On Approximate Fully Probabilistic Design of Decision Making Strategies

Miroslav Kárný

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic. Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic,
school@utia.cas.cz

Abstract. An efficient support of a single decision maker is vital in constructing scalable systems addressing complex decision-making (DM) tasks. Fully probabilistic design (FPD) of DM strategies, an extension of dynamic Bayesian DM, provides a firm basis for such a support. The limited cognitive and evaluation resources of the supported decision maker cause that theoretically optimal solutions are realised only approximately. Thus, the truly efficient support has to include reliable means for constructing approximate solutions of DM subtasks. The current paper deals with the design of the approximately optimal DM strategy for a known environment model and adequately described DM preferences. The design relies on: **a)** the explicit minimiser found within FPD; **b)** randomised nature of the strategy provided by FPD.

Keywords decision making; Bayesian learning; minimum cross-entropy principle; fully probabilistic design of DM strategies; linear-quadratic DM

1 Introduction

The paper addresses a particular problem within a research aiming at creation of a systematic support of DM. The support has to respect that any real decision maker devotes a limited cognitive and evaluation resources to single DM problem and mostly has to use an approximation of theoretically optimal DM strategy. A design of such strategy is made here for a specific but widely applicable DM.

1.1 Basic Notions

This subsection fixes basic notions, which strongly vary over different DM-inspecting domains (statistics, economy, control theory, machine learning, etc.).

The decision maker designs and uses the DM *strategy* $\mathsf{s} = (\mathsf{s}_t)_{t\in\boldsymbol{t}} \in \mathbf{s}$, $\boldsymbol{t} = \{1, 2, \ldots, T\}$[1]. The DM *rules* s_t, forming the strategy s, are indexed by the discrete time t and map non-decreasing available *knowledge* $(k_t \in \boldsymbol{k_t})_{t\in\boldsymbol{t}}$, $\boldsymbol{k}_{t-1} \subseteqq \boldsymbol{k}_t$, on *actions* $(a_t \in \boldsymbol{a_t})_{t\in\boldsymbol{t}}$, $\mathsf{s}_t : \boldsymbol{k}_{t-1} \to \boldsymbol{a}_t \neq \emptyset$.

[1] Throughout, $\boldsymbol{z}$ denotes a set of possible instances of z.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

The considered knowledge increments are *data records* $d_t \in \boldsymbol{d_t} = \boldsymbol{k}_t \setminus \boldsymbol{k}_{t-1}$ ($\setminus$ denotes substraction of sets). The data record consists of the observed *environment response* r_t and of the applied action a_t. Thus, $d_t = (r_t, a_t)$ and $k_{t-1} = (d_{t-1}, \ldots, d_1, k_0)$, where k_0 denotes *prior knowledge*.

The DM strategy is designed with the aim to satisfy decision maker's DM preferences in the best possible way. They are expressed here as preferences with respect to possible closed-loop *behaviours* $b \in \boldsymbol{b}$

$$b = (g_t,\, a_t,\, k_{t-1}). \tag{1}$$

The part g_t collects variables up to the DM *horizon* T, which are considered by the decision maker but unavailable for choosing the action a_t.

1.2 FPD Formulation of DM Under Uncertainty

The addressed *DM under uncertainty* arises whenever the available knowledge k_{t-1} and the chosen action a_t do not allow the decision maker to determine uniquely the value of g_t, at least for some $t \in \boldsymbol{t}$. The classical axiomatisation [16, 1] of DM under uncertainty leads to Bayesian DM, which selects the optimal DM strategy s^{L} as a minimiser of an *expected loss*

$$\mathsf{s}^{\mathsf{L}} \in \operatorname{Arg}\min_{\mathsf{s}\in\mathbf{s}} \mathsf{E}_{\mathsf{s}}[\mathsf{L}|k_0] = \operatorname{Arg}\min_{(\mathsf{s}_t:\boldsymbol{k}_{t-1}\to\boldsymbol{a}_t)_{t\in\boldsymbol{t}}} \int_{\boldsymbol{b}} \mathsf{L}(b)\mathsf{f}_{\mathsf{s}}(b|k_0)\ \mathrm{d}b. \tag{2}$$

Bayesian DM requires specification of a *loss* $\mathsf{L} : \boldsymbol{b} \to (-\infty, \infty]$, *quantifying decision-maker's preferences*, and of the probability distribution of the possible behaviours $b \in \boldsymbol{b}$. It serves for evaluation of the conditional *expectations* $\mathsf{E}_{\mathsf{s}}[\cdot|k_0]$ for strategies $\mathsf{s} \in \mathbf{s}$ and it is given by the *probability density* (pd, $\mathsf{f}_{\mathsf{s}}(b|k_0)$) of behaviours b conditioned on a prior knowledge k_0 with respect to a measure $\mathrm{d}b$.

