Fully Probabilistic Design of Strategies with Estimator ⋆

Miroslav Kárný

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

Abstract

The axiomatic fully probabilistic design (FDP) of decision strategies strictly extends Bayesian decision making (DM) theory. FPD also models the closed decision loop by a joint probability density (pd) of all inspected random variables, referred as behaviour. FPD expresses DM aims via an ideal pd of behaviours, unlike the usual DM. Its optimal strategy minimises Kullback-Leibler divergence (KLD) of the joint, strategy-dependent, pd of behaviours to its ideal twin. A range of FPD results confirmed its theoretical and practical strength. Curiously, no guide exists how to select a specific ideal pd for an estimator design. The paper offers it. It advocates the use of the closed-loop state notion and generalises dynamic programming so that FPD is its special case. Primarily, it provides an explorative optimised feedback that "naturally" diminishes exploration (gained in learning) as the learning progresses.

Key words: Bayes methods, Closed loop systems, Decision making, Dynamic programming, Estimation

1 Introduction

On the paper context and main result This brief paper focuses on a technical problem related to a prescriptive theory of dynamic decision making (DM). The theory is dubbed fully probabilistic design (FPD) of decision strategies [1]. It generalises methodologies developed in connection with (adaptive) control theory [1,4] and Markov decision processes [32]. Since its initial publication [13], it was broadly elaborated [17], axiomatised [14], applied [16,33] and used for supporting decision makers [9,15,18,42].

The paper deals with the evergreen known as dual control [7,22,28] or exploration-exploitation dichotomy [5]. It concerns the balance of random explorative actions, supporting parameter estimation, with actions moving the closed control loop to the desired state [2]. The main contribution of the paper is an optimised feedback that "naturally" diminishes exploration (gained in learning) as the learning progresses.

On the addressed technical problem Any estimation serves to decision making seen as the aim-focused selection and use of actions. The agent — the decision maker or the action selector, referred as "it" — acts under uncertainty. The inspected agent uses FPD. FPD models the closed-loop behaviour by the joint probability density (pd). The behaviour consists of all considered uncertain variables. The inspected estimation arises when the behaviour includes a parameter unknown to the agent. Its adopted handling as random variable coincides with bayesianism [3].

The FPD-optimal strategy minimises Kullback-Leibler divergence (KLD) [24] of the behaviours' pd to its ideal, DM-aims expressing, twin. The estimation has the parameter estimates as (a part of) agent's actions. The wish to obtain good estimates of the unknown parameter is the generic agent's aim. The key question is: *What ideal pd expresses this wish?* A universal conversion of a usual loss into the ideal pd exists, Prop. 3. in [14]. It often violates the dictum [20]: *Select an ambitious but reachable ideal pd!* Our solution meets this dictum and leads to the mentioned main result.

Layout Sec. 2 recalls FPD, embeds it into a slightly generalised dynamic programming and advocates the use of *closed-loop* states. Core Sec. 3 solves FPD with an estimator. It proposes the relevant ideal pd and finds the FPD-optimal estimator. Sec. 4 summarises properties of the proposed strategy and outlines open problems.

Notation $\{x\}$ marks the set of xs defined if needed. Sanserif fonts denote mappings. The superscripts $^{\mathsf{i}}$, $^{\mathsf{o}}$ refer to the ideal pd and optimality, respectively. The symbol $:=$ defines by assigning; $\propto$ is proportionality; $\tilde{}$ marks interim objects. The time subscript $t \in \{t\}$ of a function f_t on $\{x\}$ drops if the function argument has it, $\mathsf{f}(x_t) := \mathsf{f}_t(x_t)$. The text prefers mnemonic identifiers.

⋆ This paper was not presented at any IFAC meeting.

Email address: school@utia.cas.cz (Miroslav Kárný).

[1] It has overlaps with KL control [6,8,12,30,39], with minimum relative entropy principle [29,35], softmax, etc.

[2] The exploration counteracts the positive probability of non-optimality connected with certainty-equivalence-based strategies, see e.g. [25] the example on p. 347 and Theorem 7.1

2 Fully Probabilistic Design

FPD deals with the closed DM loop. An agent and its environment form it. The agent's actions $a_t \in \{a\} \neq \emptyset$, at time moments tagged by $t \in \{t\} := \{1, \ldots, h\}$, $h < \infty$, influence transitions of states $s_{t-1} \in \{s\} \neq \emptyset$ to states $s_t \in \{s\}$. The inspected transition model depends on an unknown, time-invariant, parameter $p \in \{p\}$. The closed-loop states $(s_t)_{t\in\{t\}}$ are gradually observed and constructed. A fixed, known initial state s_0 implicitly conditions all used pds. The case with internal states is left aside to keep the paper simple. The fact that s_t is *closed-loop* state also simplifies the text. The state s_t includes the value w_t of the sufficient statistic w_t and has the structure

$$s_t := (o_t, w_t) := (\text{observation of the environment, value of the sufficient statistic } \mathsf{w}_t),\ w_t := \mathsf{w}(o_t, a_t, w_{t-1}). \tag{1}$$

