2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

A User Comfort Model and Index Policy for Personalizing Discrete Controller Decisions

Marcel Menner and Melanie N. Zeilinger

Abstract— **User feedback allows for tailoring system operation to ensure individual user satisfaction. A major challenge in personalized decision-making is the systematic construction of a user model during operation while maintaining control performance. This paper presents both an index-based control policy to smartly collect and process user feedback and a user comfort model in the form of a Markov decision process with a priori unknown user-specific state transition probabilities. The control policy utilizes explicit user feedback to optimize a reward measure reflecting user comfort and addresses the exploration-exploitation trade-off in a multi-armed bandit framework. The proposed approach combines restless bandits and upper confidence bound algorithms. It introduces an exploration term into the restless bandit formulation, utilizes user feedback to identify the user model, and is shown to be indexable. We demonstrate its capabilities with a simulation for learning a user's trade-off between comfort and energy usage.**

I. Introduction

User feedback is essential for the control of personalized devices to ensure user satisfaction. Feedback can be provided implicitly, by gathering measurements from human interaction, or explicitly in the form of ratings provided by the user, where the latter is considered in this paper. Ratings are usually of discrete nature, e.g. an integer number on a scale between 1 and 10. The challenges of user feedback lie in smart collection and processing of the accumulated ratings and the lack of a dynamical model for how actions affect the user's comfort reflected in the ratings. In this paper, we address these challenges and present a learning technique to collect and process explicit user feedback and build a user comfort model. User comfort is modeled as a Markov decision process (MDP), in which the states correspond to the user's comfort levels and the a priori unknown user-specific state transition probabilities originate from discrete controller decisions. Collection and processing of online data to select the optimal control action maximizing user comfort is addressed in a multi-armed bandit (MAB) setting, a technique ideally suited to address the arising exploration-exploitation trade-off [1].

The MAB framework offers a scheduling policy for control actions, which are represented by multiple arms, trading-off immediate reward with acquiring information about rarely applied control actions [1]. Each arm can be operated (active action) or rested (passive action). The fundamental research in MAB was introduced in [2] deriving an index for each arm, which quantifies the expected reward for choosing the optimal action/arm. The required assumption in a classical MAB problem is that non-operated arms remain static.

Important progress in relaxing this assumption is made in [3] with the restless bandit (RB). The RB is an index policy which allows for restless states of non-operated arms by replacing the constraint on the number of active arms at each time instance by a constraint on average activity. An index is defined based on the Lagrange multiplier of the relaxed optimization problem, which can be interpreted as a price for operation of the arm. Furthermore, [3] states conditions under which the RB index allocation is optimal. The most fundamental condition in an RB approach is indexability. Indexability is a monotonicity criterion required in order for the index to provide a meaningful measure for the price. General conditions on indexability are derived in [4]–[10], where simple sufficient conditions are stated and a geometric interpretation of indexability is provided. RB models are promising for control problems due to their scalability properties. For example, [11] makes use of RBs for demand response to estimate the state of ressource availability, or [12] proposes its application to sensor management and dynamic routing. Pioneering work on regret bounds is presented in [13] and further research on upper confidence bounds (UCB) and MAB algorithms can be found in [14]–[20].

Related work in user modeling is presented e.g. in [21], [22]. In [21], a trust-based consumer decision-making model is proposed in order to analyze the impact of trust and perceived risk of a website for a user's purchasing decisions. In [22], a method is introduced to support decision making with sparse ratings. The approach takes previously rated, related objects into consideration in order to obtain an initial estimate for sparsely rated, new objects. Related work in learning MDP models can be found e.g. in [23].

In this paper, we introduce a mathematical framework to model user comfort as a function of discrete control actions and propose an index policy, the exploratory restless bandit (ERB), to address the exploration-exploitation trade-off in designing the optimal control policy. The index policy is based on the RB framework presented in [3], the UCB1 in [15], and the index-computation proposed in [10] as foundation for the control design. The ERB allows for learning probabilities such as UCB1 and for restless non-operated arms such as RB. ERB augments RB with an exploration term that is inspired by UCB1 [15]. The ERB is utilized to smartly collect and process user feedback in order to learn the user's comfort model. We prove indexability of the considered problem under a simple condition.

M. Menner and M. N. Zeilinger are with the Institute for Dynamic Systems and Control, ETH Zürich, 8092 Zürich, Switzerland {mmenner,mzeilinger}@ethz.ch

This work was supported by the Swiss National Science Foundation under grant no. PP00P2 157601 / 1.

The paper is structured as follows. In Section II, we state the problem and present the user comfort model. Section III presents the ERB-based index policy for maximizing user comfort based on the defined model. Section IV proofs indexability of the considered problem. Section V presents a simulation applying ERB for learning user preferences in the form of a desired comfort-energy trade-off, e.g., in the control of an energy system. We conclude with Section VI.

II. Problem Statement

We consider the problem of maximizing user comfort by means of a controller utilizing user feedback in order to determine a tailored decision policy. We propose a model for addressing this problem where states relate to the user's comfort and control actions induce state transitions. Figure 1 illustrates the control scheme. The user provides feedback about their current comfort level. The learning-based controller computes a decision by collecting and processing feedback to decide on a control action. The applied control action affects the user's comfort, which is modeled as a stochastic state transition.

