Information Sciences 690 (2025) 121578

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Discounted fully probabilistic design of decision rules ☆

Miroslav Kárný *, Soňa Molnárová

The Czech Academy of Sciences, Institute of Information Theory and Automation, Department of Adaptive Systems, vodárenskou věží 4, Prague 8, 182 00, Czech Republic

ARTICLE INFO

Keywords:
Design principles
Kullback-Leibler's divergence
Probabilistic techniques
Discounting
Closed loop

ABSTRACT

Axiomatic fully probabilistic design (FPD) of optimal decision rules strictly extends the decision making (DM) theory represented by Markov decision processes (MDP). This means that any MDP task can be approximated by an explicitly found FPD task whereas many FPD tasks have no MDP equivalent. MDP and FPD model the closed loop — the coupling of an agent and its environment — via a joint probability density (pd) relating the involved random variables, referred to as behaviour. Unlike MDP, FPD quantifies agent's aims and constraints by an *ideal pd*. The ideal pd is high on the desired behaviours, small on undesired behaviours and zero on forbidden ones. FPD selects the optimal decision rules as the minimiser of Kullback-Leibler's divergence of the closed-loop-modelling pd to its ideal twin. The proximity measure choice follows from the FPD axiomatics.

MDP minimises the expected total loss, which is usually the sum of discounted partial losses. The discounting reflects the decreasing importance of future losses. It also diminishes the influence of errors caused by:

- the imperfection of the employed environment model;
- roughly-expressed aims;
- the approximate learning and decision-rules design.

The established FPD cannot currently account for these important features. The paper elaborates the missing discounted version of FPD. This non-trivial filling of the gap in FPD also employs an extension of dynamic programming, which is of an independent interest.

1. Introduction

An agent — a human, a device or a mixed group of both, referred to as "it" — chooses its actions in order to meet its aims. This is the core of any decision making that always runs under uncertainty. The inspected prescriptive Bayesian DM theory has its roots in [1,2]. It underlies the theory of Markov decision processes [3], which is the standard way of designing optimal, action-generating, decision rules. Stochastic control does the same [4,5] but its stress and vocabulary differ. They are often used interchangeably or jointly[1] [6].

☆ This research was supported by EU-COST Action CA21169.

* Corresponding author.
E-mail addresses: school@utia.cas.cz (M. Kárný), molnarova@utia.cas.cz (S. Molnárová).
URLs: https://www.utia.cas.cz/people/karny (M. Kárný), https://www.utia.cas.cz/people/moln-rov (S. Molnárová).

[1] References are just samples from a much more extensive set.

https://doi.org/10.1016/j.ins.2024.121578
Received 2 January 2024; Received in revised form 27 August 2024; Accepted 19 October 2024

Available online 22 October 2024

The MDP design often employs discounting, which perceives DM results at a distant future as less important [7–9]. This weakens the adverse impact of imprecision (of any origin) on the design of decision rules [10,11]. It also simplifies the error analysis of various approximate designs of decision rules [12,13].

This text and [14] show that FPD, an abstract axiomatic version [14,15] of the model reference control [16], strictly extends MDP. Discounted FPD has not been yet developed regardless of its rich history [17–21], its tight connection with the independently proposed KL control[2] [23–26] and with the KL regularised control [27,28]. The paper fills this gap by elaborating the discounted FPD.

MDP tasks mapped on FPD tasks, see Proposition 3 below, naturally come with discounting. However, FPD tasks with no MDP equivalents lack it. The discounting extension to *any* FPD task is highly desirable to reach its positive effects.

The choice of the discounting rate is hard even if it has a monetary interpretation [29]. A sound choice is vital especially when DM is connected with sensitive aspects, say health [9]. The choice is even harder when discounting copes with doubts about persistency of the used environment model or of the quantified aims.

The importance of a sound choice of the discounting rate lies in the impact of the effective shortening of the design horizon. It may decrease the DM quality up to the closed-loop instability [30]. On the other hand, the optimisation without discounting may excite modelling errors and lead to poor DM quality up to the instability [31]. Sec. 4 illustrates the positive effect of discounting on diminishing the consequences of modelling errors.

The choice of the discounting rate is expected to be easier for FPD. The unified probabilistic language of FPD concerns both the closed-loop modelling and the aim expression. Thus FPD inputs, including the discounting rate, are quantified by probabilities. The meaning and role of the discounting rate in the design part is fully analogical to the data and time varying forgetting rate used in learning [32,33]. At the same time, the on-line learning of the forgetting rate is feasible [34]. These facts open a way to a data-based choice of the discounting rate within FPD. Due to the inclusion of MDP tasks in the set of FPD tasks this will help to choose the discounting rate for MDP as well. Initial attempts are in [35].

Layout

Sec. 2 recalls FPD and refines its relation to MDP. A generalisation of the usual dynamic programming [4] arises during this refinement. This overture serves to core Sec. 3, which presents the discounted FPD. Sec. 4 illustrates the impact of the theory numerically. Sec. 5 provides the concluding remarks.

Notation

The text uses decorated mnemonic labels: a action, D and d divergences, c closed-loop model, E expectation, h horizon, m model, N and n normalisations, L loss, p pointer, r decision rule(s), s state, t time, v value function, w weight. Sanserif fonts mark mappings. c, m and r are probability densities (pds[3]). Bold **fonts** mark the set: $\boldsymbol{x}$ is the set of xs and is defined only if needed. $\boldsymbol{x}^h$ is Cartesian product of h sets $\boldsymbol{x}$, h is a natural number. $:=$ defines the left-hand side by assigning the right-hand side. $\propto$ is proportionality. $\mathsf{f}(x_t) := \mathsf{f}_t(x_t)$, the double time subscript $t \in \boldsymbol{t}$ is dropped. $\mathsf{r}_t := (\mathsf{r}(a_t|s_{t-1}))_{a_t \in \boldsymbol{a}, s_{t-1} \in \boldsymbol{s}}$ are decision rules.[4] The initial state $s_0 \in \boldsymbol{s}$ implicitly conditions all pds. Superscript $^{\mathsf{i}}$ concerns the ideal pd, $^{\mathsf{o}}$ indicates optimality.

2. Preliminaries

The FPD recall makes the paper self-reliant. The proved version of the dynamic programming generalises the usual one. It is applicable to a total loss with partial losses dependent on the optimised decision rules.

Note that works [24,37,38] provide rich commented lists of references to a wide range of FPD-related results.

2.1. Fully probabilistic design

DM concerns the closed loop consisting of an agent and its uncertain environment. The agent gradually applies actions $a_t \in \boldsymbol{a} \neq \emptyset$ at time epochs $t \in \boldsymbol{t} := \{1, \dots, h\}$, $h \leq \infty$, influencing the closed loop. They stimulate a transition of the closed-loop state $s_{t-1} \in \boldsymbol{s} \neq \emptyset$ to the state $s_t \in \boldsymbol{s}$. The states and actions, thought up to the decision horizon h, form the closed-loop behaviours

$$b := (s_h, a_h, \dots, s_1, a_1) \in \boldsymbol{b} := (\boldsymbol{s}, \boldsymbol{a})^h.$$

The agent selects chooses $(a_t)_{t \in \boldsymbol{t}}$ via randomised decision rules

$$\mathsf{r} \in \mathbf{r} := \left\{ \mathsf{r}(b) := \prod_{t \in \boldsymbol{t}} \mathsf{r}(a_t|s_{t-1}) \right\}.$$

The conditional pds $\mathsf{r}(a_t|s_{t-1})$ model the causal decision rules r. The decision-rules-dependent joint pd $\mathsf{c}^{\mathsf{r}}(b)$ completely describes random behaviours $b \in \boldsymbol{b}$. The state definition and the chain rule for pds [39] imply the factorised closed-loop model

$$\mathsf{c}^{\mathsf{r}}(b) = \prod_{t \in \boldsymbol{t}} \mathsf{m}(s_t|a_t, s_{t-1}) \mathsf{r}(a_t|s_{t-1}) := \mathsf{m}(b)\mathsf{r}(b). \tag{1}$$

[2] KL stands for Kullback-Leibler's divergence [22]. The use of KL control is adopted.

[3] Pd is Radon-Nikodým's derivative with respect to either Lebesgue's or counting measure [36]. Lebesgue's notation is taken as the generic one.

[4] As usual, | separates conditions in pds as well as in expectations.

The known environment model $\mathsf{m}(b) := \prod_{t\in\boldsymbol{t}} \mathsf{m}(s_t|a_t, s_{t-1})$ consists of the conditional pds $\mathsf{m}(s_t|a_t, s_{t-1})$, $s_t, s_{t-1} \in \boldsymbol{s}$, $a_t \in \boldsymbol{a}$, $t \in \boldsymbol{t}$, modelling state transitions.

