Balancing Exploitation and Exploration via Fully Probabilistic Design of Decision Policies

Miroslav Kárný and František Hůla
The Czech Academy of Sciences, Inst. of Inf. Theory and Automation, POB 18, 182 08 Prague 8, Czech Republic

Keywords: Exploitation, Exploration, Bayesian Estimation, Adaptive Systems, Fully Probabilistic Design, Markov Decision Process.

Abstract: Adaptive decision making learns an environment model serving a design of a decision policy. The policy-generated actions influence both the acquired reward and the future knowledge. The optimal policy properly balances exploitation with exploration. The inherent dimensionality curse of decision making under incomplete knowledge prevents the realisation of the optimal design. This has stimulated repetitive attempts to reach this balance at least approximately. Usually, either: (a) the exploitative reward is enriched by a part reflecting the exploration quality and a feasible approximate certainty-equivalent design is made; or (b) an explorative random noise is added to the purely exploitative actions. This paper avoids the inauspicious (a) and improves (b) by employing the non-standard fully probabilistic design (FPD) of decision policies, which naturally generates random actions. Monte-Carlo experiments confirm its achieved quality. The quality stems from methodological contributions, which include: (i) an improvement of the relation between FPD and standard Markov decision processes; (ii) a design of an adaptive tuning of an FPD-parameter. The latter also suits for the tuning of the temperature in both simulated annealing and Boltzmann's machine.

1 INTRODUCTION

The inspected decision making is close to the traditional Markov decision process (MDP, (Puterman, 2005)). The next summary of known basic facts allows us to formulate and solve the addressed problem. In order to focus on the paper's topic, we restrict ourselves to a finite amount of possible agent's actions[1] $a_t \in \mathbf{A} = \{1,\ldots,k\}$, $k \in \mathbb{N}$, $k < \infty$. They are selected in a finite amount of epochs $t \in \mathbf{T} = \{1,\ldots,l\}$, $l \in \mathbb{N}$, $l < \infty$. The agent's environment responds to actions by discrete-valued observable states $s_t \in \mathbf{S} = \{1,\ldots,m\}$, $m \in \mathbb{N}$, $m < \infty$. A given real reward $\mathsf{r} = (\mathsf{r}_t(\tilde{s},a,s),\ \tilde{s},s \in \mathbf{S},\ a \in \mathbf{A})_{t\in\mathbf{T}}$ quantifies the agent's preferences. The sequence of transition probabilities

$$\mathsf{p} = (\mathsf{p}_t(\tilde{s}|a,s),\ \tilde{s},s \in \mathbf{S}, a \in \mathbf{A})_{t\in\mathbf{T}}, \tag{1}$$

models the assumed Markov random environment. A sequence of probabilities $\pi = (\pi_t(a|s),\ a \in \mathbf{A},\ s \in \mathbf{S})_{t\in\mathbf{T}}$ describes the agent's optional, randomised and Markov policy. The MDP-optimal policy π^{MDP} maximises the expected cumulative reward

$$\pi^{\mathsf{MDP}} \in \operatorname{Arg}\max_{\pi\in\mathbf{\Pi}} \mathsf{E}^{\pi}\Big[\sum_{t\in\mathbf{T}} \mathsf{r}_t(s_t,a_t,s_{t-1})\Big]. \tag{2}$$

The strategy-dependent expectation E^{π} is implicitly conditioned on a known initial state. The optimisation runs over the set $\mathbf{\Pi}$ of Markov policies

$$\mathbf{\Pi} = \Big\{\Big(\pi_t(a|s) \geq 0,\ \sum_{a\in\mathbf{A}} \pi_t(a|s) = 1,\ \forall s \in \mathbf{S}\Big)_{t\in\mathbf{T}}\Big\}. \tag{3}$$

Dynamic programming (DP) provides the MDP-optimal policy consisting of *deterministic* decision rules $(\pi_t(a|s))_{t\in\mathbf{T}}$ selecting the maximisers $a_t^{\mathsf{MDP}}(s)$ in

$$\mathsf{v}_{t-1}(s) = \max_{a\in\mathbf{A}} \mathsf{E}^{\pi}[\mathsf{r}_t(\tilde{s},a,s) + \mathsf{v}_t(\tilde{s})|a,s],\ s \in \mathbf{S}, t \in \mathbf{T}. \tag{4}$$

The functional equation (4) evolves the value functions $\mathsf{v}_t(s)$, $s \in \mathbf{S}$, and provides the used maximising arguments $a_t^{\mathsf{MDP}}(s)$, $s \in \mathbf{S}$. It is solved backwards starting with $\mathsf{v}_{|\mathbf{T}|}(s) = 0$, $\forall s \in \mathbf{S}$. This standard solution extends to the case with the incompletely known environment model parameterised by the transition-probability values

$$\mathsf{p}_t(\tilde{s}|a,s,\theta) = \theta(\tilde{s}|a,s),\ \theta \in \mathbf{\Theta}. \tag{5}$$

[*] This research has been supported by GAČR, grants GA16-09848S and GA18-15970S.

[1] Throughout, $\mathbb{N}$ denotes set of positive integers.

