Model-Based Preference Quantification ⋆

Miroslav Kárný, Tereza Sivákova

The Czech Academy of Sciences, Institute of Information Theory and Automation, POB 18, 182 08 Prague 8, Czech Republic

Abstract

Any prescriptive theory of decision-making (DM) has to cope with the common DM agents' inability to fully specify their preferences dependent on several attributes. The paper provides the needed preference completion and quantification for *f*ully *p*robabilistic *d*esign (FPD) of DM strategies. FPD (covering the usual Bayesian DM) probabilistically models the agent's environment and quantifies its preferences via an *ideal* probabilistic model of the closed DM loop. The *p*robability *d*ensity (pd) models (closed-loop) behaviour, a collection of involved random variables. Its ideal twin is high on desired behaviours, small on undesired and zero on forbidden ones. The FPD-optimal strategy minimises the *K*ullback-*L*eibler *d*ivergence (KLD) of the closed-loop modelling pd to the ideal twin. The exposed preference quantification chooses the optimal ideal pd from the set of pds compatible with partially-specified agent's preferences. The optimal ideal pd minimises the KLD minima reached by the optimal strategies for respective imminent ideal pds. This preference-focused twin of the minimum KLD principle was applied to special sets of ideal pds. The paper extends them towards exploration and balancing contradictory wishes on states and actions.

Key words: dynamic performance; probabilistic; preferences; optimal strategy; preference elicitation; exploration;

1 Introduction

Decision making (DM) is an agent's targeted choice and use of actions that try to meet the agent's preferences [1] . The choice of an optimal, action-opting, strategy needs a quantification of the faced problem. It deals with the agent's beliefs about responses to its actions [40] and its wishes [46], both concerning the closed loop. The adopted Bayesian paradigm elicits prior beliefs [11,32], modifies them by data via Bayes' rule [3,36] or its generalisation [38]. The minimum KLD principle [41] completes them. The situation is less mature concerning the agent's wishes. It is a harder problem as human agents or strategy designers are: ▶ unable quantify fully their wishes in multi-attribute DM tasks [13]; ▶ prone to conflicts [16]; ▶ unwilling to spend too much deliberation effort on this hard, but unavoidable, DM subtask [19].

The paper contributes to the remedy of this state. It deals with the preference quantification as [24] (aka preference elicitation, PE [10]). It processes the state-transition model and semi-verbal expression of agent's wishes pointing to imminent ideal pds. It serves the usual PE as the state-transition model and thus preferences are learnt while the wishes-inspecting queries and the agent's answers enter the behaviour [8,27].

FPD [17,18,20,45] models wishes probabilistically. It chooses the strategy making the behaviour-modelling pd the closest one to the ideal pd modelling the closed loop that behaves in harmony with the agent's wishes. Kullback-Leibler divergence [29] measures the pds' (di)similarity. In FPD, PE consists of: ▶ a translation of agent's wishes into a non-empty set of imminent ideal pds; ▶ a choice of the optimal ideal pd within this set that adds as few extra wishes or constraints as possible.

The developed solution: ▶ combines multiple attributes in a clear-cut way; ▶ provides an ambitious, but potentially reachable, goal of the strategy design; ▶ suppresses conflicts; ▶ omits no standard DM [20]; ▶ unifies the expression and handling of beliefs and preferences; ▶ simplifies PE based on queries [9,12,27,44].

Sec. 2 recalls FPD and the used PE principle. Core Sec. 3 extends former uses of this principle. Sec. 4 applies the result to DM with a finite amount of possible behaviours. Sec. 5 illustrates the theory. Sec. 6 adds comments.

Throughout, $\{x\}$ marks the set of xs. It is a subset of a finite-dimensional real space or a set of pds. It is specified only if needed. $\equiv$ defines by assigning, e.g. $|\{x\}| \equiv \int_{\{x\}} \mathrm{d}x$. The index $^{\mathrm{i}}$ marks ideals and $^{\mathrm{o}}$ optimality. $\propto$ is proportionality. Sanserif fonts mark mappings. p-norm, $p \geq 1$, of a real-valued function $\mathsf{f}(x)$ is $||\mathsf{f}||_p \equiv \left[\int_{\{x\}} |\mathsf{f}(x)|^p \mathrm{d}x\right]^{1/p}$, [39]. $\chi_{\{x\}}(x)$ is the indicator of $\{x\}$ at x. Minimisers of $\mathsf{f}(x)$ are in $\mathrm{Arg}\min_{x\in\{x\}} \mathsf{f}(x) \subset \{x\}$.

⋆ This paper was not presented at any IFAC meeting. MK corresponds at school@utia.cas.cz, TS at sivakova@utia.cas.cz.

[1] The agent of any nature is referred to as "it". Agent's wishes are taken as synonyms of formally treated preferences [46].

2 Preliminaries

The next recall of FPD and the employed PE principle make the paper self-reliant.

2.1 Decision Making via Fully Probabilistic Design

DM couples an agent with its environment. The agent applies actions $a_t \in \{a\} \neq \emptyset$ at discrete time $t \in \{t\} \equiv \{1,\dots,T\}$, $T \leq \infty$. Actions $a_t \in \{a\}$ stimulate transitions of the (closed loop) states $s_{t-1} \in \{s\} \neq \emptyset$ to states $s_t \in \{s\}$. The actions and states up to the horizon T form the behaviours $b \in \{b\}$. The agent selects actions via a randomised DM strategy [2] $\mathsf{r} \in \{\mathsf{r}\} \equiv \{\mathsf{r}(a_t|s_{t-1}),\, t \in \{t\},\, a_t \in \{a\}, s_{t-1} \in \{s\}\}$. The pds $\mathsf{r}(a_t|s_{t-1})$, called r-factors, model the DM rules forming the strategy. The r-dependent joint pd $\mathsf{c}^{\mathsf{r}}(b)$ fully models, generically random, behaviours $b \in \{b\}$. The chain rule for pds [36] and the meaning of state provide the factorisation, $\forall b \in \{b\} \equiv \{b = (s_t, a_t)_{t\in\{t\}}\}$,

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

The model $\mathsf{m} \in \{\mathsf{m}\} \equiv \{\mathsf{m}(s_t|a_t, s_{t-1}), t \in \{t\}, s_t, s_{t-1} \in \{s\}, a_t \in \{a\}\}$, form conditional pds $\mathsf{m}(s_t|a_t,\, s_{t-1})$ called m-factors, describes the state transitions. The m-factors are known. Often, grey-box modelling [7] and Bayesian learning [36] provide them. Then, the states include the used statistic values [15]. FPD quantifies the agent's wishes by an ideal closed-loop model. It is a joint pd $\mathsf{c}^{\mathsf{i}}(b)$, $b \in \{b\}$, with high probability values on preferred behaviours, small values on undesired ones and zero on forbidden behaviours. It also factorises

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

The m^{i}- and r^{i}-factors model the *desired* (ideal) state transitions and the *desired* ways of the action choice. Their product at given time moment is called c^{i}-factor. The FPD-optimal DM strategy $\mathsf{r}^{\mathsf{o}} \in \{\mathsf{r}\}$ minimises [3] KLD $\mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}})$ of c^{r} to c^{i}, see (1), (2),

$$\mathsf{r}^{\mathsf{o}} \in \mathrm{Arg}\min_{\mathsf{r}\in\{\mathsf{r}\}} \mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}}) \equiv \mathrm{Arg}\min_{\mathsf{r}\in\{\mathsf{r}\}} \int_{\{b\}} \mathsf{c}^{\mathsf{r}}(b) \ln\Big(\frac{\mathsf{c}^{\mathsf{r}}(b)}{\mathsf{c}^{\mathsf{i}}(b)}\Big)\mathrm{d}b. \tag{3}$$

The next proposition provides the FPD-optimal strategy. Its proof mimics dynamic programming [5] and provides the optimal r^{o}-factors. Its general version is in [22].

