

Human-in-the-loop: Real-time Preference Optimization

Wenbin Wang, Wenjie Xu, Colin N. Jones

Abstract—Optimization with preference feedback is an active research area with many applications in engineering systems where humans play a central role, such as building control and autonomous vehicles. While most existing studies focus on optimizing a static user utility, few have investigated its closed-loop behavior that accounts for system transients. In this work, we propose an online feedback optimization controller that can optimize user utility using pairwise comparison feedback with both optimality and closed-loop stability guarantees. By adding a random exploration signal, the controller estimates the gradient based on the binary utility comparison feedback between two consecutive time steps. We analyze its closed-loop behavior when interacting with a nonlinear plant and show that, under mild assumptions, the controller converges to the optimal point without inducing instability. Theoretical findings are further validated through numerical experiments.

I. INTRODUCTION

Humans are the key components in engineering systems and the primary beneficiaries of many leading technologies, such as building control [1] and human-robot collaboration [2]. It is essential to design a human-aware controller capable of regulating the system in real time to optimize user latent utility. Existing controllers typically track a predefined reference, which is often derived from large population models, e.g., indoor room temperature for building control [3]. While being simple and easy to implement, this can introduce bias and lead to suboptimal performance, as it fails to account for individual differences. Moreover, without real-time human feedback, such a controller cannot respond to time-varying utility and is not robust to external disturbances.

As an emerging real-time control technique, Online Feedback Optimization (OFO) [4] has been effective in applications such as grid control [5] and robot coordination [6]. By taking real-time system output, OFO can navigate the system toward its optimal point without requiring precise knowledge of the plant dynamics or measurement noise. Theoretical guarantees have been established for both first-order [7] and zeroth-order [8] formulations. However, people are generally more adept at making relative comparisons than providing absolute evaluations of their utility [9]. As a result, human feedback often appears as pairwise comparisons, making it difficult to directly apply existing OFO schemes to design human-aware controllers with closed-loop guarantees.

Research on offline optimization with preference feedback has been rapidly expanding. In the finite action setting, preferences can be encoded in a preference matrix, where

each entry gives the probability that one option is preferred over another. Notions of optimality such as the Copeland winner [10] and the Borda winner [11] have been studied, and algorithms based on random exploration [12] and greedy search [13] are employed to identify the optimal choice. For continuous action spaces, a common assumption is that preferences are induced by a latent utility, which is often modeled with Gaussian Processes (GP). Heuristic strategies for sequential decision-making balance exploration and exploitation [14], [15], while regret guarantees can be established when the utility lies in a Reproducing Kernel Hilbert Space (RKHS) [16], [17]. In parallel, gradient-estimation methods from preference feedback have been explored [18], [19], with optimality guarantees under regularity assumptions such as smoothness and convexity [19].

Despite recent advancements, most existing work considers a static problem and neglects transient system behavior. Many open challenges remain in developing online preference optimization algorithms that account for system transients. First, the stability of the closed-loop system is hard to quantify, as the binary nature of preference feedback renders the overall dynamics highly nonlinear. Moreover, people typically have limited knowledge about the plant dynamics, implying that directly following their preferences can drive the system toward instability. Second, it is difficult for a controller trained on offline data to track the optimal point in real time, as the utility function can be time-varying, subject to external disturbances, and environment-dependent [9]. The way different options are presented also has an impact on how people express their preferences, given that people are not always rational [9]. A carefully designed mechanism for how humans interact with the controller is necessary to ensure stable and efficient system operation.

Inspired by the recently proposed model-free OFO with one-point residual estimation [8], we introduce a novel OFO controller that leverages binary preference feedback to optimize user utility while ensuring closed-loop stability. To the best of our knowledge, it is the first work addressing the real-time preference optimization problem with closed-loop guarantees. Our approach employs a randomized exploration scheme, in which a random exploration signal is added into the system at each time step. Preference feedback is then collected based on the perturbed utilities between consecutive time steps. Unlike existing approaches [18], [19] that require two function evaluations at each time step, our method requires only one evaluation, making it well-suited for online implementation. Under mild assumptions, we show that the resulting update imitates a gradient descent step and provide theoretical guarantees on the stability and optimality of the

The authors are with Automatic Control Laboratory, EPFL, Switzerland. Email: wenbin.wang@epfl.ch, wenjie.xu@epfl.ch, colin.jones@epfl.ch. This work was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545 and the Swiss Federal Office of Energy SFOE as part of the SWEET consortium SWICE.

closed-loop system. Theoretical results are further supported through numerical experiments on a thermal comfort optimization problem.

II. PROBLEM FORMULATION AND PRELIMINARIES

A. Notation

Let $\mathbb{R}^n$ be the n -dimensional Euclidean space. For a real-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we denote its gradient at $x \in \mathbb{R}^n$ by $\nabla f(x)$. For a single input function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, we denote its derivative by $\sigma'(t)$. The value of vector $x \in \mathbb{R}^n$ at time step k is written as x_k .

B. Problem Formulation

Consider an exponentially stable plant with $n_x \in \mathbb{N}$ states and $n_u \in \mathbb{N}$ inputs

$$x_{k+1} = f(x_k, u_k), \quad (1)$$

where $f : \mathbb{R}^{n_x \times n_u} \rightarrow \mathbb{R}^{n_x}$ is the state transition function, $x \in \mathbb{R}^{n_x}$ is the plant state, $u_k \in \mathbb{R}^{n_u}$ is the plant input. We assume that there exists a unique steady-state input-state map $h : \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$ such that $\forall u \in \mathbb{R}^{n_u}$, $h(u) = f(h(u), u)$.

