IJMLC manuscript No.
(will be inserted by the editor)

Towards On-Line Tuning of Adaptive-Agent's Multivariate Meta-Parameter

Miroslav Kárný

Received: March 11, 2021/ Accepted: date

Abstract A decision-making (DM) agent models its environment and quantifies its DM preferences. An *adaptive agent* models them locally nearby the realisation of the behaviour of the closed DM loop. Due to this, a simple tool set often suffices for solving complex dynamic DM tasks. The inspected Bayesian agent relies on a unified learning and optimisation framework, which works well when tailored by making a range of case-specific options. Many of them can be made off-line. These options concern the sets of involved variables, the knowledge and preference elicitation, structure estimation, etc. Still, some meta-parameters need an on-line choice. This concerns, for instance, a weight balancing exploration with exploitation, a weight reflecting agent's willingness to cooperate, a discounting factor, etc. Such options influence, often vitally, DM quality and their adaptive tuning is needed. Specific ways exist, for instance, a data-dependent choice of a forgetting factor serving to tracking of parameter changes. A general methodology is, however, missing. The paper opens a pathway to it. The solution uses a *hierarchical feedback exploiting a generic, DM-related, observable, mismodelling indicator*. The paper presents and justifies the theoretical concept, outlines and illustrates its use.

Keywords Bayesian learning · Adaptive agent · Meta-parameter tuning · Fully probabilistic design · Kullback-Leibler divergence · Dynamic decision making

1 Introduction

The paper concerns a prescriptive tuning of meta-parameters[1] of adaptive agents solving dynamic DM tasks. The survey, Hospedales et al. (2020), confirms that meta-tuning is a hot topic in machine learning and recalls how much were done. This

The Czech Academy of Sciences, Institute of Information Theory and Automation, 182 00 Prague 8, Czech Republic, https://www.utia.cas.cz/people/karny, E-mail: school@utia.cas.cz

[1] The prefix "meta" marks a task about a task, DM about DM, an option about an option, etc. Note that all abbreviations are summarised in Table 2 at the paper end.

induces natural reader's questions: *Q1: Why should I go through another paper on this topic? Q2: What I will gain? Q3: What is novel in it? Q4: What is useful in it?*

The possible answers are:

Q1 The paper approaches the common meta-parameter choice in a non-standard way. It proposes an on-line tuning that exploits a novel, *generic, predictable and observable quality indicator* of the closed DM loop. It aligns with the spreading awareness that machine learning ultimately serves to a DM, Schweighofer and Doya (2003), Ghavamzadeh et al. (2015). Many insightful experts shift accordingly their attention. For instance, they combine the stream and on-line processing with the quest for a balanced exploration and exploitation, Klenske and Hennig (2016), or a care about the closed-loop stability, Beckenbach et al. (2020), etc. All need DM-tailored meta-tuning.

Q2 The offered meta-tuning relies on the axiomatic DM theory called fully probabilistic design of decision strategies (FDP), Kárný and Kroupa (2012), Kárný (2020a). It is worth to be aware of FPD as it strictly extends the usual maximisation of an expected reward, Savage (1954), Puterman (2005). FPD unifies *probabilistic modelling* of environments, decision strategies, *and*, unusually, *DM preferences*. This provides a new way to meta-tuning and other tasks too, e.g. Quinn et al. (2016).

Q3 The key novelties are: a) a meta-parameter tuning based on the minimisation of a *mismodelling indicator that respects the solved DM problem*; b) a refined analysis of *Bayesian-estimation asymptotic*; c) a use of the novel preference-elicitation principle, Kárný and Guy (2019), *stoping an infinite regress* of hierarchies needing to tune a meta-meta-parameter of the meta-level DM, etc.

Q4 Users of advanced algorithms in machine learning, control, computational statistics, and other domains know how hard the choice of proper parameters can be. The problem is repeatedly addressed, usually, in a domain-dependent way or for a specific algorithmic class, Duvenaud (2014). No universal solution exists. No free lunch theorem, Wolpert and Macready (1997), confirms that it cannot exist. It does not preclude a search for quite general solution ways. The paper offers one. It uses weak assumptions and a generic methodology that *does not rely on big informative data*. Its *narrowing down to a specific case follows from the solved DM problem*.

The outlined gains, hopefully, make the effort spent on non-standard notions and notations worthwhile. The paper could be quite attractive to readers searching for interesting research problems worth of their inventive abilities.

2 Technical Introduction

Agents are building blocks of distributed artificial intelligence, Sandholm (1999), of cyber-physical systems, Bogdan and Pedram (2018), of industry 4.0, Liao et al. (2017), etc. Their demanding tuning makes attractive the adaptive-control concept, Åström and Wittenmark (1994). The dreamt generic, feasible, self-tuning controller, however, never materialised. Almost always an important (multivariate) meta-parameter has to be chosen. Numerous ways exist but they are case-dependent and they demand a

substantial deliberation effort, Ishii et al. (2002). The paper contributes to a change of this state. It adds a relatively universal hierarchical feedback to an adaptive agent, Fig. 1. This feedback selects the meta-parameter so that the influence of the DM-specific *mismodelling error*, inherent to the local modelling, *is adaptively minimised*.

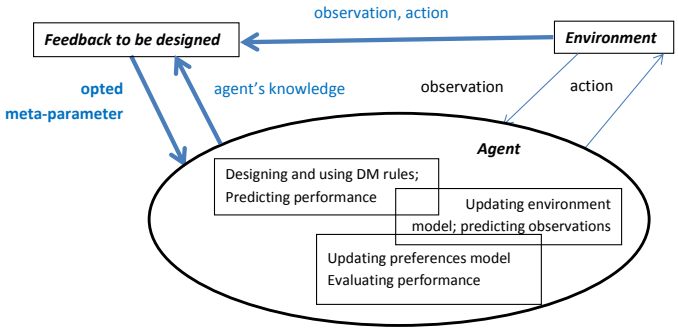

Fig. 1: The addressed task and its solution via a meta-level feedback.
The bold (blue) arrows and "***Feedback to be designed***" reflect the solved meta-DM.

2.1 Guide

Sec. 3.1 recalls the used theory of dynamic DM. Sec. 3.2 provides examples of tuned meta-parameters. They motivate the meta-tuning assumptions adopted in Sec. 3.3. Sec. 3.4 splits the DM optimality criterion into an $\mathscr{E}$-term, influenced by estimation, and an $\mathscr{M}$-term, reflecting mismodelling. Sec. 4 proofs that Bayesian estimation cares about the $\mathscr{E}$-term. The analysis refines known results. The $\mathscr{M}$-term reflects *observable* mismodelling effects on DM quality. It enables the feedback design sketched in Fig. 1. It is proposed in Sec. 5 by: a) interpreting the tuned parameter as meta-action, Sec. 5.1; b) a black-box modelling at meta-level, Sec. 5.2; and c) quantifying meta-preferences for diminishing the mismodelling impact, Sec. 5.3. Sec. 6 offers an illustrative experiment. Sec. 7 adds closing remarks. The text focuses on the addressed problem. It gives up generality, just samples related works, and tries to be concise. Comments and examples, Sec. 3.2, primarily point to challenging research tasks.

2.2 Notation, Agreements and Assumptions

$\{x\}$ denotes a set of x's. Its detailed description is given only if needed. $\equiv$ is defining equality, $\propto$ marks proportionality and $'$ denotes transposition. Sanserif fonts denote mappings. They are mostly probability densities (pd), their existence is assumed, Rao (1987). The time index is dropped if mapping arguments have it. The dependence on the decision horizon is made explicit only when needed. Mnemonic labels are preferred. The agents deal with the observed closed-DM-loop behaviours. Those using stochastic filtering, Jazwinski (1970), are left aside to focus on the key paper ideas.

3 Preliminaries

The recall of the used DM theory makes the paper self-reliant. The meta-parameter examples lead to the assumptions of our on-line tuning. The section also reveals the mismodelling impact on DM quality. It is of an independent interest.

3.1 Decision Making Under Uncertainty in Nutshell

The agent selects and uses a sequence of DM rules $(\mathsf{r}_t)_{t\in\{t\}}$. It does it up to a horizon $h \leq \infty$ at discrete time moments labelled by $t \in \{t\} \equiv \{1,\ldots,h\}$. The DM rule r_t (randomly) maps the agent's knowledge $k^{t-1} \in \{k^{t-1}\}$ about the closed DM loop on the opted action $a_t \in \{a\}$. The closed DM loop consists of the agent and its environment. An observation $o_t \in \{o\}$, made after applying the action a_t, enriches the agent's knowledge[2]. The knowledge k^t at time $t \in \{t\}$ is

$$\begin{aligned} k^t &\equiv ((o_\tau,a_\tau)_{\tau=1}^t,k^0) = (o_t,a_t,k^{t-1}) \in \{k^t\} \\ k^0 &\equiv \text{the used prior knowledge.} \end{aligned} \tag{1}$$

The agent has only a partial information about the impact of its actions on the closed DM loop. It selects them using its belief about their influence. The agent has to model the final knowledge k^h as a multivariate random variable and to express its belief via a joint pd $\mathsf{j} \equiv (\mathsf{j}(k^h))_{k^h\in\{k^h\}}$, Savage (1954). The pd j serves as the closed-DM-loop model, Ullrich (1964). It depends on the environment model and the used DM rules.

