MixUCB: Enhancing Safe Exploration in Contextual Bandits with Human Oversight

Anonymous authors
Paper under double-blind review

Keywords: Safe Exploration, human-in-the-loop contextual bandit

Summary

The integration of AI into high-stakes decision-making domains demands safety and accountability. Traditional contextual bandit algorithms for online and adaptive decision-making must balance exploration and exploitation, posing significant risks when applied to critical environments where exploratory actions can lead to severe consequences. To address these challenges, we propose MixUCB, a flexible human-in-the-loop contextual bandit framework that enhances safe exploration by incorporating human expertise and oversight with machine automation. Based on the model's confidence and the associated risks, MixUCB intelligently determines when to seek human intervention. The reliance on human input gradually reduces as the system learns and gains confidence. Theoretically, we analyzed the regret and query complexity in order to rigorously answer the question of when to query. Empirically, we validate the effectiveness through extensive experiments on both synthetic and real-world datasets. Our findings underscore the importance of designing decision-making frameworks that are not only theoretically and technically sound, but also align with societal expectations of accountability and safety.

Contribution(s)

1. We introduce **MixUCB**, a novel human-in-the-loop contextual bandit framework that dynamically determines when to seek human intervention based on uncertainty, enhancing safe exploration in high-stakes decision-making tasks. MixUCB is flexible in accepting various types of expert feedback.
   Context: Our approach unifies learning from experts (as in active learning, imitation learning, etc.) with learning from experience (as in reinforcement learning).
2. We provide a theoretical analysis of our framework, offering guarantees on regret and query complexity. This addresses the fundamental question of when to rely on expert input while balancing the cost and quality of the feedback.
   Context: While traditional online learning or bandit algorithms focus on fixed feedback settings, our analysis demonstrates MixUCB's adaptability to varying levels of expert involvement.
3. We demonstrate the practical effectiveness of MixUCB through experiments on both synthetic and real-world datasets, showcasing the superiority of combining human expertise and AI in comparison to fully automated decision-making. We highlight the importance of designing AI systems that are not only technically sound but also emphasize safety, accountability, and human-centric decision-making, setting a new standard for safe exploration in contextual bandit problems.
   Context: Our experiments cover a range of feedback settings, showcasing MixUCB's ability to maintain high performance even when expert feedback is limited or noisy.

MixUCB: Enhancing Safe Exploration in Contextual Bandits with Human Oversight

Anonymous authors
Paper under double-blind review

Abstract

The integration of AI into high-stakes decision-making domains demands safety and accountability. Traditional contextual bandit algorithms for online and adaptive decision-making must balance exploration and exploitation, posing significant risks when applied to critical environments where exploratory actions can lead to severe consequences. To address these challenges, we propose MixUCB, a flexible human-in-the-loop contextual bandit framework that enhances safe exploration by incorporating human expertise and oversight with machine automation. Based on the model's confidence and the associated risks, MixUCB intelligently determines when to seek human intervention. The reliance on human input gradually reduces as the system learns and gains confidence. Theoretically, we analyze the regret and query complexity in order to rigorously answer the question of when to query. Empirically, we validate the effectiveness through extensive experiments on both synthetic and real-world datasets. Our findings underscore the importance of designing decision-making frameworks that are not only theoretically and technically sound, but also align with societal expectations of accountability and safety.

1 Introduction

Distinct from typical machine learning applications that focus on tasks with limited risks, the deployment of AI algorithms in high-stakes decision-making domains—such as self-driving cars (Sikar et al., 2024), medical diagnostics (Esteva et al., 2017), and criminal justice (Dressel & Farid, 2018)—can have profound impacts and carry much greater responsibility (Amodei et al., 2016). The potential consequences of actions taken in these domains are far-reaching, spanning from life-and-death situations for individuals, to the broader societal, ethical, and legal challenges that affect humanity as a whole. Therefore, it is crucial that AI decision-making processes are built upon safety, accountability, responsibility, trustworthiness, and transparency, instead of excessively pursuing maximum efficiency.

However, despite the necessity of safe, reliable, and responsible AI systems, implementing them in high-stakes environments presents significant challenges. Traditional learning and decision-making algorithms, such as the contextual bandits (Wang et al., 2005), rely on balancing exploration and exploitation. While this exploration is acceptable and often beneficial in lower-risk domains like recommendation systems (Li et al., 2010), in high-stakes settings, exploratory actions can lead to unacceptable risks and severe consequences. For example, a self-driving car experimenting with unfamiliar maneuvers could result in accidents, endangering human lives.

To address these challenges, we propose a human-in-the-loop contextual bandit framework (Figure 1) that can balance the benefits of automation with the need for human expertise and oversight in critical situations. In particular, our approach allows for human intervention when the AI model lacks confidence or when decisions carry significant risk, preventing potential catastrophic errors and ensuring *safe exploration*. One of the key strengths of our framework is its ability to incorporate

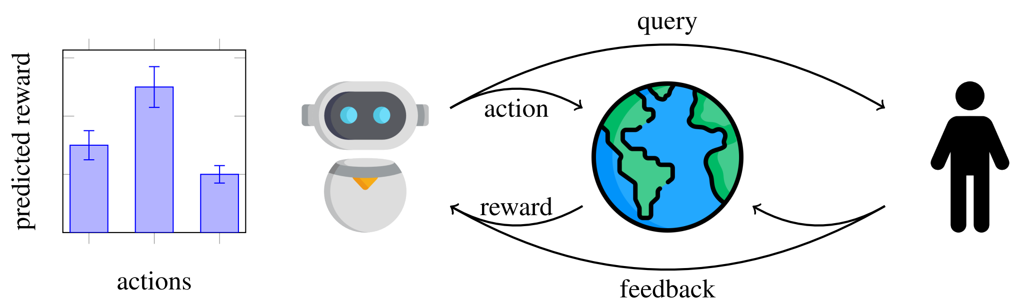


Figure 1: Illustration of our setting, which augments the traditional feedback loop between algorithm (left) and environment (middle) to include the presence of a human expert (right).

both observed consequences and expert advice. As the learner interacts more with the environment and gathers data—both from autonomous actions and expert interventions—it becomes more confident so the reliance on expert intervention reduces over time. Beyond the immediate benefits of safety, our framework offers several additional advantages. Firstly, the high-quality data collected during expert interventions/feedback can significantly accelerate the model's learning process. Secondly, actively involving humans in the decision-making process allows for a clearer assignment of responsibility, clarifying liability in cases of failure or harm.

In summary, our main contributions are as follows: (1) We develop a flexible human-in-the-loop contextual bandit algorithm MixUCB that dynamically determines when to seek human intervention. MixUCB accepts various types of expert advice. (2) We provide theoretical analyses on the regret and query complexity, answering the question of when to rely on expert advice. (3) We validate our approach through experiments on both synthetic and real-world datasets, showcasing the practical applicability and benefits of MixUCB. (4) A key finding is that combining AI and human expertise outperforms alternatives, underscoring the importance of complementing AI and human to achieve more robust and effective decision-making.

2 Related Work

Contextual bandits The standard setting in contextual bandit does not assume the existence of human experts and the learner can only learn from the feedback (i.e., reward signals) by interacting with the environment by herself (Langford & Zhang, 2007; Beygelzimer et al., 2011; Dani et al., 2008; Abbasi-Yadkori et al., 2011; Li et al., 2010). While these algorithms achieve near-optimal regret bounds in the long term, they can play potentially unsafe actions during their exploration phases. Thus, these algorithms cannot be directly applied to safety-critical applications.

Selective sampling and active learning Active learning or selective sampling is a learning paradigm that is designed to reduce query complexity by only querying for labels at selected data points (Cesa-Bianchi et al., 2005; Dekel et al., 2012; Agarwal, 2013; Hanneke & Yang, 2015; 2021; Zhu & Nowak, 2022; Sekhari et al., 2024b;a). These prior work do not assume the learner can receive reward feedback at the rounds where they do not query experts.

Interactive learning from humans Querying human experts for inputs has been studied in the context of imitation learning (Ross et al., 2011; Ross & Bagnell, 2014; Sun et al., 2017; Pan et al., 2017). While these prior works focus on the more general Markov Decision Processes, they do not study how to reduce the number of expert queries using active learning techniques. While we focus on the contextual bandit setting (i.e., RL with horizon being one), our technique can be potentially extended to the full MDP setting by treating each step in the MDP as a contextual bandit problem (Sekhari et al., 2024b).

Learning to defer Madras et al. (2018) proposed learning to defer, demonstrating its effects in improving system accuracy and fairness. Follow up works such as those by Raghu et al. (2019); Keswani et al. (2021); Narasimhan et al. (2022); Mozannar & Sontag (2020); Joshi et al. (2021); Sikar et al. (2024) studied when to defer to human judgment and when to accept automated predictions in standard ML and supervised learning settings, rather than an active learning setting.

3 Problem Formulation

3.1 Contextual Bandit

We consider the following contextual bandit setting with arbitrary (potentially adversarial) contexts and stochastic rewards. At each round $t \in [T]$, the learner observes the contextual information $x_t \in \mathcal{X}$ for the context space $\mathcal{X}$, which it may use to inform its choice of action. For example, in recommendation system, context x_t could be features of a user logging onto the system. The learner chooses an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the learner's action space. We assume that $\mathcal{A}$ is a finite set with cardinality K. Then, only the reward $r_t \sim R(x_t, a_t)$ of the chosen action a_t is observed, where $R : \mathcal{X} \times \mathcal{A} \to \Delta([0, 1])$ is the reward function.

Assume that the learner has access to a class of functions $\mathcal{F} \subset (\mathcal{X} \times \mathcal{A} \to [0, 1])$ that model the mean of the reward function, such as linear functions or neural networks. Assume there exists $f^* \in \mathcal{F}$ such that $f^*(x, a) = \mathbb{E}_{r \sim R(x,a)}[r]$, i.e., the class $\mathcal{F}$ is rich enough to contain a function that can perfectly predict the expected reward of any action under any context. This realizability assumption is rather standard and has been used in many previous works (Chu et al., 2011; Foster & Rakhlin, 2020; Foster et al., 2018a; Agarwal et al., 2012).