Assumption 1. *The input-state map h is L_h -Lipschitz continuous.*

The assumption of an exponentially stable plant with Lipschitz continuous input-state map is satisfied for many systems, e.g., linear system $x_{k+1} = Ax_k + Bu_k$ with A being invertible and stable. The results of this work can also be extended to include the output measurement $y_k = g(x_k, u_k)$, where additional assumptions such as Lipschitz continuous input-output map are necessary.

According to the converse Lyapunov theorem [8], [20, p. 194], for an exponentially stable plant, there exists a Lyapunov function $V : \mathbb{R}^{n_x \times n_u} \rightarrow \mathbb{R}$ and positive parameters $\alpha_1, \alpha_2, \alpha_3$ such that $\forall u \in \mathbb{R}^{n_u}$,

$$\alpha_1 \|x - h(u)\|^2 \leq V(x, u) \leq \alpha_2 \|x - h(u)\|^2, \quad (2)$$

$$V(f(x, u), u) - V(x, u) \leq -\alpha_3 \|x - h(u)\|^2. \quad (3)$$

The parameter α_3 characterizes how fast the system stabilizes. A larger α_3 indicates the system stabilizes to $h(u)$ with a larger rate.

In this work, we want to optimize user utility when the plant is at the steady state

$$\begin{aligned} \min_{x, u} \quad & \Phi(x, u), \\ \text{s.t.} \quad & x = h(u), \end{aligned} \quad (4)$$

where $\Phi : \mathbb{R}^{n_x \times n_u} \rightarrow \mathbb{R}$ is the latent utility function. This utility function depends on both the system state and input, capturing real-world scenarios where these factors jointly have influences. For example, in building energy management, there is an inherent trade-off between thermal comfort and control effort [1].

Assumption 2. $\forall u \in \mathbb{R}^{n_u}$, the latent utility $\Phi(x, u)$ is L_x -Lipschitz with respect to the first argument, i.e., $\|\Phi(x_1, u) - \Phi(x_2, u)\| \leq L_x \|x_1 - x_2\|, \forall x_1, x_2 \in \mathbb{R}^{n_x}$, where L_x is a uniform Lipschitz constant.

We can reformulate (4) as an unconstrained problem by replacing x with $x = h(u)$ in $\Phi(x, u)$ and get

$$\min_u \tilde{\Phi}(u), \quad (5)$$

where $\tilde{\Phi}(u) = \Phi(h(u), u)$.

Assumption 3. *The differentiable function $\tilde{\Phi}(u)$ is L_0 -Lipschitz, L_1 -smooth and m -strongly convex.*

Assumption 3 is commonly employed for theoretical analysis [21]–[23] and satisfied in applications, e.g., building control [1] and frequency regulation [24]. We denote the solution of (5) by u^* . The assumption on strongly convexity assures that u^* is unique.

Classical numerical optimization methods, such as gradient descent, require the knowledge of h to find the optimal solution [4]. However, this is impractical for systems where constructing an accurate model is costly. For this, zeroth-order feedback optimization methods have been developed to solve (4). These methods estimate the gradient using finite differences, either by considering plant transients [8] or assuming an algebraic plant [22].

Nevertheless, finite difference fails with preference feedback since the true value of $\Phi(x, u)$ is not accessible. Instead, preference feedback often appears in the form of pairwise comparisons, which are binary signals. For the problem to be properly defined, we need to link such binary feedback with the utility through a user model.

C. User model

Given u_1 and u_2 , we denote the event ' $\tilde{\Phi}(u_1) < \tilde{\Phi}(u_2)$ ' by $u_1 \succ u_2$, i.e., u_1 is better than u_2 . The corresponding preference feedback is written as $\mathbf{1}_{u_1 \succ u_2}$, which is

$$\mathbf{1}_{u_1 \succ u_2} = \begin{cases} 1, & \text{if } u_1 \text{ is better,} \\ -1, & \text{if } u_2 \text{ is better.} \end{cases} \quad (6)$$

To link the utility $\tilde{\Phi}$ with preference feedback, we adopt a probabilistic model as shown in Assumption 4.

Assumption 4. *The preference feedback $\mathbf{1}_{u_1 \succ u_2}$ follows a Bernoulli distribution, i.e., $\mathbb{P}(\mathbf{1}_{u_1 \succ u_2} = 1) = \sigma(\tilde{\Phi}(u_2) - \tilde{\Phi}(u_1))$, where $\sigma(t) = \frac{1}{1+e^{-t}}$.*

This model, i.e., the Bradley-Terry Model, among other probabilistic models such as the Thurstone-Mosteller model [25, Section 2.2.3], are commonly found in the literature for preferential learning [16], [18]. The idea is that when $\tilde{\Phi}(u_1)$ is small, the probability of reporting $\mathbf{1}_{u_1 \succ u_2} = 1$ is large. We denote the Lipschitz constant and smoothness constant of $\sigma(t)$ by $L_{\sigma,0}, L_{\sigma,1}$, respectively. The result in this work also holds with other $\sigma(t)$, provided they are monotonically increasing, rotation-symmetric, satisfy $\sigma(-\infty) = 0, \sigma(\infty) = 1$, and exhibit convexity for $x \leq 0$, essentially behaving like cumulative distribution functions [18].