The agent specifies the desired behaviour of the closed DM loop. The employed fully probabilistic design of decision strategies (FPD), Kárný and Kroupa (2012), uses the ideal pd $\mathsf{j}_\mathsf{i} \equiv (\mathsf{j}_\mathsf{i}(k^h))_{k^h\in\{k^h\}}$ as the desiderata descriptor. The agent sets high values $\mathsf{j}_\mathsf{i}(k^h)$ of the ideal pd j_i to the desired $k^h \in \{k^h\}$ and low values to the undesired ones. The FPD-optimal DM rules r_o minimise Kullback-Leibler divergence (KLD[3]), Kullback and Leibler (1951), $\mathsf{D}(\mathsf{j}||\mathsf{j}_\mathsf{i})$ of the joint pd j to the ideal pd j_i

$$\begin{aligned} \mathsf{r}_\mathsf{o} &\in \operatorname{Arg}\min_{\mathsf{r}\in\{\mathsf{r}\}} \mathsf{D}(\mathsf{j}||\mathsf{j}_\mathsf{i}) \\ \mathsf{D}(\mathsf{j}||\mathsf{j}_\mathsf{i}) &\equiv \int_{\{k^h\}} \mathsf{j}(k^h) \ln\left(\frac{\mathsf{j}(k^h)}{\mathsf{j}_\mathsf{i}(k^h)}\right) \mathrm{d}k^h \equiv \mathsf{E}\left[\ln\left(\frac{\mathsf{j}}{\mathsf{j}_\mathsf{i}}\right)\right]. \end{aligned} \tag{2}$$

The chain rule for pds, Peterka (1981), and (1) explicate the dependence of the closed-DM-loop model $\mathsf{j}(k^h)$ on the environment model $\mathsf{m}(k^h) \equiv \prod_{t\in\{t\}} \mathsf{m}(o_t|a_t,k^{t-1})$ and on the DM rules $\mathsf{r}(a_h,k^{h-1}) \equiv \prod_{t\in\{t\}} \mathsf{r}(a_t|k^{t-1})$. It holds

$$\begin{aligned} \mathsf{j}(k^h) &= \prod_{t\in\{t\}} \mathsf{j}(o_t|a_t,k^{t-1})\mathsf{j}(a_t|k^{t-1}) \\ &\equiv \prod_{t\in\{t\}} \mathsf{m}(o_t|a_t,k^{t-1}) \times \prod_{t\in\{t\}} \mathsf{r}(a_t|k^{t-1}) \equiv \mathsf{m}(k^h) \times \mathsf{r}(a_h,k^{h-1}). \end{aligned} \tag{3}$$

[2] The agent's prior knowledge k^0 implicitly conditions all pds involved. The knowledge k^t is also called information state. $(o_t,a_t)_{t\in\{t\}}$ is often referred as (closed DM loop) trajectory or the observed behaviour.

[3] KLD, formerly called cross-entropy, Shore and Johnson (1980), now relative entropy, is the *DM-rules-dependent* expectation of the loss $\ln(\mathsf{j}/\mathsf{j}_\mathsf{i})$.

Formula (3): a) renames factors of the chain-rule factorisation in mnemonic way; b) assumes that actions influence the environment response irrespectively of the rules generating them. The ideal closed-DM-loop model $\mathsf{j}_\mathsf{i} = \mathsf{j}_\mathsf{i}(k^h)$ factorises similarly

$$\mathsf{j}_\mathsf{i}(k^h) \equiv \prod_{t\in\{t\}} \mathsf{m}_\mathsf{i}(o_t|a_t,k^{t-1}) \times \prod_{t\in\{t\}} \mathsf{r}_\mathsf{i}(a_t|k^{t-1}) \equiv \mathsf{m}_\mathsf{i}(k^h) \times \mathsf{r}_\mathsf{i}(a_h,k^{h-1}). \quad (4)$$

The next proposition, proved, for instance, in Kárný et al. (2006), yields the FPD-optimal DM rules. It also relates FPD to the standard minimisation of expected loss.

Proposition 1 (FPD-Optimal DM Rules) *The backward functional recursion, performed on functions acting on $k^t \in \{k^t\}$, run for $t = h,\dots,1$ and initiated by $\mathsf{d}(k^h) = 1$, provides the unique FPD-optimal DM rules r_o (2)*

$$\begin{aligned}
\mathsf{n}(a_t,k^{t-1}) &\equiv \int_{\{o\}} \mathsf{m}(o_t|a_t,k^{t-1}) \ln\left(\frac{\mathsf{m}(o_t|a_t,k^{t-1})}{\mathsf{d}(o_t,a_t,k^{t-1})\mathsf{m}_\mathsf{i}(o_t|a_t,k^{t-1})}\right) \mathrm{d}o_t \\
\mathsf{d}(k^{t-1}) &\equiv \int_{\{a\}} \mathsf{r}_\mathsf{i}(a_t|k^{t-1}) \exp(-\mathsf{n}(a_t,k^{t-1}))\mathrm{d}a_t \\
\mathsf{r}_\mathsf{o}(a_t|k^{t-1}) &= \mathsf{r}_\mathsf{i}(a_t|k^{t-1}) \frac{\exp(-\mathsf{n}(a_t,k^{t-1}))}{\mathsf{d}(k^{t-1})}.
\end{aligned}$$

When equating the ideal DM rules r_i (4) to the opted DM rules r (leave-to-the fate option, Kárný et al. (2006)), the optimisation reduces to the stochastic dynamic programming, Bertsekas (2017), for the additive loss with the tth summand, $t \in \{t\}$,

$$\mathsf{L}(o_t,a_t,k^{t-1}) \equiv \ln\left(\frac{\mathsf{m}(o_t|a_t,k^{t-1})}{\mathsf{m}_\mathsf{i}(o_t|a_t,k^{t-1})}\right). \quad (5)$$

The dynamic programming is the functional recursion for value functions on $\{k^t\}$, run for $t = h,\dots,1$ and initiated by $\mathsf{v}(k^h) = 0$. The optimal value function v meets

$$\mathsf{v}(k^{t-1}) = \min_{a_t\in\{a\}} \int_{\{o\}} \mathsf{m}(o_t|a_t,k^{t-1})[\mathsf{L}(o_t,a_t,k^{t-1}) + \mathsf{v}(k^t)]\mathrm{d}o_t. \quad (6)$$

The deterministic *optimal DM rules choose a minimiser in (6).*

On Prop. 1:

- ✓ Markov decision process (MDP), Puterman (2005), is a frequent DM framework. It is covered by FPD with the leave-to-the-fate option for: a) the Markov environment model $\mathsf{m}(o_t|a_t,k^{t-1}) = \mathsf{m}(o_t|a_t,o_{t-1})$; and b) the ideal environment, see (5)

$$\mathsf{m}_\mathsf{i}(o_t|a_t,o_{t-1}) \propto \mathsf{m}(o_t|a_t,o_{t-1}) \exp[-\mathsf{L}(o_t,a_t,o_{t-1})], \quad (7)$$

  where $\mathsf{L}(o_t,a_t,o_{t-1})$ is the loss supplied by the agent[4].
- ✓ KLD regularisation of the loss in MDP, Guan et al. (2014), Kárný and Kroupa (2012), Larsson et al. (2017), leads to the randomised optimal decision rules

$$\mathsf{r}_\mathsf{o} \in \mathrm{Arg}\min_{\mathsf{r}\in\{\mathsf{r}\}} \mathsf{E}\Big[\sum_{t\in\{t\}} \mathsf{L}(o_t,a_t,o_{t-1}) + A\ln\Big(\frac{\mathsf{r}(a_t|o_{t-1})}{\mathsf{r}_\mathsf{i}(a_t|o_{t-1})}\Big)\Big]. \quad (8)$$

[4] The usual MDP deals with the reward $-\mathsf{L}$ and maximises its expectation.

There E is the expectation with respect to decision-rules-dependent joint pd (3). The optimisation (8) exploits a given reference (ideal) decision rule r_i with its support including the set of possible actions $\{a\}$. The optional regularisation weight A is positive. The solution of (8) coincides with the FPD-optimal DM rules for the ideal pd given by the chosen r_i and by

$$\mathsf{m}_\mathsf{i}(o_t|a_t,o_{t-1}) \propto \mathsf{m}(o_t|a_t,o_{t-1})\exp[-\mathsf{L}(o_t,a_t,o_{t-1})/A].$$

- ✓ The rules r_o (8) concentrate on the deterministic optima of MDP when relaxing the regularisation, when $A \to 0^+$. The meta-parameter A determines a sort of soft-min operation. Sec. 3.2 touches its role in exploitation-exploration dichotomy.
- ✓ Bayesian paradigm deals with pds and its key learning mechanism, Bayes' rule, is a functional recursion. This causes its complexity. It is counteracted by the use of local models converting (approximately) Bayes' rule into an algebraic update. The purposeful DM-driven local modelling pays back in this respect.
- ✓ The complexity of decision-rules design stems (mainly) from the functional nature of dynamic programming that evolves *value function*. FPD is also described by the functional recursion, which is (surprisingly?) simpler. Instead of repetitive (minimisation, expectation) of dynamic programming, it deals with a sequence of expectations only. Thus, the approximate dynamic programming, Si et al. (2004), or neurodynamic programming, Bertsekas (2017), are expected to be simpler within the FPD set-up.
- ✓ FPD operates on solely pds. It is a version of Bayesian optimisation. The orientation of FPD on *dynamic* DM extends usual Bayesian optimisations that mostly deal with complex but *static* tasks, e.g. Kandasamy et al. (2015).

3.2 DM Tasks Motivating Meta-Tuning Assumptions

This part outlines DM tasks with an optional meta-parameter, generically denoted[5] A. The examples are biased to the instances we dealt with. They allowed us to extract desirable properties of a "universal" opting mechanism and tailor assumptions accordingly. The other numerous published cases are just sampled at the section end.

Exploration in FPD The second part of Prop. 1 makes FPD equivalent to MDP

$$\mathsf{r}_\mathsf{o} \in \operatorname{Arg}\min_{\mathsf{r}\in\{\mathsf{r}\}} \int_{\{k^h\}} \mathsf{j}(k^h) \sum_{t\in\{t\}} \mathsf{L}(o_t,a_t,k^{t-1})\mathrm{d}k^h$$

with a real-valued loss L. Adaptive agents combine DM with parameter estimation. A qualitative inspection implies that the *optimal infeasible DM rules* have dual exploitation-exploration nature, Feldbaum (1961). Certainty-equivalent DM rules, Jacobs and Patchell (1972), Åström and Wittenmark (1994), mostly serve as an approximation of the infeasible optimal rules. MDP's certainty-equivalent deterministic DM rules are not explorative and fail with a positive probability, Kumar (1985). On the other hand, certainty-equivalent FPD rules are explorative, Lee et al. (2019).