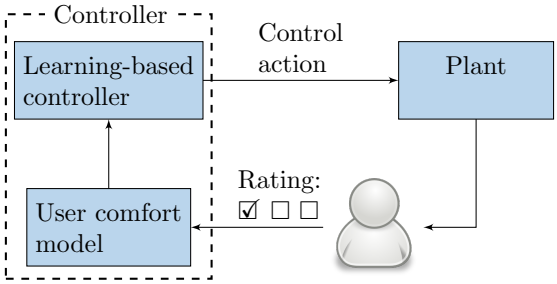

Fig. 1. Control scheme for personalized decision making.

A. User comfort model

Consider a Markov decision process with N discrete states $x^n(t) \in \{0,1\}$ for $n = 1, ..., N$ and M available discrete control actions. The states are defined such that $x^n(t)$ relates to comfort level n at time t. We define $c^m(t) \in \{0,1\}$ for $m = 1, ..., M$ as the binary variable selecting control action m, i.e. $c^m(t) = 1$ and $c^m(t) = 0$ if control action m is engaged and disengaged, respectively. We assume that there is one comfort level n active, i.e.

$$\sum_{n=1}^{N} x^n(t) = 1,$$

and that one control action is employed at each time t, i.e.

$$\sum_{m=1}^{M} c^m(t) = 1. \tag{1}$$

Each control action m causes a probabilistic state transition such that the state dynamics for each m are given by

$$\hat{x}^j(t+1) = \sum_{m=0}^{M} c^m(t) p_{jn}^m \hat{x}^n(t), \tag{2}$$

where $\hat{x}^j(t+1) \in [0,1]$ denotes the expected comfort, i.e. the probability that the comfort level at time $t+1$ is j. The probabilities p_{jn}^m are assumed to be initially unknown as they are user-specific and will be identified from user feedback. Figure 2 illustrates an example of a user comfort model for $N = 3$, i.e. there are three comfort levels $n = 1, 2, 3$.

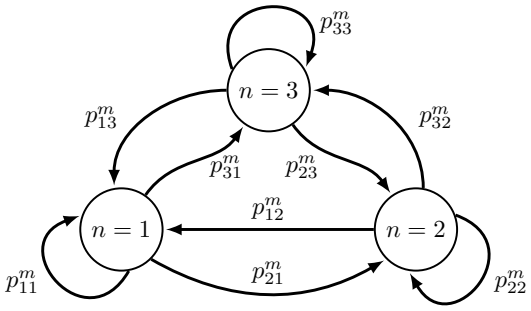

Fig. 2. User comfort model with three states $n = 1, 2, 3$.

After applying a control action, the comfort level n is reported by the user. The estimated probabilities $\hat{p}_{jn}^m$ are given by

$$\hat{p}_{jn}^m = \frac{\mathcal{N}_{jn}^m}{\sum_{\eta=1}^{N} \mathcal{N}_{j\eta}^m}, \tag{3}$$

where $\mathcal{N}_{jn}^m$ is the number of times, the comfort state transitioned from $x^n(t)$ to $x^j(t+1)$ when control action m was engaged. Eq. (3) is used to approximate (2) and yields the estimated state evolution as

$$\hat{X}(t+1) = \sum_{m=0}^{M} c^m(t) \hat{P}^m \hat{X}(t) \tag{4}$$

with

$$\hat{X}(t) := \begin{bmatrix} \hat{x}^1(t) \\ \vdots \\ \hat{x}^N(t) \end{bmatrix}, \; \hat{P}^m := \begin{bmatrix} \hat{p}_{11}^m & \cdots & \hat{p}_{1N}^m \\ \vdots & \ddots & \\ \hat{p}_{N1}^m & & \hat{p}_{NN}^m \end{bmatrix}.$$

B. Control objective

The aim is to find a policy $\pi : \{0,1\}^N \to \{0,1\}^M$, which allocates one control action m to each state $n = 1, ..., N$:

$$\begin{bmatrix} c^1(t) \\ \vdots \\ c^M(t) \end{bmatrix} = \pi\left(X(t)\right), \text{ where } X(t) := \begin{bmatrix} x^1(t) \\ \vdots \\ x^N(t) \end{bmatrix}. \tag{5}$$

We use $c^m(t) = \pi^m(X(t))$ to select one element of the vector-valued function π. The policy is designed to optimize the discounted cost $\mathcal{J}(X(0))$ with

$$\begin{aligned} \mathcal{J}(X(0)) = \max_{\hat{X}, c^m} & \sum_{t=0}^{\infty} \sum_{m=1}^{M} \beta^t R\left(\hat{X}(t), c^m(t)\right) \\ \text{s.t. } & (1) \\ & (4) \\ & \hat{X}(0) = X(0), \end{aligned} \tag{6}$$

where $R(\hat{X}(t), c^m(t))$ is the expected reward at time t, and β with $0 \leq \beta < 1$ is the discount factor. In order to enhance readability, we omit the dependance of the reward on the state $\hat{X}(t)$ in the remainder of the paper.

III. Index Policies for User Comfort

The aim of finding a control policy (5) to maximize the discounted cost (6) is addressed in a MAB framework, where each control action m is considered as an arm that can be engaged and is assigned an index to quantify the expected cost of arm m. In the considered optimization problem (6), there are multiple, mutually exclusive control actions which yield a change of states according to the stochastic model (2). In a classical MAB framework, indices are assigned under the assumption that the state of each arm does not change if this single arm is not engaged. This assumption is not satisfied in the considered case as the state evolves also if a particular discrete control action is not employed. Hence, we formulate (6) as a restless bandit problem, where each control action m corresponds to one arm m, and identify (1) as constraint on activity.