The learner's goal is to compete against the optimal policy $\pi^* : \mathcal{X} \to \mathcal{A}$ that picks the action with the highest expected reward, i.e., $a^* = \arg\max_{a \in \mathcal{A}} f^*(x, a)$. Formally, the learner's goal is to minimize the expected regret

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} f^*(x_t, a_t^*) - f^*(x_t, a_t)\,. \tag{1}$$

3.2 Expert Feedback

We augment the decision-making setting by considering the presence of human experts who can be queried for guidance. In addition to selecting an action a_t, the learner can opt to query a human expert ($Z_t = 1$) or take an action autonomously ($Z_t = 0$). Different human experts may offer different types of feedback, either directly suggesting an action or predicting the rewards associated with each action. In particular, we explore three types of expert feedback. These types of feedback vary in the level of information provided to the learner and the cognitive or computational burden placed on the expert.

I: Action Only The expert selects and takes an action $\tilde{a}^*$. The learner observes the action but does not observe the resulting reward.
II: Action + Associated Reward The expert selects and takes an action $\tilde{a}^*$. The learner observes both the action and the resulting reward r_t.
III: Rewards for All Actions The expert provides predicted rewards $\tilde{r}_{t,a}$ for all actions $a \in \mathcal{A}$.

These three types of feedback capture the fact that experts vary in their level of expertise and access to information, which influences the quality and depth of the feedback they can offer. *Type-I* feedback is applicable in situations where reward feedback is not available once the human expert takes over. For example, in a medical setting, once a doctor takes over selecting a treatment, the learner may never observe the patient's outcome. *Type-II* feedback is slightly more informative since the learner is able to observe the outcome of the expert's action. For example, a robot may be guided by an expert operator who suggests manipulation actions. The robot can then observe whether this action successfully picks up an object. *Type-III* feedback is applicable in situations where an expert has full information and can analyze all potential outcomes. By providing information about not only the action taken but also the alternatives, the expert provides the learner with a comprehensive view of the reward landscape. This type of feedback is highly informative, but it comes at a significant cost.

Beyond the type of feedback, experts vary in the quality of feedback. Humans often exhibit bounded rationality in decision-making, so the expert action $\tilde{a}^*$ is not necessarily equal to the optimal action. We model the *Type-I* and *Type-II* expert choices using the a **reward-rational choice model**, in particular the Boltzmann-rational model (Luce, 1959; 1977; Ziebart et al., 2010) with rationality parameter $\alpha \geq 0$:

$$P(\tilde{a}_t^* = a | x_t) \propto \exp(\alpha f^*(x_t, a)). \tag{2}$$

When $\alpha \to \infty$, the expert behaves perfectly rationally, always selecting the optimal action; when $\alpha = 0$, the expert chooses actions at random, independent of the rewards. This model allows us to capture the natural variability in human decision-making and reflect different levels of competence across experts.

For *Type-III* feedback, we assume that the expert predicted rewards are bounded and unbiased, i.e. that they satisfy $\mathbb{E}[\tilde{r}_{t,a}|x_t] = f^*(x_t, a)$.

3.3 Online Regression Oracles

For a contextual bandit learner to be successful, it is necessary to learn efficiently from interactions with the environment and the human expert. This is formalized by the following definition.

Definition 1 (Online Regression Oracle)**.** *An online regression oracle for a convex loss ℓ w.r.t. the class $\mathcal{F}$, provides, for any sequence $\{(z_1, y_1), \cdots, (z_T, y_T)\}$, predictors $f_t \in \mathcal{F}$ such that the prediction regret is bounded:*

$$\sum_{t=1}^{T} \ell(f_t(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq \mathrm{Reg}^{\ell}(\mathcal{F}; T)$$

Different regression oracles are appropriate for different types of feedback available to the learner. The **square loss online regression oracle** is appropriate for learning from observed rewards. In this setting, ℓ is the standard square loss, and the sequence contains context, action, reward tuples $\{((x_1, a_1), r_1), \cdots, ((x_t, a_t), r_t), \cdots ((x_T, a_T), r_T)\}$. If the learner has *Type-III* expert feedback, the predicted rewards for all actions can be incorporated into this sequence as well. The square loss oracle regret bound $\mathrm{Reg}^{sq}(\mathcal{F}, T)$ typically grows sublinearly with T and can be implemented efficiently (Krishnamurthy et al., 2019; Foster et al., 2018a; Rakhlin & Sridharan, 2014). For example, for finite function classes $\mathcal{F}$, the regret bound is $\mathrm{Reg}^{sq}(\mathcal{F}; T) = O(\log(T) \log(|\mathcal{F}|))$, while $\mathrm{Reg}^{sq}(\mathcal{F}; T) = O(d \log(T))$ when $\mathcal{F}$ is a d-dimensional linear class as in (5).

The **online logistic regression oracle** is appropriate for learning from actions selected by bounded-rational experts. In this setting, ℓ is the logistic loss, and the sequence contains context and action tuples $\{(x_1, a_1), \cdots, (x_T, a_T)\}$ observed through either *Type-I* or *Type-II* feedback. Similar to the square loss oracle, when $\mathcal{F}$ is finite, we have a regret bound $\mathrm{Reg}^{lr}(\mathcal{F}; T) = O(\log(T) \log(|\mathcal{F}|))$ (Cesa-Bianchi & Lugosi, 2006), while for $\mathcal{F}_{\text{lin}}$, there exists efficient improper learner with regret bound $\mathrm{Reg}^{lr}(\mathcal{F}; T) = O(d \log(T))$ (Foster et al., 2018b).

4 Human-in-the-loop Contextual Bandit Framework

We present a framework for seeking and incorporating expert advice in a contextual bandit setting. We call this framework MixUCB. In Algorithm 1, we present the typical scenario where experts recommend actions directly (*Type-I* or *Type-II*) according to a Boltzmann-rational model. This

Algorithm 1 MixUCB (*Type-I* and *II* feedback)

Input: Query threshold Δ, total rounds T, function class $\mathcal{F}$, initial confidence set $\mathcal{E}_1^{sq} = \mathcal{E}_1^{lr}$
for $t = 1, \cdots, T$ **do**
  Let $\mathcal{E}_t = \mathcal{E}_t^{sq} \cap \mathcal{E}_t^{lr}$
  $a_t^{ucb} = \arg\max_{a \in \mathcal{A}} \max_{f \in \mathcal{E}_t} f(x_t, a)$ and $w_t = \max_{f \in \mathcal{E}_t} f(x_t, a_t^{ucb}) - \min_{f \in \mathcal{E}_t} f(x_t, a_t^{ucb})$
  if $w_t \geq \Delta$ **then**
    Query ($Z_t = 1$) and play expert action $\tilde{a}_t^*$. Update $\mathcal{D}_t^{lr}, \mathcal{E}_t^{lr}$ with $(x_t, \tilde{a}_t^*)$ according to (4).
    if Type-II Feedback **then**
      Observe $r_t \sim r(x_t, \tilde{a}_t^*)$ and update $\mathcal{D}_t^{sq}$ and $\mathcal{E}_t^{sq}$ with $(x_t, \tilde{a}_t^*, r_t)$ according to (3).
  else
    Set $Z_t = 0$. Play a_t^{ucb} and observe $r_t \sim r(x_t, a_t^{ucb})$. Update $\mathcal{D}_t^{sq}$ and $\mathcal{E}_t^{sq}$ with (x_t, a_t^{ucb}, r_t) according to (3).

setting highlights the key challenges in leveraging diverse types of feedback. An extension to *Type-III* feedback is presented in the appendix and investigated in numerical experiments in Section 5.

Designing a human-in-the-loop contextual bandit framework presents two primary challenges: deciding when to query and effectively learning from feedback. To address the first challenge, our algorithm uses a measure of uncertainty. First, the learner follows the standard "optimism in the face of uncertainty" principle to compute the upper confidence bound (UCB) action, a_t^{ucb}. Then, the learner computes a pessimistic lower bound on the reward of this action. The uncertainty is defined as the difference between the optimistic upper bound and the pessimistic lower bound. If the learner's uncertainty in a_t^{ucb} falls above a predefined threshold Δ, i.e., the learner is not confident about this action, it queries the expert for the optimal action rather than taking the risk.

The second challenge is to integrate various types of feedback to enhance learning. Accurate confidence sets are crucial for optimism/pessimism during action selection and the querying decision. Ideally, the learner should become more confident over time through interaction with the environment or expert. In the standard bandit setting, only autonomous environment interactions are considered, while in active learning settings, only expert advice is considered. Our approach combines these two sources of information to construct confidence sets from both expert advice and observed rewards. In the next section, we discuss how to overcome a key challenge of *Type-I* and *Type-II* feedback, which is that experts don't provide information on rewards directly, but rather provide a (noisy) suggested action.

4.1 Constructing Confidence Sets

In Algorithm 1, we construct two confidence sets: one based on rewards observed after interaction with the environment, and another based on expert feedback.

Given a sequence of context-action-reward data observed up to time t, $\mathcal{D}_t^{sq} = \{(x_k, a_k, r_k)\}$, the estimated reward function f_t^{sq} is given by the square loss oracle. Then the confidence set is defined

$$\mathcal{E}_t^{sq} = \{f \in \mathcal{F} \mid \sum_{x,a \in \mathcal{D}_t^{sq}} (f_t^{sq}(x_t, a_t) - f(x_t, a_t))^2 \leq \beta_t^{sq}\} . \tag{3}$$

This expression is justified because for stochastic rewards following the realizability assumption, Foster & Rakhlin (2020) show that when $\beta_t^{sq} = \mathrm{Reg}^{sq}(\mathcal{F}; t)$ from the online regression oracle (Definition 1), $f^* \in \mathcal{E}_t^{sq}$ with high probability.