[5] This reflects its interpretation as a meta-action at the upper-level feedback, cf. Fig. 1 and Sec. 5.

As said, the non-explorative MDP is a limiting case of FPD with the loss-related ideal joint pd j_i given by the meta-parameter $A > 0$. The paper, Kárný and Hůla (2019), focuses on the FPD duality. It uses another ideal pd connecting FPD and MDP

$$\mathsf{j}_\mathsf{i}(k^h) = \prod_{t\in\{t\}} \frac{\exp\left[-\mathsf{L}(o_t,a_t,o_{t-1})/A\right]}{\int_{\{o,a\}} \exp\left[-\mathsf{L}(o_t,a_t,o_{t-1})/A\right] \mathrm{d}o_t \mathrm{d}a_t}. \tag{9}$$

The opted A in (9) balances exploitation and exploration efforts of the certainty-equivalent version of FPD. It is a non-trivial example of the tuned meta-parameter[6].

The certainty-equivalent FPD with the ideal pd j_i (9) mimics the optimal design giving the dual DM rules. This supports the conjecture that an optimal value $A_\mathfrak{o}$ exists. The meta-parameter $A > 0$ in (9), resembling the temperature in simulating annealing, Tanner (1993), guarantees that for $A \to 0^+$ the FPD-optimal DM rules provide the MDP-optimal actions. Thus, the FPD-focused meta-tuning also cares about the exploration-exploitation balance of the MDP-based adaptive agents.

Discounting Factor MDP often uses the discounted loss $\sum_{t\in\{t\}} A^t \mathsf{L}(o_t,a_t,o_{t-1})$. The discounting factor $A \in (0,1)$ makes the loss in a distant future less important. It corresponds with a monetary interpretation of the factor A. Even in the economical domain, its systematic choice is still questionable, Doyle (2013). Generally, the discounting factor A reflects doubts about persistency of the employed belief and aim descriptions. Notably, the effective shortening of the design horizon brought by discounting may cause instability of the closed DM loop, Gaitsgory et al. (2018). The need to adapt A thus arises. An optimal compromise $A_\mathfrak{o}$ between too short-sighted optimisation and the damaging uncertainty level surely exists.

Meta-Parameter in Tracking Loss In tracking, the observation o_t is to follow a given ideal trajectory $o_{t;\mathsf{i}}$, $t \in \{t\}$, Tao (2014). An elimination of abrupt action changes, that may destabilise the closed DM loop by exciting modelling errors, Rohrs et al. (1982), is the key requirement. The next loss quantifies this

$$\mathsf{L}(k^h) \equiv \sum_{t\in\{t\}} \text{distance of } o_t \text{ to } o_{t;\mathsf{i}} + A^2 \times \text{ norm of } (a_t - a_{t-1}). \tag{10}$$

The loss (10) depends on the weight $A^2 \geq 0$. The distance of the observation o_t and its ideal (desired) value $o_{t;\mathsf{i}}$ can be parameterised. Entries of the action increments $(a_t - a_{t-1})$ may be individually weighted in multivariate cases. Then, the open problem of the choice of weighting matrices in linear-quadratic tracking is faced, Kumar et al. (2014). Again, the existence of a fixed optimal meta-parameter $A_\mathfrak{o}$ is expected.

Trust in Predictive Pd Serving to Estimation[7] Bayesian estimation operates on parametric models

$$\mathsf{m}_\theta(k^h) \equiv \prod_{t\in\{t\}} \mathsf{m}_\theta(o_t|a_t,k^{t-1}),\ \theta \in \{\theta\}.$$

[6] The same choice is faced when dealing with usual exploration techniques, Ouyang et al. (2017).

[7] The term trust has narrower meaning than numerous studies focused on it, Li and Song (2016).

Bayes' rule[8] updates the posterior pd $\mathsf{p}^t_\theta \equiv \mathsf{p}(\theta|k^t)$ of an unknown parameter $\theta \in \{\theta\}$

$$\mathsf{p}^t_\theta \propto \mathsf{m}_\theta(o_t|a_t,k^{t-1})\mathsf{p}^{t-1}_\theta \tag{11}$$

starting from a prior pd $\mathsf{p}^0_\theta \equiv \mathsf{p}(\theta|k^0)$.

The predictive pd $\mathsf{m}(o_t|a_t,k^{t-1})$ then serves as the environment model. It reads

$$\mathsf{m}(o_t|a_t,k^{t-1}) = \int_{\{\theta\}} \mathsf{m}_\theta(o_t|a_t,k^{t-1})\mathsf{p}^{t-1}_\theta \mathrm{d}\theta. \tag{12}$$

Kracík and Kárný (2005) and Quinn et al. (2016) generalised Bayes' rule. They assume that an external predictive pd $\mathsf{e}(o_t|a_t,k^{t-1})$ is at disposal instead of an observed pair (o_t,a_t). The predictive pd $\mathsf{e}(o_t|a_t,k^{t-1})$ updates the posterior pd via the recipe

$$\mathsf{p}^t_\theta \propto \exp\left[A\int_{\{o\}} \mathsf{e}(o_t|a_t,k^{t-1})\ln(\mathsf{m}_\theta(o_t|a_t,k^{t-1}))\mathrm{d}o_t\right]\mathsf{p}^{t-1}_\theta,\ \ A\in[0,1]. \tag{13}$$

The formula (13) reduces to Bayes' rule (11) if the weight $A=1$ and the external predictive pd $\mathsf{e}(o_t|a_t,k^{t-1})$ shrinks on an observed o_t.

The meta-parameter $A\in[0,1]$ expresses the tunable trust into the external predictive pd $\mathsf{e}(o_t|a_t,k^{t-1})$. The trust choice strongly influences the processing (13). This makes its tuning desirable. The trust level to the source providing the external predictive pd $\mathsf{e}(o_t|a_t,k^{t-1})$ is usually stable. Then, a fixed, optimal trust-weight $A_\mathfrak{o}$ exists.

Mixing Weight Within a soft cooperation of FPD-agents, an agent α combines its ideal pd $\mathsf{j}^\alpha_\mathfrak{i}$ with the ideal pd $\mathsf{j}^\beta_\mathfrak{i}$ of an agent β. The convex combination

$$\mathsf{j}^A_\mathfrak{i} = A\mathsf{j}^\alpha_\mathfrak{i} + (1-A)\mathsf{j}^\beta_\mathfrak{i},\qquad A\in[0,1], \tag{14}$$

serves the agent α to the design of the FPD-optimal rules $\mathsf{r}^A_\mathfrak{o}$. This may respect, possibly opposite, aims of the agent β. The mixing meta-parameter A in (14) quantifies the degree of this respect. It strongly influences the achieved DM quality. The agent α selfishly measures it by the KLD $\mathsf{D}(\mathsf{j}^A_\mathfrak{o}||\mathsf{j}^\alpha_\mathfrak{i})$. The case-dependent, on-line tuning of the mixing weight A converts this academic solution into a quite practical cooperation way. The cooperation of heaters in adjacent rooms, taken as a working example in Kárný and Alizadeh (2019)[9], indicates this well. It also supports the hypothesis on the generic existence of the optimal weight $A_\mathfrak{o}$.

[8] This form of Bayes' rule is valid for the considered DM rules for which the parameter pointing to the "best" model, Berec and Kárný (1997), is unknown, cf. natural conditions of control in Peterka (1981).

[9] Extensive references on the whole approach can be found in the cited paper. The chapter, Dietrich and List (2016), is a good starting point to pooling problems that are in the core of such a cooperation.

Action Period Continuous-time signals exploited in DM are periodically sampled. The action a_t holds in the real-time interval $[At, A(t+1))$ given by a period $A > 0$.

Technology (sampling and computing rates, capacity of the involved information channels, time constants of actuators, etc.) determines a lower bound $\underline{A}$ on the period A. The smaller A is the richer is the knowledge k^{t-1} available for the a_t choice. At the same time, the period shortening calls for more complex models. It increases computational costs. The information gain of a short sampling is finally erased by the impact of modelling errors. This reveals the known fact that the DM-optimal action period[10] $A_o > \underline{A}$ exists.

To our best knowledge no algorithmic choice of the vital *action* period exists. Attempts like, Kárný (1991), are unsatisfactory and just rules-of-the-thumb are applied.

The action period enters the environment model unlike other discussed meta-parameters. In fact, this dependence is avoidable. It suffices to sample data with the smallest technically feasible period and optimise under the constraint that actions may change only with a larger optional period A, Peterka (1991). The meta-parameter then enters DM rules and the model operates on observations "averaged" over the action period. A sort of local data filter, say based on spline modelling of underlying observations, Guy and Kárný (2000), becomes relevant.

The said is important as *we shall assume that no meta-parameter enters the environment model in on-line mode and Bayes' rule suffices to its parameter estimation.*

Other Samples Multitude of meta-options include:

- ✓ the optional rates that drive Hebbian's learning, Hebb (2005);
- ✓ the step size within numerical optimisation, Yang et al. (2019);
- ✓ the receding horizon that influences DM quality, Mayne (2014);
- ✓ a meta-parameter controlling reinforcement learning, Schweighofer and Doya (2003), say related to emotions, Moerland et al. (2018), or a meta-parameter in its parametric version, Kober and Peters (2011);
- ✓ a multivariate meta-parameter in Gaussian-process-based learning, Duvenaud (2014);
- ✓ the regularisation weight in LASSO-type learning Diebold and Shin (2019);
- ✓ ...

3.3 Adopted Meta-Tuning Assumptions

The assumptions, supported by examples of the previous section, which are respected by our generic solution, are:

- ✓ the meta-parameter A may enter the ideal environment model m_i, the ideal decision rule r_i and indirectly the decision rule r but not the environment model m;
- ✓ the influence of the meta-parameter A on closed DM loop is significant;
- ✓ the optimal meta-parameter value A_o exists;
- ✓ the optimal meta-parameter value A_o may vary at most slowly;
- ✓ the influence of the A-choice on the closed DM loop is smooth but too complex to be described in a detail.

