induction where the index of induction is t. It is obvious that (21) holds when $t = Z$. Suppose now that (21) holds for $t + 1$, ···, Z. Then by using (A1) and the induction hypothesis, we have

$$\begin{aligned} U_t^\pi(h_t) &= \overline{r}_{s_t}^{x_t(h_t)} + \gamma E_{h_t}^\pi \left\{ E_{h_{t+1}}^\pi \left[\sum_{n=t+1}^{Z-1} \gamma^{n-(t+1)} \overline{r}_{s_n}^{x_n(h_n)} + \gamma^{Z-(t+1)} \overline{r}_{s_Z}^0 \right] \right\} \\ &= \overline{r}_{s_t}^{x_t(h_t)} + \gamma E_{h_t}^\pi \left\{ \sum_{n=t+1}^{Z-1} \gamma^{n-(t+1)} \overline{r}_{s_n}^{x_n(h_n)} + \gamma^{Z-(t+1)} \overline{r}_{s_Z}^0 \right\} = E_{h_t}^\pi \left\{ \sum_{n=t}^{Z-1} \gamma^{n-t} \overline{r}_{s_n}^{x_n(h_n)} + \gamma^{Z-t} \overline{r}_{s_Z}^0 \right\}. \end{aligned}$$

It is true that (21) holds for t. Therefore, $V^\pi\left(s_0, \boldsymbol{b}_{s_0}\right) = U_0^\pi(h_0)$ when $t = 0$.

Proof of Theorem 3

The proof is in two parts. Firstly, we establish by induction that $U_t(h_t) \geqslant U_t^*(h_t)$ for all $h_t \in H_t$ and $t = 0,1, \cdots, Z$. Obviously, $U_Z(h_Z) = \overline{r}_{s_Z}^0 = U_Z^\pi(h_Z)$ for all $h_Z \in H_Z$ and $\pi \in \Pi^{\mathrm{DHR}}$. Therefore $U_Z(h_Z) = U_Z^*(h_Z)$ for all $h_Z \in H_Z$. Now assume that $U_t(h_t) \geqslant U_t^*(h_t)$ for all $h_t \in H_t$ for $t = n + 1, \cdots, Z$. Let $\pi' = \left(x_0', x_1', \cdots, x_{Z-1}'\right)$ be an arbitrary policy in Π^{DHR}. So for $t = n$, we have

$$\begin{aligned} U_n(h_n) &= \max_{a_n \in A} \left\{ \overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n,s_{n+1}}^{a_n} U_{n+1}\left(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle\right) \right\} \\ &\geqslant \max_{a_n \in A} \left\{ \overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n,s_{n+1}}^{a_n} U_{n+1}^*\left(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle\right) \right\} \end{aligned} \tag{A2}$$

$$\geqslant \max_{a_n \in A} \left\{ \overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n, s_{n+1}}^{a_n} U_{n+1}^{\pi'} \big(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle \big) \right\} \tag{A3}$$

$$\geqslant \sum_{a_n \in A} \mu_{h_n}^{'}(a_n) \left\{ \overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n, s_{n+1}}^{a_n} U_{n+1}^{\pi'} \big(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle \big) \right\} \tag{A4}$$

$$= U_n^{\pi'}(h_n).$$

Line (A2) holds because of the induction hypothesis and non-negativity of $\overline{\boldsymbol{p}}$. Line (A3) holds because of the definition of U_{n+1}^*. Line (A4) follows from Lemma 1. The last equality follows from (23) and Theorem 2.

Because π' is arbitrary, we have

$$U_n(h_n) \geqslant U_n^{\pi}(h_n)$$

for all $\pi \in \Pi^{\text{DHR}}$. Thus $U_t(h_t) \geqslant U_t^*(h_t)$ and the induction hypothesis holds. The first part finishes.