Proposition 1 (FPD) *The backward functional recursion on* $\mathsf{h}(s_t) \in [0,1]$, $s_t \in \{s\}$; $\mathsf{h}(s_T) \equiv 1$, $a_t \in \{a\}$,

$$\mathsf{h}(s_{t-1}) \equiv \int_{\{a\}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}) \exp[-\mathsf{d}(a_t|s_{t-1})]\mathrm{d}a_t,\ t \in \{t\} \tag{4}$$

$$\mathsf{d}(a_t|s_{t-1}) \equiv \int_{\{s\}} \mathsf{m}(s_t|a_t, s_{t-1}) \ln\Big[\frac{\mathsf{m}(s_t|a_t, s_{t-1})}{\mathsf{h}(s_t)\mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1})}\Big]\mathrm{d}s_t$$

gives the optimal r^{o}*-factors and the reached minima* $-\ln(\mathsf{h}(s_{t-1}))$*, i.e. the value functions, [5]. It holds*

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

$$\min_{\mathsf{r}\in\{\mathsf{r}\}} \mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}}) = -\ln(\mathsf{h}(s_0)),\ \textit{cf. footnote}\ [2]. \tag{5}$$

2.2 Optimal Preference-Elicitation Principle

The ideal pd c^{i} (2) quantifies the agent's wishes. Once it is chosen, the minimisation (3) gives the optimal DM strategy r^{o}. Thus, PE within FPD reduces to the choice of the pd c^{io} that expresses the agent's wishes in the best way. The specific wishes delimit the imminent ideal joint pds. Their usually incomplete description implies, that the set $\{\mathsf{c}^{\mathsf{i}}\}$ of ideal pds c^{i}, acting on $\{b\}$,

$$\{\mathsf{c}^{\mathsf{i}}\} \equiv \{\text{ideal pds } \mathsf{c}^{\mathsf{i}}(b) \text{ meeting agent's wishes}\} \tag{6}$$

contains many (often, infinitely many) pds. The set (6) may be empty due to the agent's inconsistencies. PE deals with an amenable choice of: ▶ the *non-empty* set $\{\mathsf{c}^{\mathsf{i}}\}$ (6) that copes with inconsistencies of agent's wishes, and ▶ the *optimal* ideal pd c^{io} from this set. The PE principle [24] chooses as the optimal ideal pd

$$\mathsf{c}^{\mathsf{io}} \in \mathrm{Arg}\min_{\mathsf{c}^{\mathsf{i}}\in\{\mathsf{c}^{\mathsf{i}}\}} \Big[\min_{\mathsf{r}\in\{\mathsf{r}\}} \mathsf{D}(\mathsf{c}^{\mathsf{r}}||\mathsf{c}^{\mathsf{i}})\Big]. \tag{7}$$

Its use guarantees that no extra wish and no extra constraint are added to those expressed by the agent.

Predecessors of this work [24,26] gave up the optimisation (7) and searched for an approximate solution. They exploited that the amenable greedy strategy provides an upper bound on the value function. They minimise this bound over $\{\mathsf{c}^{\mathsf{i}}\}$. Here, this way is unused. It is possible whenever the value function can be well evaluated [4].

The minimisations over c^{i}-factors at any time $t \in \{t\}$ and for any realised state s_{t-1} are formally identical. Thus, we can hide t, s_{t-1} and deal with $\mathsf{m}(s|a) \equiv \mathsf{m}(s_t = s|a_t = a, s_{t-1})$, $\mathsf{m}^{\mathsf{i}}(s|a) \equiv \mathsf{m}^{\mathsf{i}}(s_t = s|a_t = a, s_{t-1})$, $\mathsf{r}(a) \equiv \mathsf{r}(a_t = a|s_{t-1})$, $\mathsf{r}^{\mathsf{i}}(a) \equiv \mathsf{r}^{\mathsf{i}}(a_t = a|s_{t-1})$ and $\mathsf{h}(s) \equiv \mathsf{h}(s_t = s)$, $s_{t-1}, s_t, s \in \{s\}$, $a_t, a \in \{a\}$.

The optimal $\mathsf{c}^{\mathsf{io}} \equiv \mathsf{m}^{\mathsf{io}}\mathsf{r}^{\mathsf{io}}$-factor, is, cf. (4), (5), (7),

$$\mathsf{c}^{\mathsf{io}} \in \mathrm{Arg}\max_{\mathsf{r}^{\mathsf{i}}\in\{\mathsf{r}^{\mathsf{i}}\}} \Big[\max_{\mathsf{m}^{\mathsf{i}}\in\{\mathsf{m}^{\mathsf{i}}\}} \int_{\{a\}} \mathsf{r}^{\mathsf{i}}(a) \exp[-\mathsf{d}(a)]\mathrm{d}a\Big]$$

$$\mathsf{d}(a) = \int_{\{s\}} \mathsf{m}(s|a) \ln\Big(\frac{\mathsf{m}(s|a)}{\mathsf{h}(s)\mathsf{m}^{\mathsf{i}}(s|a)}\Big)\mathrm{d}s \tag{8}$$

$$\in \Big[-\int_{\{s\}} \mathsf{m}(s|a) \ln[\mathsf{h}(s)]\mathrm{d}s, \infty\Big] \subset [0,\infty].$$

$\mathsf{D}(\mathsf{m}||\mathsf{m}^{\mathsf{i}}) \geq 0 = \mathsf{D}(\mathsf{m}||\mathsf{m})$ [29] implies $\mathsf{d}(a)$-range.

The evaluation (8) uses the given $\mathsf{h} : \{s\} \to [0,1]$ gained in the previous design step in (4). It runs over a cross-section $\{\mathsf{m}^{\mathsf{i}}\text{-factors}\}$ of $\{\mathsf{c}^{\mathsf{i}}\text{-factors}\}$ given by an r^{i}-factor. Thus, the evaluation (8) runs over

$$\{\mathsf{c}^{\mathsf{i}}\} \equiv \{\mathsf{m}^{\mathsf{i}}\mathsf{r}^{\mathsf{i}} = \mathsf{c}^{\mathsf{i}}\text{-factor meeting agent's wishes}\}. \tag{9}$$

[2] A fixed initial state s_0 is an implicit part of all conditions.

[3] It resembles the concept of model reference control [31]. The FPD-optimality has an axiomatic basis that covers all usual formulations of DM under uncertainty [20]. FPD has a range of non-trivial uses, e.g. [21,25]. The related KLD-based reinforcement learning [34] provides an extra insight.

[4] This fact was recognised by our colleague Marko Ruman.

3 Preference Quantification

First, a generic choice of c^{i}-factors is made. Then, a specific, but quite universal, non-empty set (9) is chosen. The task (8) gives the optimal ideal $\mathsf{c}^{\mathsf{i}\circ}$-factors. It runs over the m^{i}-factors for a fixed r^{i}-factor and then over the r^{i}-factors. The reversed order in [24,26] is more complex.

3.1 Generic Optimal Ideal $\mathsf{m}^{\mathsf{i}\circ}$-Factor

This part operationally specifies the optimal ideal $\mathsf{m}^{\mathsf{i}\circ}$.

Proposition 2 (Optimal $\mathsf{m}^{\mathsf{i}\circ}$-Factor) *Let an $\mathsf{r}^{\mathsf{i}} \in \{\mathsf{r}^{\mathsf{i}}\}$ define a non-empty cross-section $\{\mathsf{m}^{\mathsf{i}}\}$ of (9). Let $\mathsf{m}^{\mathsf{i}}(s|a) \in \{\mathsf{m}^{\mathsf{i}}\}$ exist giving*[5] *$\mathsf{d}(a) < \infty$, $\forall a \in \{a\}$. Then, the optimal ideal $\mathsf{m}^{\mathsf{i}\circ}$-factor minimises $\mathsf{d}(a)$ (8)*

$$\mathsf{m}^{\mathsf{i}\circ}(s|a) \in \operatorname{Arg}\max_{\mathsf{m}^{\mathsf{i}}\in\{\mathsf{m}^{\mathsf{i}}\}} \int_{\{a\}} \mathsf{r}^{\mathsf{i}}(a)\exp[-\mathsf{d}(a)]\mathrm{d}a \qquad (10)$$
$$= \operatorname{Arg}\min_{\mathsf{m}^{\mathsf{i}}\in\{\mathsf{m}^{\mathsf{i}}\}} \mathsf{d}(a),\ s \in \{s\},\ a \in \{a\}.$$

Proof For $\{\mathsf{m}^{\mathsf{i}}\} \neq \emptyset$ and any $a \in \{a\}$, a minimiser $\mathsf{m}^{\mathsf{i}\star} \in \{\mathsf{m}^{\mathsf{i}}\}$ of $\mathsf{d}(a) \geq 0$ gives the value $\mathsf{d}^{\star}(a) \leq \mathsf{d}(a)$. There, $\mathsf{d}(a)$ is given by (8) for an arbitrary $\mathsf{m}^{\mathsf{i}} \in \{\mathsf{m}^{\mathsf{i}}\}$ and the same h. This implies that $\mathsf{d}^{\star}(a) < \infty$ and $\exp[-\mathsf{d}^{\star}(a)] \geq \exp[-\mathsf{d}(a)]$. Multiplication of this inequality by $\mathsf{r}^{\mathsf{i}}(a) \geq 0$ (not being identically zero) and the integration over the set $\{a\}$ imply that $\mathsf{m}^{\mathsf{i}\circ} = \mathsf{m}^{\mathsf{i}\star}$. □

Remarks 1

▶ *The minimiser of $\mathsf{d}(a)$ (10) is uninfluenced by h.*

▶ *$\mathsf{m}^{\mathsf{i}\circ}(s|a)$ minimises the KLD given by $a = a_t$ and s_{t-1} of the state-transition pd to $\mathsf{m}(s|a)$. Thus,* the optimal ideal state-transition model is the best approximant of the state-transition model meeting agent's wishes*, [4,23].*

▶ *The found $\mathsf{m}^{\mathsf{i}\circ}$ makes m absolutely continuous with respect to $\mathsf{m}^{\mathsf{i}\circ}$ (otherwise $\mathsf{d}^{\circ} = \infty$). Thus,* the state $s \in \{s\}$ that may occur due to $\mathsf{m}(s|a) > 0$ for the allowed $a \in \{a\}$ is not forbidden by $\mathsf{m}^{\mathsf{i}\circ}(s|a)$.