III. ONLINE GRADIENT ESTIMATION WITH PREFERENCE FEEDBACK

We build on the existing result in model-free feedback optimization [8] and design a controller that establishes a

Algorithm 1 Controller with comparison feedback

Input: step size η , smoothing parameter δ , $u_0 \in \mathbb{R}^{n_u}$, number of time steps T

```
1: for  $k = 1, \dots, T - 1$  do
2:    $x_{k+1} = f(x_k, u_k + \delta v_k)$ ;
3:   ask users to express their preference between
      $\Phi(x_{k+1}, u_k + \delta v_k)$  and  $\Phi(x_k, u_{k-1} + \delta v_{k-1})$ , sample
      $\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})}$ ;
4:   update the input via
      $u_{k+1} = u_k + \frac{\eta}{2\delta} \mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} v_k$ ;
5: end for
```

Output: u_T

gradient estimator with preference feedback from real-time state measurement. Different from existing preference optimization methods, our formulation takes system transients into consideration. We assume that the binary preference feedback is collected from the user between two consecutive utility evaluations. The controller then regulates the system toward the point where the utility is minimized. This is summarized in Algorithm 1.

At time step k , a random exploration signal v_k is added to the system as shown in Line 2 of Algorithm 1. It is drawn i.i.d. from the $(n_u - 1)$ -dimensional unit sphere $\mathbb{S}^{n_u-1}$ uniformly. We then use x_{k+1} as an approximation of $h(u_k + \delta v_k)$. Eventually, this allows us to approximate $\tilde{\Phi}(u_k + \delta v_k)$ with $\Phi(x_{k+1}, u_k + \delta v_k)$.

In Line 3, the user is asked to provide a preference between the pairs $\Phi(x_{k+1}, u_k + \delta v_k)$ and $\Phi(x_k, u_{k-1} + \delta v_{k-1})$, i.e., sampling the random variable $\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})}$. We slightly abuse the notation here, denoting the event ' $\Phi(x_{k+1}, u_k + \delta v_k) < \Phi(x_k, u_{k-1} + \delta v_{k-1})$ ' by $(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})$ and representing the preference feedback via the random variable $\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})}$ with probability $\mathbb{P}(\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} = 1) = \sigma(\Phi(x_k, u_{k-1} + \delta v_{k-1}) - \Phi(x_{k+1}, u_k + \delta v_k))$.

Finally, the control input u_k is updated as shown in Line 4. When $\Phi(x_{k+1}, u_k + \delta v_k)$ has a lower value, $\mathbb{P}(\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} = 1)$ is higher, implying that, with high probability, the algorithm updates u_{k+1} toward $u_k + \delta v_k$, where the utility attains a smaller value.

The resulting closed-loop system takes the form

$$x_{k+1} = f(x_k, u_k + \delta v_k), \quad (7)$$

$$u_{k+1} = u_k + \frac{\eta}{2\delta} \mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} v_k. \quad (8)$$

The random perturbation v_k facilitates exploration of multiple directions at each time step. The collected preference feedback indicates the descent direction of the unknown utility, effectively imitating a stochastic gradient descent step. Although (8) does not explicitly estimate $\nabla \tilde{\Phi}(u)$ since the preference feedback lacks information about utility values, we will demonstrate that (8) effectively performs gradient

descent on a probability function, ultimately converging toward the optimal point in Section IV.

IV. PERFORMANCE ANALYSIS

We first show that the closed-loop system is stable.

Lemma 1. *Let Assumptions 1-4 hold. $\mathbb{E}[V(x_k, u_k + \delta v_k)] \leq \mu^k \mathbb{E}[V(x_0, u_0 + \delta v_0)] + \frac{a_1}{1-\mu} (2\delta^2 + \eta + (\frac{\eta}{2\delta})^2)$, where $\mu = \frac{2\alpha_2}{\alpha_1} (1 - \frac{\alpha_3}{\alpha_2})$, and $a_1 = 4\alpha_2 L_h^2$.*

Proof. Proof can be found in Appendix A. $\square$

Remark 1. *We consider the expected Lyapunov function $\mathbb{E}[V(x_k, u_k + \delta v_k)]$ as a stability indicator. The decay rate μ is a function of $\alpha_1, \alpha_2, \alpha_3$, which characterizes how quickly the system stabilizes to an equilibrium state under a constant input. A smaller μ implies that the system stabilizes more rapidly to the equilibrium point. According to Lemma 1, $\mathbb{E}[V(x_k, u_k + \delta v_k)]$ is upper-bounded by an exponentially decaying term and a constant. This follows from the fact that the increment of $\|u_k\|$ is bounded by $\frac{\eta}{2\delta}$, which is summable for an exponentially stable plant. Consequently, the system states remain bounded.*

To analyze optimality, we write

$$p_{u'}(u) = \mathbb{P}(\mathbf{1}_{u' \succ u} = 1) = \sigma(\tilde{\Phi}(u) - \tilde{\Phi}(u'))$$

to simplify the notation. A smaller value of $\tilde{\Phi}(u)$ yields a lower $p_{u'}(u)$. This monotonic relationship allows us to equivalently minimize $p_{u'}(u)$, rather than minimizing $\tilde{\Phi}(u)$ directly. This is further supported by Lemma 2.