Kárný, M. and Hůla, F.
Balancing Exploitation and Exploration via Fully Probabilistic Design of Decision Policies.
DOI: 10.5220/0007587208570864
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 857-864
ISBN: 978-989-758-350-6

The set $\mathbf{\Theta}$ is given the meaning of the parameter θ

$$\mathbf{\Theta} = \left\{ \theta(\tilde{s}|a,s) \geq 0,\ \sum_{\tilde{s}\in\mathbf{S}} \theta(\tilde{s}|a,s) = 1,\ \forall a \in \mathbf{A}, \forall s \in \mathbf{S} \right\}. \tag{6}$$

The parametric model (5) belongs to exponential family (Barndorff-Nielsen, 1978) and possesses Dirichlet's distribution $\mathcal{D}_\theta(V_0)$, given by the finite-dimensional occurrence array

$$V_0 = (V_0(\tilde{s}|a,s))_{\tilde{s},s\in\mathbf{S},\,a\in\mathbf{A}},\ \ V_0(\tilde{s}|a,s) > 0,$$

as its conjugate prior. With the chosen $\mathcal{D}_\theta(V_0)$, Bayesian learning (Berger, 1985) reproduces Dirichlet's form. It reduces to the updating of the occurrence array

$$V_t(s_t|a_t,s_{t-1}) = V_{t-1}(s_t|a_t,s_{t-1}) + 1,\ \text{ initiated by } V_0, \tag{7}$$

where (s_t,a_t,s_{t-1}) is the realised triple. This recursion, together with the predictive probabilities

$$\mathsf{p}(\tilde{s}|a,s,V) = \frac{V(\tilde{s}|a,s)}{\sum_{\tilde{s}\in\mathbf{S}} V(\tilde{s}|a,s)} = \hat{\theta}(\tilde{s}|a,s),\ \tilde{s},s\in\mathbf{S},\, a\in\mathbf{A}, \tag{8}$$

provides the Markov transition probability of the *information state* (s_t,V_t). Thus, the MDP-optimal policy can *formally* be computed via DP (4) where s is replaced by (s,V). Such an MDP-optimal policy $(\pi_t(a|s,V))_{t\in\mathbf{T}}$ inevitably optimally balances the explorative effort, regarding the evolution of s_t, and the exploitative effort, regarding the evolution of V_t, cf. (Feldbaum, 1961). The number of possible information states however, blows up exponentially. This prevents the evaluation and storing of the value functions $(\mathsf{v}_t(s,V))_{t\in\mathbf{T}}$.

The common remedy uses of the frozen point estimate $\hat{\theta}$ instead of θ in DP. This certainty-equivalent approximation diminishes the curse of dimensionality (Bellman, 1961). The approximation, however, gives up the care about the intentional exploration. It provably diverges from the optimal policy with a positive probability (Kumar, 1985). This experimentally well-confirmed fact has led to a range of attempts to recover the intentional exploration. The active exploration is mostly reached by introducing a random constituent into actions (Črepinšek et al., 2013; Duff, 2002; Wu et al., 2017). Good results are often achieved but the proper balance between exploration and exploitation is hard to find. This manifests itself in, repeatedly admitted, sensitivity to the choice of parameters determining the noise added to the exploitative actions.

This paper introduces the proper exploration by employing the fully probabilistic design of decision policies (FPD, (Kárný and Guy, 2006; Kárný and Kroupa, 2012)). FPD is closely related to the Kullback-Leibler control (Gómez and Kappen, 2012; Guan et al., 2012; Kappen, 2005). In the paper context, it is important that FPD leads to the *randomised*, and thus *explorative policy* unlike the usual MDP.

Methodologically, the paper relates MDP and FPD in a better way than the axiomatisation (Kárný and Kroupa, 2012). It also proposes the adaptation of an optional FPD-parameter, similar to the temperature in simulated annealing (Tanner, 1993) or Boltzmann's machine (Witten et al., 2017). Practically, it presents Monte Carlo experiments, which show that the certainty-equivalent version of FPD is indeed adequately explorative.

Layout: Section 2 recalls basic facts about the ingredients of the advocated decision policy. It formalises and solves the addressed problem. Section 3 summarises the results of extensive simulations reflecting the properties of the proposed policy. Section 4 adds concluding remarks. Appendix contains data used in simulations so that our results can be reproduced.

2 FPD AND ITS RELATION TO MDP

The environment model p (1) and any fixed policy π in (3) determine the joint probability c^π of states and actions (implicitly conditioned on the initial state)

$$\mathsf{c}^\pi(\overbrace{s_{|\mathbf{T}|},a_{|\mathbf{T}|},s_{|\mathbf{T}|-1},a_{|\mathbf{T}|-1}\ldots,s_1,a_1}^{\text{behaviour } b\in\mathbf{B} = \mathsf{X}_{t\in\mathbf{T}}(\mathbf{S}\times\mathbf{A})}) \\ = \prod_{t\in\mathbf{T}} \mathsf{p}_t(s_t|a_t,s_{t-1})\pi_t(a_t|s_{t-1}). \tag{9}$$