[10] In this context, Shannon's sampling theorem, Shannon (1948), provides no guide.

3.4 Mismodelling

This part recognises an observable, DM-related, mismodelling indicator that suits to the feedback, which tunes the meta-parameter, see Fig. 1.

The discussion deals with a factual counterpart $\mathsf{f}(k^h) \equiv \prod_{t\in\{t\}} \mathsf{f}(o_t|a_t,k^{t-1})$ of the environment model, $\mathsf{m}(k^h)$. The *factual pd* f objectively describes the environment. It means that it serves to all agents operating on the same set $\{k^h\}$, Kárný and Kroupa (2012). The joint pd

$$\mathsf{c}(k^h) \equiv \mathsf{f}(k^h) \times \mathsf{r}(a_h,k^{h-1}) = \prod_{t\in\{t\}} \mathsf{f}(o_t|a_t,k^{t-1}) \times \prod_{t\in\{t\}} \mathsf{r}(a_t|k^{t-1}) \tag{15}$$

models the closed DM loop formed by the *unknown factual environment model* and by the used DM rules. With it, the KLD expressing the truly reached DM quality algebraically decomposes

$$\mathsf{D}(\mathsf{c}||\mathsf{j}_\mathfrak{i}) = \mathsf{D}(\mathsf{c}||\mathsf{j}) + \int_{\{k^h\}} \mathsf{c}(k^h) \ln\left(\frac{\mathsf{j}(k^h)}{\mathsf{j}_\mathfrak{i}(k^h)}\right) \mathrm{d}k^h \equiv \mathscr{E} + \mathscr{M}. \tag{16}$$

The first non-negative summand $\mathscr{E}$ measures the proximity of the factual closed-DM-loop model c (15) to the closed-DM-loop model j (3). It is called $\mathscr{E}$-term. The second summand $\mathscr{M}$ uses the environment model m gained by Bayesian learning and employed in the design of the optimal DM rules. Both the pd c and the pd j contain the same DM rules r. Sec. 4 indicates that the $\mathscr{E}$-term depends on the set $\{\mathsf{m}_\theta\}_{\theta\in\{\theta\}}$ of the parametric environment models but not on the tuned meta-parameter. The used DM rules influence it only weakly via their exploration and stabilisation abilities.

The second summand $\mathscr{M}$ in (16) expresses how well the DM aims are achieved on the data objectively described by the factual environment model and by the used decision rules. The $\mathscr{M}$-term is strongly influenced by the opted meta-parameter A and serves us for its tuning, see Sec. 5.

4 Bayesian Estimation Minimises the $\mathscr{E}$-Term

As said in Sec. 3.2, the environment model is the predictive pd (12) resulting from Bayesian estimation. It gradually updates a prior pd $\mathsf{p}^0_\theta = \mathsf{p}(\theta|k^0)$ to the posterior pd $\mathsf{p}^h_\theta = \mathsf{p}(\theta|k^h)$ given by the knowledge $k^h = ((o_t,a_t)_{t=1}^h,k^0)$ collected up to the horizon[11] h. Bayes' rule (11) provides the posterior pd $\mathsf{p}^h_\theta \propto \mathsf{m}_\theta(k^h)\mathsf{r}(a_h,k^{h-1})\mathsf{p}^0_\theta \propto \mathsf{m}_\theta(k^h)\mathsf{p}^0_\theta$. Its analysis shows how the $\mathscr{E}$-term in (16) behaves for $h\to\infty$.

Proposition 2 (Bayesian Estimation Asymptotics, $h\to\infty$) *Let*[12]*, for all h*

$$\emptyset \neq \{\theta\}_\cap \equiv \mathrm{supp}[\mathsf{p}^0_\theta] \cap \left\{\theta\in\{\theta\} : \mathsf{m}^h_\theta \mathsf{r}^h = 0 \Rightarrow \mathsf{c}^h = \mathsf{f}^h\mathsf{r}^h = 0 \ \textit{on}\ \{k^h\}\right\}. \tag{17}$$

$$\textit{Then,}\ \mathrm{supp}[\mathsf{p}^\infty_\theta] \subseteq \mathrm{Arg}\inf_{\theta\in\{\theta\}_\cap}[\mathsf{D}^\infty_\theta] \equiv \mathrm{Arg}\inf_{\theta\in\{\theta\}_\cap}\left[\lim_{h\to\infty}\frac{1}{h}\mathsf{D}\left(\mathsf{c}^h\middle|\middle|\mathsf{m}^h_\theta\mathsf{r}^h\right)\right]. \tag{18}$$

[11] The dependence of pds on the horizon h is made explicit here.

[12] For a pd s on $\{x\}$, its support $\mathrm{supp}[\mathsf{s}] \equiv \{x\in\{x\} : \mathsf{s}(x) > 0\}$.

Proof[13] Let us take a *fixed value* $\theta \in \{\theta\}_\cap$ (17) and define, on the set $\{k^h\}$, the ratio

$$\mathsf{g}_\theta^h \equiv \mathsf{g}_\theta(k^h) \equiv \frac{\mathsf{m}_\theta(k^h)\mathsf{r}(a_h,k^{h-1})}{\mathsf{c}(k^h)} = \frac{\mathsf{m}_\theta(k^h)}{\mathsf{f}(k^h)} = \prod_{t=1}^{h} \frac{\mathsf{m}_\theta(o_t|a_t,k^{t-1})}{\mathsf{f}(o_t|a_t,k^{t-1})}. \tag{19}$$

The ratio (19) is the integrable non-negative martingale, Doob (1953), with respect to (sigma algebra generated by) the knowledge k^h. The integrability is directly seen

$$\mathsf{E}[\mathsf{g}_\theta^h] \equiv \int_{\{k^h\}} \frac{\mathsf{m}_\theta(k^h)\mathsf{r}(a_h,k^{h-1})}{\mathsf{c}(k^h)}\mathsf{c}(k^h)\mathrm{d}k^h = \int_{\{k^h\}} \mathsf{m}_\theta(k^h)\mathsf{r}(a_h,k^{h-1})\mathrm{d}k^h = 1.$$

The martingale property also demonstrates directly

$$\mathsf{E}[\mathsf{g}_\theta^h|k^{h-1},\theta] \equiv \int_{\{o_h,a_h\}} \prod_{t=1}^{h} \frac{\mathsf{m}_\theta(o_t|a_t,k^{t-1})}{\mathsf{f}(o_t|a_t,k^{t-1})}\mathsf{f}(o_h|a_h,k^{h-1})\mathsf{r}(a_h|k^{h-1})\mathrm{d}o_h\mathrm{d}a_h$$

$$= \int_{\{o_h,a_h\}} \mathsf{m}_\theta(o_h|a_h,k^{h-1})\mathsf{r}(a_h|k^{h-1})\mathrm{d}o_h\mathrm{d}a_h \times \prod_{t=1}^{h-1} \frac{\mathsf{m}_\theta(o_t|a_t,k^{t-1})}{\mathsf{f}(o_t|a_t,k^{t-1})} = 1 \times \mathsf{g}_\theta^{h-1}.$$

This guarantees, Doob (1953), Prop. 4.1 (i), that $\lim_{h\to\infty} \mathsf{g}_\theta^h = \mathsf{g}_\theta^\infty$ exists with c-probability 1 on the set $\{k^\infty\}$. Moreover, with the same probability, it exists

$$\lim_{h\to\infty} \frac{1}{h} \ln(\mathsf{g}_\theta(k^h)) = -\mathsf{D}_\theta^\infty \le 0. \tag{20}$$

Indeed, the proved convergence g_θ^h implies that a limit of a smooth function of g_θ^h exists. Let us select arbitrary $\varepsilon > 0$ and inspect the probabilities

$$\mathsf{Prob}_{\theta\varepsilon}^h \equiv \int_{\{k^h:\frac{1}{h}\ln(\mathsf{g}_\theta(k^h))\ge\varepsilon\}} \mathsf{c}(k^h)\mathrm{d}k^h \le \int_{\{k^h:\mathsf{g}_\theta(k^h)\exp(-h\varepsilon)\ge 1\}} \mathsf{c}(k^h)\mathsf{g}_\theta(k^h)\exp(-h\varepsilon)\mathrm{d}k^h$$

$$\le \int_{\{k^h\}} \mathsf{c}(k^h)\frac{\mathsf{m}_\theta(k^h)\mathsf{r}(a_h,k^{h-1})}{\mathsf{c}(k^h)}\mathrm{d}k^h \times \exp(-h\varepsilon) = \exp(-h\varepsilon).$$

Thus, probabilities $\mathsf{Prob}_{\theta\varepsilon}^h$ of the sets $\{k^h : \frac{1}{h}\ln(\mathsf{g}_\theta(k^h)) \ge \varepsilon\}$ converge to zero for the extending horizon $h \to \infty$. Arbitrariness of $\varepsilon > 0$ implies non-positivity of the limit $\lim_{h\to\infty} \frac{1}{h}\ln(\mathsf{g}_\theta(k^h))$. It coincides with its expectation $-\mathsf{D}_\theta^\infty$.