Lemma 2. *Let Assumptions 1-4 hold. $\forall u' \in \mathbb{R}^{n_u}$, $p_{u'}(u)$ is $L_{p,0}$ -Lipschitz, and $L_{p,1}$ -smooth with respect to u , where $L_{p,0} = L_{\sigma,0} L_0$, $L_{p,1} = \sigma'(0) L_1 + L_{\sigma,1} L_0^2$. Furthermore, $p_{u'}(u)$ is partially convex with respect to u if $\tilde{\Phi}(u) \leq \tilde{\Phi}(u')$.*

Proof. Proof on partially-convexity can be found in [18]. The rest of the proof can be found in Appendix B. $\square$

Remark 2. *Since $\sigma(t)$ is a bounded and smooth function, the Lipschitz continuity and smoothness properties are preserved for the composed function $p_{u'}(u)$. Convexity is also partially preserved since $\sigma(t)$ is partially convex.*

Under the assumption of strong convexity, we show that, for any fixed u' , the minimizer of $p_{u'}(u)$ is unique and coincides with u^* .

Lemma 3. *Let Assumptions 1-4 hold and u^* be the unique solution of (5), then $\forall u' \in \mathbb{R}^{n_u}$, u^* is also the unique minimizer of $p_{u'}(u)$.*

Proof. Proof can be found in Appendix C. $\square$

Based on Lemmas 2 and 3, together with standard techniques from convex analysis, we can show that gradient descent with respect to $p_{u'}(u)$

$$u_{k+1} = u_k - \eta \nabla p_{u_k}(u_k) \quad (9)$$

converges to u^* , where $\nabla p_{u_k}(u_k)$ denotes the gradient of $p_{u_k}(u)$ with respect to u evaluated at u_k . Update (9)

corresponds to the ideal case in which the function h is known. In other words, it can be shown that $\nabla\Phi(u_k)$ and $\nabla p_{u_k}(u_k)$ are proportional up to a constant scaling factor. Building on this evidence, we compactly write (8) as

$$u_{k+1} = u_k - \eta(\nabla p_{u_k}(u_k) + e_k), \quad (10)$$

where the error term e_k is defined as

$$e_k = -\frac{1}{2\delta} \mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} v_k - \nabla p_{u_k}(u_k). \quad (11)$$

In (10), we interpret (8) as an instance of (9) with an additional error e_k . It exists because $h(u_k + \delta v_k)$ is approximated by x_{k+1} , and the gradient is estimated via random perturbations. The boundedness of e_k is established in Lemma 4.

Lemma 4. *Let Assumptions 1-4 hold. $\|\mathbb{E}[e_k | \mathcal{F}_k]\| \leq \sqrt{R_1 V(x_{k-1}, u_{k-1} + \delta v_{k-1}) + R_2}$, where $R_1 = \frac{2L_{\sigma,0}^2 L_x^2 (\mu+1)}{\alpha_2 \delta^2} \mu$, and $R_2 = \frac{2L_{\sigma,0}^2 L_x^2 a_1}{\alpha_2 \delta^2} (2\delta^2 + \eta + (\frac{\eta}{2\delta})^2) \mu + 2a_2^2 \delta^2$, $a_2 = L_{p,1} \sqrt{n} + (\sigma'(0)L_1 + L_{p,1})(1 + \frac{\eta}{\delta})$*

Proof. Proof can be found in Appendix D. $\square$

In Lemma 4, we observe that at time step k , conditioned on the natural filtration $\mathcal{F}_k$, the error e_k is bounded by $R_1 V(x_{k-1}, u_{k-1} + \delta v_{k-1})$ and R_2 , where $V(x_{k-1}, u_{k-1} + \delta v_{k-1})$ is bounded as established in Lemma 1. Additionally, we observe that $R_1 = \mathcal{O}(\mu)$ and $R_2 = \mathcal{O}(\mu, \delta^2)$. Both of them decrease as μ decreases. This result is consistent with the approximation discussed in Section III, as a smaller μ enables x_{k+1} to more accurately approximate $h(u_k + \delta v_k)$.

Now we are ready to present the main convergence result.

Theorem 1. *$\forall k' \text{ and } \forall k > k'$, the expected distance to u^* is bounded, i.e., $\mathbb{E}[\|u_k - u^*\|^2] \leq (\frac{1+\rho}{2})^{k-k'} \mathbb{E}[\|u_{k'} - u^*\|^2] + \mathcal{O}(\mu, \mu^{k'}, \delta)$, where $\rho = 1 - 2\sigma'(0)m\eta$.*

Proof. Proof can be found in Appendix E. $\square$

We adopt the expected distance to u^* , i.e., $\mathbb{E}[\|u_k - u^*\|^2]$, as the error metric, which is commonly used in the literature on online zeroth-order optimization [22], [26]. Our results indicate that, for any fixed time step k' , $\mathbb{E}[\|u_k - u^*\|^2]$ is bounded by $\mathbb{E}[\|u_{k'} - u^*\|^2]$, scaled by an exponentially decaying factor, and a constant term of order $\mathcal{O}(\mu, \mu^{k'}, \delta)$ that depends on k' . In steady state, $\mathbb{E}[\|u_{k'} - u^*\|^2]$ vanishes, so that the bound reduces to $\mathcal{O}(\mu, \mu^{k'}, \delta)$. Since $\mu^{k'}$ decreases as k' increases for $\mu < 1$, the steady-state error is eventually fully characterized by $\mathcal{O}(\mu, \delta)$.

V. NUMERICAL SIMULATION

To demonstrate that the controller (8) is capable of identifying u^* , we perform numerical simulations on an LTI system defined by

$$x_{k+1} = Ax_k + Bu_k, \quad (12)$$

where $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$ are the system matrices. This system has an invertible steady-state input-output map $H = (I - A)^{-1}B$ and is pre-stabilized by a lower-level controller, ensuring that the spectral radius of A is less than one. Such systems frequently arise in applications, including building control [27], and power systems [24].