Remark 1 *(On the Closed-Loop State Relevance).* Many DM setups neglect the need to use the *closed-loop* state whenever the environment model results from on-line estimation [39]. Indeed, if learning and acting run in parallel, the state includes values of the used statistic as actions influence them. The optimal decision rules thus balance exploration and exploitation efforts [40]. Without a reflection of this fact, the exploration becomes the extra challenging task requiring sophisticated ad hoc techniques [41]. □

FPD quantifies agent's aims and constraints by an ideal (desired) joint pd $\mathsf{c}^{\mathsf{i}}(b)$, $b \in \boldsymbol{b}$. The pd $\mathsf{c}^{\mathsf{i}}(b)$ assigns high values to desired behaviours b, small values to undesired bs and zero to forbidden bs. The ideal closed-loop model factorises in the same way as the pd $\mathsf{c}^{\mathsf{r}}(b)$ in (1)

$$\mathsf{c}^{\mathsf{i}}(b) = \prod_{t\in\boldsymbol{t}} \mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}) := \mathsf{m}^{\mathsf{i}}(b)\mathsf{r}^{\mathsf{i}}(b). \tag{2}$$

The agent-selected ideal environment model

$$\mathsf{m}^{\mathsf{i}}(b) := \prod_{t\in\boldsymbol{t}} \mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1}),\ s_t, s_{t-1} \in \boldsymbol{s},\ a_t \in \boldsymbol{a},$$

combines the conditional pds $\mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})$, $t \in \boldsymbol{t}$, modelling the desired state transitions. The agent-selected ideal decision rules

$$\mathsf{r}^{\mathsf{i}}(b) := \prod_{t\in\boldsymbol{t}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}),\ a_t \in \boldsymbol{a},\ s_{t-1} \in \boldsymbol{s}$$

consist of the desired decision rules given by the pds $\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})$. Note that the definition of $\mathsf{m}^{\mathsf{i}}(b)$ in (2) implies the normalisation used in Sec. 2.2

$$\int_{\boldsymbol{s}^h} \mathsf{m}^{\mathsf{i}}(b)\,\mathrm{d}(s_h, \dots, s_1) = 1. \tag{3}$$

The FPD-optimal decision rules $\mathsf{r}^{\mathsf{o}} \in \mathbf{r}$ [14] minimise Kullback-Leibler's divergence (KL, [22]) $\mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}})$ of c^{r} to c^{i}

$$\mathsf{r}^{\mathsf{o}} \in \operatorname{Arg}\min_{\mathsf{r}\in\mathbf{r}} \mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}}) := \operatorname{Arg}\min_{\mathsf{r}\in\mathbf{r}} \int_{\boldsymbol{b}} \mathsf{c}^{\mathsf{r}}(b) \ln\left(\frac{\mathsf{c}^{\mathsf{r}}(b)}{\mathsf{c}^{\mathsf{i}}(b)}\right) \mathrm{d}b. \tag{4}$$

Dynamic programming [4] provides the optimal decision rules r^{o} (4). They are designed in a few steps using the following lemmas.

Lemma 1 *(Additive Form of KL). For decision rules* $\mathsf{r} \in \mathbf{r}$ *and the function*

$$\mathsf{L}^{\mathsf{r}}_{\mathsf{D}}(s_t, a_t, s_{t-1}) := \ln\left(\frac{\mathsf{m}(s_t|a_t, s_{t-1})\mathsf{r}(a_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\right),$$

the functional D *in (4) is the decision-rules-dependent expectation* (E^{r}) *of* the total loss $\mathsf{L}^{hr}_{\mathsf{D}}$ *with the decision-rules-dependent addends* $\mathsf{L}^{\mathsf{r}}_{\mathsf{D}}$. *KL in (4) reads*

$$\begin{aligned}\mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}}) &= \sum_{t\in\boldsymbol{t}} \int_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s})} \mathsf{c}^{\mathsf{r}}(s_t, a_t, s_{t-1})\mathsf{L}^{\mathsf{r}}_{\mathsf{D}}(s_t, a_t, s_{t-1})\,\mathrm{d}(s_t, a_t, s_{t-1}) \\ &= \sum_{t\in\boldsymbol{t}} \int_{\boldsymbol{s}} \mathsf{c}^{\mathsf{r}}(s_{t-1}) \Big[\int_{(\boldsymbol{s},\boldsymbol{a})} \mathsf{m}(s_t|a_t, s_{t-1})\mathsf{r}(a_t|s_{t-1}) \\ &\quad \times \ln\left(\frac{\mathsf{m}(s_t|a_t, s_{t-1})\mathsf{r}(a_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\right) \mathrm{d}(s_t, a_t) \Big]\,\mathrm{d}s_{t-1} := \mathsf{E}^{\mathsf{r}}[\mathsf{L}^{hr}_{\mathsf{D}}].\end{aligned} \tag{5}$$

The employed marginal pd $\mathsf{c}^{\mathsf{r}}(s_{t-1})$ *of* $\mathsf{c}^{\mathsf{r}}(b)$ *is independent of the decision rules* $\mathsf{r}(a_\tau|s_{\tau-1})$ *with time indices* $\tau \geq t$.

Proof. Formula (5) follows from:
- the product in the ratio $\mathsf{c}^{\mathsf{r}}(b)/\mathsf{c}^{\mathsf{i}}(b)$, see (1), (2), yielding the sum of logarithms;
- linearity of the integration in its argument;
- the dependence of $\mathsf{L}^{\mathsf{r}}_{\mathsf{D}}(s_t, a_t, s_{t-1})$ on (s_t, a_t, s_{t-1}) allows other entries of $b \in \boldsymbol{b}$ to integrate out of the joint pd $\mathsf{c}^{\mathsf{r}}(b)$ reducing it to the marginal pd $\mathsf{c}^{\mathsf{r}}(s_t, a_t, s_{t-1})$;
- the chain rule for pds applied to $\mathsf{c}^{\mathsf{r}}(s_t, a_t, s_{t-1})$ using definitions of $\mathsf{m}(s_t|a_t, s_{t-1})$, $\mathsf{r}(a_t|s_{t-1})$, see (1), and of $\mathsf{c}^{\mathsf{r}}(s_{t-1})$;
- Fubini's theorem for multiple integrals [36].

The marginal pd $\mathsf{c}^{\mathsf{r}}(s_{t-1})$ results from the next multiple integration that runs over variables listed in differentials

$$\mathsf{c}^{\mathsf{r}}(s_{t-1}) = \int\limits_{(s^{h-1},a^h)} \prod_{\tau\in t} \mathsf{m}(s_\tau|a_\tau,s_{\tau-1}) \times \mathsf{r}(a_\tau|s_{\tau-1})\,\mathrm{d}(s_h,a_h,\dots,s_t,a_t,a_{t-1},s_{t-2},a_{t-2},\dots,s_1,a_1).$$

The pds for $\tau \geq t$ integrate to unity. This shows the independence of $\mathsf{c}^{\mathsf{r}}(s_{t-1})$ of $\mathsf{r}(a_\tau|s_{\tau-1}), \tau \geq t$. □

The expression (5) and the proved independence allow us to find the optimal decision rules (4) using the backward induction known as dynamic programming [4]. It provides the desired, causal, optimal decision rules. We need its slight generalisation, which considers the minimisation of the decision-rules–dependent expectation of the total loss L^{hr} with its *addends dependent on the optimised decision rules*. The optimal randomised decision rules r^{o} are

$$\mathsf{r}^{\mathsf{o}} \in \mathrm{Arg}\min_{\mathsf{r}\in\mathbf{r}} \int\limits_{\boldsymbol{b}} \sum_{t\in\boldsymbol{t}} \mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1})\mathsf{c}^{\mathsf{r}}(b)\,\mathrm{d}b := \mathrm{Arg}\min_{\mathsf{r}\in\mathbf{r}} \mathsf{E}^{\mathsf{r}}[\mathsf{L}^{hr}], \tag{6}$$

with the partial loss $\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1})$ dependent on the rules $\mathsf{r}(a_\tau|s_{\tau-1})$ with $\tau \leq t$.

Lemma 1 shows that the optimisation (4) is a subcase of (6). The existence of other cases might be important, see Remark 2. The next lemma solves the task (6).

Lemma 2 *(Dynamic programming). Let there exist stabilising decision rules $\bar{\mathsf{r}}$, making the expected total loss $\mathsf{E}^{\bar{\mathsf{r}}}[\mathsf{L}^{h\bar{r}}]$ in (6) finite. Let us define the value functions*

$$\mathsf{v}(s_{t-1}) := \min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau\geq t}} \sum_{\tau\geq t} \mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_\tau,a_\tau,s_{\tau-1})|s_{t-1}], \quad s_{t-1}\in\boldsymbol{s},\ \ t\in\boldsymbol{t}. \tag{7}$$

Then, $\mathsf{v}(s_{t-1})$ is bounded, $\mathsf{v}(s_h)=0$, and the next functional recursion holds

$$\mathsf{v}(s_{t-1}) = \min_{\mathsf{r}_t\in\mathbf{r}} \int\limits_{(s,a)} \mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})\left[\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1}) + \mathsf{v}(s_t)\right]\mathrm{d}(s_t,a_t). \tag{8}$$

The optimal decision rules r^{o} (6) are minimisers $\mathsf{r}^{\mathsf{o}}_t$ in (8) and

$$\mathsf{v}(s_0) = \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}\left[\mathsf{L}^{h\mathsf{r}^{\mathsf{o}}}\right]. \tag{9}$$

Proof. The existence of stabilising decision rules validates the next manipulations. For any $t\in\boldsymbol{t}$, the minimisation in (6) unfolds as $\min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau<t}}\left[\min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau\geq t}} \mathsf{E}^{\mathsf{r}}\left[\mathsf{L}^{hr}\right]\right]$. The inner minimisation reduces to

$$\zeta_{t-1} := \min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau\geq t}} \mathsf{E}^{\mathsf{r}}\Big[\sum_{\tau\geq t}\mathsf{L}^{\mathsf{r}}(s_\tau,a_\tau,s_{\tau-1})\Big] = \mathsf{E}^{\mathsf{r}}[\mathsf{v}(s_{t-1})]. \tag{10}$$

The second equality in (10) follows from:

- the chain rule for expectation $\mathsf{E}^{\mathsf{r}}[\bullet] = \mathsf{E}^{\mathsf{r}}\left[\mathsf{E}^{\mathsf{r}}[\bullet|s_{t-1}]\right]$ [36];
- the independence of $\mathsf{c}^{\mathsf{r}}(s_{t-1})$ from $(\mathsf{r}_\tau)_{\tau\geq t}$, which implies the commutativity of the outer unconditional expectation and the minimisation over $(\mathsf{r}_\tau)_{\tau\geq t}$;
- the value-function definition (7).