The exploited *fully probabilistic design (FPD)* of DM strategies [8, 20] quantifies DM preferences via an *ideal pd* $\mathsf{f}_I(b|k_0)$, which expresses desirability of possible behaviours $b \in \boldsymbol{b}$. FPD selects the strategy-dependent loss $\mathsf{L}_{\mathsf{s}} = \ln(\mathsf{f}_{\mathsf{s}}/\mathsf{f}_I)$. With this loss, the optimal DM strategy s^o becomes the minimiser of the Kullback-Leibler divergence $\mathcal{D}(\mathsf{f}_{\mathsf{s}}||\mathsf{f}_I)$ (KLD, [12])

$$\begin{aligned}\mathsf{s}^o \in \operatorname{Arg}\min_{\mathsf{s}\in\mathbf{s}} \mathsf{E}[\ln(\mathsf{f}_{\mathsf{s}}/\mathsf{f}_I)|k_0] &= \operatorname{Arg}\min_{\mathsf{s}\in\mathbf{s}} \int_{\boldsymbol{b}} \mathsf{f}_{\mathsf{s}}(b|k_0) \ln\left(\frac{\mathsf{f}_{\mathsf{s}}(b|k_0)}{\mathsf{f}_I(b|k_0)}\right)\ \mathrm{d}b \\ &= \operatorname{Arg}\min_{(\mathsf{s}_t:\boldsymbol{k}_{t-1}\to\boldsymbol{a}_t)_{t\in\boldsymbol{t}}} \mathcal{D}(\mathsf{f}_{\mathsf{s}}||\mathsf{f}_I).\end{aligned} \tag{3}$$

It is always possible to construct explicitly a FPD problem formulation, which is arbitrarily close to the given Bayesian DM task [10] and there are practically significant FPD tasks having no Bayesian counterpart [9].

1.3 The Addressed Problem and Solution Idea

The design of the optimal DM strategy (2) reduces to *dynamic programming* [2]. It gives *deterministic* strategy s^{L} generating actions, which are minimising

arguments in the functional equation evolving *value function* $\zeta^{\mathsf{L}}(k_t)$ against time

$$\zeta^{\mathsf{L}}(k_{t-1}) = \min_{a_t \in \boldsymbol{a_t}} \mathsf{E}[\zeta^{\mathsf{L}}(k_t)|a_t, k_{t-1}],\ \zeta^{\mathsf{L}}(k_T) = \mathsf{L}(b). \tag{4}$$

The existence of its analytical solution is an exception and a version of approximate dynamic programming [18] is inevitable.

The design of the optimal DM strategy s^o (3) is similar to (4) and also calls for an approximation in generic case. Its design is addressed in this paper. The proposed approximation exploits that: **i)** FPD has an explicit minimiser [8], **ii)** the *value function* $\zeta(k_t)$ *in FPD* solves a nonlinear integral equation, which determines the unique *randomised optimal DM strategy*.

The proposed approximation exploits the fact that the integral equation for the value function $\zeta(k_t)$ has to hold for any knowledge k_t even if it resulted from an application of non-optimal actions. Thus, it suffices to find a function, which solves the discussed equation on a sufficiently rich subset of $\boldsymbol{k}_t$ and then we surely get an approximation of the value function.

Technically, the integral equation is converted into a probabilistic model of a parametric approximation of the value function. Then, parameter estimates are updated via the Bayes rule on realised (non-optimal) past. The application of the corresponding randomised DM strategy makes the acquired knowledge sufficiently rich. The inevitable approximation errors can be and should be taken into account by employing stabilised forgetting [11]. This measure is advisable to any approximate sequential learning [6].

1.4 Layout

Section 2 specifies the assumptions delimiting the supported DM tasks and recalls the exploited information about FPD. Section 3 forming the core of the paper proposes the approximation of the optimal FPD strategy. Section 4 applies the general result to a widely used linear-quadratic dynamic DM (control, [13]). Section 5 provides a numerical illustration. Section 6 concludes the text.

2 Stationary FPD Caring about Observable Behaviour

In this preparatory section, the DM task leading to a stationary version of FPD is formulated and solved. For the sake of presentation simplicity, it deals with preferences specified for observable behaviours only. Thus, the part g_t of the behaviour b in (1) consists of yet unobserved environment responses $(r_\tau)_{\tau \geq t}$ and non-applied actions $(a_\tau)_{\tau > t}$.

The pd $\mathsf{f}_{\mathsf{s}}(b|k_0)$, describing behaviours $b \in \boldsymbol{b}$ under a DM strategy $\mathsf{s} \in \mathbf{s}$, can be factorised via the chain rule [15]

$$\mathsf{f}_{\mathsf{s}}(b|k_0) = \prod_{t \in \boldsymbol{t}} \underbrace{\mathsf{f}_{\mathsf{s}}(r_t|a_t, k_{t-1})}_{\text{environment model}} \times \underbrace{\mathsf{f}_{\mathsf{s}}(a_t|k_{t-1})}_{\text{DM-rule model}} \tag{5}$$

$k_t = ((r_t, a_t), k_{t-1}) = (d_t, k_{t-1})$ is the knowledge available for choosing a_{t+1}.