A. Formulation as restless bandit problem

The RB setting is used to compute an index for each control action/arm with the following three steps [3]:

i) The constraint on activity (1) is relaxed into a constraint on average activity.
ii) The Lagrangian of the relaxed optimization problem is constructed, where the constraint on average activity introduces one Lagrange multiplier per arm.
iii) The Lagrange multipliers are computed by comparing the expected cost of the relaxed optimization problem for engaging and disengaging each arm.

The steps are outlined in the following, details can be found in [3].

Step i) Constraint on average activity: Using the results in [3], (1) is relaxed such that the constraint on activity only has to be fulfilled on average:

$$0 = \frac{1}{1-\beta} - \sum_{t=0}^{\infty}\sum_{m=1}^{M} \beta^t c^m(t), \tag{7}$$

where we use $\sum_{t=0}^{\infty} \beta^t = 1/(1-\beta)$ for $0 \leq \beta < 1$.

Step ii) Relaxed optimization problem: The relaxed optimization problem is then given by substituting (7) for (1) in problem (6). The corresponding Lagrangian results in

$$\sum_{t=0}^{\infty}\sum_{m=1}^{M} \beta^t R\left(c^m(t)\right) + \gamma\left(\frac{1}{1-\beta} - \sum_{t=0}^{\infty}\sum_{m=1}^{M}\beta^t c^m(t)\right),$$

where γ is the Lagrange multiplier. The resulting dual optimization problem is given by

$$\begin{aligned}\mathcal{L}(X(0)) = \max_{\hat{X},c^m,\gamma} & \sum_{m=1}^{M} \mathcal{L}_m(c^m,\gamma) + \gamma\frac{1}{1-\beta} \\ \text{s.t. } & (4) \\ & \hat{X}(0) = X(0)\end{aligned} \tag{8}$$

where

$$\mathcal{L}_m(c^m,\gamma) = \sum_{t=0}^{\infty} \beta^t \left(R(c^m(t)) - \gamma c^m(t)\right). \tag{9}$$

The change in the order of summations from (6) to (8) is valid since the discounted cost is absolutely summable. The proposed formulation multiplies γ with $c^m(t)$ indicating activity. Hence, γ in (9) can be identified as price for activity and will be used as index in the proposed bandit formulation.

Step iii) Index computation for each arm: As proposed in [3], problem (6) is approximated by comparing the expected cost (9) for each arm. As a measure for the price, the index γ plays the leading role in the RB control framework and is conceptually computed as follows. Let the expected cost in (9) for both employing ($c^m(0) = 1$) and not employing ($c^m(0) = 0$) arm m at time $t = 0$ be equal, i.e.

$$\begin{aligned}&\sum_{t=1}^{\infty} \beta^t \left(R(c^m(t)) - \gamma c^m(t)\right)\big|_{c^m(t=0)=1} + R(1) - \gamma \\ &= \sum_{t=1}^{\infty} \beta^t \left(R(c^m(t)) - \gamma c^m(t)\right)\big|_{c^m(t=0)=0} + R(0).\end{aligned} \tag{10}$$

If $\gamma > 0$, the reward for employing arm m at time $t = 0$ is expected to be higher than not employing arm m; if $\gamma = 0$, the controller is indifferent as to whether or not arm m is employed; and if $\gamma < 0$, not employing arm m is beneficial. Note that the reward at time $t \geq 0$ is dependent on the control action taken at $t = 0$ because the predicted state $\hat{X}(1)$ depends on the employed control action m at $t = 0$. The detailed computation of the index according to (10) is presented in Section III-C. Due to the relaxation (7), it may be beneficial for more than one control action to be employed. In order to enforce a unique control law and reinforce constraint (1), the arm yielding the highest γ is employed. More details can be found in [3].

B. Exploratory restless bandit

The previously introduced RB formulation can be used in order to obtain a control policy, which exploits gathered knowledge of the model. In order to generate a control policy that also acquires new knowledge and prioritizes control actions with badly explored transition probabilities, we introduce the so-called exploratory restless bandit, which introduces a subsidy for exploration ρ in problem (9). This heuristic subsidy increases or decreases the price γ and is motivated by the UCB1 algorithm [15]. We define the exploratory cost $\mathcal{L}_m^{\rho}$ of arm m as

$$\mathcal{L}_m^{\rho}(c^m,\gamma) = \sum_{t=0}^{\infty} \beta^t \left(R(c^m(t)) - (\gamma-\rho)c^m(t)\right). \tag{11}$$

We select the exploration measure ρ along the lines of [15] and present the condition under which ERB reduces to UCB1 in the following.