For the second part, we establish that for any $\varepsilon > 0$, there always exists a $\pi' \in \Pi^{\text{DHD}}$ so that

$$U_t^{\pi'}(h_t) + (Z - t)\varepsilon \geqslant U_t(h_t) \tag{A5}$$

for all $h_t \in H_t$ and $t = 0,1, \cdots, Z$. Since $U_Z^{\pi'}(h_Z) = U_Z(h_Z) = \overline{r}_{s_Z}^0$, the induction hypothesis holds for $t = Z$. Assuming that $U_t^{\pi'}(h_t) + (Z - t)\varepsilon \geqslant U_t(h_t)$ for $t = n + 1, \cdots, Z$, we have

$$\begin{aligned} U_n^{\pi'}(h_n) &= \overline{r}_{s_n}^{x_n(h_n)} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n, s_{n+1}}^{x_n(h_n)} U_{n+1}^{\pi'} \big(h_n, x_n(h_n), \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle \big) \\ &\geqslant \overline{r}_{s_n}^{x_n(h_n)} + \gamma \sum_{s_{n+1} \in S} \overline{p}_{s_n, s_{n+1}}^{x_n(h_n)} U_{n+1} \big(h_n, x_n(h_n), \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}} \rangle \big) - (Z - n - 1)\varepsilon \\ &\geqslant U_n(h_n) - (Z - n)\varepsilon. \end{aligned} \tag{A6}$$

The second line in (A6) holds because of the induction hypothesis and $\varepsilon > 0$. Thus the inductive hypothesis is satisfied and (A5) holds for $t = 0,1, \cdots, Z$.

Thus for any $\varepsilon > 0$, there exists a $\pi' \in \Pi^{\text{DHR}}$ for which

$$U_t^*(h_t) + (Z - t)\varepsilon \geqslant U_t^{\pi'}(h_t) + (Z - t)\varepsilon \geqslant U_t(h_t) \geqslant U_t^*(h_t),$$

then part (i) in the theorem holds. Part (ii) in the theorem holds because

$$U_0(h_0) = U_0^*(h_0) = \max_{\pi \in \Pi^{\text{DHR}}} U_0^{\pi}(h_0) = \max_{\pi \in \Pi^{\text{DHR}}} V^{\pi}\big(s_0, \boldsymbol{b}_{s_0}\big) = V^*\big(s_0, \boldsymbol{b}_{s_0}\big).$$

Proof of Theorem 4

For $t = Z$, clearly $U_Z^{\pi^*}(h_Z) = U_Z^*(h_Z)$ for all $h_Z \in H_Z$. Assume argument in part (i) in the theorem is true for $t = n + 1, \cdots, Z$. Then, for $t = n$ and $h_n = \left(h_{n-1}, x_{n-1}^*(h_{n-1}), \langle s_n, \boldsymbol{b}_{s_n}\rangle\right)$,

$$\begin{aligned} U_n^*(h_n) &= \max_{a_n \in A}\left\{\overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1}\in S} \overline{p}_{s_n,s_{n+1}}^{a_n} U_{n+1}^*\big(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}}\rangle\big)\right\} \\ &= \overline{r}_{s_n}^{x_n^*(h_n)} + \gamma \sum_{s_{n+1}\in S} \overline{p}_{s_n,s_{n+1}}^{x_n^*(h_n)} U_{n+1}^{\pi^*}\big(h_n, x_n^*(h_n), \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}}\rangle\big) = U_n^{\pi^*}(h_n). \end{aligned}$$

Thus the induction hypothesis is true. Part (ii) in the theorem follows from Theorem 1 and Theorem 3-(ii).

Proof of Theorem 6

We show that (i) holds by induction. Since $U_Z^*(h_Z) = U_Z^*\big(h_{Z-1}, a_{Z-1}, \langle s_Z, \boldsymbol{b}_{s_Z}\rangle\big) = \overline{r}_{s_Z}^0$ for all $h_{Z-1} \in H_{Z-1}$ and $a_{Z-1} \in A$, $U_Z^*(h_Z) = U_Z^*\big(s_Z, \boldsymbol{b}_{s_Z}\big)$. Assume now that (i) is true for $t = n + 1, \cdots, Z$. Then let $t = n$. For any $h_n = \big(h_{n-1}, a_{n-1}, \langle s_n, \boldsymbol{b}_{s_n}\rangle\big) \in H_n$, it follows from (24), the induction hypothesis and (14) that