This closed-loop model $\mathsf{c}^\pi(b)$ *completely* describes (closed-loop) behaviours $b\in\mathbf{B}$ (9) consisting of observed and opted variables. Thus, all design ways, e.g. MDPs with different rewards, leading to the same c^π are equivalent. This observation (Ullrich, 1964) implies that decision objectives can generally be expressed via an ideal (desired) closed-loop model $\mathsf{c}^{\mathsf{i}}(b)$, $b\in\mathbf{B}$. Informally, the ideal assigns high values to desired behaviours and small values to undesired behaviours. With the ideal closed-loop model chosen, the FPD-optimal policy π^{FPD} makes $\mathsf{c}^{\pi^{\mathsf{FPD}}}$ closest to c^{i}. The FPD axiomatisation (Kárný and Kroupa, 2012) specifies widely-acceptable conditions under which the Kullback-Leibler divergence $\mathsf{D}(\mathsf{c}^\pi||\mathsf{c}^{\mathsf{i}})$, (Kullback and Leibler, 1951), is the adequate proximity measure. The FPD-optimal policy

π^{FPD} is thus

$$\begin{aligned} \pi^{\mathsf{FPD}} &\in \mathrm{Arg}\min_{\pi\in\mathbf{\Pi}} \mathsf{D}(\mathsf{c}^{\pi}||\mathsf{c}^{\mathsf{i}}) \\ &= \mathrm{Arg}\min_{\pi\in\mathbf{\Pi}} \sum_{b\in\mathbf{B}} \mathsf{c}^{\pi}(b)\ln\left(\frac{\mathsf{c}^{\pi}(b)}{\mathsf{c}^{\mathsf{i}}(b)}\right). \end{aligned} \tag{10}$$

Proposition 1 presented below describes the FPD-optimal decision rules. The proposition is a direct counterpart of stochastic DP (Åström, 1970; Bertsekas, 2001). It uses the chain-rule factorisation of c^{i}, which delimits: (a) the ideal environment model $\mathsf{p}^{\mathsf{i}}_t(\tilde{s}|a,s)$, which is the ideal counterpart of the transition probability $\mathsf{p}_t(\tilde{s}|a,s)$, and (b) the ideal decision rules $\pi^{\mathsf{i}}_t(a|s)$ of the ideal policy.

Proof of Proposition 1 is, for instance, in (Šindelář et al., 2008). The general FPD with the state estimation, corresponding to the partially observable MDP, is in (Kárný and Guy, 2006).

Proposition 1 (FPD-Optimal Policy)**.** *Decision rules $\pi^{\mathsf{FPD}}_t(a|s)$, $t \in \mathbf{T}$, forming the FPD-optimal policy (10) result from the following backward recursion, $t = |\mathbf{T}|, |\mathbf{T}|-1, \ldots, 1$, initiated by $\mathsf{w}_{|\mathbf{T}|}(s) = 1$, $\forall s \in \mathbf{S}$,*

$$\begin{aligned} \pi^{\mathsf{FPD}}_t(a|s) &= \frac{\pi^{\mathsf{i}}_t(a|s)\exp[-\omega_t(a,s)]}{\underbrace{\sum_{a\in\mathbf{A}} \pi^{\mathsf{i}}_t(a|s)\exp[-\omega_t(a,s)]}_{\mathsf{w}_{t-1}(s)}},\ s\in\mathbf{S}, \\ \omega_t(a,s) &= \sum_{\tilde{s}\in\mathbf{S}} \mathsf{p}_t(\tilde{s}|a,s)\ln\left(\frac{\mathsf{p}_t(\tilde{s}|a,s)}{\mathsf{p}^{\mathsf{i}}_t(\tilde{s}|a,s)\mathsf{w}_t(\tilde{s})}\right). \end{aligned} \tag{11}$$

The work (Kárný and Kroupa, 2012) containing axiomatisation of FPD also proved that: (i) any Bayesian decision making can be arbitrarily well approximated by the FPD formulation (10); (ii) there are FPD tasks having no standard counterpart. In other words, FPD tasks represent the proper dense extension of Bayesian decision making. Here, we modify the constructive way in which this result was shown. The construction explicitly relates the standard MDP to the less usual FPD. Importantly, it serves the purpose of this paper. It shows how the MDP-optimal deterministic policy is arbitrarily-well approximated by the naturally explorative, FPD-optimal, randomised policy. The construction uses the standard notion of entropy H^{π} (Cover and Thomas, 1991) of the closed-loop model c^{π} and the given cumulative reward R

$$\begin{aligned} \mathsf{H}^{\pi} &= -\sum_{b\in\mathbf{B}} \mathsf{c}^{\pi}(b)\ln(\mathsf{c}^{\pi}(b)) \\ \mathsf{R}(b) &= \sum_{t\in\mathbf{T}} \mathsf{r}_t(s_t,a_t,s_{t-1}),\ b\in\mathbf{B}. \end{aligned} \tag{12}$$