Reduction to UCB1: UCB1 allocates an index to each arm m calculated as the benefit of employing arm m and an exploration measure ρ:

$$\text{UCB1} = R(1) - R(0) + \rho, \tag{12}$$

which represents a trade-off between the averagely gained reward and exploration represented by ρ [15]. As we will illustrate in Section V, a disadvantage of UCB1 is the lack of

prediction capabilities. UCB1 estimates the expected future reward (12) with a horizon of one time step. This drawback becomes evident by letting $\beta = 0$ in (11), in which case (11) accounts for only one time step:

$$\mathcal{L}_m^\rho(c^m, \gamma) = R(c^m(0)) - (\gamma - \rho)c^m(0).$$

Let $\mathcal{L}_m^\rho$ for employing m, i.e. $c^m(0) = 1$, and not employing m, i.e. $c^m(0) = 0$, be equal, cf. Step iii) in Section III-A:

$$R(1) - (\gamma - \rho) = R(0).$$

Then, it can be seen that $\gamma =$ UCB1 in (12), which shows that UCB1 corresponds to the resulting control policy maximizing (11) for $\beta = 0$.

C. Index computation for the proposed ERB formulation

As in RB, the index γ for the ERB is computed by comparing the cost of engaging and disengaging a control action. In the considered problem, disengaging one control action, however, can induce $M-1$ different control actions, i.e. $M-1$ different state transition probabilities. In order to ease notation, this section presents the computation of indices under the assumption of two available control actions m and $\neg m$, i.e. the reward for disengaging control action m is unique. The extension to multiple control actions is addressed in Section III-D. We extend the computation of the index γ in [10] to incorporate the exploration measure ρ in the following. The computation makes use of the reward measure $f_n^{\pi^m}(\rho)$ and activity measure $g_n^{\pi^m}$ of arm m defined as

$$\begin{aligned} f_n^{\pi^m}(\rho) &:= \sum_{t=0}^{\infty} \beta^t \left(R(c^m(t)) + \rho c^m(t)\right) \\ g_n^{\pi^m} &:= \sum_{t=0}^{\infty} \beta^t c^m(t). \end{aligned} \tag{13}$$

It follows that (11) is equivalent to

$$\mathcal{L}_m^\rho(c^m, \gamma) = f_n^{\pi^m}(\rho) - \gamma g_n^{\pi^m}.$$

The procedure to compute the index for each arm makes use of the notion of active sets S_k [10]. An active set can be used as a control policy and indicates activity of a control action for all states in the active set.

Definition 1 (Active Set S_k)**.** *Considering an arm m, an active set S_k is defined as the set of comfort states for which control action m is employed with $S_0 := \varnothing$ and*

$$S_k := \{n_1, n_2, ..., \ n_k\}. \tag{14}$$

We define $\langle S_k \rangle$ as a control policy employing the control action m ($c^m(t) = 1$) if the current comfort level $n \in S_k$.

The active sets S_k with threshold comfort state n_k are successively built up to induce a consistent ordering of states, for which the control action is employed, i.e.

$$S_k = S_{k-1} \cup \{n_k\}. \tag{15}$$

The optimal control action at a current comfort state $\bar{n}$ is computed as follows: Consider the empty set S_0, i.e. the control action m is not employed for any comfort level n. First, one index $\gamma_n^{S_0}(\rho)$ is computed for every comfort level n as

$$\gamma_n^{S_0}(\rho) = \frac{r_n^{S_0}(\rho)}{w_n^{S_0}} \tag{16}$$

with the marginal reward measure and marginal activity measure defined as

$$\begin{aligned} r_n^{S_0}(\rho) &:= f_n^{\langle 1, S_0 \rangle}(\rho) - f_n^{\langle 0, S_0 \rangle}(\rho) \\ w_n^{S_0} &:= g_n^{\langle 1, S_0 \rangle} - g_n^{\langle 0, S_0 \rangle}. \end{aligned} \tag{17}$$

The control policies $\langle 1, S \rangle$ and $\langle 0, S \rangle$ are defined such that, at time $t = 0$, the control action m is employed ($c^m(0) = 1, c^{\neg m}(0) = 0$) and not employed ($c^m(0) = 0, c^{\neg m}(0) = 1$), respectively, followed by the control policy $\langle S \rangle$. Hence, the two policies differ only by virtue of the control action taken at time $t = 0$. Eq. (17) results to

$$\begin{aligned} r_n^{S_0}(\rho) &= R(1) + \rho - R(0) + \beta \sum_{j=1}^{N} (\hat{p}_{jn}^m - \hat{p}_{jn}^{\neg m}) f_j^{\langle S_0 \rangle}(\rho) \\ w_n^{S_0} &= 1 + \beta \sum_{j=1}^{N} (\hat{p}_{jn}^m - \hat{p}_{jn}^{\neg m}) g_j^{\langle S_0 \rangle}. \end{aligned} \tag{18}$$

Note that $f_j^{\langle S_0 \rangle}(\rho)$ and $g_j^{\langle S_0 \rangle}$ define the reward measure and the activity measure for applying $\neg m$ at every time step.

The estimated state transition probabilities $\hat{p}_{jn}^m$ result from the control action m, i.e. $c^m(0) = 1$, and $\hat{p}_{jn}^{\neg m}$ result from the control action $\neg m$, i.e. $c^m(0) = 0$. The threshold state n_1, cf. (15), results from the highest value $\gamma_n^{S_0}(\rho)$, i.e. $n_1 = \arg\max\{\gamma_n^{S_0}(\rho)\}$. Then, $S_1 := \{n_1\}$ and $\gamma_n^{S_1}(\rho)$ is computed for all n except n_1 in order to find n_2. The procedure is iterated until the current comfort level $\bar{n}$ matches the threshold state n_k, i.e. $\bar{n} = n_k$. If $\gamma_{n_k}^{S_{k-1}}(\rho) > 0$, control action m is employed in comfort state $\bar{n}$ and $\neg m$ otherwise. Algorithm 1 summarizes the index computation.