The smallest total expected loss in (6) coincides with ζ_0. Thus, the minimising rules in the ζ_0 definition (10) form the optimal decision rules. Moreover, the implicit conditioning by the known s_0 makes $\mathsf{v}(s_0)=\zeta_0$. This proves (9). It remains to prove the validity of the recursion (8).

The independence of $\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1})$ from rules $(\mathsf{r})_{\tau>t}$ implies

$$\begin{aligned}\mathsf{v}(s_{t-1}) &= \min_{\mathsf{r}_t\in\mathbf{r}}\left[\mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1})|s_{t-1}] + \min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau>t}}\sum_{\tau>t}\mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_\tau,a_\tau,s_{\tau-1})|s_{t-1}]\right]\\ &= \min_{\mathsf{r}_t\in\mathbf{r}}\left[\mathsf{E}^{\mathsf{r}}\left[\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1}) + \min_{(\mathsf{r}_\tau\in\mathbf{r})_{\tau>t}}\sum_{\tau>t}\mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_\tau,a_\tau,s_{\tau-1})|s_t]\Big|s_{t-1}\right]\right]\\ &= \min_{\mathsf{r}_t\in\mathbf{r}}\mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1}) + \mathsf{v}(s_t)\Big|s_{t-1}].\end{aligned}$$

The justification of the respective equalities above is as follows.

- The first equality uses (7) and the independence of $\mathsf{E}^{\mathsf{r}}[\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1})|s_{t-1}]$ from the rules $(\mathsf{r}_\tau)_{\tau>t}$.
- The second equality utilises the chain rule for expectations and the state definition. They imply $\mathsf{E}^{\mathsf{r}}[\bullet|s_{t-1}] = \mathsf{E}^{\mathsf{r}}\left[\mathsf{E}^{\mathsf{r}}[\bullet|s_t,s_{t-1}]\Big|s_{t-1}\right] = \mathsf{E}^{\mathsf{r}}\left[\mathsf{E}^{\mathsf{r}}[\bullet|s_t]\Big|s_{t-1}\right]$.

▶ The third equality results from the fact that the minimisation over $(\mathsf{r}_\tau)_{\tau>t}$ and the outer expectation in $\mathsf{E}^{\mathsf{r}}\left[\mathsf{E}^{\mathsf{r}}[\bullet|s_t]\middle|s_{t-1}\right]$ commute, and repetitive application of the definition (7). □

Remark 2 *(On Possible FPD Generalisations).* The FPD axiomatics [14] singled out KL among all f-divergences [42]. It avoids an unjustified coupling of optimal decision rules of two completely independent DMs solved as a single DM task. If this degenerate case is avoided by the problem formulation, then the axiomatics and Lemma 2 allow us to use any divergence in the role of the expected partial loss. This opens a way to a rich generalisation of FPD [43]. □

Lemmas 1, 2 serve to the *novel proof* of the solution of FPD. For an alternative proof, see [17].

Proposition 1 *(FPD). The backward induction reduces to the next backward functional recursion for* $\mathsf{n}(s_{t-1})\in[0,1]$, $s_{t-1}\in\boldsymbol{s}$, $t\in\boldsymbol{t}$, *with* $\mathsf{n}(s_h):=1$, *and* $a_t\in\boldsymbol{a}$,

$$\begin{aligned}\mathsf{d}(a_t,s_{t-1}) &:= \int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left(\frac{\mathsf{m}(s_t|a_t,s_{t-1})}{\mathsf{n}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})}\right)\mathrm{d}s_t\\ \mathsf{n}(s_{t-1}) &:= \int_{\boldsymbol{a}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})]\,\mathrm{d}a_t.\end{aligned}\tag{11}$$

The reached minimum is $\mathsf{v}(s_0):=-\ln(\mathsf{n}(s_0))=\mathsf{D}(\mathsf{c}^{\mathsf{r}^{\mathsf{o}}}||\mathsf{c}^{\mathsf{i}})$ *in (4). The optimal decision rules are*

$$\mathsf{r}^{\mathsf{o}}(a_t|s_{t-1})=\frac{\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})]}{\mathsf{n}(s_{t-1})}\propto\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})].\tag{12}$$

Proof. For non-negative KL, the value function $\mathsf{v}(s_{t-1})\geq 0$. Thus, it can be expressed in the form $-\ln(\mathsf{n}(s_{t-1})):=\mathsf{v}(s_{t-1})$. The non-negativity of $\mathsf{v}(s_{t-1})$ and the identity $\mathsf{v}(s_h)=0$ give $\mathsf{n}(s_t)\in[0,1]$ and $\mathsf{n}(s_h)=1$. For the addends $\mathsf{L}^{\mathsf{r}}(s_t,a_t,s_{t-1}):=\mathsf{L}^{\mathsf{r}}_{\mathsf{D}}(s_t,a_t,s_{t-1})$ in the expression (5), and for $\mathsf{r}_t:=\mathsf{r}(a_t|s_{t-1})$, $a_t\in\boldsymbol{a}$, $s_{t-1}\in\boldsymbol{s}$, the recursion (8) gets the form

$$\begin{aligned}\mathsf{v}(s_{t-1})=\min_{\mathsf{r}_t\in\mathsf{r}}\int_{(\boldsymbol{s},\boldsymbol{a})}&\mathsf{m}(s_t|a_t,s_{t-1})\\ &\times\mathsf{r}(a_t|s_{t-1})\left[\ln\left(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\right)+\mathsf{v}(s_t)\right]\mathrm{d}(s_t,a_t).\end{aligned}$$

The definition $\mathsf{v}(s_t):=-\ln(\mathsf{n}(s_t))$ and the fact that the pd $\mathsf{m}(s_t|a_t,s_{t-1})$ integrates to one over the states s_t in the set $\boldsymbol{s}$ imply

$$\begin{aligned}\mathsf{v}(s_{t-1})=\min_{\mathsf{r}_t\in\mathsf{r}}\int_{\boldsymbol{a}}&\mathsf{r}(a_t|s_{t-1})\left[\ln\left(\frac{\mathsf{r}(a_t|s_{t-1})}{\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\right)\right.\\ &\left.+\underbrace{\int_{\boldsymbol{s}}\mathsf{m}(s_t|a_t,s_{t-1})\ln\left(\frac{\mathsf{m}(s_t|a_t,s_{t-1})}{\mathsf{n}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})}\right)\mathrm{d}s_t}_{\mathsf{d}(a_t,s_{t-1}):=}\right]\mathrm{d}a_t.\end{aligned}$$

The second integral above is the function $\mathsf{d}(a_t,s_{t-1})$ used in the proposition formulation. The final recursion for the value function uses this function and a few elementary manipulations

$$\begin{aligned}\mathsf{v}(s_{t-1})&=\min_{\mathsf{r}_t\in\mathsf{r}}\left[\int_{\boldsymbol{a}}\mathsf{r}(a_t|s_{t-1})\ln\left(\frac{\mathsf{r}(a_t|s_{t-1})}{\mathsf{r}^{\mathsf{o}}(a_t|s_{t-1})}\right)\mathrm{d}a_t\right.\\ &\qquad\left.-\ln\Bigg(\underbrace{\int_{\boldsymbol{a}}\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp(-\mathsf{d}(a_t,s_{t-1}))\,\mathrm{d}a_t}_{\mathsf{n}(s_{t-1}):=}\Bigg)\right]\\ &=\min_{\mathsf{r}_t\in\mathsf{r}}\int_{\boldsymbol{a}}\mathsf{r}(a_t|s_{t-1})\ln\left(\frac{\mathsf{r}(a_t|s_{t-1})}{\mathsf{r}^{\mathsf{o}}(a_t|s_{t-1})}\right)\mathrm{d}a_t-\mathsf{v}(s_{t-1}).\end{aligned}$$

In the first equality above, just the logarithm $\ln(\mathsf{n}(s_{t-1}))$ of the $\mathsf{r}(a_t|s_{t-1})$-independent normalisation factor $\mathsf{n}(s_{t-1})$ is added and subtracted. Then, KL of the optimised pd $\mathsf{r}(a_t|s_{t-1})$ to $\mathsf{r}^{\mathsf{o}}(a_t|s_{t-1})$ (12) of $a_t\in\boldsymbol{a}$, conditioned on $s_{t-1}\in\boldsymbol{s}$, appears. It gets its smallest zero value for $\mathsf{r}(a_t|s_{t-1})=\mathsf{r}^{\mathsf{o}}(a_t|s_{t-1})$ and provides the value function $\mathsf{v}(s_{t-1})=-\ln(\mathsf{n}(s_{t-1}))\geq 0$. □

2.2. Mutual relationship of FPD and MDP

Introduction claimed that the FPD tasks strictly extend the class of MDP tasks. This subsection proves this claim and primarily points to a "normalisation problem" similar to that faced when formulating the discounted FPD in Sec. 3.