▶ *These facts confirm that the ideal pd $\mathsf{m}^{\mathsf{i}\circ}$ (10) is the realistic option. Its optimality (7) makes it ambitious.*

3.2 Generic Optimal Ideal $\mathsf{r}^{\mathsf{i}\circ}$-Factor

This part inspects universally desirable r^{i}-factors.

The support $\mathrm{supp}[\mathsf{r}] \equiv \{a : \mathsf{r}(a) > 0\}$ of the opted r-factor is to be included in the set $\{a\}$ accessible to the agent. The form of the FPD-optimal r°-factor (5) implies $\mathrm{supp}[\mathsf{r}^{\circ}] \subseteqq \mathrm{supp}[\mathsf{r}^{\mathsf{i}}]$. Hence, only the ideal r^{i}-factors

$$\mathsf{r}^{\mathsf{i}} \in \{\mathsf{r}^{\mathsf{i}}\} \equiv \{\mathsf{r}^{\mathsf{i}} : \mathrm{supp}[\mathsf{r}^{\mathsf{i}}] = \{a\}\} \qquad (11)$$

keep actions in $\{a\}$ and exclude none. This makes (11) the generic constraint on the set $\{\mathsf{r}^{\mathsf{i}}\}$.

Proposition 3 (Optimal $\mathsf{r}^{\mathsf{i}\circ}$-Factor Meeting (11)) *Let $\{\mathsf{r}^{\mathsf{i}}\}$ be given by an opted $p > 1$ as follows*

$$\{\mathsf{r}^{\mathsf{i}}\} \equiv \{\mathsf{r}^{\mathsf{i}} : \mathrm{supp}[\mathsf{r}^{\mathsf{i}}] = \{a\} \text{ and } ||\mathsf{r}^{\mathsf{i}}||_p < \infty\} \qquad (12)$$
$$\text{while assuming} \quad |\{a\}| < \infty. \qquad (13)$$

Let each $\mathsf{r}^{\mathsf{i}} \in \{\mathsf{r}^{\mathsf{i}}\}$ (12) define a non-empty cross-section $\{\mathsf{m}^{\mathsf{i}}\}$ of (9). Let $\mathsf{m}^{\mathsf{i}}(s|a) \in \{\mathsf{m}^{\mathsf{i}}\}$ exist such that $\mathsf{d}(a) < \infty$, $\forall a \in \{a\}$ (8). Then, the optimal ideal $\mathsf{r}^{\mathsf{i}\circ}$-factor is

$$\mathsf{r}^{\mathsf{i}\circ}(a) \propto \chi_{\{a\}}(a)\exp[-\nu\mathsf{d}^{\circ}(a)],\ \nu \equiv 1/(p-1) \qquad (14)$$
$$\mathsf{d}^{\circ}(a) \equiv \int_{\{s\}} \mathsf{m}(s|a)\ln\left(\frac{\mathsf{m}(s|a)}{\mathsf{h}(s)\mathsf{m}^{\mathsf{i}\circ}(s|a)}\right)\mathrm{d}s \overbrace{\leq}^{(10)} \mathsf{d}(a).$$

The $\mathsf{r}^{\mathsf{i}\circ}$-factor (14) is in (12) and thus it meets (11).

Proof The range of $\mathsf{d}(a)$ (8) implies $\exp[-\mathsf{d}(a)] \in [0,1]$. This and (13) make $||\exp[-\mathsf{d}^{\circ}]||_q < \infty$ for $q \equiv \frac{p}{p-1} = p\nu$. Hölder's inequality [39] implies that maximiser of (8) is $\mathsf{r}^{\mathsf{i}\circ}(a) \propto \chi_{\{a\}}(a)\exp\big(-\frac{q}{p}\mathsf{d}^{\circ}(a)\big)$, i.e. (14) hold. The repeated assumptions of Prop. 2 guarantee that $\mathsf{d}^{\circ}(a) < \infty$ on $\{a\}$ so that $\exp[-\nu\mathsf{d}^{\circ}(a)] > 0$ on $\{a\}$, i.e. $\mathrm{supp}[\mathsf{r}^{\mathsf{i}\circ}] = \{a\}$ and constraints (12), (11) are met. □

Remarks 2

▶ *The equality (5) implies similarity of r° and $\mathsf{r}^{\mathsf{i}\circ}$ as $\mathsf{r}^{\circ} \propto \mathsf{r}^{\mathsf{i}\circ}\exp[-\mathsf{d}^{\circ}] \propto \exp[-\nu\mathsf{d}^{\circ}]\exp[-\mathsf{d}^{\circ}]$, $\nu \in (0,\infty)$.*

▶ *The constraint (11) guarantees that the ideal r^{i}-factor supports exploration (the key dual feature of the optimal learning strategy [6,15]) as it* a priori *forbids no action.*

▶ *The opted value $\nu > 0$ controls exploration. The higher the parameter ν, the weaker the exploration becomes.*

▶ *The requirement (12), implying (11), is almost unrestrictive for $p \to 1^+$, i.e. for a large ν (14).*

▶ *The value function $-\ln(\mathsf{h}(s))$, Prop. 1, influences the $\mathsf{r}^{\mathsf{i}\circ}$-factor (14) via $\mathsf{d}^{\circ}(a)$, but, as said, not the $\mathsf{m}^{\mathsf{i}\circ}$-factor.*

▶ *The strong assumption (13) avoids technicalities. Any weaker one giving $||\exp[-\mathsf{d}^{\circ}]||_q < \infty$ suffices.*

▶ *Other wishes on $a \in \{a\}$ can be met by enforcing $\exp[-\mathsf{g}(a)]$, with a given $\mathsf{g}(a) \in [0,\infty)$, into r^{i}, i.e. using*

$$\mathsf{r}^{\mathsf{i}} \in \{\mathsf{r}^{\mathsf{i}}\} \equiv \{\mathsf{r}^{\mathsf{i}} : \mathsf{r}^{\mathsf{i}}(a) \propto \mathsf{f}(a)\exp[-\mathsf{g}(a)] \qquad (15)$$
$$\text{with an opted } \mathsf{f}(a) > 0 \text{ on } \{a\} \text{ giving } ||\mathsf{r}^{\mathsf{i}}||_p < \infty\}.$$

The function $\mathsf{g}(a)$ may express the quest for low cost of DM or deliberation. As a rule, it is loosely given and a domain-specific PE is needed [35].

▶ *The replacement of (12) by (15) in Prop. 3 yields*

$$\mathsf{r}^{\mathsf{i}\circ} \propto \chi_{\{a\}}(a)\exp[-(\nu+1)\mathsf{g}(a) - \nu\mathsf{d}^{\circ}(a))].$$

3.3 Specific Optimal Ideal $\mathsf{c}^{\mathsf{i}\circ}$-Factor

This part specialises Sec. 3.1, 3.2. Its result is useful per se. It hints how to proceed in other cases.

The optimal ideal $\mathsf{r}^{\mathsf{i}\circ}$-factor is uniquely given by[6] $\mathsf{m}^{\mathsf{i}\circ}$, $\mathsf{r}^{\mathsf{i}\circ} = \mathsf{r}^{\mathsf{i}\circ}(\mathsf{m}^{\mathsf{i}\circ})$, and by the opted $\nu > 1$, see (14). This allows us to meet agent's wishes by opting $\mathsf{m}^{\mathsf{i}\circ} \in \{\mathsf{m}^{\mathsf{i}}\}$. The treated version deals with the agent's wish:

$$\text{Reach ideal sets } \emptyset \neq \{s^{\mathsf{i}}\} \subseteq \{s\},\ \emptyset \neq \{a^{\mathsf{i}}\} \subseteq \{a\}\,! \qquad (16)$$

This is taken as the wish to assign high probabilities to the sets of ideal states $\{s^{\mathsf{i}}\}$ and of ideal actions $\{a^{\mathsf{i}}\}$ (16). The probabilities arise by closing the loop of the

[5] It needs absolute continuity of m with respect to a potential m^{i} [39], e.g., $\mathsf{m}^{\mathsf{i}} > 0$ on $\{s\}$ for all conditions meets it.