2.1 Considered Class of DM Tasks

The supported DM tasks are delimited by the following conditions.

- The environment model is a time-invariant, strategy-independent, state-space model $\mathsf{m}(x_t|a_t, x_{t-1})$ with the finite-dimensional real state $x_t \in \boldsymbol{x_t}$ and action $a_t \in \boldsymbol{a_t}$. The state x_t is a known image of its previous value x_{t-1} and of the observed data record $d_t = (r_t, a_t)$. Thus, for all $t \in \boldsymbol{t}$,

$$\mathsf{f}_{\mathsf{s}}(r_t|a_t, k_{t-1}) = \mathsf{m}(x_t|a_t, x_{t-1}).$$

- The initial state x_0 is assumed to be a part of the prior knowledge k_0.
- The DM rules s_t having the same model (5) are operationally equivalent and they are formally identified with their model. Thus, for all $t \in \boldsymbol{t}$,

$$\mathsf{s}_t(a_t|k_{t-1}) = \mathsf{f}_{\mathsf{s}}(a_t|k_{t-1}).$$

- The ideal pd $\mathsf{f}_I(b|k_0)$ only cares about preferences on the observed states and actions and thus it can be factorised as follows

$$\mathsf{f}_I(b|k_0) = \prod_{t\in\boldsymbol{t}} \mathsf{m}_I(x_t|a_t, x_{t-1})\mathsf{s}_I(a_t|x_{t-1}), \tag{6}$$

  where the given pds m_I, s_I in (6) are assumed to be time-invariant.
- The design is performed for the DM horizon $T \to \infty$.

2.2 Optimal DM Strategy To Be Approximated

Proposition 1 (Solution of Stationary FPD) *Let a stabilising DM strategy $\mathsf{s}^s \in \boldsymbol{s}$ exist, which means that*

$$c_{\mathsf{s}^s} = \lim_{T\to\infty} \frac{1}{T}\mathcal{D}(\mathsf{f}_{\mathsf{s}^s}||\mathsf{f}_I) < \infty.$$

Then, the optimal DM strategy, minimising the KLD (3), is stabilising, stationary $\mathsf{s}^o(b|k_0) = \prod_{t\in\boldsymbol{t}} \mathsf{s}^o(a_t|x_{t-1})$ and determined by the time-invariant DM rule

$$\mathsf{s}^o(a_t|x_{t-1}) = \frac{\mathsf{s}_I(a_t|x_{t-1})\exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1})\right]}{\underbrace{\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1})\exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1})\right]\ \mathrm{d}a_t}_{\exp[-\mathsf{h}(x_{t-1})]}} \tag{7}$$

$$\mathsf{D}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1}) \ln\left(\frac{\mathsf{m}(x_t|a_t, x_{t-1})}{\mathsf{m}_I(x_t|a_t, x_{t-1})}\right)\ \mathrm{d}x_t \geq 0 \tag{8}$$

$$\mathsf{H}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1})\mathsf{h}(x_t)\ \mathrm{d}x_t \geq -c \tag{9}$$

$$c = \lim_{T\to\infty} \frac{1}{T}\mathcal{D}(\mathsf{f}_{\mathsf{s}^o}||\mathsf{f}_I) \in [0, c_{\mathsf{s}^s}] \ \Rightarrow\ \zeta(x_t) = c + \mathsf{h}(x_t) \geq 0. \tag{10}$$

Proof It is omitted as it follows steps used in proving standard dynamic programming [2]. It only exploits the fact that the KLD reaches its minimum zero value for coinciding arguments. You can consult [3] containing the proof concerning general case with preferences specified also for unobserved states. □

Remarks

- The functional equations (7) – (9) rarely have an analytical solution. Their approximate solution is proposed in Section 3.
- The function $\exp[-\mathsf{h}(x)]$ is proportional to the stationary pd of the state when the optimal strategy is used. It is seen from (7) and the conditioning rule $\mathsf{pd}(a|x) = \mathsf{pd}(a,x)/\mathsf{pd}(x)$. This interpretation should be respected when selecting the set of functions in which its approximation is searched for.
- The decisive function $\mathsf{h}(x_t)$ is the shifted value of the non-negative value function $\zeta(x_t)$ (10). Its non-negativity implies $\mathsf{H}(a_t, x_{t-1}) \geq -c$ (9).
- The function $\mathsf{D}(a_t, x_{t-1})$ (8) is non-negative as it is the conditional KLD of the environment model m from its ideal counterpart m_I.
- All involved functions are assumed to be time invariant. The time-invariance of the environment model $\mathsf{m}(x_t|a_t, x_{t-1})$ is asymptotically guaranteed if it is obtained as the predictive pd resulting from Bayesian learning, [15]. Thus, the presented treatment is extendable to this case.

3 Approximation of the Optimal Strategy

Here, the approximation of the optimal DM strategy is searched for. It consists of approximations of the functions D, H defining the pd s^o (7), cf. Proposition 1.

