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

Parisa Mansourifard
Farrokh Jazizadeh, Bhaskar Krishnamachari, Burcin Becerik-Gerber
Ming Hsieh Dept. of Electrical Engineering, and
Sony Astani Dept. of Civil and Environmental Engineering
University of Southern California

November 14, 2013
Introduction

- **problem**: automatically learning the optimal thermal control in a room
- **goal**: to maximize the expected average satisfaction among occupants
- **Occupants provide stochastic feedback on their comfort through a participatory sensing application**
- **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.**
Problem Formulation

Assumptions:

- the discrete time steps: \( n = 1, 2, \ldots, N \).
- number of occupants \( M \) in the building
- occupant’s thermal comfort profile unknown to us
- Goal: learn their profiles during the thermal control of the building
- two cases:
  - (i) all the \( M \) occupants are present at all time steps
  - (ii) a subset of them are present and give their thermal feedback, denoted by \( P(n) \subset \{1, \ldots, M\} \)
At each time step, we adjust the temperature to one of the possible points \( t \in \{ t_l, \ldots, t_h \} \)
- \( t_l \) and \( t_h \) are the lowest and the highest temperature points

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
- 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,
- a value in the range of $[0, 1]$
- 0 means the occupant is uncomfortable (the condition is too hot or too cold for them)
- 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

$$S_t(n) = \sum_{m \in P(n)} S_{m,t}(n) / |P(n)|$$

- $|P(n)|$ 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.

- Total satisfaction is defined as the summation of the average comfort proportions in all time steps.

- We model this problem as a multi-armed bandit problem
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.
Data Collection Process

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

Figure: A sample of the collected data for one of the participants
The temperature range in the room was manually set between 20 and 27 °C at different times.

The data is collected in the form of thermal comfort preferences (in the range of $[-50, 50]$)

if we call the thermal comfort preference $x$, the comfort proportion is assumed to be $S_{m,t}(n) = f(x) = (1 - \frac{|x|}{30})^+$. 

Data analysis

Figure: The mean value of comfort proportion, $\mu_t$, versus the temperature for different occupants.

From the collected data, we extract the mean value of the comfort proportion for each occupant $\mu_{m,t}$. 
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 unknown distribution corresponded to that arm.

The 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.
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.

In this problem we are faced with a trade-off between exploration and exploitation.

The distribution-aware genie’s policy is the repetitive sequence of the action with the highest mean value, $\mu^*$.

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:

$$\mu^* \cdot n - \sum_{j=1}^{K} \mu_j E[n_j(n)]$$ (1)
We use UCB1 algorithm, proposed by Auer et al., 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 $t \in \{t_l, \ldots, t_h\}$. Therefore, the total number of arms are $T = t_h - t_l + 1$.

The reward corresponding to the temperature (arm) $t$ collected at time step $n$ is equal to the average comfort proportion, $S_t(n)$.

---

**Algorithm 1**: Policy UCB1

1: // Initialization
2: Play each arm once. Update $\bar{S}_t, n_t$ accordingly; $n \leftarrow T$
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; $n \leftarrow n + 1$
7: end while
Simulation Results: Constant Occupancy

Figure: Total satisfaction and Regret of UCB1 algorithm and the distribution-aware genie’s policy (playing always the best arm) versus the passed time for $M = 4$. 
Simulation Results: Constant Occupancy

Figure: The time fraction of selection that the temperature $t$ is selected, $n_t/N$ versus the mean value of comfort proportion (average of $\mu_{m,ts}$).
Proposed Algorithm for dynamic occupancy

- use the Learning with Linear Rewards (LLR) algorithm, proposed recently by Gai et. al.
- 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.

**Algorithm 2: Policy LLR**

1: // Initialization
2: Play each arm $t$ once. Update $\bar{S}_{m,t}$, $n_{m,t}$ for all $m \in P(i)$, $i = 1, \ldots, T$ accordingly; $n \leftarrow T$
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; $n \leftarrow n + 1$
7: end while
Simulation Results: Dynamic Occupancy

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.

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

- 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 (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.

- We have empirically validated this claim via simulations based on real user data.
Conclusions and Future Work

Future work

- to consider objectives other than maximizing the average user satisfaction.
- in some cases a prioritized, possible even non-linear, may be preferable (e.g., giving higher consideration to the comfort of longer-term occupants).
- to refine the user satisfaction model based on sensed user activity.
- 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.
- to deploy the proposed thermal control system live in a real environment and conduct experiments to test its performance.
Thank you for your attention