The data $d := (d_t)_{t\in\{t\}} := (s_t, a_t)_{t\in\{t\}}$ consists of states, $s_t \in \{s\}$, and actions, $a_t \in \{a\}$, up to the horizon h. The data, $d \in \{d\}$, and the unknown parameter, $p \in \{p\}$, form the random behaviour $b := (d, p) \in \{b\} := (\{d\}, \{p\})$. The agent opts its actions via a randomised causal strategy [3]

$$\mathsf{s} \in \{\mathsf{s}\} := \left\{ \mathsf{s}(b) := \prod_{t\in\{t\}} \mathsf{r}(a_t|s_{t-1}) = \mathsf{s}(d) \right\}. \tag{2}$$

Pds $(\mathsf{r}(a_t|s_{t-1}))_{t\in\{t\}}$ describe the decision rules of the strategy s. They meet natural conditions of control (NCC) [31]

$$\mathsf{r}(a_t|s_{t-1}) = \mathsf{r}(a_t|s_{t-1}, p), \quad a_t \in \{a\},\ s_{t-1} \in \{s\},\ p \in \{p\}. \tag{3}$$

NCC express the adopted assumption that the parameter is unknown to the agent. The s-dependent joint pd $\mathsf{c}^{\mathsf{s}}(b) = \mathsf{c}^{\mathsf{s}}(d, p)$ fully models random behaviours. The chain rule for pds [31], the state notion and NCC (3) factorise this closed-loop model of the behaviours, $b \in \{b\} = \{b = (d, p)\}$,

$$\mathsf{c}^{\mathsf{s}}(b) := \mathsf{c}^{\mathsf{s}}(d|p)\mathsf{p}(p) := \prod_{t\in\{t\}} \mathsf{m}(s_t|a_t, s_{t-1}, p)\mathsf{r}(a_t|s_{t-1})\mathsf{p}(p) := \mathsf{m}(d|p)\mathsf{s}(d)\mathsf{p}(p). \tag{4}$$

The parametric models $\mathsf{m}(d|p)$, $p \in \{p\}$,

$$\mathsf{m}(d|p) \in \{\mathsf{m}(d|p)\} := \left\{ \mathsf{m}(d|p) := \prod_{t\in\{t\}} \mathsf{m}(s_t|a_t,\ s_{t-1}, p),\ d = (s_t, a_t)_{t\in\{t\}} \right\} \tag{5}$$

serve the agent to describe its environment. The factors $\mathsf{m}(s_t|a_t,\ s_{t-1}, p)$ are pds of $s_t \in \{s\}$ conditioned on $a_t \in \{a\}$. They model the state transition from $s_{t-1} \in \{s\}$. They do it for each $p \in \{p\}$ and all $t \in \{t\}$. The strategy s influences the agent's environment solely via the actions it generates. This explains the dropped superscript $^{\mathsf{s}}$ in (4). The strategy s has no influence on the prior pd $\mathsf{p}(p)$ modelling the unknown parameter $p \in \{p\}$. It drops $^{\mathsf{s}}$, too.

The value of the prior pd $\mathsf{p}(p)$, $p \in \{p\}$, expresses agent's belief that $\mathsf{m}(d|p)$ is the best projection of the objective environment description on the set of parametric models $\{\mathsf{m}(d|p),\ d \in \{d\}\}_{p\in\{p\}}$, [15]. Under NCC (3), the value of the prior pd $\mathsf{p}(p)$ also quantifies the belief that $\mathsf{m}(d|p)\mathsf{s}(d)$ is the best projection of the objective closed-loop model.

FPD quantifies agent's aims and constraints by an ideal closed-loop model c^{i}. It is a joint pd $\mathsf{c}^{\mathsf{i}}(b)$ of ideally distributed behaviours $b \in \{b\}$. The agent chooses the ideal pd $\mathsf{c}^{\mathsf{i}}(b)$. It assigns high values to preferred behaviours b, small values to unwanted ones and zero to forbidden bs. The pd $\mathsf{c}^{\mathsf{i}}(b)$ factorises as $\mathsf{c}^{\mathsf{s}}(b)$, $b \in \{b\}$,

$$\begin{aligned}\mathsf{c}^{\mathsf{i}}(b) &:= \mathsf{c}^{\mathsf{i}}(d|p)\mathsf{p}^{\mathsf{i}}(p) := \\ &:= \prod_{t\in\{t\}} \mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1}, p)\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}, p)\mathsf{p}^{\mathsf{i}}(p) \\ &:= \mathsf{m}^{\mathsf{i}}(d|p)\mathsf{s}^{\mathsf{i}}(d|p)\mathsf{p}(p).\end{aligned} \tag{6}$$

The factors of the ideal parametric environment model, $\mathsf{m}^{\mathsf{i}}(s_t|a_t,\ s_{t-1}, p)$, $s_t, s_{t-1} \in \{s\}$, $a_t \in \{a\}$, $t \in \{t\}$, describe the *desired* state transitions. The ideal strategy $\mathsf{s}^{\mathsf{i}}(d|p) := \prod_{t\in\{t\}} \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}, p)$ consists of the preferred decision rules. Unlike the optimised decision rules (3), they may depend on the unknown $p \in \{p\}$ as they model agent's wishes. The strategy s has no influence on the prior pd, which is left-to-its-fate [16] by setting $\mathsf{p}^{\mathsf{i}}(p) := \mathsf{p}(p)$. FPD axiomatics [14] implies that the FPD-optimal strategy $\mathsf{s}^{\mathsf{o}} \in \{\mathsf{s}\}$ minimises the KLD $\mathsf{D}(\mathsf{c}^{\mathsf{s}}||\mathsf{c}^{\mathsf{i}})$ of c^{s} to c^{i}

$$\begin{aligned}\mathsf{s}^{\mathsf{o}} &\in \operatorname{Arg}\min_{\mathsf{s}\in\{\mathsf{s}\}} \mathsf{D}(\mathsf{c}^{\mathsf{s}}||\mathsf{c}^{\mathsf{i}}) \\ &:= \operatorname{Arg}\min_{\mathsf{s}\in\{\mathsf{s}\}} \int_{\{b\}} \mathsf{c}^{\mathsf{s}}(b) \ln\left(\frac{\mathsf{c}^{\mathsf{s}}(b)}{\mathsf{c}^{\mathsf{i}}(b)}\right) \mathrm{d}b.\end{aligned} \tag{7}$$