Proposition 2 (FPD from MDP)**.** *The optimisation (2) over policies $\pi\in\mathbf{\Pi}$ (3), restricted by the additional requirement, determined by an optional $h>0$,*

$$\mathsf{H}^{\pi} \geq h > \mathsf{H}^{\pi^{\mathsf{MDP}}}, \tag{13}$$

leads to the FPD-optimal policy (10) with respect to the ideal closed-loop model[2]

$$\mathsf{c}^{\mathsf{i}}(b) \propto \exp[\mathsf{R}(b)/\lambda]. \tag{14}$$

The corresponding ideal environment model and the ideal decision rules are

$$\begin{aligned} \mathsf{p}^{\mathsf{i}}_t(\tilde{s}|a,s) &\propto \exp[\mathsf{r}_t(\tilde{s},a,s)/\lambda] \\ \pi^{\mathsf{i}}_t(a|s) &\propto \sum_{\tilde{s}\in\mathbf{S}} \exp[\mathsf{r}_t(\tilde{s},a,s)/\lambda]. \end{aligned} \tag{15}$$

The optional bound h in (13) determines the scalar parameter $\lambda = \lambda(h) > 0$ and

$$\lim_{h\to\mathsf{H}^{\pi^{\mathsf{MDP}}}} \lambda(h) = 0. \tag{16}$$

Proof. It can be directly verified that any policy, which replaces some deterministic rules of the policy π^{MDP} by randomised ones has a higher entropy. Thus, when maximising the expected accumulated reward (2), under the inequality constraint (13), the constraint becomes active. The maximisation, equivalent to the negative-reward minimisation, reduces to the unconstrained minimisation of the Kuhn-Tucker functional (Kuhn and Tucker, 1951), given by the multiplier $\lambda = \lambda(h) > 0$,

$$\begin{aligned} \pi^{\mathsf{FPD}} &\in \mathrm{Arg}\min_{\pi\in\mathbf{\Pi}} \sum_{b\in\mathbf{B}} \mathsf{c}^{\pi}(b)[-\mathsf{R}(b) + \lambda\ln(\mathsf{c}^{\pi}(b))] \\ &= \mathrm{Arg}\min_{\pi\in\mathbf{\Pi}} \sum_{b\in\mathbf{B}} \mathsf{c}^{\pi}(b)\ln\left(\frac{\mathsf{c}^{\pi}(b)}{\exp[\mathsf{R}(b)/\lambda]}\right) \\ &= \mathrm{Arg}\min_{\pi\in\mathbf{\Pi}} \mathsf{D}(\mathsf{c}^{\pi}||\mathsf{c}^{\mathsf{i}}). \end{aligned}$$

The additive form of the cumulative reward (12), standard conditioning and marginalisation imply the forms of the ideal factors (15). The limiting property (16) corresponds with the relaxation of the constraint (13). □

Remarks

✓ The role of the *ideal* decision rule (15) differs from the closely-related Bolzmann's machine, which uses the decision rules

$$\pi_t(a|s) \propto \exp\Big(\sum_{\tilde{s}\in\mathbf{S}} \mathsf{r}_t(\tilde{s},a,s)\mathsf{p}_t(\tilde{s}|a,s)/\lambda\Big),\quad \lambda > 0. \tag{17}$$

[2] $\propto$ means proportionality.

✓ The original, less general, relation of FPD and MDP (Kárný and Kroupa, 2012) led to the ideal closed-loop model c^{iorig} that exploited the environment model $\mathsf{p} = \prod_{t\in\mathbf{T}} \mathsf{p}_t$

$$\mathsf{c}^{iorig}(b) \propto \mathsf{p}(b)\exp[\mathsf{R}(b)/\lambda] \overbrace{=}^{(14)} \mathsf{p}(b)\mathsf{c}^{\mathsf{i}}(b),\ b\in\mathbf{B}. \tag{18}$$

Recovering the explorative nature of the certainty-equivalent MDP-optimal policy is the main reason for employing the constraint (13). The following accounting of the influence of the incomplete knowledge on resulting policy brings an additional insight into the exploration problem. Primarily, it guides the adaptive choice of $\lambda = \lambda(h) > 0$ parameterising the ideal closed-loop model (14).

The policy $\pi^{\mathsf{MDP}}(\theta)$, which maximises the expected cumulative reward while using a given parameter $\theta \in \mathbf{\Theta}$, consists of the MDP-optimal deterministic rules

$$\begin{aligned} \pi_t^{\mathsf{MDP}}(a|s,\theta) &= 1 \text{ if } a = a_t^{\mathsf{MDP}}(s,\theta) \\ \pi_t^{\mathsf{MDP}}(a|s,\theta) &= 0 \text{ otherwise.} \end{aligned} \tag{19}$$

There $a_t^{\mathsf{MDP}}(s,\theta)$ is the maximising argument in the tth step of DP (4) modified by the explicit conditioning on $\theta \in \mathbf{\Theta}$. The decision rules[3] $\pi_t^{\mathsf{FPD}}(a|s,V,\lambda)$ of the constructed approximation of the FPD-optimal policy should approximate the policy $\pi^{\mathsf{MDP}}(\theta)$ made of the rules (19) exploiting also the knowledge of the parameter θ

$$\pi^{\mathsf{MDP}}(\theta) = (\pi_t^{\mathsf{MDP}}(a_t|s_{t-1},\theta))_{t\in\mathbf{T}},\ a_t\in\mathbf{A},\ s_{t-1}\in\mathbf{S}.$$

The approximate policy has the posterior probability $\mathsf{p}(\theta|s,V)$ as the only information about $\theta \in \mathbf{\Theta}$.