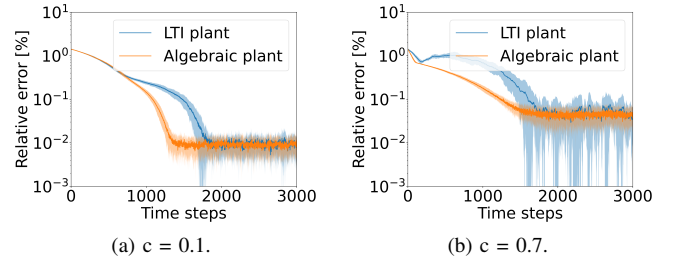


Fig. 1. Quadratic problem.

A. A simple example

To analyze the impact of plant transients on the algorithm's performance, we first consider the quadratic problem

$$\begin{aligned} \min \quad & (x - x_{\text{ref}})^\top (x - x_{\text{ref}}), \\ \text{s.t.} \quad & x = Hu, \end{aligned}$$

where the solution is $u = H^{-1}x_{\text{ref}}$. We set $A = \begin{bmatrix} c & 1 \\ 0 & c \end{bmatrix}$, B to be the identity matrix and $x_{\text{ref}} = [100, 100]^\top$. The parameter c varies between $c = 0.1$ and $c = 0.7$ to represent plants with different decay rates. The simulation parameters are chosen as $\eta = 0.1$ and $\delta = 0.5$. The results are shown in Fig. 1.

In Fig. 1, the relative error $\|x_k - x_{\text{ref}}\|/\|x_{\text{ref}}\|$ on a logarithmic scale is shown for $c = 0.1$ and $c = 0.7$. The solid line represents the mean value over 20 simulations, while the shaded region indicates one standard deviation. The orange line corresponds to the algebraic plant, which assumes H is known and samples $\mathbf{1}_{u_k + \delta v_k \succ u_{k-1} + \delta v_{k-1}}$ directly, whereas the blue line represents (8). At the steady state, both (8) and its steady-state counterpart achieve a comparable level of accuracy. However, for the slower system ($c = 0.7$), (8) exhibits a larger overshoot and higher steady-state variance.

B. Thermal comfort optimization

Next, we consider a thermal comfort optimization problem, in which a building is represented by an LTI system with 13 states [27]. An occupant's thermal comfort utility is represented with the well-known Predictive Mean Vote (PMV) model [3]. The PMV output is typically expressed in terms of the Predicted Percentage of Dissatisfied (PPD) index, which is a nonconvex function of room temperature. Preference feedback is generated based on this model, with the goal of identifying the room temperature that minimizes PPD using the controller (8). For other PMV parameters, we assume the occupant is typing, wearing sweatpants, T-shirt and shoes or sandals (default setting in the PMV model).

In Fig. 2, the mean indoor temperature with (8) over 20 simulations (blue line) is plotted, while the orange line represents the case with a noise-free user model, i.e., $\mathbf{1}_{(x_{k+1}, u_k + \delta v_k) \succ (x_k, u_{k-1} + \delta v_{k-1})} = \text{sign}(\Phi(x_k, u_{k-1} + \delta v_{k-1}) - \Phi(x_{k+1}, u_k + \delta v_k))$. The shaded region represents one standard deviation, and the black horizontal line represents the true optimal temperature. With careful tuning of η and δ , controller (8) can track the optimal point effectively without large overshoot, which demonstrates its potential for

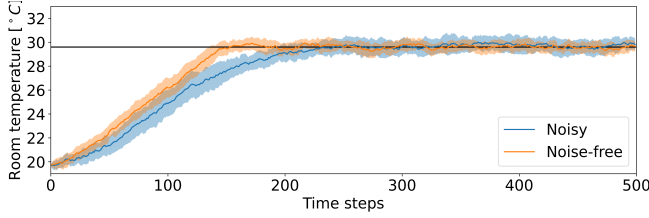


Fig. 2. Thermal comfort optimization.

learning a human’s utility in real-world applications. Meanwhile, the algorithm exhibits a higher convergence rate when there is no noise in the preference feedback. Quantifying its closed-loop behavior from a theoretical perspective remains an interesting direction for future work.

VI. CONCLUSION

In this work, we developed a human-aware controller that utilizes preference feedback to optimize utility while accounting for system transients. We derived an explicit upper bound on the error introduced by approximating the steady-state input-state map using the real-time state measurements. We analyzed the impact of both the system decay rate and smoothing parameters on the stability and optimality of the closed-loop system. Numerical experiments on a thermal comfort optimization task demonstrate its potential for solving real-world problems. Further research directions include extending the theoretical framework to alternative user models (e.g., noise-free model) and exploring real-world applications such as product design and chemical selection.