[6] The abusing symbol $\mathsf{r}^{\mathsf{i}\circ}(\mathsf{m}^{\mathsf{i}\circ})$ stresses the used dependence.

given, un-mutable, state-transition model with the optimal ideal DM rule $\mathsf{r}^{\mathsf{i}\mathsf{o}} = \mathsf{r}^{\mathsf{i}\mathsf{o}}(\mathsf{m}^{\mathsf{i}\mathsf{o}})$ (14)

$$\mathsf{r}^{\mathsf{i}\mathsf{o}}(\mathsf{m}^{\mathsf{i}\mathsf{o}}) \in \underset{\mathsf{m}^{\mathsf{i}}\in\{\mathsf{m}^{\mathsf{i}}\}}{\mathrm{Arg\,max}} \int_{\{a\}} \rho(a)\mathsf{r}^{\mathsf{i}}(a)\mathrm{d}a \equiv \tag{17}$$

$$\underset{\mathsf{m}^{\mathsf{i}}\in\{\mathsf{m}^{\mathsf{i}}\}}{\mathrm{Arg\,max}} \int_{\{a\}} \Big[\int_{\{s\}} \chi_{\{s^{\mathsf{i}}\}}(s)\mathsf{m}(s|a)\mathrm{d}s + w\chi_{\{a^{\mathsf{i}}\}}(a) \Big] \mathsf{r}^{\mathsf{i}}(a)\mathrm{d}a.$$

The weight $w \geq 0$ assigns the importance of acting in $\{a^{\mathsf{i}}\} \subset \{a\}$ relatively to reaching $\{s^{\mathsf{i}}\} \subset \{s\}$. The task (17) has a meaningful solution if the sets $\{s^{\mathsf{i}}\}$, $\{a^{\mathsf{i}}\}$ are

"probabilistically reachable" $\Leftrightarrow$ $\rho(a) > 0$ on $\{a\}$. (18)

Remarks 3

▶ *If $\rho(a) > 0$ only on a non-empty proper subset of $\{a\}$ then it makes no sense to choose an action out of it. Then, $\{a\}$ reduces to this subset.*

▶ *The rule r^{o} (4) could replace the ideal DM rule $\mathsf{r}^{\mathsf{i}\mathsf{o}}$ (17) giving the almost same result, cf. the 1st item in Rem. 2.*

▶ *The paper uses a fixed weight $w \geq 0$. Its fine tuning is made via the agent's marking of the seen quality [27].*

▶ *Any reward of the usual DM [14] may serve as the function of (s, a) defining $\rho(a)$ in (17). As said in Rem. 1, our construction quantifies the agent's wishes in an ambitious but realistic way and cares about the exploration.*

The solution of (17) gives the optimal values of d^{o} (14).

Proposition 4 (Optimal Values $\mathsf{d}^{\mathsf{o}}(a)$, $a \in \{a\}$) *Let* $\mathrm{supp}[\{\mathsf{r}^{\mathsf{i}}\}] \equiv \{a\}$, $||\mathsf{r}^{\mathsf{i}}||_p < \infty, p > 1$, *and* $|\{a\}| < \infty$. *Let*[7] *each* $\mathsf{r}^{\mathsf{i}} \in \{\mathsf{r}^{\mathsf{i}}\}$ *(12) define a non-empty cross-section* $\{\mathsf{m}^{\mathsf{i}}\}$ *of (9) and* $\mathsf{m}^{\mathsf{i}}(s|a) \in \{\mathsf{m}^{\mathsf{i}}\}$ *exist giving* $\mathsf{d}(a) < \infty$, $\forall a \in \{a\}$. *Let the assumption (18) be met. Then, the optimal ideal* $\mathsf{m}^{\mathsf{i}\mathsf{o}}$ *meeting (17) provides* $\mathsf{d}^{\mathsf{o}}(a)$, *giving* $\mathsf{r}^{\mathsf{i}\mathsf{o}} = \mathsf{r}^{\mathsf{i}}(\mathsf{m}^{\mathsf{i}\mathsf{o}})$ *(14), as the next function*

$$\mathsf{d}^{\mathsf{o}}(a) = \mathsf{d}^{\mathsf{o}}(\bar{a}) + \ln\left[\frac{\rho(\bar{a})}{\rho(a)}\right],\ \bar{a} \in \underset{a\in\{a\}}{\mathrm{Arg\,max}}[\rho(a)],\ a \in \{a\}.$$

Proof For $p > 1$, $||\mathsf{r}^{\mathsf{i}\mathsf{o}}||_p < \infty$ and $|\{a\}| < \infty$, (13) gives $||\rho||_q < \infty$ with $q \equiv \frac{p}{p-1} \equiv p\nu$. Hölder's inequality [39] applied to the functional (17) provides its maximiser

$$\mathsf{r}^{\mathsf{i}\mathsf{o}}(a) = \kappa^{\nu}\rho^{\nu}(a) \overbrace{=}^{(14)} \frac{\exp[-\nu\mathsf{d}^{\mathsf{o}}(a)]}{||\exp[-\mathsf{d}^{\mathsf{o}}]||_{\nu}^{\nu}}. \tag{19}$$

The normalization of $\mathsf{r}^{\mathsf{i}\mathsf{o}}$ gives the a-independent factor κ^{ν}. $\kappa > 0$ due to the finite volume of the action set (13), i.e. $\kappa = ||\rho||_{\nu}^{-1} > 0$. Also the norm-defining integral $\int \exp(-\nu\mathsf{d}^{\mathsf{o}}(a))da \in (0, \infty)$ due to $\mathsf{d}^{\mathsf{o}} \in (0, \infty)$ and the finite volume of $\{a\}$ (13). The logarithmic version of the equation (19) gives $\forall a \in \{a\}$,

$$\begin{aligned}\mathsf{d}^{\mathsf{o}}(a) &= -\ln(\kappa||\exp[-\mathsf{d}^{\mathsf{o}}]||_{\nu}) - \ln[\rho(a)] \\ &\equiv \Phi(\kappa) - \ln[\rho(a)].\end{aligned} \tag{20}$$

As $\Phi(\kappa)$ is independent of a, the equality (20) gives

$$\Phi(\kappa) = \mathsf{d}^{\mathsf{o}}(\bar{a}) + \ln[\rho(\bar{a})] \ \text{ for } \ \bar{a} \in \underset{a\in\{a\}}{\mathrm{Arg\,max}}(\rho(a)) \Leftrightarrow \tag{21}$$

$$\bar{a} \in \underset{a\in\{a\}}{\mathrm{Arg\,min}}\, \mathsf{d}^{\mathsf{o}}(a), \ \text{which yields (19).} \qquad \square$$

[7] These sentences are the assumptions of Props. 2 and 3.

Proposition 5 (Solvability of (19)) *Under (18) and $|\{a\}| < \infty$, the smallest $\mathsf{d}^{\mathsf{o}}(\bar{a})$ exists such that (19) has a solution $\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)$, $s \in \{s\}$, $\forall a \in \{a\}$, (21).*

Proof Properties of the KLD conditioned on $a \in \{a\}$ (and implicitly on s_{t-1}) imply that the values $\mathsf{d}^{\mathsf{o}}(a) \in [-\int_{\{s\}} \mathsf{m}(s|a)\ln(\mathsf{h}(s))\mathrm{d}s, \infty] \subset [0, \infty]$. Indeed, the option $\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a) \equiv \mathsf{m}(s|a)$ attains the lower bound. The upper bound is reached for $\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)$ singular with respect to $\mathsf{m}(s|a)$, i.e. being zero on a subset of $\{s\}$ to which $\mathsf{m}(s|a)$ assigns a positive probability. Thus, the smallest $\mathsf{d}^{\mathsf{o}}(\bar{a})$ guaranteeing solvability of (19) $\forall a \in \{a\}$ is

$$\mathsf{d}^{\mathsf{o}}(\bar{a}) = \max\left[0, \max_{a\in\{a\}} \int_{\{s\}} \mathsf{m}(s|a)\ln\left[\frac{\rho(a)}{\rho(\bar{a})\mathsf{h}(s)}\right]\mathrm{d}s\right]. \tag{22}$$

For $|\{a\}| < \infty$, $\mathsf{h}(s) \in (0, 1]$ and $\rho(a) > 0$, the maximum in (22) is finite and the range of $\mathsf{d}^{\mathsf{o}}(\bar{a})$ implies existence of $\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|\bar{a})$ with $\mathsf{d}^{\mathsf{o}}(\bar{a})$ (22). $\square$

The ideal $\mathsf{m}^{\mathsf{i}\mathsf{o}}$ gives $\mathsf{d}^{\mathsf{o}}(a)$ (8) and $\mathsf{r}^{\mathsf{i}\mathsf{o}}(\mathsf{m}^{\mathsf{i}\mathsf{o}})$ via (14). The next proposition provides it for generic pds $\mathsf{m}(s|a)$.

