Online Learning for Personalized Room-Level Thermal Control: A Multi-Armed Bandit Framework

Parisa Mansourifard
Ming Hsieh Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
parisama@usc.edu

Farrokh Jazizadeh
Sony Astani Dept. of Civil and Environmental Engineering
University of Southern California
Los Angeles, CA, USA
jazizade@usc.edu

Bhaskar Krishnamachari
Ming Hsieh Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
bkrishna@usc.edu

Burcin Becerik-Gerber
Sony Astani Dept. of Civil and Environmental Engineering
University of Southern California
Los Angeles, CA, USA
becerik@usc.edu

ABSTRACT

We consider the problem of automatically learning the optimal thermal control in a room in order to maximize the expected average satisfaction among occupants providing stochastic feedback on their comfort through a participatory sensing application. Not assuming any prior knowledge or modeling of user comfort, we first apply the classic UCB1 online learning policy for multi-armed bandits (MAB), that combines exploration (testing out certain temperatures to understand better the user preferences) with exploitation (spending more time setting temperatures that maximize average-satisfaction) for the case when the total occupancy is constant. When occupancy is time-varying, the number of possible scenarios (i.e., which particular set of occupants are present in the room) becomes exponentially large, posing a combinatorial challenge. However, we show that LLR, a recently-developed combinatorial MAB online learning algorithm that requires recording and computation of only a polynomial number of quantities can be applied to this setting, yielding a regret (cumulative gap in average satisfaction with respect to a distribution aware genie) that grows only polynomially in the number of users, and logarithmically with time. This in turn indicates that difference in unit-time satisfaction obtained by the learning policy compared to the optimal tends to 0. We quantify the performance of these online learning algorithms using real data collected from users of a participatory sensing iPhone app in a multi-occupancy room in an office building in Southern California.

BuildSys '13, November 13 - 14 2013, Roma, Italy

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems

General Terms

Human Factors

Keywords

Thermal Comfort, Personalized Control, Online Learning

1. INTRODUCTION

A key ingredient for building smart, interactive, energy-efficient spaces is automated, personalized, thermal control. Indeed industry developments such as the Nest thermostat show that there is a great demand for thermal-control systems that can learn and adapt themselves to the preference of users.

For relatively simple single-user environments such a control system can either explicitly ask a user for her desired temperature, try to infer or directly observe the user's presence/absence and set the thermal controls accordingly. In this paper, we tackle a somewhat different setting, in which user preferences are learned over time through trials with stochastic feedback.

We consider initially a single space with multiple occupants that are always present. However, we do not assume that the occupants directly report their desired temperature. Instead, they are presented over time with different temperature settings, and asked to give feedback on the temperature as an integer between -50 (too hot, reduce temperature) to +50 (too cold, increase temperature). Their feedback is then translated and normalized to a satisfaction level from 0 (highly unsatisfied) to 1 (highly satisfied). This feedback is solicited through a participatory-sensing application - we have ourselves developed an app that does precisely this for our raw data collection experiments.

For this problem, we propose to use a well-known online learning policy called UCB1 [1] from the framework of

multi-armed bandit theory for selecting the temperature levels over time. Each temperature setting can be seen as an arm, yielding a stochastic reward (the average of the satisfaction reported by occupants). Over time, UCB1 keeps track of the sample mean of the reward obtained from each arm (temperature setting), as well as the number of times that each arm has been played, and combines these quantities into an index for each arm, which on the one-hand tries to give higher preference to arms with higher mean rewards (for exploitation of arms that have thus far shown to be good candidates), and on the other hand gives preference to arms that have been insufficiently well sampled (for exploration, to take sufficiently many samples from all arms to make sure a good arm does not go undetected due to a bad "streak"). Over time, it learns to spend most of its time on the best arm.

The advantage of adopting the MAB framework for this problem is that the online learning policy has some provable bounds on performance. In particular, Auer *et al.* have proved mathematically that the regret (gap in cumulative reward compared to an arm-distribution-aware genie) achieved by the UCB1 policy is bounded by a function that is logarithmic in time and linear in the number of arms. This, in turn, means that the instantaneous reward obtained by this online learning policy asymptotically tends to the optimal reward obtained by the genie. We evaluate UCB1 on a real data set obtained from four users in a multi-occupancy office building and show that this is indeed the case.

We next consider a more general and challenging case, when the occupancy of the room is dynamic. For this case, we make the assumption that each time the identity of the current occupants is known, specifically that the feedback is obtained from all occupants at each time, and tagged with their ID. Over time, different users occupy the space. For each combination of users a different temperature setting may maximize the average user satisfaction. Applying UCB1 to this setting naively, however, would yield a suboptimal outcome - a single temperature setting that is best for the *average* occupant distribution. Therefore we seek a more sophisticated solution that learns to adapt the temperature setting to the particular users in the space, for any combination. however, there are in principle exponentially many distinct cases to consider - if there are n users, there are 2^n different combinations in which they could be present.