REFERENCES

- [1] A. Eichler, G. Darivianakis, and J. Lygeros, “Humans-in-the-loop: A game-theoretic perspective on adaptive building energy systems,” in *European Control Conference*, 2018, pp. 1322–1327.
- [2] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kotsuge, and O. Khatib, “Progress and prospects of the human–robot collaboration,” *Autonomous Robots*, vol. 42, pp. 957–975, 2018.
- [3] P. O. Fanger, *Thermal comfort: Analysis and applications in environmental engineering*. Danish Technical Press, 1970.
- [4] A. Hauswirth, Z. He, S. Bolognani, G. Hug, and F. Dörfler, “Optimization algorithms as robust feedback controllers,” *Annual Reviews in Control*, vol. 57, p. 100941, 2024.
- [5] L. Ortmann, A. Hauswirth, I. Caduff, F. Dörfler, and S. Bolognani, “Experimental validation of feedback optimization in power distribution grids,” *Electric Power Systems Research*, vol. 189, p. 106782, 2020.
- [6] A. Terpin, S. Fricker, M. Perez, M. H. de Badyn, and F. Dörfler, “Distributed feedback optimization for robotic coordination,” in *American Control Conference*, 2022, pp. 3710–3715.
- [7] M. Colombino, E. Dall’Anese, and A. Bernstein, “Online optimization as a feedback controller: Stability and tracking,” *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 422–432, 2020.
- [8] Z. He, S. Bolognani, J. He, F. Dörfler, and X. Guan, “Model-free nonlinear feedback optimization,” *IEEE Transactions on Automatic Control*, vol. 69, no. 7, pp. 4554–4569, 2023.
- [9] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.
- [10] M. Zoghi, Z. S. Karnin, S. Whiteson, and M. De Rijke, “Copeland dueling bandits,” in *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [11] T. Urvoy, F. Clerot, R. Féraud, and S. Naamane, “Generic exploration and k-armed voting bandits,” in *International Conference on Machine Learning*. PMLR, 2013, pp. 91–99.

- [12] M. Zoghi, S. Whiteson, R. Munos, and M. Rijke, “Relative upper confidence bound for the k-armed dueling bandit problem,” in *International Conference on Machine Learning*. PMLR, 2014, pp. 10–18.
- [13] B. Chen and P. I. Frazier, “Dueling bandits with weak regret,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 731–739.
- [14] J. González, Z. Dai, A. Damianou, and N. D. Lawrence, “Preferential Bayesian optimization,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1282–1291.
- [15] S. Takeno, M. Nomura, and M. Karasuyama, “Towards practical preferential Bayesian optimization with skew Gaussian processes,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 33 516–33 533.
- [16] W. Xu, W. Wang, Y. Jiang, B. Svetozarevic, and C. N. Jones, “Principled preferential Bayesian optimization,” *arXiv preprint arXiv:2402.05367*, 2024.
- [17] B. Pásztor, P. Kassraie, and A. Krause, “Bandits with preference feedback: A stackelberg game perspective,” *arXiv preprint arXiv:2406.16745*, 2024.
- [18] Y. Yue and T. Joachims, “Interactively optimizing information retrieval systems as a dueling bandits problem,” in *International Conference on Machine Learning*, 2009, pp. 1201–1208.
- [19] A. Saha, T. Koren, and Y. Mansour, “Dueling convex optimization,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 9245–9254.
- [20] H. K. Khalil and J. W. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.
- [21] J. W. Simpson-Porco, “Analysis and synthesis of low-gain integral controllers for nonlinear systems,” *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4148–4159, 2020.
- [22] W. Wang, Z. He, G. Belgioioso, S. Bolognani, and F. Dörfler, “Online feedback optimization over networks: A distributed model-free approach,” *arXiv preprint arXiv:2403.19834*, 2024.
- [23] W. Wang, Z. He, G. Belgioioso, S. Bolognani, and F. Dörfler, “Decentralized feedback optimization via sensitivity decoupling: Stability and sub-optimality,” in *European Control Conference*, 2024, pp. 3201–3206.
- [24] J. Zhao and F. Dörfler, “Distributed control and optimization in DC microgrids,” *Automatica*, vol. 61, pp. 18–26, 2015.
- [25] P. Mikkola, “Humans as information sources in Bayesian optimization,” Ph.D. dissertation, Aalto University, 2024.
- [26] Y. Tang, Z. Ren, and N. Li, “Zeroth-order feedback optimization for cooperative multi-agent systems,” *Automatica*, vol. 148, p. 110741, 2023.
- [27] Y. Lian, J. Shi, M. Koch, and C. N. Jones, “Adaptive robust data-driven building control via bilevel reformulation: An experimental result,” *IEEE Transactions on Control Systems Technology*, vol. 31, no. 6, pp. 2420–2436, 2023.

APPENDIX

A. Proof of Lemma 1

Proof. We first consider $\mathbb{E}[V(x_k, u_k + \delta v_k) | \mathcal{F}_k]$, where $\mathcal{F}_k$ is the natural filtration at time step k .

$$\begin{aligned}
 & \mathbb{E}[V(x_k, u_k + \delta v_k) | \mathcal{F}_k] \\
 & \leq \mathbb{E}[\alpha_2 \|x_k - h(u_{k-1} + \delta v_{k-1}) \\
 & \quad + h(u_{k-1} + \delta v_{k-1}) - h(u_k + \delta v_k)\|^2 | \mathcal{F}_k] \\
 & \leq \mathbb{E}[2\alpha_2 \|x_k - h(u_{k-1} + \delta v_{k-1})\|^2 | \mathcal{F}_k] \\
 & \quad + \mathbb{E}[2\alpha_2 \|h(u_{k-1} + \delta v_{k-1}) - h(u_k + \delta v_k)\|^2 | \mathcal{F}_k] \\
 & \leq 2 \frac{\alpha_2}{\alpha_1} V(x_k, u_{k-1} + \delta v_{k-1}) \\
 & \quad + \mathbb{E}[2\alpha_2 L_h^2 \|u_{k-1} + \delta v_{k-1} - u_k - \delta v_k\|^2 | \mathcal{F}_k] \\
 & \leq 2 \frac{\alpha_2}{\alpha_1} (1 - \frac{\alpha_3}{\alpha_2}) V(x_{k-1}, u_{k-1} + \delta v_{k-1}) + 4\alpha_2 L_h^2 (\delta^2 + (\delta + \frac{\eta}{2\delta})^2) \\
 & = \mu V(x_{k-1}, u_{k-1} + \delta v_{k-1}) + a_1 (2\delta^2 + \eta + (\frac{\eta}{2\delta})^2).
 \end{aligned}$$