Propositions 2, 3 below motivate the way to the discounted FPD, Sec. 3. They are of an independent interest. Proposition 2 was proved in [44]. It shows how FPD reduces to MDP. The novel observation warns that this reduction cannot be directly inverted. Proposition 3 slightly refines the claim of [14] that the set of FPD tasks densely extends the set of MDP tasks. Thus, an FPD task can be found to each MDP task providing the decision rules arbitrarily close to the MDP-optimal ones.

FPD uses the ideal pd c^{i}, which is the product (2) of the ideal environment model m^{i} and the ideal decision rules r^{i}. Often, the agent is unwilling or unable to specify some factors forming them. In this case, it is wise to let FPD to identify these factors with their counterparts in c^{r}. When applied to a decision rule, then this *leave-to-the-fate* option (LTF, [44]) reads

$$\mathsf{r}^{\mathrm{i}} := \mathsf{r} \Leftrightarrow \text{ the ideal decision rule equals to the optimised one.} \tag{13}$$

Proposition 2 *(FPD under LTF (13) becomes MDP). Under LTF, FPD (4) reduces to MDP with the total, decision-rules-independent,*[5] *loss*

$$\mathsf{L}^h_{\mathsf{D}}(b) := \sum_{t\in\boldsymbol{t}} \mathsf{L}_{\mathsf{D}}(s_t, a_t, s_{t-1}) := \sum_{t\in\boldsymbol{t}} \ln\left(\frac{\mathsf{m}(s_t|a_t, s_{t-1})}{\mathsf{m}^{\mathrm{i}}(s_t|a_t, s_{t-1})}\right). \tag{14}$$

The MDP-optimal decision rules r^{o}, assigned to the sum of arbitrary decision-rules-independent partial losses $\mathsf{L}(s_t, a_t, s_{t-1})$ (including L_{D}), are deterministic. They generate the minimising arguments $a^{\mathrm{o}}(s_{t-1})$ in

$$\mathsf{v}(s_{t-1}) := \min_{a_t\in\boldsymbol{a}} \int_{\boldsymbol{s}} [\mathsf{L}(s_t, a_t, s_{t-1}) + \mathsf{v}(s_t)]\mathsf{m}(s_t|a_t, s_{t-1})\,\mathrm{d}s_t. \tag{15}$$

The backward functional recursion (15) coincides with the usual dynamic programming [4] and $\mathsf{v}(s_h) = 0$ initiates it.

Proof. Under LTF, decision rules r and ideal decision rules r^{i} cancel in the logarithm used in the definition (4). The minimised KL becomes the expectation (given by the closed-loop model $\mathsf{c}^{\mathrm{r}} = \mathsf{m}\mathsf{r}$ (1)) of the total loss $\mathsf{L}^h_{\mathsf{D}}(b)$ (14). The formulation coincides with the MDP in which the optimised functional (8) is linear in the optimised r_t. This implies the standard deterministic solution (15), [4]. □

Proposition 2 *seemingly* shows that any MDP, determined by the environment model $\mathsf{m}(b)$ and by its given total loss $\mathsf{L}^h(b)$, can be seen as a special case of FPD that:

- applies LTF (13), and;
- uses the ideal environment model m^{i} of the form

$$\mathsf{m}^{\mathrm{i}}(b) \propto \mathsf{m}(b)\exp\left[-\mathsf{L}^h(b)\right]. \tag{16}$$

This conclusion *is invalid* as the normalisation factor N in (16), guaranteeing the normalisation (3), depends on actions

$$\mathsf{N}(a_h, \ldots, a_1) = \int_{\boldsymbol{s}^h} \mathsf{m}(b)\exp\left[-\mathsf{L}^h(b)\right]\mathrm{d}(s_h, \ldots, s_1).$$

Thus, the choice of the ideal model (16) for LTF (13) leads to the minimisation of

$$\mathsf{D}(\mathsf{c}^{\mathrm{r}}||\mathsf{m}^{\mathrm{i}}\mathsf{r}) = \mathsf{E}^{\mathrm{r}}[\mathsf{L}^h] + \mathsf{E}^{\mathrm{r}}[\ln(\mathsf{N})], \tag{17}$$

instead of the desired minimisation of $\mathsf{E}^{\mathrm{r}}[\mathsf{L}^h]$ made within MDP. The good news is that any MDP task, given by an environment model $\mathsf{m}(b)$ and a total loss $\mathsf{L}^h(b)$, $b \in \boldsymbol{b}$, can be arbitrarily well approximated by a specific FPD task.

Proposition 3 *(FPD Strictly and Densely Extends MDP).*

1. *There are FPD tasks having no MDP equivalent.*
2. *Any MDP with a decision-rules-independent total loss L^h and stabilising decision rules $\bar{\mathsf{r}}$ making $\mathsf{E}^{\bar{\mathsf{r}}}[\mathsf{L}^h] < \infty$ can be approximated to an arbitrary precision by an FPD task with the same environment model $\mathsf{m}(b)$ and the ideal pd*

$$\mathsf{c}^{\mathrm{i}\lambda}(b) := \frac{\tilde{\mathsf{c}}(b)\exp[-\mathsf{L}^h(b)/\lambda]}{\int_{\boldsymbol{b}} \tilde{\mathsf{c}}(b)\exp[-\mathsf{L}^h(b)/\lambda]\,\mathrm{d}b}. \tag{18}$$

It is given by $\lambda > 0$ and a pd $\tilde{\mathsf{c}}(b) > 0$ on $\boldsymbol{b}$ making

[5] It explicitly depends on the behaviour realisation, not on the decision rules that influenced it.

i) *the pd* $\mathsf{c}^{\mathsf{i}\lambda}(b)$ *(18) a proper one with* $\int_{\boldsymbol{b}} \tilde{\mathsf{c}}(b)\exp[-\mathsf{L}^h(b)/\lambda]\,\mathrm{d}b < \infty$

ii) $\mathsf{D}(\mathsf{mr}||\tilde{c}) < \infty$ *for deterministic decision rules* r *with* $\mathsf{E}^{\mathsf{r}}[L^h] < \infty$. (19)

Proof. **ad 1.** The randomised nature of the FPD optimal decision rules, see Proposition 1, implies that they cannot be optimal for any MDP with a unique minimiser $a^{\mathsf{o}}(s_{t-1})$ in dynamic programming, see Proposition 2.

ad 2. The following decision rules r^{o}, $\mathsf{r}^{\mathsf{o}\lambda} \in \mathbf{r}$ are well-defined, for $\lambda > 0$ and the pd $\tilde{\mathsf{c}}(b) > 0$ used in (18),

$$\mathsf{r}^{\mathsf{o}} \in \mathrm{Arg}\min_{\mathsf{r}\in\mathbf{r}} \mathsf{E}^{\mathsf{r}}[\mathsf{L}^h], \text{ which guarantees } \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h] \le \mathsf{E}^{\tilde{\mathsf{r}}}[\mathsf{L}^h] < \infty$$
$$\mathsf{r}^{\mathsf{o}\lambda} \in \mathrm{Arg}\min_{\mathsf{r}\in\mathbf{r}} \left[\mathsf{E}^{\mathsf{r}}[\mathsf{L}^h] + \lambda\mathsf{D}(\mathsf{mr}||\tilde{\mathsf{c}})\right] \underbrace{=}_{(18)} \mathrm{Arg}\min_{\mathsf{r}\in\mathbf{r}} \mathsf{D}(\mathsf{mr}||\mathsf{c}^{\mathsf{i}\lambda}). \tag{20}$$

The joint pd $\mathsf{c}^{\mathsf{i}\lambda}(b)$ (18) exists due to the assumption (19) *i*). The definitions (20) of r^{o} and $\mathsf{r}^{\mathsf{o}\lambda}$ imply the next inequalities for the induced expected losses

$$0 \overbrace{\le}^{(20)} \mathsf{E}^{\mathsf{r}^{\mathsf{o}\lambda}}[\mathsf{L}^h] - \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h] \overbrace{\le}^{\lambda\mathsf{D}(\bullet||\bullet)\ge 0} \mathsf{E}^{\mathsf{r}^{\mathsf{o}\lambda}}[\mathsf{L}^h] + \lambda\mathsf{D}(\mathsf{mr}^{\mathsf{o}\lambda}||\tilde{\mathsf{c}}) - \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h]$$
$$\overbrace{\le}^{(20)} \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h] + \lambda\mathsf{D}(\mathsf{mr}^{\mathsf{o}}||\tilde{\mathsf{c}}) - \mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h] = \lambda\mathsf{D}(\mathsf{mr}^{\mathsf{o}}||\tilde{\mathsf{c}}) \to_{\lambda\to 0^+} 0.$$