Remark 1 (On Knowledge Accumulation)

✓ *Having a parametric model $\tilde{\mathsf{m}}(o_t|a_t, s_{t-1}, p)$ relating the observations $o_t \in \{o\}$ (1) to the actions $a_t \in \{a\}$, to the past states $s_{t-1} \in \{s\}$, and to the parameters $p \in \{p\}$, Bayes' rule [31] gradually updates the prior pd $\mathsf{p}(p) := \mathsf{p}(p|s_0)$ into the posterior pds $\mathsf{p}(p|s_t)$*

$$\begin{aligned}\mathsf{p}(p|s_t) &= \frac{\tilde{\mathsf{m}}(o_t|a_t, s_{t-1}, p)\mathsf{p}(p|s_{t-1})}{\int_{\{p\}} \tilde{\mathsf{m}}(o_t|a_t, s_{t-1}, p)\mathsf{p}(p|s_{t-1})\mathrm{d}p} \\ &\propto \tilde{\mathsf{m}}(o_t|a_t, s_{t-1}, p)\mathsf{p}(p|s_{t-1}).\end{aligned} \tag{8}$$

This knowledge accumulation is correct under NCC (3) that have the equivalent expression

$$\mathsf{p}(p|a_t, s_{t-1}) = \mathsf{p}(p|s_{t-1}) \Leftrightarrow \mathsf{r}(a_t|s_{t-1}) = \mathsf{r}(a_t|s_{t-1}, p) \quad a_t \in \{a\},\ s_{t-1} \in \{s\},\ p \in \{p\}. \tag{9}$$

[3] A decision function [40], act [34], policy [32], etc.

✓ *Bayes' rule (8) updates the value w_t of the sufficient statistic w_t that is a part of the state (1). Thus, the parametric observation model $\tilde{\mathsf{m}}_t$ and Bayes' rule describe the parametric model m_t (5).*

The full tth step of the knowledge collection starts with $(\mathsf{p}(p|s_{t-1}), s_{t-1})$, uses the action $a_t \sim \mathsf{r}(a_t|s_{t-1})$ and the observation o_t made on agent's environment in Bayes' rule (8) giving the statistic value $w_t = \mathsf{w}(o_t, a_t, s_{t-1})$. This completes the state $s_t = (o_t, w_t)$, see (1).

✓ *The use of the closed-loop state:*
- ⋆ *opens a way to a better estimation as the options like forgetting [23] or a trust weight in Bayes' rule [21] just extend the actions;*
- ⋆ *unifies inference with tasks influencing the environment;*
- ⋆ *allows to model human in DM cycle [38] and to respect agent's attitudes and emotions [11];*
- ⋆ *gives a specific view on (ir)rationality of DM [19];*
- ⋆ *fits conceptually to dual control [7] and to exploration-exploitation dichotomy [5,37].*

3 FPD with Estimator

Part 3.1 constructs the FPD-optimal strategy. It relies on a slight extension of stochastic dynamic programming [4] that minimises [4] the strategy-dependent expectation $\mathsf{E}^{\mathsf{s}}[\mathsf{L}^{\mathsf{s}}]$ of the s-*dependent* additive loss $\mathsf{L}^{\mathsf{s}}(b)$, $\{b\}$. The optimal strategy s^{o} minimises the expectation of the loss

$$\mathsf{L}^{\mathsf{s}}(b) := \sum_{t\in\{t\}} \mathsf{l}^{\mathsf{r}_t}(s_{t-1}), \quad \mathsf{l}^{\mathsf{r}_t} : \{s\} \to (-\infty, \infty)$$

$$\mathsf{s}^{\mathsf{o}} \in \operatorname{Arg}\min_{\{\mathsf{s}\}} \mathsf{E}^{\mathsf{s}}[\mathsf{L}^{\mathsf{s}}] := \operatorname{Arg}\min_{\{\mathsf{s}\}} \int_{\{b\}} \mathsf{L}^{\mathsf{s}}(b)\mathsf{c}^{\mathsf{s}}(b)\mathrm{d}b. \qquad (10)$$

The dependence of the loss L^{s} on s makes the optimised functional (10) non-linear in the strategy $\mathsf{s} \in \{\mathsf{s}\}$. This leads to the non-standard optimisation giving the optimal *randomised* strategy. FPD is a special case of (10). It has an explicit form of the FPD-optimal strategy s^{o}.

Part 3.2 finds the FPD-optimal strategy s^{o} with the optimal estimator by using the results of Part 3.1.

3.1 Dynamic Programming on Strategy-Dependent Loss

The next proposition prepares the solution of (10).