Works (Bernardo, 1979; Kárný and Guy, 2012) imply that the expected Kullback-Leibler divergence of $\pi^{\mathsf{MDP}}(\theta)$ from π^{FPD} is the adequate proximity measure to be minimised by

$$\pi^{\mathsf{FPD}} = (\pi_t^{\mathsf{FPD}}(a_t|s_{t-1},V_{t-1},\lambda_t^{\mathsf{FPD}}))_{t\in\mathbf{T}},\ a_t\in\mathbf{A}$$

via the adequately chosen $\lambda^{\mathsf{FPD}} = \lambda^{\mathsf{FPD}}(s_{t-1},V_{t-1})$. This dictates the selection

$$\lambda^{\mathsf{FPD}}(s_{t-1},V_{t-1}) \in \operatorname*{Arg\,min}_{\lambda>0} \int_{\mathbf{\Theta}} \sum_{a\in\mathbf{A}} \pi_t^{\mathsf{MDP}}(a|s_{t-1},\theta) \tag{20}$$

$$\times \ln\left(\frac{\pi_t^{\mathsf{MDP}}(a|s_{t-1},\theta)}{\pi_t^{\mathsf{FPD}}(a|s_{t-1},V_{t-1},\lambda)}\right)\mathsf{p}(\theta|s_{t-1},V_{t-1})\,\mathrm{d}\theta.$$

The optimal actions $a_t^{\mathsf{MDP}}(s,\theta)$ depend on the parameter $\theta \in \mathbf{\Theta}$ in a quite complex way. This makes us to solve (20) for greedy (one-stage-ahead) FPD. Importantly, the resulting ideal factors with the frozen $\lambda^{\mathsf{FPD}} = \lambda^{\mathsf{FPD}}(s_{t-1},V_{t-1})$ (15) are used in the multi-step policy design. Thus, the dynamic nature of the policy design is not compromised unlike in the wide-spread solutions of the exploration problem (Wu et al., 2017).

For choosing $\lambda^{\mathsf{FPD}}(s,V)$, at the observed $s = s_{t-1}$ and given $V = V_{t-1}$, let us define, cf. (2), (6),

$$\begin{aligned} \mathbf{\Theta}_a \;=\; & \Big\{\theta\in\mathbf{\Theta}: \sum_{\tilde{s}\in\mathbf{S}} \mathsf{r}_t(\tilde{s},a,s)\theta(\tilde{s}|a,s) \\ & \geq \sum_{\tilde{s}\in\mathbf{S}} \mathsf{r}_t(\tilde{s},\tilde{a},s)\theta(\tilde{s}|\tilde{a},s),\ \forall \tilde{a}\in\mathbf{A}\Big\},\ \forall a\in\mathbf{A}. \end{aligned} \tag{21}$$

On $\mathbf{\Theta}_a$, the action $a = a^{\mathsf{MDP}}(\theta)$ is optimal. For the FPD-optimal greedy decision rule (11) and the ideal factors (15), the optimisation (20) reads[4] $\lambda^{\mathsf{FPD}}(s,V)$

$$\begin{aligned} &\in \operatorname*{Arg\,min}_{\lambda>0} \sum_{a\in\mathbf{A}} \mathsf{p}(\mathbf{\Theta}_a|s,V)\Bigg[-\bar{r}(a,s,V)/\lambda - \mathsf{H}(a,s,V) \\ &+ \ln\Big(\sum_{\tilde{a}\in\mathbf{A}} \exp\big(+\bar{\mathsf{r}}_t(\tilde{a},s,V)/\lambda + \mathsf{H}(\tilde{a},s,V)\big)\Big)\Bigg] \\ &\bar{\mathsf{r}}_t(a,s,V) = \sum_{\tilde{s}\in\mathbf{S}} \mathsf{r}_t(\tilde{s},a,s)\mathsf{p}(\tilde{s}|a,s,V), \\ &\mathsf{H}(a,s,V) = -\sum_{\tilde{s}\in\mathbf{S}} \mathsf{p}_t(\tilde{s}|a,s,V)\ln(\mathsf{p}_t(\tilde{s}|a,s,V)), \\ &\mathsf{p}(\mathbf{\Theta}_a|s,V) = \int_{\mathbf{\Theta}_a} \mathsf{p}(\theta|s,V)\,\mathrm{d}\theta. \end{aligned} \tag{22}$$

Numerical solution of the scalar minimisation (22) is simple and can be done by any off-the-shelf software. The evaluation of probabilities $\mathsf{p}(\mathbf{\Theta}_a|s,V)$ (22) of the sets $\mathbf{\Theta}_a$, $a\in\mathbf{A}$ (21) is the only more involved step. Even it can be made by a direct Monte Carlo integration without excessive demands on its precision.

3 EXPERIMENTS

This part provides a representative sample of made Monte Carlo studies.