Apply this inequality k times from $V(x_k, u_k + \delta v_k)$ to $V(x_0, u_0 + \delta v_0)$, we obtain the desired result. $\square$

B. Proof of Lemma 2

Proof. For the Lipschitz property, we have $\forall u'$

$$|p_{u'}(u_1) - p_{u'}(u_2)| = |\sigma(\tilde{\Phi}(u_1) - \tilde{\Phi}(u')) - \sigma(\tilde{\Phi}(u_2) - \tilde{\Phi}(u'))| \\ \leq L_{\sigma,0} L_0 \|u_1 - u_2\|.$$

For the smoothness property,

$$\|\nabla p_{u'}(u_1) - \nabla p_{u'}(u_2)\| \\ = \|\sigma'(\tilde{\Phi}(u_1) - \tilde{\Phi}(u')) \nabla \tilde{\Phi}(u_1) - \sigma'(\tilde{\Phi}(u_1) - \tilde{\Phi}(u')) \nabla \tilde{\Phi}(u_2) \\ + \sigma'(\tilde{\Phi}(u_1) - \tilde{\Phi}(u')) \nabla \tilde{\Phi}(u_2) - \sigma'(\tilde{\Phi}(u_2) - \tilde{\Phi}(u')) \nabla \tilde{\Phi}(u_2)\| \\ \leq \|\sigma'(\tilde{\Phi}(u_1) - \tilde{\Phi}(u'))\| L_1 \|u_1 - u_2\| + L_{\sigma,1} L_0 \|\nabla \tilde{\Phi}(u_2)\| \|u_1 - u_2\| \\ \leq (\sigma'(0) L_1 + L_{\sigma,1} L_0^2) \|u_1 - u_2\|. \quad \square$$

C. Proof of Lemma 3

Proof. Since $\sigma(t)$ is a strictly increasing function, u^* that minimizes $\tilde{\Phi}(u)$ also minimizes $p_{u'}(u)$. Suppose, for contradiction, there exists two distinct minimizers of $p_{u'}(u)$, u_1 and u_2 such that $u_1 \neq u_2$. Because $\sigma(t)$ is strictly increasing, it follows that $\tilde{\Phi}(u_1) = \tilde{\Phi}(u_2)$, implying that $\tilde{\Phi}(u)$ attains the minimum value at two different points. This contradicts the assumption that $\tilde{\Phi}(u)$ is strongly convex, which guarantees a unique minimizer. This concludes the proof. $\square$

D. Proof of Lemma 4

Proof. Let us consider $\|\mathbb{E}[e_k | \mathcal{F}_k]\|^2$.

$$\|\mathbb{E}[e_k | \mathcal{F}_k]\|^2 \\ = \left\| \frac{1}{\delta} \mathbb{E}[\sigma(\Phi(x_{k+1}, u_k + \delta v_k) - \Phi(x_k, u_{k-1} + \delta v_{k-1})) v_k \right. \\ \left. - \sigma(\Phi(h(u_k + \delta v_k), u_k + \delta v_k) - \Phi(h(u_{k-1} + \delta v_{k-1}), u_{k-1} + \delta v_{k-1})) v_k \right. \\ \left. + \sigma(\Phi(h(u_k + \delta v_k), u_k + \delta v_k) - \Phi(h(u_{k-1} + \delta v_{k-1}), u_{k-1} + \delta v_{k-1})) v_k | \mathcal{F}_k] - \nabla p_{u_k}(u_k) \right\|^2 \\ \leq \frac{2}{\delta^2} \mathbb{E}[\|\sigma(\Phi(x_{k+1}, u_k + \delta v_k) - \Phi(x_k, u_{k-1} + \delta v_{k-1})) v_k \\ - \sigma(\Phi(h(u_k + \delta v_k), u_k + \delta v_k) - \Phi(h(u_{k-1} + \delta v_{k-1}), u_{k-1} + \delta v_{k-1})) v_k\|^2 | \mathcal{F}_k] \\ + 2 \|\nabla \tilde{p}_{u_{k-1} + \delta v_{k-1}}(u_k) - \nabla p_{u_{k-1} + \delta v_{k-1}}(u_k) \\ + \nabla p_{u_{k-1} + \delta v_{k-1}}(u_k) - \nabla p_{u_k}(u_k)\|^2 \\ \stackrel{(s.1)}{\leq} \frac{4L_{\sigma,0}^2 L_x^2}{\delta^2} \|x_k - h(u_{k-1} + \delta v_{k-1})\|^2 \\ + \frac{4L_{\sigma,0}^2 L_x^2}{\delta^2} \mathbb{E}[\|x_{k+1} - h(u_k + \delta v_k)\|^2 | \mathcal{F}_k] + 2a_2^2 \delta^2 \\ \leq \frac{2L_{\sigma,0}^2 L_x^2 (\mu + 1)}{\alpha_2 \delta^2} \mu V(x_{k-1}, u_{k-1} + \delta v_{k-1}) \\ + \frac{2L_{\sigma,0}^2 L_x^2 a_1}{\alpha_2 \delta^2} (2\delta^2 + 2\eta + (\frac{\eta}{\delta})^2) \mu + 2a_2^2 \delta^2,$$

where in (s.1) we apply the bound from [26, Lemma 1], and perform analysis under the assumptions of smoothness and Lipschitz continuity. $\square$

E. Proof of Theorem 1

We first present a useful lemma.