The assumption *ii*) in (19) yields $\mathsf{D}(\mathsf{mr}^{\mathsf{o}}||\tilde{\mathsf{c}}) < \infty$. Thus, the expected loss $\mathsf{E}^{\mathsf{r}^{\mathsf{o}\lambda}}[\mathsf{L}^h]$, for the FPD-optimal decision rules $\mathsf{r}^{\mathsf{o}\lambda}$ (20), is arbitrarily close (for $\lambda \to 0^+$) to the expected loss $\mathsf{E}^{\mathsf{r}^{\mathsf{o}}}[\mathsf{L}^h]$ with the optimal decision rules r^{o} (20). □

Remark 3 *(On Generalisation of Proposition 3).* The proof of Proposition 3 adds the positive term $\lambda\mathsf{D}(\bullet||\bullet)$. λ-multiple of another f-divergence [42], say α-divergence [45,46], could be used to prove that the gained FPD generalisation, see Remark 2, densely extends MDP. The extension will be strict due to the non-linearity of the optimised criterion with respect to the decision rules making the optimal decision rules randomised. □

3. Main result: discounted FPD

First, the quest for the discounted FPD is shown to be nontrivial. Then, the presentation provides the DM formulation leading to the discounted FPD.

3.1. Unsuccessful attempt

The additive expression of the optimised KL, employing the expectation functional $\mathsf{E}^{\mathsf{r}}[\bullet]$ introduced in (6),

$$\mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}}) = \mathsf{E}^{\mathsf{r}}\Big[\sum_{t\in\boldsymbol{t}} \ln\Big(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\Big)\Big] \tag{21}$$

together with LTF (13), hint a way to the discounted FPD. It seems that it suffices to apply a *partial* LTF to factors forming the ideal closed-loop model and to use the weighted ideal pd

$$\begin{aligned}\mathsf{c}^{\mathsf{i}\mathsf{w}}(b) :=& \prod_{t\in\boldsymbol{t}} \mathsf{n}^{\mathsf{r}\mathsf{w}}(s_{t-1})^{-1}[\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})]^{\mathsf{w}(s_{t-1})}\\ &\times[\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})]^{1-\mathsf{w}(s_{t-1})}\\ \mathsf{n}^{\mathsf{r}\mathsf{w}}(s_{t-1}) :=& \int_{(\boldsymbol{s},\boldsymbol{a})} [\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})]^{\mathsf{w}(s_{t-1})}\\ &\times[\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})]^{1-\mathsf{w}(s_{t-1})}\,\mathrm{d}(s_t,a_t)\end{aligned} \tag{22}$$

with weights $\mathsf{w} := \big(\mathsf{w}(s_{t-1}) \in [0,1], s_{t-1} \in \boldsymbol{s}\big)_{t\in\boldsymbol{t}}$. It uses the ideal pds $\mathsf{m}^{\mathsf{i}}(b), \mathsf{r}^{\mathsf{i}}(b)$ from (2). Inserting (22) into (21) in the place of $\mathsf{c}^{\mathsf{i}}(b)$ yields

$$\mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}\mathsf{w}}) = \mathsf{E}^{\mathsf{r}}\Big[\sum_{t\in\boldsymbol{t}} \mathsf{w}(s_{t-1})\ln\Big(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\Big) + \ln(\mathsf{n}^{\mathsf{r}\mathsf{w}}(s_{t-1}))\Big]. \tag{23}$$

Above, the first addend after the summation sign has the desired form of the weighted partial loss but the unwanted logarithm of the normalisation $\mathsf{n}^{\mathsf{r}\mathsf{w}}(s_{t-1})$ appears as the second addend. Thus, the same problem as that met in connection with Proposition 2 (17), appears: *careless handling of the ideal pd changes the optimised functional in an undesired way.*

3.2. Successful approach

This part provides the remedy of the above problem. It formulates and solves the relevant FPD.

The adopted approach stems from the recognition that discounting partially *gives up the full optimisation* in order to avoid amplifying errors caused either by an imprecise model [31] and (or) imprecisely quantified aims [29].

Here, this "giving-up the full optimisation" is achieved by introducing a random optional pointer. In each optimisation epoch, it decides whether the selected decision rule should take the current partial loss into account or not. The LTF option allows to neglect the partial loss, see Sec. 2.2. It will be shown that the probability of the employed optimisation acts as the discounting rate. The found optimal solution exhibits the expected properties of DM under discounting.

Let us proceed in the outlined way. Let us introduce optional random pointers $p_t \in \boldsymbol{p} := \{0,1\}$, $t \in \boldsymbol{t}$, influencing the factors of the constructed ideal pd $\mathsf{c}^{\mathsf{i}}(b)$ (2) in the way leading to the discounted FPD

$$\mathsf{m}^{\mathsf{i}}(s_t|a_t,p_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|p_t,s_{t-1}) := \begin{cases} \mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}) & \text{if} \quad p_t=1 \\ \mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t,s_{t-1}) & \text{if} \quad p_t=0 \end{cases}. \tag{24}$$

The value $p_t = 1$ says that the t-th addend in (21) is to be optimised and $p_t = 0$ omits the addend from the optimisation.

Pointers $(p_t)_{t\in\boldsymbol{t}}$ extend the behaviour $b \in \boldsymbol{b}$ to $b := (s_t, a_t, p_t)_{t\in\boldsymbol{t}}$. The closed-loop model of the extended behaviour becomes

$$\mathsf{c}^{\mathsf{rw}}(b) := \prod_{t\in\boldsymbol{t}} \mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t,s_{t-1})\mathsf{w}^{p_t}(s_{t-1})(1-\mathsf{w}(s_{t-1}))^{1-p_t}, \text{ with}$$
$$\mathsf{w}(s_{t-1}) := \mathsf{r}(p_t=1|s_{t-1}) \in [0,1]. \tag{25}$$

The form (25) uses the assumption that s_{t-1} is the state, i.e. the pointer p_t brings no information regarding the state transition

$$\mathsf{m}(s_t|a_t,p_t,s_{t-1}) := \mathsf{m}(s_t|a_t,s_{t-1}). \tag{26}$$

The assumption (26) of no influence of p_t on s_t, if a_t, s_{t-1} are given, is a version of natural conditions of control [39].

The pointer $p_t \in \{1,0\}$ should ideally behave according to a chosen ideal pd $\mathsf{w}^{\mathsf{i}}(s_{t-1}) := \mathsf{r}^{\mathsf{i}}(p_t=1|s_{t-1}) \in [0,1]$, $\mathsf{r}^{\mathsf{i}}(p_t=0|s_{t-1}) = 1-\mathsf{w}^{\mathsf{i}}(s_{t-1})$. The ideal pd on the *extended* behaviour $b \in \boldsymbol{b}$ then gets the form implied by (24)

$$\mathsf{c}^{\mathsf{iw}^{\mathsf{i}}}(b) := \prod_{t\in\boldsymbol{t}} \mathsf{m}^{\mathsf{i}}(s_t|a_t,p_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|p_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(p_t|s_{t-1}) = \prod_{t\in\boldsymbol{t}} \Big[\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})$$
$$\times \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\mathsf{w}^{\mathsf{i}}(s_{t-1})\Big]^{p_t}\Big[\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t,s_{t-1})(1-\mathsf{w}^{\mathsf{i}}(s_{t-1}))\Big]^{1-p_t}. \tag{27}$$

Using (25), (27) and the definitions of $\mathsf{w}(s_{t-1}), \mathsf{w}^{\mathsf{i}}(s_{t-1})$, the optimised KL reads

$$\begin{aligned}\mathsf{D}(\mathsf{c}^{\mathsf{rw}}||\mathsf{c}^{\mathsf{iw}^{\mathsf{i}}}) &= \mathsf{E}^{\mathsf{r}}\Big[\sum_{t\in\boldsymbol{t}} \ln\Big(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t,s_{t-1})\mathsf{r}(p_t|s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,p_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|p_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(p_t|s_{t-1})}\Big)\Big] \\ &= \mathsf{E}^{\mathsf{r}}\Big[\sum_{t\in\boldsymbol{t}} \ln\Big(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t=1,s_{t-1})\mathsf{w}(s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\mathsf{w}^{\mathsf{i}}(s_{t-1})}\Big)\mathsf{w}(s_{t-1}) \\ &\quad + \ln\Big(\frac{1-\mathsf{w}(s_{t-1})}{1-\mathsf{w}^{\mathsf{i}}(s_{t-1})}\Big)(1-\mathsf{w}(s_{t-1}))\Big].\end{aligned}$$