The simulated environment corresponded to MDP with $|\mathbf{S}| = 10$ possible states and $|\mathbf{A}| = 5$ possible actions. These options balanced the wish to deal with a non-trivial example and to perform extensive Monte Carlo experiments within a reasonable time even in the experimental Matlab implementation. Numerical values of the time-invariant simulated environment model p and of the time-invariant reward r are in Appendix.

[3] The dependence on λ is stressed by the condition.

[4] In experiments, λ^{FPD} was also optimised for the original ideal closed-loop model (18). Then, λ^{FPD} minimises an appropriate analogy of (22).

Table 1: The compared policies: CE is the certainty equivalent version of the policy and model means the environment model.

Label	Characterisation	Reference
DPknownPar	MDP, known model	(1), (2), (4)
DP	MPD, learnt model, CE	(2), (4), (7), (8)
FPD	FPD, learnt model, CE, former ideal (18) given λ	(11), (18), (7), (8)
FPDAdaptive	FPD, learnt model, CE, former ideal (18) adapting λ	(11), (18), (7), (8), (22)
FPDExp	FPD, learnt model, CE, proposed ideal (14) given λ	(11), (14), (7), (8)
FPDExpAdaptive	FPD, learnt model, CE, proposed ideal (14) adapting λ	(11), (14), (7), (8), (22)
Boltzmann	Greedy MDP, the learnt model, CE Boltzmann's machine, learnt model, given λ	(7), (8), (17)
eps-Greedy	Greedy MDP, learnt model, CE, uniform noise injected with probability $\varepsilon = 0.3$	(Vermorel and Mohri, 2005)
UCB1	Greedy MDP, learnt model, CE, noise tuned according an upper confidence bound	(Auer et al., 2002; Tang and et al, 2017)

The considered number of epochs was $|\mathbf{T}| = 10 << |\mathbf{\Theta}| \approx |\mathbf{S}|^2 \times |\mathbf{A}| = 500$. As already said, the proper balancing of exploration with exploitation is vital under the conditions of this type.

The compared policies are summarised in Table 1, which provides their labels, under which they are referred to in the figures. The table briefly characterises them and refers to their detailed descriptions.

Policies depending on a fixed λ were judged on the uniform grid

$$\lambda \in \{0.15, 0.20, 0.25, \ldots, 3.60\}. \qquad (23)$$

The policy quality was quantified by the sample mean (referred to as the average profit) of sampled cumulative rewards R (12) evaluated for 10^5 Monte Carlo runs. Preliminary experiments verified that this number is more than sufficient to guarantee the representability of the results.

Results showing that the exploration is not necessarily helpful are in Figure 1 with abbreviations referring to labels in Table 1. They were obtained within *the first experiment* where the corresponding environment model and the reward are described in Table 2. The policy DPknownPar, designed under the complete knowledge, reached the average profit of 72.11. Its variance $\sigma = 51.72$ quantifies its volatility. Straight lines correspond to policies independent of λ varied on the considered grid (23).

The results in which exploration was significant, were gained within *the second experiment*. They are summarised in Figure 2 with abbreviations again referring to labels in Table 1. The corresponding environment model and reward are described in Table 3. The policy DPknownPar, designed under the complete knowledge, reached the average profit of 62.84 and variance $\sigma = 149.78$.

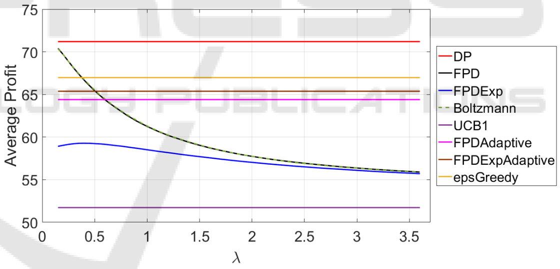

Figure 1: *The results of the first experiment*. The average profit is the sample mean of cumulative rewards (12) for the compared policies, Table 1, and different λ values on the grid (23).

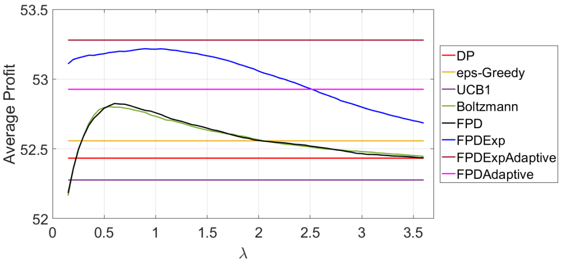

Figure 2: *The results of the second experiment*. The average profit is the sample mean of cumulative rewards (12) for compared policies, Table 1, and different λ values on the grid (23).

Discussion starts with stressing that the inspected small number of epochs $|\mathbf{T}|$ respects that the exploration-exploitation balance is vital in this case. Otherwise, even a rare adding of random deviations from exploitive actions, whose non-optimal character with respect to the exploitation has negligible influence, guarantees convergence of learning and thus the policy optimality. This distinguishes our experiments from usual tests, e.g. (Ouyang et al., 2017), and makes them relevant.

The experiments dealt with structurally same static DM. The numerical choice of their parameters was based on the following, qualitatively obvious fact. The need for exploration (within the considered short-horizon scenario) depends on the mutual relation of the prior probability $\mathsf{p}(\theta|V_0)$, see Table 4, and the parameter $\theta_{simulated}$ of the simulated environment model determining transition probability, see Tables 2, 3. The influence of this relation is enhanced or attenuated by the considered reward r.