The conditional KLD $\mathsf{D}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1}) \ln\left(\frac{\mathsf{m}(x_t|a_t,x_{t-1})}{\mathsf{m}_I(x_t|a_t,x_{t-1})}\right) \mathrm{d}x_t$ in (7) is time-invariant and can be, at least approximately, evaluated off-line. Thus, the approximation concerns primarily the shifted value function $\mathsf{h}(x_t)$ and its expectation $\mathsf{H}(a_t, x_{t-1})$ with respect to the environment model

$$\mathsf{H}(a_t, x_{t-1}) = \mathsf{E}[\mathsf{h}|a_t, x_{t-1}] = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1})\mathsf{h}(x_t)\ \mathrm{d}x_t.$$

3.1 Technical Elaboration

The proposed approximation uses:

- parametric approximation of $\mathsf{h}(x) \approx \mathsf{h}(x, \Theta)$ inducing the approximation
$$\mathsf{H}(a, x_{t-1}) = \mathsf{E}[\mathsf{h}|a, x_{t-1}] \approx \mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{E}[\mathsf{h}(\cdot, \Theta)|a, x_{t-1}, \Theta];$$
- mean-value theorem applied to the integral over $\boldsymbol{a}_t$, Proposition 1 & (11);
- decomposition of expectation $\mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{E}[\mathsf{h}(\cdot, \Theta)|a, x_{t-1}, \Theta]$ in $\mathsf{h}(x_t, \Theta)$ and *innovation* $\varepsilon(a, x_{t-1}, \Theta)$: $\mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{h}(x_t, \Theta) + \varepsilon(x_t, a, x_{t-1}, \Theta)$ [15];
- minimum KLD (cross-entropy) principle [17], which extends a partial information about a pd into the complete pd.

Proposition 1 implies that the function $\mathsf{h}(x, \Theta)$ should solve the equation

$$\exp[-\mathsf{h}(x_{t-1}, \Theta)] = \int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1}, \Theta)\right] \, \mathrm{d}a_t, \quad (11)$$

which *has to hold for any state* $x_{t-1} \in \boldsymbol{x_{t-1}}$ *even if it resulted from use of non-optimal past actions*. An application of mean-value theorem to this equation, introduction of innovations and logarithmic transformation provide

$$\begin{aligned} -\mathsf{h}(x_{t-1}, \Theta) &= \underbrace{\ln \left[\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp[-\mathsf{D}(a_t, x_{t-1})] \, \mathrm{d}a_t \right]}_{\phi(x_{t-1})<0} \\ &\quad + \ln \left(\exp \left[- \int_{\boldsymbol{x}_t} \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \mathsf{h}(x_t, \Theta) \, \mathrm{d}x_t \right] \right) \\ &= \phi(x_{t-1}) - \mathsf{h}(x_t, \Theta) + \varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta), \end{aligned} \quad (12)$$

where $\underline{a}(x_{t-1}, \Theta)$ denotes the action resulting from the mean-value theorem.

The time and Θ invariant function $\phi(x_{t-1})$ is *negative*. It can be prepared off-line and thus it is fully determined by the knowledge k_{t-1}. The innovations

$$\varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta) = \mathsf{h}(x_t, \Theta) - \int_{\boldsymbol{x}_t} \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \mathsf{h}(x_t, \Theta) \, \mathrm{d}x_t$$

are, by construction, zero mean and uncorrelated with their past values, [15],

$$\int_{\boldsymbol{x_t}} \varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta) \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \, \mathrm{d}x_t = 0.$$

This property and (12) imply that the positive random value of the value function $\zeta(x_t, c, \Theta) = c + \mathsf{h}(x_t, \Theta)$ has conditional expectation $\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1}) \in (0, \infty)$, i.e.

$$\mathsf{E}[\zeta(x_t, c, \Theta) - \zeta(x_{t-1}, c, \Theta)|\zeta(x_{t-1}, \Theta), k_{t-1}, \Theta] = \phi(x_{t-1}) < 0. \quad (13)$$

The minimum KLD principle [17] completes this information about the conditional expectation into the exponential distribution

$$\mathsf{f}(\zeta(x_t, c, \Theta)|\zeta(x_{t-1}, c, \Theta), k_{t-1}, c, \Theta) = \frac{\exp\left[-\frac{\zeta(x_t, c, \Theta)}{\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1})}\right]}{\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1})}. \quad (14)$$

In order to avoid discussion of non-linear Bayesian learning, which is out our scope, we also adopt the approximation

$$\zeta(x_{t-1}, c, \Theta) = c + \mathsf{h}(x_{t-1}, \Theta) \approx (\hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}))\chi(c + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) \geq 0), \quad (15)$$

where $\hat{c}_{t-1}, \hat{\Theta}_{t-1}$ are point estimates of c, Θ based on k_{t-1} and $\chi(\cdot)$ is an indicator function of the set in its argument. In this way, the parametric model relating $\zeta(x_t, c, \Theta)$ to the knowledge k_{t-1} and unknown c, Θ is obtained

$$\begin{aligned}
\mathsf{f}(\zeta(x_t, c, \Theta)|k_{t-1}, c, \Theta) &= \alpha_{t-1}^{-1} \exp\left[-(c + \mathsf{h}(x_t, \Theta))\alpha_{t-1}\right] \chi(c + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) \geq 0) \\
\alpha_{t-1}^{-1} &= \hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) + \phi_{t-1} \qquad (16) \\
&= \hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) + \ln\left(\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp[-\mathsf{D}(a_t, x_{t-1})] \, \mathrm{d}a_t\right).
\end{aligned}$$