Learning efficiently in the face of such dynamics and combinatorial explosion is a potentially difficult task. For this problem, we adapt the recently designed LLR algorithm for combinatorial multi-armed bandits [7], which can be applied to problems involved constrained combinatorial arm selections so long as the reward is a linear combination of component rewards, and the component rewards are individually observed. Applied to our problem, the LLR policy maintains sample means and number of times played, not for each temperature alone as with UCB1, or with each possible combination of occupants and temperature, but for each occupant individually for each temperature. As a result, it incurs only polynomial storage and computation, and has been mathematically shown to yield regret that is polynomial in the underlying variables and logarithmic in time. In our setting, we again illustrate this behaviour based on the real data set. This policy, incidentally, has a nice additional feature —- it is flexible enough to handle the presence of new users and it does so gracefully. Any new users detected are simply added to the table of tracked users, with 0 prior observations. If the number of newer users is large compared to the old users, it will be more likely to select a temperature for exploration, whereas when the number of new users is relatively small, it is more likely to select a temperature that is consistent with the preferences of the old users.

The remainder of the paper is organized as follows: Related works are presented in Section 2. In Section 3, the problem formulation is introduced. In Section 4, we show the data collection process and in Section 5 we analyze the collected data. A background on Multi-Armed Bandit is presented in Section Section 6 and the proposed algorithms are given in Section 7. Simulation results are presented in Section 8. Finally, Section 9 concludes the paper.

2. RELATED WORK

Thermal comfort is defined as the state of mind in which occupants express satisfaction about the indoor environment. Accordingly, thermal comfort is a subjective factor that is best described by individual occupants. However, the operational settings of the HVAC systems are set based on predefined standard models for thermal comfort. These models include standard recommended thermal comfort ranges for different seasons in the simplest form to heat balance model, in which thermal comfort of the occupants is determined based on a number of environmental and human related factors. PMV-PPD (predicted mean vote and predicted percentage dissatisfied) [6] is the standard recommended heat balance model in which the collective vote of a group of occupants is represented by PMV index (a value between -3 to 3 representing cold to cool on a seven ASHARE thermal sensation scale). The PMV index has been used in several studies as the metric for user comfort integration [11, 4, 3]. As noted in the literature, the PMV index depends on a number of parameters including environmental and human related variables. Assumptions for human related variables are used in the absence of information about building occupants [4, 3]. Incorporation of these assumptions causes the index to be less representative of the dynamic occupancy characteristics in buildings. Therefore, a number of studies proposed that user provided information to be used for obtaining the metric for thermal comfort perceptions in the control logic of building systems. Controlling building systems through user provided set points has the drawback that set points in buildings are not necessarily equal to the perceived room temperatures. Moreover, user defined set points might not always lead to user comfort. Accordingly, a number of studies proposed mechanisms for learning users' comfort ranges. Guillemin and Morel used occupants' preferences in the form of temperature set points through keyboards in each room [8]. They have proposed a self-adapting control system that learns specific occupant wishes through user input in the form of set points and an artificial neural network for thermal and lighting conditions. Murakami et al. used user input for combination of binary preferences of warmer and cooler along with ASHRAE thermal sensation scale through a user interface along with their proposed logic for energy and comfort optimization called "Logic for Building a Consensus" [12]. Daum et al. used user input in the form of too hot/too cold complaints along with a probabilistic approach for determination of user comfort profiles in which, for each user, they update a default probabilistic representation of user comfort profiles using user-provided

information [5]. The profiles had three zones: comfortable, too cold, or too hot - allowing occupants to address (vote) the two extreme indoor conditions. In another approach, Bermejo et al. proposed to use static fuzzy rules for the PMV and update the thermal comfort index as occupants interacted with the thermostats in their rooms [2]. Jazizadeh et al. used a participatory sensing approach in which occupants provide their thermal preferences through a smartphone (or similar mobile device) interface and they learn occupant's comfort profile using a fuzzy pattern recognition approach as occupants interact with the interface to adjust their desired indoor conditions [10]. Although these models address the notion of personalized context dependent thermal comfort model, the problem of controlling based on the average satisfaction is a challenging task. In this study, we proposed the approach to learn personalized thermal preferences and maximize the expected collective satisfaction.

3. PROBLEM FORMULATION

We consider the thermal control problem in office buildings where the satisfaction feedback received from the occupants could be used for decision-making.