Similarly, given a sequence of expert context-action data observed up to time t, $\mathcal{D}_t^{lr} = \{(x_k, a_k)\}$, the estimated reward function f_t^{lr} is given by the logistic regression oracle. Then the confidence set is defined as

$$\mathcal{E}_t^{lr} = \{f \in \mathcal{F} \mid \sum_{x,a \in \mathcal{D}_t^{lr}} (f_t^{lr}(x_t, a_t) - f(x_t, a_t))^2 \leq \beta_t^{lr}\} . \tag{4}$$

This expression is justified because for bounded-rational experts and rewards following the realizability assumption, Sekhari et al. (2024b) show that when $\beta_t^{lr} = \text{Reg}^{lr}(\mathcal{F}; t)$ from the online regression oracle (Definition 1), $f^* \in \mathcal{E}_t^{lr}$ with high probability.

Therefore, with high probability, the true reward function lies in the intersection of these sets $f^* \in \mathcal{E}_t^{sq} \cap \mathcal{E}_t^{lr}$. Algorithm 1 makes use of both estimates and both confidence sets, to combine bandit feedback with expert advice.

Linear Contextual Bandits (Chu et al., 2011) We focus on the special case of linear contextual bandits, where the online regression oracles and confidence sets can be written concretely. Consider a featurization of context-action pairs $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, and a linear function class,

$$\mathcal{F}_{\text{lin}} = \{(x, a) \to \theta^\top \phi(x, a) \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1\}\,. \tag{5}$$

Linear contextual bandit operates under the linear realizability assumption, i.e that there exists weight vector $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\| \leq 1$ and $\mathbb{E}[r_t | x_t, a_t] = \phi(x_t, a_t)^\top \theta^*$ for all x_t and a_t.

In this case, the regression oracles are simply standard linear and logistic regression algorithms with regularization parameters λ^{sq} and λ^{lr}. The regression oracle regret scales as $O(d \log(T))$. The confidence sets over linear functions are equivalent to ellipsoidal confidence sets over parameters θ, taking the form

$$\|\theta - \theta_t\|_{V_t}^2 \leq \beta_t, \quad V_t = \sum_{x,a \in \mathcal{D}_t} \phi(x, a)\phi(x, a)^\top + \lambda I$$

Therefore, the optimistic/pessimistic computation in algorithm 1 involves solving a conic optimization problem over possible parameters θ: the objective function is linear and there are two ellipsoidal constraints. While this problem does not have a clean closed-form solution, it is computationally feasible to solve to high precision with modern optimizers.

4.2 Theoretical results

We provide a theoretical analysis of Algorithm 1 that characterizes its safety, performance, and querying behavior. For ease of exposition, the theoretical results focus on the linear contextual bandit setting. We present all proofs in the appendix.

Assumption 1. *The reward function is linear as in* (5) *with dimension* d*, and the feature function satisfies* $\|\phi(x_t, a)\|_2 \leq L, \forall t \in [T], a \in \mathcal{A}$.

The above assumption is standard in linear bandits (Abbasi-Yadkori et al., 2011). Next, we assume that the confidence sets $\mathcal{E}_t^{sq}$ and $\mathcal{E}_t^{lr}$ are valid, i.e. that they contain the true reward function. In the appendix, we use results from Foster & Rakhlin (2020); Sekhari et al. (2024b) to define β_t^{sq} and β_t^{lr} such that this assumption holds with high probability.

Assumption 2. *The confidence sets satisfy*

1. $1 \leq \beta_1^{sq} \leq \beta_2^{sq} \leq \cdots \leq \beta_T^{sq}$ *and* $1 \leq \beta_1^{lr} \leq \beta_2^{lr} \leq \cdots \leq \beta_T^{lr}$.
2. $\forall t \in [T]$, $f^* \in \mathcal{E}_t^{sq} \cap \mathcal{E}_t^{lr}$.

Under these assumptions, we characterize the performance of MixUCB. First, we show that the query condition prevents the learner from autonomously taking highly sub-optimal actions. As such, MixUCB guarantees that autonomous actions are always safe.

Lemma 1 (Autonomous Sub-optimality)**.** *Under Assumptions 1 and 2, a learner following Algorithm 1 never autonomously takes an action* a_t^{ucb} *with sub-optimality greater than* Δ.

Next, we consider the fact that experts may take sub-optimal actions due to their bounded rationality. The following lemma bounds the cost of the expert's bounded rationality.

Lemma 2 (Expert Sub-optimality)**.** *Let* $R_\infty = \max_{x\in\mathcal{X},a\in\mathcal{A}} f^*(x,a)$*. Then under the Boltzmann-rational model, the expected sub-optimality of an* α *rational expert is bounded by*

$$c(\alpha) \leq \frac{R_\infty(K-1)}{\exp(\alpha R_\infty) + K - 1} \tag{6}$$

The cost of bounded rationality increases as the rationality α decreases. It also increases with the number of actions K. Combining these results, we characterize the regret of MixUCB (Algorithm 1) in terms of the total number of queries that it makes.

Proposition 1 (MixUCB Regret)**.** *Under Assumptions 1 and 2, the expected regret of Algorithm 1 satisfies*

$$\mathrm{Reg}(T) \leq \frac{2\Delta\beta_T^{sq}\sqrt{T-Q}}{\sqrt{\log_2(1+\Delta^2)}}\sqrt{d\log_2(1+\tfrac{(T-Q)L^2}{\lambda d})} + Qc(\alpha) \tag{7}$$

where $Q = \sum_{t=1}^T Z_t$ *is the total number of queries made by the algorithm.*

Next, we upper bound the query complexity Q.

Theorem 1 (Query complexity)**.** *Under Assumptions 1 and 2, the query complexity of Algorithm 1 is bounded:*

$$Q = \sum_{t=1}^T Z_t \leq \frac{10\max\{1, \beta_T^{sq}, \beta_T^{lr}\}d}{\Delta^2}\,. \tag{8}$$

Note that $\max\{\beta_T^{sq}, \beta_T^{lr}\} = O(d\log T)$, therefore, the query complexity has only a weak dependence on the horizon T. In other words, expert feedback will be sought for a small, almost constant, portion of the interaction horizon. The proof of this result crucially relies on the fact that MixUCB uses the logistic regression oracle to learn from expert feedback. In the absence of incorporating expert advice, it is possible that the learner would never shrink the confidence set and would thus query indefinitely. We therefore emphasize that observing the expert's action is crucial to this online bandit setting. Interestingly, observing the outcome of the expert's action is not so important—the above results hold for either *Type-I* or *Type-II* feedback.

Finally, we address the question of how to set the query threshold Δ. In some applications, this threshold may be determined purely by safety considerations (Lemma 1). In such settings, it is undesirable to allow a learner to try sub-optimal actions. In other applications, the overall performance may be the main criterion. Our final result is a summary theorem which provides guidance on setting Δ. We also characterize when MixUCB will outperform the purely autonomous LinUCB (Abbasi-Yadkori et al., 2011), which is equivalent to MixUCB with $\Delta \to \infty$.

Theorem 2. *Assume that* $\max\{1, \beta_T^{sq}, \beta_T^{lr}\} = O(d\log T)$ *and Assumptions 1 and 2 holds. Then by setting* $\Delta = \Theta(\sqrt[3]{\frac{d^2c(\alpha)}{T}})$*, the regret of MixUCB bounded by*

$$\mathit{Reg}(T) = O(\sqrt[3]{c(\alpha)d^2T^2}) \tag{9}$$

Moreover, if $c(\alpha) \leq O(\frac{d}{\sqrt{T}})$*, the regret is no worse than LinUCB.*

Proof. By Lemmas 1 and 2, the total regret on the rounds that we don't query is bounded by Δ, while the regret on the rounds that we query is bounded by $c(\alpha)$, thus, MixUCB-I regret is at most

$$c(\alpha)Q + \Delta(T-Q) = (c(\alpha)-\Delta)Q + \Delta T = O\left(\frac{d^2c(\alpha)}{\Delta^2} + \Delta T\right) = O(\sqrt[3]{c(\alpha)d^2T^2}) \tag{10}$$

where we take $\Delta = \Theta(\sqrt[3]{\frac{d^2c(\alpha)}{T}})$. To ensure that this is no worse than the LinUCB regret $O(d\sqrt{T})$ (Abbasi-Yadkori et al., 2011), we need $c(\alpha) \leq O(\frac{d}{\sqrt{T}})$. □

Categories	Algorithms	Action taken		Information/Feedback	
		$Z_t = 0$	$Z_t = 1$	$Z_t = 0$	$Z_t = 1$
Human-AI hybrid	MixUCB-I	a_t^{ucb}	$\tilde{a}_t^*$	$r(x_t, a_t^{ucb})$	$\tilde{a}_t^*$
	MixUCB-II		$\tilde{a}_t^*$		$r(x_t, \tilde{a}_t^*)$ and $\tilde{a}_t^*$
	MixUCB-III		a_t^*		$f^*(x_t, a), \forall a \in \mathcal{A}$
AI	LinUCB	a_t^{ucb}		$r(x_t, a_t^{ucb})$	
Linear Oracle	Classification	$\arg\max_a \hat{\theta}_{lr}^\top \phi(x_t, a)$		-	
	Regression	$\arg\max_a \hat{\theta}_{sq}^\top \phi(x_t, a)$		-	

Table 1: Summary of the algorithms and baselines.

This theorem shows that the querying threshold should increase for higher dimensions or expert costs (i.e. noisier experts), and decrease for longer interaction horizons. Furthermore, by comparing against the performance of LinUCB, this result justifies the intuition that MixUCB performs best when the cost is sufficiently small. In particular, the cost should be small compared with the dimension of the reward function, and inversely with the interaction horizon.

As a final remark, we note that the cost of bounded rationality $c(\alpha)$ could be replaced with $c(\alpha) + c$ where c is some additional cost of obtaining expert advice, e.g. due to monetary payment or degraded user experience.

5 Experiments

In this section, we conduct experiments in multiple settings to illustrate the effectiveness of our approach using both synthetic and real world datasets.