The gained parametric model is used in Bayesian learning, which evolves the posterior pd $\mathsf{f}(c, \Theta|k_t)$ on the unknown c, Θ. The evolution has the form, cf. (14), (15) and (16)

$$\mathsf{f}(c, \Theta|k_t) \propto \mathsf{f}(c, \Theta|k_{t-1}) \exp\left[-(c + \mathsf{h}(x_t, \Theta))\alpha_{t-1}\right] \qquad (17)$$

Non-negativity of $\zeta(x_t, c, \Theta)$ and its conditional expectation is the key information about c, which subtracts in (13). It implies that $c \geq \max_{\tau \leq t-1}(-\mathsf{h}_\tau)$.

3.2 Discussion

Here, explanatory comments are added to the above technical manipulations. In summary, the proposed learning of DM strategy relies on the heuristic steps:

- *The parametric expression $c + \mathsf{h}(x, \Theta)$ of the value function $\zeta(x)$ is supposed to approximate well the value function for some $\Theta \in \boldsymbol{\Theta}$.*
  This assumption can be met by an appropriate choice of the function $\mathsf{h}(x, \Theta)$, by exploiting a "universal approximation property" [4]. In this respect, it is important that, by construction, the approximated function $\exp[-\mathsf{h}(x)]$ is proportional to the stationary distribution of x_t for the optimal strategy.
- *The pd of the approximate value function $\zeta(x_t, c, \Theta) = c + \mathsf{h}(x_t, \Theta)$ has been derived via maximum entropy while neglecting a direct information about the environment model $\mathsf{m}(x_t|a_t, x_{t-1})$.*
  This assumption is unrestrictive as the scalar $\mathsf{h}(x_t, \Theta)$ depends on multivariate x_t and the adopted DM strategy in a complex unknown way. It means that a negligible amount of useful and truly available knowledge is neglected. The dependence on the environment model and the ideal counterpart of the applied DM strategy projects into the weight α_{t-1} (16).
- *The crude approximation (15) is adopted.*
  This assumption is generally unnecessary. It has helped us in suppressing the need to discuss non-linear Bayesian learning, which is out of our scope.

The following points are also worth discussing.

- The posterior pd on Θ serves for approximating the optimal DM strategy, i.e. for estimation of $\mathsf{H}(a_{t+1}, x_t) = \mathsf{E}[\mathsf{h}(\cdot)|a_{t+1}, x_t]$. It hints to take

$$\begin{aligned}
\mathsf{H}(a_{t+1}, x_t) &\approx \int_{\boldsymbol{x_{t+1}}} \mathsf{m}(x_{t+1}|a_{t+1}, x_t) \int_{\boldsymbol{\Theta}} \mathsf{h}(x_{t+1}, \Theta)\mathsf{f}(\Theta|k_t) \, \mathrm{d}\Theta \, \mathrm{d}x_{t+1} \\
&\approx \int_{\boldsymbol{x_{t+1}}} \mathsf{m}(x_{t+1}|a_{t+1}, x_t)\mathsf{h}(x_{t+1}, \hat{\Theta}_t) \, \mathrm{d}x_{t+1} = \hat{\mathsf{H}}(a_{t+1}, x_t).
\end{aligned}$$

The last approximate equality delimits the needed point estimate $\hat{\Theta}_t$ of Θ.

- The function $-\mathsf{D}(a_{t+1}, x_t) - \mathsf{H}(a_{t+1}, x_t|k)$ is immediately used for generating a_{t+1}, see (7). After applying a_{t+1} and recording x_{t+1}, the learning process can continue. The randomised nature of the optimal DM strategy in FPD makes the used DM strategy explorative. The higher is uncertainty about the parameter Θ the flatter is the pd used for generating a_{t+1}. It indicates qualitatively plausible variations of the exploration extent.
- The actions $\underline{a}(x_{t-1}, \Theta)$ considered in approximate evaluations (12) origin from the ideal counterpart of the DM strategy, neither from the optimal nor the used DM strategy. The basic idea of the construction indicates that the learning running on non-optimal states, caused by the applied non-optimal actions, is counteracted by the weight α_{t-1} (16) determined by the combination environment model – ideal strategy.
- It is possible to introduce additional weighting suitable whenever learning contains some approximation error [6]. We shall not employ it in order to check whether the proposed learning copes with the "incorrect data".