Let $n = 1, ..., N$ indicate the time steps with the horizon of N. We assume that there are M number of occupants, in the building. Each occupant has a thermal comfort profile which is unknown to us and our goal is to learn these profiles during the thermal control of the building. We consider two cases: (i) All the M occupants are present at all time steps; (ii) at each time step a subset of these M occupants are present and give their thermal feedback. In the later case, we denote the subset of present occupants at time step n by $P(n) \subset \{1, 2, ..., M\}$.

At each time step we adjust the temperature of the building to one of the possible points $t \in \{t_l, ..., t_h\}$ where t_l and t_h are the lowest and the highest temperature points, respectively. After adjusting the temperature we receive feedback from all present occupants. The feedbacks are in the form of thermal comfort preferences which are integer values such that larger positive/negative values means the occupant prefer the temperature to be warmer/cooler. These thermal comfort preferences can be converted to another measurement, the comfort proportion, which is a value in the range of $[0, 1]$ such that 0 means the occupant is uncomfortable (the condition is too hot or too cold for them) and 1 means that this is a perfect thermal condition for the occupant. Let $S_{m,t}(n)$ indicate the comfort proportion of the occupant $m = 1, ..., M$ obtained by selecting the temperature t at time step n.

The average comfort proportion at time step n is given by $S_t(n) = \sum_{m \in P(n)} S_{m,t}(n)/|P(n)|$, where $|P(n)|$ is the number of present occupants at time step n. Based on the history of the selected temperatures and the comfort proportions computed up to the current time step, a suitable temperature will be selected to adjust at the next time step. The goal is to find the optimal sequence of temperatures to select over time in order to maximize the total satisfactions. Satisfaction is defined as the summation of the comfort proportions in all time steps. We model this problem as a multi-armed bandit problem which we will explain more in details in Section 6.

4. DATA COLLECTION PROCESS

An experimental field study was conducted in order to collect user provided data for validation of the proposed algorithm through simulation of control process. The data, collected in this experiment, is used to develop user thermal comfort profiles, which are consequently used for simulating users' reaction to different temperatures in the room. In this experiment, four occupants of a multi occupancy room in an office building were asked to participate in the data collection. The building is located in Southern California. The indoor air of the room is conditioned through a variable air volume (VAV) box which is controlled through a thermostat in the room. Having access to the thermostat enabled us to expose the participants to different indoor thermal conditions. The user feedback about the indoor environment was collected through a thermal preference scale embedded in a custom user interface designed for user-HVAC interaction. The thermal preference scale is a preference slider, which enables users to provide their feedback in the form of a preference for warmer or cooler indoor environment. The thermal comfort preference scale and its asscoiated interface has been designed and evaluated to increase the consistency of occupants' votes (details could be found in [9]). The user interface is illustrated in Figure.1.

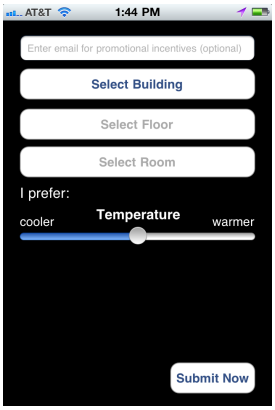

Figure 1: The user feedback interface implemented as a smartphone application

The interface was implemented as an iPhone participatory sensing application to facilitate user interaction during the day. Moreover, in order to determine the associated temperature with user feedback a temperature sensor was installed in the room. Sensor was located in the middle of the room on top of a desk to provide a representative temperature that users perceive. The layout of the room including the location of the mechanical systems and the location of the occupants is presented in Figure 2. MaxDetect, RHT03 temperature/humidity sensor was used. Temperature measurement accuracy is $\pm 0.2^{\circ}$C and the resolution (sensitivity) is 0.1°C. The sensor system uses an Arduino Black Widow stand-alone single-board microcontroller with integrated support for 802.11 WiFi communications. The data collection was carried out in September and the outside temperature was perceived as warm during the experiment. The requirement in conducting the experiment was to cover a wide range of temperature to ensure that the algorithm is tested against different conditions of perceptions.

The data was collected for four weeks during the working hours from morning to evening. The temperature range in

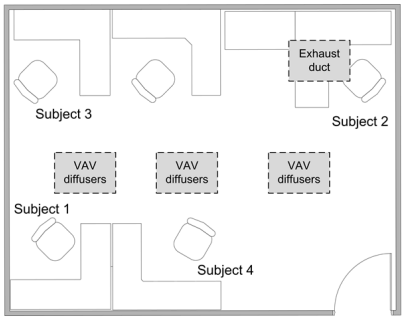

Figure 2: The layout of the office set up used as the test bed

the room was manually set between 20 and 27 °C at different times. A sample of collected data is shown in Figure 3. In this figure, the horizontal axis shows the feedback submitted through the interface and the y axis shows the ambient temperature in °C. Having collected data for about four weeks, 114, 77, 76, and 61 data points were provided by subjects 1 to 4, respectively.