Proposition 1 (Static Design) *Let us search for the optimal strategy s^{o} consisting of single decision rule r^{o}*

$$\mathsf{s}^{\mathsf{o}} := (\mathsf{r}^{\mathsf{o}}(a|s))_{a\in\{a\},s\in\{s\}} \in \operatorname{Arg}\min_{\{\mathsf{s}\}} \mathsf{E}^{\mathsf{s}}[\mathsf{L}^{\mathsf{s}}] := \operatorname{Arg}\min_{\{\mathsf{r}\}} \int_{\{s\}} \mathsf{l}^{\mathsf{r}}(s)\mu(s)\mathrm{d}s,$$

where the marginal pd μ on $\{s\}$ does not depend on r. Let $\mathsf{l}^{\mathsf{r}}(s) : \{\mathsf{r}\} := \{pds\ \mathsf{r}(a|s)\ on\ \{a\}\} \to (-\infty, \infty)$ have a minimiser $\mathsf{r}^\star$ for μ-almost all $s \in \{s\}$.

Then, $\mathsf{s}^{\mathsf{o}} = \mathsf{r}^{\mathsf{o}} = \mathsf{r}^\star$. Symbolically,

$$(\mathsf{r}^{\mathsf{o}}(a|s))_{a\in\{a\},s\in\{s\}} \in \operatorname{Arg}\min_{\{\mathsf{r}(a|s)\}} \int_{\{s\}} \mathsf{l}^{\mathsf{r}}(s)\mu(s)\mathrm{d}s = \operatorname{Arg}\min_{\{\mathsf{r}(a|s)\}} \mathsf{l}^{\mathsf{r}}(s).$$

The minimum is $\int_{\{s\}} \mathsf{l}^{\mathsf{r}^{\mathsf{o}}}(s)\mu(s)\mathrm{d}s$.

Proof By definition, $\mathsf{l}^{\mathsf{r}^\star}(s) \le \mathsf{l}^{\mathsf{r}}(s)$ for any $\mathsf{r} \in \{\mathsf{r}\}$ and for μ-almost all $s \in \{s\}$. Multiplication of this inequality by the positive values of $\mu(s)$ and integration over $\{s\}$ preserve it. This proofs the claim. □

The solution of (10) uses the optimal value functions, [4], $\mathsf{v}_{t-1} : \{s\} \to (-\infty, \infty)$, $t \in \{t\}$,

$$\mathsf{v}(s_{t-1}) := \min_{\{(\mathsf{r}_\tau)_{\tau\ge t}\}} \mathsf{E}^{\mathsf{s}}\Big[\sum_{\tau=t}^{h} \mathsf{l}^{\mathsf{r}_\tau}(s_{\tau-1})\Big|s_{t-1}\Big]. \qquad (11)$$

Proposition 2 (Dynamic Programming) *The backward, $t = h, \dots, 1$, functional recursion, initiated by $\mathsf{v}(s_h) = 0$, evolves the optimal value functions $(\mathsf{v}_t)_{t\in\{t\}}$, $s_{t-1} \in \{s\}$,*

$$\mathsf{v}(s_{t-1}) = \min_{\{\mathsf{r}_t\}} \big\{\mathsf{l}^{\mathsf{r}_t}(s_{t-1}) + \mathsf{E}^{\mathsf{s}}[\mathsf{v}(s_t)|s_{t-1}]\big\}. \qquad (12)$$

The minimising arguments $(\mathsf{r}^{\mathsf{o}}_t)_{t\in\{t\}}$ in (12) form the optimal strategy s^{o} (10). The minimum of (10) is $\mathsf{v}(s_0)$.

Proof It exploits the "backward" induction. For $t = h$ and $\mathsf{v}(s_h) = 0$, $\forall s_h \in \{s\}$, (12) holds. In a generic step,

$$\mathsf{v}(s_{t-1}) \overset{(11)}{=} \min_{\{(\mathsf{r}_\tau)_{\tau\ge t}\}} \mathsf{E}^{\mathsf{s}}\Big[\sum_{\tau=t}^{h} \mathsf{l}^{\mathsf{r}_\tau}(s_{\tau-1})\Big|s_{t-1}\Big] = \min_{\{\mathsf{r}_t\}} \Bigg\{ \min_{\{(\mathsf{r}_\tau)_{\tau\ge t+1}\}} \Big\{\mathsf{l}^{\mathsf{r}_t}(s_{t-1}) + \mathsf{E}^{\mathsf{s}}\Big[\mathsf{E}^{\mathsf{s}}\Big[\sum_{\tau=t+1}^{h} \mathsf{l}^{\mathsf{r}_\tau}(s_{\tau-1})\Big|s_t\Big]\Big|s_{t-1}\Big]\Big\}\Bigg\}$$
$$\overset{(11),\,Prop.\,1}{=} \min_{\{\mathsf{r}_t\}} \big\{\mathsf{l}^{\mathsf{r}_t}(s_{t-1}) + \mathsf{E}^{\mathsf{s}}[\mathsf{v}(s_t)|s_{t-1}]\big\}.$$