The first experiment, reflected in Figure 1, in which the DP policy is the best one warns that exploration need not be always helpful. Notably, FPD and Boltzmann's machine with sufficiently small λ can be arbitrarily close to its best behaviour. Due to the lack of exploration significance no other conclusions concerning the quality of the tested policies can be made. But it calls for an improvement of λ-tuning, which should converge to zero if the exploration is superfluous.

The second experiment, reflected in Figure 2, is more informative. The policy based on the newly proposed relation of FPD with MDP and an adaptive choice of λ (FPDExpAdaptive) brings the highest improvement (about 2%). A similar performance can be reached for a fixed but properly chosen λ (FPDExp). The adaptive FPD is worse (FPDAdaptive) but still outperforms the remaining competitors. The similarity of the results for the λ-dependent FPD and Boltzmann's machine supports the conjecture that the performance of Boltzmann's machine can be improved by adapting λ. This may be important in its other applications.

4 CONCLUDING REMARKS

The paper has arisen from inspecting the conjecture that the certainty-equivalent version of non-traditional fully probabilistic design (FPD) of decision policies properly balances exploitation with exploration. The achieved results support it. Moreover the paper: (a) established a better relation of FPD to the wide-spread Markov decision processes; (b) proposed an adaptive tuning of the involved parameter, which can be used in the closely-related simulated annealing and Boltzmann's machine; (c) provided a sample of extensive experiments, which confirmed that standard exploration techniques are outperformed by the FPD-based policies.

The future work will concern: (i) an algorithmic recognition of cases in which exploration is unnecessary; (ii) inspection of a tuning mechanism based on extremum-seeking control; (iii) an efficient implementation of λ-tuning; (iv) application of the proposed ideas to continuous-valued MDP; (v) real-life problems, especially those in which a short, but non-unit, decision horizon is vital as in environmental decision making (Springborn, 2014).

REFERENCES

Åström, K. (1970). *Introduction to Stochastic Control*. Academic Press, NY.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, NY.

Bellman, R. (1961). *Adaptive Control Processes*. Princeton U. Press, NJ.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, NY.

Bernardo, J. (1979). Expected information as expected utility. *The An. of Stat.*, 7(3):686–690.

Bertsekas, D. (2001). *Dynamic Programming and Optimal Control*. Athena Scientific, US.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley. 2nd edition.

Črepinšek, M., Liu, S., and Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Survey*, 45(3):37–44.

Duff, M. O. (2002). *Optimal Learning; Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst.

Feldbaum, A. (1960,61). Theory of dual control. *Autom. Remote Control*, 21,22(9,2).

Gómez, A. G. V. and Kappen, H. (2012). Dynamic policy programming. *The J. of Machine Learning Research*, 30:3207–3245.

Guan, P., Raginsky, M., and Willett, R. (2012). Online Markov decision processes with Kullback-Leibler control cost. In *American Control Conference*, pages 1388–1393. IEEE.

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

Kárný, M. and Guy, T. (2012). On support of imperfect Bayesian participants. In Guy, T. and et al, editors, *Decision Making with Imperfect Decision Makers*, volume 28. Springer, Berlin. Intelligent Systems Reference Library.

Kárný, M. and Guy, T. V. (2006). Fully probabilistic control design. *Systems & Control Letters*, 55(4):259–265.

Kárný, M. and Kroupa, T. (2012). Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1):105–113.

Kuhn, H. and Tucker, A. (1951). Nonlinear programming. In *Proc. of 2nd Berkeley Symposium*, pages 481–492. Univ. of California Press.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–87.

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

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown Markov decision processes: A Thompson sampling approach. In et al, I. G., editor, *Advances in Neural Information Processing Systems 30*, pages 1333–1342. Curran Associates, Inc.

Puterman, M. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.

Šindelář, J., Vajda, I., and Kárný, M. (2008). Stochastic control optimal in the Kullback sense. *Kybernetika*, 44(1):53–60.

Springborn, M. (2014). Risk aversion and adaptive management: Insights from a multi-armed bandit model of invasive species risk. *Journal of Environmental Economics and Management*, 68:226–242.

Tang, H. and et al (2017). #Exploration: A study of count-based exploration for deep reinforcement learning. In et al, I. G., editor, *Advances in Neural Information Processing Systems 30*, pages 2753–2762. Curran Associates, Inc.

Tanner, M. (1993). *Tools for statistical inference*. Springer Verlag, NY.

Ullrich, M. (1964). Optimum control of some stochastic systems. In *Preprints of the VIII-th conference ETAN*. Beograd.

Vermorel, J. and Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer.

Witten, I., Frank, E., Hall, M., and Pal, C. (2017). *Data Mining: Practical machine learning tools and techniques*. 4th edition, Elsevier.

Wu, H., Guo, X., and Liu, X. (2017). Adaptive exploration-exploitation trade off for opportunistic bandits. preprint arXiv:1709.04004.