Algorithm 1 Index computation [10]

Output: $\{n_k, \gamma_{n_k}^{S_{k-1}}\}_{k=1}^N$

```
S_0 = ∅
for k = 1 to N do
    n_k = arg max{γ_n^{S_{k-1}} ∀ n ∈ {1,...N}\S_{k-1}}
    S_k = S_{k-1} ∪ {n_k}
end for
```

D. Extension to more than two available control actions

This section extends the index computation in Section III-C to the case where more than two control actions are available. The availability of multiple control actions requires additional steps for the index computation as both $R(0)$ and $\hat{p}_{jn}^{\neg m}$ in (17) are not unique but a result of one of the other $M-1$ control actions. Available techniques only allow for considering each control action as active or passive [3], [10], similarly to Section III-C. We therefore present simple techniques for reducing the problem to a sequence of

selections between two policies, which we denote as π_a and π_p, for which the technique in Section III-C applies.

The quantities in (18) which are required for the selection between two policies π_a and π_p are computed as follows: Let m_a and m_p be the control actions that the policies π_a and π_p employ when in state $\bar{n}$, respectively. Then, $R(1)$ is computed as the reward of employing control action m_a, i.e. $R(1) = R(c^{m_a}(0) = 1)$, and $\hat{p}_{jn}^{m} = \hat{p}_{jn}^{m_a}$. Similarly, $R(0) = R(c^{m_p}(0) = 1)$ and $\hat{p}_{jn}^{\neg m} = \hat{p}_{jn}^{m_p}$.

Let $\mathcal{A}$ be the operator to determine the optimal control action at the current state given two possible policies as inputs by applying the procedure in Section III-C:

$$\pi(\hat{X}) = \mathcal{A}\left(\pi_a(\hat{X}), \pi_p(\hat{X})\right). \tag{19}$$

1) Enumeration: A naive approach is to enumerate all possible combinations of state to action allocations to select the best policy. The procedure is initialized by defining the active policy π_a such that the control action $m = 1$ is chosen for every comfort state n. The procedure iterates over all M^N possible combinations for defining a passive policy. If π_p is superior to the current active policy π_a, π_p is considered as active policy for the remaining iterations.

2) Branch-and-bound: The idea of the branch-and-bound algorithm is to use a bound on the maximum achievable index to exclude branches from the enumeration tree, which cannot be optimal [24]. A bound is computed by means of an artificial control action which directly leads to maximum comfort with probability 1 and this bound is used to disregard branches. This method is in the worst case identical to enumeration but offers the potential of excluding certain combinations, or branches, to save computation time.

3) Approximation with subset of policies: A computationally inexpensive heuristic is to select a small subset of possible policies in order to find a sub-optimal policy.

Let $\bar{\pi}_m$ be the scheduling policy, which allocates control action m to every comfort state n. First, the method in Section III-C is utilized to determine a control policy π_2:

$$\pi_2(\hat{X}) = \mathcal{A}\left(\bar{\pi}_2(\hat{X}), \bar{\pi}_1(\hat{X})\right),$$

i.e. π_2 chooses the best control action out of $m \in \{1, 2\}$, for each state n. Then, π_2 is considered as passive policy $\pi_p = \pi_2$ in (19) to determine π_3 with $\pi_a = \bar{\pi}_3$: $\pi_3 = \mathcal{A}(\bar{\pi}_3, \pi_2)$. This procedure is iterated $M - 1$ times to determine

$$\pi_M(\hat{X}) = \mathcal{A}\left(\bar{\pi}_M(\hat{X}), \pi_{M-1}(\hat{X})\right).$$

While reducing the complexity of the problem, this greedy approach is not guaranteed to find an optimal control policy.

IV. Indexability

In order for the index γ to be a meaningful measure for the price, a consistent ordering of the control actions must be induced [3], i.e. as γ for a control action m decreases, it becomes profitable to engage m in more and more states n. This implies that once the threshold price of engaging m for comfort level n is reached, m is always engaged in n as γ decreases, which is formalized in the following definition.

Definition 2 (Indexability [3])**.** *Let S be the active set of control action m. Then, the optimization problem* (9) *is indexable, if S increases monotonically from the empty set $\varnothing$ to the entire state space $\{1, ..., N\}$ for γ decreasing from ∞ to $-\infty$.*

Simple sufficient conditions for indexability are presented in [10] using the notion of partial conservation laws (PCL). A key result from [10] is summarized in the following that is applied to prove indexability of the ERB problem in (11).

Definition 3 ([10])**.** *A bandit is PCL-indexable if*

(i) the marginal work expenditure is positive:

$$g_n^{\langle \pi^a, S\rangle} - g_n^{\langle \pi^p, S\rangle} > 0 \ \forall \ n = 1, ...N$$

(ii) monotonically nonincreasing indices can be found:

$$\gamma_{n_1}^{S_0}(\rho) \geq \gamma_{n_2}^{S_1}(\rho) \geq ... \geq \gamma_{n_N}^{S_{N-1}}(\rho). \tag{20}$$

Theorem 1 ([10])**.** *A PCL-indexable bandit is indexable.*

Theorem 2. *The ERB design in* (11) *is indexable if $\beta < 0.5$ and the indices are computed with Algorithm 1.*

Proof. The proof can be found in the Appendix. □

V. Simulation Example

The proposed user comfort model and exploratory restless bandit is applied to the problem of learning a user's individual trade-off between comfort and energy consumption and optimize their comfort. To demonstrate the control method, we optimize (6) solely of one user, however consideration of multiple users can be similarly addressed.