Baselines We compare MixUCB (I, II, III) with LinUCB, the standard purely autonomous algorithm which always takes a_t^{ucb} and corresponds to MixUCB with $\Delta \to \infty$, and two Linear Oracles, which select actions according to the best linear model in hindsight. The Oracles represent the performance of (unrealistically) having access to all information about the contexts and rewards ahead of time. The Linear Classification Oracle computes the best linear classifier $\hat{\theta}_{lr}$ for action selection using the (multiclass) logistic loss. The Linear Regression Oracle computes the best linear predictor $\hat{\theta}_{lr}$ of rewards using the squared loss. The algorithms and baselines are summarized in Table 1.

Online Regression and Confidence Sets For all methods and datasets, we define $\phi(x, a) = x \otimes e_a$ where e_a is a standard basis vector, so that $d = d_x K$ and we can write $\theta = (\theta_1, \dots, \theta_K)$. For computational simplicity, we define a joint estimate and confidence set which directly combines the squared and logistic losses on the datasets $\mathcal{D}_t^{sq}$ and $\mathcal{D}_t^{lr}$ respectively. This formulation results in a single estimate $\hat{\theta}_t$ and an ellipsoidal confidence set. The advantage of this joint formulation is that the optimistic/pessimistic optimization has a closed form solution. Further details are provided in the appendix.

Evaluation Metrics We report *Cumulative Reward* and *Average Autonomous Reward*. Cumulative reward measures the actual rewards accumulated over time (thus mixing autonomous and expert actions), while average autonomous reward is the reward averaged over the time steps in which the algorithm didn't query. Additionally, we evaluate the cost of MixUCB with *Cumulative Queries*.

5.1 Synthetic Experiments

For synthetic data, we set $d_x = 2$ and fix a true parameter $\theta_a^* \sim \mathcal{N}(0, I)$ for $a = 1, 2, 3$ and define $f^*(x_t, a) = \langle \theta_a^*, x_t \rangle$. The observed reward is $r(x_t, a) = f^*(x_t, a) + \mathcal{N}(0, \sigma^2)$. For Type I and II feedback, the expert selects an action according to (2) with rationality $\alpha = 1$. For Type III feedback, the expert reveals $f^*(x_t, a)$ for $a = 1, ..., K$. We sample $x_t \sim \mathcal{N}(0, I)$ at each time step. We present results for a variety of query thresholds Δ in the appendix.

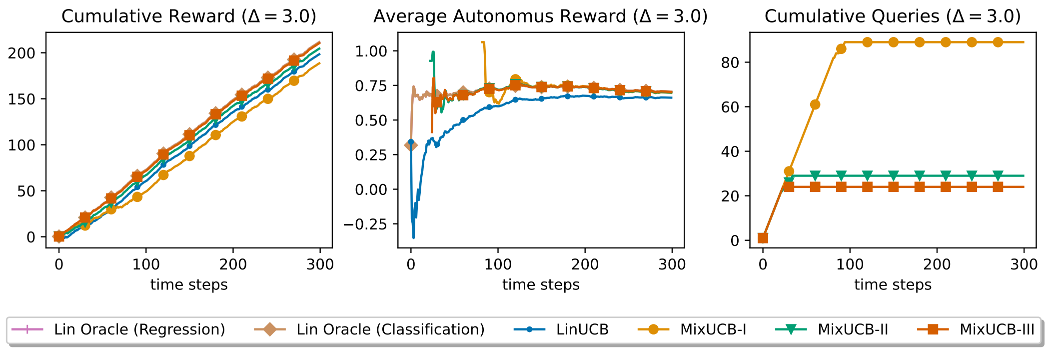


Figure 2: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB (with query threshold $\Delta = 3.0$) on synthetic data.

As shown in Figure 2, MixUCB-III achieves a cumulative reward comparable to that of the linear oracles, while MixUCB-II attains a slightly lower cumulative reward. Both outperform LinUCB in terms of cumulative reward, whereas MixUCB-I performs worse than LinUCB. However, despite the limited initial information, MixUCB-I eventually achieves autonomous rewards similar to MixUCB-II and III and the linear oracles. This indicates that the poor overall performance of MixUCB-I arises from the fact that the noisy expert takes suboptimal actions. Also notice that, unlike LinUCB, the MixUCB algorithms never attain very low or negative autonomous reward, highlighting the safety guarantees. The cumulative queries plot further illustrates the efficiency of the MixUCB variants: MixUCB-I stops querying after approximately 100 time steps, whereas MixUCB-II and MixUCB-III cease querying in fewer than 30 steps. So, all the MixUCB variants efficiently reduce their dependence on queries while achieving strong performance. This demonstrates that all MixUCB variants effectively balance expert feedback with autonomous learning, reducing reliance on queries while maintaining strong performance. Additionally, MixUCB-II and MixUCB-III leverage expert feedback more efficiently, quickly transitioning to autonomous decision-making.

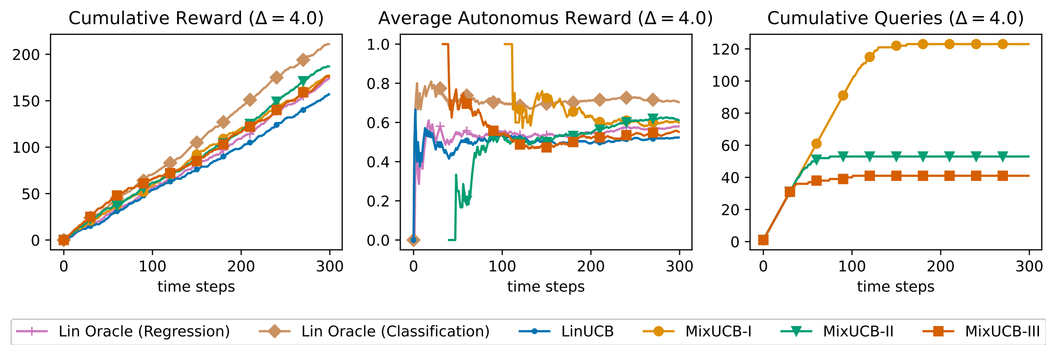
5.2 Real Data Experiments

Full details on data preprocessing are in the appendix.

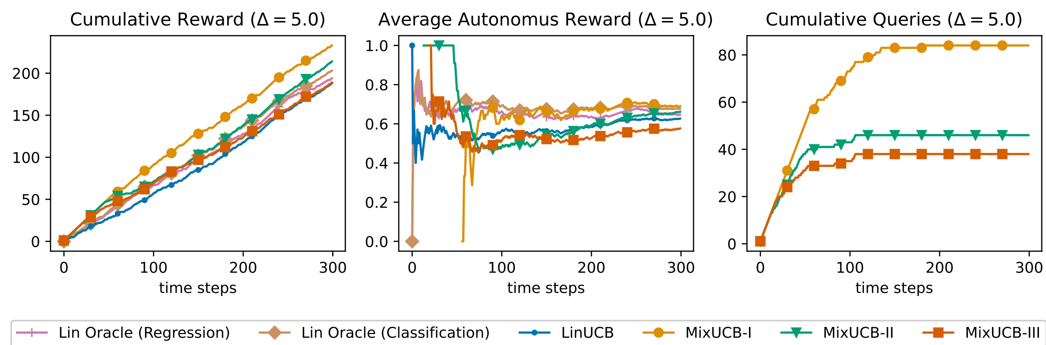
Robot Manipulation We consider a robot-assisted bite acquisition setting where contexts are pieces of food, $K = 6$ actions are different orientations of the end-effector, and rewards are successful acquisition. We use a dataset from Feng et al. (2019) which contains images of food and success rates of the actions. We perform PCA on the embeddings of the images to define contexts with $d_x = 5$. We define $f^*(x_t, a)$ as the success rate and sample the observed reward $r(x_t, a)$ from a Bernoulli distribution. We define expert feedback using $f^*(x_t, a)$ as in the synthetic setting.

Medical Classification Datasets We define additional settings using medical classification datasets: heart disease (Bou Rjeily et al., 2019) and MedNIST (Yang et al., 2023). We use PCA on the features to define contexts with $d_x = 6$, define each class as an action ($K = 2$ and 6 respectively), and define the observed reward $r(x_t, a)$ as 1 when a is the correct classification and 0 otherwise. Since we do not have access to the expected reward $f^*(x_t, a)$, we define expert feedback based on the observed rewards for Type-III, and give the true class label for Types-I and II.

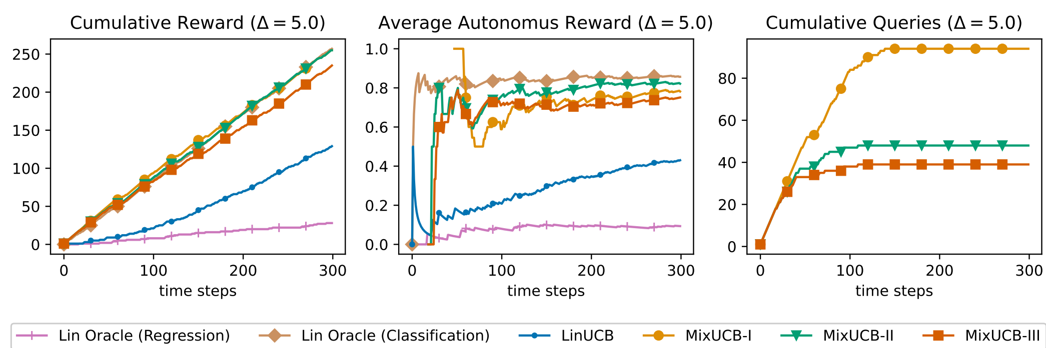
Results We present the results for all the three real world dataset (Robot Manipulation, Heart Disease, and MedNIST) in Figure 3. Unlike in the synthetic setting, the rewards are not necessarily linearly realizable. This is illustrated by the performance of the Linear Oracles: the regression oracle (which attempts to predict rewards) performs poorly compared with the classification oracle (which need only distinguish between actions). As a result, methods that rely most heavily on



(a) Robot Manipulation Dataset (Feng et al., 2019).