The proved limit (20) implies that the deviations $\rho_\theta^h \equiv \mathsf{D}_\theta^\infty - \frac{1}{h}\ln(\mathsf{g}_\theta^h)$ converge to zero for the horizon $h \to \infty$ with c-probability 1. This together with the meaning of the proportionally $\propto$ allows the next expression of the pd p_θ^h at any fixed value $\theta \in \{\theta\}$

$$\mathsf{p}_\theta^h \propto p_\theta^0 \exp\left[-h\left(-\frac{\ln(\mathsf{g}_\theta^h)}{h})\right)\right] \propto \mathsf{p}_\theta^0 \exp\Big\{-h\underbrace{[(\overbrace{\mathsf{D}_\theta^\infty - \inf_{\theta\in\{\theta\}_\cap}[\mathsf{D}_\theta^\infty]}^{\equiv\gamma_\theta}) + \rho_\theta^h]}_{\equiv\delta_\theta^h}\Big\}. \tag{21}$$

For a value $\theta \notin \mathrm{Arg}\inf_{\theta\in\{\theta\}_\cap}[\mathsf{D}_\theta^\infty]$, the 1st summand γ_θ in the exponent of (21) is positive. Let us take any such θ and define $\varepsilon = \gamma_\theta/2$. Then, an $h_\theta \in (0,\infty)$ exists

[13] The proof tailors and refines results in Algoet and Cover (1988), Berec and Kárný (1997).

such that $\forall h > h_\theta$ it holds $\mathsf{abs}(\rho_\theta^h) < \varepsilon$ and thus $\delta_\theta^h > 0$ in the exponent of (21) $\forall h > h_\theta$. This implies $\exp\left\{-h\delta_\theta^h\right\} \to 0$ for $\theta \notin \mathrm{Arg}\inf_{\theta\in\{\theta\}_\cap}[\mathsf{D}_\theta^\infty]$, which gives (18). Assumption (17) excludes singularity $\mathsf{D}_\theta^\infty = \infty$, $\forall \theta \in \mathrm{supp}[\mathsf{p}_\theta^0] \cap \{\theta\}$. □

On Prop. 2:

- ✓ Informally, Prop. 2 states that the posterior pd asymptotically concentrates on models, which are nearest to the factual pd describing data-generating environment. KLD is the "adequate" proximity measure, cf. with Bernardo (1979) and Kárný and Guy (2012). These *best projections of reality on the considered model set* are equivalent. The equivalence set contains a single model if the chosen parametrisation is identifiable under the realised experimental conditions.
- ✓ A general discussion Prop. 2 that enlightens general properties of Bayesian estimation is in Berec and Kárný (1997). Among others, Prop. 2 implies:
  - if there is a "true" parameter value $\theta_\mathfrak{T} \in \{\theta\}_\cap$ such that the factual pd $\mathsf{f} = \mathsf{m}_{\theta_\mathfrak{T}}$ then the value $\theta_\mathfrak{T}$ is in the support of the asymptotic posterior pd p_θ^∞;
  - if, moreover, the model is identifiable with the used DM rules, i.e. if the asymptotic support contains a single parameter value, then the posterior pd concentrates on the "true" parameter value $\theta_\mathfrak{T}$.
- ✓ Formally, the existence of the "true" parameter $\theta_\mathfrak{T}$ means that is a function of k^∞. It implies: *if a consistent estimator of $\theta_\mathfrak{T}$ exists then Bayesian estimation is consistent*. Thus, the useful (in)consistency studies as, Grünwald and Langford (2007), in fact do not blame Bayesian framework but mismodelling.
- ✓ Prop. 2 and the predictive-pd form (12) imply that Bayesian estimation asymptotically minimises the $\mathscr{E}$-term.
- ✓ The same DM rules r enter the factual closed DM loop and the closed DM loop considered for their optimisation. The DM rules influence closed-loop stability and the rate with which the posterior pd delimits the best models (18).
- ✓ The FPD-optimal rules r_o, which are by their construction optimally explorative, Feldbaum (1961), Wu et al. (2017), are mostly infeasible. It is important that their used certainty-equivalent approximators, Klenske and Hennig (2016), are explorative too, Kárný and Hůla (2019). This guarantees a high learning rate and the smallness of the support $\mathrm{supp}[\mathsf{p}_\theta^\infty]$.
- ✓ The value $\inf_{\theta\in\{\theta\}_\cap}[\mathsf{D}_\theta^\infty]$ is generically positive as the factual closed loop c is out of the convex hull of the modelled closed loops $(\mathsf{m}_\theta\mathsf{r})_{\theta\in\{\theta\}}$. The value depends on the model set $(\mathsf{m}_\theta)_{\theta\in\{\theta\}}$. Its members have to meet (17), i.e. to assign a positive probability to factually probable data.
- ✓ The individual values D_θ^∞, $\theta \in \{\theta\}$, are unavailable as the factual pd f is unknown. They depend on the position of the factual pd f and of the off-line-chosen set of parametric models. This formally underlines importance of this choice, i.e. the importance of modelling.
- ✓ *In summary*, the $\mathscr{E}$-term is unsuitable for tuning of the optional meta-parameter $A \in \{A\}$. Importantly, ignoring $\mathscr{E}$-term in the meta-tuning makes no harm as the meta-parameter enters it via DM rules. The explorativeness and the (in)ability to stabilise the factual closed DM loop of DM rules only influence the $\mathscr{E}$-term.

5 Hierarchical Feedback Based on the $\mathscr{M}$-term

This section proposes the hierarchical feedback promised in Fig. 1. It tries to minimise mismodelling error as then the meta-parameter is appropriately tuned.

The feedback design relies on the adaptive framework by using a simple model estimated on-line via Bayes' rule. It avoids the trap of infinite regress, i.e. the need to select a meta-meta-parameter of the hierarchical feedback. In harmony with no-free-lunch theorem, Wolpert and Macready (1997), the solution exploits the specific structure of the addressed problem, see below. It relies on:

- ✓ the minimum cross-entropy principle, Shore and Johnson (1980), which serves to a justified construction of the meta-parametric model;
- ✓ an extension of this principle to preference elicitation, Kárný and Guy (2019), which provides the relevant ideal pd to FPD at the inspected hierarchical level.

To distinguish elements of DM at the meta-level of the hierarchy, they are denoted by capital counterparts of those used at the basic level. For instance, $O \in \{O\}, A \in \{A\}$, $\Theta \in \{\Theta\}$ denote observations, actions and parameters at the meta-level.

5.1 Meta-Parameter as a Meta-Action

Mismodelling means that the factual pd f is out of the convex hull of the parametric model $(\mathsf{m}_\theta)_{\theta\in\{\theta\}}$. This makes the factual closed loop different from the modelled one, $\mathsf{c} \neq \mathsf{j}$. Consequently, the FPD-optimal rules, r_o, computed for the learnt model, m, are not optimal for the factual model, f. The KLD value $\mathsf{D}(\mathsf{c}||\mathsf{j}_\mathsf{i})$ (16) grows with the mismodelling-reflecting $\mathscr{M}$-term. The deterioration depends on the meta-parameter $A_t \in \{A\}$ that is used in the design of the optimal DM rules r_o.

The meta-parameter A_t is opted at time $t \in \{t\}$, and may enter the ideal environment model m_i and the ideal decision rules r_i, see Sec. 3.3. As seen in Sec. 3.2, their dependencies on the meta-parameter A_t at time $t \in \{t\}$ may have the known functional forms. Their dependence on A_t may also be given numerically. This is the case of the roughly optimal DM decision rules, which use m_i, r_i depending on A_t. By adopted assumptions, Sec. 3.3, the meta-parameter A_t does not enter the parametric model. The notation $A_t \in \{A\}$ stresses that it is in fact the *additional* agent's (meta-)action. It has to be chosen by a causal randomised decision rule R_t, which realises the hierarchical feedback shown in Fig. 1.

In the used adaptive context, with the on-line-estimated environment model, the (sub)optimal DM rules r_o are designed in the receding-horizon mode, Mayne (2014), Mesbah (2018). They freeze the knowledge about the unknown parameter $\theta \in \{\theta\}$. This naturally applies to the meta-parameter, too.

Thus, the agent optimises the rules r_o, prescribing the actions $a_\tau \in \{a\}$, $\tau \in [t, t+h]$. The rules are designed for the environment model m, being the predictive pd given by the fixed posterior pd p_θ^{t-1}, and for the fixed meta-parameter sample $A_t \sim \mathsf{R}_t$. The (meta-)rule R_t is designed by FPD using a model M. The agent applies the action generated by the rule $\mathsf{r}_\mathsf{o}(a_t|A_t, k^{t-1})$ and the environment responds. Bayes' rule (11) and the related parametric pds fed into (12) provide the new models m, M. Then, the procedure repeats.

The choice of the meta-parameter A_t is made under the assumption that *there is an optimal, at most slowly varying, value* $A_{\mathfrak{o}} \in \{A\}$ for which the mismodelling error is the smallest one, Sec. 3.3. The on-line option A_t is thus a guess of the optimal meta-parameter $A_{\mathfrak{o}}$. $A_{\mathfrak{o}}$ is often low-dimensional and $\{A\}$ is partially bounded, cf. Sec. 3.2.

5.2 Modelling and Estimation at the Meta-Level

The $\mathscr{M}$-term (16) offers itself as the operational mean for checking the DM quality even under mismodelling. The $\mathscr{M}$-term is the *factual expected value* of the sum of

$$\eta(A_\tau, k^\tau) \equiv \ln\left(\frac{\mathsf{m}(o_\tau|a_\tau, k^{\tau-1})\mathsf{r}_{\mathfrak{o}}(a_\tau|A_\tau, k^{\tau-1})}{\mathsf{m}_{\mathfrak{i}}(o_\tau|a_\tau, A_\tau, k^{\tau-1})\mathsf{r}_{\mathfrak{i}}(a_\tau|A_\tau, k^{\tau-1})}\right) \tag{22}$$

and its smallest, *factually expected*, value is desirable.

The used receding-horizon design is to select h so that the dominant dynamic changes of the closed DM loop are covered, Mayne (2014). This implies that it suffices to select the action A_t, which only cares about the expectation of

$$O_t \equiv \frac{1}{h+1}\sum_{\tau=t-h}^{t} \eta(A_\tau, k^\tau) \text{ with } A_\tau = A_t. \tag{23}$$

The design needs to model how the (meta-)observation O_t depends on the (meta-)action A_t, and the accumulated knowledge k^{t-1}. The use of:

- ✓ zero-mean innovations $O_t - \mathsf{E}[O_t|A_t, k^{t-1}]$ uncorrelated with A_t, k^{t-1}, Peterka (1981), and
- ✓ the local linear expansion around the unknown time-invariant $A_{\mathfrak{o}}$, see Sec. 3.3, leads to the simple regression model for the scalar observation O_t (23)

$$O_t = \Theta_{0\{A\}}A_t + \Theta_{1\{A\}}A_{t-1} + \Theta_{\{O\}}O_{t-1} + \Theta_1 + E_t \equiv \Theta\Psi_t + E_t, \text{ where} \tag{24}$$

$\Theta \equiv [\Theta_{0\{A\}}, \Theta_{1\{A\}}, \Theta_{\{O\}}, \Theta_1]$ is the row vector of regression coefficients,

$\Psi_t \equiv \begin{bmatrix} A_t \\ A_{t-1} \\ O_{t-1} \\ 1 \end{bmatrix}$ is the regression vector with the column-wise ordered A_t.