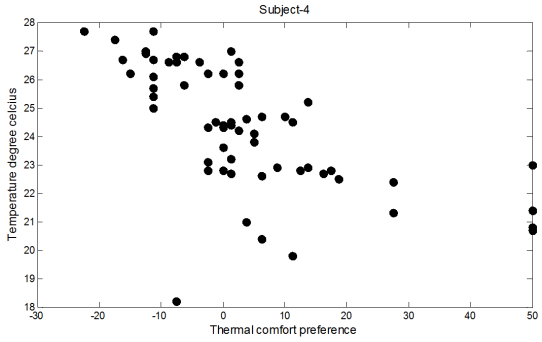

Figure 3: A sample of the collected data for one of the participants

5. DATA ANALYSES

The data is collected in the form of thermal comfort preferences (in the range of $[-50, 50]$) for temperatures in the range of $[20, 27]$ °C. We convert the thermal comfort preferences to the comfort proportions which is in the range of $[0, 1]$. Therefore if we call the thermal comfort preference x, the comfort proportation is assumed to be $S_{m,t}(n) = f(x) = (1 - \frac{|x|}{30})^+$, where $f(0) = 1$ and $f(x) = 0$ for all x lower than -30 or higher than +30.

From the collected data, we extract the mean value of the comfort proportion for each occupant $\mu_{m,t}$. Fig. 4 shows the mean values versus the temperature for all four occupants. As it's obvious from the figure, the occupant 2 is more comfortable at lower temperatures, but occupants 1 and 4 prefer higher temperatures.

For the collected data, we compute the probability mass functions of the total comfort proportions (the average of the comfort proportions of all occupants) for different temperatures, as shown in Fig. 5. This figure shows that in general the low or high temperatures are less preferable and the occupants are more comfortable at the moderate temperatures, 23-25 °C.

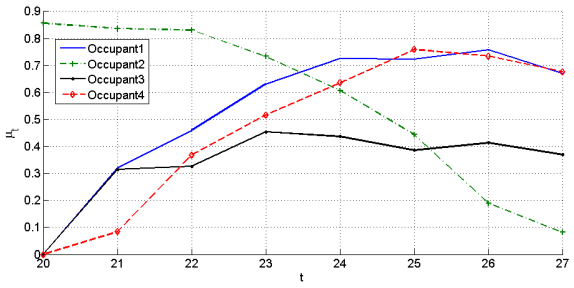

Figure 4: The mean value of satisfaction μ_t versus the temperature for different occupants.

6. MULTI-ARMED BANDIT BACKGROUND

Multi-Armed Bandit (MAB) problems are a class of sequential resource allocation problems where the resource is being selected among several alternatives. Each resource alternative is called an arm and selecting (playing) the arm results in a reward which is generated from an unkonwn distribution corresponded to that arm.

A multi-armed bandit is a type of decision-making problem where: (i) The goal is to find the best or most rewardable action (ii) The reward distribution can be updated as the experiment progresses. Therefore, this decision making problem is about which arms and in what order we should select such that the total reward collected over time horizon is maximized. In this problem we are faced with a trade-off between exploration and exploitation. Exploitation means to select the arm which has given the maximum value of reward so far, as often as possible, and exploration means to try the arms, which have not been selected enough times, to explore their reward distributions. In other words, there is a fundamental conflict between making decisions that gives high immediate reward or sacrificing current reward in order to get more information helpful for future decisions.

The classical multi-armed (K-armed) bandit process consists of one decision-maker who selects only one arm at each time step and all other arms stay frozen. This problem is defined by random variables $X_{i,n}$ for $1 \leq i \leq K$ and $n \geq 1$, where each i is the index of an arm. Sequential plays of arm i yields rewards $X_{i,1}, X_{i,2}, ...$ which are assumed to be independent and identically distributed according to an unknown distribution with unknown expectation μ_i.

A policy or allocation strategy is an algorithm that chooses the next arm to play based on the sequence of past plays and collected rewards. Let $n_i(n)$ be the number of times that arm i has been played during the first n plays. The distribuion-aware genie's policy is the repetitive sequence of the action with the highest mean value, μ^*, which achieves the highest possible expected reward. Therefore the total collected reward of the policies (without the knowledge about the distributions) is less than the maximum possible value achieved by the distribution-aware genie.