The way to the final formula exploits: a) $\mathsf{l}^{\mathsf{r}_t}(s_{t-1})$ behaves as a constant with respect to conditional expectation $\mathsf{E}^{\mathsf{s}}[\bullet|s_{t-1}]$; b) the chain rule for expectation $\mathsf{E}^{\mathsf{s}}[\bullet|\diamondsuit] = \mathsf{E}^{\mathsf{s}}\big[\mathsf{E}^{\mathsf{s}}[\bullet|\star,\diamondsuit]|\diamondsuit\big]$, valid for arbitrary $\bullet, \star, \diamondsuit$; c) equality $\mathsf{E}^{\mathsf{s}}[\bullet|s_t, s_{t-1}] = \mathsf{E}^{\mathsf{s}}[\bullet|s_t]$ holds as s_t is the state; d) both $\mathsf{l}^{\mathsf{r}_t}$ and $\mathsf{c}^{\mathsf{s}}(s_t|s_{t-1})$ (used in the outer expectation) are independent of the rules r_τ, $\tau > t$. The final claim of the

[4] The assumed generic existence of minima keeps the text simple.

proposition follows from the identity $\min_{\{\mathsf{s}\}} \mathsf{E}^{\mathsf{s}}[\mathsf{L}^{\mathsf{s}}] = \mathsf{v}(s_0)$ for the given s_0, cf. (11), (12). □

The next proposition provides the FPD-optimal strategy s^{o}.

Proposition 3 (FPD with an Unknown Parameter) *The backward, $t = h, h-1, \ldots, 1$, functional recursion for $\mathsf{n}(s_{t-1}) \in [0,1]$, $\mathsf{d}(a_t, s_{t-1}) \geq 0$, $s_{t-1} \in \{s\}$, $a_t \in \{a\}$,*

$$\begin{aligned}\mathsf{d}(a_t, s_{t-1}) &:= \int_{\{p\}} \mathsf{p}(p|s_{t-1}) \int_{\{s\}} \mathsf{m}(s_t|a_t, s_{t-1}, p) \qquad (13)\\ &\quad \times \ln\Big(\frac{\mathsf{m}(s_t|a_t, s_{t-1}, p)}{\mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1}, p)\mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}, p)\mathsf{n}(s_t)}\Big)\mathrm{d}s_t\mathrm{d}p\\ \mathsf{n}(s_{t-1}) &:= \int_{\{a\}} \exp[-\mathsf{d}(a_t, s_{t-1})]\mathrm{d}a_t, \quad \mathsf{n}(s_h) := 1,\end{aligned}$$

gives the minimum of (7) $\mathsf{v}(s_0) := -\ln(\mathsf{n}(s_0)) = \mathsf{D}(\mathsf{c}^{\mathsf{s}^{\mathsf{o}}}||\mathsf{c}^{\mathsf{i}})$. The rules $\mathsf{r}^{\mathsf{o}}_t$ of the strategy (7) in (2) are

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

Proof The definition of KLD $\mathsf{D}(\mathsf{c}^{\mathsf{s}}||\mathsf{c}^{\mathsf{i}})$ (7), the factorised forms of the involved pds (4), (6) and NCC (9) imply that KLD is of the additive form (10) $\mathsf{D}(\mathsf{c}^{\mathsf{s}}||\mathsf{c}^{\mathsf{i}}) =$

$$\begin{aligned}\mathsf{E}^{\mathsf{s}}[\mathsf{L}^{\mathsf{s}}] = \mathsf{E}^{\mathsf{s}}\Bigg[\sum_{t\in\{t\}} \int_{(\{s\},\{a\},\{p\})} &\mathsf{m}_t(\tilde{s}|a, s, p)\mathsf{r}_t(a|s)\mathsf{p}_{t-1}(p|s)\\ &\times \ln\left(\frac{\mathsf{m}_t(\tilde{s}|a, s, p)\mathsf{r}_t(a|s)}{\mathsf{m}^{\mathsf{i}}_t(\tilde{s}|a, s, p)\mathsf{r}^{\mathsf{i}}_t(a|s,p)}\right)\mathrm{d}(\tilde{s}, a, p)\Bigg].\end{aligned}$$

The summands, denoted $\mathsf{l}^{\mathsf{r}_t}(s)$, are obviously DM-rule dependent. The related value function $\mathsf{v}(s_{t-1})$ is non-negative as it is the sum of conditional KLD's. Thus, the transformation $\mathsf{v}(s_{t-1}) := -\ln(\mathsf{n}(s_{t-1}))$ implies that the function $\mathsf{n}(s_{t-1}) \in [0,1]$, $\mathsf{n}(s_h) = 1 \Leftrightarrow \mathsf{v}(s_h) = 0$. The minimised right-hand side of (12) reads

$$\begin{aligned}\mathsf{l}^{\mathsf{r}_t}(s) + \mathsf{E}^{\mathsf{s}}[\mathsf{v}_t(\tilde{s})|s] = \int_{\{a\}} \mathsf{r}_t(a|s)\Bigg[&\ln(\mathsf{r}_t(a|s))\\ &+ \int_{\{p\}} \mathsf{p}_{t-1}(p|s) \int_{\{s\}} \mathsf{m}_t(\tilde{s}|a, s, p)\\ &\times \ln\left(\frac{\mathsf{m}_t(\tilde{s}|a, s, p)}{\mathsf{m}^{\mathsf{i}}_t(\tilde{s}|a, s, p)\mathsf{r}^{\mathsf{i}}_t(a|s, p)\mathsf{n}_t(\tilde{s})}\right)\mathrm{d}\tilde{s}\mathrm{d}p\Bigg]\mathrm{d}a\end{aligned}$$

$$\begin{aligned}&= \int_{\{a\}} \mathsf{r}_t(a|s) \ln\Bigg(\frac{\mathsf{r}_t(a|s)}{\underbrace{\frac{\exp[-\mathsf{d}_t(a,s)]}{\int_{\{a\}}\exp[-\mathsf{d}_t(\tilde{a},s)]\mathrm{d}\tilde{a}}}_{:=\mathsf{r}^{\star}_t(a|s)}}\Bigg)\mathrm{d}a\\ &\quad - \ln\Big(\int_{\{a\}} \exp[-\mathsf{d}_t(\tilde{a}, s)]\mathrm{d}\tilde{a}\Big).\end{aligned}$$