$$\begin{aligned} U_n^*(h_n) &= \max_{a_n\in A}\left\{\overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1}\in S} \overline{p}_{s_n,s_{n+1}}^{a_n} U_{n+1}^*\big(h_n, a_n, \langle s_{n+1}, \boldsymbol{b}_{s_{n+1}}\rangle\big)\right\} \\ &= \max_{a_n\in A}\left\{\overline{r}_{s_n}^{a_n} + \gamma \sum_{s_{n+1}\in S} \overline{p}_{s_n,s_{n+1}}^{a_n} U_{n+1}^*\big(s_{n+1}, \boldsymbol{b}_{s_{n+1}}\big)\right\} \\ &= \max_{a_n\in A}\left\{\sum_{k\in C} b_{s_n,k} r_{s_n,k}^{a_n} + \gamma \sum_{s_{n+1}\in S}\sum_{k\in C} b_{s_n,k} p_{s_n,s_{n+1},k}^{a_n} U_{n+1}^*\big(s_{n+1}, \tau\big(\langle s_n, \boldsymbol{b}_{s_n}\rangle, a_n, s_{n+1}\big)\big)\right\}. \end{aligned} \tag{A7}$$

Because the quantities within brackets in the last line of (A7) depends on h_n only through $\langle s_n, \boldsymbol{b}_{s_n}\rangle$, we have $U_n^*(h_n) = U_n^*\big(s_n, \boldsymbol{b}_{s_n}\big)$. Thus part (i) in the theorem is true for $t = 0,1, \cdots, Z$.

When S and A are finite, there exists a policy $\pi^* = \big(x_0^*, x_1^*, \ldots, x_{Z-1}^*\big) \in \Pi^{\text{DMD}}$ derived from

$$x_t^*\big(s_t, \boldsymbol{b}_{s_t}\big) \in \operatorname*{argmax}_{a_t\in A}\left\{\sum_{k\in C} b_{s_t,k} r_{s_t,k}^{a_t} + \gamma \sum_{s_{t+1}\in S}\sum_{k\in C} b_{s_t,k} p_{s_t,s_{t+1},k}^{a_t} U_{t+1}^*\big(s_{t+1}, \tau\big(\langle s_t, \boldsymbol{b}_{s_t}\rangle, a_t, s_{t+1}\big)\big)\right\}.$$

Therefore, by part (i) in the theorem and Theorem 4-(ii), π^* is optimal.

Proof of Theorem 7

We will show that $W^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right)$ in (4) is equal to $V^{\pi}(s_0, \boldsymbol{b}_{s_0})$ in (16). Firstly, we provide another calculation of scenario belief $\boldsymbol{b}_{s_t}$ in state s_t at decision epoch t. For any $h_0 = \langle s_0, \boldsymbol{b}_{s_0} \rangle \in S \times \Delta^C$, using history recursion $h_t = (h_{t-1}, a_{t-1}, \langle s_t, \tau(\langle s_{t-1}, \boldsymbol{b}_{s_{t-1}} \rangle, a_{t-1}, s_t) \rangle)$ repeatedly, the belief $\boldsymbol{b}_{s_t}$ within $\langle s_t, \boldsymbol{b}_{s_t} \rangle$ in $h_t \in H_t$ can be expressed by

$$b_{s_t,k} = \frac{b_{s_0,k} p^{a_0}_{s_0,s_1,k} p^{a_1}_{s_1,s_2,k} \cdots p^{a_{t-1}}_{s_{t-1},s_t,k}}{\sum\limits_{\mathrm{k}' \in \mathrm{C}} b_{s_0,k'} p^{a_0}_{s_0,s_1,k'} p^{a_1}_{s_1,s_2,k'} \cdots p^{a_{t-1}}_{s_{t-1},s_t,k'}}, \forall k \in C, \tag{A8}$$

where s_0, a_0, s_1, ···, a_{t-1}, s_t are the state realizations and actions taken up to t.