(b) Heart Disease Dataset (Bou Rjeily et al., 2019).



(c) MedNIST Dataset (Yang et al., 2023).

Figure 3: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB on Robot Manipulation Dataset (3a), Heart Disease Dataset (3b), and MedNIST Dataset (3c) using $\Delta = 4.0$, $\Delta = 5.0$, and $\Delta = 5.0$ respectively.

linear regression (LinUCB and the Linear Regression Oracle) do not perform well. On the other hand, methods that follow the experts advice and learn from classification feedback (MixUCB and the Linear Classification Oracle) perform better. In the MedNIST dataset, the realizability issue is particularly pronounced: the Linear Regression Oracle attains 10% performance of the Linear Classification Oracle. The violation of the linear realizability assumption is worse for algorithms that rely on linear regression, like LinUCB and MixUCB-III. The effect on total reward is mitigated for MixUCB-III because of the high rewards from expert actions.

MixUCB-I and II fair better in the real data settings due to 1) learning from classification style feedback and 2) gaining high rewards from expert actions. This second point is particularly pronounced for the classification datasets, where we do not directly simulate the noisiness of the expert—as a result, for the heart disease data, MixUCB-I outperforms the Linear Classification Oracle in terms of total reward. However, MixUCB-I and II are still able to perform well even with noisy expert advice in the robot manipulation setting.

Among the three MixUCB variants, MixUCB-I queries the most frequently, while MixUCB-III queries the least, with MixUCB-II falling in between. This aligns with expectations—MixUCB-III gains more information per query, while MixUCB-I obtains the least. In all cases, the algorithm queries the most in the beginning, but then slowly stops querying. Finally, we observe that when MixUCB stops querying, there is a brief period of performance fluctuation before stabilization. This can be attributed to the sudden shift from relying on expert feedback to autonomous decision-making. However, within 100 steps, the model effectively adapts, demonstrating its ability to generalize from the acquired knowledge.

6 Conclusion

In this paper, we propose MixUCB, a flexible human-in-the-loop contextual bandit framework that enhances safe exploration by integrating human expertise with machine automation. Our results demonstrate that human and AI can complement each other to enable safer and more effective decision-making. Our experiments highlight the effectiveness of MixUCB in balancing query efficiency and reward maximization. Compared with LinUCB, MixUCB consistently achieves a favorable trade-off, efficiently navigating between querying experts and autonomous decision-making.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pp. 1220–1228. PMLR, 2013.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Carine Bou Rjeily, Georges Badr, Amir Hajjarm El Hassani, and Emmanuel Andres. Medical data mining for heart diseases and the future of sequential mining in medical field. *Machine learning paradigms: Advances in data analytics*, pp. 71–99, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pp. 355–366, 2008.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple experts. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Ryan Feng, Youngsun Kim, Gilwoo Lee, Ethan K Gordon, Matt Schmittle, Shivaum Kumar, Tapomayukh Bhattacharjee, and Siddhartha S Srinivasa. Robot-assisted feeding: Generalizing skewering strategies across food items on a plate. In *The International Symposium of Robotics Research*, pp. 427–442. Springer, 2019.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 1539–1548. PMLR, 2018a.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pp. 167–208. PMLR, 2018b.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(1): 3487–3602, 2015.

Steve Hanneke and Liu Yang. Toward a general theory of online selective sampling: Trading off mistakes and queries. In *International Conference on Artificial Intelligence and Statistics*, pp. 3997–4005. PMLR, 2021.

Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*, 2021.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20(65): 1–50, 2019.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959.

R Duncan Luce. The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3): 215–233, 1977.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087. PMLR, 2020.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35:29292–29304, 2022.

Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pp. 1232–1264. PMLR, 2014.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, 36, 2024a.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression. *Advances in Neural Information Processing Systems*, 36, 2024b.

Daniel Sikar, Artur Garcez, Tillman Weyde, Robin Bloomfield, and Kaleem Peeroo. When to accept automated predictions and when to defer to human judgment? *arXiv preprint arXiv:2407.07821*, 2024.

Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *International conference on machine learning*, pp. 3309–3318. PMLR, 2017.

Chih-Chun Wang, Sanjeev R Kulkarni, and H Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. *Advances in Neural Information Processing Systems*, 35:35379–35391, 2022.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. 2010.

Supplementary Materials

The following content was not necessarily subject to peer review.

7 Main Proofs

Proof of Lemma 2. Let $R_\infty = \max_{x\in\mathcal{X},a\in\mathcal{A}} f^*(x,a)$,

$$\begin{aligned} c(\alpha) &= \max_{x\in\mathcal{X}} \left(\max_{a\in\mathcal{A}} f^*(x,a) \right) - \mathbb{E}_a[f^*(x,a)] \\ &\le \max_{x\in\mathcal{X}} R_\infty - \sum_{a\in\mathcal{A}} \frac{\exp(\alpha f^*(x,a))}{\sum_{a'\in\mathcal{A}} \exp(\alpha f^*(x,a'))} f^*(x,a) \\ &\le R_\infty - \min_{\|\vec{r}\|_\infty = R_\infty} \frac{\langle \exp(\alpha\vec{r}), \vec{r}\rangle}{\langle \exp(\alpha\vec{r}), 1\rangle} \\ &= R_\infty - \frac{R_\infty \exp(\alpha R_\infty)}{\exp(\alpha R_\infty) + K - 1} \end{aligned}$$

where the final equality holds when $\vec{r}$ has one element being R_∞ while the rest being 0. (For example, when $\vec{r} = [R_\infty, 0, \cdots, 0]$). Such $\vec{r}$ attain the minimum, as the element-wise derivative of $\frac{\langle \exp(\alpha\vec{r}), \vec{r}\rangle}{\langle \exp(\alpha\vec{r}), 1\rangle}$ is increasing. The final expression holds by simplifying the difference of fractions.

□

Proof of Proposition 1. Let $\mathcal{E}_t^{sq} = \{\theta \in \mathbb{R}^d, \|\theta\| \le 1 : \|\theta - \theta_{t-1}^{sq}\|^2_{V_{t-1}^{sq}} \le \beta_t^{sq}\}$ and $\mathcal{E}_t^{lr} = \{\theta \in \mathbb{R}^d, \|\theta\| \le 1 : \|\theta - \theta_{t-1}^{lr}\|^2_{V_{t-1}^{lr}} \le \beta_t^{lr}\}$ be the confidence set from square loss oracle and logistic regression oracle, respectively, and let $\mathcal{E}_t = \mathcal{E}_t^{lr} \cap \mathcal{E}_t^{sq}$ be the confidence set that contains the true parameter θ^*. Recall from Algorithm 1 that the UCB action $a_t^{ucb} = \arg\max_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a)$ and the confidence width of a_t^{ucb} is $w_t = \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) - \min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$.

Case 1. The algorithm is not confident about its predicted action, i.e., $w_t \ge \Delta$, which satisfies the query condition. In this case, the algorithm takes action from noisy expert $\tilde{a}_t^*$, and incurs regrets $R_t^{ExP}(\tilde{a}_t^*)$, which is controlled by how noisy the expert is.

Case 2. the algorithm is confidence about its predicted action a_t^{ucb}, i.e, $w_t < \Delta$, so it will play the UCB action a_t^{ucb}. Let a_t^* be the optimal action at round t, i.e., $a_t^* = \arg\max_{a\in\mathcal{A}} \langle \theta^*, \phi(x_t, a)\rangle$, the regret of playing this action is bounded as

$$\begin{aligned} R_t^{NoE} &= \langle \theta^*, \phi(x_t, a_t^*)\rangle - \langle \theta^*, \phi(x_t, a_t^{ucb})\rangle \\ &\le \max_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a) - \min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) \\ &= \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) - \min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) \\ &= w_t < \Delta \end{aligned} \tag{11}$$

On the other hand, let $\bar{\theta}_t = \arg\max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$ and $\underline{\theta}_t = \arg\min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$, it holds that

$$\begin{aligned} R_t^{NoE} &\le \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) - \min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) \\ &= \bar{\theta}_t^\top \phi(x_t, a_t^{ucb}) - \underline{\theta}_t^\top \phi(x_t, a_t^{ucb}) \\ &= \langle \bar{\theta}_t - \underline{\theta}_t, \phi(x_t, a_t^{ucb})\rangle \\ &\le \|\bar{\theta}_t - \underline{\theta}_t\|_{V_{t-1}^{sq}} \cdot \|\phi(x_t, a_t^{ucb})\|_{(V_{t-1}^{sq})^{-1}} \\ &\le 2\sqrt{\beta_t^{sq}} \cdot \|\phi(x_t, a_t^{ucb})\|_{(V_{t-1}^{sq})^{-1}} \end{aligned} \tag{12}$$

Putting them together, we have

$$R_t^{NoE} \leq \min\{\Delta, 2\sqrt{\beta_t^{sq}} \cdot \|\phi(x_t, a_t^{ucb})\|_{(V_{t-1}^{sq})^{-1}}\} \tag{13}$$

From assumption 2, we have that $\beta_T^{sq} \geq \max\{1, \beta_t^{sq}\}$, and thus

$$R_t^{NoE} \leq 2\sqrt{\beta_T^{sq}} \min\{\Delta, \|\phi(x_t, a_t^{ucb})\|_{(V_{t-1}^{sq})^{-1}}\} \tag{14}$$

and

$$\begin{aligned}(R_t^{NoE})^2 &\leq 4\beta_T^{sq} \min\{\Delta^2, \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}}\} \\ &\leq 4\beta_T^{sq} \cdot \frac{\Delta^2}{\log_2(1+\Delta^2)} \cdot \log_2(1 + \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}})\end{aligned} \tag{15}$$

where we used the fact that for any $\Delta < 1$ and $u \geq 0$, $\min\{\Delta^2, u\} \leq \log_v(1+u) = \frac{\log_2(1+u)}{\log_2 v}$ with $\log_2 v = \frac{\log_2(1+\Delta^2)}{\Delta^2}$.