The 1^{st} summand after the last equality is the KLD (conditioned on s) of the decision rules' pair. It reaches its smallest zero value for $\mathsf{r}_t = \mathsf{r}^{\star}_t \overset{Prop.2}{=} \mathsf{r}^{\mathsf{o}}_t$. The 2^{nd} summand, $-\ln(\text{normaliser})$ is the reached minimum. □

Remark 2 (On Generality and Constraints)

- ✓ *The paper [17] deals with the general case containing internal states (time-varying unknown "parameters") but unlike Prop. 3 does not consider the admissible dependence of r^{i} (6) on them.*
- ✓ *The support of $\mathsf{m}^{\mathsf{i}}_t\mathsf{r}^{\mathsf{i}}_t$ includes the support of $\mathsf{m}_t\mathsf{r}^{\mathsf{o}}_t$ as, otherwise infinite, $\mathsf{d}(a_t, s_{t-1})$ (13) guarantees this. This shows that the ideal pd provides also hard constraints on the optimal closed-loop model $\mathsf{c}^{\mathsf{s}^{\mathsf{o}}}$.*

3.2 FPD with Estimation

DM with estimation deals with an *action a_t that has a parameter estimate $\hat{p}_t \in \{p\}$ as its part*. It means

$$a_t := (\alpha_t, \hat{p}_t) \in \{a\} := (\{\alpha\}, \{p\}). \qquad (14)$$

The possibly empty part $\alpha_t \in \{\alpha\}$ serves to other DM aim than the estimation. This structures the used pds (the explanations of the choices follow their presentation in (15))

$$\begin{aligned}\text{(a): } \mathsf{m}(s_t|a_t, s_{t-1}, p) &\overset{(14)}{=} \mathsf{m}(s_t|\alpha_t, \hat{p}_t, s_{t-1}, p)\\ &:= \mathsf{m}(s_t|\alpha_t, s_{t-1}, p) \qquad (15)\\ \text{(b): } \mathsf{m}^{\mathsf{i}}(s_t|a_t, s_{t-1}, p) &\overset{(14)}{=} \mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, \hat{p}_t, s_{t-1}, p)\\ &:= \mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, \hat{p}_t)\\ \mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, \hat{p}_t) &:= \mathsf{m}(s_t|\alpha_t, s_{t-1}, \hat{p}_t) \text{ if } \alpha_t \text{ is free}\\ \text{(c): } \quad \mathsf{r}(a_t|s_{t-1}) &\overset{(14)}{=} \mathsf{r}(\alpha_t|\hat{p}_t, s_{t-1})\mathsf{r}(\hat{p}_t|s_{t-1})\\ \mathsf{r}^{\mathsf{i}}(a_t|s_{t-1}, p) &\overset{(14)}{=} \mathsf{r}^{\mathsf{i}}(\alpha_t|\hat{p}_t, s_{t-1}, p)\,\mathsf{r}^{\mathsf{i}}(\hat{p}_t|s_{t-1}, p)\\ &:= \mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t) \times \mathsf{p}(\hat{p}_t|s_{t-1}).\end{aligned}$$

The initial equalities follow from (14). The last equalities in (15) quantify conditions and aims of DM with estimation. They respectively express:

(a): independence of the state of the parameter estimate: the environment does not care about agent's estimate and it also does not influence the sufficient statistic;

(b): *the wish to choose $\hat{p}_t$ that ideally replaces the unknown p* in $\mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, p)$ expressing the α-driving aim; $\mathsf{m}^{\mathsf{i}} := \mathsf{m}$ when α_t is not optimised (left-to-its-fate [16]);

(c): the estimate $\hat{p}_t$ *ideally replaces the unknown* $p \in \{p\}$ also in $\mathsf{r}^{\mathsf{i}}_t$ and in $\mathsf{p}(p|s_{t-1})$ collecting the knowledge on p.

Proposition 4 (FPD with Estimator) *The FPD-optimal strategy with an estimator, given by options (14), (15), results from the backward, $t = h, \ldots, 1$, recursion for $\mathsf{n}(s_{t-1}) \in [0,1]$ and $\Delta(\alpha_t, \hat{p}_t, s_{t-1}) \geq 0$, $s_{t-1} \in \{s\}$, $\alpha_t \in \{\alpha\}$, $\hat{p}_t \in \{p\}$, $t \in \{t\}$, with $\mathsf{n}(s_h) := 1$,*

$$\Delta(\alpha_t, \hat{p}_t, s_{t-1}) := \int_{\{p\}} \mathsf{p}(p|s_{t-1}) \tag{16}$$
$$\times \int_{\{s\}} \mathsf{m}(s_t|\alpha_t, s_{t-1}, p) \ln\Big[\frac{\mathsf{m}(s_t|\alpha_t, s_{t-1}, p)}{\mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, \hat{p}_t)\mathsf{n}(s_t)}\Big] \mathrm{d}s_t \mathrm{d}p$$
$$\mathsf{n}(s_{t-1}) := \int_{(\{p\},\{\alpha\})} \mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1}) \times \exp[-\Delta(\alpha_t, \hat{p}_t, s_{t-1})] \mathrm{d}(\hat{p}_t, \alpha_t).$$