A popular measure of a policy's success is the regret, defined as the gap between the expected accumulated reward over time obtained by this policy and the one achieved by the distribution-aware genie. The regret of a policy after n plays is given by:

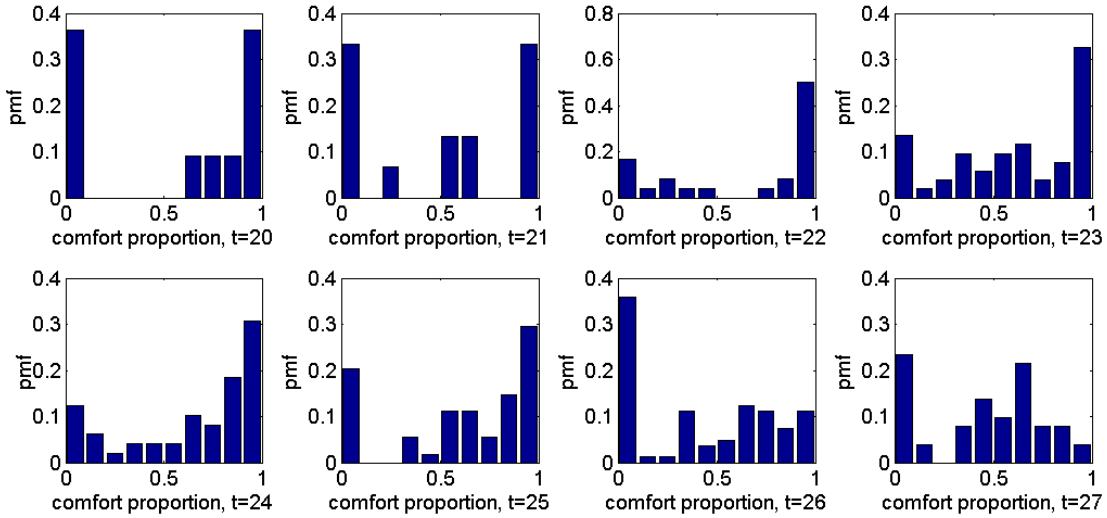

Figure 5: Probability mass function of the total comfort proportions for different temperatures.

Algotihm 1: Policy UCB1 from Auer et al. [1]

1: // Initialization
2: Play each arm once. Update $\bar{S}_t$, n_t accordingly;
3: // Main Loop
4: while 1 do
5: Play the arm t that maximizes $\bar{S}_t + \sqrt{\frac{2\ln n}{n_t}}$;
6: Update $\bar{S}_t$, n_t accordingly;
7: end while

$$\mu^* \cdot n - \sum_{j=1}^{K} \mu_j E[n_j(n)] \tag{1}$$

where $\mu^* = \max_{1 \le i \le K} \mu_i$ and $E[.]$ indicates expectation.

7. PROPOSED ALGORITHM

We use UCB1 algorithm, proposed by Auer et al. [1], in our problem to find the best temperature at each time step in order to maximize the total satisfactions over time horizon. In our Multi-Armed Bandit problem, the arms are the temperatures $t \in \{t_l, ..., t_h\}$. Therefore, the total number of arms are $t_h - t_l + 1$. The reward corresponding to the temperature (arm) t collected at time step n is equal to the comfort proportion, $S_t(n)$. The UCB1 algorithm is shown in Algorithm 1.

In this policy two variables are stored and updated each time step as an arm is selected:

(i) $n_t(n)$, the number of times that the temperature t has been selected up to the time step n.

(ii) $\bar{S}_t(n)$, the sample mean of the comfort proportion rewards collected by selecting temperature t up to the current time step n.

$\bar{S}_t$ and n_t are both initialized to 0 and updated as follows:

$$\bar{S}_t(n) = \begin{cases} \frac{\bar{S}_t(n-1)n_t(n-1)+S_t(n)}{n_t(n-1)+1} & \text{if arm t is played,} \\ \bar{S}_t(n-1) & \text{else.} \end{cases} \tag{2}$$

$$n_t(n) = \begin{cases} n_t(n-1)+1 & \text{if arm t is played,} \\ n_t(n-1) & \text{else.} \end{cases} \tag{3}$$

The trade-off between exploration and exploitation is taken into account in the line 5 of the above algorithm where the first term tries to select the arm with higher average reward (exploitation) and the second term tries to select the arm with smaller n_t, *i.e.* less number of selection (exploration). Therefore the best temperature to select at time step $n+1$ is given by:

$$t^* = \arg\max_t [\bar{S}_t + \sqrt{\frac{2\ln(n)}{n_t}}], \tag{4}$$

where $\bar{S}_t$ and n_t are the updated value at time steps n as given by (2) and (3).

For the case of dynamic occupancy, we propose to use the Learning with Linear Rewards (LLR) algorithm, proposed recently by Gai *et. al.* [7]. The LLR algorithm is shown in Algorithm 2. The basic idea is to track the sample means and number of times played not for the temperature values, but for each user and each temperature value. Then, for a given set of occupants at any time (it is assumed that the current occupants of the room can be correctly identified, for instance, through their mobile device ID's or through some other identifying mechanism), a corresponding index is computed only for those occupants at each temperature, to decide, on the fly, which temperature should be selected.