Now, we will bound the sum over $\log_2(1 + \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}})$:

For any $t \geq 1$, we have

$$\begin{aligned}V_t^{sq} =& V_{t-1}^{sq} + \bar{Z}_t \cdot \phi(x_t, a_t^{ucb})\phi(x_t, a_t^{ucb})^\top \\ =& (V_{t-1}^{sq})^{1/2}(I + \bar{Z}_t(V_{t-1}^{sq})^{-1/2}\phi(x_t, a_t^{ucb})\phi(x_t, a_t^{ucb})^\top(V_{t-1}^{sq})^{-1/2})(V_{t-1}^{sq})^{1/2}\end{aligned} \tag{16}$$

and thus

$$\begin{aligned}\det(V_t^{sq}) &= \det(V_{t-1}^{sq})\det(I + \bar{Z}_t(V_{t-1}^{sq})^{-1/2}\phi(x_t, a_t^{ucb})\phi(x_t, a_t^{ucb})^\top(V_{t-1}^{sq})^{-1/2}) \\ &= \det(V_{t-1}^{sq}) \cdot \left(1 + \bar{Z}_t\|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}}\right)\end{aligned} \tag{17}$$

where it follows because matrix $I + yy^\top$ has eigenvalues $1 + \|y\|_2^2$ and 1, as well as the fact that the determinant of a matrix is the product of its eigenvalues.

$$\begin{aligned}&\sum_{t=1}^T \bar{Z}_t \cdot \log_2(1 + \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}}) \\ =& \sum_{t=1}^T \log_2(1 + \bar{Z}_t\|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}}) \\ =& \sum_{t=1}^T \log \frac{\det(V_t^{sq})}{\det(V_{t-1}^{sq})} \\ =& \log \frac{\det(V_T^{sq})}{\det(V_0^{sq})} \\ \leq& \log \frac{\Pi_{i=1}^d \lambda_i^{sq}}{\det(V_0^{sq})} \\ \leq& \log \frac{(\frac{1}{d}Tr(V_T^{sq}))^d}{\det(V_0^{sq})} \\ \leq& \log \frac{(\frac{1}{d}(d\lambda + \sum_{t=1}^T \bar{Z}_t L^2))^d}{\lambda^d} \\ \leq& d\log(1 + \frac{(\sum_{t=1}^T \bar{Z}_t)L^2}{\lambda d})\end{aligned} \tag{18}$$

where $\lambda_1^{sq}, \cdots, \lambda_d^{sq}$ are the eigenvalues of V_T^{sq}.

The total regret on the rounds that we don't query is

$$
\begin{aligned}
\sum_{t=1}^{T} R_t^{NoE} \bar{Z}_t &\leq \sqrt{\left(\sum_{t=1}^{T} \bar{Z}_t\right) \cdot \left(\sum_{t=1}^{T} \bar{Z}_t \cdot (R_t^{NoE})^2\right)} \\
&\leq \sqrt{\left(\sum_{t=1}^{T} \bar{Z}_t\right) \cdot \left(\sum_{t=1}^{T} \bar{Z}_t \cdot 4\beta_T^{sq} \cdot \frac{\Delta^2}{\log_2(1+\Delta^2)} \cdot \log_2(1 + \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}})\right)} \\
&= \frac{2\Delta\beta_T^{sq}}{\sqrt{\log_2(1+\Delta^2)}} \sqrt{\sum_{t=1}^{T} \bar{Z}_t} \sqrt{\sum_{t=1}^{T} \bar{Z}_t \log_2(1 + \|\phi(x_t, a_t^{ucb})\|^2_{(V_{t-1}^{sq})^{-1}})} \\
&\leq \frac{2\Delta\beta_T^{sq}}{\sqrt{\log_2(1+\Delta^2)}} \sqrt{\sum_{t=1}^{T} \bar{Z}_t} \sqrt{d \log_2(1 + \frac{(\sum_{t=1}^{T} \bar{Z}_t) L^2}{\lambda d})}
\end{aligned}
\tag{19}
$$

Putting case 1 and case 2 together, we have the overall regret

$$
\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \bar{Z}_t R_t^{NoE} + \sum_{t=1}^{T} Z_t R_t^{ExP} \\
&\leq \frac{2\Delta\beta_T^{sq}}{\sqrt{\log_2(1+\Delta^2)}} \sqrt{\sum_{t=1}^{T} \bar{Z}_t} \sqrt{d \log_2(1 + \frac{(\sum_{t=1}^{T} \bar{Z}_t) L^2}{\lambda d})} + \sum_{t=1}^{T} Z_t R_t^{ExP}(\tilde{a}_t^*)
\end{aligned}
\tag{20}
$$

□

Proof of Theorem 1. let $\bar{\theta}_t = \arg\max_{\theta \in \mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$, $\underline{\theta}_t = \arg\min_{\theta \in \mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$, and $a_t^* = \arg\max_{a \in \mathcal{A}} \langle \theta^*, \phi(x_t, a) \rangle$. Recall that $w_t = \max_{\theta \in \mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) - \min_{\theta \in \mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$

$$
\begin{aligned}
\sum_{t=1}^{T} Z_t &= \sum_{t=1}^{T} \mathbb{1}\{w_t \geq \Delta\} \\
&= \sum_{t=1}^{T} \mathbb{1}\{\langle \bar{\theta}_t - \underline{\theta}_t, \phi(x_t, a_t^{ucb}) \rangle \geq \Delta\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\{\langle \bar{\theta}_t - \theta^*, \phi(x_t, a_t^{ucb}) \rangle \geq \frac{\Delta}{2}\} + \sum_{t=1}^{T} \mathbb{1}\{\langle \theta^* - \underline{\theta}_t, \phi(x_t, a_t^{ucb}) \rangle \geq \frac{\Delta}{2}\}
\end{aligned}
\tag{21}
$$

Using Lemma 7 from Sekhari et al. (2024b), we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{1}\{\langle \bar{\theta}_t - \theta^*, \phi(x_t, a_t^{ucb}) \rangle \geq \frac{\Delta}{2}\} \\
=&\sum_{t=1}^{T} Z_t \mathbb{1}\{\langle \bar{\theta}_t - \theta^*, \phi(x_t, a_t^{ucb}) \rangle \geq \frac{\Delta}{2}\} + \sum_{t=1}^{T} \bar{Z}_t \mathbb{1}\{\langle \bar{\theta}_t - \theta^*, \phi(x_t, a_t^{ucb}) \rangle \geq \frac{\Delta}{2}\} \\
\leq& \left(\frac{4\beta_T^{sq}}{\Delta^2} + 1\right) d + \left(\frac{4\beta_T^{lr}}{\Delta^2} + 1\right) d \\
\leq& \frac{10\beta_T d}{\Delta^2}
\end{aligned}
\tag{22}
$$

□

Algorithm 2 MixUCB-I (Detailed)

Input: Query threshold Δ, total rounds T.
Let $V_0^{sq} = V_0^{lr} = \lambda I$, the initial confidence set $\mathcal{E}_1^{sq} = \mathcal{E}_1^{lr} = \{\theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$
for $t = 1, \cdots, T$ **do**
  Construct the current parameter space $\mathcal{E}_t = \mathcal{E}_t^{sq} \cap \mathcal{E}_t^{lr}$
  Learner predict the UCB action $a_t^{ucb} = \arg\max_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a)$
  Compute the confidence of a_t^{ucb}: $w_t = \max_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb}) - \min_{\theta\in\mathcal{E}_t} \theta^\top \phi(x_t, a_t^{ucb})$
  if $w_t \geq \Delta$ **then**
    Query the expert to get the noisy optimal action $\tilde{a}_t^*$ and play $\tilde{a}_t^*$, and $Z_t = 1$.
  else
    Play the UCB action a_t^{ucb} and observe the reward r_t and $Z_t = 0$.
  Update $V_t^{sq} = V_{t-1}^{sq} + \bar{Z}_t \cdot x^t_{a_t^{ucb}} (x^t_{a_t^{ucb}})^\top$ and $V_t^{lr} = V_{t-1}^{lr} + Z_t \cdot \sum_{a\in\mathcal{A}} x_a^t (x_a^t)^\top$, where $x_a^t = \phi(x_t, a)$
  \\ Update the square loss oracle and its confidence set
  Update the square loss parameter estimation $\theta_t^{sq} = (V_t^{sq})^{-1} (\sum_{s=1}^{t-1} x^s_{a_s^{ucb}} r_s + \bar{Z}_t \cdot x^t_{a_t^{ucb}} r_t)$ and confidence set $\mathcal{E}_{t+1}^{sq} = \{\theta \in \mathbb{R}^d, \|\theta\| \leq 1 : \|\theta - \theta_t^{sq}\|^2_{V_t^{sq}} \leq \beta_t^{sq}\}$
  \\ Update the logistic loss oracle and its confidence set
  Update logistic regression oracle and get the new parameter estimation $\theta_t^{lr} = \mathcal{O}_{\theta_{t-1}^{lr}}(\{x_t, \tilde{a}_t^*\}) Z_t + \mathcal{O}_{\theta_{t-1}^{lr}}(\emptyset) \bar{Z}_t$, then update the confidence set $\mathcal{E}_{t+1}^{lr} = \{\theta \in \mathbb{R}^d, \|\theta\| \leq 1 : \|\theta - \theta_t^{lr}\|^2_{V_t^{lr}} \leq \beta_t^{lr}\}$
Return

8 Detailed Algorithms

Let $x_a^t = \phi(x, a)$ be the feature vector of action a at step t.