Proposition 6 ($\mathsf{m}^{\mathsf{i}\mathsf{o}}$ Meeting (17), Generic $\mathsf{m}(s|a)$) *Let $\mathsf{m}(s|a)$, $a \in \{a\}$, be non-uniform on $\{s\}$ and Prop. 3 hold. Then, the $\mathsf{m}^{\mathsf{i}\mathsf{o}}$-factor meeting (17) has the form*

$$\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a) = \frac{\mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]}{\int_{\{s\}} \mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]\mathrm{d}s} \tag{23}$$

$$\text{defined under the adopted assumption } |\{s\}| < \infty. \tag{24}$$

The real-valued $\mathsf{e}(a)$ in (23) is the existing solution of $\mathsf{L}(\mathsf{e}(a)) = \mathsf{R}(a)$. For $\mathsf{d}^{\mathsf{o}}(\bar{a})$ meeting (22) with $\bar{a} \in \mathrm{Arg\,max}_{a\in\{a\}}[\rho(a)]$, the left- and right-hand sides of this equation are

$$\mathsf{L}(\mathsf{e}(a)) \equiv \mathsf{e}(a)\Lambda(a) + \ln\left[\int_{\{s\}} \mathsf{m}(s|a)\exp[-\mathsf{e}(a)\mathsf{m}(s|a)]\mathrm{d}s\right]$$

$$\Lambda(a) \equiv \int_{\{s\}} \mathsf{m}^2(s|a)\mathrm{d}s > 0 \tag{25}$$

$$\mathsf{R}(a) \equiv \int_{\{s\}} \mathsf{m}(s|a)\ln(\mathsf{h}(s))\mathrm{d}s + \mathsf{d}^{\mathsf{o}}(\bar{a}) + \ln\left[\frac{\rho(\bar{a})}{\rho(a)}\right] \geq 0.$$

Proof Let us fix $a \in \{a\}$ with a non-uniform $\mathsf{m}(s|a)$. Then, (19) with $\mathsf{d}^{\mathsf{o}}(\bar{a})$ given by (22) is Fredholm's integral equation for the unknown function $\ln(\mathsf{m}(s|a)/\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a))$, $s \in \{s\}$. Its particular solution $\ln(\mathsf{m}(s|a)/\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)) = \mathsf{e}(a)\mathsf{m}(s|a) + \mathsf{v}(a)$ with a proper, scalar-valued, $\mathsf{e}(a)$ and $\mathsf{v}(a)$ suffices. The searched function could have a summand $\mathsf{o}(s|a)$ orthogonal to $\mathsf{m}(s|a)$, i.e. with $\int_{\{s\}} \mathsf{m}(s|a)\mathsf{o}(s|a)\mathrm{d}s = 0$. Its use is here unnecessary. The exploited solution form and normalisation of $\mathsf{m}^{\mathsf{i}\mathsf{o}}$ give (23). It remains to find $\mathsf{e}(a)$.

Equation (19) for $\mathsf{d}^{\mathsf{o}}(a)$ and the $\mathsf{d}(a)$ definition (4) give $\mathsf{e}(a)$ in (23) as a solution of $\mathsf{L}(\mathsf{e}(a)) = \mathsf{R}(a)$ with left- and right-hand sides given by (25).

The 1$^{\mathrm{st}}$ derivative of $\mathsf{L}(\mathsf{e}(a))$ with respect to $\mathsf{e}(a)$, $a \in \{a\}$, reads ($\mathsf{m}^{\mathsf{i}\mathsf{o}}$ means the expression (23))

$$\frac{\mathrm{d}\mathsf{L}(\mathsf{e}(a))}{\mathrm{d}\mathsf{e}(a)} = \Lambda(a) - \int_{\{s\}} \mathsf{m}(s|a)\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)\mathrm{d}s.$$

The 2^{nd} derivative is the positive variance of the non-constant $\mathsf{m}(s|a)$ with respect to $\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)$

$$\frac{\mathrm{d}^2\mathsf{L}(\mathsf{e}(a))}{\mathrm{d}\mathsf{e}^2(a)} = \int_{\{s\}} \mathsf{m}^2(s|a)\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)\mathrm{d}s - \left[\int_{\{s\}} \mathsf{m}(s|a)\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a)\mathrm{d}s\right]^2 > 0.$$

Thus, $\mathsf{L}(\mathsf{e}(a))$ is strictly convex in $\mathsf{e}(a)$.
For $\mathsf{e}(a) = 0$, $\mathsf{L}(0) = 0 \leq \mathsf{R}(a)$ as $\mathsf{R}(a) \geq 0$ due to (22) and (19). For the non-constant $\mathsf{m}(s|a)$, $\lim_{\mathsf{e}(a)\to\infty} \mathsf{L}(\mathsf{e}(a)) = \infty$ as $\Lambda(a) > 0$. The case $\Lambda(a) = 0$ is excluded by the normalisation $\int_{\{s\}} \mathsf{m}(s|a)\mathrm{d}s = 1$.

Thus, the left-hand side $\mathsf{L}(\mathsf{e}(a))$, continuously dependent on $\mathsf{e}(a)$, intersects $\mathsf{R}(a)$ at most for two values of $\mathsf{e}(a)$ solving the inspected equation. The solution leading to the smaller (non-negative) value $\mathsf{d}^{\mathsf{o}}(a)$ is the proper one. The strict convexity guarantees that the numerical search for the solution is trouble-less. □

The next proposition addresses the yet unsolved case.

Proposition 7 ($\mathsf{m}^{\mathsf{i}\mathsf{o}}$ Meeting (17), Uniform $\mathsf{m}(s|a)$) *For uniform pd $\mathsf{m}(s|a)$ on $\{s\}$ with $|\{s\}| < \infty$, the optimal $\mathsf{m}^{\mathsf{i}\mathsf{o}}$-factor meeting (19) has the form*

$$\mathsf{m}^{\mathsf{i}}(s|a) = \frac{\exp[-\mathsf{e}(a)\mathsf{o}(s|a)]}{\int_{\{s\}} \exp[-\mathsf{e}(a)\mathsf{o}(s|a)]\mathrm{d}s} \tag{26}$$

for an arbitrary non-zero $\mathsf{o}(s|a)$ with $\int_s \mathsf{o}(s|a)\mathrm{d}s = 0$. The real-valued $\mathsf{e}(a)$ is that of the pair existing solutions of (27), which makes the corresponding $\mathsf{d}^{\mathsf{o}}(a)$ smaller.

$$\mathsf{L}(\mathsf{e}(a)) \equiv \ln\left[\frac{\int_{\{s\}} \exp[-\mathsf{e}(a)\mathsf{o}(s|a)]\mathrm{d}s}{|\{s\}|}\right] = \mathsf{R}(a)$$
$$\equiv \mathsf{d}^{\mathsf{o}}(\bar{a}) + \int_{\{s\}} \mathsf{m}(s|a) \ln\left[\frac{\mathsf{h}(s)\rho(\bar{a})}{\rho(a)}\right] \mathrm{d}s. \tag{27}$$

Proof Let us consider $a \in \{a\}$ with a uniform $\mathsf{m}(s|a)$. Then, (19) with $\mathsf{d}^{\mathsf{o}}(\bar{a})$ given by (22) is Fredholm's integral equation for the unknown function $\ln(\mathsf{m}(s|a)/\mathsf{m}^{\mathsf{i}\mathsf{o}}(s|a))$, $s \in \{s\}$. Its particular solution is searched in the form $\ln(\mathsf{m}(s|a)/\mathsf{m}^{\mathsf{i}}(s|a)) = \mathsf{e}(a)\mathsf{o}(s|a) + \mathsf{v}(a)$. The choice $\int_{\{s\}} \mathsf{o}(s|a)\mathrm{d}s = 0$ makes $\mathsf{o}(s|a)$ orthogonal to the uniform $\mathsf{m}(s|a)$ and gives (26). The definition of $\mathsf{d}(a)$ (8) and equation (19) provide (27).