Lemma 5. For non-negative sequence a_k satisfying $\forall k, a_{k+1}^2 \leq \rho a_k^2 + b_k a_k + c$, where $\rho < 1$, and b_k is a non-increasing positive sequence, we have $\forall k' \geq 0$ and $\forall k > k', a_k^2 \leq \rho'^{k-k'} a_{k'}^2 + (a_{k'}^*)^2$, where $\rho' = (1 - \frac{\sqrt{b_{k'}^2 + 4(1-\rho)c}}{2a_{k'}^*})$, $a_{k'}^* = \frac{b_{k'} + \sqrt{b_{k'}^2 + 4(1-\rho)c}}{2(1-\rho)}$.

Proof. $\forall k' \geq 0$, if $a_i \geq a_{k'}^*, \forall i \in \{k', \dots, k-1\}$,

$$a_{i+1}^2 \leq \rho a_i^2 + b_i a_i + c \leq (\rho - 1) a_i^2 + b_{k'} a_i + c + a_i^2 \\ \leq (1 - \frac{\sqrt{b_{k'}^2 + 4(1-\rho)c}}{2a_{k'}^*}) a_i^2 + \frac{\sqrt{b_{k'}^2 + 4(1-\rho)c}}{2} a_{k'}^*.$$

Applying this inequality recursively, we have

$$a_k^2 \leq \rho'^{k-k'} a_{k'}^2 + \frac{1}{1-\rho'} \frac{\sqrt{b_{k'}^2 + 4(1-\rho)c}}{2} a_{k'}^* \\ = \rho'^{k-k'} a_{k'}^2 + (a_{k'}^*)^2.$$

If $\exists i \in \{k', \dots, k-1\}$ such that $a_i \leq a_{k'}^*$,

$$a_{i+1}^2 \leq \rho a_i^2 + b_i a_i + c \leq \rho a_i^2 + b_{k'} a_i + c \leq (a_{k'}^*)^2.$$

Since $\rho a_i^2 + b_{k'} a_i + c$ is an increasing function when $a_i \geq 0$, we have $a_k^2 \leq (a_{k'}^*)^2 \leq \rho'^{k-k'} a_{k'}^2 + (a_{k'}^*)^2$. $\square$

Now, we are ready to present the main proof.

Proof.

$$\mathbb{E}[\|u_{k+1} - u^*\|^2 | \mathcal{F}_k] \\ = \mathbb{E}[\|u_k - u^*\|^2 - 2\eta(\nabla p_{u_k}(u_k) + e_k)^\top (u_k - u^*) + \eta^2 | \mathcal{F}_k] \\ \leq (1 - 2\sigma'(0)m\eta) \|u_k - u^*\|^2 + 2\eta \|\mathbb{E}[e_k | \mathcal{F}_k]\| \|u_k - u^*\| + \eta^2 \\ \leq (1 - 2\sigma'(0)m\eta) \|u_k - u^*\|^2 \\ + 2\eta \sqrt{R_1 V(x_{k-1}, u_{k-1} + \delta v_{k-1}) + R_2} \|u_k - u^*\| + \eta^2. \\ \mathbb{E}[\|u_{k+1} - u^*\|^2] = \mathbb{E}[\mathbb{E}[\|u_{k+1} - u^*\|^2 | \mathcal{F}_k]] \\ \leq (1 - 2\sigma'(0)m\eta) \mathbb{E}[\|u_k - u^*\|^2] \\ + 2\eta \mathbb{E}[\sqrt{R_1 V(x_{k-1}, u_{k-1} + \delta v_{k-1}) + R_2} \|u_k - u^*\|] + \eta^2 \\ \leq (1 - 2\sigma'(0)m\eta) \mathbb{E}[\|u_k - u^*\|^2] \\ + 2\eta \sqrt{R_1 \mathbb{E}[V(x_{k-1}, u_{k-1} + \delta v_{k-1})] + R_2} \sqrt{\mathbb{E}[\|u_k - u^*\|^2]} + \eta^2 \\ \leq (1 - 2\sigma'(0)m\eta) \mathbb{E}[\|u_k - u^*\|^2] \\ + 2\eta \sqrt{R_1 \mathbb{E}[V(x_0, u_0 + \delta v_0)] \mu^{k-1} + R_1 \frac{a_1}{1-\mu} (2\delta^2 + \eta + (\frac{\eta}{2\delta})^2) + R_2} \\ \sqrt{\mathbb{E}[\|u_k - u^*\|^2]} + \eta^2.$$

Writing $\rho = 1 - 2\sigma'(0)m\eta$, $b_k = 2\eta \sqrt{R_1 \mu^{k-1} + b_2}$, $c = \eta^2$ and applying Lemma 5, we have

$$\mathbb{E}[\|u_k - u^*\|^2] \\ \leq \rho'^{k-k'} \mathbb{E}[\|u_{k'} - u^*\|^2] + (a_{k'}^*)^2 \\ \leq (\frac{1+\rho}{2})^{k-k'} \mathbb{E}[\|u_{k'} - u^*\|^2] + \frac{b_1 \mu^{k'-1} + b_2 + 2\sigma'(0)m\eta}{\sigma'(0)^2 m^2},$$

where $\rho' = \frac{b_{k'} + \rho \sqrt{b_{k'}^2 + 4(1-\rho)c}}{b_{k'} + \sqrt{b_{k'}^2 + 4(1-\rho)c}}$. $\square$