4 Application to Linear-Gaussian DM

The influence and extent of the applicability of adopted approximations as well as of heuristic assumptions, see Section 3, are yet unclear. Thus, it makes sense to check the proposed procedure on a case with a known solution. Linear-Gaussian DM treated here serves to this purpose. It is given by the following assumptions.

- The environment model is linear Gaussian

$$\mathsf{m}(x_t|a_t, x_{t-1}) = \mathcal{N}_{x_t}(Ax_{t-1} + Ba_t, R)$$
$$\mathcal{N}_x(\mu, R) = |2\pi R|^{-0.5} \exp\left[-0.5(x-\mu)' R^{-1}(x-\mu)\right],$$

  where A, B, determining its conditional expectation, as well as positive definite covariance $R > 0$ are known matrices of dimensions compatible with the vectorial state x_t and action a_t. $'$ denotes transposition.
- The ideal counterpart of the environment model is chosen also Gaussian

$$\mathsf{m}_I(x_t|a_t, x_{t-1}) = \mathcal{N}_{x_t}(0, R).$$

  It reflects the wish to push the state to zero (so called regulation problem, [13]). The equality of covariances of the environment model and its ideal counterpart respects the fact that R represents the lowest reachable covariance. The ideal counterpart of the DM strategy is chosen also Gaussian

$$\mathsf{s}_I(a_t|x_{t-1}) = \mathcal{N}_{a_t}(0, q).$$

  This ideal pd represents the wish to spare acting energy $0.5a_t' q^{-1} a_t$.

In this case, the exact solution of FPD is known, [5]. It holds

$\exp[-\mathsf{h}(x)] = \mathcal{N}_x(0, S)$ the covariance $S > 0$ is known as Riccati matrix
$\mathsf{s}^o(a_t|x_{t-1}) = \mathcal{N}_{a_t}(-L'x_{t-1}, Q)$,

where the matrices $S > 0$, $Q > 0$ as well the matrix L (control law) are determined by parameters of the environment model and those of ideal pds.

The proposed procedure specialises to this case as follows.

The function $\mathsf{D}(a_t, x_{t-1})$ (8), the conditional KLD of the pd m from the pd m_I, can be computed analytically, e.g. [7],

$$\mathsf{D}(a_t, x_{t-1}) = 0.5\,(Ax_{t-1} + Ba_t)'\,R^{-1}\,(Ax_{t-1} + Ba_t)\,.$$

The function $\phi(x_{t-1})$ (12) is also given analytically

$$\begin{aligned}\phi(x_{t-1}) &= \ln\left[\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1})\exp[-\mathsf{D}(a_t, x_{t-1})]\;\mathrm{d}a_t\right] \\ &= 0.5\left[\ln\left(\left|I + qB'R^{-1}B\right|\right) - x'_{t-1}A'(R + BqB')^{-1}Ax_{t-1}\right],\end{aligned} \tag{18}$$

where I denotes unit matrix. The neat final form is obtained by employing so called Woodbury formula.

The next approximation $\mathsf{h}(x, \Theta)$ of $\mathsf{h}(x)$ admits the needed comparisons

$$\exp[-\mathsf{h}(x,\Theta)] = \mathcal{N}_x(0,\Theta),\ \Theta > 0\ \Rightarrow\ \mathsf{h}(x,\Theta) = 0.5(\ln(|\Theta|) + x'\Theta^{-1}x) + \textit{constant}. \tag{19}$$

The forms of $\phi(x_{t-1})$ (18) and of $\mathsf{h}(x)$ (19) specialise the learning (17) to

$$\begin{aligned}&\mathsf{f}(\Theta|k_t) \propto \mathsf{f}(\Theta|k_{t-1})\exp\left[-(c + \mathsf{h}(x_t,\Theta))\alpha_{t-1}\right]\chi(c + \mathsf{h}(x_t,\hat{\Theta}_t) \geq 0) \\ &\propto |\Theta|^{-0.5\nu_t}\exp\left[-(c\nu_t + 0.5\mathrm{tr}\left(\Theta^{-1}V_t\right))\right]\chi(c \geq \bar{c}_t) \\ &\bar{c}_t = \max\left[\bar{c}_{t-1}, -0.5\left(\ln\left(\left|\hat{\Theta}_t\right|\right) + x'_t\hat{\Theta}_t x_t\right)\right] \\ &V_t = V_{t-1} + \alpha_{t-1}x_t x'_t,\quad \nu_t = \nu_{t-1} + \alpha_{t-1},\quad \alpha_{t-1} = \\ &\frac{2}{2\hat{c}_{t-1} + \ln\left(|I + qB'R^{-1}B|\,|\hat{\Theta}_{t-1}|\right) + x'_{t-1}(\hat{\Theta}^{-1}_{t-1} - A'(R + BqB')^{-1}A)x_{t-1}} \\ &V_0 > 0, \nu_0, \bar{c}_0 > 0 \text{ determine the prior pd in the self-reproducing form} \\ &\mathsf{f}(c,\Theta|k_0) \propto |\Theta|^{-0.5\nu_0}\exp\left[-\left(c\nu_0 + 0.5\mathrm{tr}\left(\Theta^{-1}V_0\right)\right)\right]\chi(c \geq \bar{c}_0).\end{aligned} \tag{20}$$