Inspection of the 1^{st} and 2^{nd} derivatives of $\mathsf{L}(\mathsf{e}(a))$ in (27) with respect to $\mathsf{e}(a)$ shows that it is convex function for the inevitably non-constant $\mathsf{o}(s|a)$.
The left-hand side $\mathsf{L}(\mathsf{e}(a))$ of (27) is zero for $\mathsf{e}(a) = 0$, while right-hand side is non-negative for $\mathsf{d}^{\mathsf{o}}(\bar{a})$ (22). Also, $\lim_{\mathsf{e}(a)\to\pm\infty} \mathsf{L}(\mathsf{e}(a)) = \infty$ as $\mathsf{o}(s)$ must be negative (positive) on a subset of $\{s\}$ of a positive volume. This implies the nature and existence of the solution of (27). □

4 Algorithmic Summary

Algorithm 1 summarises the results. It does it for the closed loop with a finite amount of state and action values. This simple but useful case shows the evaluation structure while avoiding hard integrations and potential infiniteness of the volumes $|\{s\}|$, $|\{a\}|$. The conditioning state $\tilde{s} = s_{t-1}$ is there explicit.
The proposed PE becomes practical by using Bayesian estimation of unknown but time-invariant values of transition probabilities Θ. This parametric model $\mathsf{m}(s_t|a_t, s_{t-1}, \Theta)$ belongs to exponential family [2] and makes Dirichlet's prior pd self-reproducing. Its degrees of freedom, counting the observed transitions $s_{t-1} = \tilde{s} \in \{s\}$, $a_t = a \in \{a\}$ to $s_t = s \in \{s\}$, form the sufficient statistic for learning the unknown values $\Theta_{s|a,\tilde{s}} \equiv \mathsf{m}(s|a, \tilde{s}, \Theta)$, $s, \tilde{s} \in \{s\}$, $a \in \{a\}$, [21].
The explorative nature of FPD allows us to employ a certainty-equivalent strategy that uses point estimates of transition probabilities instead of them. With a forgetting, [28] an adaptive agent's strategy arises.

5 Illustrative Experiments

The experiments illustrate the theory. They apply Algorithm 1 with the following common options:
▶ The recursive point estimation of transition probabilities provides us with the environment model

$$\mathsf{m}(s|a, \tilde{s}) \propto \text{number of occurences of } (s, a, \tilde{s}) + 1.$$

The added 1 reflects the used uniform prior pd of the estimated probabilities. Let us stress that the gained approximate certainty-equivalent strategy is explorative as the FPD-optimal strategy is randomised.
▶ The receding horizon in the design cycle was set $T = 100$ and exploration parameter $\nu = 1 \Leftrightarrow p = 2$.
▶ The initial guess of the exponent in (23), (26), solving (25) or (27) was set $\mathsf{e}(a|\tilde{s}) = 1.2$, $\forall a \in \{a\}$, $\tilde{s} \in \{s\}$.
▶ The initial state was always $s_0 = 1$ and the seed of pseudo-random generators was fixed.
Figures show the amount of visits of states and actions. Their numbering follows that of inspected *cases*.

Toy Example The double-stochastic, static, transition pd with dominating diagonal was simulated for five hundred time steps while considering 3 states and 3 actions

$$[\mathsf{Probability}(s|a, \tilde{s})]_{s,\tilde{s}\in\{s\},a\in\{a\}} \equiv$$
$$[\mathsf{Probability}(s|a)]_{s\in\{s\},a\in\{a\}} \equiv \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{bmatrix}. \tag{28}$$

This choice makes the expected optimal behaviour quite intuitive and the influence of the weight w (17), favouring the preferred actions, predictable.
Case 1: No Preference on Actions The experiments show in Fig. 1 the counts of states and actions for three different wishes on states. The optimisation seeks the highest counts of the preferred state. The environment (28) allows this as all states may appear with a positive probability. The desired state is $\{s^{\mathsf{i}} = k\}$ in Experiment no $k \in \{1, 2, 3\}$. No action is preferred, $\{a^{\mathsf{i}}\} = \{a\} = \{1, 2, 3\}$.
Discussion The results in Fig. 1 confirm the expected behaviour for the diagonally dominated, static environment (28) with no wishes on actions. The counts of the preferred state are the highest ones. The desired explorative actions make them slightly lower than possible.

Algorithm 1 FPD with Preference Quantification for Behaviours with a Finite Number of Realisations

Inputs

✓ Finite sets of states $\{s\}$ & actions $\{a\}$, sets of ideal states $\{s^{\mathsf{i}}\} \subset \{s\}$ & actions $\{a^{\mathsf{i}}\} \subset \{a\}$

✓ Relative weight $w \geq 0$ of $\{s^{\mathsf{i}}\}$, $\{a^{\mathsf{i}}\}$ (17)

✓ Environment model $\mathsf{m}(s|a,\tilde{s})$, $s,\tilde{s} \in \{s\}, a \in \{a\}$

✓ Design horizon T, exploration controlling $\nu > 1$ and the value function $\mathsf{h}(s) \equiv 1, \forall s \in \{s\}$ (4)

Evaluation of h-independent variables

For $\tilde{s} \in \{s\}$ **do**

For $a \in \{a\}$ **do**

$\rho(a|\tilde{s}) = \sum_{s\in\{s^{\mathsf{i}}\}} \mathsf{m}(s|a,\tilde{s}) + \chi_{\{a^{\mathsf{i}}\}}(a)w$ (17)

$\Lambda(a|\tilde{s}) \equiv \sum_{s\in\{s\}} \mathsf{m}^2(s|a,\tilde{s})$ (25)

end $a \in \{a\}$

$\bar{a}(\tilde{s}) \in \operatorname{Arg}\max_{a\in\{a\}} \rho(a|\tilde{s})$ (19)

$\bar{\rho}(\tilde{s}) \equiv \rho(\bar{a}(\tilde{s})|\tilde{s})$ (19)

end $\tilde{s} \in \{s\}$

Design cycle for $t = T, T-1, \ldots, 1$

For $\tilde{s} \in \{s\}$ **do**

$\mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s})) \equiv \max\Big\{0,$

$\max_{a\in\{a\}}\Big[\sum_{s\in\{s\}} \mathsf{m}(s|a,\tilde{s}) \ln\Big[\frac{\rho(a|\tilde{s})}{\bar{\rho}(\tilde{s})\mathsf{h}(s)}\Big]\Big]\Big\}$ (22)

For $a \in \{a\}$ **do**

$\mathsf{d}^{\mathsf{o}}(a|\tilde{s}) = \mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s})) + \ln\Big(\frac{\bar{\rho}(\tilde{s})}{\bar{\rho}(\tilde{s})}\Big)$ (19)

If $\mathsf{m}(s|a,\tilde{s})$ is not uniform

$\mathsf{R}(a|\tilde{s}) = \mathsf{d}^{\mathsf{o}}(a|\tilde{s})$
$+ \sum_{s\in\{s\}} \mathsf{m}(s|a,\tilde{s}) \ln(\mathsf{h}(s))$ (25)

Find $\mathsf{e}(a|\tilde{s})$ in $\mathsf{R}(a|\tilde{s}) = \mathsf{e}(a|\tilde{s})\Lambda(a|\tilde{s})$ (25)
$+ \ln\big(\sum_{\{s\}} \mathsf{m}(s|a,\tilde{s}) \exp[-\mathsf{e}(a|\tilde{s})\mathsf{m}(s|a,\tilde{s})]\big)$

$\mathsf{m}^{\mathsf{io}}(s|a,\tilde{s}) \propto$
$\mathsf{m}(s|a,\tilde{s}) \exp[-\mathsf{e}(a|\tilde{s})\mathsf{m}(s|a,\tilde{s})]$ (23)

else

Choose $\mathsf{o}(s)$ such that $\sum_{s\in\{s\}} \mathsf{o}(s) = 0$

Find $\mathsf{e}(a|\tilde{s})$ in $\ln\Big[\sum_{s\in\{s\}} \frac{\exp[-\mathsf{e}(a|\tilde{s})\mathsf{o}(s)]}{|\{s\}|}\Big] =$
$\mathsf{d}^{\mathsf{o}}(\bar{a}(\tilde{s}) + \frac{1}{|\{s\}|}\sum_{s\in\{s\}} \ln\Big[\frac{\mathsf{h}(s)\bar{\rho}(\tilde{s})}{\rho(a|\tilde{s})}\Big]$

Set $\mathsf{m}^{\mathsf{io}}(s|a) \propto \exp[-\mathsf{e}(a|\tilde{s})\mathsf{o}(s)]$. (26)

end if on uniform m

$\mathsf{r}^{\mathsf{io}}(a|\tilde{s}) = \exp\big[-\nu\mathsf{d}^{\mathsf{o}}(a|\tilde{s})\big]$ (14)

end $a \in \{a\}$

$\mathsf{r}^{\mathsf{io}}(a|\tilde{s}) = \frac{\mathsf{r}^{\mathsf{io}}(a|\tilde{s})}{\sum_{\{a\}} \mathsf{r}^{\mathsf{io}}(a|\tilde{s})}$, $a \in \{a\}$ (14)