Consider the user comfort model in Section II with $N = 3$ comfort states $n = 1, ..., 3$ and $M = 5$ control actions $m = 1, ..., 5$, which represent five levels of energy consumption. The objective is to maximize (6) with the expected reward

$$R(c^m(t)) = 1/M\left(\begin{bmatrix}1 & 2 & 3\end{bmatrix}\hat{X}(t)\right) - 0.16c^m(t)m$$

by achieving a high comfort level n, represented by $\hat{X}(t)$, and low energy usage, represented by $c^m(t)m$. The user's initial comfort level for every simulation is chosen as $n = 1$ ($x^1(0) = 1$). The comfort state $x^n(t)$ evolves as in (2) where the true state transition probability matrices are given by

$$\begin{aligned} P^1 &= \begin{bmatrix}.65 & .05 & .40\\ .25 & .05 & .40\\ .10 & .90 & .20\end{bmatrix}, P^2 = \begin{bmatrix}.05 & .50 & .30\\ .90 & .30 & .35\\ .05 & .20 & .35\end{bmatrix}\\ P^3 &= \begin{bmatrix}.25 & .40 & .90\\ .70 & .40 & .05\\ .05 & .20 & .05\end{bmatrix}, P^4 = \begin{bmatrix}.65 & .70 & .20\\ .25 & .20 & .20\\ .10 & .10 & .60\end{bmatrix}\\ P^5 &= \begin{bmatrix}.20 & .40 & .70\\ .10 & .50 & .20\\ .70 & .10 & .10\end{bmatrix}. \end{aligned} \tag{21}$$

We choose $\beta = 0.499$ to ensure indexability, cf. Theorem 2 and hence, an approximation of (6) with a finite horizon of $t = [0, 50]$ is sufficient as β^{50} is of the order of computational accuracy. We carried out 1000 simulations and analyze the mean and standard deviation of the accumulated

reward of different policies, where the accumulated reward with $t \geq 1$ is defined as

$$\bar{R}(t) = \sum_{\tau=1}^{t} \left(\sum_{n=1}^{N} \frac{x^n(\tau) n}{t} - 0.16 \sum_{m=1}^{M} \frac{c^m(\tau) m}{t} \right). \quad (22)$$

A. Baseline performance with known probabilities

First, we assume that the user-specific Markov state transition probabilities are known and we compare the reward of a baseline ERB (ERB_{bl}) and baseline UCB1 (UCB1_{bl}) policy. The approximation procedure in Section III-D.3 is used for the ERB control policy. As a result, ERB_{bl} employs control action $m = 2$, 1, and 2 when in state $n = 1$, 2, and 3, respectively. UCB1_{bl} employs control action $m = 5$, 1, and 4 when in state $n = 1$, 2, and 3, respectively. The average reward $\bar{R}(t)$ for ERB_{bl} is higher than for UCB1_{bl} throughout the simulation with stationary values of 2.05 and 1.89, respectively, cf. Figure 3.

B. Learning performance with unknown probabilities

We define ρ for both ERB and UCB1 as

$$\rho = \sqrt{2\ln(\sigma)/\sigma_a} - \sqrt{2\ln(\sigma)/\sigma_p},$$

where

$$\sigma = t, \;\; \sigma_a = \sum_{n=1}^{N} \mathcal{N}_{jn}^{m(\pi_a(X(t)))}, \;\; \sigma_p = \sum_{n=1}^{N} \mathcal{N}_{jn}^{m(\pi_p(X(t)))}$$

with $\mathcal{N}_{jn}^{m}$ as in (3). This exploration term measures how well the probabilities of the control actions, which the active and passive policies π_a and π_p would employ, are explored if the system is in comfort state n. In addition, a naive approach to trade-off exploration and exploitation is used for comparison, which is referred to as ϵ-PI. It is defined such that it chooses with probability $1-\epsilon_t$ the action with highest reward obtained from a policy iteration (PI) method [25] and a random action with probability $\epsilon_t = \min\{1\ ,\ 4M/t\ \}$. The initial state transition probabilities are chosen as $\hat{p}_{jn}^{m} = 1/N \;\forall\; n, j, m$.

Figure 3 shows the average reward (22) for $t = [1, 250] \cup [10250, 10400]$ for ERB_{bl}, UCB1_{bl}, ERB, UCB1, and ϵ-PI. It can be seen that ERB has a higher average reward than UCB1 of around 0.02 in the learning phase $t < 250$. The average rewards of both ERB and UCB1 are significantly higher than ϵ-PI for $t < 250$. For $t > 10250$, the average reward of ERB is higher than ϵ-PI and UCB1. The plot demonstrates a robust learning performance of ERB since the mean converges to a value close to the maximum achievable average reward of 2.05 (cf. ERB_{bl}) and the standard deviation is small. Both ϵ-PI and UCB1 have a lower mean and larger standard deviation as ERB, which suggests a wider spread of learned policies. Note that UCB1 gained more reward than UCB1_{bl}, which is best explained with the high standard deviation of UCB1. Once UCB1 has learned the probabilities perfectly, the average reward would converge to UCB1_{bl} with low standard deviation. To summarize, ERB achieves a good trade-off between exploration and exploitation (cf. $t < 250$) with robust learning performance (cf. $t > 10250$) compared to both UCB1 and ϵ-PI.