This algorithm stores and updates $\bar{S}_{m,t}$ and $n_{m,t}$ for different occupants $m \in \{1, ..., M\}$ and different temperature $t \in \{t-l, t_{l+1}, ..., t_h\}$ which requires only polynomial storage. Updating of $\bar{S}_{m,t}$ and $n_{m,t}$ at time step n are similar to (2) and (3) for only the occupants $m \in P(n)$ and for the rest of occupants these values stay fixed. The index that is computed in line 5 for each temperature now considers the sum of corresponding indices for each present occupant. M is a parameter in this problem that can be conservatively

Algotihm 2: Policy LLR from Gai *et. al.* [7]

1: // Initialization
2: Play each arm t once. Update $\bar{S}_{m,t}$, $n_{m,t}$ for all $m \in P(i), i = 1, ..., t$ accordingly;
3: // Main Loop
4: while 1 do
5: Play the arm t that maximizes $\sum_{m \in P(n)} \bar{S}_{m,t} + \sqrt{\frac{M \ln n}{n_{m,t}}}$;
6: Update $\bar{S}_{m,t}$, $n_{m,t}$, for all $m \in P(n)$ accordingly;
7: end while

set to the maximum number of users.

8. SIMULATION RESULTS

We apply the UCB1 and LLR algorithms, given in previous section, to the thermal control problem for constant and dynamic occupancy, respectively. We use the real data to regenerate the data samples during the simulation.

8.1 Constant Occupancy

If all of occupants are present in all time steps to give us feedback about their comfort, we can use UCB1 algorithm. Fig. 6 shows the total satisfaction collected up to time n versus the time n, for two policies: (i) the UCB1 algorithm and (ii) the distribution-aware genie which has the knowledge about the mean values of comfort proportions. The distribution-aware genie always selects the arm with the highest mean value, which for our data set is $t = 25$ °C.

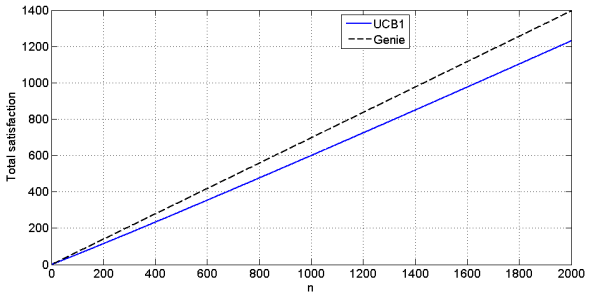

Figure 6: Total satisfaction of UCB1 algorithm and the distribution-aware genie's policy (playing always the best arm) versus the passed time.

Fig. 7 shows the regret of the UCB1 algorithm versus time steps n which is the gap between total satisfactions of UCB1 and genie's policy given in Fig. 6. Fig. 8 shows the loss of the UCB1 algorithm which is the difference in unit-time satisfaction obtained by this learning policy compared to genie' optimal policy, versus time steps n. The loss function at time step n is given by $E[S_{t^*}(n) - S_t(n)]$. As figure shows, the loss (dissatisfaction) will decrease to less than 10% after passing 400 time steps and goes to zero by increasing the time steps.

Fig. 9 shows the fraction of times that a temperature has been selected in UCB1 algorithm versus the mean value of comfort proportion. As it is obvious from the figure, the temperature with higher mean value has a higher chance to be selected.

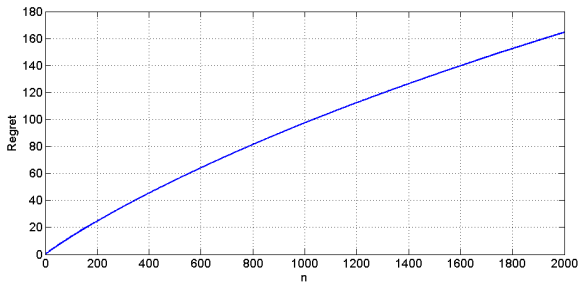

Figure 7: Regret of UCB1 algorithm versus the time index.

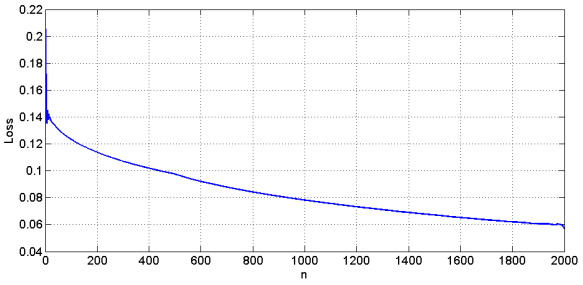

Figure 8: Unit-time loss of UCB1 algorithm versus the time index.