Algorithm 3 MixUCB (*Type-III* feedback)

Input: Query threshold Δ, total rounds T, function class $\mathcal{F}$, initial confidence set $\mathcal{E}_1$
for $t = 1, \cdots, T$ **do**
  $a_t^{ucb} = \arg\max_{a\in\mathcal{A}} \max_{f\in\mathcal{E}_t} f(x_t, a)$
  $w_t = \max_{f\in\mathcal{E}_t} f(x_t, a_t^{ucb}) - \min_{f\in\mathcal{E}_t} f(x_t, a_t^{ucb})$
  if $w_t \geq \Delta$ **then**
    Set $Z_t = 1$. Query the experts and observe the rewards for all the actions $r_{t,a} \sim r(x_t, a), \forall a \in \mathcal{A}$ and play optimal action $a_t^* = \arg\max_{a\in\mathcal{A}} r(x_t, a)$.
    Update $\mathcal{D}_t$ and $\mathcal{E}_t$ with $(x_t, a, r_{t,a})$ according to

$$\mathcal{E}_t = \{f \in \mathcal{F} \mid \sum_{x,a\in\mathcal{D}_t} (f_t(x_t, a_t) - f(x_t, a_t))^2 \leq \beta_t\} . \tag{23}$$

    .
  else
    Set $Z_t = 0$. Play a_t^{ucb} and observe $r_t \sim r(x_t, a_t^{ucb})$.
    Update $\mathcal{D}_t$ and $\mathcal{E}_t$ with (x_t, a_t^{ucb}, r_t) according to (23).
Return

9 Experimental details

Online regression and confidence sets The joint loss is defined as

$$\sum_{x,a\in\mathcal{D}_t^{lr}} \ell_{lr}(\theta, x, a) + \sum_{x,a,r\in\mathcal{D}_t^{sq}} \ell_{sq}(\theta, (x, a), r) + \lambda\|\theta\|_2^2$$

where ℓ_{lr} is the cross entropy loss and ℓ_{sq} is the squared loss. Then we define $\hat{\theta}_t$ for all algorithms as the minimizer of this loss and the confidence set as $\mathcal{E}_t = \{\theta \mid \|\theta - \hat{\theta}_t\|^2_{V_t(\beta)} \leq 1\}$ where

$$V_t(\beta) = \frac{1}{(\beta^{lr})^2} \sum_{x,a \in \mathcal{D}_t^{lr}} \phi(x,a)\phi(x,a)^\top + \frac{1}{(\beta^{sq})^2} \sum_{x,a \in \mathcal{D}_t^{sq}} \phi(x,a)\phi(x,a)^\top + \frac{1}{(\beta^{sq})^2}\lambda I \,.$$

The advantage of this joint definition is that the optimistic/pessimistic optimization has a closed form solution: $\max_{f \in \mathcal{E}_t} f(x,a) = \hat{\theta}_t^\top \phi(x,a) + \|x\|_{V_t(\beta)}$.

Robotics dataset We consider a dataset from the challenging robot manipulation problem of robot-assisted bite acquisition (Feng et al., 2019), in which the task of the robotic agent is to acquire bite-sized food items. The dataset include images from 16 different food types. In this setting, the raw observation space $\mathcal{O}$ consists of RGB images of the bite-sized food items. We derive a context space $\mathcal{X} \subset \mathbb{R}^5$ by first extracting a lower-dimensional intermediate context $x_{int} \in \mathbb{R}^{2048}$ by passing the each image through the SPANet network (a supervised network developed in (Feng et al., 2019) for this domain) and extracting the penultimate layer (which is a linear layer). We then run PCA with $n = 5$ components to get the final context $x \in \mathbb{R}^5$. The action space $\mathcal{A}$ consists of 6 discrete actions, corresponding to different orientations of the robot end-effector. The rewards $r \in \mathbb{R}$ represent the probability of a successful acquisition.

Medical datasets In this study, we utilize a heart disease dataset sourced from the UCI Machine Learning Repository, which is publicly available (Bou Rjeily et al., 2019). The dataset comprises 297 instances and 14 attributes. These attributes include age, sex, cholesterol levels, chest pain type (e.g., typical or non-anginal), resting blood pressure, maximum heart rate, and results from tests like resting ECG and Thallium stress tests. Additional variables such as exercise-induced angina and ST depression assess heart performance under stress. The dataset also includes attributes like the number of major vessels and fasting blood sugar. The target variable, 'Diagnosis,' indicates whether a patient has heart disease (1 = yes, 0 = no), and serves as the dependent variable, while the remaining 13 attributes act as independent variables. No personally identifiable information is included. We derive a context space $x \in \mathbb{R}^6$ by running PCA with $n = 6$ components from the original context $x_{int} \in \mathbb{R}^{13}$. The action space $\mathcal{A}$ consists of 2 discrete actions.

MedNIST (Yang et al., 2023) consists of 28×28 images with corresponding classification labels. We randomly select 20 samples from each of the 6 classes: 'BreastMRI', 'HeadCT', 'CXR', 'ChestCT', 'Hand', and 'AbdomenCT'. We derive a context space $x \in \mathbb{R}^6$ by running PCA with $n = 6$ components. The action space $\mathcal{A}$ consists of 6 discrete actions.

10 Complete experimental results

In Figure 4, Figure 5, Figure 6 and Figure 7, we show the complete result of different query threshold Δ for synthetic data, robot manipulation dataset, MedNIST dataset and Heart Disease dataset, respectively.

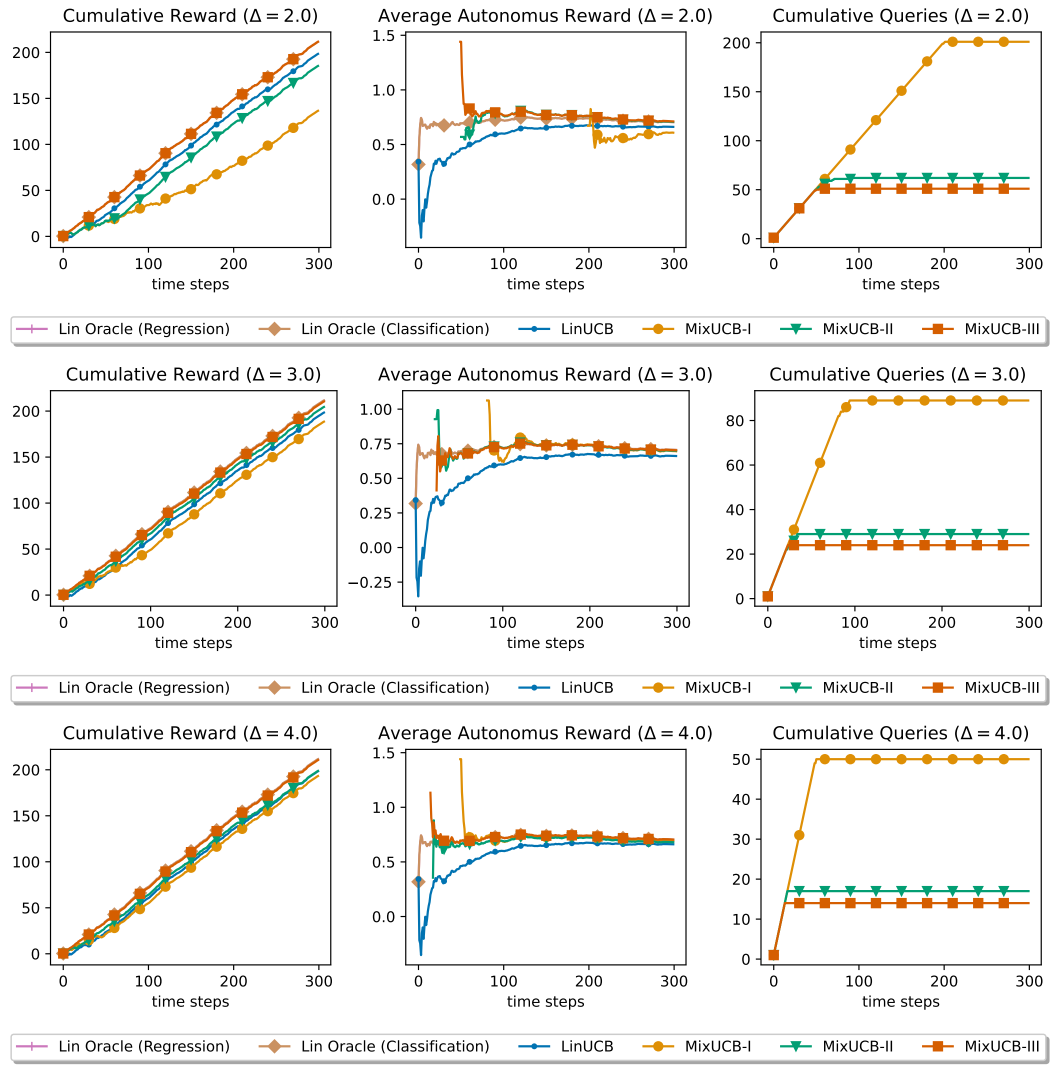


Figure 4: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB on synthetic data with different query threshold $\Delta = \{2, 3, 4\}$.