$\mathsf{n}(\tilde{s}) = \sum_{a\in\{a\}} \mathsf{r}^{\mathsf{io}}(a|\tilde{s}) \exp[-\mathsf{d}^{\mathsf{o}}(a|\tilde{s})]$ (4)

$\mathsf{r}^{\mathsf{o}}(a|\tilde{s}) = \frac{\exp[-(\nu+1)\mathsf{d}^{\mathsf{o}}(a|\tilde{s})]}{\mathsf{n}(\tilde{s})}$, $a \in \{a\}$ (5)

end $\tilde{s} \in \{s\}$

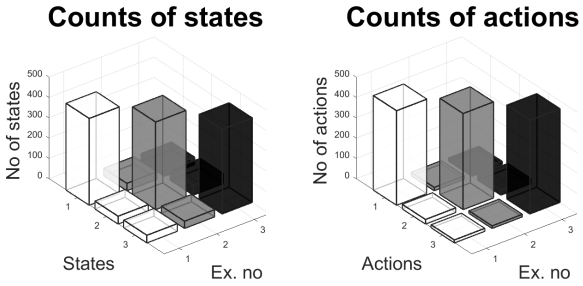
$\mathsf{h}(s) = \mathsf{n}(s), \forall s \in \{s\}$ (4)

end of the design cycle

Outputs

✓ All optimal ideal m^{io}-, r^{io}- and r^{o}-factors

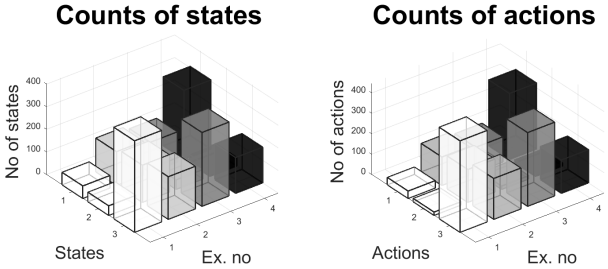
Case 2: Influence of Learning and Balancing Wishes on States and Actions Experiments cover two aspects documented in Fig. 2. Firstly, the need for learning during DM is stressed. The learnt environment model serves also the preference quantification. This combination results in the data-based preference elicitation. The results with learning switched off serve us for comparison. The used fixed model $\mathsf{m}(s|a,\tilde{s})$ was the discretised normal pd with the mean $a+\tilde{s}$ and the variance 3.



(a) States (b) Actions

Fig. 1. Counts of states and actions with different preferences on states for Ex. no 1: $\{s^{\mathsf{i}}\} = \{1\}$, Ex. no 2: $\{s^{\mathsf{i}}\} = \{2\}$, Ex. no 3: $\{s^{\mathsf{i}}\} = \{3\}$ and no preference on actions.

Secondly, nontrivial wishes on actions are inspected. Experiments indicate how the weight w (17) balances wishes with respect to states and actions.



(a) States (17) (b) Actions (17)

Fig. 2. Counts of states and actions with wishes on states $\{s^{\mathsf{i}}\} = \{3\}$ and actions $\{a^{\mathsf{i}}\} = \{1\}$ with different weights (17). Ex. no 1: $w = 0$ (no wish on actions) with learning, Ex. no 2: $w = 0$ no learning, Ex. no 3: $w = 0.4$, Ex. no 4: $w = 1$.

Discussion Fig. 2 shows that the designs with learning, Ex. no 1, and without learning, Ex. no 2, behave as expected. No-learning design deals with the wrong model. Thus, the exploitative role of actions is almost given up. Their exploratory role dominates.

The results of Ex. no 1, 3, 4 confirm predictable properties. The desired state $\{3\}$ is the less visited the more important the "unhelpful" action $\{a^{\mathsf{i}}\} = \{1\}$ is asked.

Experience from non-presented experiments

They confirmed: ▶ the positive influence of the optimal design; ▶ significant improvements of quality compared to various heuristic preference quantification; ▶ a smooth but non-linear dependency of the resulting performance on the weight w and non-triviality of its plausible numerical choice in the "reward" (17); ▶ the predictable influence of the set $\{a^{\mathsf{i}}\}$ and of the weight w.

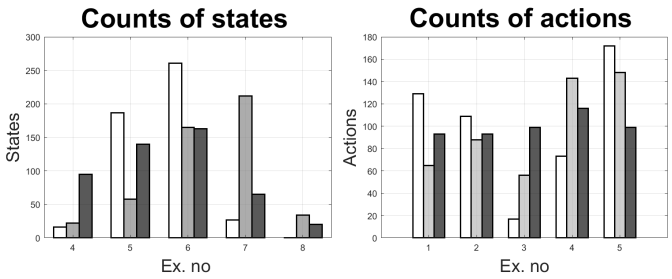
Realistic Example Here, experiments with a more realistic dynamic environment having eleven states and

five actions are presented. The simulated environment was gained by learning the transition pd from observed 1000 states $\{s\} \equiv \{1, \ldots, 11\}$ stimulated by independently generated discrete actions $\{a\} \equiv \{1, \ldots, 5\}$. The states were gained via an affine mapping of discretised values $s_t = \mathsf{floor}(2.6311 y_t - 1.2838)$ of y_t observed on the simulated normal linear model

$y_t = 0.99 y_{t-1} + 0.05 a_t - 0.125 + 0.05 e_t$, where the white noise e_t has the constant zero mean and unit variance. The stationary expected level $y \approx 5.5$, reached for the expected $a \approx 2.5$, is interpreted as zero "spent energy".

Remarks 4 *Figures reflecting the realistic example only show the counts of the states from 4 to 8 as the counts of the other states were negligible.*

Case 3: Varying Ideal States, No Preference on Actions These experiments, documented in Fig. 3, show the behaviour of the closed loop with the $11 \times 5 \times 11$ environment. The preferred states are the reachable ideal states $\{s^{\mathsf{i}}\} = \{6\}$, $\{s^{\mathsf{i}}\} = \{7\}$ and the unrealistic ideal state $\{s^{\mathsf{i}}\} = \{11\}$, whose probability approaches zero. No wishes are put on actions, i.e. $\{a^{\mathsf{i}}\} = \{a\} = \{1, \ldots, 5\}$.



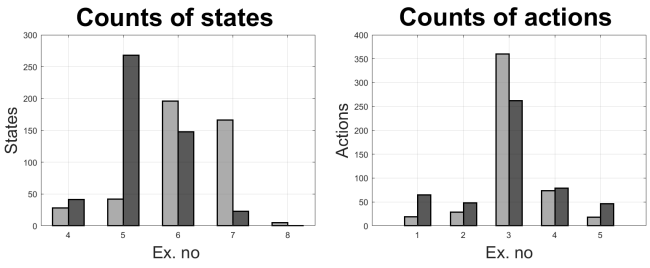
(a) States (b) Actions

Fig. 3. Counts of states and actions with different preferences on states demonstrate the differences in achievable quality for realistic and unrealistic wishes. Ex. no 1: (white) $\{s^{\mathsf{i}}\} = \{6\}$, Ex. no 2: (light gray) $\{s^{\mathsf{i}}\} = \{7\}$, Ex. no 3: (dark gray) $\{s^{\mathsf{i}}\} = \{11\}$. No extra wish is put on actions.

Discussion The state $\{s^{\mathsf{i}}\} = \{6\}$ is easier to reach than the state $\{s^{\mathsf{i}}\} = \{7\}$. It is seen in a higher number of occurrences of the ideal state. No optimisation can cope with practically unreachable state $\{s^{\mathsf{i}}\} = \{11\}$. Its occurrences were negligibly small and the optimised actions were almost uniformly distributed as driven by the built-in exploration (11).

Case 4: Extension of the Set of Ideal States The aim of this case is to show that an extension of the set of ideal states may have a huge impact on the reached distribution of states and actions. The considered extended set $\{s^{\mathsf{i}}\} = \{5, 6, 7\}$ includes the originally preferred state $\{s^{\mathsf{i}}\} = \{6\}$. The desired action $\{a^{\mathsf{i}}\} = \{3\}$ is the same.

Discussion Fig. 4 confirms that the extension of the set $\{s^{\mathsf{i}}\} = \{6\}$ to the set $\{s^{\mathsf{i}}\} = \{5, 6, 7\}$ had the significant impacts on the states. The counts of the preferred state for the unextended set is 196. For the extended set, it is 439. The preferred action predictably appears less often.