The sample mean of comfort proportions corresponded to different temperatures is updated as (2) and it is going to converge to the mean value of the total comfort proportion, μ_t, shown in Fig. 10. Note that μ_t is the average of the mean values of the comfort proportions of all occupants, $\mu_{m,t}, m = 1, ..., M$ shown in Fig. 4, *i.e.,* $\mu_t = \sum_{m=1}^{M} \mu_{m,t}/M$. Fig. 9 shows $\bar{S}_t$ achieved by UCB1 algorithm at the horizon N versus μ_t and it confirms that the sample means converge to the comfort proportion mean values. Note that in this figure, each pair of $(\mu_t, \bar{S}_t)$ corresponds to one of the temperatures in the range of $[t_l, t_h]$

8.2 Dynamic Occupancy

For the case that the number of occupants are varying over time, the UCB1 algorithm is not applicabale, because the data from some of the occupants are not available. We could use LLR algorithm given in previous section. To simulate the dynamic occupancy using the data collected from $M = 4$ occupants, we assume that at each time step all of 2^4 subset of occupants are possible and we choose one of them uniformly. The feedback received from this subset of occupants will determine the best temperature for the next time step.

Fig. 11 shows the total satisfaction collected up to time n versus the time n, for two policies: (i) the LLR algorithm and (ii) the policy of the genie who is aware of the subset of present occupants and the mean values of their comfort proportions at each time step. The distribution-aware genie always selects the arm with the highest mean value for the present occupants in $P(n)$.

Fig. 12 and Fig. 13 show the regret and the unit-time loss of the LLR algorithm compared to the genie's policy, respectively. Comparing them with Fig. 7 and Fig. 8 shows

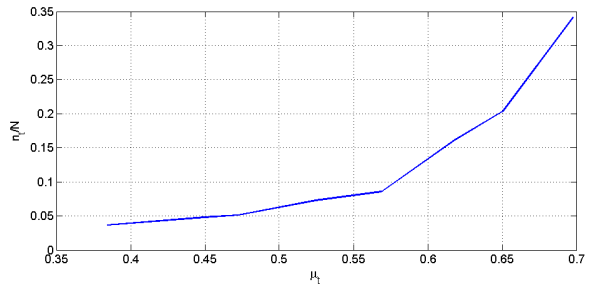

Figure 9: The number of selection of the temperature, n_t versus the mean value of comfort proportion.

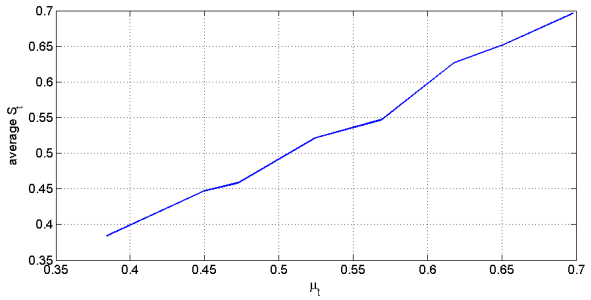

Figure 10: The sample mean of satisfaction $\bar{S}_t(N)$ versus the mean value of comfort proportion, μ_t.

that the regret and loss of the dynamic occupancy case are lower than that of the constant occupancy.

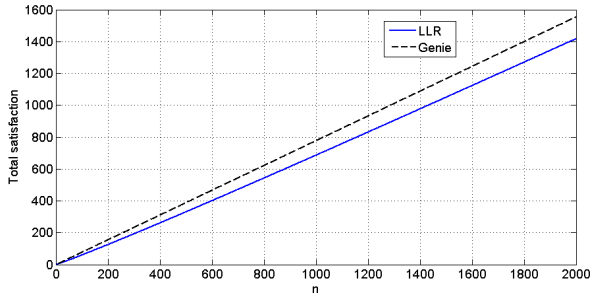

Figure 11: Total satisfaction of LLR algorithm and the distribution-aware genie's policy for dynamic occupancy versus the passed time.

9. CONCLUSION

This study has shown how feedback obtained from users of a participatory sensing app deployed in multi-occupant spaces can be used to automatically learn the best temperature setting to maximize average user satisfaction. We have shown that this can be done in a personalized fashion, taking into account the individual preference of each user. Our primary contribution is to show that the problem of online learning of thermal control settings for even a dynamic population of users with exponential combinations can be handled efficiently, by applying a state-of-the-art combinatorial

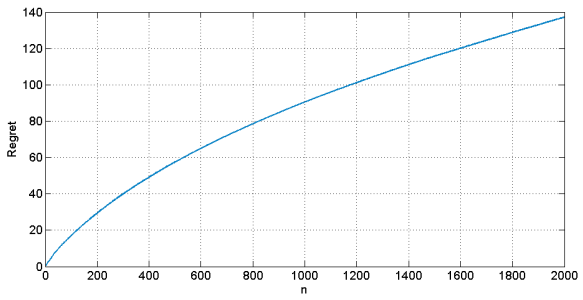

Figure 12: Regret of LLR algorithm for dynamic occupancy versus the time index.