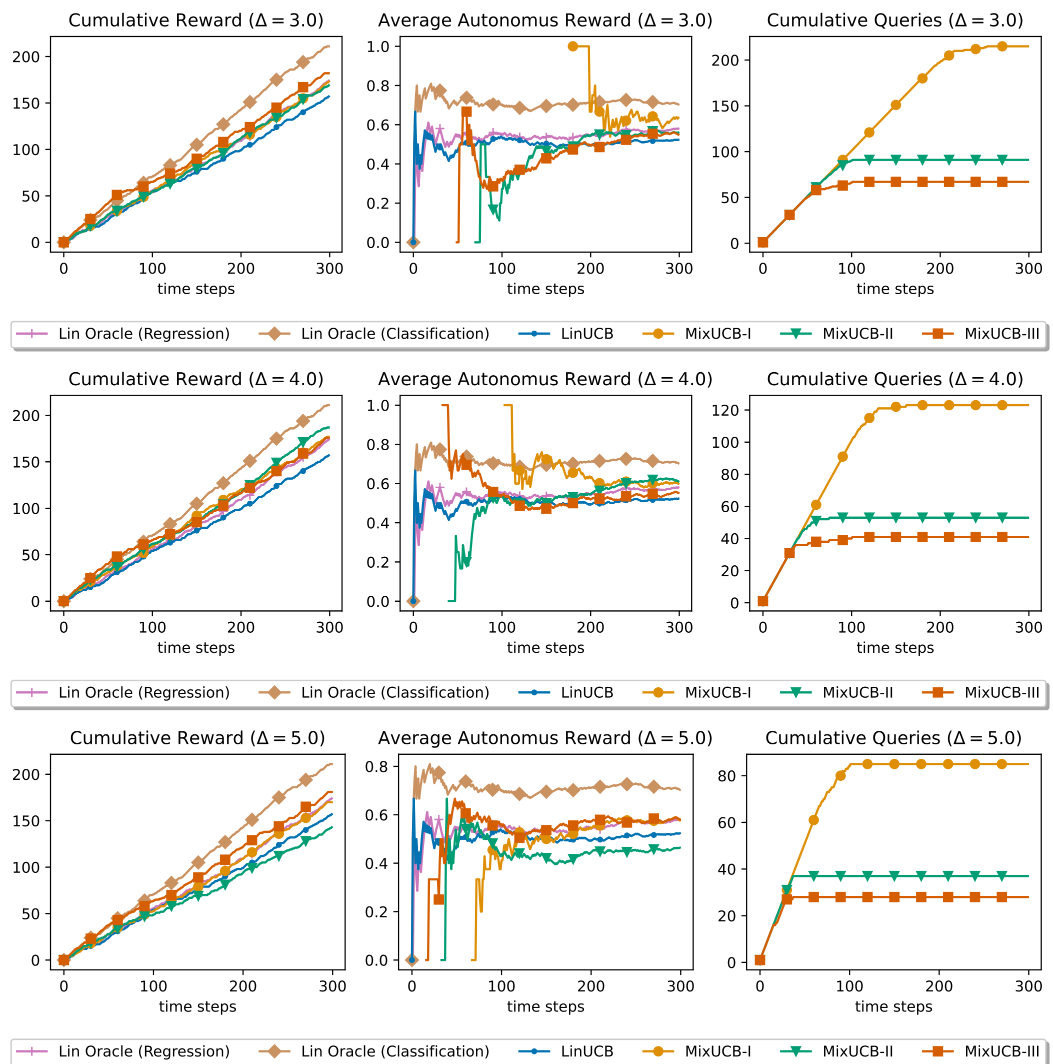


Figure 5: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB on Robot manipulation dataset with different query threshold $\Delta = \{3, 4, 5\}$.

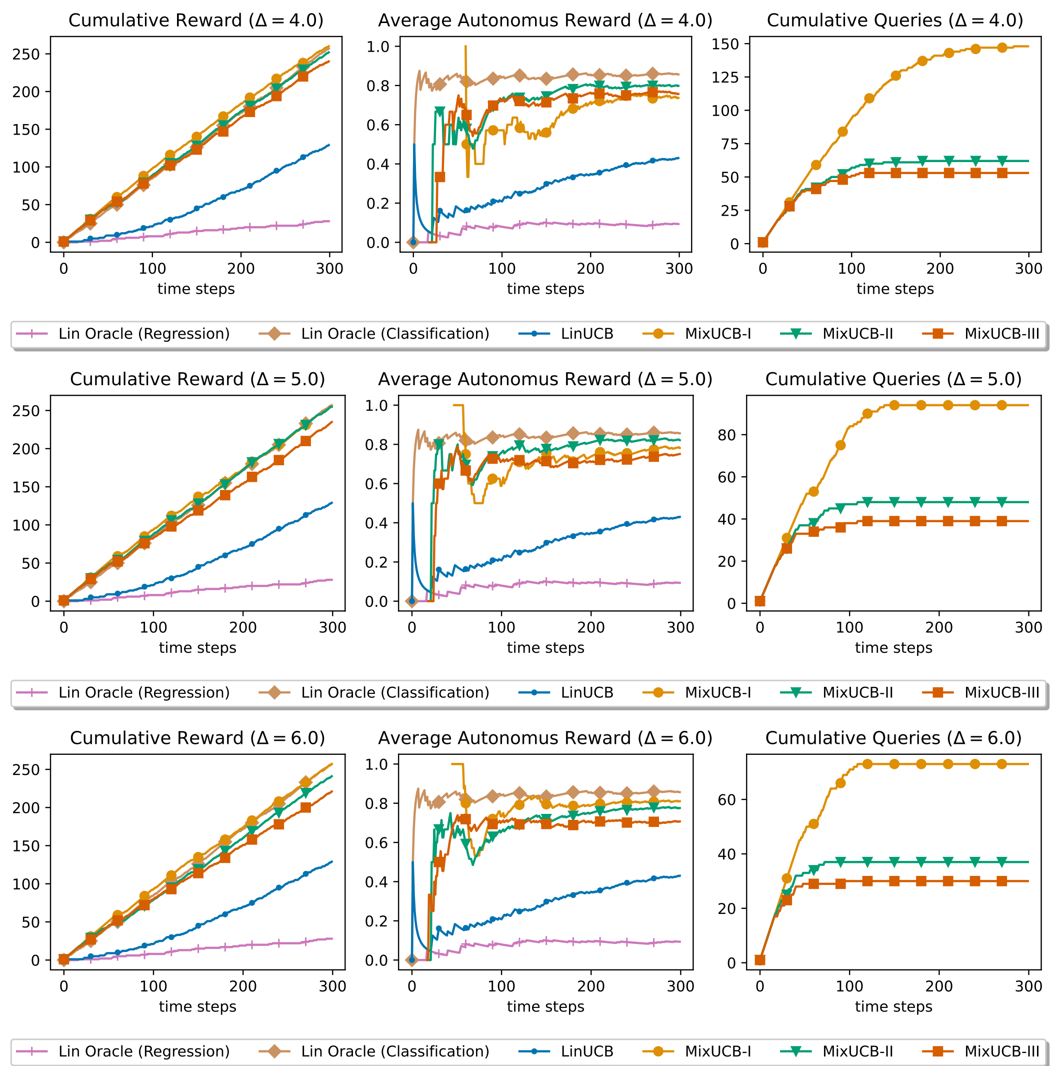


Figure 6: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB on MedNIST dataset with different query threshold $\Delta = \{4, 5, 6\}$.

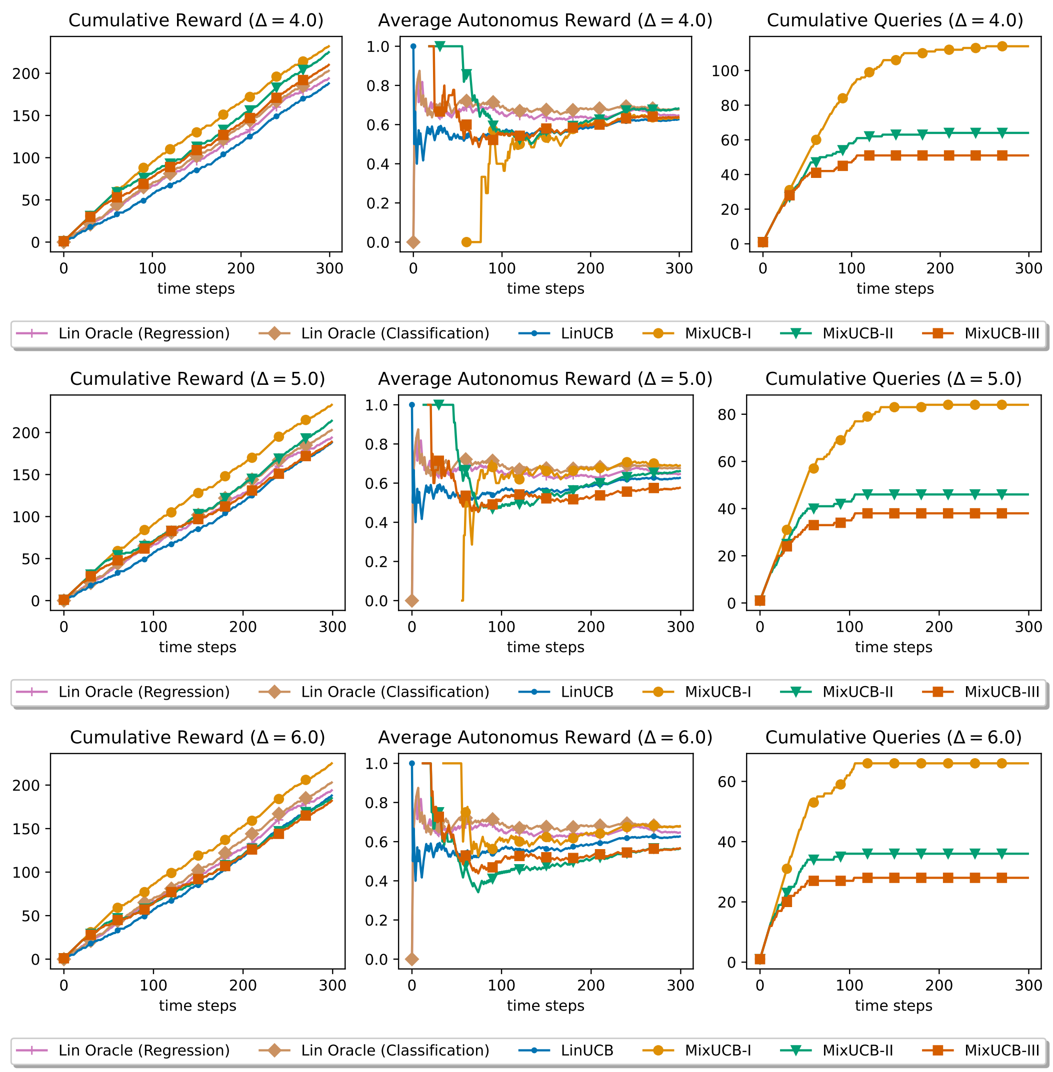


Figure 7: Cumulative Reward, Average Autonomous Reward, and Cumulative Queries for MixUCB on Heart disease dataset with different query threshold $\Delta = \{4, 5, 6\}$.