For a fixed policy $\pi \in \Pi^{\mathrm{DHD}}$, starting from the boundary condition $U^{\pi}_{Z}(h_Z) = \overline{r}^{0}_{s_Z}$, we evaluate recursively U^{π}_{t} for $t = Z - 1, \cdots, 1, 0$ by (22), that is,

$$\begin{aligned} U^{\pi}_{Z-1}(h_{Z-1}) &= \overline{r}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1}} + \gamma \sum_{s_Z \in S} \overline{p}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},s_Z} U^{\pi}_{Z}\left(h_{Z-1}, x_{Z-1}(h_{Z-1}), \langle s_Z, \boldsymbol{b}_{s_Z} \rangle\right) \\ &= \overline{r}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1}} + \gamma \sum_{s_Z \in S} \overline{p}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},s_Z} \overline{r}^{0}_{s_Z}, \forall h_{Z-1} \in H^{\pi}_{Z-1}, \end{aligned}$$

$$\begin{aligned} U^{\pi}_{Z-2}(h_{Z-2}) &= \overline{r}^{x_{Z-2}(h_{Z-2})}_{s_{Z-2}} + \gamma \sum_{s_{Z-1} \in S} \overline{p}^{x_{Z-2}(h_{Z-2})}_{s_{Z-2},s_{Z-1}} U^{\pi}_{Z-1}\left(h_{Z-2}, x_{Z-2}(h_{Z-2}), \langle s_{Z-1}, \boldsymbol{b}_{s_{Z-1}} \rangle\right) \\ &= \overline{r}^{x_{Z-2}(h_{Z-2})}_{s_{Z-2}} + \gamma \sum_{s_{Z-1} \in S} \overline{p}^{x_{Z-2}(h_{Z-2})}_{s_{Z-2},s_{Z-1}} \left\{ \overline{r}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1}} + \gamma \sum_{s_Z \in S} \overline{p}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},s_Z} \overline{r}^{0}_{s_Z} \right\}, \forall h_{Z-2} \in H^{\pi}_{Z-2}, \end{aligned}$$

$\cdots,$

$$\begin{aligned} U^{\pi}_{0}(h_0) = \overline{r}^{x_0(h_0)}_{s_0} + \gamma \sum_{s_1 \in S} \overline{p}^{x_0(h_0)}_{s_0,s_1} \Bigg\{ \overline{r}^{x_1(h_1)}_{s_1} + \gamma \sum_{s_2 \in S} \overline{p}^{x_1(h_1)}_{s_1,s_2} \\ \cdot \left\{ \overline{r}^{x_2(h_2)}_{s_2} + \ldots + \gamma \sum_{s_{Z-1} \in S} \overline{p}^{x_{Z-2}(h_{Z-2})}_{s_{Z-2},s_{Z-1}} \left\{ \overline{r}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1}} + \gamma \sum_{s_Z \in S} \overline{p}^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},s_Z} \overline{r}^{0}_{s_Z} \right\} \ldots \right\} \Bigg\}. \end{aligned} \tag{A9}$$

Submitting (14) into the last equation in (A9) and using (A8), we have

$$\begin{aligned} U^{\pi}_{0}(h_0) = \sum_{k \in C} b_{s_0,k} \Bigg\{ & r^{x_0(h_0)}_{s_0,k} + \gamma \sum_{s_1 \in S} p^{x_0(h_0)}_{s_0,s_1,k} r^{x_1(h_1)}_{s_1,k} + \gamma^2 \sum_{s_1 \in S} \sum_{s_2 \in S} p^{x_0(h_0)}_{s_0,s_1,k} p^{x_1(h_1)}_{s_1,s_2,k} r^{x_2(h_2)}_{s_2,k} \\ & + \ldots + \gamma^{Z-1} \sum_{s_1 \in S} \sum_{s_2 \in S} \cdots \sum_{s_{Z-1} \in S} p^{x_0(h_0)}_{s_0,s_1,k} p^{x_1(h_1)}_{s_1,s_2,k} \cdots p^{x_{Z-2}(h_{Z-2})}_{s_{Z-2},s_{Z-1},k} r^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},k} \\ & + \gamma^{Z} \sum_{s_1 \in S} \sum_{s_2 \in S} \cdots \sum_{s_{Z-1} \in S} \sum_{s_Z \in S} p^{x_0(h_0)}_{s_0,s_1,k} p^{x_1(h_1)}_{s_1,s_2,k} \cdots p^{x_{Z-2}(h_{Z-2})}_{s_{Z-2},s_{Z-1},k} p^{x_{Z-1}(h_{Z-1})}_{s_{Z-1},s_Z,k} r^{0}_{s_Z,k} \Bigg\}. \end{aligned} \tag{A10}$$