The final formula (20) is intuitively plausible as:

- Θ should estimate the Riccati matrix, which is covariance matrix of the state in the closed loop with the optimal DM strategy. The proposed learning provides such estimate in the form of the *weighted* covariance. The adopted maximum-likelihood estimates $\hat{c}_t$, $\hat{\Theta}_t$ of c, Θ for the knowledge k_t are

$$\hat{c}_t = \bar{c}_t,\ \hat{\Theta}_t = \frac{V_t}{\nu_t}. \tag{21}$$

- The weight of the dyad increment $x_t x'_t$ is the higher the closer is x_{t-1} to 0.
- The relative closeness of x_{t-1} to zero is determined by relations between properties of the controlled environment (A, B, R) and the cost q of actions.

The approximately optimal DM strategy corresponding to (7) and (21) is

$$\begin{aligned}
&\hat{\mathsf{s}}_{t+1}(a_{t+1}|x_t) \propto \mathcal{N}_{a_t}(0,q)\exp[-\mathsf{D}(a_{t+1},x_t)-\hat{\mathsf{H}}(a_{t+1},x_t)]\\
&\propto \exp\left[-0.5\left(a'_{t+1}q^{-1}a_{t+1}+(Ax_t+Ba_{t+1})'\left(\hat{\Theta}_t^{-1}+R^{-1}\right)(Ax_t+Ba_{t+1})\right)\right]\\
&=\mathcal{N}_{a_{t+1}}(-\hat{L}_t x_t,\hat{Q}_t)\\
&\hat{Q}_t=(q^{-1}+B'(\hat{\Theta}_t^{-1}+R^{-1})B)^{-1},\quad \hat{L}_t=\hat{Q}_t B'(\hat{\Theta}_t^{-1}+R^{-1})A.
\end{aligned}$$

Structurally, it corresponds with the optimal DM strategy. Limited experimental experience indicates that the procedure often approaches the optimal DM strategy. The approximation quality can be improved by employing the stabilised forgetting counteracting the accumulation of approximation errors [6].

5 Numerical Example

This section illustrates numerically behaviour of the algorithm in linear-Gaussian case described in the previous section. The environment model is specified by

$$A=\begin{bmatrix}0.70 & -0.30 & 0.80\\ 0.70 & 0.95 & 0.20\\ 0.20 & 0.00 & 0.90\end{bmatrix},\quad B=\begin{bmatrix}1.00\\ 0.50\\ 0.00\end{bmatrix},\quad R=\begin{bmatrix}1.00 & -0.20 & 0.20\\ -0.20 & 0.29 & 0.11\\ 0.20 & 0.11 & 0.17\end{bmatrix},$$

where the covariance is positive definite as it was generated as product of its Choleski factors. In the inspected regulation problem and for the scalar action, preferences are specified just via the ideal action variance $q = 10$. The results are shown for $T = 100$ allowing to display time trajectories. Non-presented runs up to $T = 50000$ confirmed stability of the solution and of the closed DM loop.

The optimal stationary strategy is given by the Gaussian pd

$$\mathcal{N}_{a_t}(-[0.817,\ 0.788,\ -0.409]x_{t-1},0.069),$$

while the proposed procedure provides

$$\mathcal{N}_{a_t}(-[0.788,\ 0.781,\ -0.531]x_{t-1},0.068).$$

Closeness of sample moments of states and actions with optimal and approximate strategy indicates that the found strategy approximates well the optimal one. Importantly, the essentially same approximate strategy has been obtained

$$\mathcal{N}_{a_t}(-[0.794,\ 0.787,\ -0.512]x_{t-1},0.064)$$

when the learning run with the optimal controller. The learning was also run with enforced zero action. It lead to the controller

$$\mathcal{N}_{a_t}(-[0.746,\ 0.722,\ -0.622]x_{t-1},0.071),$$

with poorer, but still quite-reasonable, closed-loop behaviour. The mild deterioration of quality can be attributed to the lack of exploration.

The possibility to learn reasonable strategy from non-optimal closed-loop behaviour is the focal feature of the example as it indicates that the adopted concept is sound. Numerically, it manifests on time course of the weights α_t (16). They become (relatively) large if the closed-loop behaviour is locally (even by chance) close to the optimal one. Fig. 1 illustrates this statement by showing time-courses of this weight in all described configurations of experiments.

For completeness, Figure 2 shows state evolutions in all configurations.