(a) States (b) Actions

Fig. 4. Relaxed target states change the reached distribution. Counts of states and actions for different targets. Ex. no 1: (light gray) $\{s^{\mathsf{i}}\} = \{6\}$, Ex. no 2: (dark gray) $\{s^{\mathsf{i}}\} = \{5, 6, 7\}$ with the ideal action $\{a^{\mathsf{i}}\} = \{3\}$ with weight $w = 0.3$.

The result appeals to agents to specify the ideal states in a realistic way. Note that the observed plausible results can hardly be gained by a priori chosen ideal pd.

6 Concluding Remarks

The paper advances the completion and quantification of preferences within the fully probabilistic design of DM strategies. It derives the optimal ideal closed-loop model c^{io} from: ▶ the set of allowed actions; ▶ the agent's wish to get closed-loop states and actions (*novel*) from the ideal sets with the highest probability; ▶ the query-tunable [27] exploration-controlling scalar ν and the scalar weight w balancing the importance of the ideal states and actions; ▶ the learnt state transitions.

The paper extends the former uses [24,26] of the universal preference elicitation principle by *allowing to control exploration and to balance (often contradictory) wishes on states and actions*. The papers [27,44] complement this one by experiments. It has required refining the used optimisation techniques as mentioned on the fly. Compared to forerunners, the *solution avoids the greedy approximation and cares both about the target states and actions*. It is of extreme practical importance: please think, e.g., about balancing the economic costs and the population's health state during Covid 19 [33,30].

Methodologically, the elaborated principle is the wishes-focused twin to the minimum KLD principle serving the knowledge elicitation. The reached state of its elaboration provides a firm basis for trial real-life uses.

The solution combined with on-line learning of the state-transition model achieves the dreamt learning of preference [37]. It is worth stressing that the quantified preferences are both ambitious and realistic. For instance, the quantification of wishes does not demand avoiding unavoidable closed-loop behaviours.

The presented research is an open-ended story, which surely requires dealing with: ▶ other sets of wishes, say, balancing the importance of state entries as needed in multi-attribute DM [1]; ▶ tailoring query-based preference elicitation [8]; ▶ weaker existence conditions than finiteness of state and action sets' volumes (13), (24); ▶ inspection how much the results challenge the claim that the quest for absolute optimality is unrealistic [42]

or what is a proper level of inattention [43]; ▶ specific application cases like [35], etc.

References

[1] C.R.B. Azevedo and F.J. von Zuben. Learning to anticipate flexible choices in multiple criteria decision-making under uncertainty. *IEEE Tran. on Cyb.*, 46(3):778–791, 2016.

[2] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, N.Y., 1978.

[3] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[4] J. Bernardo. Expected information as expected utility. *The An. of Stat.*, 7:686–690, 1979.

[5] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Sci., 2001.

[6] O. Besbes, Y. Gur, and A. Zeevi. Optimal exploration – exploitation in a multi-armed bandit problem with nonstationary rewards. *Stoch. Syst.*, 9(4):319–337, 2019.

[7] T. Bohlin. *Interactive System Identification: Prospects and Pitfalls*. Springer, 1991.

[8] C. Boutilier. A POMDP formulation of preference elicitation problems. In *Proc. of the 18th National Conf. on AI, AAAI-2002*, pages 239–246, Edmonton, AB, 2002.

[9] J. Branke and et al. Efficient pairwise preference elicitation allowing for indifference. *Computers & Oper. Res.*, 88(Suppl. C):175 – 186, 2017.

[10] L. Chen and P. Pu. Survey of preference elicitation methods. Technical Report IC/2004/67, HCI Group Ecole Politechnique Federale de Lausanne, Switzerland, 2004.

[11] P. Daee, T. Peltola, M. Soare, and S. Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Mach. Learn.*, 106:1599–1620, 2017.

[12] J. Drummond and C. Boutilier. Preference elicitation and interview minimization in stable matchings. In *Proc. of 28th AAAI Conf. on AI*, pages 645 – 653, 2014.

[13] J.S. Dyer, P.C. Fishburn, and et al. Multiple criteria decision making, multiattribute utility theory: The next ten years. *Man. Sci.*, 38(5):645–654, 1992.

[14] E.A. Feinberg and A. Shwartz. *Handbook of Markov Decision Processes: Methods & Applications*. Kluwer, 2002.

[15] A.A. Feldbaum. Theory of dual control. *Autom. Remote Control*, 22:3–19, 1961.

[16] P.C. Fishburn. Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty*, 4:113–134, 1991.

[17] D. Gagliardi and G. Russo. On a probabilistic approach to synthesize control policies from example datasets. *Automatica*, 137:110121, 2022.

[18] P. Guan, M. Raginsky, and R.M. Willett. Online Markov decision processes with Kullback Leibler control cost. *IEEE Trans. on AC*, 59(6):1423–1438, 2014.

[19] T.V. Guy, S. Fakhimi Derakhshan, and J. Štěch. Lazy fully probabilistic design: Application potential. In F. Belardinelli, editor, *Multi-Agent Systems & Agreement Technologies*, 2018.

[20] M. Kárný. Axiomatisation of fully probabilistic design revisited. *SCL*, 141:104719, 2020.

[21] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, UK, 2006.

[22] M. Kárný and T.V. Guy. Fully probabilistic control design. *SCL*, 55:259–265, 2006.

[23] M. Kárný and T.V. Guy. On support of imperfect Bayesian participants. In T.V. Guy and et al, editors, *Decision Making with Imperfect Decision Makers*, volume 28, pages 29–56. Springer Int. Syst. Ref. Lib., 2012.

[24] M. Kárný and T.V. Guy. Preference elicitation within framework of fully probabilistic design of decision strategies. In *IFAC Workshop ALCOS*, volume 52, pages 239–244, 2019.

[25] M. Kárný and R. Herzallah. Scalable harmonization of complex networks with local adaptive controllers. *IEEE Trans. on SMC: Systems*, 47(3):394–404, 2017.

[26] M. Kárný and M. Ruman. Preference elicitation for Markov decision processes in fully probabilistic design set up. *Annals of Operation Research*, 2022. submitted.

[27] M. Kárný and T. Sováková. Agent's feedback in preference elicitation. In *20th Int. Conf. on Ubiquitous Computing and Communications, IUCC*, pages 421–429, 2021.

[28] R. Kulhavý and M. B. Zarrop. On a general concept of forgetting. *Int. J. of Control*, 58(4):905–924, 1993.

[29] S. Kullback and R. Leibler. On information and sufficiency. *Ann Math Stat*, 22:79–87, 1951.

[30] G.H. Kwak, L. Ling, and P. Hui. Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. *PLOS ONE*, 16(5):1–15, 05 2021.

[31] I.D. Landau. A survey of model reference adaptive techniques. *Automatica*, 10(4):353 – 379, 1974.

[32] A. O'Hagan. *Uncertain Judgement: Eliciting Experts' Probabilities*. J. Wiley, 2006.

[33] R. Padhan and K.P. Prabheesh. The economics of COVID-19 pandemic: A survey. *Economic Analysis and Policy*, 70:220–237, 2021.

[34] D. Palenicek. A survey on constraining policy updates using the KL divergence. In B. Belousov and et al, editors, *Reinforcement Learning Algorithms: Analysis and Applications*. Springer, Cham, 2021.

[35] A. Perrault and C. Boutilier. Experiential preference elicitation for autonomous heating & cooling systems. In *Proc. of the 18th Int. Conf. AAMAS*, pages 431–439, 2019.

[36] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends & Progress in System Identification*, pages 239–304. Perg. Press, 1981.

[37] G. Pigozzi, , A. Tsoukiàs, , and P. Viappiani. Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 77(3):361–401, 2016.

[38] A. Quinn, M. Kárný, and T.V. Guy. Fully probabilistic design of hierarchical Bayesian models. *Inf. Sci.*, 369:532–547, 2016.

[39] M.M. Rao. *Measure Theory and Integration*. J. Wiley, 1987.

[40] L.J. Savage. *Foundations of Statistics*. Wiley, 1954.

[41] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. *IEEE Tran. on Inf. Th.*, 26(1):26–37, 1980.

[42] H.A. Simon. *Models of Bounded Rationality*. MacMillan, 1997.

[43] C. A. Sims. Rational inattention: Beyond the linear-quadratic case. *The Am. Econ. Rev.*, 96(2):158–163, 2006.

[44] T. Sováková and M. Kárný. Experiments with the user's feedback in preference elicitation. In *AIBAI Workshop, Proc. CEUR Workshop*. Udine, 2022.

[45] E. Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf and et al, editors, *Adv. in Neur. Inf. Proc.*, pages 1369 – 1376. MIT Press, 2006.

[46] J. Wallenius, J.S. Dyer, P.C. Fishburn, and et al. Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Manag. Sci.*, 54(7):1336–1349, 2008.