On the other hand, Steimle et al. [4] infer that there are no optimal policies that are Markovian for the problem in (5). Then, we consider the history-dependent deterministic policies of standard MDPs, $\pi = \{x_t(h_t) : h_t \in S \times A \times \cdots \times A \times S, t \in T\}$. Using the finite horizon policy evaluation algorithm presented in [1] and the same logic as in (A9), we obtain $V_k^{\pi}(s_0)$ in (3) as follows

$$\begin{aligned} V_k^{\pi}(s_0) = {} & r_{s_0,k}^{x_0(h_0)} + \gamma \sum_{s_1 \in S} p_{s_0,s_1,k}^{x_0(h_0)} r_{s_1,k}^{x_1(h_1)} + \gamma^2 \sum_{s_1 \in S} p_{s_0,s_1,k}^{x_0(h_0)} \sum_{s_2 \in S} p_{s_1,s_2,k}^{x_1(h_1)} r_{s_2,k}^{x_2(h_2)} \\ & + \ldots + \gamma^{Z-1} \sum_{s_1 \in S} p_{s_0,s_1,k}^{x_0(h_0)} \sum_{s_2 \in S} p_{s_1,s_2,k}^{x_1(h_1)} \cdots \sum_{s_{Z-1} \in S} p_{s_{Z-2},s_{Z-1},k}^{x_{Z-2}(h_{Z-2})} r_{s_{Z-1},k}^{x_{Z-1}(h_{Z-1})} \\ & + \gamma^Z \sum_{s_1 \in S} p_{s_0,s_1,k}^{x_0(h_0)} \sum_{s_2 \in S} p_{s_1,s_2,k}^{x_1(h_1)} \cdots \sum_{s_{Z-1} \in S} p_{s_{Z-2},s_{Z-1},k}^{x_{Z-2}(h_{Z-2})} \sum_{s_Z \in S} p_{s_{Z-1},s_Z,k}^{x_{Z-1}(h_{Z-1})} r_{s_Z,k}^{0}. \end{aligned} \tag{A11}$$

It holds from (4), (A11) and (A10) that

$$W^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right) = \sum_{k \in C} b_{s_0,k} V_k^{\pi}(s_0) = U_0^{\pi}(h_0) = V^{\pi}\left(s_0, \boldsymbol{b}_{s_0}\right).$$

Then it follows from (5) and (29) that the finite horizon DFMDP is equivalent to the WVP of the multi-scenario MDP.

Proof of Theorem 8

For a given $\langle s_0, \boldsymbol{b}_{s_0} \rangle$, (10) becomes

$$b_{s_t,k} = \frac{b_{s_{t-1},k} p_{s_{t-1},s_t,0}^{a_{t-1}}}{p_{s_{t-1},s_t,0}^{a_{t-1}} \sum_{k' \in \mathrm{C}} b_{s_{t-1},k'}} = b_{s_{t-1},k}, \forall a_{t-1} \in A, s_{t-1}, s_t \in S, k \in C,$$

since $\boldsymbol{p}_1 = \ldots = \boldsymbol{p}_{|C|} = \boldsymbol{p}_0$. This implies that the all scenario beliefs for any state at any time are constant and equal to $\boldsymbol{b}_{s_0}$. As a result, $\overline{r}_{s_t}^{a_t} = \sum_{k \in C} b_{s_0,k} r_{s_t,k}^{a_t}, \forall s_t \in S, a_t \in A, t \in T$, i.e., $\overline{\boldsymbol{r}} = \sum_{k \in C} b_{s_0,k} \boldsymbol{r}_k$ which means that the scenario expected reward does not change over time. Therefore, the DFMDP with the type-I of scenarios can be reduced to a standard MDP with parameter pair ($\boldsymbol{p}_0$,$\overline{\boldsymbol{r}}$).