E_t in (24) includes innovations, expansion errors and deviations caused by the fact that the observation O_t is evaluated for evolving actions $A_\tau \neq A_t$ in (23).

Θ_1 is the important offset that preserves zero mean of E_t.

$\Theta_{1\{A\}}A_{t-1}$, $\Theta_{\{O\}}O_{t-1}$ are delayed terms that compensate the correlations caused by the expansion. The first order auto-regression suffices to cope with the slow modelled dynamics, again see Sec. 3.3.

E_t is thus *zero mean sequence, uncorrelated* with A_t, k^{t-1}. Moreover, its variance $\Omega \in (0, \infty)$ is (approximately) constant.

Bayesian learning of the unknown Θ and Ω needs to specify the pd of E_t. With the listed properties, the minimum cross-entropy principle, Shore and Johnson (1980), leads to the Gaussian model of E_t. This gives the parametric (meta-)model

$$\mathsf{M}_{\Theta,\Omega}(O_t|A_t,k^{t-1}) = \mathsf{G}_{O_t}(\Theta\Psi_t,\Omega) \equiv \frac{\exp\left[-\frac{(O_t-\Theta\Psi_t)^2}{2\Omega}\right]}{\sqrt{2\pi\Omega}}. \tag{25}$$

The model (25) has the conjugated Gauss inverse-Gamma prior pd $\mathsf{P}^0_{\Theta,\Omega} = \mathsf{P}(\Theta,\Omega|k^0)$, Berger (1985). Bayes' rule reproduces its form. This reduces Bayesian learning to the algebraic updating of values of the finite-dimensional sufficient statistic.

The updating coincides with recursive least squares (RLS) initiated by values given by the prior pd. The statistic consist of degrees of freedom ν, the a priori increased amount of the used pairs (O,A), Peterka (1981), Kárný et al. (2006), and of

$$\begin{aligned}
\hat{\Theta} &\equiv \mathsf{E}[\Theta|k] = \text{least-squares estimate of } \Theta \\
\hat{\Omega} &\equiv \mathsf{E}[\Omega|k] = \frac{\text{least-squares reminder}}{\nu-2} \\
C &\equiv \frac{\text{covariance}[\Theta|k]}{\hat{\Omega}} = \text{least-squares covariance factor.}
\end{aligned} \tag{26}$$

The predictive pd $\mathsf{M}(O_t|A_t,k^{t-1})$ has Student's form. The adopted certainty-equivalent design approximates it by the Gaussian pd to get a simple model. The approximation principle, Bernardo (1979), reduces to the moment fitting in this case. The approximation is known to be tight for $\nu \approx 10$. This gives the (meta-)environment model $\mathsf{M}(O_t|A_t,k^{t-1}) = \mathsf{G}_{O_t}\left(\hat{\Theta}\Psi_t,\hat{\Omega}\right)$, cf. (23), (26).

Comments on Meta-Learning:

- ✓ The value of the prior sufficient statistic initiating RLS is a meta-meta-parameter. It can be well chosen in off-line mode, Kárný et al. (2014). Other options are fixed using the arguments outlined above. Thus, the infinite regress (each hierarchical level needs its meta-parameter) is avoided in the meta-estimation part.
- ✓ For the treated *scalar* observation O_t and for the action A_t of at most mild-dimension, RLS algorithm is computationally cheap. One step of robust, square-root version of RLS needs a small multiple of $[\text{length}(\Psi)]^2$ elementary operations Peterka (1975).
- ✓ RLS can robustly track slow variations of the estimated parameters Θ and Ω via forgetting, Kulhavý and Zarrop (1993). Even non-informativeness and abrupt changes can be well counteracted, Kárný (2020b).

5.3 Choice of the Ideal Pd and FPD Solution

The formulated aim, to choose the actions $A_t \in \{A\}$ so that $\mathsf{E}[O_t] = 0$, $t \in \{t\}$, incompletely determines the ideal pd $\mathsf{J}_\mathsf{i} = \mathsf{M}_\mathsf{i}\mathsf{R}_\mathsf{i}$ needed for the intended use of FPD. Recently, its optimal choice has been proposed, Kárný and Guy (2019). The solution is an analogy of the minimum cross-entropy principle, Shore and Johnson (1980). It

states that *having a set* $\{\mathsf{J_i}\}$ *of ideal pds compatible with the agent's aim, the optimal ideal joint pd* $\mathsf{J_{io}}$ *minimises the minimum found by FPD*

$$\mathsf{J_{io}} \in \mathrm{Arg}\min_{\mathsf{J_i}\in\{\mathsf{J_i}\}}\min_{\{\mathsf{R}\}}\mathsf{D}(\mathsf{J}||\mathsf{J_i}). \tag{27}$$

The example presented in Kárný and Guy (2019) covers exactly the situation faced here. For the Gaussian environment model and the wish $\mathsf{E}[O_t]=0$, the optimal ideal environment model, implied by (26) and (27), reads

$$\mathsf{M_{io}}(O_t|A_t,k^{t-1}) = \mathsf{G}_{O_t}(0,\hat{\Omega}). \tag{28}$$

The optimal ideal DM rule has the intuitively appealing form

$$\mathsf{R_{io}}(A_t|k^{t-1}) \propto \mathsf{M}(O_t=0|A_t,k^{t-1}) \propto \mathsf{G}_{O_t=0}(\hat{\Theta}\Psi_t,\hat{\Omega}). \tag{29}$$

The paper Kárný (1996), applying Prop. 1 to this linear-Gaussian case, shown that FPD provides the optimal DM rule

$$\mathsf{R_o}(A_t|k^{t-1}) = \mathsf{G}_{A_t}(-L[A_{t-1},O_{t-1},1]',\Omega_{\{a\}}). \tag{30}$$

The vector L and variance $\Omega_{\{a\}}$ are obtained by an algebraic evaluation of Riccati equation, known from the classical linear-quadratic (LQ) control, Meditch (1969).

Algorithmic Summary Algorithm 1 below describes the adaptive agent with the gained hierarchical feedback. It just summarises the derived relations that are cross-referred at its individual steps.

Comments on Meta-Design:

- ✓ The adopted preference-elicitation principle, Kárný and Guy (2019), allows us to avoid infinite regress also in the meta-design.
- ✓ Maximiser of the optimal decision rule $\mathsf{R_o}$ (30) coincides with the action of LQ controller, Meditch (1969). This MDP-type minimiser of the expected quadratic loss realises deterministic linear feedback. FPD-implied sampling around it cares about exploration. A cautious version, Peterka and Astrom (1973), can be also used. It respects uncertainty of point estimates replacing the unknown parameters.
- ✓ Complexity of a single step of robust square-root solution of Riccati equation coincides with that of RLS. A very small number of such steps is needed. Even one step per one estimation period may suffice with the strategy called iterations-spread-in time, Kárný et al. (1985).
- ✓ The tuning is elaborated for continuous-valued meta parameter A. Discrete-valued cases can be addressed quite similarly using a logistic or Markov-chain model M.
- ✓ *In summary*, the meta-feedback requires a small multiple of $[\mathrm{length}(\Psi)]^2$ operations and practically avoids infinite regress.

Algorithm 1 Conceptual Implementation of Scheme of Fig. 1

Inputs:

- Initial set-up of the solved DM problem:
  - ✓ the environment-describing parametric model m_θ and prior pd p^0_θ (11) with observation $\{o\}$, action $\{a\}$, parameter $\{\theta\}$ and meta-parameter $\{A\}$ sets
  - ✓ the DM-aims describing ideal pd j_i (4)
- Prior knowledge of the meta-design at time $t = 1$:
  - ✓ a guess of the horizon h (23) covering environment dynamics
  - ✓ prior statistic of the meta-model, ν, $\hat{\Theta}$, $\hat{\Omega}$, C (26)
  - ✓ the filled register $[\eta_{t-1},\dots,\eta_{t-h}]$, (22),
  - ✓ the regression vector $\Psi'_t = [A'_t, A'_{t-1}, O_{t-1}, 1]$ filled by prior guesses of meta-parameter $A_1, A_0 \in \{A\}$ of the meta-parameter to be tuned and of the scalar meta-observation O_0 (23)

Time (on-line) cycle:

- Perform h-step-ahead certainty-equivalent FPD (at basic level) with the frozen posterior pd p^{t-1}_θ and the frozen action A_t, i.e. apply Prop. 1
- Apply action $a_t \sim \mathsf{r}_\mathsf{o}(a_t|A_t, k^{t-1})$ and observe o_t on the environment
- Update the posterior pd p^{t-1}_θ to the pd p^t_θ via Bayes' rule (11)
- Evaluate the mismodelling indicator $\eta_t \equiv \ln\left(\frac{\mathsf{m}(o_t|a_t,k^{t-1})\mathsf{r}_\mathsf{o}(a_t|A_t,k^{t-1})}{\mathsf{m}_\mathsf{i}(o_t|a_t,k^{t-1})\mathsf{r}_\mathsf{i}(a_t|A_t,k^{t-1})}\right)$ (22)
- Set observation O_t (23) equal to sample mean of $h+1$ newest η_τ, (22).
- Update (meta-)posterior pd P^{t-1}_Θ by data (O_t, A_t) to the pd P^t_Θ, i.e. perform RLS step with the scalar observation O_t, and the regression vector $\Psi'_t \equiv [A'_t, A'_{t-1}, O_{t-1}, 1]'$; this updates statistics (26)
- Increment time $t \equiv t+1$
- Perform the certainty equivalent meta-level LQ design corresponding to the ideal environment model $\mathsf{M}_{\mathsf{i}\mathsf{o}}$ (28) and the ideal (meta-)rule $\mathsf{R}_{\mathsf{i}\mathsf{o}}(A_t|k^{t-1})$ (29)
- Sample the meta-action (meta-parameter) A from the optimal (meta-)rule R_o (30)

end of time cycle

Outputs:

- observations o_t, actions a_t, observations O_t, meta-parameters A_t, $t \in \{t\}$, learning and design results at both hierarchical levels

6 Illustrative Experiment

The experiment indicates contribution of the proposed meta-parameter tuning. It deals with the FDP version of *meta-parameter in tracking loss*, Sec. 3.2. It deliberately introduces a structural modelling error.