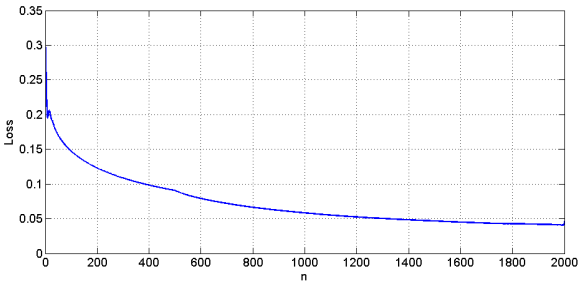

Figure 13: Unit-time loss of LLR algorithm for dynamic occupancy versus the time index.

multi-armed bandit algorithm. Furthermore, we have empirically validated this claim via simulations based on real user data. Since users provide their feedback overtime and they have control over the environment the proposed method enables the integration of the contextual information such as user clothing level and metabolic rates benefiting from the user adaptiveness ability. Moreover, this approach, in its current format, is suited for building zones with permanent occupancy. In this way, the preferences of the occupants could be learned over time while the occupants comfort is preserved. However, this method requires user cooperation in provision of satisfaction and dis-satisfaction feedback during the training for a number of times per day. Our observations in the field experiments show that if the users are provided with a personalized control system, they are willing to participate in feedback provision.

For future work, we would like to consider objectives other than maximizing the average user satisfaction. Depending on the setting, in some cases a prioritized, possible even nonlinear, may be preferable (e.g., giving higher consideration to the comfort of longer-term occupants). Another possibility is to refine the user satisfaction model based on sensed user activity — for instance, a user may have a different temperature preference after having just performed vigorous exercise than when engaged in sedentary activity. Integrating information from automated personal activity state sensing (e.g. as proposed in [13]) into the smartphone application could allow for even more fine-grained sensing and control.

Also, the current policies assume no prior information is available about the user preferences. It is conceivable that previously learned information about the users' preferences in other spaces could be used to speed up learning. In future

work, we also plan to consider a Bayesian approach that allows the prior preferences to be taken into account in the learning process. Finally, we hope to be able to deploy the proposed thermal control system live in a real environment and conduct experiments to test its performance.

Acknowledgment and Disclaimer

This material is based upon work supported by the National Science Foundation under Grant No. 1201198. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2-3 (2002), pp. 235–256.

[2] Pablo Bermejo et al. "Design and simulation of a thermal comfort adaptive system based on fuzzy logic and on-line learning". In: *Energy and Buildings* 49 (2012), pp. 367–379.

[3] Francesco Calvino et al. "The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller". In: *Energy and Buildings* 36.2 (2004), pp. 97–102.

[4] K Dalamagkidis et al. "Reinforcement learning for energy conservation and comfort in buildings". In: *Building and environment* 42.7 (2007), pp. 2686–2698.

[5] David Daum, Frédéric Haldi, and Nicolas Morel. "A personalized measure of thermal comfort for building controls". In: *Building and Environment* 46.1 (2011), pp. 3–11.

[6] Poul O Fanger et al. "Thermal comfort. Analysis and applications in environmental engineering." In: *Thermal comfort. Analysis and applications in environmental engineering.* (1970).

[7] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations". In: *IEEE/ACM Transactions on Networking (TON)* 20.5 (2012), pp. 1466–1478.

[8] A Guillemin and N Morel. "Experimental results of a self-adaptive integrated control system in buildings: a pilot study". In: *Solar Energy* 72.5 (2002), pp. 397–403.

[9] Farrokh Jazizadeh, Franco Moiso Marin, and Burcin Becerik-Gerber. "A thermal preference scale for personalized comfort profile identification via participatory sensing". In: *Building and Environment* 68.0 (2013), pp. 140 –149.

[10] Farrokh Jazizadeh et al. "A Human-Building Interaction Framework for Personalized Thermal Comfort Driven Systems in Office Buildings". In: *Journal of Computing in Civil Engineering* (2013).

[11] D Kolokotsa et al. "Design and installation of an advanced EIB fuzzy indoor comfort controller using Matlab". In: *Energy and buildings* 38.9 (2006), pp. 1084–1092.

[12] Yoshifumi Murakami et al. "Field experiments on energy consumption and thermal comfort in the office environment controlled by occupantsâĂŹ requirements from PC terminal". In: *Building and Environment* 42.12 (2007), pp. 4022–4027.

[13] Yi Wang et al. "A framework of energy efficient mobile sensing for automatic user state recognition". In: *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM. 2009, pp. 179–192.