Proof of Theorem 9

Firstly, we will prove that H is an isotone mapping. Let V, $U \in \mathcal{V}$ and $U \geqslant V$. Suppose that

$$a^*(s, \boldsymbol{b}_s) \in \operatorname*{argmax}_{a \in A} \left\{ \sum_{k \in C} b_{s,k} r_{s,k}^{a} + \gamma \sum_{s' \in S} \sum_{k \in C} b_{s,k} p_{s,s',k}^{a} V\left(s', \tau\left(\langle s, \boldsymbol{b}_s \rangle, a, s'\right)\right) \right\},$$

then for any $\langle s, \boldsymbol{b}_s\rangle \in S \times \Delta^C$,

$$\begin{aligned}
&(HU)(s, \boldsymbol{b}_s) - (HV)(s, \boldsymbol{b}_s) \\
&\geqslant \sum_{k\in C} b_{s,k} r_{s,k}^{a^*(s,\boldsymbol{b}_s)} + \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} U\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) \\
&\quad - \sum_{k\in C} b_{s,k} r_{s,k}^{a^*(s,\boldsymbol{b}_s)} - \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} V\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) \\
&= \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} \big\{\text{argmax}\, U\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) - V\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big)\big\} \geqslant 0.
\end{aligned}$$

It shows that $HU \geqslant HV$ when $U \geqslant V$. So H is an isotone mapping.
Then we will prove that for all V, $U \in \mathcal{V}$, that $\|HV - HU\| \leqslant \gamma\|V - U\|$ is true for any $0 < \gamma < 1$. Let $V, U \in \mathcal{V}$ and $HV(s, \boldsymbol{b}_s) \geqslant HU(s, \boldsymbol{b}_s)$ for a fixed $\langle s, \boldsymbol{b}_s\rangle \in S \times \Delta^C$. Again, suppose that

$$a^*(s, \boldsymbol{b}_s) \in \underset{a\in A}{\text{argmax}} \left\{ \sum_{k\in C} b_{s,k} r_{s,k}^a + \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^a V\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a, s'\big)\big) \right\}.$$

Then

$$\begin{aligned}
0 &\leqslant (HV)(s, \boldsymbol{b}_s) - (HU)(s, \boldsymbol{b}_s) \\
&\leqslant \sum_{k\in C} b_{s,k} r_{s,k}^{a^*(s,\boldsymbol{b}_s)} + \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} V\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) \\
&\quad - \sum_{k\in C} b_{s,k} r_{s,k}^{a^*(s,\boldsymbol{b}_s)} - \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} U\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) \\
&= \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} \big\{V\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big) - U\big(s', \tau\big(\langle s, \boldsymbol{b}_s\rangle, a^*(s, \boldsymbol{b}_s), s'\big)\big)\big\} \\
&\leqslant \gamma \sum_{s'\in S}\sum_{k\in C} b_{s,k} p_{s,s',k}^{a^*(s,\boldsymbol{b}_s)} \|V - U\| \leqslant \gamma\|V - U\|.
\end{aligned}$$

If we assume $HU(s, \boldsymbol{b}_s) \geqslant HV(s, \boldsymbol{b}_s)$, the same logic will imply that

$$|(HV)(s, \boldsymbol{b}_s) - (HU)(s, \boldsymbol{b}_s)| \leqslant \gamma\|V - U\|$$

for any $\langle s, \boldsymbol{b}_s\rangle \in S \times \Delta^C$. This results in $\| HV - HU\| \leqslant \gamma\|V - U\|$ So H is a contraction mapping.