It gives the minimum $\mathsf{v}(s_0) := -\ln(\mathsf{n}(s_0)) = \mathsf{D}(\mathsf{c}^{\mathsf{r}^{\mathsf{o}}}||\mathsf{c}^{\mathsf{i}})$. *The rules $\mathsf{r}^{\mathsf{o}}_t$ of the optimal strategy are*

$$\mathsf{r}^{\mathsf{o}}(\alpha_t, \hat{p}_t|s_{t-1}) \propto \mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1}) \exp[-\Delta(\alpha_t, \hat{p}_t, s_{t-1})]. \tag{17}$$

Proof It uses Prop. 3 specialised by (14), (15). The exponent $\mathsf{d}(a_t, s_{t-1}) \overset{(14)}{=} \mathsf{d}(\alpha_t, \hat{p}_t, s_{t-1})$ in (13) under (15) gets the form, with $\mathsf{n}(s_h) = 1$ and $\Delta(\alpha_t, \hat{p}_t, s_{t-1})$ defined in (16),

$$\mathsf{d}(\alpha_t, \hat{p}_t, s_{t-1}) := \int_{\{p\}} \mathsf{p}(p|s_{t-1}) \int_{\{s\}} \mathsf{m}(s_t|\alpha_t, s_{t-1}, p) \times \ln\Big(\frac{\mathsf{m}(s_t|\alpha_t, s_{t-1}, p)}{\mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, \hat{p}_t)\mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1})\mathsf{n}(s_t)}\Big) \mathrm{d}s_t \mathrm{d}p$$
$$= -\ln\Bigg(\mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1}) \times \exp\Bigg[\underbrace{-\int_{\{p\}} \mathsf{p}(p|s_{t-1}) \int_{\{s\}} \mathsf{m}(s_t|\alpha_t, s_{t-1}, p) \times \ln\Big(\frac{\mathsf{m}(s_t|\alpha_t, s_{t-1}, p)}{\mathsf{m}^{\mathsf{i}}(s_t|\alpha_t, s_{t-1}, \hat{p}_t)\mathsf{n}(s_t)}\Big) \mathrm{d}s_t \mathrm{d}p}_{:= -\Delta(\alpha_t, \hat{p}_t, s_{t-1})}\Bigg]\Bigg)$$
$$\mathsf{n}(s_{t-1}) := \int_{(\{\alpha\},\{p\})} \mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1}) \times \exp[-\Delta(\alpha_t, \hat{p}_t, s_{t-1})] \mathrm{d}(\alpha_t, \hat{p}_t)$$
$$\mathsf{r}^{\mathsf{o}}(\alpha_t, \hat{p}_t|s_{t-1}) = \mathsf{r}^{\mathsf{o}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{r}^{\mathsf{o}}(\hat{p}_t|s_{t-1}) \propto \mathsf{r}^{\mathsf{i}}(\alpha_t|s_{t-1}, \hat{p}_t)\mathsf{p}(\hat{p}_t|s_{t-1}) \exp[-\Delta(\alpha_t, \hat{p}_t, s_{t-1})]. \qquad \square$$

Remark 3 (Complexity and Propositions 3 & 4)

✓ *The dynamic programming generally needs approximations [36]. This fact is behind need for model-based predictive control [26]. The proposed strategy has a potential to enhance such approximate strategies as it supports exploration (dual control [7]) in a novel powerful way.*

✓ *Propositions 3 and 4 deal with an unknown parameter but only Prop. 4 formulates its estimation as a part of the DM aim. It allows to solve sole estimation task. Primarily, it makes the optimal strategy more amenable to approximation as seen from the discussion in Sec. 4.*

✓ *The posterior pd $\mathsf{p}(\hat{p}_t|s_{t-1})$ influences complexity of any strategy with estimation. This implies a wide use of models from the exponential family (EF) [2] for which the functional Bayes' rule reduces to algebraic operations.*

✓ *Linear Gaussian autoregressive-regressive model is a prominent member of EF. Its estimation reduces to recursive least squares [31]. Markov decision process [32] with discrete-valued states and actions is the other EF member with the estimation that simply counts the number of observed configurations (s_t, a_t, s_{t-1}). Many extensions rely on these EF members. Their mixtures [16] are typical and serve as universal approximators [27].*

4 On the Proposed Strategy

The novel FPD-optimal strategy with estimator:

✓ respects both the knowledge collected in the posterior pd p_{t-1} (8) and influence of the parameter estimate $\hat{p}$ on α-driven DM via the function $\Delta_t(\alpha, \hat{p}, s)$ (16), which is the expected (weighted) divergence of the environment model $\mathsf{m}_t(\tilde{s}|\alpha, s, p)$ to its ideal twin $\mathsf{m}^{\mathsf{i}}_t(\tilde{s}|\alpha, s, \hat{p})$;

✓ correlates, due to the previous property, usual actions α_t with estimates $\hat{p}_t$ more deeply [5] than the common "certainty equivalence" that uses an estimate $\hat{p}_t$ based only on the past knowledge accumulated in the posterior pd p_{t-1} and unrelated with the DM dynamics;

✓ shows that the estimation is dynamic DM task influenced by the value function even when the "ordinary" action α_t is un-optimised.