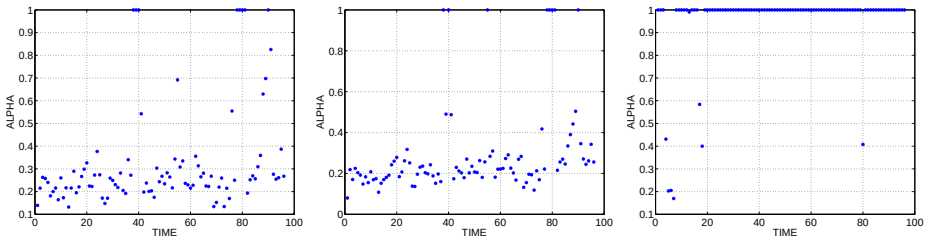

Fig. 1. Time courses of α_t (16). The left one corresponds to the closed-loop with the proposed strategy. The middle one reflects learning with the optimally closed loop. The right one concerns learning while action is fixed at zero value: the relatively high values of α_t are caused by the lack of informative data.

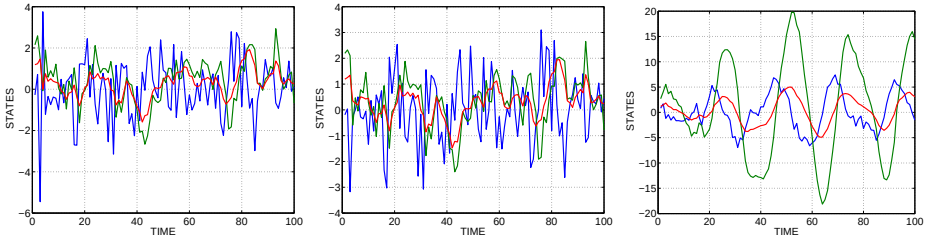

Fig. 2. Time courses of state x_t. The left one corresponds to the closed-loop with the proposed strategy. The middle one concerns the optimally closed loop. The right one concerns the loop with the action fixed at zero value. *Scales reflect regulation quality.*

6 Conclusions

The paper tries tailor approximate dynamic programming to fully probabilistic design of DM strategies. The presented preliminary results indicate that the addressed problem is solvable in the outlined way but otherwise the paper is an open-ended story. The logical necessity of respective development steps is the weakest conceptual point. Technically, the future work should focus on:

- analysing the proposed solution (at least via simulations);
- guiding in parameterisations of the function $\exp[-\mathsf{h}(x_t)]$ (universal approximation [4], probably by dynamic mixtures [14, 19]);
- combining with Bayesian learning of the environment model, [15];
- addressing the DM problem with indirectly observed state, [8];
- applying forgetting as a universal counter-measure against accumulation of approximation errors [6].

Acknowledgements This research has been supported by GAČR 13-13502S. The text has been substantially influenced by discussions with Dr. T.V. Guy.

References

1. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
2. Bertsekas, D.: Dynamic Programming and Optimal Control. Athena Scientific, Nashua, US (2001), 2nd edition
3. Guy, T.V., Kárný, M.: Stationary fully probabilistic control design. In: Filipe, J., Cetto, J.A., Ferrier, J.L. (eds.) Proc. of the Second Int. Conference on Informatics in Control, Automation and Robotics. pp. 109–112. INSTICC, Barcelona (2005)
4. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)
5. Kárný, M.: Towards fully probabilistic control design. Automatica 32(12), 1719–1722 (1996)
6. Kárný, M.: Approximate Bayesian recursive estimation. Inf. Sci. (2013), submitted
7. Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer (2006)
8. Kárný, M., Guy, T.V.: Fully probabilistic control design. Systems & Control Letters 55(4), 259–265 (2006)
9. Kárný, M., Guy, T.: On support of imperfect Bayesian participants. In: Guy, T., Kárný, M., Wolpert, D. (eds.) Decision Making with Imperfect Decision Makers, vol. 28. Springer, Berlin (2012), Intelligent Systems Reference Library
10. Kárný, M., Kroupa, T.: Axiomatisation of fully probabilistic design. Information Sciences 186(1), 105–113 (2012)
11. Kulhavý, R., Zarrop, M.B.: On a general concept of forgetting. Int. J. of Control 58(4), 905–924 (1993)
12. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–87 (1951)
13. Meditch, J.: Stochastic Optimal Linear Estimation and Control. Mc. Graw Hill (1969)
14. Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T.: Bayesian estimation of dynamic finite mixtures. Int. Journal of Adaptive Control and Signal Processing 25(9), 765–787 (2011)
15. Peterka, V.: Bayesian system identification. In: Eykhoff, P. (ed.) Trends and Progress in System Identification, pp. 239–304. Pergamon Press, Oxford (1981)
16. Savage, L.: Foundations of Statistics. Wiley, New York (1954)
17. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. 26(1), 26–37 (1980)
18. Si, J., Barto, A., Powell, W., Wunsch, D. (eds.): Handbook of Learning and Approximate Dynamic Programming. Wiley-IEEE Press, Danvers (May 2004)
19. Titterington, D., Smith, A., Makov, U.: Statistical Analysis of Finite Mixtures. John Wiley, New York (1985)
20. Todorov, E.: Linearly-solvable Markov decision problems. In: Schölkopf, B., et al (eds.) Advances in Neural Inf. Processing, pp. 1369 – 1376. MIT Press, NY (2006)