By separating the weights from the first logarithm the KL of the weights-vector to its ideal counterpart appears. The overall KL gets the discounted form

$$\begin{aligned}\mathsf{D}(\mathsf{c}^{\mathsf{rw}}||\mathsf{c}^{\mathsf{iw}^{\mathsf{i}}}) &= \mathsf{E}^{\mathsf{r}}\Big[\sum_{t\in\boldsymbol{t}} \mathsf{w}(s_{t-1})\ln\Big(\frac{\mathsf{m}(s_t|a_t,s_{t-1})\mathsf{r}(a_t|p_t=1,s_{t-1})}{\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})}\Big) \\ &\quad + \mathsf{D}\Big([\mathsf{w}(s_{t-1}),1-\mathsf{w}(s_{t-1})]\Big|\Big|[\mathsf{w}^{\mathsf{i}}(s_{t-1}),1-\mathsf{w}^{\mathsf{i}}(s_{t-1})]\Big)\Big].\end{aligned} \tag{28}$$

The possible choice $\mathsf{r}^{\mathsf{i}}(p_t=1|s_{t-1}) := \mathsf{w}^{\mathsf{i}}(s_{t-1}) := \mathsf{w}(s_{t-1}) = \mathsf{r}(p_t=1|s_{t-1})$ mimics LTF introduced in (13). It eliminates KL divergence of the rule generating the pointer $\mathsf{r}(p_t|s_{t-1})$ to $\mathsf{r}^{\mathsf{i}}(p_t|s_{t-1})$. Then, (28) gives the discounted FPD with the weight $\mathsf{w}(s_{t-1}) := \mathsf{r}(p_t = 1|s_{t-1}) = \mathsf{w}^{\mathsf{i}}(s_{t-1})$.

Generally, the pointer-generating rule $\mathsf{r}(p_t|s_{t-1}) := \mathsf{w}(s_{t-1})^{p_t}(1-\mathsf{w}(s_{t-1}))^{1-p_t}$ can be optimised, as well. It suffices to apply Proposition 2 for finding the optimal rules $\mathsf{r}^{\mathsf{o}}(a_t,p_t|s_{t-1}) = \mathsf{r}^{\mathsf{o}}(a_t|p_t,s_{t-1})\mathsf{r}^{\mathsf{o}}(p_t|s_{t-1})$ for the extended actions (a_t,p_t). The optimal weight $\mathsf{w}^{\mathsf{o}}(s_{t-1}) = \mathsf{r}^{\mathsf{o}}(p_t=1|s_{t-1})$ is obtained by integrating $\mathsf{r}^{\mathsf{o}}(a_t,p_t|s_{t-1})$ over the actions $a_t \in \boldsymbol{a}$.

Proposition 4 *(Main Result: Discounted FPD). Let the closed-loop model* $\mathsf{c}^{\mathsf{rw}}(b)$ *(25) operate on the extended behaviour* b *made of the preserved states* $s_t \in \boldsymbol{s}$ *(26) and of the extended actions* $(a_t,p_t) \in (\boldsymbol{a},\{0,1\})$, $t \in \boldsymbol{t}$. *Let the ideal pd* $\mathsf{c}^{\mathsf{iw}^{\mathsf{i}}}(b)$ *have the form (27). Then (28) holds. It implies that:*

- *the term* $-\mathsf{E}^{\mathsf{r}}[\ln(\mathsf{n}^{\mathsf{rw}}(s_{t-1}))]$, *as seen in (23), does not appear;*
- *the extra KL in (28) expresses the proximity of* $\mathsf{w}(s_{t-1})$ *to* $\mathsf{w}^{\mathsf{i}}(s_{t-1})$;

▶ *the optimal rule* $\mathsf{r}^{\mathsf{o}}(a_t|p_t,s_{t-1})$ *and the weight (the pointer-generating rule)* $\mathsf{r}^{\mathsf{o}}(p_t=1|s_{t-1}) := \mathsf{w}^{\mathsf{o}}(s_{t-1})$ *evaluate as follows*

$$\mathsf{r}^{\mathsf{o}}(a_t|p_t=1,s_{t-1}) = \frac{\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})]}{\mathsf{n}^{\mathsf{i}}(s_{t-1})}$$

$$\mathsf{n}^{\mathsf{i}}(s_{t-1}) := \int_{\boldsymbol{a}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})]\,\mathrm{d}a_t$$

$$\mathsf{d}(a_t,s_{t-1}) := \int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left[\frac{\mathsf{m}(s_t|a_t,s_{t-1})}{\mathsf{n}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})}\right]\mathrm{d}s_t.$$

$\mathsf{r}^{\mathsf{o}}(a_t|p_t=0,s_{t-1})$ *selects* $a^{\mathsf{o}}(s_{t-1})\in \operatorname{Arg}\min_{a_t\in\boldsymbol{a}} \mathsf{K}(a_t,s_{t-1})$ *with*

$$\mathsf{K}(a_t,s_{t-1}) := \int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left[\frac{1}{\mathsf{n}(s_t)}\right]\mathrm{d}s_t$$

$$\mathsf{n}(s_{t-1}) = \mathsf{n}^{\mathsf{i}}(s_{t-1})\mathsf{w}^{\mathsf{i}}(s_{t-1}) + \mathsf{k}(s_{t-1})(1-\mathsf{w}^{\mathsf{i}}(s_{t-1})),$$

$$\mathsf{k}(s_{t-1}) := \exp[-\mathsf{K}(a^{\mathsf{o}}(s_{t-1}),s_{t-1})],$$

$$\mathsf{w}^{\mathsf{o}}(s_{t-1}) = \frac{\mathsf{n}^{\mathsf{i}}(s_{t-1})\mathsf{w}^{\mathsf{i}}(s_{t-1})}{\mathsf{n}^{\mathsf{i}}(s_{t-1})\mathsf{w}^{\mathsf{i}}(s_{t-1}) + \mathsf{k}(s_{t-1})(1-\mathsf{w}^{\mathsf{i}}(s_{t-1}))}.$$

This backward recursion, $t=h,h-1,\dots,1$*, is initiated by* $\mathsf{n}(s_h)=\mathsf{n}^{\mathsf{i}}(s_h)=1$.[6]

Proof. The optimal decision rules, dealing with extended actions a_t,p_t, have to be found. Proposition 1 provides the optimal rules

$$\begin{aligned}\mathsf{d}(a_t,p_t,s_{t-1}) &\overbrace{=}^{(11)} \int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,p_t,s_{t-1})\ln\left[\frac{\mathsf{m}(s_t|a_t,p_t,s_{t-1})}{\mathsf{n}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t,p_t,s_{t-1})}\right]\mathrm{d}s_t\\ &\overbrace{=}^{(24),(26)} p_t \underbrace{\int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left[\frac{\mathsf{m}(s_t|a_t,s_{t-1})}{\mathsf{n}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t,s_{t-1})}\right]\mathrm{d}s_t}_{:=\mathsf{d}(a_t,s_{t-1})}\\ &\quad + (1-p_t)\underbrace{\int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left[\frac{1}{\mathsf{n}(s_t)}\right]\mathrm{d}s_t}_{\mathsf{K}(a_t,s_{t-1}):=}.\end{aligned} \tag{29}$$

The weight $\mathsf{w}^{\mathsf{i}}(s_{t-1})$ influences the value-function $\mathsf{v}(s_{t-1})=-\ln(\mathsf{n}(s_{t-1}))$ (11) via

$$\begin{aligned}\mathsf{n}(s_{t-1}) &= \int_{(\boldsymbol{a},\boldsymbol{p})} \mathsf{r}^{\mathsf{i}}(a_t,p_t|s_{t-1})\exp[-\mathsf{d}(a_t,p_t,s_{t-1})]\,\mathrm{d}(a_t,p_t)\\ &\overbrace{=}^{(24),(29)} \mathsf{w}^{\mathsf{i}}(s_{t-1})\underbrace{\int_{\boldsymbol{a}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t,s_{t-1})]\,\mathrm{d}a_t}_{\mathsf{n}^{\mathsf{i}}(s_{t-1}):=}\\ &\quad + (1-\mathsf{w}^{\mathsf{i}}(s_{t-1}))\int_{\boldsymbol{a}} \mathsf{r}^{\mathsf{o}}(a_t|p_t=0,s_{t-1})\exp\left[-\mathsf{K}(a_t,s_{t-1})\right]\mathrm{d}a_t.\end{aligned}$$

LTF applied for $p_t=0$ implies that the optimal decision rule is deterministic and selects

$$a^{\mathsf{o}}(s_{t-1})\in \operatorname{Arg}\min_{a_t\in\boldsymbol{a}} \mathsf{K}(a_t,s_{t-1}) = \operatorname{Arg}\min_{a_t\in\boldsymbol{a}} \int_{\boldsymbol{s}} \mathsf{m}(s_t|a_t,s_{t-1})\ln\left[\frac{1}{\mathsf{n}(s_t)}\right]\mathrm{d}s_t.$$

The recursion for the function $\mathsf{n}(s_{t-1})$ then gets its final form

[6] K resembles Kerridge's inaccuracy [47] and k is its transformation.