Moreover, Prop. 4 opens ways to:

✓ the combination of the classical estimation with sequential stopping [41] with a dynamic DM;

✓ the radically novel approach to feasible exploration: non-explorative, certainty-equivalent, α_t-driven DM will be stimulated by random sampling of the parameter estimate $\hat{p}_t$ from the optimal rule (17): it is influenced by the posterior pd $\mathsf{p}(\hat{p}_t|s_{t-1})$ (8), which "naturally" shrinks.

The presented result reflects a part of an open-ended research, which surely requires to deal with:

✓ the analysis, almost surely simulation-based, of the proposed strategy and its model-based predictive variants;

✓ the opened research ways outlined above;

✓ specific cases and important real-life problems;

✓ the potential indicated in connection with the use of closed-loop states;

✓ the general FPD with unobserved states in the ideal pd.

[5] The paper [10], indicates, within a quite different set up, how useful such a dependence can be.

The foreseeable positive consequences make this research worthwhile. You are invited to join it.

Acknowledgements

MŠMT ČR LTC18075 and EU-COST Action CA16228 support this research.

References

[1] K.J. Åström and B. Wittenmark. *Adaptive Control*. Addison-Wesley, 1994.

[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] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2017.

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

[6] A. Bušič and S. Meyn. Action-constrained Markov decision processes with Kullback-Leibler cost. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proc. of Machine Learning Research*, volume 75, pages 1–14. MLR Press, 2018.

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

[8] P. Guan, M. Raginsky, and R. Willett. Online Markov decision processes with Kullback-Leibler control cost. In *Am. Control Conference*, pages 1388–1393. IEEE, June 2012.

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

[10] T.A.N. Heirung, B.E. Ydstie, and B. Foss. Dual adaptive model predictive control. *Automatica*, 80:340–348, 2017.

[11] J. Homolová, A. Černěcká, T.V. Guy, and M. Kárný. Affective decision-making in ultimatum game: Responder. In M. Slavkovik, editor, *Multi-Agent Systems, EUMAS 2018*, volume LNAI 11450, pages 127–139. Springer Nature Switzerland AG, 2019.

[12] H.J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20):200201, 2005.

[13] M. Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719–1722, 1996.

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

[15] M. Kárný. Towards on-line tuning of adaptive-agent's multivariate meta-parameter. *Int. J. of Machine Learning and Cybernetics*, 12(9):2717–2731, 2021.

[16] 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.

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

[18] 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.

[19] M. Kárný and T.V. Guy. On the origins of imperfection and apparent non-rationality. In T.V. Guy and et al, editors, *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*, Studies in Comput. Intelligence, pages 57–91. Springer, Switzerland, 2015.

[20] M. Kárný and T.V. Guy. Preference elicitation within framework of fully probabilistic design of decision strategies. In *IFAC Int. Workshop on Adaptive and Learning Control Systems*, volume 52, pages 239–244, 2019.

[21] M. Kárný and F. Hůla. Fusion of probabilistic unreliable indirect information into estimation serving to decision making. *Int. Journal of Machine Learning and Cybernetics*, 12:33673378, 2021.

[22] E.D. Klenske and P. Hennig. Dual control for approximate Bayesian reinforcement learning. *JMLR*, 17:1–30, 2016.

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

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

[25] P.R. Kumar. A survey on some results in stochastic adaptive control. *SIAM J. Control and Applications*, 23:399–409, 1985.

[26] D.Q. Mayne. Model predictive control: Recent developments and future promise. *Automatica*, pages 2967–2986, 2014.

[27] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probab. & Stat. Wiley, N.Y., 2000.

[28] Ali Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annual Reviews in Control*, 45:107 – 117, 2018.

[29] P.A. Ortega and D.A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[30] D. Palenicek. A survey on constraining policy updates using the KL divergence. In B. Belousov, H. Abdulsamad, P. Klink, S. Parisi, and J. Peters, editors, *Reinforcement Learning Algorithms: Analysis and Applications*. Springer, Cham, 2021.

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

[32] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[33] A. Quinn, P. Ettler, L. Jirsa, I. Nagy, and P. Nedoma. Probabilistic advisory systems for data-intensive applications. *Int. J. Adapt Control Signal Process.*, 17(2):133–148, 2003.

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

[35] 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.

[36] J. Si, A.G. Barto, W.B. Powell, and D. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*. Wiley, 2004.

[37] H. Tang and et al. #Exploration: A study of count-based exploration for deep reinforcement learning. In I. Guyon et al, editor, *Adv. in Neur. Inf. Proc. 30*, pages 2753–2762. Curran Associates, Inc., 2017.

[38] N. Tishby and D. Polani. Information theory of decisions and actions. In V. Cutsuridis and et al, editors, *Perception-Action Cycle*, pages 601–636. Springer, N.Y., 2011.

[39] 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.

[40] A. Wald. *Statistical Decision Functions*. J. Wiley, 1950.

[41] A. Wald. *Sequential Analysis*. Dover Publications, 2013.

[42] E. Zugarová and T.V. Guy. Similarity-based transfer learning of decision policies. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 37–44. IEEE, 2020.