The simulated non-minimum-phase environment (31) cannot be well-handled by myoptic (greedy) decision rule Peterka (1972). Its high static gain make it sensitive to action amplitudes.

Simulation Set Up

The 2nd order linear Gaussian environment was simulated with observations

$$o_t \sim \mathsf{G}_{o_t}([1, 1.1, 0, 1.8, -0.81, 0][a_t, a_{t-1}, a_{t-2}, o_{t-1}, o_{t-2}, 1]', 1), \quad t \in \{t\}. \tag{31}$$

It is non-minimum phase linear system with the static gain 210 and the double real pole 0.9. The regulation problem, the tracking with the constant ideal trajectory $o_{t;\mathsf{i}} = 0$ (10), was solved. The results were gained for the receding horizon $h = 10$, which safely covers the environment dynamics.

The basic DM loop, Fig. 1, learnt the 1st order model in order to see the mismodelling influence. Algorithmically, the estimation run RLS, Sec. 5.2, Peterka (1981).

The regulation aim was given by the ideal joint pd $\mathsf{j_i}$. It consisted of the product of time-invariant ideal models $\mathsf{m_i}(o_t|a_t,k^{t-1}) = \mathsf{G}_{o_t}(0,\hat{\omega})$, Kárný (1996); Kárný and Guy (2019). The used estimate $\hat{\omega}$ of the observation variance $\omega_{\{o\}} = 1$ was gained when running RLS.

The product of time-invariant ideal DM rules

$$\mathsf{r_i}(a_t|k^{t-1}) = \mathsf{G}_{a_t}(a_{t-1},A^2) \tag{32}$$

expressed the wish to limit action changes, cf. (10). The optional *variance A^2 was the DM meta-parameter*. The meta-level feedback solved the same problem but without a meta-meta-parameter. It run RLS and LQ, operating on O_t, A_t, see Algorithm 1.

Simulation Results

The obvious contribution of the meta-level tuning is expressed numerically in Table 1, which contains sample statistics of the realised data.

Table 1: Sample statistics of observations o_t and actions a_t *without* and *with* the meta-level feedback. A fixed seed of the random generator makes the comparison fair.

case	variable	mean	median	minimum	maximum	2-norm
without	o_t	-8.677	-5.956	-282.710	255.030	1.734
meta-level	a_t	-0.011	-0.036	-20.000	20.000	0.274
with	o_t	2.434	1.843	-12.373	38.490	0.238
meta-level	a_t	0.028	-0.012	-20.000	20.000	0.125

Fig. 2 complements these numbers and shows typical simulation results *without* the hierarchical feedback. The meta-parameter in (32) was $A^2 = 1$.

Fig. 3 shows the results *with* the proposed hierarchical feedback. Its contribution is obvious when noticing rather different observation scale comparing to Fig. 2.

Fig. 4 provides an additional insight by showing time course of the meta-observation O_t (23). It is positive as expected. The course of the meta-parameter A_t demonstrates sensitivity of the problem to the proper choice of the ideal variance A^2 in (32). It suffices to notice that the tested fixed option $A^2 = 1$ is mostly not too distant from the varying values of A_t^2 but it produced much worse closed-DM-loop behaviour.

7 Concluding Remarks

The paper proposes a unified methodology of the meta-parameter tuning. It is applicable to a range of disparate DM tasks. It relies on a meta-feedback, Fig. 1, that tries to counteract mismodelling error while practically avoiding an infinite regress.

Sec. 1 already advertised the reported achievements. Thus, it remains to comment the preposition "Towards" in the title. It indicates that an open-ended research is presented. While the adopted conceptual building blocks — FPD, minimum cross-entropy principle and its preference elicitation counterpart — have firm theoretical bases, the full solution relies on heuristics steps. It uses certainty-equivalent, receding-horizon design of DM rules. The feeling that the theory is advanced enough and the

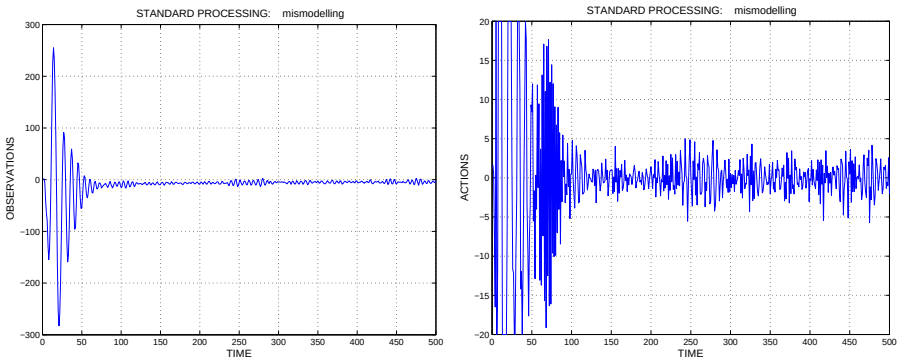

Fig. 2: Observed o_t, left, opted a_t, right, *without* a meta-level feedback. The observations and action increments should be ideally zero mean and have unit variance.

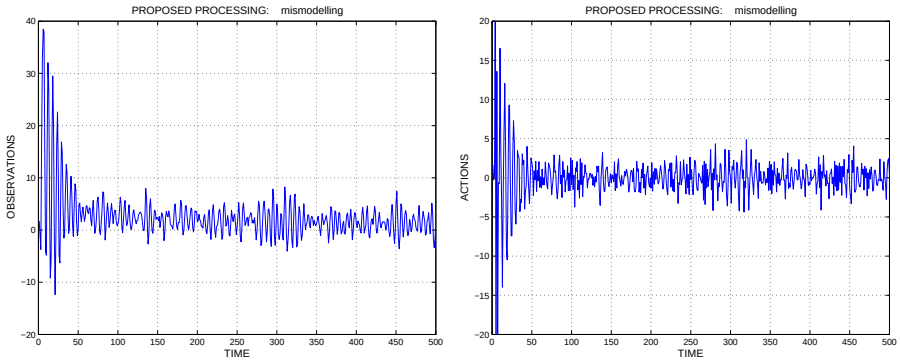

Fig. 3: Observed o_t, left, opted a_t, right, *with* the meta-level feedback. The observations should be ideally zero mean with unit variance. The action increments should be zero mean and have the variance tuned to counteract mismodelling.

wish to stimulate its thorough development and real-life tests have made us to present the solution as it is. Definitely, it is not panacea but it may help in a range up to now unsolved problems.

Implementation as well as evaluation costs connected with the additional feedback are low. Preliminary experience, represented by the illustrative example, confirms that the adopted approach is worth of further elaborating, ideally, by readers of this paper.

Declaration

Funding The reported research has been supported by MŠMT ČR LTC18075 and EU-COST Action CA16228.

Conflict of interests The author has no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

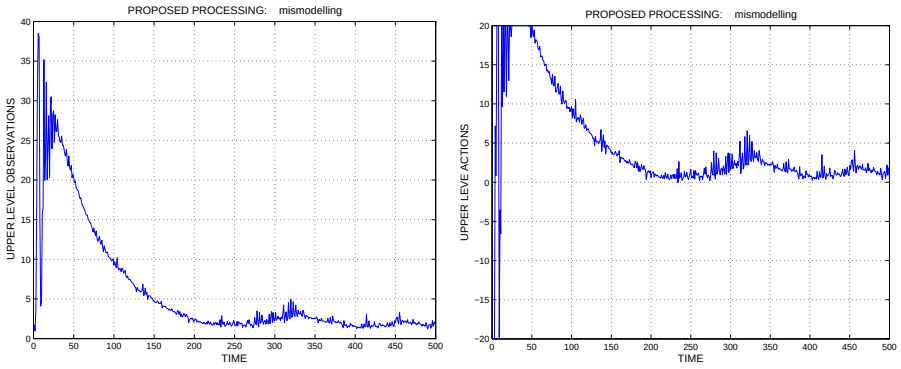

Fig. 4: Meta-observed O_t (23) reflecting mismodelling, left, meta-opted ideal standard deviation of action increments A_t, right.

Availability of data and material Not applicable

Code availability The source code of the example is available upon a request.