$$\mathsf{n}(s_{t-1}) := \mathsf{w}^{\mathrm{i}}(s_{t-1})\mathsf{n}^{\mathrm{i}}(s_{t-1}) + (1-\mathsf{w}^{\mathrm{i}}(s_{t-1}))\mathsf{k}(s_{t-1}) \text{ with}$$

$$\mathsf{k}(s_{t-1}) := \exp[-\mathsf{K}(a^{\mathrm{o}}(s_{t-1}), s_{t-1})].$$

The optimal rule generating the extended actions (a_t, p_t) reads

$$\begin{aligned}\mathsf{r}^{\mathrm{o}}(a_t, p_t|s_{t-1}) &\overbrace{=}^{(12)} \frac{\mathsf{r}^{\mathrm{i}}(a_t, p_t|s_{t-1})\exp[-\mathsf{d}(a_t, p_t, s_{t-1})]}{\mathsf{n}(s_{t-1})} \\ &= \frac{p_t \mathsf{r}^{\mathrm{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t, s_{t-1})]\mathsf{w}^{\mathrm{i}}(s_{t-1})}{\mathsf{n}(s_{t-1})} \\ &\quad + \frac{(1-p_t)\mathsf{r}^{\mathrm{o}}(a_t|p_t=0, s_{t-1})\exp[-\mathsf{K}(a_t, s_{t-1})](1-\mathsf{w}^{\mathrm{i}}(s_{t-1}))}{\mathsf{n}(s_{t-1})}.\end{aligned} \tag{30}$$

The second equality uses the chain rule $\mathsf{r}^{\mathrm{i}}(a_t, p_t|s_{t-1}) = \mathsf{r}^{\mathrm{i}}(a_t|p_t, s_{t-1})\mathsf{r}^{\mathrm{i}}(p_t|s_{t-1})$ and the definition $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = \mathsf{r}^{\mathrm{i}}(p_t=1|s_{t-1})$.

Let us consider $p_t = 1$. The marginalisation of (30) yields

$$\begin{aligned}\mathsf{r}^{\mathrm{o}}(p_t=1|s_{t-1}) &:= \mathsf{w}^{\mathrm{o}}(s_{t-1}) = \frac{\mathsf{n}^{\mathrm{i}}(s_{t-1})}{\mathsf{n}(s_{t-1})}\mathsf{w}^{\mathrm{i}}(s_{t-1}) \\ &= \frac{\mathsf{n}^{\mathrm{i}}(s_{t-1})\mathsf{w}^{\mathrm{i}}(s_{t-1})}{\mathsf{n}^{\mathrm{i}}(s_{t-1})\mathsf{w}^{\mathrm{i}}(s_{t-1}) + \mathsf{k}(s_{t-1})(1-\mathsf{w}^{\mathrm{i}}(s_{t-1}))}, \quad \text{while}\end{aligned}$$

$$\mathsf{r}^{\mathrm{o}}(a_t|p_t=1, s_{t-1}) \propto \mathsf{r}^{\mathrm{o}}(a_t, p_t=1|s_{t-1}) \quad \text{implies}$$

$$\mathsf{r}^{\mathrm{o}}(a_t|p_t=1, s_{t-1}) = \frac{\mathsf{r}^{\mathrm{i}}(a_t|s_{t-1})\exp[-\mathsf{d}(a_t, s_{t-1})]}{\mathsf{n}^{\mathrm{i}}(s_{t-1})}.$$

It remains to find explicitly the deterministic optimal decision rule for $p_t = 0$. A direct check confirms that the equality

$$\mathsf{r}^{\mathrm{o}}(a_t|p_t=0, s_{t-1}) = \frac{\mathsf{r}^{\mathrm{o}}(a_t, p_t=0|s_{t-1})}{\mathsf{r}(p_t=0|s_{t-1})}$$

is met for any $a^{\mathrm{o}}(s_{t-1})$ minimising $\mathsf{K}(a_t, s_{t-1})$. □

Remark 4 *(On Proposition 4).* ▶ For $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = 1$, the last addend in (28) is finite only for

$$\mathsf{w}^{\mathrm{o}}(s_{t-1}) = \mathsf{w}^{\mathrm{i}}(s_{t-1}) = 1 \text{ as otherwise } \mathsf{D}\Big([\mathsf{w}(s_{t-1}), 1-\mathsf{w}(s_{t-1})]\Big|\Big|[1,0]\Big)$$

$= \mathsf{w}(s_{t-1})\ln(\mathsf{w}(s_{t-1})) + (1-\mathsf{w}(s_{t-1}))\ln\big((1-\mathsf{w}(s_{t-1}))/0\big) = \infty$. Thus, the non-discounted FPD, see Proposition 1, recovers for $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = 1$.

▶ The optimal rule $\mathsf{r}^{\mathrm{o}}(a_t|p_t=1, s_{t-1})$ coincides with the rule in the non-discounted FPD, see Proposition 1. Discounting influences the value function via $\mathsf{n}(s_{t-1})$ as expected. □

4. Illustrative experiment

The example *illustrates* that the designed, randomly switching off, the optimisation leads to the discounted FPD that *counteracts adverse consequences of imprecise modelling* and acts as expected and desirable.

The simulation had three state values $s \in \boldsymbol{s} := \{1,2,3\}$ and started from $s_0 = 1$. The environment was stimulated by actions with two values $a \in \boldsymbol{a} := \{1,2\}$ generated by stationary discounted FPD-optimal decision rules applied up to the horizon $h = 200$. Table 1 describes the used ideal environment model and the ideal decision rule, both time-invariant. The FPD ran with the environment model $\mathsf{m}(s_t|a_t, s_{t-1}) \propto \sqrt{\mathsf{m}^{\mathrm{i}}(s_t|a_t, s_{t-1})}$. The modelling error was introduced by simulating the environment with the transition pd derived from $\mathsf{m}(s_t|a_t, s_{t-1})$ by swapping the pd values with entries $(s_t|a_t, s_{t-1}) = (3|2,2)$ and $(s_t|a_t, s_{t-1}) = (2|2,3)$.

Fig. 1 shows the expected influence of the discounting by comparing the cases with $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = 0.4$ and $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = 1.0$ (no discounting). The strongly preferred state $s_t = 3$ is reached more often with the discounting rate $\mathsf{w}^{\mathrm{i}}(s_{t-1}) = 0.4$.

All simulation options are presented to allow a cross-check of our results. Their choice has no deeper meaning. It just fits the illustrative purpose.

The practical choice of the discounting rate is non-trivial and the discounted FPD is expected to help in this respect. It is necessary to conduct a further applied research. Thesis [35] reflects its current state. At present, it is too soon for a meaningful comparative study. It will be done and published elsewhere after combining the derived decision rules with on-line learning [48] and the related quantification of the ideal closed-loop model [21].

Table 1
The Ideal Environment Model and the Ideal Decision Rule.

$\mathsf{m}^{\mathrm{i}}(s_t\|a_t, s_{t-1})$	$a_t=1$	$a_t=2$
$s_t=1,\ s_{t-1}=1$	0.316	0.043
$s_t=1,\ s_{t-1}=2$	0.307	0.080
$s_t=1,\ s_{t-1}=3$	0.377	0.016
$s_t=2,\ s_{t-1}=1$	0.263	0.053
$s_t=2,\ s_{t-1}=2$	0.373	0.069
$s_t=2,\ s_{t-1}=3$	0.330	0.027
$s_t=3,\ s_{t-1}=1$	0.421	0.904
$s_t=3,\ s_{t-1}=2$	0.320	0.851
$s_t=3,\ s_{t-1}=3$	0.292	0.957

$\mathsf{r}^{\mathrm{i}}(a_t\|s_{t-1})$	$a_t=2$
$s_{t-1}=1$	0.933
$s_{t-1}=2$	0.930
$s_{t-1}=3$	0.957

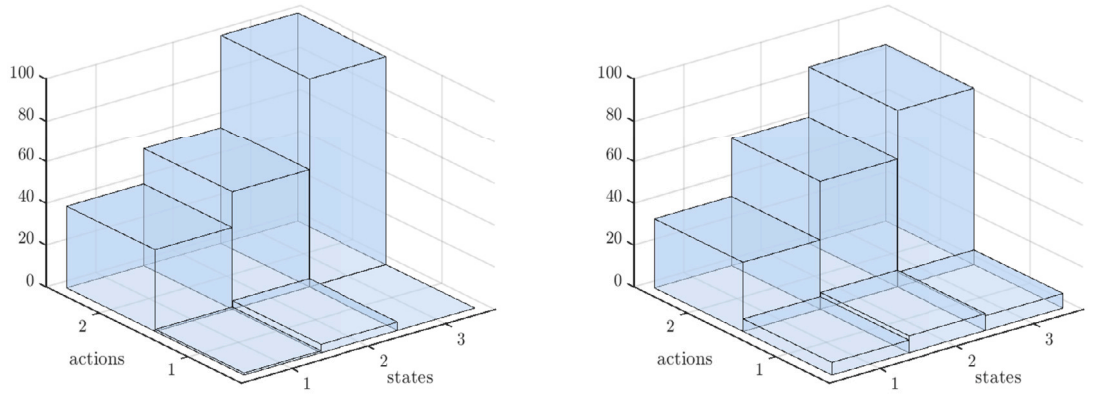

Fig. 1. The demonstration of the discounting influence under mismodelling. The left panel reflects the discounting with $\mathsf{w}^{\mathrm{i}}(s_{t-1})=0.4$, the right one reflects no discounting, $\mathsf{w}^{\mathrm{i}}(s_{t-1})=1.0$. Recall: the most desired state is $s=3$ and the most desired action is $a=2$.