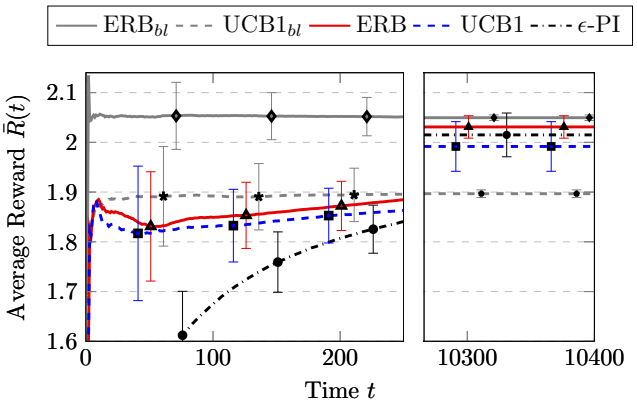

Fig. 3. The average reward of ERB_{bl}, UCB1_{bl}, ERB, UCB1, and ϵ-PI are shown together with their standard deviation as solid gray, dashed gray, solid red, dashed blue, and dash-dotted black lines, respectively.

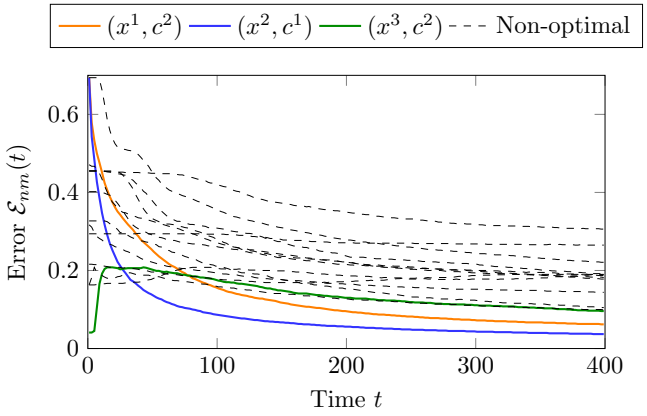

Fig. 4. Errors of learned transition probabilities from true values in (23), where the errors of non-optimal allocations are plotted in dashed black and of the optimal allocations in solid orange, blue, and green.

Figure 4 shows the learning performance of ERB measured as the errors $\mathcal{E}_{nm}(t)$ of the estimated transition probabilities $\hat{p}_{jn}^{m}(t)$ with respect to the true values in (21)

$$\mathcal{E}_{nm}(t) = \sqrt{\textstyle\sum_{j=1}^{N} \left(\hat{p}_{jn}^{m}(t) - p_{jn}^{m}\right)^2}. \quad (23)$$

The mean of 1000 simulations is shown for $n \in \{1, 2, 3\}$ and $m \in \{1, 2, 3, 4, 5\}$. It can be seen that ERB decides quickly on actions that have to be explored more carefully and disregards unfavorable control actions at an early stage.

VI. Conclusion

This paper has proposed a multi-armed bandit framework for selecting the optimal controller from a set of discrete controller decisions, which maximizes user comfort by incorporating user feedback. A new model for user comfort in the form of a Markov decision process with a-priori unknown transition probabilities was presented. An index-based control policy, the exploratory restless bandit, to trade-off exploration with exploitation was introduced. The proposed exploratory restless bandit was shown to be indexable and to outperform other scheduling policies such as UCB1 and a greedy policy iteration method in a simulation example.

APPENDIX: PROOF OF THEOREM 2

Proof of Theorem 2. Lemma 1 shows that the ERB in (11) for all ρ is indexable if the ERB with $\rho = 0$ is indexable. Using Theorem 1, it follows that the ERB with $\rho = 0$ is indexable if it is PCL-indexable. What remains to show is that the ERB in (11) with $\rho = 0$ is PCL-indexable, i.e. both (i) and (ii) in Definition 3 are satisfied.

Condition (i) requires a positive marginal work $w_n^S > 0$:

$$w_n^S = 1 + \beta \sum_{j=1}^{N} \left(p_{jn}^{m(\pi_a X(0))} - p_{jn}^{m(\pi_p X(0))} \right) g_j^{\langle S \rangle} > 0.$$

With $\sum_{t=0}^{\infty} \beta^t = 1/(1-\beta)$, bounds for $g_j^{\langle S \rangle}$ can be stated, cf. (13) with $c^{m(\pi_p(X(t)))} = 0$ for all t and $c^{m(\pi_a(X(t)))} = 1$ for all t:

$$g_j^{\langle S_0 \rangle} = 0 \leq g_j^{\langle S \rangle} \leq g_j^{\langle S_N \rangle} = \frac{1}{1-\beta}. \tag{24}$$

Thus, the lower bound of w_n^S evaluates with (24) to

$$1 + \beta \left(g_j^{\langle S_0 \rangle} - g_j^{\langle S_N \rangle} \right) \leq w_n^S. \tag{25}$$

The term in parenthesis in (25) originates from the worst case scenario: Taking the active policy at $t = 0$, the passive policy will be activated with probability 1 for $t \in [1, \infty)$; taking the passive policy at $t = 0$, the active policy will be activated with probability 1 for $t \in [1, \infty)$. With (24) and (25), Condition (i) in Definition 3 is therefore satisfied if

$$\left(\frac{1-2\beta}{1-\beta} \right) \leq w_n^S.$$

Hence, for any $\beta < 0.5$, $w_n^S > 0$.