Table 2: Used Abbreviations

abbreviation	meaning	reference
DM	decision making	Berger (1985)
pd	probability density, Radon-Nikodým derivative	Rao (1987)
FPD	fully probabilistic design of DM strategies	Kárný and Kroupa (2012)
KLD	Kullback-Leibler divergence, also cross or relative entropy	Kullback and Leibler (1951) Shore and Johnson (1980)
MDP	Markov decision process	Puterman (2005)
LASSO	least absolute shrinkage and selection operator	Diebold and Shin (2019)
RLS	recursive least squares	Peterka (1981)
LQ	linear-quadratic control (design)	Meditch (1969)

References

Algoet P, Cover T (1988) A sandwich proof of the Shannon-McMillan-Breiman theorem. The Annals of Probability 16:899–909

Åström K, Wittenmark B (1994) Adaptive Control. Addison-Wesley, 2nd Edition

Beckenbach L, Osinenko P, Streif S (2020) A Q-learning predictive control scheme with guaranteed stability. European Journal of Control

Berec L, Kárný M (1997) Identification of reality in Bayesian context. In: Warwick K, Kárný M (eds) Computer-Intensive Methods in Control and Signal Processing, Birkhäuser, pp 181–193

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

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

Bertsekas D (2017) Dynamic Programming and Optimal Control. Athena Scientific
Bogdan P, Pedram M (2018) Toward enabling automated cognition and decision-making in complex cyber-physical systems. In: 2018 IEEE ISCAS, pp 1–4
Diebold F, Shin M (2019) Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. International Journal of Forecasting 35:1679–1691
Dietrich F, List C (2016) Probabilistic opinion pooling. In: Hajek A, Hitchcock C (eds) Oxford Handbook of Philosphy and Probability, Oxford University Press
Doob J (1953) Stochastic Processes. Wiley
Doyle J (2013) Survey of time preference, delay discounting models. Judgement and Decision Making 8:116–135
Duvenaud D (2014) Automatic model construction with Gaussian processes. PhD thesis, Pembroke College, Univ. of Cambridge
Feldbaum A (1961) Theory of dual control. Autom Remote Control 22:3–19
Gaitsgory V, Grüne L, Höger M, Kellett C, Weller S (2018) Stabilization of strictly dissipative discrete time systems with discounted optimal control. Automatica 93:311 – 320, DOI https://doi.org/10.1016/j.automatica.2018.03.076
Ghavamzadeh M, Mannor S, Pineau J, Tamar A (2015) Bayesian reinforcement learning: A survey. Foundations and Trends in Machine Learning 8(5–6):359 – 483, DOI 10.1561/2200000049
Grünwald P, Langford J (2007) Suboptimal behavior of Bayes and MDL in classification under misspecification. Machine Learning 66(2-3):119–149
Guan P, Raginsky M, Willett R (2014) Online Markov decision processes with Kullback Leibler control cost. IEEE Trans on AC 59(6):1423–1438
Guy TV, Kárný M (2000) Design of an adaptive controller of LQG type: spline-based approach. Kybernetika 36(2):255–262
Hebb D (2005) The Organization of Behavior: A Neuropsychological Theory. Taylor & Francis, URL https://books.google.cz/books?id=uyV5AgAAQBAJ
Hospedales T, Antoniou A, Micaelli P, Storkey A (2020) Meta-learning in neural networks: A survey ArXiv:2004.05439v1 [cs.LG] 11 Apr 2020
Ishii S, Yoshida W, Yoshimoto J (2002) Control of exploitation-exploration meta-parameter in reinforcement learning. Neural Networks 15(4-6):665–687
Jacobs O, Patchell J (1972) Caution and probing in stochastic control. Intern Journal of Control 16(1):189–199
Jazwinski A (1970) Stochastic Processes and Filtering Theory. Ac. Press
Kandasamy K, Schneider J, Póczo B (2015) High dimensional Bayesian optimisation and bandits via additive models. In: International Conference on Machine Learning, proceedings mlr.press
Kárný M (1991) Estimation of control period for selftuners. Automatica 27(2):339–348, extended version of the paper presented at 11th IFAC World Congr. , Tallinn
Kárný M (1996) Towards fully probabilistic control design. Automatica 32(12):1719–1722
Kárný M (2020a) Axiomatisation of fully probabilistic design revisited. Syst & Con Lett DOI 10.1016/j.sysconle.2020.104719, 104719
Kárný M (2020b) Minimum expected relative entropy principle. In: Proc. of the 18th ECC, IFAC, Sankt Petersburg, pp 35–40

Kárný M, Alizadeh Z (2019) Towards fully probabilistic cooperative decision making. In: Slavkovik M (ed) Multi-Agent Systems, EUMAS 2018, Springer Nature Switzerland AG, vol LNAI 11450, pp 1–16

Kárný M, Guy T (2012) On support of imperfect Bayesian participants. In: Guy T, et al (eds) Decision Making with Imperfect Decision Makers, vol 28, Springer, Int. Syst. Ref. Lib., pp 29–56

Kárný M, Guy T (2019) Preference elicitation within framework of fully probabilistic design of decision strategies. In: IFAC Int. Workshop on Adaptive and Learning Control Systems, vol 52, pp 239–244

Kárný M, Hůla F (2019) Balancing exploitation and exploration via fully probabilistic design of decision policies. In: Proc. of the 11th Int. Conf. on Agents and Artificial Intelligence: ICAART, vol 2, pp 857–864

Kárný M, Kroupa T (2012) Axiomatisation of fully probabilistic design. Inf Sci 186(1):105–113

Kárný M, Halousková A, Böhm J, Kulhavý R, Nedoma P (1985) Design of linear quadratic adaptive control: Theory and algorithms for practice. Kybernetika 21, supp. Nos 3–6

Kárný M, Böhm J, Guy T, Jirsa L, Nagy I, Nedoma P, Tesař L (2006) Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London, UK

Kárný M, Bodini A, Guy T, Kracík J, Nedoma P, Ruggeri F (2014) Fully probabilistic knowledge expression and incorporation. Statistics and Its Interface 7(4):503–515

Klenske E, Hennig P (2016) Dual control for approximate Bayesian reinforcement learning. Journal of Machine Learning Research 17:1–30

Kober J, Peters J (2011) Policy search for motor primitives in robotics. Machine Learning 84(1):171–203, DOI 10.1007/s10994-010-5223-6

Kracík J, Kárný M (2005) Merging of data knowledge in Bayesian estimation. In: Filipe J, et al (eds) Proc. of the 2nd Int. Conf. on Informatics in Control, Automation and Robotics, Barcelona, pp 229–232

Kulhavý R, Zarrop MB (1993) On a general concept of forgetting. Int J of Control 58(4):905–924

Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22:79–87

Kumar EV, Jerome J, Srikanth K (2014) Algebraic approach for selecting the weighting matrices of linear quadratic regulator. In: 2014 Intern. Conf. on Green Computing Communication and Electrical Engineering (ICGCCEE), pp 1–6, DOI 10.1109/ICGCCEE.2014.6922382

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

Larsson D, Braun D, Tsiotrasz P (2017) Hierarchical state abstractions for decision-making problems with computational constraints ArXiv:1710.07990v1[cs.AI], 22 Oct 2017

Lee K, Kim G, Ortega P, Lee D, Kim K (2019) Bayesian optimistic Kullback–Leibler exploration. Machine Learning 108(5):765–783, DOI 10.1007/s10994-018-5767-4

Li W, Song H (2016) ART: An attack-resistant trust management scheme for securing vehicular ad hoc networks. IEEE Transactions On Intelligent Transportation Systems 17:960–969

Liao Y, Deschamps F, Loures E, Ramos L (2017) Past, present and future of industry 4.0 – A systematic literature review and research agenda proposal. Int J of Production Res 55(12):3609–3629

Mayne D (2014) Model predictive control: Recent developments and future promise. Automatica pp 2967–2986

Meditch J (1969) Stochastic Optimal Linear Estimation and Control. McGraw Hill

Mesbah A (2018) Stochastic model predictive control with active uncertainty learning: A survey on dual control. Annual Reviews in Control 45:107 – 117, DOI https://doi.org/10.1016/j.arcontrol.2017.11.001, URL http://www.sciencedirect.com/science/article/pii/S1367578817301232

Moerland TM, Broekens J, Jonker CM (2018) Emotion in reinforcement learning agents and robots: a survey. Machine Learning 107(2):443–480, DOI 10.1007/s10994-017-5666-0

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

Peterka V (1972) On steady-state minimum variance control strategy. Kybernetika 8:219–231

Peterka V (1975) A square-root filter for real-time multivariable regression. Kybernetika 11:53–67

Peterka V (1981) Bayesian system identification. In: Eykhoff P (ed) Trends and Progress in System Identification, Perg. Press, pp 239–304

Peterka V (1991) Adaptation for LQG control design to engineering needs. In: Warwick K, Kárný M, Halousková A (eds) Lecture Notes: Adv. Methods in Adaptive Control for Industrial Application; Joint UK-CS seminar, vol 158, Springer-Verlag, N.Y.

Peterka V, Astrom K (1973) Control of multivariable systems with unknown but constant parameters. In: Prepr. of the 3rd IFAC Symp. on Identification and Process Parameter Estimation, IFAC, Hague, Delft, pp 534–544

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

Quinn A, Kárný M, Guy T (2016) Fully probabilistic design of hierarchical Bayesian models. Inf Sci 369:532–547

Rao M (1987) Measure Theory and Integration. J. Wiley

Rohrs C, Valavani L, Athans M, Stein G (1982) Robustness of adaptive control algorithms in the presence of unmodeled dynamics. In: IEEE Conf. on Decision and Control, Orlando, FL, vol 1, pp 3–11

Sandholm T (1999) Distributed rational decision making. In: Weiss G (ed) Multiagent Systems – A Modern Approach to Distributed Artificial Intelligence, MIT Press, pp 201–258

Savage L (1954) Foundations of Statistics. Wiley

Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural Networks 16(1):5 – 9, DOI https://doi.org/10.1016/S0893-6080(02)00228-9

Shannon C (1948) A mathematical theory of communication. Bell System Tech J 27:379–423, 623–656

Shore J, Johnson R (1980) Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran on Inf Th 26(1):26–37
Si J, Barto A, Powell W, Wunsch D (eds) (2004) Handbook of Learning and Approximate Dynamic Programming, Wiley-IEEE Press
Tanner M (1993) Tools for statistical inference. Springer Verlag, N.Y.
Tao G (2014) Multivariable adaptive control: A survey. Automatica 50(11):2737 – 2764
Ullrich M (1964) Optimum control of some stochastic systems. In: Prepr. of the VIII-th conf. ETAN, Beograd
Wolpert D, Macready W (1997) No free lunch theorems for optimization. IEEE Trans on Evolutionary Computation 1(1):67–82
Wu H, Guo X, Liu X (2017) Adaptive exploration-exploitation trade off for opportunistic bandits. Preprint arXiv:1709.04004
Yang Z, Wang C, Zhang Z, Li J (2019) Mini-batch algorithms with online step size. Knowledge-Based Systems 165:228–240