5. Concluding remarks

The paper pushes further on the support of DM of agents exploiting FPD and thus MDP. It offers them its novel discounted version, Proposition 4. Its usefulness relies on an operational specification of the ideal weights $\{\mathsf{w}^{\mathrm{i}}(s_{t-1})\}_{t\in\mathbf{t}}$. The unified, pd-based, formulation of FPD, opens a direct way of their choice.

The weights act as forgetting rates in estimation [32]. They can directly be estimated using Bayesian estimation paradigm, at least on discrete grid [33]. The probabilistic interpretation, so useful in preference elicitation [49,50], then allows us to use such estimates as data-dependent choices of the discounting weight.

The running research [35] tries to:

- relate tightly the discounting to forgetting;
- admit different weights to the environment models and decision rules;
- find the conditions under which the designed decision rules converge and stabilise closed loop [13] if the horizon is unbounded.

Authors will be happy if the paper will encourage readers to convert the presented advanced theory into a practically useful tool.

CRediT authorship contribution statement

Miroslav Kárný: Writing – original draft, Investigation. **Soňa Molnárová:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] L. Savage, Foundations of Statistics, Wiley, 1954.

[2] A. Wald, Statistical Decision Functions, J. Wiley, 1950.
[3] E. Feinberg, A. Shwartz, Handbook of Markov Decision Processes: Methods & Applications, Kluwer, 2002.
[4] D. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, 2017.
[5] J. van Schuppen, Stochastic Control Theory, Springer International Publishing, Cham, 2021, pp. 617–624.
[6] Y. Yang, G. Hu, C. Spanos, Stochastic optimal control of HVAC system for energy-efficient buildings, IEEE Trans. Control Syst. Technol. 30 (1) (2022) 376–383.
[7] C. Gollier, Discounting an uncertain future, J. Public Econ. 85 (2) (2002) 149–166.
[8] P. Dasgupta, Discounting climate change, J. Risk Uncertain. 37 (2008) 141–169.
[9] A. Attema, W. Brouwer, K. Claxton, Discounting in economic evaluations, PharmacoEconomics 36 (2018) 745–758.
[10] Y. Gu, Y. Cheng, C. Chen, X. Wang, Proximal policy optimization with policy feedback, IEEE Trans. Syst. Man Cybern. Syst. 52 (7) (2022) 4600–4610.
[11] L. Pan, Y. Deng, K. Cheong, Dynamical Markov decision-making model based on mass function to quantitatively predict interference effects, Inf. Sci. 648 (2023) 119482.
[12] P. Yan, D. Wang, H. Li, D. Liu, Error bound analysis of Q-function for discounted optimal control problems with policy iteration, IEEE Trans. Syst. Man Cybern. Syst. 47 (7) (2017) 1207–1216.
[13] D. Wang, J. Ren, M. Ha, Discounted linear Q-learning control with novel tracking cost and its stability, Inf. Sci. 626 (2023) 339–353.
[14] M. Kárný, Axiomatisation of fully probabilistic design revisited, Syst. Control Lett. 141 (2020) 104719.
[15] M. Kárný, T. Kroupa, Axiomatisation of FPD, Inf. Sci. 186 (1) (2012) 105–113.
[16] I. Landau, A survey of MRAS techniques, Automatica 10 (4) (1974) 353–379.
[17] M. Kárný, Towards fully probabilistic control design, Automatica 32 (12) (1996) 1719–1722.
[18] M. Kárný, T. Guy, Fully probabilistic control design, SCL 55 (2006) 259–265.
[19] A. Quinn, M. Kárný, T. Guy, Fully probabilistic design of hierarchical Bayesian models, Inf. Sci. 369 (2016) 532–547.
[20] E. Garrabé, G. Russo, Probabilistic design of optimal sequential decision-making algorithms in learning and control, Annu. Rev. Control 54 (2022) 81–102.
[21] M. Kárný, T. Sováková, Model-based preference quantification, Automatica 156 (2023) 111185.
[22] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–87.
[23] H. Kappen, Linear theory for control of nonlinear stochastic systems, Phys. Rev. Lett. 95 (20) (2005) 200201.
[24] P. Guan, M. Raginsky, R. Willett, Online Markov decision processes with Kullback Leibler control cost, IEEE Trans. Automat. Control 59 (6) (2014) 1423–1438.
[25] Y. Abbasi-Yadkori, P. Bartlett, X. Chen, A. Malek, Large-Scale Markov Decision Problems with KL Control Cost and Its Application to Crowdsourcing, Proc. of the 32nd Intern. Conf. on Machine Learning, vol. 37, JMLR, Lille, France, 2015, pp. 19–28.
[26] D. Palenicek, A survey on constraining policy updates using the KLD, in: B. Belousov, et al. (Eds.), Reinforcement Learning Algorithms: Analysis and Applications, Springer, 2021.
[27] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, M. Geist, Leverage the average: an analysis of KL regularization in reinforcement learning, Adv. Neural Inf. Process. Syst. 33 (2020) 12163–12174.
[28] K. Rana, V. Dasagi, J. Haviland, B. Talbot, M. Milford, N. Sünderhauf, Bayesian controller fusion: leveraging control priors in deep reinforcement learning for robotics, Int. J. Robot. Res. 42 (3) (2023) 123–146.
[29] J. Doyle, Survey of time preference, delay discounting models, Judg. Dec. Making 8 (2013) 116–135.
[30] V. Gaitsgory, L. Grüne, M. Höger, C. Kellett, S. Weller, Stabilization of strictly dissipative discrete time systems with discounted optimal control, Automatica 93 (2018) 311–320.
[31] C. Rohrs, L. Valavani, M. Athans, G. Stein, Robustness of adaptive control algorithms in the presence of unmodeled dynamics, in: IEEE Conf. on Decision and Control, vol. 1, Orlando, FL, 1982, pp. 3–11.
[32] R. Kulhavý, M.B. Zarrop, On a general concept of forgetting, Int. J. Control 58 (4) (1993) 905–924.
[33] K. Dedecius, I. Nagy, M. Kárný, Parameter tracking with partial forgetting method, Int. J. Adapt. Control Signal Process. 26 (1) (2012) 1–12.
[34] M. Aguayo, L. Bellido, C. Lentisco, E. Pastor, DASH adaptation algorithm based on adaptive forgetting factor estimation, IEEE Trans. Multimed. 20 (5) (2018) 1224–1232.
[35] S. Molnárová, Applicable Adaptive Discounted Fully Probabilistic Design of Decision Strategy, FNSPE, CTU, Prague, 2024, http://hdl.handle.net/10467/114570.
[36] M. Rao, Measure Theory and Integration, J. Wiley, 1987.
[37] N. Cammardella, A. Bušić, S. Meyn, Kullback-Leibler-quadratic optimal control in a stochastic environment, in: 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 158–165.
[38] D. Gagliardi, G. Russo, On a probabilistic approach to synthesize control policies from example datasets, Automatica 137 (2022) 110121.
[39] V. Peterka, Bayesian system identification, in: P. Eykhoff (Ed.), Trends & Progress in System Identification, 1981, pp. 239–304.
[40] A. Feldbaum, Theory of dual control, Autom. Remote Control 22 (1961) 3–19.
[41] H. Hu, S. Song, G. Huang, Self-attention-based temporary curiosity in reinforcement learning exploration, IEEE Trans. Syst. Man Cybern. Syst. 51 (9) (2021) 5773–5784.
[42] I. Sason, On f-divergences: integral representations, local behavior, and inequalities, Entropy 20 (5) (2018) 383.
[43] M. Kárný, Renyi's extension of fully probabilistic design of decision-making rules, IEEE Trans. Syst. Man Cybern. Syst. (2024), submitted for publication.
[44] M. Kárný, J. Böhm, T. Guy, L. Jirsa, I. Nagy, P. Nedoma, L. Tesař, Optimized Bayesian Dynamic Advising: Theory and Algorithms, Springer, London, UK, 2006.
[45] S. Amari, α - divergence is unique, belonging to both f-divergence and Bregman divergence classes, IEEE Trans. Inf. Theory 55 (11) (2009) 4925–4931.
[46] T. van Erven, P. Harremos, Rényi divergence and Kullback-Leibler divergence, IEEE Trans. Inf. Theory 60 (7) (2014) 3797–3820.
[47] D. Kerridge, Inaccuracy and inference, J. R. Stat. Soc. B 23 (1961) 284–294.
[48] M. Kárný, FPD of strategies with estimator, Automatica 141 (2022) 110269.
[49] U. Chajewska, D. Koller, Utilities as random variables: density estimation and structure discovery, in: Proc. UAI–00, 2000, pp. 63–71.
[50] M. Kárný, T. Guy, Preference elicitation within framework of fully probabilistic design of decision strategies, in: IFAC Workshop ALCOS, vol. 52, 2019, pp. 239–244.