Condition ii) is satisfied as the indices $\gamma_{n_k}^{S_{k-1}}$ defined by Algorithm 1 are monotonically nonincreasing. □

Lemma 1. *If (11) with $\rho = 0$ is PCL-indexable, then (11) with any $\rho \neq 0$ is indexable.*

Proof of Lemma 1. It is immediate from (13) that

$$f_n^{\langle S \rangle}(\rho) = f_n^{\langle S \rangle}(0) + \rho g_n^{\langle S \rangle}$$

and thus

$$\begin{aligned} r_n^S(\rho) &= R\left(c^{m(\pi_a X(0))}(0) \right) + \rho - R\left(c^{m(\pi_p X(0))}(0) \right) \\ &\quad + \beta \sum_{j=1}^{N} \left(p_{jn}^{m(\pi_a X(0))} - p_{jn}^{m(\pi_p X(0))} \right) \left(f_j^{\langle S \rangle}(0) + \rho g_j^{\langle S \rangle} \right) \\ &= r_n^S(0) + \rho w_n^S. \end{aligned} \tag{26}$$

The indices in (16), with (17) and (26), become

$$\gamma_n^S(\rho) = \frac{r_n^S(0) + \rho w_n^S}{w_n^S} = \gamma_n^S(0) + \rho.$$

Hence, if indices $\gamma_{n_k}^{S_{k-1}}(\rho = 0)$ in (20) can be found, then indices $\gamma_{n_k}^{S_{k-1}}(\rho)$ also fulfill (20), and condition (ii) in Definition 3 is satisfied. □

REFERENCES

[1] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments (Monographs on statistics and applied probability)*. London: Chapman & Hall, 1985.

[2] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 41, no. 2, pp. 148–177, 1979.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probability*, vol. 25, pp. 287–298, 1988.

[4] D. Bertsimas and J. Niño-Mora, "Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems," *Math. Operations Research*, vol. 21, no. 2, pp. 257–306, 1996.

[5] ——, "Restless bandits, linear programming relaxations and a primal-dual index heuristic," *Operations Research*, vol. 48, no. 1, pp. 80–90, 2000.

[6] J. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Advances Appl. Probability*, vol. 33, no. 1, pp. 76–98, 2001.

[7] ——, "Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach," *Math. Programming*, vol. 93, no. 3, pp. 361–413, 2002.

[8] ——, "A marginal productivity index policy for the finite-horizon multiarmed bandit problem," in *Proc. 44th IEEE Conf. Decision and Control*, Seville, Spain, 2005, pp. 1718–1722.

[9] ——, "Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock m/g/1 queues," *Math. Operations Research*, vol. 31, no. 1, pp. 50–84, 2006.

[10] ——, "Dynamic priority allocation via restless bandit marginal productivity indices," *Top*, vol. 15, no. 2, pp. 161–198, 2007.

[11] J. A. Taylor and J. L. Mathieu, "Index policies for demand response," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1287–1295, May 2014.

[12] J. L. Ny, M. Dahleh, and E. Feron, "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," in *Proc. American Control Conf.*, Seattle, WA, 2008, pp. 4220–4225.

[13] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[14] R. Agrawal, "Sample mean based index policies with O(log n) regret for the multi-armed bandit problem," *Advances Appl. Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[15] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[16] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proc. 27th Int. Conf. Machine Learning*, Haifa, Israel, 2010, pp. 1015–1022.

[17] A. Krause and C. S. Ong, "Contextual Gaussian process bandit optimization," in *Conf. Neural Inf. Process. Systems*, Granada, Spain, 2011.

[18] T. Desautels, A. Krause, and J. W. Burdick, "Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization," *J. Machine Learning Research*, vol. 15, pp. 3873–3923, 2014.

[19] I. Bogunovic, J. Scarlett, and V. Cevher, "Time-varying Gaussian process bandit optimization," in *Proc. 19th Int. Conf. Artifical Intelligence and Statistics*, Cadiz, Spain, 2016, pp. 314–323.

[20] H. Abdelrahman, F. Berkenkamp, J. Poland, and A. Krause, "Bayesian optimization for maximum power point tracking in photovoltaic power plants," in *Proc. European Control Conf.*, Aalborg, Denmark, 2016, pp. 2078–2083.

[21] D. J. Kim, D. L. Ferrin, and H. R. Rao, "A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents," *Decision support systems*, vol. 44, no. 2, pp. 544–564, 2008.

[22] M. Jamali and M. Ester, "TrustWalker: A random walk model for combining trust-based and item-based recommendation," in *Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Paris, France, 2009, pp. 397–406.

[23] H. Itoh, H. Fukumoto, H. Wakuya, and T. Furukawa, "Bottom-up learning of hierarchical models in a class of deterministic POMDP environments," *Int. J. of Appl. Mathematics and Comput. Sci.*, vol. 25, no. 3, pp. 597–615, 2015.

[24] E. L. Lawler and D. E. Wood, "Branch-and-bound methods: A survey," *Operations research*, vol. 14, no. 4, pp. 699–719, 1966.

[25] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: The MIT Press, 1960.