References

[1] Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming, 2nd edn. Wiley, New Jersey (2005)

[2] Beemsterboer, B.J., Land, M.J., Teunter, R.H.: Flexible lot sizing in hybrid make-to-order/make-to-stock production planning. Eur. J. Oper. Res. **260**(3), 1014–1023 (2017)

[3] Chen, N., Teven, K., Wang, C.: A partitioning algorithm for markov decision processes with applications to market microstructure. Manag. Sci. **64**(2), 784–803 (2018)

[4] Steimle, L.N., Kaufman, D.L., Denton, B.T.: Multi-model Markov decision processes. IISE Trans. **53**(10), 1124–1139 (2022)

[5] Buchholz, P., Scheftelowitsch, D.: Computation of weighted sums of rewards for concurrent MDPs. Math. Methods Oper. Res. **89**(1), 1–42 (2019)

[6] Givan, R., Leach, S., Dean, T.: Bounded-parameter Markov decision processes. Artif. Intell. **122**(1–2), 71–109 (2000)

[7] Delgado, K.V., Sanner, S., de Barros, L.N.: Efficient solutions to factored MDPs with imprecise transition probabilities. Artif. Intell. **175**(9–10), 1498–1527 (2011)

[8] Witwicki, S.J., Melo, F.S., Capitan, J., Spaan, M.T.J.: A flexible approach to modeling unpredictable events in MDPs. In: Proceedings of Twenty-Third International Conference on Automated Planning and Scheduling ICAPS2013, pp. 260–268 (2013)

[9] Duff, M.: Optimal learning: computational procedures for bayes-adaptive markov decision processes. Ph.D. thesis, University of Massachusetts Amherst, Amherst, MA (2002)

[10] Castro, P. S., Precup, D.: Using linear programming for Bayesian exploration in Markov decision processes. In: International Joint Conference on Artificial Intelligence IJCAI2007, pp. 2437–2442 (2007)

[11] Kumar, P.: Information theoretic learning methods for Markov decision processes with parametric uncertainty. Ph.D. thesis, University of Washington (2018).

[12] Iyengar, G.: Robust dynamic programming. Math. Oper. Res. **30**(2), 257–280 (2005)

[13] Nilim, A., El Ghaoui, L.: Robust control of Markov decision processes with uncertain transition matrices. Oper. Res. **53**(5), 780–798 (2005)

[14] Delgado, K.V., de Barros, L.N., Dias, D.B., Sanner, S.: Real-time dynamic programming for Markov decision process with imprecise probabilities. Artif. Intell. **230**(8), 192–223 (2016)

[15] Moreira, D.A.M., Delgado, K.V., de Barros, L.N.: Robust probabilistic planning with ilao. Appl. Intell. **45**(3), 662–672 (2016)

[16] Delage, E., Shie, M.: Percentile optimization for markov decision processes with parameter uncertainty. Oper. Res. **58**(1), 203–213 (2010)

[17] Adulyasak, Y., Varakantham, P., Ahmed, A., Jaillet, P.: Solving uncertain MDPs with objectives that are separable over instantiations of model uncertainty. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, AAAI Press, pp. 3454–3460 (2015)

[18] Ahmed, A., Varakantham, P., Lowalekar, M., Adulyasak, Y., Jaillet, P.: Sampling based approaches for minimizing regret in uncertain Markov decision processes (MDPs). J. Artif. Intell. Res. **59**, 229–264 (2017)

[19] Meraklı, M., Küçükyavuz, S.: Risk aversion to parameter uncertainty in Markov decision processes with an application to slow-onset disaster relief. IISE Trans. **52**(8), 811–831 (2019)

[20] Shani, G., Heckerman, D., Brafman, R.: An MDP-based recommender system. J. Mach. Learn. Res. **6**(43), 1265–1295 (2005)

[21] Chen, Q., Ayer, T., Chhatwal, J.: Sensitivity analysis in sequential decision models: a probabilistic approach. Med. Decis. Making **37**(2), 243–252 (2017)

[22] Bala, M.V., Mauskopf, J.A.: Optimal assignment of treatments to health states using a Markov decision model. Pharmacoeconomics **24**(4), 345–354 (2006)