APPENDIX

This section provides considered rewards and transition probabilities used in experiments, Section 3. Static, time-invariant cases are considered. Their transition probabilities $\mathsf{p}(\tilde{s}|a,s) = \mathsf{p}(\tilde{s}|a)$ modelling the environment and rewards $\mathsf{r}(\tilde{s},a,s) = \mathsf{r}(\tilde{s},a)$ determining the cumulative reward (12) are the same $\forall s \in \mathbf{S}$.

Table 2: The data used in the first experiment. Explicit values of the reward $\mathsf{r}_t(\tilde{s},a,s) = \mathsf{r}(\tilde{s},a)$, on the left-hand side and of the transition probabilities $\mathsf{p}_t(\tilde{s}|a,s) = \mathsf{r}(\tilde{s},a)$ on the right-hand side. They are constant $\forall s \in \mathbf{S},\ t \in \mathbf{T}$ and $|\mathbf{S}| = 10$, $|\mathbf{A}| = 5$. Rows and columns correspond to states $\tilde{s} \in \mathbf{S}$ and actions $a \in \mathbf{A}$, respectively.

	The reward r_t					The transition probability p				
	actions $a \in \mathbf{A}$					actions $a \in \mathbf{A}$				
states $\tilde{s} \in \mathbf{S}$	5	7	6	5	10	0.12	0.16	0.12	0.12	0.08
	1	6	1	3	6	0.02	0.13	0.08	0.02	0.02
	6	2	5	7	9	0.08	0.16	0.06	0.14	0.15
	5	6	1	5	4	0.18	0.04	0.08	0.08	0.13
	5	2	2	6	6	0.10	0.06	0.18	0.10	0.06
	4	8	6	4	5	0.02	0.10	0.16	0.10	0.09
	3	9	3	8	5	0.06	0.07	0.08	0.08	0.13
	7	5	2	6	8	0.02	0.02	0.02	0.12	0.13
	3	9	3	2	6	0.20	0.18	0.16	0.20	0.04
	3	1	4	8	10	0.20	0.09	0.06	0.04	0.17

Table 3: The data used in the second experiment. Explicit values of the reward $\mathsf{r}_t(\tilde{s},a,s) = \mathsf{r}(\tilde{s},a)$, on the left-hand side and of the transition probabilities $\mathsf{p}_t(\tilde{s}|a,s) = \mathsf{r}(\tilde{s},a)$ on the right-hand side. They are constant $\forall s \in \mathbf{S},\ t \in \mathbf{T}$ and $|\mathbf{S}| = 10$, $|\mathbf{A}| = 5$. Rows and columns correspond to states $\tilde{s} \in \mathbf{S}$ and actions $a \in \mathbf{A}$, respectively.

	The reward r_t					The transition probability p				
	actions $a \in \mathbf{A}$					actions $a \in \mathbf{A}$				
states $\tilde{s} \in \mathbf{S}$	1	1	1	1	1	0.03	0.05	0.03	0.02	0.08
	2	2	2	2	2	0.05	0.07	0.09	0.05	0.05
	3	3	3	3	3	0.08	0.12	0.14	0.07	0.08
	3	3	3	3	3	0.08	0.07	0.09	0.07	0.05
	5	5	5	5	5	0.11	0.17	0.11	0.12	0.11
	6	6	6	6	6	0.29	0.31	0.20	0.15	0.30
	12	12	12	12	12	0.13	0.07	0.09	0.29	0.16
	4	4	4	4	4	0.11	0.05	0.11	0.10	0.08
	3	3	3	3	3	0.08	0.07	0.09	0.07	0.05
	2	2	2	2	2	0.05	0.02	0.06	0.05	0.03

Table 4: The occurrence array V_0 determining the prior probability $\mathsf{p}(\theta|V_0)$ (7) for the first experiment on the left-hand side and for the second experiment on the right-hand side. The occurrence arrays are constant $\forall s \in \mathbf{S}$. Rows and columns correspond to states $\tilde{s} \in \mathbf{S}$, $|\mathbf{S}| = 10$, and actions $a \in \mathbf{A}$, $|\mathbf{A}| = 5$.

	The first experiment					The second experiment				
	actions $a \in \mathbf{A}$					actions $a \in \mathbf{A}$				
states $\tilde{s} \in \mathbf{S}$	0.1	0.1	0.1	0.1	0.1	0.03	0.04	0.02	0.06	0.08
	0.1	0.1	0.1	0.1	0.1	0.05	0.06	0.05	0.09	0.05
	0.1	0.1	0.1	0.1	0.1	0.08	0.11	0.07	0.09	0.08
	0.1	0.1	0.1	0.1	0.1	0.08	0.06	0.07	0.09	0.05
	0.1	0.1	0.1	0.1	0.1	0.11	0.15	0.10	0.11	0.11
	0.1	0.1	0.1	0.1	0.1	0.29	0.19	0.36	0.26	0.30
	0.1	0.1	0.1	0.1	0.1	0.13	0.06	0.12	0.14	0.16
	0.1	0.1	0.1	0.1	0.1	0.11	0.11	0.10	0.06	0.08
	0.1	0.1	0.1	0.1	0.1	0.08	0.09	0.07	0.06	0.05
	0.1	0.1	0.1	0.1	0.1	0.05	0.13	0.05	0.06	0.03